ALLDATA 2015

The First International Conference on Big Data, Small Data, Linked Data and Open Data

ISBN: 978-1-61208-445-9

KESA 2015

The International Workshop on Knowledge Extraction and Semantic Annotation

April 19 - 24, 2015

Barcelona, Spain

ALLDATA 2015 Editors

Jerzy Grzymala-Busse, University of Kansas, USA
Ingo Schwab, University of Applied Sciences Karlsruhe, Germany
Maria Pia di Buono, University of Salerno, Italy

ALLDATA 2015

Foreword

The First International Conference on Big Data, Small Data, Linked Data, and Open Data (ALLDATA 2015), held between April 19th-24th, 2015 in Barcelona, Spain, is an inaugural event bridging the concepts and the communities devoted to each of data categories for a better understanding of data semantics and their use, by taking advantage from the development of Semantic Web, Deep Web, Internet, non-SQL and SQL structures, progresses in data processing, and the new tendency for acceptance of open environments.

The volume and the complexity of available information overwhelm human and computing resources. Several approaches, technologies and tools are dealing with different types of data when searching, mining, learning and managing existing and increasingly growing information. From understanding Small data, the academia and industry recently embraced Big data, Linked data, and Open data. Each of these concepts carries specific foundations, algorithms and techniques, and is suitable and successful for different kinds of application. While approaching each concept from a silo point of view allows a better understanding (and potential optimization), no application or service can be developed without considering all data types mentioned above.

ALLDATA 2015 also featured the following Symposium:

- KESA 2015, The International Workshop on Knowledge Extraction and Semantic Annotation

We take here the opportunity to warmly thank all the members of the ALLDATA 2015 Technical Program Committee, as well as the numerous reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to ALLDATA 2015. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the ALLDATA 2015 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that ALLDATA 2015 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the domain of data management.

We also hope Barcelona provided a pleasant environment during the conference and everyone saved some time for exploring this beautiful city.

ALLDATA 2015 Advisory Committee:

Mark Balas, Embry-Riddle Aeronautical University in Daytona Beach, USA
Yeh-Ching Chung, National Tsing Hua University, Taiwan
Andreas Schmidt, University of Applied Sciences Karlsruhe | Karlsruhe Institute of Technology, Germany
Fabrice Mourlin, Université Paris-Est Créteil Val de Marne, France
Dan Tamir, Texas State University, USA

ALLDATA 2015

Committee

ALLDATA 2015 Chairs

Mark Balas, Embry-Riddle Aeronautical University in Daytona Beach, USA
Yeh-Ching Chung, National Tsing Hua University, Taiwan
Andreas Schmidt, University of Applied Sciences Karlsruhe | Karlsruhe Institute of Technology, Germany
Fabrice Mourlin, Université Paris-Est Créteil Val de Marne, France
Dan Tamir, Texas State University, USA

ALLDATA 2015 Technical Program Committee

Babak Abbasi, RMIT University, Australia
Mary Akinyemi, University of Lagos, Nigeria
Valentina E. Balas, Aurel Vlaicu University of Arad, Romania
Sandjai Bhulai, VU University Amsterdam, Netherlands
Simone Braun, CAS Software AG, Germany
Borut Čampelj, Ministry of Education, Science and Sport / School of Business and Management Novo mesto, Slovenia
Lijun Chang, University of New South Wales, Australia
Rachid Chelouah, EISTI, France
Haifeng Chen, NEC Laboratories America, USA
Chi-Hung Chi, CSIRO, Australia
Tsan-Ming Choi (Jason), Hong Kong Polytechnic University, Hong Kong
Esma Nur Cinicioglu, Istanbul University, Turkey
Alexandru Costan, INRIA / INSA Rennes, France
Cinzia Daraio, Sapienza University of Rome, Italy
Maria Cristina De Cola, IRCCS Centro Neurolesi "Bonino-Pulejo", Italy
Noel De Palma, University Joseph Fourier, France
Dorien DeTombe, International Research Society on Methodology of Societal Complexity, Netherlands
Mohamed Y. Eltabakh, Worcester Polytechnic Institute (WPI), USA
Serpil Erol, Gazi University, Turkey
Gustav Feichtinger, Vienna University of Technology, Austria
Denise Beatriz Ferrari, Instituto Tecnológico de Aeronáutica, Brazil
Paola Festa, Universita' degli Studi di Napoli "FEDERICO II", Italy
Yangchun Fu, University of Texas at Dallas, USA
Fausto P. Garcia, Universidad Castilla-La Mancha, Spain
Clemens Grelck, University of Amsterdam, Netherlands
Thomas Gottron, Universität Koblenz-Landau, Germany
Venkat N. Gudivada, Marshall University, USA
Ali Fuat Guneri, Yildiz Technical University, Turkey
Gür Emre Güraksin, Afyon Kocatepe University, Turkey
Wen-Chi Hou, Southern Illinois University, USA

Nilesh Jain, Intel Labs, USA
Hai Jiang, Arkansas State University, USA
David Kaeli, Northeastern University, USA
Jinho Kim, Kangwon National University, South Korea
Alexander Lazovik, University of Groningen, Netherlands
Kun Liu, KTH Royal Institute of Technology, Sweden
Claudio Lucchese, ISTI-CNR, Italy
Victor E. Malyshkin, Russian Academy of Science, Russia
Armando B. Mendes, Azores University, Portugal
Pablo Moscato, University of Newcastle, Australia
Ingo Müller, Karlsruhe Institute of Technology (KIT) / SAP SE, Germany
Hidemoto Nakada, National Institute of Advanced Industrial Science and Technology, Japan
Mirco Nanni, KDD Lab - ISTI-CNR, Italy
Sadegh Nobari, Skolkovo Institute of Science and Technology, Russia
Mario Pavone, University of Catania, Italy
Yonghong Peng, University of Bradford, UK
Serge G. Petiton, University of Lille 1, France
Jaroslav Pokorny, Charles University, Czech Republic
Meikel Poess, Oracle Corporation, USA
Filip Radulovic, Universidad Politécnica de Madrid, Spain
Paolo Romano, University of Lisbon / INESC-ID, Portugal
Ismael Sanz, Universitat Jaume I, Spain
Stefanie Scherzinger, Regensburg University of Applied Sciences (OTH Regensburg), Germany
Suzanne Michelle Shontz, University of Kansas, USA
Abhishek Sharma, NEC Laboratories America, USA
Patrick Siarry, Université de Paris 12, France
Dave Snowden, Cognitive Edge Pte Ltd, UK
Srivathsan Srinivasagopalan, Cognizant, USA
Jacek Sroka, University of Warsaw, Poland
Bela Stantic, Griffith University, Australia
Arthur Tórgo Gómez, Universidade do Vale do Rio dos Sinos (UNISINOS), Brazil
Henry Tufo, University of Colorado at Boulder, USA
Antonino Tumeo, Pacific Northwest National Laboratory, USA
Bhekisipho Twala, University of Johannesburg, South Africa
Liqiang Wang, University of Wyoming, USA
Ouri Wolfson, University of Illinois, USA
Chase Qishi Wu, University of Memphis, USA
Yinglong Xia, IBM Research, USA
Feng Yan, College of William and Mary, USA
Hongzhi Yin, University of Queensland, Australia
Feng Yu, Youngstown University, USA
Stefan Zander, FZI Research Center for Information Technology, Germany
Daqiang Zhang, School of Software Engineering - Tongji University, China
Vincent Zheng, Advanced Digital Sciences Center, Singapore

KESA 2015 Co-Chairs

Maria Pia di Buono, University of Salerno, Italy
Mario Monteleone, University of Salerno, Italy
Annibale Elia, University of Salerno, Italy

KESA 2015 Communication and Coordination

Alessandro Maisto, University of Salerno, Italy
Serena Pelosi, University of Salerno, Italy

Technical Program Committee

Rodrigo Agerri, University of the Basque Country, Spain
Ahmet Aker, University of Sheffield, UK
Flora Amato, University of Naples, Italy
Diego Ceccarelli, ISTI-CNR, Italy
Kavallieratou Ergina, University of the Aegean, Greece
Xavier Blanco Escoda, Universitat Autònoma de Barcelona, Spain
Kristina Kocijan, University of Zagreb, Croatia
Gijs Koot, TNO, Netherlands
Giuseppe Laquidara, X23 Ltd., Italy
Kun Lu, University of Oklahoma, USA
Antonino Mazzeo, University of Naples, Italy
Thiago Pardo, University of São Paulo, Brazil
Jan Radimsky, University of South Bohemia, Czech Republic
Giovanni Semeraro, University of Bari, Italy
Max Silberztein, University de Franche-Comté, France

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Numerical Groundwater Model Results As Linked Open Data

Xavier Almolda Cardona, Francisco Batlle Pifarré
Hydromodel Host, S.L.
Barcelona, Spain
xavi.almolda@hydromodelhost.com, francisco.batlle@hydromodelhost.com

Abstract— **A service has been developed that facilitates the management of groundwater bodies (aquifers) using linked open data. Numerical models simulating the behavior of each aquifer are hosted in Cloud Services, both for data model as for the model execution itself in a flexible virtual dedicated computer. Data are periodically and automatically obtained from open databases (mainly climatic conditions data) and public agencies (water level observations and detractions) in order to update the model. An ontology has been designed to describe groundwater data coming either from measuring sites or model results. This ontology has been applied to an area of interest with an associated numerical model and transformed its results to Linked Open Data (RDF) files. This set of Linked Open Data files has been stored in a RDF store (Strabon) located in a cloud platform. Finally, a set of Web Services has been designed to query the above mentioned database as a public interface. This methodology has been applied to the Delta del Llobregat Aquifer located south of Barcelona (Spain).**

Keywords-Linked Open Data; cloud computing; groundwater; numerical models; ontology; Web Service.

I. INTRODUCTION

There is an increasing environmental awareness and quantitatively based management of natural resources, but specifically on the Water Framework Directive (WFD), which demands not only quantitative assessment of the status of water bodies, but also their expected evolution. Moreover, it requires active engagement of the public, which requires providing access to model results. The goal is to facilitate frequent updating of the model by means of integration of the large volume of observation data being collected by Water Agencies, and communication of model results. Numerical groundwater models allow assessing the current status of those bodies, their evolution under natural conditions and their sensitivity to human actions.

Many projects and initiatives are trying to maximize the exploitation of linked data. The W3C Linking Open Data Project [1] aims at making data freely available to everyone by publishing open data sets on the Web and by linking items from different data sources, and the Geospatial Semantic Web [2] goal is to use the Web as a platform for geospatial knowledge integration.

In the context of the EU Project MELODIES (Maximizing the Exploitation of Linked Open Data In Enterprise and Science) [3] a service has been developed that facilitates the management of groundwater bodies by means of numerical models using linked open data.

In this paper, a methodology is described in which the numerical models are hosted in Cloud Services that provide space and computing capabilities for executing model simulations [4]. This project is also using Strabon [5], an open-source system for storing and analyzing time-evolving geospatial Linked Data, like dynamic meshes of values that change over time.

Numerical models typically discretize spatial geometry using grids or meshes. Although some ontologies for groundwater data exist [6], an ontology from the model results has been derived since no ontologies have been found related to groundwater data derived from numerical models; and it has been complemented with an ontology for times series observations. The linked open data that results from applying this ontology has been stored in the above mentioned Strabon storage and made available via web services.

II. GROUNDWATER MODEL RESULTS DATA

The groundwater model data that will be published as Linked Open Data consist of time series, either of point values or mesh values. A mesh value is a collection of values over a space discretization, such a as a finite element mesh.

A. Time series of mesh values

Time series of mesh values consist of a list of mesh values for a given variable, each one with its date and time property. Each mesh value consists of a space discretization and a set of values for each node or element of this space discretization. The space discretization is defined by a set of points in space called nodes, and a set of elements that connect these nodes forming polygons. For the time series of mesh values, each new value has been stored as a mesh with its geometry and values for nodes and/or elements.

Meshes have been stored as Resource Description Framework (RDF) triples in an RDF store [7].

B. Time series of point values

Time series of point values consists of a list of time and value pairs, for a given location in space and a specific variable. The locations of these time series are sites where the measurements are made, like wells or meteorological stations.

 ISBN: 978-1-61208-445-9

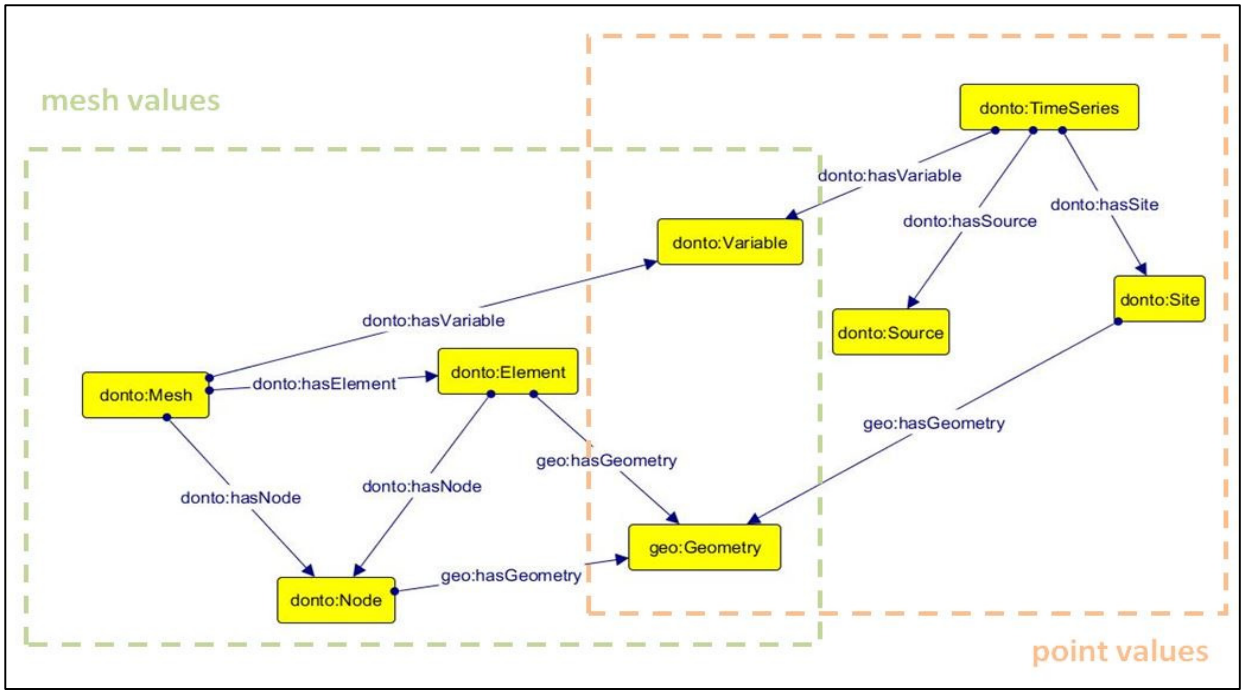


Figure 1. Groundwater results ontology diagram

Due to the large amount of time series data of point values, the individual values have been stored in an Observations Data Model (ODM) database [8], and Representational State Transfer (REST) web services have been provided to query and retrieve this data as WaterML [9].

For the purpose of linking these time series of point values with other Open Data datasets and to facilitate the discovery of new time series, a catalog of available time series for the model has been stored as RDF data in an RDF store, with start time, end time and location among other properties.

III. MODEL ONTOLOGY

In order to represent the data from the model as RDF triples, an ontology [10] has been created to describe the objects that contain this information.

Time series of values defined over a spatial discretization are defined as a list of meshes, with values on its elements or nodes, and with a *hasTimestamp* property that indicates the date and time of the simulated values.

In order to expose the time series of point values as Open Linked Data, the ontology describes the elements necessary to represent the catalog of time series available for the model. This catalog allows the user to retrieve the time series information for a given location, variable or source and use this data to retrieve the values from web services.

A diagram of the main classes of the ontology is shown in Figure 1. The elements shown have two different prefixes, the prefix *geo* for the elements of the geoSPARQL namespace, and the prefix *donto* for the elements of our ontology.

A. Mesh

A mesh is an unstructured grid that consists of nodes and elements. The elements are sets of connected nodes. The property *hasVariable* links this mesh to a Variable object that indicates the type of values for the nodes and/or elements. The data property *hasTimestamp* represents the date and time of the simulated values for this mesh.

B. Element

An element is a geometric element connecting a list of nodes. Each element has a geometry property. This geometry property is redundant because it can be obtained from the nodes of the element, but it can be useful for spatial queries involving elements.

C. Node

Each node of the unstructured mesh has a geometry property with its location of space and another property with its calculated value.

D. Variable

A variable is the type of property calculated for the nodes or elements, such as water level or chloride concentration.

The data properties for a Variable are:

- *hasName*: Name of the variable.
- *hasCode*: Code used to identify the variable.
- *hasGeneralCategory*: Category of the variable.
- *hasValueType*: An indication of whether the value represents an actual measurement, a calculated value, or is the result of a model simulation.
- *hasDataType*: An indication of the kind of quantity being measured, such as (according to WaterML

 ISBN: 978-1-61208-445-9

controlled vocabulary): Average, Best Easy Systematic Estimator, Categorical, Constant Over Interval, Continuous, Cumulative, Incremental, Maximum, Median, Minimum, Sporadic, Standard Deviation, Variance and Unknown.

E. TimeSeries

Each Time Series is identified by three characteristics:

- The *variable* measured.
- The *site* where the measurements were taken.
- The *source* that collected the measurements.

Each Time Series has the following data properties:

- *hasBeginDateTime*: Date and time of the first value of the series.
- *hasBeginDateTimeUTC*: Date and time of the first value of the series in UTC time.
- *hasEndDateTime*: Date and time of the last value of the series.
- *hasEndDateTimeUTC*: Date and time of the last value of the series in UTC time.
- *hasVariableUnitsName*: Name of the variable observed.
- *hasTimeUnitsName*: Name of the time units of the variable.
- *hasQCLCode*: Code that indicates the Quality Control Level of the measurements.
- *hasMethodDescription*: Description of the method used for obtaining the values.
- *hasValueCount*: Number of values in this series.
- *hasValues*: link to the web service with the values for this series.

F. Site

A site is the location in space where measurements are made. It has a geometry property indicating its geographic location, that is, a point.

It also has the following data properties:

- *hasName*: Name of the site.
- *hasCode*: Unique code that identifies the site.

G. Source

A source element contains the information on the original source of the observation.

It has the following properties:

- *hasContactName*: Name of contact, or title of organization.
- *hasDescription*: Full text description of the source of the data.
- *hasOrganization*: Name of the organization that collected the data.

IV. LINKED OPEN DATA

In order to make the groundwater model data available as linked data, it has been transformed to RDF triples and these triples have been stored in an RDF store.

The data from the time series of mesh values have been encoded as RDF triples and stored in Strabon, a RDF store with spatial and temporal capabilities that allows the use of stSPARQL [5] to query this data.

Data can be retrieved by accessing the web page of the Strabon endpoint or by making REST web service calls to the Strabon endpoint. The result can be obtained in a variety of formats, such as RDF/XML [11] or KML [12]. An example of a stSPARQL query on the Strabon endpoint web page can be seen in Figure 2.

The main purpose of linking time series data is not necessary to access each of the individual value of the time series of point data as RDF but the whole series, hence only the catalog of the available time series has been encoded as RDF data.

For each time series available, the catalog provides, among other information, the location in space (point), the variable measured, the first and last date of the values, the original source of the data, and a link to a web service with the necessary parameters to retrieve the set of values of this time series as WaterML.

A REST web service has been implemented, and is part of a family of web services that provide access to the time

```
SELECT DISTINCT ?elemWKT ?tstamp
WHERE {
?mesh rdf:type donto:Mesh;
      donto:hasElement ?elem;
      donto:hasTimestamp ?tstamp.
?node rdf:type donto:Node;
      donto:hasValue ?nodeValue.
?elem rdf:type donto:Element;
      geo:hasGeometry ?eg;
      donto:hasNode ?node.
?eg rdf:type geo:Geometry;
   geo:asWKT ?elemWKT.
FILTER ( ?nodeValue > 2000 )
}
```

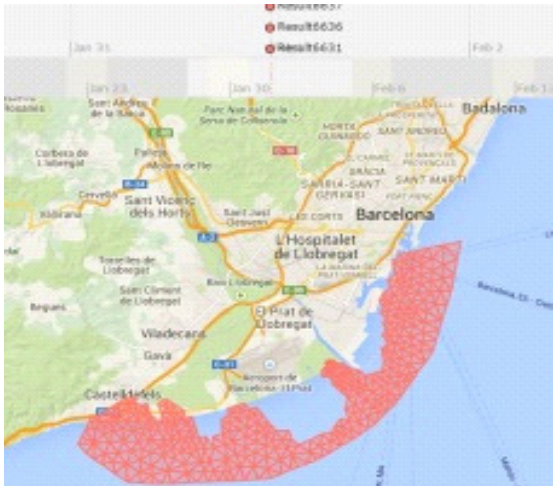


Figure 2. stSPARQL query and results (mesh elements filtered by a certain condition) on the Strabon endpoint

series data as WaterML. The main methods implemented are the following:

- *GetSites*(): Returns a list of sites with their basic information.
- *GetSiteInfo*(): It returns the information of one site.
- *GetVariableInfo*(): Returns information about a time series variable.
- *GetValues*(): It returns a time series for a given variable, at a given location and for a given time interval.

The diagram in Figure 3 shows the architecture of the linked data service.

V. APPLICATION

This technology has been tested and applied to the groundwater model of the Llobregat Delta Aquifer [13].

This model has a two-layered mesh of 2920 nodes and 5538 elements and it calculates flow and transport (chloride) data. The first 11 monthly values of the model simulation have been converted to RDF and uploaded to Strabon as a time series of mesh values. A catalog of more than 7000 time series of point data has also been uploaded as RDF to Strabon. The result is a dataset with more than 800,000 RDF triples for both model simulation and a time series catalog. A web portal with public access has also been created to test data [14].

VI. CONCLUSIONS

An ontology has been defined to describe the elements of groundwater model results and the time series of observations used by the model. This ontology describes how to convert this information into RDF. The use of RDF allows other services or applications to cross their data with the model results.

By choosing a spatiotemporal RDF store like Strabon, topological or temporal relations can be established to link the results from the numerical model with other existing datasets.

By publishing the catalog of time series and the links to obtain the values as Linked Open Data, the time series can be linked with other spatial or temporal information without having to store each of the individual values in the RDF store.

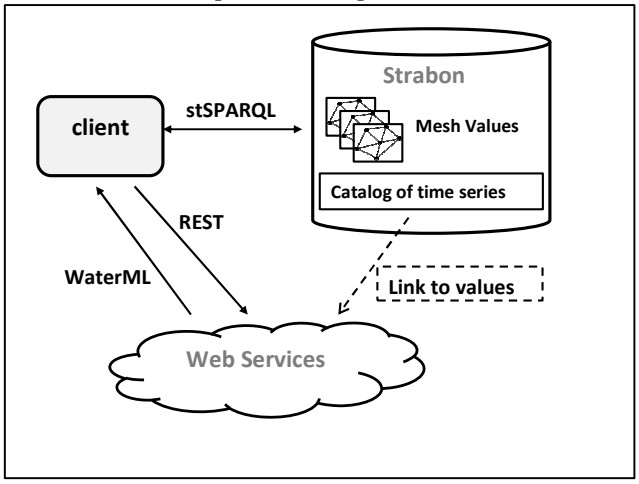


Figure 3. Linked Open Data Service architecture schema

Finally, this methodology has been tested and applied to a real model located near Barcelona (Spain).

ACKNOWLEDGMENT

This work has been funded by the FP7 project MELODIES under grant agreement number 603525.

REFERENCES

[1] C. Bizer, T. Heath, and T. Berners-Lee, "Linked Data - The Story So Far", IJSWIS, Vol. 5, Issue 3, pp. 1-22, 2009

[2] "Geospatial Semantic Web Community Group", Retrieved from https://www.w3.org/community/geosemweb/ [retrieved: March, 2015]

[3] J. D. Blower, D. Clifford, P. Gonçalves, and M. Koubarakis, "The MELODIES Project: Integrating diverse data using Linked Data and Cloud Computing" in ESA Big Data From Space Conference, 2014, pp. 244-247.

[4] J. Jódar, X. Almolda, F. Batlle, and J. Carrera, "Model Hosting for continuous updating and transparent Water Resources Management", Geophysical Research Abstracts (15), EGU2013-13009-1, 2013.

[5] K. Kyzirakos, M. Karpathiotakis, and M. Koubarakis, "Strabon: A semantic geospatial DBMS", in International Semantic Web Conference (1), 2012, pp. 295-311.

[6] B. Brodaric and T. Hahmann. "Towards a Foundational Hydro Ontology for Water Data Interoperability", Proc. of the 11th Int. Conference on Hydroinformatics (HIC-2014).

[7] R. Cyganiak, D. Wood, and M. Lanthaler, "RDF 1.1 Concepts and Abstract Syntax", W3C Recommendation, 25 February 2014. URL: http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/. The latest edition is available at http://www.w3.org/TR/rdf11-concepts. [retrieved: March, 2015]

[8] D. G. Tarboton, J. S. Horsburgh, and D. R. Maidment, "CUAHSI Community Observations Data Model (ODM)", May 2008. Available from: http://his.cuahsi.org/documents/odm1.1designspecifications.pdf. [retrieved: March, 2015]

[9] D. Valentine and I. Zaslavsky, "CUAHSI WaterML 1.1. Specification. Introduciotn to WaterML Schema", June 2009. Available from: http://his.cuahsi.org/documents/WaterML_1_1_part1_v2.docx. [retrieved: March, 2015]

[10] N. Shadbolt, W. Hall, and T. Berners-Lee, "The Semantic Web Revisited", IEEE Intelligent Systems Journal, May/June 2006, pp. 96-101.

[11] F. Gandon and G. Schreiber, "RDF 1.1 XML Syntax", W3C Recommendation, 25 February 2014. URL: http://www.w3.org/TR/2014/REC-rdf-syntax-grammar-20140225/. The latest published version is available at http://www.w3.org/TR/rdf-syntax-grammar. [retrieved: March, 2015]

[12] T. Wilson, "OGC KML 2.2.0", Document #07-147r2, Open Geospatial Consortium. April 2008. Available from: http://www.opengeospatial.org/standards/kml. [retrieved: March, 2015]

[13] E. Abarca, E. Vázquez-Suñé, J. Carrera, B. Capino, D. Gámez, and F. Batlle, "Optimal design of measures to correct seawater intrusion", Water Resources Research, Vol.42, W09415, doi:10.1029/2005WR004524, 2006.

[14] "Hydromodel Host Open Data Portal", Available from: http://h2-lod.cloudapp.net. [retrieved: March, 2015]

 ISBN: 978-1-61208-445-9

Big Data for Demand Management Programs Designing for Colombia's Industrial Sector

Jairo Pineda Agudelo, Sandra Ximena Carvajal Quintero, Daniela Valencia Lopez
Universidad Nacional de Colombia
Manizales, Colombia
e-mails: {jpinedaa, sxcarvajalq, davalencialo}@unal.edu.co

***Abstract*—Demand Management Programs is not a new concept; moreover, the key technologies for their implementation are already successful. However, Demand Management Programs applications in a worldwide context have been slow, especially in the industrial sector. Despite this, emerging countries like Colombia have great opportunities to internalize these policies in their energy planning and economic growth programs, so as to maximize their use as a tool integrated to energy markets. Demand Management Programs may allow to deal with the risks associated with system demand and to satisfy the reliability needs of an active and dynamic energy market. For this to take place, one should migrate to active and dynamic demand response, under reliability criteria based on the smart grid paradigm. This paper uses a big data analysis for planning industrial demand-management programs, based on the mechanisms and instruments of demand management and integration processes in smart grids.**

Keywords- All Data; Big Data; Demand Management Programs; Industrial demand; Energy Efficiency; Open Data.

I. INTRODUCTION

Global consumption of electricity has increased 45% since 1980 [1] and is projected to grow 70% by 2030; this is mainly due to growth in emerging countries like China and India [2]. Meanwhile, mature markets, such as North America, Europe and Japan will also face increasing demand and limited fossil resources [3].

In Colombia, industrial demand is the largest consumer of electrical energy from fossil fuels [4] and this causes problems to the environment by greenhouse - gas emissions.

The creation of methodologies to efficiently manage electricity consumption is internationally known as Demand Management Programs (DMP). DMP have environmental benefits, allow postponing investment in electricity infrastructure and increase electricity service coverage, because an efficient management increases transformers capability near the end user and improve voltage profiles [5].

DMP design depends on the accuracy in predicting the power consumption behavior of the target demand [5]. The characterization of the demand allows a prediction that might ensure a balance between generation and demand in order to achieve a more efficient operation of the electrical power system. Therefore, it is possible to avoid cost overruns in the generation, transmission and distribution as well as a resources optimization by the trader that provides the end user with quality standard conditions [6].

The objective of this research is to select industrial users to implement DMP in Colombia. Such response is estimated from data found in a virtual platform, since deregulated users have smart metering devices that send usage information via web to the network operator [7].

The paper is organized as follows: Section II describes the proposed methodology, including response and analytical prediction, Section III presents the case study and Section IV summarizes the main conclusions and proposes future work.

II. DEMAND SIDE MANAGEMENT AND BIG DATA ANALYTICS: OVERVIEW

DMP have been implemented to achieve a better participation in demand compared to electrical market prices or the need to improve reliability levels of the electric power system. Participation of demand in DMP seeks to mitigate the constraints of power grid and yield economic benefits for stakeholders [8]. Historically, the possibility of increasing the efficiency of the system and the existing investment in generation and transmission of electricity has been the key to the introduction of DMP [9].

DMPs in mature markets, like USA and Europe, show consumption reductions up to 40% in peak hours and a reduction in the need for generation reserve up to 50% [8]. The effectiveness of DMPs in USA and Europe is due to their technology implementation, which allows them to exert active demand control [9]. In the regulation, they have also incorporated various DMPs, which are based on studies of user's behavior and habits with assistance of methodologies related to Analytics.

Latin America has slowly begun integrating policies related to the efficient use of energy and the integration of demand management programs in residential, commercial and industrial sectors. Chile, Brazil and Colombia have been referring to the region for the implementation of innovative energy policies.

Particularly, in Colombia, there are initiatives that allow, in some way, implementation of some basic concepts about the possibilities of managing electricity demand in energy

 ISBN: 978-1-61208-445-9

markets; such initiatives are advances in demand management in the country, although not are demand management programs. Among these initiatives there are the hourly pricing for the case of industrial demand and the recent decision of the regulator regarding contracts load shedding, disconnecting Voluntary Demand [10], as well as Law 1715 2014 [11], which considers the participation in energy efficiency, including renewable energy.

However, ignorance of the laws and regulations, as well as the lack of adequate human and financial resources, plus the difficulty of designing realistic and clear goals with measurable results, makes it difficult to quantify the potential benefits of these policies [12].

Currently, Colombia has data provided by concentrators of measurements, but no studies have been developed based on this information. A next step is to exploit the billions of data rows to work on transforming useful knowledge in order to provide answers to operational and market issues with active participation of demand [13].

The active participation of demand refers to the study of the users consumption behavior and the changes in those habits regarding tariff incentives that may include hourly-based price discrimination on a daily basis and by use periods [14][15].

The relation between the study of the target demand active participation and the regulatory incentives that can be implemented is the basis for the design of possible successful DMPs. An analytics method that provides guidelines on the best conditions for the development of DMPs is necessary to determine such information from data obtained by the meters [16].

Big Data Analytics is the application of advanced analytics techniques to operate on large data sets. Predictive analytics, data mining, statistics and artificial intelligence are among these advanced data analytics to examine data [16].

The management of such tools to analyze giant datasets requires identifying, combining and managing multiple data sources, and the ability to build advanced analytics models for the prediction and optimization of results. The most critical component is related to the ability to transform the passive contribution of energy customers into an active contribution in real-time to achieve the purposes of DMPs for industrial users.

The information demand for these smart grids involves the consumption quantification and characterization, including the study of clusters within industrial consumers to determine, for example, economic similarities in users with similar loads. This fits the broader definition of Big Data (large data or macro data according to Fundéu BBVA Foundation), since the meters owned by the unregulated user generate an avalanche of data that must be exploited to dramatically improve DMPs performance. According to Harvard Business Review [17], the evidence is clear: “decisions controlled by data tend to make better decisions”. Particularly, electrical system operating practices need to give greater emphasis to effective real-time operation with accurate and timely information and state-of-art technology to facilitate effective contingency planning.

Currently, Management of system security needs to keep improving to maintain reliable electricity services in this more dynamic operating environment. The challenges raise fundamental issues for policymakers [18]. Big data promises to have success in to design energy policy, since the evidence is clear: decisions controlled by data tend to make better decisions.

Along with this disruptive force associated with the data growth, data analysis has evolved into what is now known as the analytics, visualization and, in particular, data mining, from the traditional disciplines of statistical data analysis. It is actually a complement of tools caused by the evolution rather than a dilemma in terms of a break since the statistical criteria remain valid in the sense of preventing failed predictions to make Big Data in Big Data Winter. For Michael Jordan [19], this may be "Due to simple-minded and statistically unsound approaches which will produce too many false positives."

True, the immense amount of data poses unprecedented challenges in terms of its analysis, due inter alia to qualitative change that implies an increase in the amount. However, a statistical point of view is essential as it contributes to consider the component of uncertainty in predictions and errors quantification. This is missing in much of the current literature of Machine Learning [19].

The massive data production growth in all modern society dimensions and the demand for prompt responses for decision-making is a powerful challenge for data scientists who are at risk of an inappropriate statistical coverage of their work and provide models with high randomized bias. Language pushed by the urgency of the decisions ends up being the result of what Douglas Merrill called the "button effect" [20], which refers to a simple expression of the pitfalls of chance and not sufficiently controlled in the model. "Button effect" occurs when surface data analyses are made and this causes erroneous results. In this paper, we propose to reiterate the rigorous use of Big Data tools to ensure correct results.

The next step after data production is to focus on analysis in order to overcome the situation, as described by Graham Williams who suggests that we are very rich in data but very poor in information [21]. This concern coincides with IDC International Data Corporation, predictions for the year 2017 [22], including the role of the CIO Chief Information Officer, that focus 80% of their time on analytics, cyber security and new revenue sources creation through digital services. This does not include recent cybersecurity issues and expanding revenue sources that, despite their importance, are not the subject of this article; the interest is to focus the discussion on the elements of Analytics and its benefit in DMPs for an industrial user.

The study of time series originated in the periodic consumption data collection is the basis for the construction

 ISBN: 978-1-61208-445-9

of the demand curve, consumption modelling and demand prediction. Advanced analytics also offers the use of decision trees, neural networks and support vector machines to predict time series problems.

The Multilayer Perceptron (MLP) [23] appears to be the most used architecture of artificial neural networks for prediction of nonlinear time series. Forecasting electricity prices and demand are counted among its many applications.

Autoregressive neural networks (ARNN) are obtained by considering the merging of a linear autoregressive model with an MLP [23]. Their initial conceptual development is based on the development of a statistical test for nonlinearity to compare the two previous models. However, ARNN is an important alternative to the use of MLP in predicting time series due to the incorporation of the linear autoregressive component [23].

Support Vector Machines (SVM) [24] are a type of neural network that was originally designed for the solution of nonlinear classification problems; but recently, they have been applied to time series regression and forecasting problems. This is due to their generalization features, which is a direct function of their structure and methodology used for their parameters estimation [24].

Similarly, it is possible to conduct a study of clusters based on the response to different price levels. This information is included as an objective in this first stage of the DMP and it is extremely important to have data available from open source like the traditionally so-called Open Data, more recently called Urban Data [25]. Subsequent claims are related to operations analysis and analytical applications.

III. CASE STUDY

As mentioned before, the aim of this work is industrial demand. In Colombia, industrial demand is classified within the so-called unregulated users, *i.e.*, those users with consumption above 0.1 MW. The current regulation allows them to purchase electricity at prices agreed freely [6], which usually causes reductions in kWh prices, compared to the price the residential user must pay for the same kWh [26].

To access tariff privileges, industrial user must install a measuring system with telemetry capacity to determine the traded energy hourly [6]. This information is recorded in a web site via the Internet that can be accessed by the user later.

The study was conducted in the west central area of Colombia and industry data pertaining to the metropolitan area corresponding to the municipalities of Manizales, Villamaría, Neira, Palestine, and Chinchiná were used.

Unregulated users in that region were identified according to the operating company of the Colombian electricity system [27] in order to model and characterize the curve of daily demand. Consumption of industrial users will be used since the meters installed send this information to a web site.

After a characterization of industrial users and knowing the main factors that influence their behavior, the next step is to create a DMP appropriate for users, which is crucial because of the change in demand by industrial activity. Another aspect that directly affects the development of a DMP is the knowledge that the end user has about the active participation benefits he gets. It is also important to know the law requirements and its participation in these programs and the technologies that will be used for remote controls and monitoring on consumption in real-time.

This paper considers the influence of the type of activity within the industrial sector in the demand behavior to develop diversified demand graphs that allow observing a typical demand curve for selected industrial activities. Figure 1 shows the graphs of diversified demand corresponding to various activities of industrial production. The demand factor in a range of a distribution system or a load is the ratio between its peak demand in the range considered and the total installed load.

The demand factor is a dimensionless number; therefore, peak demand and installed load must be considered in the same units, the demand factor is generally less than 1 and it will be unitary when, during the interval, all installed loads are operating at nominal power. Therefore, in the time intervals in which the demand profiles shown in Figure 1 exceed the unity, industrial users are considered to be operating above 100% of permissible value for the electrical installation.

Exceeding safe operating condition means overloading the electrical system of each user. Figure 1 shows that generally this behavior occurs about two standard hours, the first reflected at 8:00 when work activity begins, and the second can be seen at 16:00 before the end of the day.

The generalized overload behavior at certain intervals causes widespread power quality problems [15] since the voltage supply is reduced and there are technical losses in the power system due to electrical current increase. These technical conditions result in economic losses for industrial users since electrical machines can reduce the production and their lifetime might decrease as well [15]. Additionally, Figure 1 shows the trend of stable consumption during the day with exceptions in the activity of cement and iron and steel, which have more marked valleys at 12m throughout the workday. In order to identify the fundamental reason of that behavior, it is necessary to have detailed knowledge about the production process development.

Typical demand curves form the basis of statistical analysis for the corresponding decision-making in terms of DMP. They gather the consumption history of different industrial activities and allow defining the demand profiles through the statistical analysis methodology of time series.

The description of such curves includes identifying the long-term trend in industrial use and its seasonal component, present in all of them, given the regularity of

 ISBN: 978-1-61208-445-9

energy consumption behavior, which is typical in any industrial activity, as deduced from visual analysis of Figure1.

The input provided by the description of the curve displays the consumption through historical time in various industrial activities and, therefore, the proposals to manage demand. Table 1 shows an operation typology in which several criteria that must be clear at the time of initiating the process of analysis and data collection are identified.

TABLE I DIFFERENTIATING FACTORS IN OBTAINING DATA

DEVELOPMENT CRITERIA	OPTIONS	
User Type	Regulated User	Unregulated User
Initiative by	Power system operator	Agents of the liberalized market
Consumer type	Low Voltage (small businesses and home users)	High voltage (Industry and high trade)
Obtaining information	Installation of smart meters and creation of specialized software to gather information	Telemetry devices owned by the user, to discriminate the value of consumption on an hourly basis

The target population of this work has the features framed in column 3 of Table 1. Data collected from the target population are used to process first description and then forecasting, so that in the time horizon defined by operators and in accordance with industry, the curves are projected in time and therefore demand. Hence, it is possible to harmonize generation and consumption in relation to the forecasts provided by the data. Of course, it is always desirable to supplement the historical course provided by the demand curve with one related to the prospects for consumption in the short- and medium-term. The final model gathers past behaviors and the immediate prospects.

The core of the Analytics of typical consummation demand curves is the Analysis of Time Series whose objective is to identify those components that are present to detect its causes and to predict future series values. Table II identifies the models for the four industrial activities.

TABLE II MODELS FOR THE FOUR INDUSTRIAL ACTIVITIES

Economic Activity	Exponential Smoothing (Brown)	ARIMA (0,2,0)
Chemical	α = 0,844 R square = 0,937 Statistical Ljung-Box Q(18) = 18,22 (0,375)	
Food	α = 0,796 R square = 0,933 Statistical Ljung-Box Q(18) = 7,936 (0,968)	R square = 0,939 Statistical Ljung-Box Q(18) = 7,936 (0,905)
Drinks	α = 0,924 R square = 0,949 Statistical Ljung-Box Q(18) = 13,005(0,736)	R square = 0,955 Statistical Ljung-Box Q(18) = 15,908 (0,599)
Plastics	α = 0,864 R square = 0,901 Statistical Ljung-Box Q(18) = 9,455(0,925)	R square = 0,935 Statistical Ljung-Box Q(18) = 12,056 (0,844)

The models of exponential smoothing and ARIMA show similarities for the four activities, R square high and significant Statistical Ljung-Box Q [28] (SPSS V. 22)

Analytics combines a statistical and an algorithmic modelling approach to build the basis of DMP.

IV. CONCLUSION

Demand management programs, framed in the context of data analysis show great potential and a promising future, especially if it is possible to realize the benefits of demand modelling as a fundamental step in understanding industrial user behavior. This is important firstly to estimate the changes that might occur in the demand curves and secondly to assess the impact of these changes in the power system.

At present, it is necessary to make a use of energy resources because of overall demand growth. Thus, sound knowledge about power consumption will have environmental and social impact. By acquiring knowledge, it is possible to transform the thinking of people for efficient use of energy.

Industrial users have particular behaviors given their productive activity, which can be a downside for DMP design. However, this type of users is important because they can have more active participation in a DMP; they also have the resources and technology to provide the data to be analyzed.

Data Analytics is the answer to the exponential growth of data that the industry acquires through remote sensing in order to provide management tools of power demand.

REFERENCES

[1] "IEA International Energy Agency 1874-2014 Energy Policiy Highlight", http://www.iea.org/publications/freepublications/publication/energy_policy_highlights_2013.pdf, [Accesed: Dec, 2014].

[2] "International Energy Agency Word Energy Outlook", http://www.worldenergyoutlook.org/media/weowebsite/factsheets/WEO2013_Factsheets.pdf, [Accesed: Dec, 2014].

[3] "World Energy Investment Outlook", International Energy Agency, http://www.iea.org/publications/freepublications/publication/weio2014.pdf, © OECD/IEA, [Accesed: Dec, 2014].

[4] B. Paola and C. Angela Ines, "Benefits of Implementing a Demand Response Program in a Non-regulated Market in Colombia", Innovative Smart Grid Technologies (ISGT Latin America), 2011 IEEE PES Conference on. October 2011. Medellín, Colombia.

[5] K. Jungsuk and R. Ram, "Demand Response Targeting Using Big Data Analytics", 2013 IEEE International Conference on Big Data, doi: 10.1109/BigData.2013.6691643, Oct 2013, Santa Clara, CA, USA.

[6] Resolution 131 of 1998. Official Journal No 43.465 of December 31, 1998. Comisión de Regulación de Energía y Gas (CREG). http:www.creg.gov.co, [Accesed: Dec, 2014]..

[7] L. M. Johanna, N. P. Phillip, K. Sila, and P. Mary Ann. "Quantifying Changes in Building Electricity Use, With Application to Demand Response", IEEE Transactions On Smart Grid, vol. 2, no. 3, September 2011, pp. 507-518.

[8] J. M Victor and R. Hugh, "Design of demand response programs in emerging countries", Power System Technology (POWERCON), 2012 IEEE International Conference

on doi: 10.1109/PowerCon.2012.6401387, November 2012, Santiago de Chile, Chile.

[9] C. Adela and L. Pedro, "Estimating the benefits of active demand management. Review of the state of art and proposals", Economic Notebooks of ICE, ISSN 0210-2633, Nº 79, 2010 p.p. 187-212.

[10] Resolution 063 de 2010. Official Journal No 47.700 of april 27, 2010. Comisión de Regulación de Energía y Gas (CREG). http:www.creg.gov.co, March 20, 2015.

[11] Law 1715 of 2014. "By regulating the integration of renewable energies non conventional of energy to the system national", May 13, 2014, Gobierno Nacional de la Republica de Colombia, Bogota, http://wsp.presidencia.gov.co/Normativa/Leyes/Documents/LEY%201715%20DEL%2013%20DE%20MAYO%20DE%202014.pdf, [Accesed: Mar, 2014].

[12] B. C. Paola, "Implementation of a program to the demand response for electric energy in an unregulated customers of market in Colombia", Rev. maest. derecho econ. Bogotá (Colombia) vol. 6, no. 6, pp: 259-292, Dec 2010.

[13] Schneider electric, Energy efficiency: Solutions Manual. http:www.schneiderelectric.es/.../eficiencia-energetica/eficiencia-energetica, [Accesed: Dec, 2014]..

[14] P. Jennifer. Analytics at SMUD evolve with the smart grid, Nov 4, 2014, http://www.intelligentutility.com/article/14/11/analytics-smud-evolve-smart-grid, [Accesed: Dec, 2014].

[15] I. Toshifiimi, H. Yusuke, and T. Kiichiro, "Definitions of Power Quality Levels and the Simplest Approach for Unbundled Power Quality Services", Ninth International Conference on Harmonics and Quality of Power Proceedings, vol 2, Oct 2000, doi: 10.1109/ICHQP.2000.897711, Orlando-Florida.

[16] J. A. Luis, "Big Data Analysis of large volumes of data in organizations", Alfaomega Marcombo Ediciones Técnicas, 2013

[17] M. Andrew and B. Erik, " Big Data: The Management Revolution" October, 2012, https://hbr.org/2012/10/bigdatathemanagementrevolution/ar, [Accesed: Dec, 2014].

[18] C. Q. Sandra, S. Jean, and A. Santiago, "Colombian ancillary services and international connections: Current weaknesses and policy challenges", Energy Policy, vol 52, Jan 2013, pp 770‑778, Special Section: Transition Pathways to a Low Carbon Economy, doi:10.1016/j.enpol.2012.10.041.

[19] J. Michel, ¨Big Data Winter ahead – unless we change course¨, By Gregory Piatetsky, Oct 30, 2014. , http://www.kdnuggets.com/2014/10/big-data-winter-ahead-unless-we-change-course.html, [Accesed: Dec, 2014].

[20] Douglas M. "Careful with easy answers of Big Data". HBR, August 5, 2014.

[21] W. Graham. "Data Mining with Rattle and R, The Art the Excavation Data for Knowledge Discovery". Springer Use R!. New York 2011.

[22] The 10 predictions of the CIO's agenda in the coming years, http://www.ticbeat.com/tecnologias/10-predicciones-agenda-cio-proximos-anos-segun-idc/, October 31, 2014.

[23] V. Juan, Z. Cristian, and V. Laura, "ARNN: A packages for time series forecasting using autoregressive neural networks" , Computer systems and informatics breakthrougs magazine, vol.8, no2, Jul 2011, Medellin-Colombia, ISSN 1657-7663.

[24] V. Juan, O. Yris, and F. Carlos, **"**Time series prediction using support vector machines", Ingeniare. Rev. chil. ing. v.18 n.1 Arica abr. 2010, pp. 64-75, http://dx.doi.org/10.4067/S0718-33052010000100008.

[25] B. Luciano, P. Kien, S. Claudio, R. V. Marcos, and F. Juliana. Structured Open Urban Data. Understanding the Landscape. BIG DATA, doi: 10.1089/big.2014.0020. [Accesed: Sep, 2014].

[26] Document CREG-138: Revision of the limit unregulated user of electricity power, December 2009. Comisión de Regulación de Energía y Gas (CREG). http:www.creg.gov.co, [Accesed: Mar, 2015].

[27] Especialistas en mercado XM filial de ISA, "list of users not regulated by voltage levels", http: http://www.xm.com.co/Pages/UsuariosNoReguladosporNivelesdeTension.aspx , November 2014.

[28] P.O. Hermelinda, Estadistics II, Exponential Smoothing (Brown), Chapter 5, tips models, Facultad de Ciencias Exactas y Naturales, Universidad Nacional de Colombia http://www.virtual.unal.edu.co/cursos/sedes/manizales/4030006/lecciones/capitulocinco/5_2_3.html, [Accesed: Mar, 2014].

 ISBN: 978-1-61208-445-9

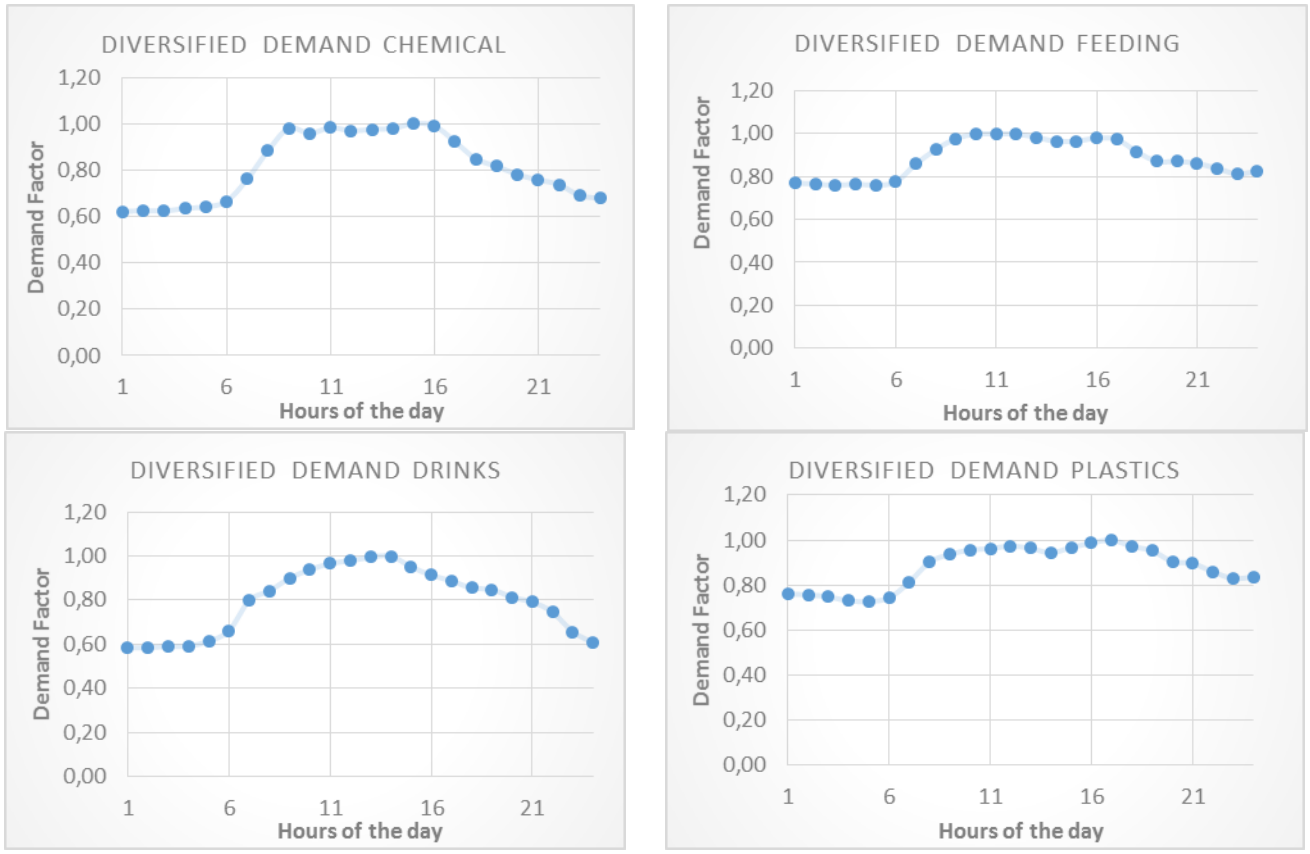


Figure 1. Typical curves of electricity demand by economic activity of industrial users.

 ISBN: 978-1-61208-445-9

Scalable Traffic Video Analytics using Hadoop MapReduce

Vaithilingam Anantha Natarajan
Department of Computer Science and Engineering
Annamalai University
Tamilnadu, India
Email: v.ananth.satyam@gmail.com

Subbaiyan Jothilakshmi
Department of Computer Science and Engineering
Annamalai University
Tamilnadu, India
Weisberg Division of Computer Science
Marshall University
West Virginia, USA
Email: jothi.sekar@gmail.com

Venkat N Gudivada
Weisberg Division of Computer Science
Marshall University
West Virginia, USA
Email: gudivada@marshall.edu

***Abstract*—Road traffic video analytics aims at using a number of techniques to achieve better traffic and road safety, control congestion and provide immediate care for accident victims. In this paper, we propose a near real-time traffic analytics system which can automatically detect road accidents from live video streams. The system alerts nearby hospitals and highway rescue teams when accidents occur. It also detects road congestion and broadcasts alternative route information to relevant commuters. We have implemented the system using Apache Hadoop. Analysis results are stored in Hive, which is a data warehouse built on top of Hadoop. This data includes overall traffic speed, traffic volume, and individual vehicle speed. Hive provides data summarization, query, and analysis. We have deployed and tested the system on a cluster computer.**

Keywords*—*Traffic Video Analytics; Road Accidents.

I. INTRODUCTION

The economy of a city greatly relies on its road network and it is important to monitor this infrastructure. Traffic jams, congestion, and accidents in city roads is a common problem in most major cities across the world. Road traffic accidents on the highways are increasing. The World Health Organization reports that by 2030 road traffic accidents will become the fifth leading cause of human death. The National Crime Records Bureau reports that every year in India more than 135,000 traffic collision-related deaths occur [1].

Real-time and historical data on road traffic is essential to effectively manage city road networks and minimize road traffic related deaths. This data can be used to ease traffic congestion by suitably programming traffic lights and suggesting alternative routes to drivers through instant messaging services. Also, historical traffic data is used to identify peak traffic hours, highly congested road intersections and accident-prone roadways.

Current generation traffic monitoring systems have the capability to capture and transmit real-time live traffic data, number of vehicles that pass through an intersection as a function of time intervals, and average speed of vehicles. The data grows in size quite rapidly and analysis entails high computational requirements. Furthermore, this data combined with Geographic Information Systems (GIS) and Global Positioning System (GPS) enable new possibilities that were not possible hitherto.

Most of the research in road traffic analytics is done in the context of developed countries where traffic follows strict lane discipline. Sensor devices are used to detect vehicles and measure road traffic variables. Sensor devices include magnetic loop sensors, speed guns, and video surveillance cameras. Magnetic loop detectors are used in large-scale for traffic monitoring [2]. Installation and maintenance of these sensor-based systems are both labor intensive and expensive. Moreover, drivers do not follow lane discipline in over 90% of the cities in the developing world. Furthermore, using sensors to detect vehicles and collect traffic data is prohibitively expensive for these cities.

Another approach to traffic monitoring and analysis is based on video image processing. In [3], an approach to vehicle detection using a rule-based reasoning is proposed. In another work [4], a Hidden Markov model (HMM)-based technique is proposed to detect accidents and other events at road intersections. Computer vision based techniques for detecting vehicles using corner features is discussed in [5]. Here again, lane abiding traffic as well as average vehicle speeds in the range 80 *km/h* to 110 *km/h* are assumed. Clearly, these approaches do not work in situations where there is no lane discipline and vehicle speed ranges vary greatly.

To circumvent the requirements of lane discipline and sensor arrays, we propose a data-driven approach to road traffic analytics using digital video. Most cities have digital video cameras installed in hundreds of locations primarily for monitoring crime and terrorist activities. They generate video data per day at the scale of terabytes. Issues involved include efficient and secure transmission and storage, processing and feature extraction, storage and retrieval of features, and performing analytics on feature data. Analytics reveal traffic patterns keyed to geographic location and time intervals, congestion and accident reports. History of feature data should be maintained to enable both descriptive and predictive analytics. The latter is critical for enhancing accuracy of traffic pattern forecasting, preventive maintenance, and proactive capacity

 ISBN: 978-1-61208-445-9

planning.

In this paper, we describe a traffic monitoring system that we developed which is suited for road traffic conditions in the developing world. Our system is based on Hadoop MapReduce framework and can capture, process, store, analyze, and retrieve video data at scale. Our system detects and tracks individual vehicles in the video frames and computes total number of vehicles that have passed through an intersection over a time interval. It also computes the speed of individual vehicles and average speed of vehicles. The system detects vehicle collisions which can be communicated to a nearby hospitals and highway rescue teams in real-time. Additional functionality of the system includes suggesting alternative routes to commuters when congestion is spotted on roadways.

The remainder of the paper is structured as follows. The traffic monitoring system is outlined in Section II. Architecture of the system is described in Section III. Section IV describes various processes involved in computing traffic analytics – video splitting technique using Apache Hadoop, vehicle detection using Haar classifiers and Support Vector Machine (SVM), and an algorithm for speed estimation. Experimental results and their analysis are presented in Section V. Finally, Section VI concludes the paper.

II. MapReduce based Video Processing

Our application uses Apache implementation of MapReduce programming paradigm, Hadoop, to chunk incoming video frames and decode them into a sequence file of image frames in parallel mode. The application can run on a local network or can be deployed on clouds that support Infrastructure as a Service (IaaS). The Hadoop Distributed File System (HDFS) is used for storing massive volumes of data across a number of physical disk clusters.

To enhance retrieval and processing speeds, the sequence files are split into smaller units known as *chunks* and distributed over HDFS nodes. When jobs are submitted to Hadoop, they are scheduled on machines that host the required data. Our system can be integrated with various traffic video capturing devices like Internet protocol (IP) cameras, closed-circuit television (CCTV), and other traffic video streaming systems.

Different video devices capture video in different file formats. A video file format can be considered as a logical container that wraps compressed video and audio. The results presented in [6] influenced the use of MapReduce framework for our system. The latter transcodes various video code formats into MPEG-4 video format.

The raw video streams are automatically partitioned into blocks. The blocks are of size 64 MB and are stored in cluster nodes. The distributed MapReduce video processing algorithm is illustrated in Figure 1. In the first mapping phase of MapReduce, the cluster nodes are assigned data from the frames of a single video [7]. The input video frames are read and vehicles are detected using a machine-learning algorithm. For each frame processed by a map, a corresponding output frame is produced and is indexed for subsequent efficient access. The index contains pairs of key values which are used to connect a frame identifier with the corresponding data.

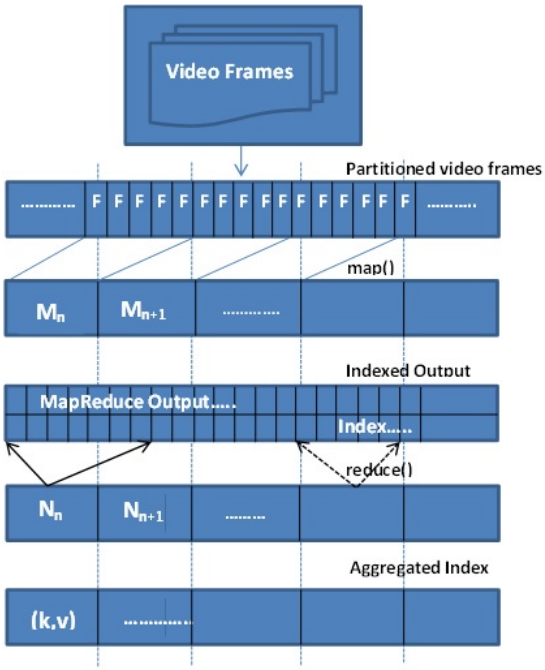


Fig. 1. MapReduce based video processing algorithm

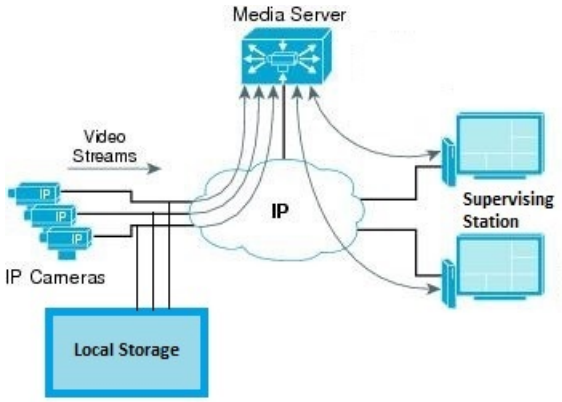


Fig. 2. System architecture

During the second phase, the index maps generated during the first phase are reduced into a single result map.

The information extracted by the MapReduce including vehicle counts, accident data and average vehicle speed are stored in Hive data warehouse for further analysis. Hive is part of the Apache Hadoop ecosystem and is optimized for analytics workloads.

III. System Architecture

The architecture of our system is shown in Figure 2. It is divided into four modules – video streams, IP wireless transmission, media server, and supervisor/control station.

The video recording module provides two functions. It streams the video to the control station. It also saves the video

 ISBN: 978-1-61208-445-9

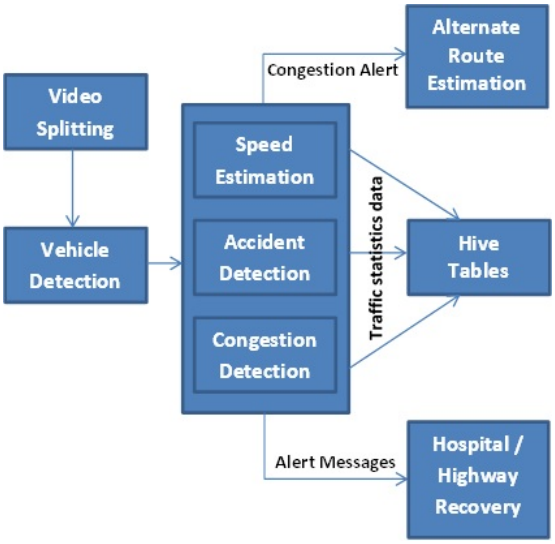


Fig. 3. Block diagram of video analytics process

in digital form in its local storage. The IP wireless transmission module routes the captured video from various spots in the city to the control station via the media server. The control station processes the live video streams and stores them in Hive data warehouse for further analysis.

The latitude and longitude of important places in and around city are stored in a Hive table. They are used to suggest alternative routes when road congestion occurs. The alternate routes are computed on-demand basis using the Google Map Directions application programming interface (API).

Our Hadoop implementation runs on a 3-node cluster. We have tested the application in three scenarios by varying the number of nodes in the cluster.

IV. MapReduce based Video Analytics

Hadoop does not have built-in capability to extract structured data from unstructured video. Shown in Figure 3 is the process our system uses for video analytics. After splitting the video into chunks, vehicles are detected from the chunks. Vehicle detection is not a trivial task and is performed in two stages. First, Haar classifiers are used for pre-detecting vehicles in the video frames [8]. This is a pre-processing step. In the second stage, a Support Vector Machine (SVM) is used to accurately detect the presence of vehicles.

The Haar classifier works fast with low-resolution images. However, SVM works slower due to algorithmic complexity involved in extracting complex features and the need for high-resolution detection window. Next, vehicle speeds are estimated, accidents and congestion are detected. All this structured data is written to Hive data warehouse. If congestion is detected, alternative route messages are sent to mobile phones of subscribed users. If an accident is detected, messages are sent to hospitals and highway rescue teams which are close by the accident site. These steps are discussed below in some detail.

A. Video Splitting

Generally in Hadoop the file is split into blocks of specified size (default is 64 MB). Each block is processed by one map process and map processes run in parallel. Larger block size offers several advantages. Video files are split into blocks in a way to avoid information loss. In other words, we do not want some video frames of an accident to go to one mapper process and remaining frames of the same accident go to another mapper.

We split video into blocks based on time units, which is specified by the users of the system. In the case of a single cluster mode, the video file is processed by one map process. For two- and three-node scenarios, the video file is processed two and three map processes, respectively. Our scalable video analytics approach is illustrated in Figure 4.

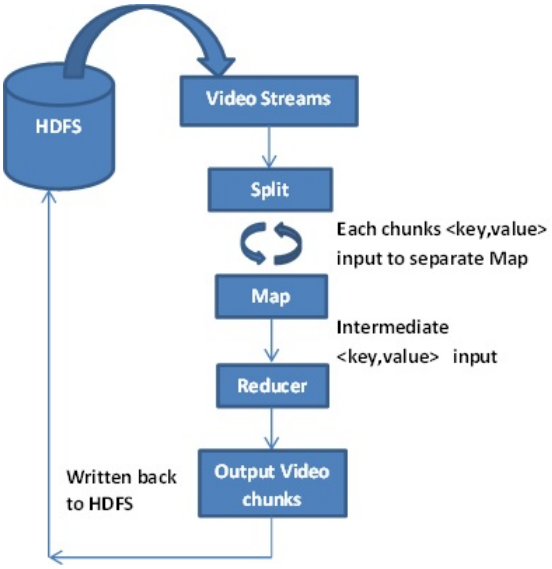


Fig. 4. Process flow for MapReduce based video processing

B. Vehicle Detection using Haar Classifier and SVM

As indicated in Section IV, vehicle detection is done using a two-step process. Haar classifiers' detection speed is high. The prediction accuracy is high with low false positives. In the second step, SVM uses Histogram of Gradients (HOG) features to further improve vehicle detection accuracy.

A Haar feature for vehicle contains a set of adjacent rectangles and the position of the rectangles is defined relative to the detection window that acts like a bounding box to the vehicle. To detect vehicles, a window is scanned over the input image, and for each subsection of the image Haar-like feature is computed. This difference is then compared to a learned threshold that separates non-objects from objects. The Haar-like features are combined into a classifier cascade to form a strong classifier.

Though vehicle detection using Haar-training is more difficult compared to SURF (Speeded Up Robust Features) [9], its robustness and execution speed compensate for this difficulty. We first train the classifier by providing both positive and negative examples. We have used 2,000 photos containing

 ISBN: 978-1-61208-445-9

vehicles (positive examples) and 2,500 background images without vehicles (negative examples) for the training step of the classifier.

A detector was used to identify Regions of Interest (ROI) in images by sliding a window across the image. The detector determines whether the detected object is present inside the window by using a cascade classifier. The size of the window is varied to detect objects at different scales, but keeping the aspect ratio constant. After coarse detection, the ROI is clipped from the original video frame and then the presence of the vehicle is verified using the SVM. If more than one vehicle is present in a video frame, all the ROIs are clipped and verified using the SVM.

C. Vehicle Speed Estimation

The speed of the moving object is computed using the displacement of centroids. To overcome the problem of perspective distortion, the camera calibration parameters are used to convert the pixel distance measured from the video frames into real world distance. Images represent the real world 3D coordinates as 2D points. Therefore, the camera calibration matrix must be known to convert the speed measurement calculated in pixels per second to the actual distance traveled by the object [10]. The accuracy of the estimated speed is calculated by comparing it to the average traffic speed.

RGB format video frames are converted into gray scale and a reference background image from video stream is generated. Then, the moving objects were extracted from the video frame by finding the difference between the current frame, I_t and the reference frame or previous frame, I_{t-1}. The frame-differencing algorithm is used to detect the motion of the detected vehicles.

When there is a change in the illumination, the reference background image is updated. The extracted vehicle moving regions in the image are converted into binary format. A series of morphological operations are performed on the binary image. Next, the objects contours are filled to remove holes inside object areas.

The speed of a vehicle is the ratio between the difference of object centroid at time t and $t + 1$ and the frame rate of the capturing device. The speed computed is in pixels per second and is converted to *km/h* using the camera calibration parameters.

Traffic congestion was detected when vehicles are traveling below the average speed for a period of time. Similarly, road accidents were detected using two measures. One is the difference between the centroids of the detected vehicles, which indicates that there may be a collision. The second one is when the vehicle speed falls close to zero. For generating rerouting messages to subscribers in case of congestion, current positions of users are calculated using the Geolocation API and an alternate route estimation is provided using Google Maps API.

V. Experiments and Results

For our experiments, we placed a handy cam just outside our lab building facing the roadside and collected videos during various lighting conditions. The camera was connected to a laptop and live video was streamed and transmitted to the MapReduce application running on a virtualization server through local campus network. The performance of the application is also tested on the road traffic videos collected from Internet sources.

Fig. 5. Result of the vehicle detection process

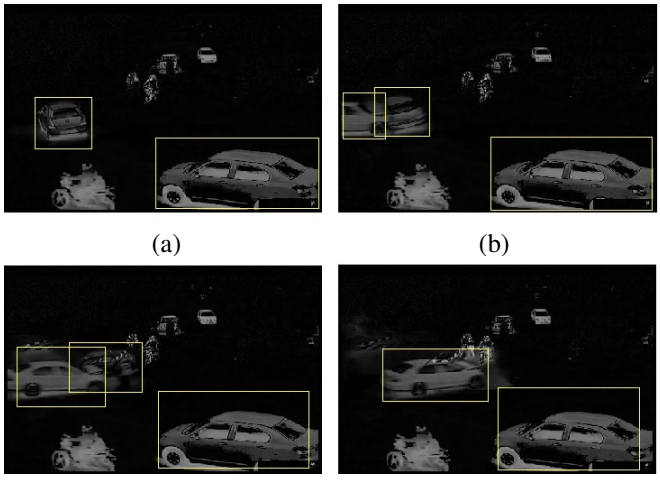


Fig. 6. (a) Single vehicle detected (b) & (c) Collision detected (d) The speed of the vehicles becomes almost zero

To detect vehicles, Haar classifier cascades have to be trained first. For training two image sets are needed – negative and positive examples. The location of vehicles within the positive images was given as input to the classifier. The SVM is then trained with HOG features extracted from the training data. The intermediate result of the vehicle analytics process in two consecutive video frames is shown in Figure 5.

Next, vehicle speed is estimated by finding the centroid of the vehicle. When there is a collision the centroid of the vehicles colliding with each other appears to be closer. This does not guarantee an accurate detection of the collision or accident since the same condition will prevail when the vehicles pass each other. When the distance between the centroids is less than a threshold value T_1, the speed of the vehicle is checked. If the speed is less than a threshold value T_2, an accident is confirmed. Figure 6 shows the output of the frame differencing process done to detect vehicle collision and speed estimation in the consecutive frames with an interval of three frames.

The result of the vehicle detection was verified manually by annotating a short video sequence and then comparing the result of the MapReduce application. The overall accuracy of

 ISBN: 978-1-61208-445-9

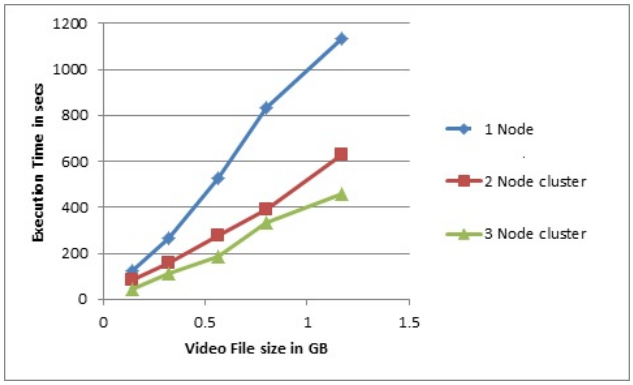


Fig. 7. Performance of MapReduce application

the vehicle detection process was found to be 88.6%. Certainly reducing the rate of false positives will increase the detection accuracy. This is achieved by limiting the region of the video frame that is analyzed.

The performance of our application is analyzed by running it on single and multiple nodes of a cluster. Figure 7 shows the performance of the application when run on one, two, and three nodes of a cluster. As expected, the time required to process video data decreases linearly with the increasing number of nodes. The scalability of our application is illustrated in Table I. When the application is run using three nodes, the processing time required to extract traffic statistics data is almost equal to the time taken by humans.

TABLE I. PERFORMANCE COMPARISON

Video file size in GB	Duration of the video file in hours	Execution time in hours		
		1 Node	2 Nodes	3 Nodes
5.60	19.20	63.60	38.40	22.10
4.80	15.30	46.20	30.10	18.16
3.43	11.15	33.80	21.90	13.30
1.17	3.80	18.15	9.89	6.05

VI. Conclusion and Future Work

In this paper, we presented an analytics solution for road traffic video using Apache MapReduce implementation, Hadoop. We have demonstrated the scalability of the system. In future, the system performance will be analyzed by incorporating more nodes. Live analysis of video data is a task operating on a stream of data. Hadoop is intended for batch processing of large volumes of data. To support real time stream computing, Storm will be considered in future instead of Hadoop. We plan to work on enhancing high-level event recognition and prediction as well as classifying vehicles. We will also investigate and validate the relationship between collision probability and safety.

Acknowledgment

S. Jothilakshmi is a postdoctoral researcher at Marshall University, USA. She is sponsored by the University Grants Commission of India under Raman Fellowship program.

References

[1] National Crime Records Bureau. Accidental deaths in India. Retrieved March 9, 2015. [Online]. Available: http://ncrb.nic.in/CD-ADSI-2012/accidental-deaths-11.pdf

[2] S.-Y. Cheung, S. Coleri, B. Dundar, S. Ganesh, C.-W. Tan, and P. Varaiya, "Traffic measurement and vehicle classification with a single magnetic sensor," *Transportation Research Record*, pp. 173–181, 2005.

[3] R. Cucchiara, M. Piccardi, and P. Mello, "Image analysis and rule-based reasoning for a traffic monitoring system," *IEEE Transactions on Intelligent Transportation Systems*, vol. 1, no. 2, pp. 119–130, Jun. 2000.

[4] S. Kamijo, Y. Matsushita, K. Ikeuchi, and M. Sakauchi, "Traffic monitoring and accident detection at intersections," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 1, no. 2, pp. 108–118, 2000.

[5] B. Coifman, D. Beymer, P. McLaughlin, and J. Malik, "A real-time computer vision system for vehicle tracking and traffic surveillance," in *Transportation Research Part C: Emerging Technologies*, 1998, vol. 6, no. 4, pp. 271–288.

[6] M. Kim, Y. Cui, S. Han, and H. Lee, "Towards efficient design and implementation of a hadoop-based distributed video transcoding system," *Multimedia and Ubiquitous Engineering*, vol. 8, no. 2, pp. 213–224, 2013.

[7] R. Schmidt and M. Rella, "An approach for processing large and non-uniform media objects on mapreduce-based clusters," *Lecture Notes in Computer Science*, vol. 7008, no. 2011, pp. 172–181, 2011.

[8] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 2001, pp. I–511–I–518.

[9] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[10] U. U. Sheikh and S. A. R. Abu-Bakar, "Three-dimensional pose estimation from two-dimensional monocular camera images for vehicle classification," in *Proc. of 6th WSEAS International Conference on Circuits,Systems, Electronics, Control and Signal Processing*, 2007, pp. 356–361.

 ISBN: 978-1-61208-445-9

Data Management Issues in Big Data Applications

Venkat N. Gudivada
Weisberg Division of Computer Science
Marshall University
Huntington, WV, USA
e-mail: gudivada@marshall.edu

Subbaiyan Jothilakshmi
Department of CS and Engineering
Annamalai University
Chidambaram, TN, India
e-mail: jothi.sekar@gmail.com

Dhana Rao
Department of Biological Sciences
Marshall University
Huntington, WV, USA
e-mail: raod@marshall.edu

***Abstract*—Big Data has the potential for groundbreaking scientific discoveries, business innovation and increased productivity. It provides as many challenges as the number of new opportunities it ushers in. However, several problems need solutions before the full potential of Big Data is realized. In this paper, we provide an overview of Big Data problems from databases perspective and elaborate on security aspects. We expect that this overview will help the reader to quickly obtain a panoramic view of research challenges in Big Data and contribute to this fast evolving discipline.**

***Keywords*—Big Data Management; NoSQL Databases; Database Security.**

I. Introduction

Wireless sensor networks, earth-orbiting satellites, social media applications, supercomputers and supercolliders, and smart phones are generating unprecedented volumes of data. In 2014, the White House commissioned a study to examine how Big Data will transform the way we live and work [1]. The report examines new business opportunities, privacy concerns, and the potential of Big Data analytics to usurp long standing civil rights of citizens. It outlines recommendations related to preserving citizens' privacy, responsible educational innovation, preventing discrimination, and judicious use in law enforcement and national security. This study attests to the role of Big Data in impacting people across the board. Big Data is a double edged sword and entails enormous implications.

It is not just the *volume* that makes this data unparalleled. Other aspects such as *velocity*, *variety*, *veracity*, and *value* bestow this data the title *Big Data*. Velocity refers to the speed at which the data is produced. For example, detecting financial fraud and real-time monitoring of cyber security requires analysis of high velocity data. Variety refers to data heterogeneity. It is often comprised of unstructured, semi-structured, and structured data of disparate types. During its life cycle, the data goes through several transformations. It is essential to be able to trace the history of such transformations to establish veracity of data. The term data provenance is used to refer to this aspect of data. Finally, collecting, cleansing, transforming, storing, analyzing, interpreting, visualizing, and querying data require substantial resources. The data should provide value and actionable information for organizations to justify investments in Big Data.

According to EMC Digital Universe with Research & Analysis by International Data Corporation (IDC) [2], data is growing at an annual rate of 40% into the next decade. What is significant is that the smart devices that are connected to the Internet – Internet of Things (IoT) – will contribute significantly to the data volumes, velocity, and heterogeneity. Furthermore, data is doubling in size every two years. The data volume will reach 44 zettabytes by 2020, from 4.4 zettabytes in 2013 [2].

Currently, much of the Big Data, especially that originating from the social media and businesses are not looked at or analyzed more than once. However, this situation is likely to change. Data becomes more useful if it is enhanced by adding meta-data and semantic annotations. According to IDC [2], by 2020, more than 35% of all data could be considered useful due to increased production of it from IoT devices.

Big Data provides challenges as well as opportunities. There are numerous challenges from databases perspective, which are discussed in Section II. The opportunities lie in creatively integrating and analyzing heterogeneous data from multiple sources to drive innovation and produce value. Big Data is also creating a new paradigm for scientific research and related applications – data-driven science. For example, many problems in natural language understanding, computer vision, and predictive data analytics are ill-posed for solution using exact algorithms. In such cases, statistical models are used to deal with the problem complexity. In his 1960 article titled *The Unreasonable Effectiveness of Mathematics in the Natural Sciences*, Wigner discusses how mathematical models developed in one context were found to be equally applicable in totally unrelated contexts [3]. In a recent article [4], Halevy, Norvig, and Pereira argue that the accurate selection of a mathematical model ceases its

 ISBN: 978-1-61208-445-9

importance when compensated by *big enough* data.

The primary goal of this paper is to provide a unified view of the various issues associated with Big Data management at a conceptual level. The intent is to help the reader quickly gain an understanding of the challenges involved in harnessing the power of Big Data. The issues we consider are data quality, data streams, dynamically evolving data, data heterogeneity and modeling, multi-model databases, client and query interfaces, data compression, data encryption, privacy, access control and authorization, and deployment on cloud-hosted cluster computers. Though not all issues are applicable to every Big Data application, often they have implications indirectly through cross interactions. For example, the complexity of client and query interfaces is directly impacted by the multi-data model.

The rest of the paper is organized as follows. Section II discusses various challenges inherent to Big Data management. One such challenge, security, is elaborated in Section III. Conclusions and future research directions are provided in Section IV.

II. Big Data Management Challenges

The challenges we discuss in this section include data quality, data streams, dynamically evolving data, data heterogeneity and data modeling, multi-model databases, client and query interfaces, data compression, data encryption, access control and authorization, and deployment on cloud-hosted cluster computers. One task that crosscuts all of the above challenges is identifying a subset of Big Data that has high value. This requires separating the data that is contaminated by spam, noise, and bias from that which is uncontaminated.

A. Data Quality

In addition to internally generated data, many organizations acquire massive datasets from diverse data vendors. Typically, the data acquired from vendors is produced without any specific application or analysis context. However, the perceived meaning of the data varies with the intended purpose [5]. This necessitates defining data validity and consistency in the context of intended data use. A related issue is inconsistency between the vendor supplied data and the same data which has been modified to conform to intended use-specific validity and consistency checks. Another issue is the need for maintaining data validity and consistency across the recent and older datasets given the long data life cycles in Big Data context.

B. Data Streams

Continuous data streams are the norm in applications, such as security surveillance, sensor networks, clickstream monitoring, and network-operations monitoring. Current approaches to data stream processing focus on application specific solutions rather than generic frameworks and approaches. For example, Najmabadi et al. [6] extracting connected component labels from image and video streams using fine grain parallel field programmable gate arrays. A low-power, many-core architecture for data stream mining applications is discussed by Kanoun et al. in [7]. Yang et al. [8] describe a cloud-based, multi-filter strategy for querying streaming XML Big Data.

Data streams pose special problems given the limited memory and CPU-time resources [9]. Unlike the *one-time database queries*, streaming data queries are *long-running and continuous*. Integrating data from multiple heterogeneous streams, mining streaming data through clustering and other unsupervised machine learning techniques, dealing with data quality issues, and real-time processing of fast moving data streams are open research issues.

C. Dynamically Evolving Data

Credit card fraud detection applications critically depend on real-time and current data. Detecting fraud in applications, such as United States (US) government sponsored health care programs Medicare and Medicaid [10] requires modeling and processing of dynamically changing data. The US Congressional Office of Management and Budget estimates that improper payments in Medicare and Medicaid programs in 2010 amounts to $50.7 billion.

Time-evolving graphs are used to model, store, process, analyze, visualize, and mine dynamically evolving data [11]. Since these graphs tend to be large and require low-latency, special hardware is used. For example, YarcData's Urika appliance was used to detect this type of fraud in real-time. The appliance memory can scale to 512 terabytes, which is shared by up to 8,192 CPUs.

D. Data Heterogeneity and Modeling

One problem that almost all organizations face is dealing with disparate data mediums and quality and extent of semantic annotations. Not all data can be stored using the relational data model and yet meet the stringent latency requirements for data access. For this reason, several data models have emerged during the last few years [12][13].

The new data models include key-value [14], column-oriented relational [15], column-family [16], document-oriented [17], and graph-based [18][19]. Furthermore, object-oriented and XML databases are reemerging in the Big Data context. Resource Description Framework (RDF) [20] data model is increasingly used for knowledge representation. The data modeling challenge in Big Data context is how to model heterogeneous data which requires multiple data models.

E. Multi-model Databases

Big Data value primarily comes through integrating massive heterogeneous datasets. As discussed in Section II-D, it is unlikely that a single model can capture the essential characteristics of heterogeneous data. It is more natural and practical to model the heterogeneous data using a collection of data models.

To provide a simple and unified user access to data, a *meta data model* should abstract all the underlying

 ISBN: 978-1-61208-445-9

data models. User queries will be specified using the meta model. Next, the query against the meta model needs to be decomposed into several queries, each of which is executed against a specific data model. Given the complexities of meta model development and query decomposition, Database-as-a-Service model [21] will simplify Big Data application development.

F. Client and Query Interfaces

Client interfaces provide programmatic access to data, whereas query interfaces are used for interactive querying and exploration of data. Efficient data access is a paramount consideration for Big Data applications. Structured Query Language (SQL) is the *de facto* standard for querying and updating relational databases. In contrast, there is no such standard query language yet which can be used to specify queries that require access to data across databases with disparate data models.

Client interfaces are typically developed by database professionals who are knowledgeable about the database schemas and how to link the data. Another issue is the number of programming languages for which client access interface is available.

The case of Statoil exploration illustrates the complexities involved in user access to Big Data [22]. Statoil is an oil and gas production company. One of its core tasks is to reduce exploration risk by developing stratigraphic models of unexplored areas using the data of previous operations at nearby sites. The data for stratigraphic modeling exceeds one petabyte and is stored under various relational database schemas. Answering queries require accessing data that is spread over 2,000 tables across different databases. As reported by Calvanese [22], answering certain queries require over four days even with the assistance of database experts. About 30 - 70% of oil and gas exploration time is spent on data gathering. Data heterogeneity will make the data access task even more time consuming.

Ontology Based Data Access (OBDA) is an approach to querying Big Data [23][24]. An ontology provides a formal representation of a domain at a conceptual level. Mappings are constructed between the ontology and data. Users specify data requests using the ontology. OBDA system translates a user data request into queries across various data sources.

Medical Literature Analysis and Retrieval System Online (MEDLINE) [25] is a bibliographic database featuring articles from academic journals covering medicine, nursing, pharmacy, dentistry, veterinary medicine, and health care. PubMed [26] stores over 24 million citations from MEDLINE and online books. Likewise, ACL Anthology is a digital archive of over 34,000 research papers in computational linguistics and natural language processing [27].

Even with advanced search tools it is often difficult to find and discover what one is looking for from large document collections. Therefore, collections like the MEDLINE and ACL Anthology need a different type of user interface for organizing, exploring, querying, visualizing, and understanding them. *Topic models* are tools just for this purpose [28]. Topic models are algorithms for discovering main themes that are present in a document collection and to organize the collection according to these themes [29].

G. Data Compression

With Big Data, secondary storage is always a premium resource. Data replication for high availability and read throughput aggravates this problem further. Compression ratios depend on the nature of the data itself as well as the compression algorithm used. Some compression algorithms are lossless, while others are lossy. If the data have been compressed using a lossless algorithm, then it is possible to recover the original data exactly from the compressed data [30]. Typically, text compression requires a lossless algorithm. Image and video data, on the other hand, may tolerate some data loss when decompressed. Application requirements dictate the choice of a compression algorithm – lossy or lossless.

Some algorithms focus on decompression speed and exploit the underlying hardware [31]. Other considerations include what additional resources are required by an algorithm. For example, some algorithms may require memory buffers for both compression and decompression [30].

Lempel-Ziv-Oberhumer (LZO) [32] is a popular, lossless data compression algorithm and its compression ratio is about 3:1. It is optimized for very fast decompression and is often used with Hadoop [33]. Relational database systems also offer options for data compression and compression ratios hover around 6:1. Gzip [34] is both a file format and a compression algorithm, whose compression ratio is about 7:1. Other products, such as RainStor database seem to provide a much higher compression ratio, 40:1 and in some cases as high as 100:1 [35].

H. Data Encryption

Big Data applications need to comply with global data security and privacy regulations to realize potential business benefits. For example, in the health care domain, patient data is made available to research and development organizations to analyze and identify emerging risks and patterns in the patient population. Health Information Portability and Accountability Act (HIPAA) and Health Information Technology for Economic and Clinical Health Act (HITECH) [36] compliance requires that the data be de-identified at the field level. If needed, the de-identified data needs to be securely re-identified for proactive treatment of the affected individuals.

Given that Big Data is stored in distributed file systems and are processed using cloud-hosted cluster computers, securing the data through encryption is extremely challenging. For example, file-system encryption is effective only for data at rest. This introduces excessive operational overhead for the continuous write encryption and read decryption. Furthermore, encryption-decryption cycle should preserve formats and referential integrity of the data. Finally, encryption/decryption should be

 ISBN: 978-1-61208-445-9

cost effective and not diminish operational flexibility or computational performance.

I. Privacy

Privacy and security are tightly integrated aspects of Big Data. Protecting rights of privacy is a great challenge. For example, in 2013, there were more than 13 million identity thefts in the United States and it is one of the fastest growing crimes [37]. Other facets such as encryption in both hardware and software, and round the clock monitoring of security infrastructure are critical to protecting privacy. The notion of personally identifiable information is difficult to define precisely. Furthermore, as data goes through various transformations, it becomes even more difficult to identify and tag personally identifiable data. It has been shown that even anonymized data can often be re-identified and attributed to specific individuals [38]. Sadly, re-identification has become a center piece for business models employed in fields such as online behavioral advertising.

J. Access Control and Authorization

Access control refers to ways in which user access to applications and databases is controlled. Databases limit access to those users who have been authenticated by the database itself or through an external authentication service, such as Kerberos [39]. Authorization controls what types of operations can an authenticated user perform. Access control and authorization capabilities of relational database systems have evolved over a period of four decades. In contrast, data management for Big Data applications is provided by a class of systems referred to as Not Only SQL (NoSQL) systems [40].

NoSQL systems principally focus on providing near real-time reads and writes in the order of billions and millions, respectively. NoSQL systems features vary widely and there are no standards yet. They use different data models, some do not provide database transactions, while others do not use SQL. They are referred to by various names including NoSQL, NewSQL, and non-RDBMS. To avoid the misconception that NoSQL systems eschew SQL, they are also referred to as *Not only SQL*.

NoSQL systems are relatively new and are evolving very rapidly. Their access control and authorization capabilities vary widely. Some NoSQL systems provide limited capabilities and some assume that the system is operating in a trusted environment. For example, initial versions of Riak, a key-value NoSQL database, provided no authentication or authorization support [41]. We elaborate on this aspect in Section III.

K. Deployment on Cloud-hosted Cluster Computers

Though Big Data applications can be developed and tested on desktop computers on a small scale, usually they are developed, tested, and deployed on cluster computers. Installing, operating, and maintaining cluster computers require specialized technical expertise in addition to significant upfront investment in hardware. For this reason, many Big Data applications are developed using cloud-hosted, cluster-powered application hosting commercial platforms such as Amazon Web Services [42] and Heroku [43]. In contrast, XSEDE is a free supercomputer platform dedicated for advancing academic science and engineering research [44].

III. Security Challenges in Managing Big Data

Database systems security has been a topic of major research interest in the database community [45]. Database security has multiple dimensions including physical, personnel, operational, and technical. The physical dimension deals with barriers to ensure physical inaccessibility to unauthorized users. The personnel facet is related to employing trustworthy people to operate the database. Policies and procedures that govern operating and maintaining aspects of databases comprise the operational dimension. These three dimensions are external to the technical aspects of database systems.

Traditionally, security aspects addressed by the database system include protecting confidentiality, ensuring data integrity, and assuring availability. Protecting confidentially involves preventing unauthorized access to sensitive information such as health records, credit history, trade secrets, marketing and sales strategies. This may require encrypting the data, authenticating users, and providing fine granular access.

Ensuring data integrity requires that data insertions, modifications, and deletions are done by authorized users in a way that none of the database integrity constraints are violated. Attacks such as data corruption through viruses and SQL injections make data integrity assurance a difficult job. High availability requires database system's resilience to attacks such as denial of service.

Big Data ushers in several more challenges. For example, initiatives by various governments, such as Right to Information [46], Freedom of Information [47], and Open Government Initiative [48] provide access to vast amounts of data to the public at large. One of the greatest challenges is privacy-preserving data mining and analytics – ensuring that deriving personally identifiable information is impossible.

The sheer volume of data can easily overwhelm the first-generation security information and event management technologies. For example, Barclays bank generated over 44 billion security events per month in 2013 [49]. Analyzing database access logs to proactively identify security breaches is also made difficult by data volume. Identifying useful data for a given context from massive datasets is a problem in itself. This problem is often referred to as *right data* in contrast to *big data*.

Data input validation and filtering, real-time regulatory compliance monitoring, and secure communications pose additional problems. Cloud-hosted, distributed cluster computing infrastructure must ensure secure computations by encrypting data during transit. Finally, data provenance [50] is an issue that received little or no attention from a security standpoint. As data goes through various transformations, metadata associated with provenance grows in complexity. The size of provenance graphs

 ISBN: 978-1-61208-445-9

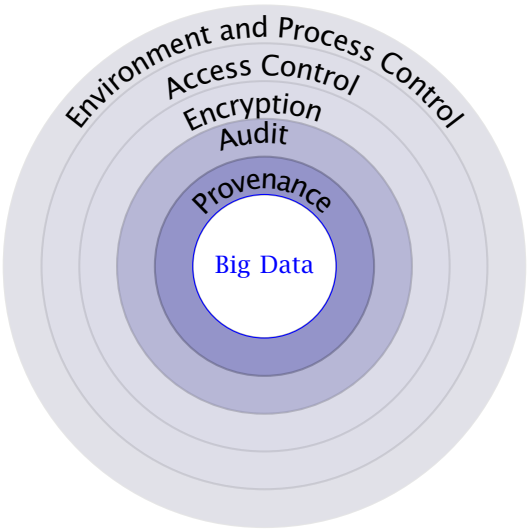


Figure: 1 Five-layer Big Data security model.

increases rapidly [51] which makes analyzing them computationally expensive.

Big Data security models entail more complexity due to their distributed nature relative to traditional database systems. We envision a five-layer security model for Big Data security as shown in Figure 1. Security controls are applied beginning with the outermost layer and sequentially progressing to the innermost layer. The outermost layer, *Environment and Process Control*, enforces physical controls in the underlying deployment environment such as firewalls, file system permissions, network configurations.

The *Access Control* layer restricts access to data through user authentication and authorization controls. The next layer provides data *encryption and decryption* services. Encryption is needed for both data-at-rest and data-in-flight. The *Provenance* layer is responsible for tracking data transformations and recording their annotations. The innermost one is the *Audit* layer which records and analyzes all database accesses in real-time to discover security breaches and to ensure compliance.

IV. Conclusion and Future Work

The ability to effectively process massive datasets has become integral to a broad range of academic disciplines. However, this does not preclude the need for deeper understanding of the theoretical foundations of scientific domains. The adage – a tool without theory is blind and a theory without tool is useless – holds in the Big Data context too. Big Data enables scientists to overcome problems associated with small data samples in ways, such as relaxing the assumptions of theoretical models, avoiding over-fitting of models to *training data*, effectively dealing with noisy training data, and providing ample *test data* to validate models.

Big Data has the potential to fundamentally affect the way we live and work. Just as a picture is worth 1000 words, Big Data analytics can help us unravel a thousand stories by analyzing and interpreting the data. Big Data offers possibilities for uncovering unexpected and hidden insights. For example, in medical and health data, these insights may lead to ground-breaking discoveries and profitable innovation. However, several problems need to be solved before this potential can be realized. As much as it sounds ironical, only technology can solve technology created problems. We expect that the research issues raised in this paper will inspire the readers to solve Big Data problems and advance this fast moving and exciting field.

Acknowledgment

S. Jothilakshmi is a postdoctoral researcher at Marshall University, USA. She is sponsored by the University Grants Commission of India under the Raman Fellowship program.

References

[1] Executive Office of the President. Big data: Seizing opportunities, preserving values. [retrieved: March, 2015]. [Online]. Available: http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_5.1.14_final_print.pdf

[2] V. Turner. The digital universe of opportunities: Rich data and the increasing value of the internet of things. [retrieved: March, 2015]. [Online]. Available: http://www.emc.com/leadership/digital-universe/2014iview/digital-universe-of-opportunities-vernon-turner.htm

[3] E. Wigner, "The unreasonable effectiveness of mathematics in the natural sciences," *Communications in Pure and Applied Mathematics*, vol. 13, no. 1, pp. 1 – 14, February 1960.

[4] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intelligent Systems*, vol. 24, no. 2, pp. 8 – 12, 2009.

[5] D. Loshin. Understanding big data quality for maximum information usability. [retrieved: March, 2015]. [Online]. Available: http://www.dataqualitybook.com

[6] S. M. Najmabadi, M. Klaiber, Z. Wang, Y. Baroud, and S. Simon, "Stream processing of scientific big data on heterogeneous platforms – image analytics on big data in motion," *2013 IEEE 16th International Conference on Computational Science and Engineering*, pp. 965–970, 2013.

[7] K. Kanoun, M. Ruggiero, D. Atienza, and M. van der Schaar, "Low power and scalable many-core architecture for big-data stream computing," *2014 IEEE Computer Society Annual Symposium on VLSI*, pp. 468–473, 2014.

[8] C. Yang, C. Liu, X. Zhang, S. Nepal, and J. Chen, "Querying streaming xml big data with multiple filters on cloud," *2013 IEEE 16th International Conference on Computational Science and Engineering*, pp. 1121–1127, 2013.

[9] L. Golab and M. T. Özsu, "Data stream management," *Synthesis Lectures on Data Management*, vol. 2, no. 1, pp. 1–73, 2010.

[10] US Government. Centers for medicare & medicaid services. [retrieved: March, 2015]. [Online]. Available: http://www.cms.gov/

[11] V. V. Raghavan. Visual analytics of time-evolving large-scale graphs. [retrieved: March, 2015]. [Online]. Available: http://grammars.grlmc.com/bigdat2015/courseDescription.php

[12] solid IT. Knowledge base of relational and NoSQL database management systems. [retrieved: March, 2015]. [Online]. Available: http://db-engines.com/en/ranking

[13] A. Schram and K. M. Anderson, "Mysql to nosql: Data modeling challenges in supporting scalability," in *Proceedings of the 3rd Annual Conference on Systems, Programming, and Applications: Software for Humanity*, ser. SPLASH '12. New York, NY, USA: ACM, 2012, pp. 191–202.

 ISBN: 978-1-61208-445-9

[14] R. Gandhi, A. Gupta, A. Povzner, W. Belluomini, and T. Kaldewey, "Mercury: Bringing efficiency to key-value stores," in *Proceedings of the 6th International Systems and Storage Conference*, ser. SYSTOR '13. New York, NY, USA: ACM, 2013, pp. 6:1–6:6.

[15] Z. Liu, S. Natarajan, B. He, H.-I. Hsiao, and Y. Chen, "Cods: Evolving data efficiently and scalably in column oriented databases," *Proc. VLDB Endow.*, vol. 3, no. 1-2, pp. 1521–1524, Sep. 2010.

[16] A. Lakshman and P. Malik, "Cassandra: A structured storage system on a p2p network," in *Proceedings of the Twenty-first Annual Symposium on Parallelism in Algorithms and Architectures*, ser. SPAA '09. New York, NY, USA: ACM, 2009, pp. 47–47.

[17] P. Murugesan and I. Ray, "Audit log management in mongodb," *2014 IEEE World Congress on Services*, pp. 53–57, 2014.

[18] R. Angles, "A comparison of current graph database models," *2014 IEEE 30th International Conference on Data Engineering Workshops*, pp. 171–177, 2012.

[19] I. Robinson, J. Webber, and E. Eifrem, *Graph Databases*. O'Reilly, 2013.

[20] Z. Kaoudi and I. Manolescu, "Rdf in the clouds: A survey," *The VLDB Journal*, vol. 24, no. 1, pp. 67–91, Feb. 2015.

[21] D. Agrawal, A. El Abbadi, F. Emekci, and A. Metwally, "Database management as a service: Challenges and opportunities," in *Data Engineering, 2009. ICDE '09. IEEE 25th International Conference on*, March 2009, pp. 1709–1716.

[22] D. Calvanese. End-user access to big data using ontologies. [retrieved: March, 2015]. [Online]. Available: http://grammars.grlmc.com/bigdat2015/courseDescription.php

[23] D. Calvanese, G. D. Giacomo, D. Lembo, M. Lenzerini, A. Poggi, and R. Rosati, "Ontology-based Database Access," in *Sistemi Evoluti per Basi di Dati*, 2007, pp. 324–331.

[24] A. Poggi, D. Lembo, D. Calvanese, G. D. Giacomo, M. Lenzerini, and R. Rosati, *Linking Data to Ontologies*, 2008, vol. 10.

[25] U.S. National Library of Medicine. Medical Literature Analysis and Retrieval System Online (MEDLINE). [retrieved: March, 2015]. [Online]. Available: http://www.nlm.nih.gov/pubs/factsheets/medline.html

[26] NCBI. PubMed, National Center for Biotechnology Information. [retrieved: March, 2015]. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed

[27] ACL, "ACL Anthology: A Digital Archive of Research Papers in Computational Linguistics," http://www.aclweb.org/anthology/, [retrieved: March, 2015].

[28] H. Soleimani and D. J. Miller, "Parsimonious topic models with salient word discovery," *IEEE Transactions on Knowledge and Data Engineering*, vol. 99, no. PrePrints, p. 1, 2014.

[29] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012.

[30] K. Sayood, *Introduction to Data Compression*, 4th ed. Morgan Kaufmann, 2012.

[31] A. Ozsoy, "Culzss-bit: A bit-vector algorithm for lossless data compression on gpgpus," in *Proceedings of the 2014 International Workshop on Data Intensive Scalable Computing Systems*, ser. DISCS '14. Piscataway, NJ, USA: IEEE Press, 2014, pp. 57–64.

[32] S. Navqi, R. Naqvi, R. A. Riaz, and F. Siddiqui, "Optimized rtl design and implementation of lzw algorithm for high bandwidth applications," *Electrical Review*, no. 4, pp. 279–285, April 2011.

[33] A. Holmes, *Hadoop in Practice*. Manning Publications Co., 2012.

[34] J.-l. Gailly, *gzip: The data compression program*. iUniverse, 2000.

[35] RainStor. Industry leading compression translates to huge cost savings. [retrieved: March, 2015]. [Online]. Available: http://rainstor.com/products/rainstor-database/compress/

[36] T. L. Murray, M. Calhoun, and N. C. Philipsen, "Privacy, confidentiality, hipaa, and hitech: Implications for the health care practitioner," *The Journal for Nurse Practitioners*, vol. 7, no. 9, pp. 747–752, October 2011.

[37] United Credit Service. Identity theft; will you be the next vitcim? [retrieved: March, 2015]. [Online]. Available: https://ucscollections.wordpress.com/2014/03/06/identity-theft-will-you-be-the-next-victim/

[38] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," *2013 IEEE Symposium on Security and Privacy*, vol. 0, pp. 111–125, 2008.

[39] S. T. F. Al-Janabi and M. A. S. Rasheed, "Public-key cryptography enabled kerberos authentication," in *Developments in E-systems Engineering (DeSE), 2011*. IEEE Computer Society, Dec 2011, pp. 209–214.

[40] V. Gudivada, D. Rao, and V. Raghavan, "Renaissance in data management systems: Sql, nosql, and newsql," *IEEE Computer, forthcoming*.

[41] solid IT. Current data security issues of nosql databases network defense & forensics insights. [retrieved: March, 2015]. [Online]. Available: http://www.fidelissecurity.com/files/NDFInsightsWhitePaper.pdf

[42] AWS. Amazon web services. [retrieved: March, 2015]. [Online]. Available: http://aws.amazon.com/

[43] Heroku. Cloud application platform. [retrieved: March, 2015]. [Online]. Available: https://www.heroku.com/

[44] XSEDE. Advanced cyberinfrastructure. [retrieved: March, 2015]. [Online]. Available: http://www.xsede.org

[45] E. Bertino and R. Sandhu, "Database security – concepts, approaches, and challenges," *IEEE Transactions on Dependable and Secure Computing*, vol. 2, no. 1, pp. 2–19, 2005.

[46] Government of India. Right to information. [retrieved: March, 2015]. [Online]. Available: http://righttoinformation.gov.in/

[47] USA Federal Government. Freedom of information act. [retrieved: March, 2015]. [Online]. Available: http://www.foia.gov/

[48] The White House. Open government initiative. [retrieved: March, 2015]. [Online]. Available: http://www.whitehouse.gov/Open/

[49] B. Glick. Information security is a big data issue. [retrieved: March, 2015]. [Online]. Available: http://www.computerweekly.com/feature/Information-security-is-a-big-data-issue

[50] U. Braun, A. Shinnar, and M. Seltzer, "Securing provenance," in *Proceedings of the 3rd Conference on Hot Topics in Security*, ser. HOTSEC'08, 2008, pp. 4:1–4:5.

[51] Y.-W. Cheah, "Quality, retrieval and analysis of provenance in large-scale data," Ph.D. dissertation, Indianapolis, IN, USA, 2014.

 ISBN: 978-1-61208-445-9

Big Data Solutions For Urban Environments

A Systematic Review

Francisco Ribeiro, Felipe Ferraz, Gustavo Alexandre
CESAR – Recife Center for Advanced Studies and Systems
Recife, Brazil
e-mail: {finr,fsf,ghsa}@cesar.org.br

Francisco Ribeiro, Felipe Ferraz,
Maria Silva, Gustavo Alexandre
Informatics Center
Federal University of Pernambuco
Recife, Brazil
e-mail: {finr, fsf3,mcts,ghsa}@cin.ufpe.br

Abstract— **Big data is a vast field of research and one that has not been completely explored. Studies involving big data, especially for solving problems in urban environments, still require more evidence. This work has the objective of identifying, evaluating, and interpreting published research that examine usages of the great amount of data generated by cities and its systems, in order to use Big Data technologies to improve city conditions. To achieve that, a systematic review of current literature was conducted. This review resulted in the finding of 1291 works of which 40 were identified as primary studies. The studies were then classified according to research focus and aspect of the city they focused on. The review investigates what it is known about the benefits and limitations in the use of big data in urban environments. The results show statistical data about big data, gaps in current research and models of successful implementation.**

Keywords— Big Data; Smart City, Systematic Review, Ubiquity

I. INTRODUCTION

Two global tendencies have been significantly influencing the information technology field in recent years: the growth in city population, marked by the migration of people from rural zones to urban ones; and the exponential rise of data generated by citizens stemming from the use of pervasive computing [1]. In this context, many challenges, opportunities for learning and for better management of cities arise through the use of new and extensive data sources, extracting data from them.

According to the United Nation Population Fund (UNFPA), 2008 was marked as the year in which more than half of the planets inhabitants, 3.3 billions, are now living in urban environments. With this unprecedented growth, many problems such as lacking transportation infrastructure, quality health assurance, citizens safety, unemployment rising, are driven cities in order to develop new and different means to try to avoid and mitigate those kind of issues, in order to provide citizens with a better living.

However, in tandem with the increase of the population also increases the amount of data that is generated by the same population. This can be leveraged to help improving city services and to better control and manage resources through the use of new technologies. Such as Big Data, data mining and/or data analysis.

The increase for data that is generated at a daily basis is explicit. According to Eron Kelly, "In the next five years we will generate more data as a human species than we have in the last 5.000 years" [2].

Develop and use new means to connect, integrate and to analyze such overwhelming amount of data is an important step to city continuity and competitiveness. It is vital to them to develop this kind of approach in order to help citizens providing them with a cohesive and intelligent view of the city.

The ubiquity of data sources and the resulting richness of information creates new research areas in computer and social sciences. Its application in diverse segments of industry is focus of many study efforts. However, the knowledge of the opportunities and challenges that arise through the use of big data technology applied to urbanization are still not sufficiently structured [1]. Albeit that many urban environments already use big quantities of data to improve infrastructure, planning and management, the term, as well as the underlying concept of big data are still lacking of more efforts. Among many different definitions, according to [4], big data refers not only to size of data but also to its speed and variety, and to obtain valuable information from this group of data, moreover, through the correct data analysis and leveraging on the size of the analyzed data it is possible to statistically increase the precision of the resulted information. In other words, a big data solution uses a group of data that is constantly increasing in order to provide its maintainers with more and more correct information that could be used to increase several aspects of a area.

The study contained herein is an effort to map out the current knowledge of the aforementioned issues. The conditions under which big data could be used strategically need to be understood, its limitations, anticipated and clarified.

This work is organized as such: In Section 2, basic concepts related to Big Data and the used methodology will be presented along with the objectives of the study. In Section 3, the methods, processes and the protocol that were used in the systematic review will be described. In Sections 4 and 5 the results related to the conducted research will be detailed presented. Finally, in Section 6 some conclusions and future works will be depicted.

II. BIG DATA

Big data is a term that is widely used in both academia and business. Despite the widespread adoption of the term, its meaning is still relatively unclear.

According to Cukier et al. [5], big data refers to the analysis of big quantities of data to find useful relationships and/or patterns. Other aspects can be seen in the definition of Michael Friedenberg – president and CEO of IDG Enterprise. According to him, the term refers to sets of data that are so big

 ISBN: 978-1-61208-445-9

-in terms of volume, speed and variety - that they become impossible to manage by conventional databases [6].

The amount of digital data that is currently being created on a daily basis by social network apps, embedded systems etc. is huge. Terabytes of information are produced from various types of sources

The application of big data technologies in urban environments can lead to truly intelligent cities that can be managed in real time with a high degree of precision [9].

III. APPLIED PROTOCOL

Based upon the guidelines for the development of systematic reviews in software engineering described by Kitchenham [10] and the analysis of the review model by Dybå et al. [11], a new methodology for revision was created. Our review methodology is composed of six steps: (1) development of the protocol, (2) identification of inclusion and exclusion criteria, (3) search for relevant studies, (4) critical assessment, (5) extraction of data, and (6) synthesis. The steps applied to the study contained herein are presented below:

The objective of this review is to identify primary studies that focus on the use of big data techniques that aim at solving urban problems. The following question helps identifying primary studies:

- *How is it possible to improve urban environment using big data and what are the challenges that accompanies the use of such technology for the creation of smart cities?*

From this central question, others secondary questions were developed of to help in the comprehension of the problem:

- *What aspects of urban environments can be optimized through the use of big data?*
- *Which solution models can the applied to cities?*
- *How can citizens and government officials benefit from the use of big data technologies?*
- *What are the main challenges in using big data in urban environments?*

A. Inclusion and Exclusion Criteria

For this review, we considered studies that aim at analyzing the use of big quantities of data to improve the efficiency and effectiveness of any indispensable city service. The studies could refer to specific sectors (e.g., traffic control or security) or have a broader scope taking into account many types of services. Since this field of research is recent, this review limited the examined studies to the ones published starting from year 2004.

Were also excluded:

- Studies not published in the English language;
- Studies that were unavailable online;
- Studies not based on research and that express only the official opinions of governments and field experts;
- Call for works, prefaces, conference annals, handouts, summaries, panels, interviews and news reports.

B. Search Strategies

The databases considered in the study is in the list below:

- ACM Digital Library;
- IEEE Xplore;
- ScienceDirect – Elsevier;
- SpringerLink.

Combinations of terms were created to guarantee that relevant information would not be excluded when querying different search engines and databases. As a result, four search strings were created:

1. "big data" AND (city OR cities);
2. "big data" AND citizen*;
3. "big data" AND urban*;
4. "big data" AND govern*;

In the process of extracting information from the databases, the search strings were used separately on each database. The searches were performed between December 2013 and January 2014. The results of each search were grouped together according to database and were, later, examined closer in order to identify duplicity. Table 1 shows the amount of studies found on each database.

TABLE I. AMOUNT OF STUDIES FOUND ON EACH DATABASE

Database	Number of studies
ACM Digital Library	396
IEEE Xplore	114
ScienceDirect – Elsevier	515
SpringerLink	290

C. Studies Selection Process

This Section describes the selection process from the beginning: from initial search using the Search Strategies described below to identification of primary studies.

At the first step, the studies that were obtained from the databases were gathered and added to Mendeley's citation management tool. This resulted in the finding of 1291 non-duplicated citations.

Secondly, the titles of all works selected in the previous step were analyzed to determine its relevance in this systematic review. At this stage, many works that did not mentioned using big data being to improve city conditions were eliminated.

Due to the use of terms related to city data, many works of geography, biology, medicine and sociology were found. In those cases, all works whose titles did not conform to the scope of the review were eliminated. In other cases, when the works titles were vague or unclear, they were put aside to be analyzed in the next step. At the end of this stage, 981 citations were excluded, thus remaining 310 items for further analysis.

In the third step, all abstracts of the works found in the

 ISBN: 978-1-61208-445-9

previous phase were assessed. Once more, many studies whose primary focus was not the analysis of large data directed to solving urban problems have been eliminated. We could conclude that the abstracts varied a lot in quality. Some items did not even had abstracts or had abstracts that did not clearly presented what the rest of the article was about. At this step, the articles that have not made it clear their conformity with the inclusion or exclusion criteria were included to be filtered out at a next step. Because of this phase, 251 studies were excluded, thus remaining 59 to be analyzed more closely.

Table 2 presents the amount of studies filtered in each step of selection process.

TABLE II. AMOUNT OF STUDIES FILTERED IN EACH STEP OF SELECTION PROCESS

Phase of Selection Process	Number of Studies
1. Databases Search	1291
2. Title Analysis	310
3. Abstract Analysis	59

D. Quality Assessment

In the quality assessment stage, works passed through a thorough critical analysis. In this stage, the complete studies were analyzed, instead of only the titles or abstracts. After this, the last studies that were considered uninteresting for the review were eliminated resulting in the final set of works. After the quality assessment, relevance grades were attributed to the remaining works. The relevance grades are going to be useful in the next stage.

Eight questions, based on [10] and [11], were used to help in the quality assessment. Those questions helped determining the relevance, rigor, and credibility of the work being analyzed. Of the eight, the first two are more useful to establish whether the work is relevant for the review, reason why they were the only ones used as exclusion criteria. The remaining questions are more useful for determining quality of the work, reason why they were used to grade the work according to quality. The questions were:

1. Does the study examine big data analysis as a way to aggregate value to citizens or rulers?
2. Is the study based on research - not merely on specialist's opinions?
3. Are the objectives of the study clearly stated?
4. Is the context of the study adequately described?
5. Were the methods for data gathering correctly used and described?
6. Was the research Project adequate to reach the research objectives?
7. Were the research results adequately validated?
8. Does the study contribute to research or citizens daily needs in any way?

Of the remaining 59 studies that were analyzed in the quality assessment stage, 40 passed to the stage of Data Extraction and Synthesis and were thus considered the primary studies. The quality assessment process will be presented in detail in the results Section along with the assessment of the 40 remaining studies.

IV. RESULTS

40 primary studies were identified [7], [12]–[50]. Each one deals with on a wide array of research topics and utilize a wide set of exploration models for each different scenario.

According to Ferraz et al. [51], it is possible to understand an intelligent city environment by interpreting six types of services: healthcare, transportation, education, security, government, and resources. Among the primary studies, we could find occurrences in four of the aforementioned groups. They were: healthcare, transportation, government and resources. The studies that did not fit into any of those groups were classified as being "general". The ones considered "general" are so for not relating to only one aspect of the city. They usually analyze the growth of data sources by focusing on the development of the city without taking into consideration the specific area they act upon.

A. Quantitative Analysis

The research process that was developed resulted in 40 primary studies. They were written by 124 authors linked to institutions based on 19 different countries, distributed on four continents, and were published between 2010 and 2013. In total, the authors identified 158 different keywords in their works.

In regards to the country of origin, most of the publications came from the United Stated (eight publications), followed by Germany, Greece, Italy, Australia and Spain (all with three works). China, Switzerland, Ireland and Korea was the third group (all with two publications). Each of the other remaining countries had only one publication. The large amounts of countries that have publications on the subject of big data show how widespread the topic is globally.

The most common keywords used in the remaining works with their respective frequency were: big data (10), smart city (8), data mining (4), social media (4), cloud computing (3), e-government (3), open data (3), open government (3), data fusion (2), innovation (2), personalized healthcare (2), twitter (2), ubiquitous computing (2). The first three keywords - big data, smart city and data mining - reflect precisely the theme of the research contained herein.

B. Quality Analysis

As was described in the previous Section, each of the primary studies was assessed according to eight quality criteria that relate to rigor and credibility as well as to relevance. If considered as a whole, these eight criteria provide a trustworthiness measure to the conclusions that a particular study can bring to the review. The classification for each of the criteria used a scale of positives and negatives.

Table 3 presents the results of the evaluation. Each row represents a primary work and the columns 'Q1' to 'Q8'

 ISBN: 978-1-61208-445-9

represent the 8 criteria defined by the questions used on quality assessment: Focus, Research, Objectives, Context, Data Gathering, Project, Validation, and Value, respectively. For each criteria, '1' represent the positive answer and '0' the negative one.

TABLE III. QUALITY ANALYSIS OF PRIMARY STUDIES

Study	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Total
[7]	1	1	0	1	0	1	0	1	5
[12]	1	1	0	1	0	0	0	1	4
[13]	1	1	1	1	0	1	0	1	6
[14]	1	1	1	1	0	1	0	1	6
[15]	1	1	1	1	1	1	1	1	8
[16]	1	1	1	1	1	1	0	1	7
[17]	1	1	0	1	1	1	1	1	7
[18]	1	1	0	1	1	0	0	1	5
[19]	1	1	0	1	0	0	0	1	4
[20]	1	1	1	1	0	1	0	1	6
[21]	1	1	1	1	1	1	0	1	7
[22]	1	1	1	1	0	1	0	1	6
[23]	1	1	1	1	1	1	1	1	8
[24]	1	1	1	1	0	1	1	1	7
[25]	1	1	0	1	1	1	1	1	7
[26]	1	1	1	1	1	1	1	1	8
[27]	1	1	1	1	1	1	1	1	8
[28]	1	1	0	1	0	0	0	1	4
[29]	1	1	1	1	1	0	0	1	6
[30]	1	1	1	1	1	1	1	1	8
[31]	1	1	1	1	1	0	0	1	6
[32]	1	1	1	1	1	1	1	1	8
[33]	1	1	1	1	1	0	0	1	6
[34]	1	1	1	1	1	0	1	1	7
[35]	1	1	1	1	0	0	1	1	6
[36]	1	1	1	1	1	1	1	1	8
[37]	1	1	0	1	1	0	1	1	6
[38]	1	1	1	1	0	1	0	1	6
[39]	1	1	1	1	1	1	1	1	8
[40]	1	1	1	1	1	1	0	1	7
[41]	1	1	1	1	0	1	0	1	6
[42]	1	1	0	1	0	1	1	1	6
[43]	1	1	1	1	1	1	1	1	8
[44]	1	1	1	1	1	1	0	1	7
[45]	1	1	0	1	1	1	1	1	7
[46]	1	1	0	1	0	0	0	1	4
[47]	1	1	0	1	1	1	1	1	7
[48]	1	1	1	1	1	1	1	1	8
[49]	1	1	1	1	0	1	1	1	7
[50]	1	1	0	1	1	1	1	1	7
Total	40	40	27	40	25	29	21	40	

All studies that were analyzed in this step had positive answers for questions 1 and 2 because, as previously stated in the research methodology part, these questions represent inclusion and exclusion criteria. Consequentially, all studies with negatives answers to at least one of these criteria were already removed during selection stage.

All studies that were analyzed provided information on the context of the work and contributed in some way to research and the field development. A small part, 13 of 40 works, did not provide an objective description of the problem. A number of 15 works did not show properly its procedures for data collection and 11 works did not present its procedures for data analysis. 19 theoretical works did not provide validations for the proposed models or research results. Another fraction, 10 of 40 studies, obtained the maximum score in quality analysis. The highest number of negative answers was found for study 4.

V. DISCUSSION

After the analysis and data extraction steps performed on the primary works, it was possible to identify some aspects relating big data solutions for urban environments. In the first place, it was possible to conclude that big data applied to urban environments is a very recent field of research. All primary works were published after 2010. Second, a great quantity of theoretical solution and conceptual models are available. This shows that the research field is very active but also that concrete efforts to implement what was theorized are rare.

The review also showed that there is a lack of standardization in government solutions. This happens because unification of data policy is a hard task due to each government possessing specific priorities. In this context, solutions based on collaborative data, despite showing a great potential, need sustainable business models to ensure that all involved parties can profit.

A. *Urban sectors that can be optimized through the use of big data.*

After the analysis of primary studies, solutions could be categorized into four main urban sectors: government, healthcare, transportation and resources.

In the government sector, problems faced by both government officials and citizens were identified. Issues faced by government officials have to do with efficiency and effectiveness of organizational sectors [18][20]. Those issues are accentuated by urban population growth. Issues faced by citizens relate to lack of transparency of public sectors [16][38]. Citizens could benefit from accessing to government data because it would give them insight into how public money is spent.

In the healthcare sector the problem relates to high costs paid by government and general population. This happens as a result of fraud crimes [39] and inadequacies in medical tools. If doctors had better tools, diagnoses would be more precise and faster [22][29]. This would, in turn, reduce the amount of times the patients needed to return for consultations, benefiting the healthcare sector as a whole.

The main problem identified in transportation sector is the intense traffic that exists in cities [26][32][43]. This calls for alternatives in transportation to be developed and for constant monitoring of the traffic conditions.

In resources, we could verify that the difficulty lies in management. For efficient distribution, it is necessary that distribution networks be constantly being monitored in search for flaws and to aid planning of infrastructure [14][15][44].

B. *Solution Models.*

The solutions found in the revisions can be classified under the following categories: conceptual models, sensors and services, social networks and visualization techniques, and pattern recognition.

 ISBN: 978-1-61208-445-9

The conceptual models consist in revisions of the topic along with the development of theoretical models of recommendations based on research and learned lessons. This was the category that had the highest amount of solutions during revision where we can highlight the healthcare and government groups. Altogether, 19 of the analyzed studies are framed in this group: [7], [12]–[14], [16], [18]–[22], [27]–[29], [31], [33], [38], [40], [44] and [48].

The solutions based on sensors and services aggregate data from various sensors (be they heterogeneous or not) and provision them to be used in the development of services that can aid government officials or citizens. We identified 5 works at this group: [23], [24], [26], [36] and [49]. This type of solution, due to them not specifying what type of sensors can be utilized, can be applied in various areas of the city. Due that they were classified as being "general".

Solutions in social networks analyze data already available to infer tendencies, events, or reception of public policies. 4 studies were categorized at this class: [25], [34], [41] and [47]. Since this type of data can be analyzed with many different objectives, this type of solution can be found mostly in the "general" category.

The solutions in pattern recognition entailed mostly the development of many data analysis, machine learning and visualization techniques. The 12 works remaining were grouped at this category: [15], [17], [30], [32], [35], [37], [39], [42], [43], [45], [46] and [50]. They intend to extract relevant information from raw data and to present them in a way that can be used as a decision support tool. Many of those solutions were applied to problems of finding better routes to help in city traffic and were consequentially classified under the "transportation" category.

C. Value created by Big Data solutions

The application of big data technologies brings several benefits to urban sectors. In the "resources" category, the development of measurement structures used in tandem with data extraction technologies helped enormously in the management of water and energy. These technologies provide a number of advantages, including lower measurement costs, resource waste reduction for customers, theft detection, increased reliability of supply methods and the possibility of custom pricing strategies.

The benefits for the healthcare category vary from diagnoses that are more efficient to reduction in costs of medical systems. Most of the technologies that were revised are used to provide a more detailed analysis of patients' health by taking into account personal and medical history, and similarities to other patients to reduce the amount of medical consultations. Other benefits that medical systems provide stem from the analysis of data to reduce frauds, and other kinds of waste.

In transportation, the analysis of traffic patterns allows a more effective investment and provides the citizen with information on routes.

In government, most research highlights benefits of big data technologies for both citizens and managers. To citizens, the open government initiatives help providing more transparency in government spending, as well as allowing innovative solutions to be developed due to the availability of city data. To managers the benefits lie mostly in improvements of analytical works to aid in decision-making and predictions on impacts on society.

D. Towards Big Data challenges.

The main challenge in working with big data refers to how to deal with the overwhelming amount, speed and variety of that data. The size, speed and variety attributes turn traditional data analysis models obsolete or insufficient. Besides the problem of size, speed and variety, our study found another set of challenges:

- Standardization: The lack of standards in gathering and provisioning of data and lack of standards in solution models makes reuse difficult. As a result, solutions become very specific which can hinder the development on the field.
- Information Security: Most solutions use personal or sensitive data of people or businesses. The assurance of privacy, integrity and availability of that data is of the outmost importance when the objective is raising quality of life.
- Reversal and recuperation: Given the large amount of data and the vulnerabilities inherent to big data solutions, it is important to assure the recuperation of data in the case when data is lost due to some unexpected event. More intelligent solutions than Backup and Restore need to be created due to the large amount of data.
- Acceptance of Change: Some sectors are more resistant to the adoptions of new technologies – or of change in general. This happens often in government where people are accustomed to bureaucracy and to certain traditional ways of doing things.
- Quality: Data used by big data solutions often lack quality due to inexistent or insufficient validation. This is a challenge due to data validation being difficult when the size of it is very large. Errors can happened during gathering, analyzing or provisioning of data.

VI. CONCLUSION

The objective of this review was to identify studies and solutions that propose solving urban problems with application of big data. In the search phase, 1291 works were found of which 40 were classified as primary studies after selection and quality criteria were applied.

The studies were first classified in according to focus of their solutions. A big number of theoretical works and conceptual models were found. This showed that concretely implemented and validated solutions are currently scarce. Many solutions based on sensors and services were also found. Those solutions, however, lacked standardization. Other works presented data analysis models in social networks and pattern recognition techniques.

 ISBN: 978-1-61208-445-9

In regards to aspects of the city that are more frequently targeted in studies, we found the following: government, healthcare, transportation, and resources. Distinct solutions were identified for each and potential benefits were presented. The main challenges in concretely applying the solutions that were found are privacy, standardization, data quality, willingness to change, and security.

A limitation of the current study is the potential bias that exists when only one researcher is responsible for deciding selection criteria and analyzing quality of the works. Even though measures were taken to keep the analysis impartial, such as defining search questions and protocol beforehand, more strict validation, especially in regards to selection criteria, would benefit future research efforts. Another limitation is the lack of an in depth analysis of the solutions proposed by each study. The main focus of this work was to find patterns in the use of big data solutions applied to urban environments to provide a general view of the current state of art.

As a future work, more effort could go into the analysis of the solutions to enable the development of standardization models.

REFERENCES

[1] L. M. A. Bettencourt, "The Uses of Big Data in Cities," 2013.

[2] W. Redmond, "The Big Bang: How the Big Data Explosion Is Changing the World," 2012. [Online]. Available: https://www.microsoft.com/en-us/news/features/2013/feb13/02-11bigdata.aspx. [retrieved: January, 2015] .

[3] J. K. Laurila, J. Blom, O. Dousse, D. Gatica-perez, O. Bornet, and M. Miettinen, "The Mobile Data Challenge : Big Data for Mobile Computing Research." Proceedings of the Workshop on the Nokia Mobile Data Challenge, in Conjunction with the 10th International Conference on Pervasive Computing, pp. 1–8, 2012.

[4] S. Dirks and M. Keeling, "A vision of smarter cities," IBM, 2009.

[5] K. N. Cukier and V. Mayer-Schoenberger, "The Rise of Big Data. How It's Changing the Way We Think About the World.," Foreign Affairs, 2013. [Online]. Available: http://www.foreignaffairs.com/articles/139104/kenneth-neil-cukier-and-viktor-mayer-schoenberger/the-rise-of-big-data. [retrieved: February, 2015].

[6] M. Friedenberg, "Why Big Data Means a Big Year for Hadoop," 2012. [Online]. Available: http://www.cio.com/article/698750/Why_Big_Data_Means_a_Big_Year_for_Hadoop. [retrieved: February, 2015].

[7] R. Kitchin, "The real-time city? Big data and smart urbanism", GeoJournal, Volume 79, Issue 1 , pp 1-14, Nov. 2013.

[8] M. S. Fox, "City Data : Big , Open and Linked," pp. 1–20, 2013.

[9] M. Batty et al., "Smart cities of the future", Eur. Phys. J. Spec. Top., vol. 214, no. 1, Dec. 2012, pp. 481–518.

[10] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering", Technical Report EBSE-2007-01, 2007.

[11] T. Dybå and T. Dingsøyr, "Empirical studies of agile software development: A systematic review," Inf. Softw. Technol., vol. 50, no. 9–10, Aug. 2008, pp. 833–859.

[12] R. Nambiar, R. Bhardwaj, A. Sethi, and R. Vargheese, "A look at challenges and opportunities of Big Data analytics in healthcare," Big Data, 2013 IEEE Int. Conf., Oct. 2013, pp. 17–22.

[13] D. Arribas-Bel, "Accidental, open and everywhere: Emerging data sources for the understanding of cities," Appl. Geogr., Oct. 2013, pp. 1–9.

[14] D. Alahakoon and X. Yu, "Advanced analytics for harnessing the power of smart meter big data," Intell. Energy Syst. (IWIES), 2013 IEEE Int. Work., Nov. 2013, pp. 40–45.

[15] K. a. Nguyen, R. a. Stewart, and H. Zhang, "An intelligent pattern recognition model to automate the categorisation of residential water end-use events," Environ. Model. Softw., vol. 47, Sep. 2013, pp. 108–127.

[16] G. Lee and Y. H. Kwak, "An Open Government Maturity Model for social media-based public engagement," Gov. Inf. Q., vol. 29, no. 4, Oct. 2012, pp. 492–503.

[17] W. Radl, S. Jäger, F. Mödritscher, and A. Komendera, "And Data for All : On the Validity and Usefulness of Open Government Data," Proc. 13th Int. Conf. Knowl. Manag. Knowl. Technol., 2013, pp. 6–9.

[18] J. Bertot and H. Choi, "Big data and e-government: issues, policies, and recommendations," Proc. 14th Annu. Int. Conf. Digit. Gov. Res., 2013, pp. 1–10.

[19] H. Moon and H. S. Cho, "Big Data and Policy Design for Data Sovereignty: A Case Study on Copyright and CCL in South Korea," Soc. Comput. (SocialCom), 2013 Int. Conf., Sep. 2013, pp. 1026–1029.

[20] R. C. Joseph, P. State, and N. A. Johnson, "Big Data and Transformational Government," IT Prof., vol. 15, 2013, pp. 43–48.

[21] I. Vilajosana and J. Llosa, "Bootstrapping smart cities through a self-sustainable model based on big data flows," Commun. Mag. IEEE, vol. 51, no. June, 2013, pp. 128–134.

[22] N. V Chawla and D. a Davis, "Bringing big data to personalized healthcare: a patient-centered framework.," J. Gen. Intern. Med., vol. 28 Suppl 3, Sep. 2013, pp. S660–5.

[23] C.-H. Lee et al., "Building a generic platform for big sensor data application," Big Data, 2013 IEEE Int. Conf., Oct. 2013, pp. 94–102.

[24] F. J. Villanueva, M. J. Santofimia, D. Villa, J. Barba, and J. C. Lopez, "Civitas: The Smart City Middleware, from Sensors to Big Data," Innov. Mob. Internet Serv. Ubiquitous Comput. (IMIS), 2013 Seventh Int. Conf., no. 3, Jul. 2013, pp. 445–450.

[25] S. K. Bista, S. Nepal, and C. Paris, "Data Abstraction and Visualisation in Next Step: Experiences from a Government Services Delivery Trial," Big Data (BigData Congr. 2013 IEEE Int. Congr., Jun. 2013, pp. 263–270.

[26] L. Calderoni, D. Maio, and S. Rovis, "Deploying a network of smart cameras for traffic monitoring on a 'city kernel,'" Expert Syst. Appl., vol. 41, no. 2, Feb. 2014, pp. 502–507.

[27] W. Maass and U. Varshney, "Design and evaluation of Ubiquitous Information Systems and use in healthcare," Decis. Support Syst., vol. 54, no. 1, Dec. 2012, pp. 597–609.

[28] A. J. Jara, Y. Bocchi, and D. Genoud, "Determining Human Dynamics through the Internet of Things," Web Intell. Intell. Agent Technol. (IAT), 2013 IEEE/WIC/ACM Int. Jt. Conf., Nov. 2013, pp. 109–113.

[29] P. Bamidis, "Enabling e-services based on affective exergaming, social media and the semantic web: A multitude of projects serving the citizen-centric vision for ICT in support of pHealth," in Bioinformatics and Bioengineering (BIBE), 2013 IEEE 13th International Conference on, 2013, pp. 1–6.

[30] C. Kaiser and A. Pozdnoukhov, "Enabling real-time city sensing with kernel stream oracles and MapReduce," Pervasive Mob. Comput., vol. 9, no. 5, Oct. 2013, pp. 708–721.

[31] K. Andreasson, J. Millard, and M. Snaprud, "Evolving e-government benchmarking to better cover technology development and emerging societal needs," Proc. 6th Int. Conf. Theory Pract. Electron. Gov. - ICEGOV '12, 2012, p. 430.

[32] E. Bouillet, B. Chen, C. Cooper, D. Dahlem, and O. Verscheure, "Fusing Traffic Sensor Data for Real-time Road Conditions," Proc. First Int. Work. Sens. Big Data Min. - SENSEMINE'13, 2013, pp. 1–6.

[33] S. Daniel, "geoSmartCity : geomatics contribution to the Smart City," Proc. 14th Annu. Int. Conf. Digit. Gov. Res., 2013, pp. 65–71.

[34] J. I. Criado, R. Sandoval-Almazan, and J. R. Gil-Garcia, "Government innovation through social media," Gov. Inf. Q., vol. 30, no. 4, Oct. 2013, pp. 319–326.

[35] A. Heise and F. Naumann, "Integrating open government data with stratosphere for more transparency," Web Semant. Sci. Serv. Agents World Wide Web, vol. 14, Jul. 2012, pp. 45–56.

[36] C. Dobre and F. Xhafa, "Intelligent services for Big Data science," Futur. Gener. Comput. Syst., Aug. 2013.

[37] X. Liu, F. Lu, H. Zhang, and P. Qiu, "Intersection delay estimation from floating car data via principal curves: a case study on Beijing's road network," Front. Earth Sci., vol. 7, no. 2, Jan. 2013, pp. 206–216.

 ISBN: 978-1-61208-445-9

[38] G. Puron-cid, J. R. Gil-garcia, L. F. Luna-reyes, and S. C. Martir, "IT-Enabled Policy Analysis : New Technologies , Sophisticated Analysis and Open Data for Better Government Decisions," Proc. 13th Annu. Int. Conf. Digit. Gov. Res., 2012, pp. 97–106.

[39] V. Chandola, S. R. Sukumar, and J. C. Schryver, "Knowledge discovery from massive healthcare claims data," Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '13, 2013, p. 1312.

[40] M. Bicking and M. Wimmer, "Need for computer-assisted qualitative data analysis in the strategic planning of e-government research," Proc. 11th Annu. Int. Digit. Gov. Res. Conf. Public Adm. Online Challenges Oppor., 2010, pp. 153–162.

[41] S. Nikolopoulos, S. Papadopoulos, and Y. Kompatsiaris, "Reality mining in urban space," in IISA 2013, 2013, pp. 1–4.

[42] A. Garzo, A. a. Benczur, C. I. Sidlo, D. Tahara, and E. F. Wyatt, "Real-time streaming mobility analytics," in 2013 IEEE International Conference on Big Data, 2013, pp. 697–702.

[43] A. Artikis, M. Weidlich, A. Gal, V. Kalogeraki, and D. Gunopulos, "Self-adaptive event recognition for intelligent transport management," Big Data, 2013 IEEE Int. Conf., Oct. 2013, pp. 319–325.

[44] D. S. Markovic, D. Zivkovic, I. Branovic, R. Popovic, and D. Cvetkovic, "Smart power grid and cloud computing," Renew. Sustain. Energy Rev., vol. 24, Aug. 2013, pp. 566–577.

[45] S. Hasan, C. M. Schneider, S. V. Ukkusuri, and M. C. González, "Spatiotemporal Patterns of Urban Human Mobility," J. Stat. Phys., vol. 151, no. 1–2, Dec. 2012, pp. 304–318.

[46] I. Carreras, S. Gabrielli, and D. Miorandi, "SUPERHUB: a user-centric perspective on sustainable urban mobility," Proc. 6th ACM Work. Next Gener. Mob. Comput. Dyn. Pers. Travel Plan., 2012, pp. 9–10.

[47] P. Metaxas and E. Mustafaraj, "The rise and the fall of a citizen reporter," Proc. 5th Annu. ACM Web Sci. Conf. - WebSci '13, 2013, pp. 248–257.

[48] J. H. Lee, M. G. Hancock, and M.-C. Hu, "Towards an effective framework for building smart cities: Lessons from Seoul and San Francisco," Technol. Forecast. Soc. Change, Oct. 2013.

[49] J. Yu and T. Zhu, "Towards Dynamic Resource Provisioning for Traffic Mining Service Cloud," Green Comput. Commun. (GreenCom), 2013 IEEE Internet Things, Aug. 2013, pp. 1296–1301.

[50] F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, and R. Trasarti, "Unveiling the complexity of human mobility by querying and mining massive trajectory data," VLDB J., vol. 20, no. 5, Jul. 2011, pp. 695–719.

[51] F. S. Ferraz et al., "Towards a Smart City Security Model Exploring Smart Cities Elements Based on Nowadays Solutions", no. c, 2013, pp. 546–550.

 ISBN: 978-1-61208-445-9

Big Data Analysis on Puerto Rico Testsite for Exploring Contamination Threats

Xiangyu Li, Leiming Yu and David Kaeli
Department of Electrical and Computer Engineering
Northeastern University
Boston, MA, USA
Email:{xili,ylm,kaeli}@ece.neu.edu

Yuanyuan Yao, Poguang Wang and Roger Giese
Department of Pharmaceutical Sciences and Barnett Institute, Bouve College
Northeastern University
Boston, MA, USA
Email:yao.yu@husky.neu.edu, {p.wang,r.giese}@neu.edu

Akram Alshawabkeh
Department of Civil and Environmental Engineering
Northeastern University
Boston, MA, USA
Email:aalsha@neu.edu

***Abstract*—In this paper, we present the use of Principal Component Analysis and customized software, to accelerate the spectral analysis of biological samples. The work is part of the mission of the National Institute of Environmental Health Sciences sponsored Puerto Rico Testsite for Exploring Contamination Threats Center, establishing linkages between environmental pollutants and preterm birth. This paper provides an overview of the data repository developed for the Center, and presents a use case analysis of biological sample data maintained in the database system.**

Keywords–non-targeted analysis; principal component analysis; environmental health.

I. INTRODUCTION

Since the early 1980's, the rate of preterm birth has been increasing worldwide [1]. Preterm birth is defined as a birth of an infant before 37 weeks of pregnancy. Preterm-related deaths accounted for 35% of all infant deaths in 2010. The rate of preterm birth in Puerto Rico is 50% higher than the average in the United States. There are a number of potential factors that can increase the probability of preterm birth. There is documented evidence that ties environmental factors to increased rates in preterm birth, as reported in several studies [2][3][4][5][6][7].

In the Puerto Rico Testsite for Exploring Contamination Threats (PROTECT) Center, we are working with a cohort of over 2000 women in northern Puerto Rico (presently 800 of the 2000 have been recruited), as part of a National Institute of Environmental Health Sciences (NIEHS) P42 Center project. We are studying linkages between a large number of potential contributing factors to premature birth. The goal is to establish a link between environmental pollution, particularly Chlorinated Volatile Organic Compounds (CVOCs) and phthalates, and birth outcomes. The project also considers the fate and transport (distribution, transport and transformation) of these pollutants into water supplies in northern Puerto Rico, as well as remediation methods.

This study is highly data driven, collecting and analyzing data from a wide range of sources, including:

- Environmental Samples and Measurements - soil samples, well and tap water samples, historical Environmental Protection Agency (EPA) data, soil samples, Superfund site data,
- Biological Samples - blood, urine, hair and placenta samples, and
- Human Subjects Information - medical history, reproductive health records, product use data surveys, and birth outcomes.

The data collected is carefully cleaned and maintained in fully-indexed relational database system. The PROTECT Database allows environmental health researchers to effectively tie any two entities present in the database together through relationships across two common indices:

1) Human Subject ID, or
2) Geographic Indexing System (GIS) coordinates.

To date, over 400 million data entries have been collected, cleaned and incorporated into the database. The repository includes a comprehensive data dictionary documenting the over 2457 data entities in the system.

In this paper, we provide an overview of our data management system, discuss our data management challenges, and present a compelling use case that evaluates the urine sample data present in the system. As part of PROTECT's research mission, selected chemicals (e.g., phthalates, bisphenol A) are measured in biological samples (i.e., in blood, urine, hair or placenta samples). Looking for the presence of a suspect chemical can be done utilizing a protocol, commonly referred to as *targeted analysis*. As an example of the power of our database system, we present results on our *non-targeted chemical analysis*, utilizing Principle Component Analysis (PCA) [8] to identify suspect chemicals present in the urine samples provided by the expectant mothers in our study.

The rest of this paper is organized as follows. Section II provides an overview of the PROTECT Database System. Section III presents an example of the richness of our data repository. Section IV covers preprocessing steps needed to precondition the data, and performs analysis of the data using Principal Component Analysis. Section V concludes the paper and outlines plans for future capabilities of our system.

II. PROTECT DATABASE

The PROTECT Database system has been built on top of EarthSoft's EQuIS software, and incorporates a Microsoft SQL Server as the database engine. A number of backend tools can work seamlessly with EQuIS, including ArcGIS [9], Surfer [10], and a number of statistical packages. Next, we discuss the different elements of the EQuIS system, which are used to maintain the large and diverse data repository maintained by PROTECT.

A. Data Repository

The Puerto Rico Testsite for Exploring Contamination Threats Center investigates exposure to environmental contamination in Puerto Rico and their causal effects with preterm birth. This program studies the high preterm birth phenomena

 ISBN: 978-1-61208-445-9

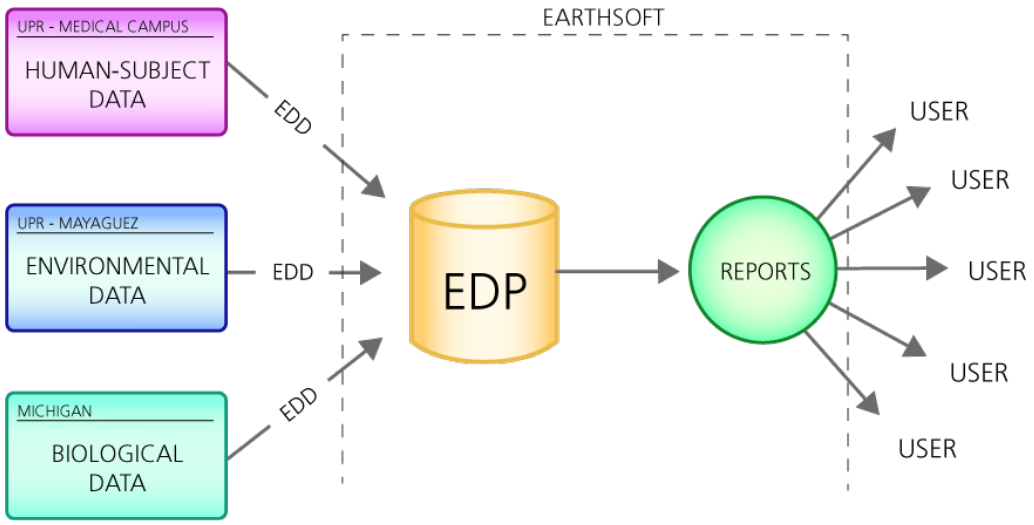


Figure 1. Data flow in the PROTECT database, where the Human Subject, Environmental and Biological Data are exported in Electrical Data Deliverable (EDD) format and verified using EQuIS Data Processor (EDP). After data are cleaned, they could be reported to different users with various permissions.

and transport of hazardous substances in karstic aquifers. In order to develop green remediation strategies to alleviate the exposure, Analytical data from various of sources are collected. Data collection includes Superfund sites, ground water, tap water, expectant mothers and birth outcome. The data repository generated supports a series of analysis activities, such as non-targeted chemical analysis, mechanistic toxicology, and targeted epidemiology. The PROTECT database delivers an efficient framework of *data management and modeling* across different disciplinary research domains.

In the current database system, data from human subjects, environmental sources and biological sampling have been collected and cleaned for further analysis. These entries exceed 400 million data points to date, as shown in Table I. More than 5 billion entries are anticipated to be housed in our system upon completion of the project. Since each data record is potentially related to adverse reproductive outcomes, understanding any correlations present across data sources is necessary. The underlying correlations could unveil important linkages between pollutants and birth outcomes. To find these linkages, machine learning techniques are applied during *data analysis*.

TABLE I. Present PROTECT database contents.

	Data Points (In millions)
Environmental	1.3
Human Subjects	1.5
Biological	400

Before incorporating any information into the database system, a careful data cleaning process is conducted. Each data export file consists of fields for different targeted analysis. For each field, the data type, format and nullability need to be verified. Its corresponding data value should stay within the range of pre-defined scope. Checking the dependencies between fields is also required. A comprehensive cleaning procedure pinpoints any corrupted data and avoids them leaking into the database. This procedure abides by the PROTECT Data Dictionary (containing detailed definitions of the 2457 different data entities in the system).

B. *EQuIS Professional*

In PROTECT database, we perform the automated cleaning by leveraging EarthSoft's EQuIS Professional [11] for the standalone development. Using EQuIS Professional, input data is first placed in an Electronic Data Deliverable (EDD) format, a format that is also supported by Microsoft Excel. The specific format of the EDD is essential for proper data checking. Each EDD entry defines the data type, range, constraints and dependencies of each individual field. The EDD format is customizable and typically includes three or four files: 1) format definition file, 2) custom handler, 3) enumeration file, and 4) reference values. The format definition file holds the definition and mapping for every field. The custom handler provides the detailed rules that apply to each data format. Common operations check for the specific data range, null data format and specific data types allowed/supported for each data field. The enumeration file is optional, and requires the EQuIS Data Processor (EDP) to execute a set of lookup values. The reference value file is needed when users need to check reference values remotely [12].

EDDs are checked according to the constraints defined in the data dictionary. First, the EDD format is verified using the format definition file in XML Schema Definition format. Whenever any conflicts occur, the corresponding fields will be highlighted by leveraging the custom handler coded in visual basic script. Additional error messages could be added to the script to facilitate the debugging process. The other two files are also necessary to make sure each field comply with the listed rules and mapping schemes. Since conflicts still exist in the EDDs, the data input are returned to the submitting project.

Once all errors are resolved, the record can be committed, and the repository is updated. Figure 1 describes the data flow of our database system. The database is frequently backed up

 ISBN: 978-1-61208-445-9

and provides the flexibility for users to produce customized reports.

C. EQuIS Enterprise

EQuIS Enterprise provides web-based access to the PROTECT data, which suits the distributed development. Online access is critical since the PROTECT team is distributed across Puerto Rico, Massachusetts, Michigan and West Virginia. Instead of managing data locally, EQuIs Enterprise automates the workflow using web-based applications.

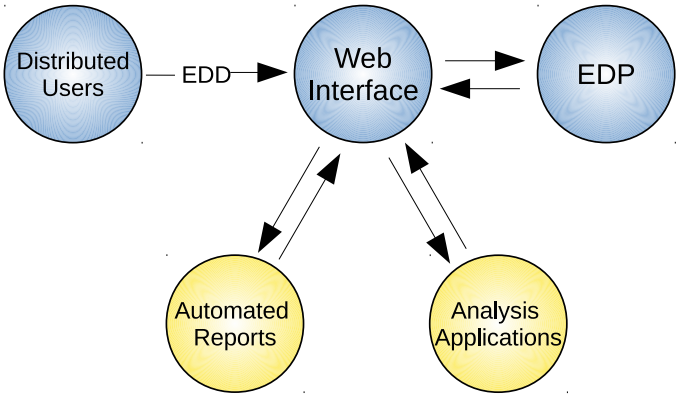


Figure 2. Data flow of EQuIS Enterprise. Distributed users can upload EDDs through the web interface where EDP is used and produce customized report accordingly.

Figure 2 describes the workflow we applied for the PROTECT database frontend. EDDs can be processed through a web interface. Users can receive status notification through File Transfer Protocol (FTP), email or web widgets. The web interface provided by Enterprise can produce standard or customized reports. It also provides the researcher with visualization of their data through Geographic Indexing System tools.

III. URINE STUDY

Next, we provide a use case of the PROTECT Database system. The goal is not to answer any particular question, but instead to demonstrate the richness and the challenges associated with the research project.

In our case study, we focus on the biological data that represents a majority of the data maintained in the database, as shown in Table I. The biological data contains samples from urine, blood, hair and placenta. In this example, we will analyze the urine samples present in the biological data, which holds 99% of the total biological data volume. The goal is to demonstrate our ability to perform big data analysis and modeling, which is supported by the PROTECT database system.

To perform non-targeted chemical analysis of the urine samples, we are employing a matrix-assisted laser desorption ionization time of flight/time of flight mass spectrometer (MALDI-TOF/TOF-MS). This instrument can detect many urine metabolites that we are looking for, while also giving us clues to which other chemicals are present.

Each urine sample extract is first separated into 240 droplets by Ultra Performance Liquid Chromatography, based on the analyte polarity Chromatography [13]. Each of the sets of samples is then mixed with a chemical reagent, which is sensitive to the specific laser wavelength in the mass spectrometer. The laser transforms the analyte and reagent in gas phase ions, and the detector registers the analyte (in units of m/z, the mass-to-charge ratio) and the corresponding signal intensity for each droplet. Each observed analyte is then subjected to the fragmentation analysis (TOF/TOF-MS) to check if it belongs to a specific metabolites group, for example, sulfate conjugate.

During the analysis stage, two problems associated with the MALDI-TOF/TOF-MS system (model 5800 from AB SCIEX is used) are encountered. The first problem is that the data exports (t2d) for MarkerView (the proprietary software used on the system) are stored in a binary format, which are not easily decoded in order to analyze metabolite data measured [14]. Hence, our own methodology to decode the binaries before porting them to the databas is developed, as shown in Figure 3. At first, the t2d file is decoded into mzXML format, and then the mzXML file is exported to a text file using ProteoWizard [15][16]. Customized python programs are written to extract the peak and intensity values from the text file and exported them to our database in the EDD format.

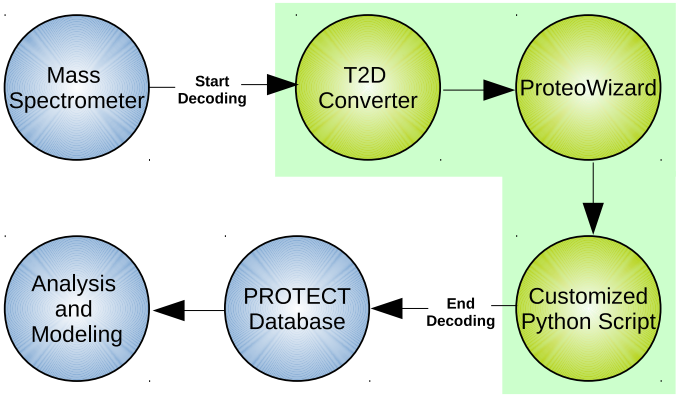


Figure 3. Urine study methodology: 1) process raw data via Mass Spectrometer 2) decode the binary using T2D converter 3) read deciphered binaries using ProteoWizard 4) extract data using Python 5) data cleaning via the database 6) support data analysis and modeling.

The second problem is the limited processing capabilities of the proprietary software that comes with the Mass Spectrometer, MarkerView. To identify metabolomic features present in the data, Principal Component Analysis (PCA) is applied on the data [8]. The software could only compute PCA on a few sets of metabolomic features (5000 highest intensity mass-to-charge features are selected), but it took 20 minutes to compute these on just 6 data sets. The processing spends 10 minutes on input processing and peak picking, and 10 minutes on computing PCA. This throughput becomes a barrier to discovery, especially when the data to be processed requires different scaling and weighting factors. A huge increase in PCA processing time when processing larger data sets is also observed. Exploring different scaling and weighting factors is severely impacted by the limited computation power of MarkerView. To accelerate the processing, our own PCA modeling scheme is developed, which is described in the next section.

 ISBN: 978-1-61208-445-9

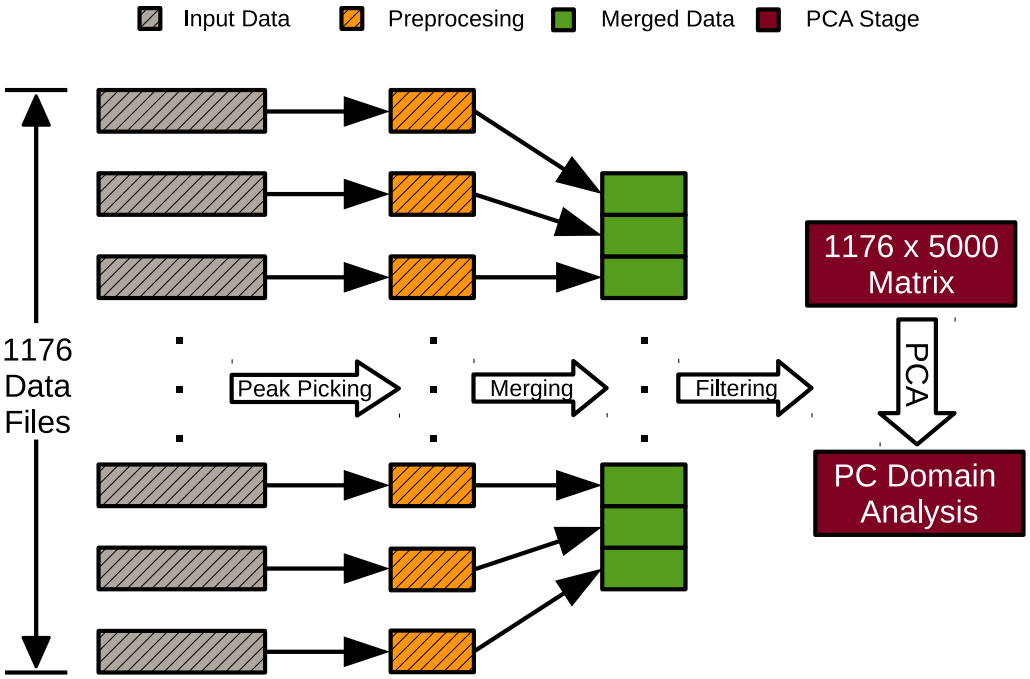


Figure 4. Urine analysis workflow. Peak picking reduces the feature dimension at the preprocessing stage. Filtered peaks are formed into a matrix for PCA.

IV. PCA OF URINE SAMPLES

The existing urine samples contain 400 million data points to be characterized, which does not include the associated TOF/TOF-MS data yet. The execution time will become unacceptable if all of them are to be directly analyzed, especially given that more than 80 urine samples are expected to be processed in the future. In addition, exploring all possible features across a large input data set may not lead to a proper classification. There may be too much correlation between the different features. Therefore, PCA is used to represent such a large number of data points with fewer uncorrelated features, while retaining the important variations present in the original dataset. In this section, we discuss how PCA is applied to analyze the patterns present in the urine samples.

A. Principal Component Analysis

PCA is often used to reduce the dimensionality of large multi-variate datasets. It transforms a set of possibly correlated samples into a set of linearly uncorrelated data points called principal components [8]. To be more specific, PCA takes as input a numerical matrix, where the rows of the matrix correspond to different input samples, and where the columns correspond to different dimensions of each sample. Then the input matrix is transformed orthogonally into the principal component domain, where each principal component is a linear combination of the input dimensions.

Principal components are sorted in a decreasing order, which captures the variance present in the input samples. These variances after orthogonal transformation are also known as eigenvalues, that describe the scaling factor of the orientation given a linear transformation. The first few principal components, usually two or three, can represent most of the variations present in the data. Therefore, PCA can significantly reduce the complexity in the data, without eliminating patterns and outliers in the data. Using PCA can greatly simplify and accelerate big data analytics by reducing the feature dimensionality.

B. Input Data

In this preliminary case, urine sample data sets collected from 6 project participants are analyzed, where each sample includes 196 urine mass spectra, resulting in 1176 urine sample files in total. In each of those urine sample files, 130K mass to charge ratio (analyte molecular weight) and associated intensities are recorded. Accurate mass to charge ratio, less than 10 ppm with internal calibration, combined with metabolome database search and the associated TOF/TOF analysis could potentially reveal the chemicals present in the urine sample. In order to detect patterns in these urine mass spectra, we need to cluster the urine samples in a 130K-dimensional space, where each dimension represents the mass of a potential chemical. However, the data dimensionality is too large for us to analyze. On the other hand the majority of data points are baseline measurements. Hence, we filter out a lot of the data before applying PCA. We describe this process in following sections.

C. Pre-processing Stage

When the mass spectrum data are decoded in the MarkerView's binaries, the 130K measurements are decoded in each file. Each measurement is separated by 0.007 *Da* (the unified atomic mass unit), which indicates the mass on an atomic or molecular scale.

The data is first pre-processed where only peaks are picked. A peak represents an analyte with the local maximum intensity value above the preselected signal to noise ratio threshold (20 in our current analysis) within a mass range . The peak-picking process eliminates the noise and baseline points that have too low intensities. The *MALDIquant* package for Quantitative

 ISBN: 978-1-61208-445-9

Analysis of Mass Spectrometry Data in R (a free software programming language for statistical computing) is applied to facilitate the process [17][18]. As illustrated in Figure 4, the peak-picking process normally reduces the size of the data from 130K data points to 300 data points. After the peak-picking process, the number of analytes is reduced from the original number of 153M to 353K. These selected peaks are then merged into a single data file and screened in the filtering stage.

D. Filtering Stage

After completing the previous stage, the 300 peaks from each input file are merged into one single peak list file. The peak width is defined as 0.01 Da, in order to filter out small experimental variation for different sample runs, where the same analyte from different data files are assigned to slightly different mass-to-charge value. As a result, analyte peaks within 0.01 *Da* are all assigned to the same mass-to-charge value, which is associated with the highest intensity within this mass range. The remaining analytes are sorted again by its intensity so that only the top 5000 chemicals with the highest intensities are kept for PCA processing. Python is used to implement the filtering stage. A sparse matrix is generated in the filtering stage, as shown in Figure 4, where a row corresponds to a spectrum data file ID and a column correspond to an analyte. A non-zero element *(i,j)* represents the intensity of an analyte *j* from spectrum data file *i*, while an empty cell *(m,n)* shows that file *m* does not contribute that particular analyte or intensity to the matrix.

E. PCA Stage

The filtering stage generates a 1176x5000 sparse matrix, where an analyte in a certain column could be identified in a urine spectrum file specified by the row number. PCA then is applied to transform the matrix into principal components. Each principal component is a linear combination of the 5000 variables in that row. In most dimensionality reduction problems, the first and second principal components usually capture 70-90% of the variance in the data. Therefore the first two principal components are considered. *scikit-learn*, a machine learning toolbox in Python, is used to carry out PCA analysis on the filtered data [19]. We discuss our preliminary clustering results of the chemicals in next section.

F. Clustering Results Analysis

Among the current 6 urine samples, 5 samples are from Puerto Rico and 1 sample is from Boston. Due to security reasons, each human subject is identified using the particular naming scheme, location plus study id.

The 1176 urine spectra defined in 5000 mass-to-charge feature space is projected onto a two-dimensional plane, where the first two principal components (PC1 and PC2) are used for the x-axis and the y-axis, respectively. The resulting projection is presented in Figure 5, showing some clustering on the right side of the plot. This clustered region is associated with the early and later droplets from the Ultra Performance Liquid Chromatography, which contain common background noise. On the left side of the plot, where higher PC1 variance appears, the red dots (PR3_10_14) and blue dots (PR7027) are grouped in clusters. It implies a possible environmental and metabolomic difference.

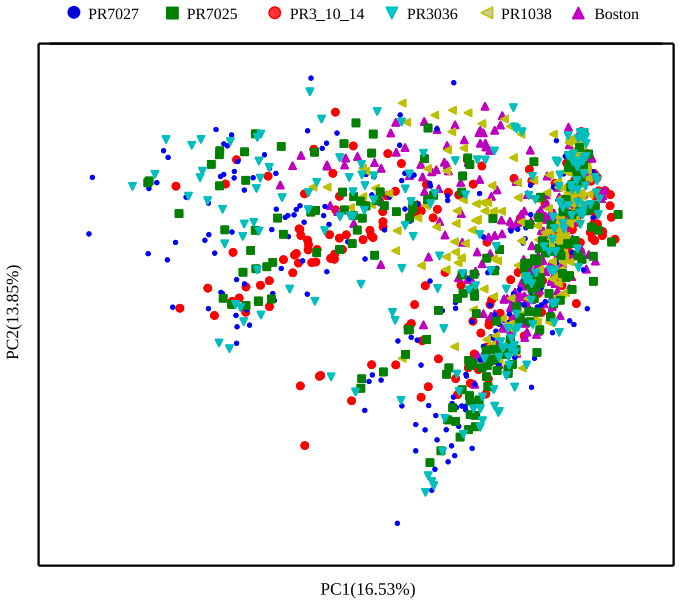


Figure 5. PCA results on six urine samples using PC1 and PC2.

Our current results show that by combining PC1 and PC2, only 30% of the total variation is captured without applying any weighting or scaling. However, if the distribution of these 1176 mass spectra are visulized using other principal components, the variance captured is only smaller with the spectra distributions being similar. Non-negligible differences can be observed using alternative principal components. As additional urine samples are included, more insights are to be learned from this statistical analysis. Our automation process has been able to accelerate urine analysis from 20 mins to 12 seconds by leveraging *MALDIquant* and *scikit-learn*. A 100x speedup has been effectively achieved by this analysis framework, which will enable more extensive analysis of our data sets as the number of participants in our study group grows.

V. Conclusion and Future Work

This paper provided an overview of the data repository developed for the PROTECT Center, and presented a use case analysis of biological sample data maintained in the database system.

Establishing linkages between environmental pollutants and preterm birth can produce important health benefits for both expectant mothers and their babies. In the PROTECT Center, we are collecting detailed information from expectant mothers during their entire 9 month pregnancy, as well as after delivery. The data collected in this study encompasses a wide variety of data types, ranging from soil and water composition to birth outcomes. We expect to be managing billions of data points over the next few years.

In the PROTECT database system, we have developed an efficient framework to handle efficient big data cleaning and entry. We have built our architecture on top of EQuIS Professional to handle the data cleaning and provide online and secure access through EQuIS Enterprise.

To demonstrate the utility of our data, as well as describe our challenges when working with such large data, we present

 ISBN: 978-1-61208-445-9

results of a preliminary urine sample study. A sophisticated decoding scheme has been proposed to extract the spectra information from mass spectrometer measurements. With the proposed PCA tools, we are capable of reducing the analysis time by 100-fold, while maintaining the same accuracy as the proprietary software. Since a customized toolset has been developed, we can directly interface the metabolite database to apply data weights. Given that there are approximately 50 associated Mass Spectrometry (MS^2) spectra for each original mass spectrum, the actual data set size we will be working with is much bigger than 1176x136K. From these MS^2 spectra, we can learn more about the detected chemicals. This additional information can then be fed into our PCA analysis, to allow this huge data set to become more manageable.

Even though a significant speedup has been achieved, there is still room for further acceleration. Currently, we are working on a standalone CPU optimization using the open-source software. However, we plan to also leverage the benefits of parallel accelerators, such as Graphics Processing Units. Prior work utilizing accelerators has provided 9x to 130x speedup on a variety of data analysis domains [20]. We are developing a GPU-accelerated PCA implementation to further accelerate our analysis tools.

In our current PCA analysis of the urine samples, the first two dominant principal components captured 30% of the variation of all the variables. As part of our ongoing research, we would like to consider additional principal components to cover more feature variations.

ACKNOWLEDGMENT

This work is supported in part by Award Number P42ES017198 from the National Institute of Environmental Health Sciences. In addition, we would like to thank EarthSoft for their generous support.

REFERENCES

[1] H. Blencowe et al., "National regional and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications," The Lancet, vol. 379, 2012, pp. 2162–2171.

[2] J. D. Meeker et al., "Urinary phthalate metabolites in relation to preterm birth in mexico city," Environ. Health Perspect., vol. 117, 2009, pp. 1587–1592.

[3] D. Cantonwine et al., "Bisphenol a exposure in Mexico City and Risk of prematurity: a pilot nested case control study," Environ. Health, vol. 9, 2010, pp. 62–68.

[4] A. P. Mucha et al., "Abstract: Association between pbde exposure and preterm birth," in 10 Annual Workshop on Brominated Flame Retardants, Victoria, BC Canada, 2008, p. 42.

[5] K. Tsukimori et al., "Long-term effects of polychlorinated biphenyls and dioxins on pregnancy outcomes in women affected by the yusho incident," Eniron. Health. Perspect., vol. 116, 2008, pp. 626–630.

[6] P. Z. Ruckart, F. J. Bove, and M. Maslia, "Evaluation of contaminated drinking water and preterm birth, small for gestational age, and birth weight at Marine Corps Base Camp Lejeune, North Carolina: a cross-sectional study," Environ. Health, vol. 13, 2014, pp. 1–10.

[7] J. D. Meeker, "Exposure to environmental endocrine disruptors and child development," Arch. Pediatr. Adolesc. Med., vol. 166, 2012, pp. 952–958.

[8] I. Jolliffe, Principal component analysis. Wiley Online Library, 2005.

[9] J. McCoy, K. Johnston, and E. systems research institute, Using ArcGIS spatial analyst: GIS by ESRI. Environmental Systems Research Institute, 2001.

[10] D. Keckler, "Surfer for windows-users guide.-golden software," Inc., Golden, CO, 1995.

[11] EarthSoft, "EQuIS Professional," http://www.earthsoft.com/products/professional/, 2015 (accessed March 1, 2015).

[12] EarthSoft, "EarthSoft: Standalone EQuIS Data Processor (EDP) User Guide," http://www.dec.ny.gov/docs/remediation_hudson_pdf/edpuserguide.pdf, 2008.

[13] R. Plumb et al., "Ultra-performance liquid chromatography coupled to quadrupole-orthogonal time-of-flight mass spectrometry," Rapid Communications in Mass Spectrometry, vol. 18, no. 19, 2004, pp. 2331–2337.

[14] M. Sugimoto, M. Kawakami, M. Robert, T. Soga, and M. Tomita, "Bioinformatics tools for mass spectroscopy-based metabolomic data processing and analysis," Curr. Bioinform., vol. 7, 2012, pp. 96–4108.

[15] Y. Gao, "T2D converter," http://www.pepchem.org/, 2013 (accessed March 1, 2015).

[16] D. Kessner, M. Chambers, R. Burke, D. Agus, and P. Mallick, "Proteowizard: open source software for rapid proteomics tools development," Bioinformatics, vol. 24, no. 21, 2008, pp. 2534–2536.

[17] S. Gibb and K. Strimmer, "Maldiquant: a versatile r package for the analysis of mass spectrometry data," Bioinformatics, vol. 28, no. 17, 2012, pp. 2270–2271.

[18] J. Tuimala and A. Kallio, "R, programming language," Encyclopedia of Systems Biology, 2013, pp. 1809–1811.

[19] F. Pedregosa et al., "Scikit-learn: Machine learning in python," The Journal of Machine Learning Research, vol. 12, 2011, pp. 2825–2830.

[20] J. Nickolls and W. J. Dally, "The gpu computing era," IEEE micro, vol. 30, no. 2, 2010, pp. 56–69.

 ISBN: 978-1-61208-445-9

Towards Big Business Process Mining

Badr Omair, Ahmed Emam
Information Systems
King Saud University, KSU
Riyadh, Saudi Arabia
e-mail:balomair@gmail.com, aemam@ksu.edu.sa

***Abstract*—Nowadays, the topic of Big Data has received much attention from researchers. Because Big Data provides excellent analytics on a broad spectrum of data, data science is currently emerging as an interesting scientific discipline as a field of study comprehensively identified with the accumulation, administration, and examination of data. Business process mining is a process-centric technique focused on the mining of data and, for this reason, Big Data can be a big help for business process mining. In this paper, we will review Big Data and business process mining and present a model for mapping between Big Data characteristics and business process mining techniques. The mapping model has discovered that Big Data can help business process mining in different areas and open the door for more help.**

Keywords-business process mining; Big Data; business process; smart business processes.

I. INTRODUCTION

Business process mining techniques are very useful for any organization in any field. The power of these techniques comes from the fact that they are based on fact-based data [1]. Most of the time, it is very difficult to know all the facts of a given situation. One of the most important reasons for this is the inability of current technology to host huge volumes of data [2]. In some fields, like health care, seismology, astronomy and finance, archive data is regularly deleted to save on storage space. This data usually contains much important information and so business process techniques are unable to work well without it. Even in fields that do not generate huge data volumes, they will, in the long run, have increasingly large event logs, which will eventually create performance problems. Machines, sensors, and surveillance devices generate huge amounts of data that are also frequently deleted due to the incapability of saving such a large volume of information. We are now approaching the Internet of things era. Business process mining techniques must be able to handle the huge amount of data logs these devices create in order to be able to manage the facts and correlate them efficiently, or they will not be able to work successfully. In addition to the volume problem, currently, data is restricted to the structured data type only in process mining techniques [3], but there are also lots of facts available in semi and unstructured data types. For these reasons, some efforts are being made to solve these issues by using emerging Big Data technologies. These efforts are scattered, disorganized and do not cover comprehensive views, and so they do not take full advantage of Big Data technologies for process mining techniques. In this paper, we will present a model that maps every characteristic of Big Data to process mining techniques. This model proves that Big Data can contribute a great deal to business process mining. Comprehensive discovery, accurate prediction abilities, visibility, efficiency, and flexibility represent the main advantages of the mapping model. This mapping model will open the door for using Big Data at different levels for business process mining and will maximize its use in business process mining. Future efforts should now be aimed at investigating each help track in detail to ascertain a practical implementation, especially as Big Data technologies are beginning to mature and become more available. The rest of this paper is organized as follows: Section II describes Big Data definitions and characteristics, Big Data versus business intelligence, the Big Data life cycle, Big Data opportunities and challenges, and Big Data architecture, business process and business process mining; Section III describes Big Data and business process mining tools; Section IV reviews the previous work in using Big Data for business process mining; Section V will present the mapping model; and Section VI shows the conclusion and future work.

II. BACKGROUND

A. Big Data definitions and characteristics

Due to the boom of generated data, Big Data has become a strong concern for many organizations. Current technologies, such as Relational Database Management System (RDBMS), data warehouses, and business intelligence, do not have the capabilities to support the Big Data business goal, which is to enable organizations to create actionable business insights in a rapidly changing environment [4].

In a book by Hurwitz et al. [5], Big Data is defined as the ability to deal with an enormous volume of divergent data, at the speed, and inside the time span to permit continuous investigation and response.

In Savitz's [6] gave a more detailed definition as: "Big Data are high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization".

Watson [2] provided the most comprehensive definition of Big Data as: "Big data is a term that is used to describe data that is high volume, high velocity, and/or high variety; requires new technologies and techniques to capture, store, and analyze it; and is used to enhance decision making, provide insight and discovery, and support and optimize processes". Big data is considered as the most recent era in the development of decision support data management [7].

Thus, from these definitions of Big Data, we can see that the main characteristics of Big Data are: volume,

 ISBN: 978-1-61208-445-9

velocity, and variety, also known as the 3Vs. The term volume alludes to the huge size of the data set, velocity demonstrates the speed of data movement, and variety refers to the different data types [8].

B. Big Data versus business intelligence

Chen et al. [9] stated that business intelligence and analytics fall into three categories: first, Database Management System (DBMS) structured content, utilizing traditional analytic tools via data warehousing, Extract, Transform, and Load (ETL), Online Analytical Processing (OLAP) and data mining; second, web-based and unstructured content, utilizing tools in information retrieval, opinion mining, question answering, web analytics, social media analytics, social network analysis, and spatial-temporal analysis; and third, mobile and sensor-based content, utilizing tools in location-awareness analysis, person-centered analysis, context-relevant analysis, and mobile visualization.

The main differences between Big Data and a data warehouse are: Big Data is stored in a distributed file system rather than on a central server; the Big Data format can be unstructured or semi-structured rather than only structured data; Big Data includes real-time data as well as offline data [10]; Big Data comes from a variety of sources, including new data sources such as web data, social media, device logs and mobile data; and finally, Big Data is mainly used for predictive analysis. Hu et al. [11] made a comparison between big data and traditional data as depicted in TABLE I.

C. Big Data life cycle

There are some applied Big Data life cycles, but Hu et al. [11] stated that the most common Big Data life cycle consists of four phases, the first of which is data generation. . Second is data acquisition, which refers to the process of obtaining information and is subdivided into data collection, data transmission, and data pre-processing. Third is data storage, which refers to the persistent storage and management of large-scale datasets. Fourth is data analysis, which leverages analytical methods or tools to inspect, transform, and model data to extract value. Figure 1 depicts these phases with exemplar technologies.

D. Big Data opportunities and challenges

Big Data has a wide variety of sources, from traditional transactional processing to processes that incorporate Internet data (e.g., clickstream and social networking), research data (e.g., reviews and industry reports), area information (e.g., cell phone information and geospatial information), pictures (e.g., observation and satellites), store network information (e.g., Electronic Data Interchange (EDI) and merchant inventories), and device data (e.g., sensors and Radio Frequency Identification (RFID) gadgets) which allow for many Big Data applications [12]. For example, by integrating a customer's profile with his habits, location and interests, which can be obtained from the Internet (e.g., Google, Twitter, Facebook, LinkedIn and other social media), we can build tailored products, conduct customer sentiment analysis and provide targeted services [13]. Integrating the huge amount of patients' historical data with medicine manufacturing data can help to provide personalized medication and gain insights into genetic and environmental causes of diseases [14]. Support companies can provide better services and improved troubleshooting, cyber security and uptime for their customers by monitoring recorded data from smart meters and machine logs [10].

In addition to retail, manufacturing, banking and finance, and healthcare, C. L. Philip Chen & Zhang [15], stated that Big Data applications also lie in many scientific disciplines, such as astronomy, atmospheric science, medicine, genomics, biology, biogeochemistry and other complex and interdisciplinary scientific researches. They also conducted a survey about Big Data benefits in the business sector, as depicted in Figure 2.

Big Data security and privacy are a big challenge. In addition, inconsistency, incompleteness, scalability and timeliness of the data are also challenges [16][17]. In a predictive analytics study in 2013 at The Data Warehousing Institute (TDWI) [18], a survey was made regarding the challenges of Big Data. It found that data integration complexity, architecting Big Data systems, lack of skills or staff, dealing with real-time data, poor data quality, and data privacy are the most important challenges.

Watson [2] stated that the clear business need, strong management support, the alignment of analytics strategies with business need, skilled people, the right analytics tools and fact-based decision making are the keys to Big Data analytics success.

E. Big Data architecture

The most famous software framework for processing Big Data is Apache Hadoop. For this reason, we will take it as an example to demonstrate Big Data architecture. The Apache Hadoop software library is a massive computing framework consisting of several modules including Hadoop Distributed File System (HDFS), Hadoop MapReduce, HBase, and Chukwa [11] as depicted in Figure 3. TABLE will briefly describe these modules. Chan [4] illustrated architecture for Big Data analytics and investigated Big Data innovations that incorporated Not only SQL (NoSQL) databases, HDFS and MapReduce. He examined running batch and real-time analytics using Hadoop. Its bidirectional association with conventional data warehouses and data mining analytics is depicted in Figure 4.

F. Business process

"A business process instance represents a concrete case in the operational business of a company, consisting of activity instances" [19]. Business processes have become progressively vital in numerous enterprises because they define the method for developing value and distributing it to customers [20]. Business processes are the key drivers behind three critical success factors—cost, quality, and time [21]. There are some well-known business processes methodologies, such as Six Sigma, Lean, BP Trends, Hammer and Rummler–Brache. The main elements of these methodologies are: (1) management and leadership, for describing how the processes will be managed, (2) process improvement, for describing the improvement process steps, (3) measurement, for describing how the processes will be measured, (4) learning, for describing training needs, (5) alignment with organizational priorities, for prioritizing process improvement projects, (6) continuous improvement, for determine how are

 ISBN: 978-1-61208-445-9

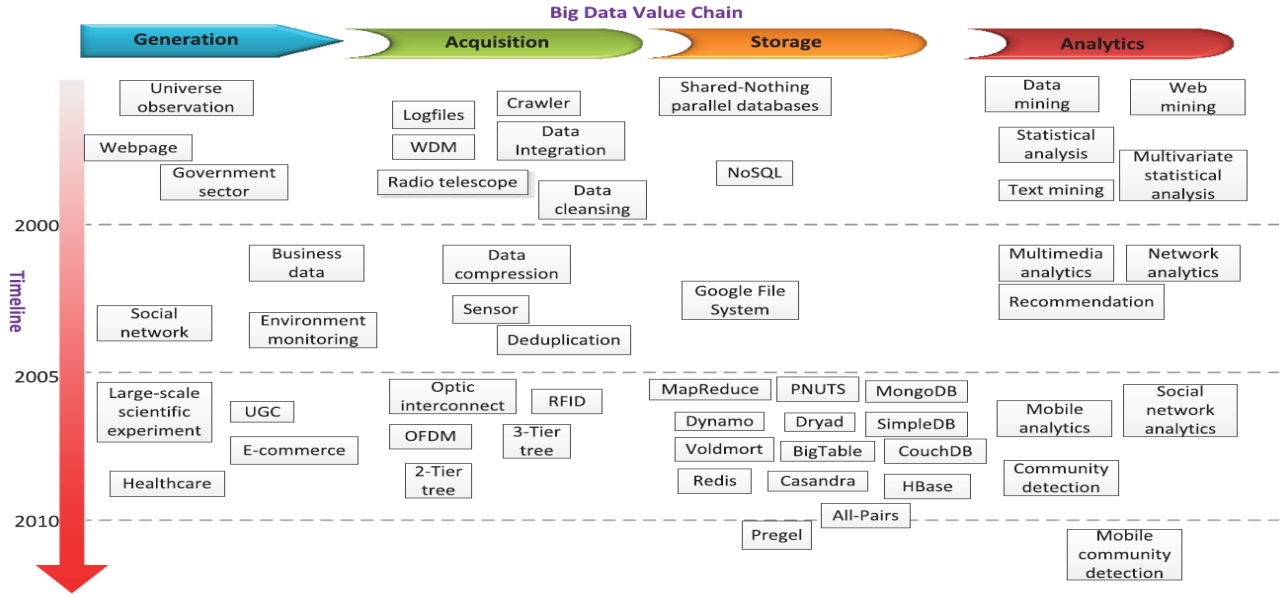


Figure 1. Big Data life cycle technology map [11].

TABLE I. BIG DATA VERSUS TRADITIONAL DATA [11].

	Traditional Data	Big Data
Volume	gigabyte	Constantly updated (terabyte or petabyte currently)
Generated rate	per hour/day	more rapid
Structure	structured	semi-structured or unstructured
Data source	centralized	fully distributed
Data integration	easy	difficult
Data store	RDBMS	HDFS, NoSQL
Access	interactive	batch or near-time

processes be monitored and by whom, (7) technology, for specifying Business Process Management (BPM) tools, (8) common practices, for specifying the organization standards, (9) change management, for describing the change management process for the business processes [22]. BPM covers how we study, distinguish, change and screen business methodologies to guarantee that they run easily and can be enhanced over the long term. It involves a constant assessment of existing processes and identification of approaches to improve them, in order to bring about global organizational enhancement [23].

G. *Business process mining*

W. M. P. van der Aalst [24] stated that there is currently a missing link between business processes and the real processes with information systems. Process mining has arisen as a new scientific discipline to provide a link between process models and event data [25]. Simeonova [26] defined process mining as techniques that help to find, screen and enhance genuine procedures by concentrating learning from event logs. Data is gathered from assorted types of systems and examined to identify deviations from standard processes and see where the bottlenecks are. Process mining is based on fact-based data and starts with an analysis of data, followed by the creation of a process model. This differs from the classic way of first modeling a control flow and then adding data to it. For example, in navigation devices, there is a link between the current reality and the models; they are not a static map, but a dynamic one that we use every day for understanding the locations of traffic jams, listening to directions, and estimating the remaining journey time. For this reason, using process mining together with traditional process documenting techniques will give more accurate results, as shown in Figure 5. Companies often use process intelligence, mining or analytics, and apply a variety of statistical and artificial intelligence techniques to measure and analyze process-related data [27]. The three types of Business Process Analysis (BPA) are validation, verification, and performance, which all require collecting and storing large volumes of process and event data [28]. In the following section, we will describe the well-known tools of Big Data and process mining.

III. TOOLS OF BIG DATA AND BUSINESS PROCESS MINING

A. *Big Data tools*

Watson [2] stated that the criteria to choose the right Big Data platform are: the applications that use the platform; the capabilities of processing the volume, velocity and variety of data; real-time or batch processing; people skills; and finally the implementation cost. As previously mentioned, Apache Hadoop is the most famous software framework for processing big data. It has the capabilities to process large amounts of data across potentially massively parallel clusters of servers (for example, Yahoo! has over 42,000 servers in its Hadoop installation) [2]. Apache Hadoop consists of client machines and clusters of loosely-coupled commodity servers. Hadoop has two main components: first, HDFS, which is the storage system and distributes data files over large server clusters; and second, MapReduce is the distributed processing framework for parallel processing of large data sets that distribute computing jobs to each server in the cluster and collects the results. There are three major categories of machines in HDFS: client machines, master nodes and slave nodes [29]. The client machines load data

 ISBN: 978-1-61208-445-9

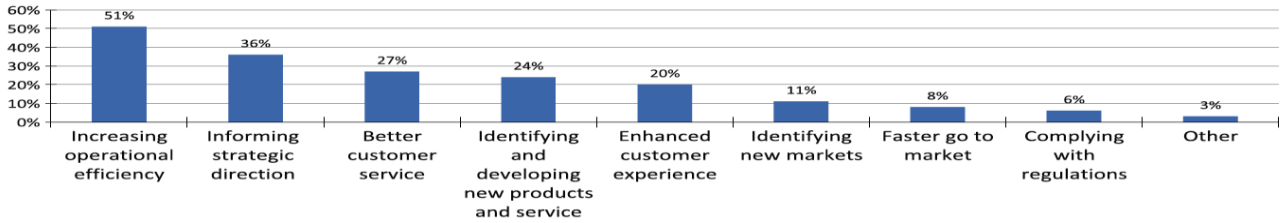


Figure 2. Operational Big Data opportunities [15].

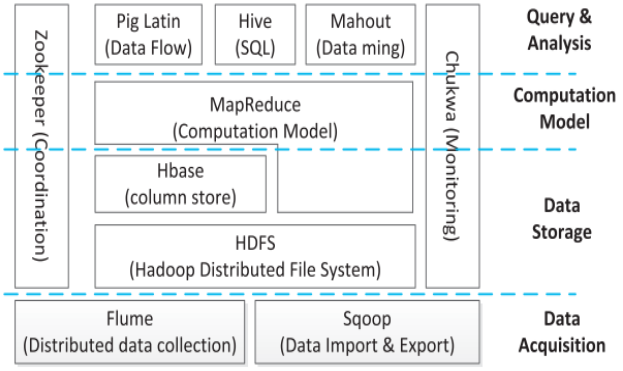


Figure 3. Hadoop architecture [11].

TABLE II. SUMMARY OF A HADOOP MODULE [11].

Function	Module	Description
Data Acquisition	Flume	Data collection from disparate sources to a centralized store
	Sqoop	Data import and export between structured stores and Hadoop
Data Storage	HDFS	Distributed file system
	HBase	Column-based data store
Computation	MapReduce	Group-aggregation computation framework
Query & Analysis	Pig Latin	SQL-like language for data flow tasks
	Hive	SQL-like language for data query
	Mahout	Data mining library
Management	Zokeeper	Service configuration, synchronization, etc.
	Chukwa	System monitoring

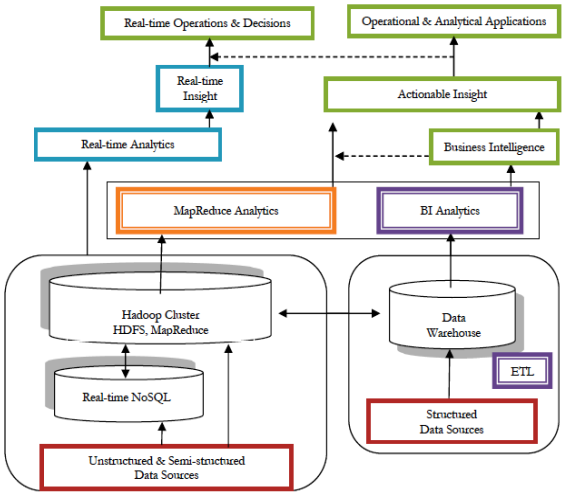


Figure 4. Big Data analytics architecture [4].

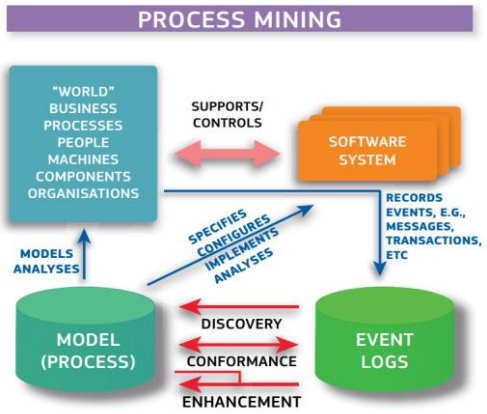


Figure 5. Process mining [25].

into servers and retrieve results. The master nodes have two types: HDFS nodes (name nodes), which are responsible for keeping the directory of all files in the HDFS file system, and MapReduce (job tracks) nodes, which are responsible for assigning MapReduce tasks to slave nodes.

B. *Business process mining tools*

W. M. P. van der Aalst [1] stated that there are three main techniques in BPM, as depicted in Figure 6. The first is the discovery technique, which takes an event log and produces a model without using a priori information. The second is the conformance technique, where an existing process model is compared to an event log of the same process. The third is the enhancement technique, which covers or improves an existing process model using information about the actual process recorded in an event log.

The most commonly used tools for process mining are ProM and Disco. ProM is a generic open-source framework for implementing process. It is very powerful but complicated. On the other hand, Disco is a commercial product and is very easy to use, but it lacks some of the process mining techniques. In the following section, we will show the most important effort made for using Big Data with business process mining.

IV. PREVIOUS WORK OF USING BIG DATA TO CONDUCT BUSINESS PROCESS MINING

Traditional Business Intelligence (BI) and Decision Support System (DSS) tools require something more than

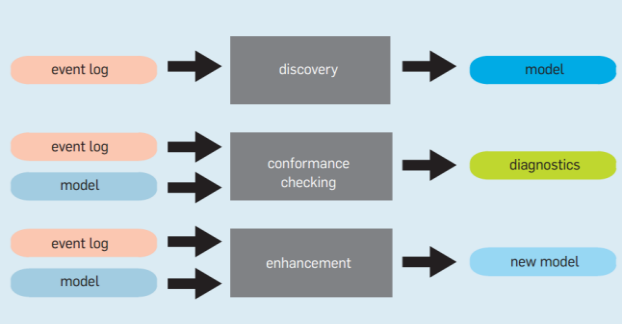


Figure 6. Types of process mining techniques in terms of input and output [1].

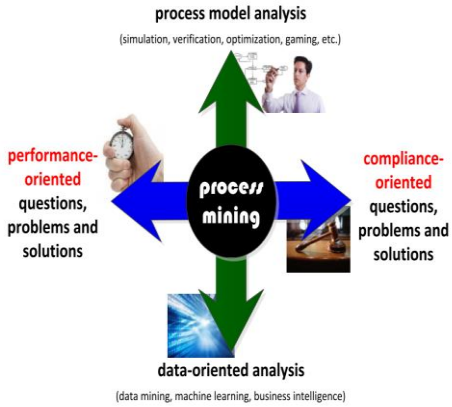


Figure 7. Process mining positions [31].

the use of mere historical data and rudimentary analysis tools to be able to predict future actions, identify trends or discover new business opportunities [30]. Real-time, low latency monitoring and analysis of business events for decision-making are required, but difficult to achieve [20]. It is a challenge to turn lots of event data ("Big Data") into valuable insights related to performance and compliance. Process mining with Big Data can do more than automate process discovery based on event data; it can also be used to answer a wide variety of performance and compliance questions in a unified and integrated manner, as shown in Figure 7. Transaction data and sensor data will enable new process mining applications to replace traditional analysis based on hand-made models [25]. Therefore, we should focus on a wide variety of different event data for mining. Process mining with Big Data will enable us to develop business processes that follow the right path in each situation.

Vera-Baquero et al. [32] introduced a business process improvement methodology for overcoming processing time and data size limitations by integrating process improvement with Big Data-based DSSs. The methodology is explained in Figure 8, and consists of five phases. The first phase is the definition phase, and it intends to not only identify and represent the business process that has a significant value to the organization, but also to have clear insight into the strategic management of the enterprise and a good understanding of the business goals being pursued. The definition phase involves four steps: (1) determine the scope and boundaries of the global business process (cross-organizational); (2) identify operational flaws within each single organization, including interactions between operational units (interdepartmental); (3) identify the level of detail that the global business process will be broken down into (level of sub-activities); and (4) develop process and activity tables. The configuration phase comes next and intends to prepare the analytical environment for receiving structural event data from the operational systems that will feed the DSS for later analysis. It includes identifying software boundaries and interdepartmental processes within business nodes, the selection of an event data format, the determination of instance correlations, and implementation of software listeners, along with a selection of metric and their threshold values. Next, the execution phase starts to capture the operational data and send business event data to the DSS. Finally, the control phase monitors and analyzes the outcomes of the DSS, and the diagnosis phase identifies deficiencies and weaknesses in the business processes identified in the definition phase. We use visualization to identify hotspots, or re-run event streams in a simulation mode in order to perform root cause analysis, among others. In the following section, we will present the mapping model, which map the Big Data characteristics to business process techniques.

V. MAPPING BIG DATA CHARACTERISTICS TO BUSINESS PROCESS TECHNIQUES

In Figure 9, we have presented the 3Vs characteristics of Big Data and showed how each of them can affect process mining techniques. The 3Vs characteristics of Big Data are volume, variety, and velocity, and they are considered the most important characteristics of Big Data as previously explained in this paper. These characteristics are mapped to process mining techniques of process discovery, process enhancement, and conformance checking; they are also explained previously in this paper. This mapping is very useful to discover how Big Data can help in process mining. We will explain this mapping in detail by describing each help track as follows:

1) Adequateness for discovery: in some application domains like the healthcare and finance sectors, it is impossible to preserve data for more than a year due to the huge volume of data generated from their systems. The same thing happens for data logs that are generated by machines and sensors. For this reason, this data is frequently deleted, even it is very important, in order to save on storage space. However, with existing Big Data technologies, we can save this data and take advantage of saving it for the long term. In process discovery techniques, we depend on event logs to discover the business process model. So, a bigger event log will ensure that we cover almost all the cases that may happen in the system. The data volume will increase the accuracy of the process discovery technique.

2) Adequateness for prediction: there are very useful probability techniques that can applied to the event logs for prediction, for example, what-if analysis, and the decision tree that can be applied to predict the path of a business process. For this reason, the more samples we have, the more we can improve the probability techniques. In addition, prediction reports will be improved with the volume of data, such as estimating the completion date.

 ISBN: 978-1-61208-445-9

3) Visibility: Event logs cannot include all the desired facts for the business process model. Big Data technologies can help us to cover all the facts by looking at extra data that was previously impossible to analyze. Big Data can help us to look for semi and unstructured data, like social media, images, emails and competitors' web sites, to feed the business process model with useful data.

4) Flexibility for efficiency: with increasing numbers of devices that are connected to the Internet, and the advent of the Internet of things era, it is very important to take advantage of these devices as much as we can. Because Big Data technologies can give us the ability analyze data in real-time by handling the huge data logs of these devices, this will help us to use these devices to increase the efficiency of the business process model. These devices can be used to extract the data from inside or outside the organization. For example, the business process model path can be changed in real-time as a quick response to use the logs of some meters like workload and inventory balance.

5) Flexibility for conformance: similar to flexibility for efficiency, flexibility for conformance can be achieved by enabling the business process model path to be changed automatically in real-time as a quick response to generated data logs from conformance devices. Temperature meters and surveillance are examples of such devices.

VI. CONCLUSION AND FUTURE WORK

As we have seen in this paper, there are some un-coordinated efforts existing for discovering Big Data applications for business process mining, but they are not yet sufficient for covering the whole picture. For this reason, we have provided a model for mapping the 3Vs characteristics of Big Data to process mining techniques. The model will help us to bring all the applicable benefits of Big Data into business process mining. Adequateness for discovery, adequateness for prediction, visibility, flexibility for efficiency, and flexibility for conformance are the main help tracks shown in the model. In future work, these tracks should be studied in detail to ascertain practical help for business process mining using Big Data technologies.

ACKNOWLEDGMENT

I would like to thank KSU-CCIS Ph.D. committee for their unlimited support.

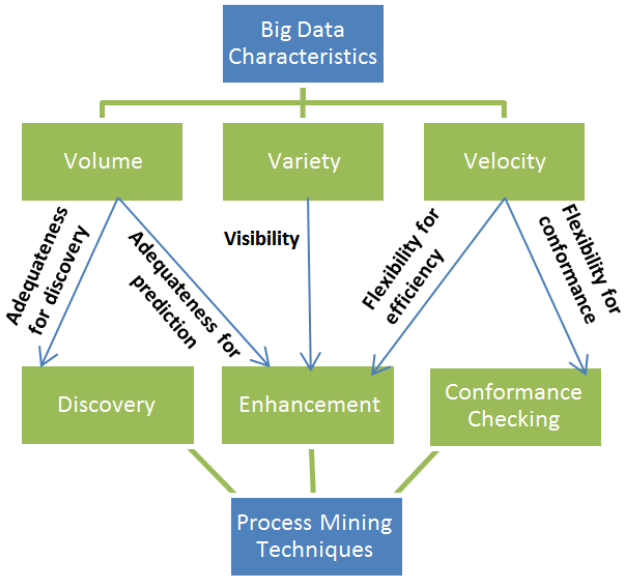


Figure 9. Mapping between Big Data characteristics and business process mining techniques.

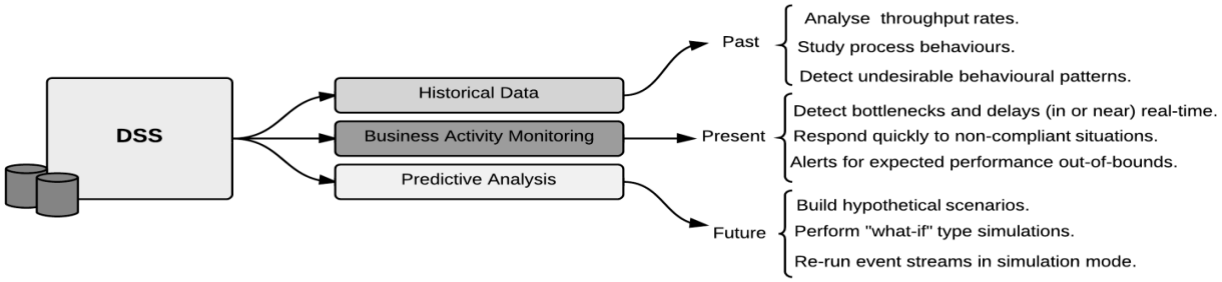


Figure 8. Business Process Analytics for Different Dimensions [32].

REFERENCES

[1] W. M. P. van der Aalst, "Process Mining," *Communications of the ACM*, Springer Berlin Heidelberg, 2012.

[2] H. J. Watson, "Tutorial: Big data analytics: Concepts, technologies, and applications," *Commun. Assoc. Inf. Syst.*, vol. 34, 2014, pp. 1247–1268.

[3] A. Rozinat, "How Big Data Relates to Process Mining – And How It Doesn't — Flux Capacitor," 2014. [Online]. Available: http://fluxicon.com/blog/2011/12/how-big-data-relates-to-process-mining-and-how-it-doesnt/. [Accessed: 13-Dec-2014].

[4] J. Chan, "An architecture for Big Data analytics," *Commun. IIMA*, vol. 13, 2013, pp. 1.

[5] J. Hurwitz, A. Nugent, F. Halper, and M. Kaufman, *Big data for dummies*. For Dummies, 2013, pp. 15.

 ISBN: 978-1-61208-445-9

[6] E. Savitz, “Gartner: 10 Critical Tech Trends For The Next Five Years,” 2012. [Online]. Available: http://www.forbes.com/sites/ericsavitz/2012/10/22/gartner-10-critical-tech-trends-for-the-next-five-years/. [Accessed: 29-Dec-2014].

[7] H. J. Watson and O. Marjanovic, “Big Data: The Fourth Data Management Generation,” *Bus. Intell. J.*, vol. 18, no. 3, 2013, pp. 4–8.

[8] C. L. Philip Chen and C.-Y. Zhang, “Data-intensive applications, challenges, techniques and technologies: A survey on Big Data,” *Inf. Sci. (Ny).*, vol. 275, 2014, pp. 314–347.

[9] H. Chen, R. H. L. Chiang, and V. C. Storey, “Business intelligence and analytics: from Big Data to big impact,” *MIS Q.*, vol. 36, 2012, pp. 1165–1188.

[10] N. Sawant and H. Shah, *Big Data Application Architecture Q&A: A Problem-Solution Approach*. Apress, 2013, pp. 2.

[11] H. Hu, Y. Wen, T. Chua, and X. Li, “Towards Scalable Systems for Big Data Analytics: A Technology Tutorial,” *IEEE Access*, 2014.

[12] M. Minelli, M. Chambers, and A. Dhiraj, *Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today’s Businesses*. 2013, pp. 5–8.

[13] K. V. N. Rajesh, “Big Data Analytics: Applications and Benefits.,” *IUP J. Inf. Technol.*, vol. 9, no. 4, 2013, pp. 41–51.

[14] P. Chandarana and M. Vijayalakshmi, “Big Data analytics frameworks,” in *Circuits, Systems, Communication and Information Technology Applications (CSCITA), 2014 International Conference on*, 2014, pp. 430–434.

[15] C. L. Philip Chen and C. Y. Zhang, “Data-intensive applications, challenges, techniques and technologies: A survey on Big Data,” *Inf. Sci. (Ny).*, vol. 275, Aug. 2014, pp. 314–347.

[16] D. Agrawal et al., “Challenges and Opportunities with Big Data 2011-1,” *Purdue e-Pubs*, 2011.

[17] R. T. Kouzes, G. A. Anderson, S. T. Elbert, I. Gorton, and D. K. Gracio, “The Changing Paradigm of Data-Intensive Computing.,” *IEEE Comput.*, vol. 42, no. 1, 2009, pp. 26–34.

[18] “TDWI predictive analytics study,” *TDWI*, 2013. [Online]. Available: http://www.business2community.com/big-data/drive-real-time-revenue-world-big-data-01109279. [Accessed: 31-Dec-2014].

[19] M. Weske, *Business process management: concepts, languages, architectures*. Springer Science & Business Media, 2012, pp. 7.

[20] A. Vera-Baquero, R. Colomo-Palacios, and O. Molloy, “Business process analytics using a big data approach,” *IT Prof.*, vol. 15, 2013, pp. 29–35.

[21] S. Adam, N. Riegel, J. Doerr, O. Uenalan, and D. Kerkow, “From business processes to software services and vice versa - An improved transition through service-oriented requirements engineering,” in *Journal of software: Evolution and Process*, 2012, vol. 24, pp. 237–258.

[22] S. Sweet, “Which BPM Methodology is Best for Us?,” *bpminstitute*. [Online]. Available: http://www.bpminstitute.org/resources/articles/which-bpm-methodology-best-us. [Accessed: 17-Feb-2015].

[23] “AIIM - What is Business Process Management?,” *AIIM*. [Online]. Available: http://www.aiim.org/What-is-BPM-Business-Process-Management. [Accessed: 30-Dec-2014].

[24] W. M. P. van der Aalst, “Business Process Management: A Comprehensive Survey,” *ISRN Softw. Eng.*, 2013, pp. 1–37.

[25] W. M. P. van der Aalst, *Process Mining: Discovery, Conformance and Enhancement of Business Processes*, vol. 136. 2011, pp. 80.

[26] D. Simeonova, “BIG DATA AND PROCESS MINING,” *DG DIGIT*, 2014. [Online]. Available: https://ec.europa.eu/digit-ict/sites/digit-ict/files/ictinterview.pdf. [Accessed: 13-Dec-2014].

[27] C. Janiesch, M. Matzner, and O. Müller, “Beyond process monitoring: a proof-of-concept of event-driven business activity management,” *Business Process Management Journal*, vol. 18. 2012, pp. 625–643.

[28] M. P. van der A. Wil, W. Mathias, and W. Guido, “Advanced Topics In Workflow Management: Issues, Requirements, And Solutions,” *J. Integr. Des. Process Sci.*, vol. 7, 2003, pp. 49–77.

[29] B. Hedlund, “Understanding Hadoop Clusters and the Network,” 2011. [Online]. Available: http://bradhedlund.com/2011/09/10/understanding-hadoop-clusters-and-the-network/. [Accessed: 29-Dec-2014].

[30] A. V. Baquero and O. Molloy, “A Framework to Support Business Process Analytics.,” in *KMIS*, 2012, pp. 321–332.

[31] W. van der Aalst, “‘ Mine your own business ’: using process mining to turn big data into real value,” *Proc. Eur. Conf. Inf. Syst.*, 2013, pp. 1–9.

[32] A. Vera-Baquero, R. Colomo-Palacios, and O. Molloy, “Towards a Process to Guide Big Data Based Decision Support Systems for Business Processes,” *Procedia Technol.*, vol. 16, 2014, pp. 11–21.

 ISBN: 978-1-61208-445-9

Forecasting Hourly Electricity Demand in Egypt

Using Double Seasonal Autoregressive Integrated Moving Average Model

Mohamed A. Ismail Alyaa R. Zahran Eman M. Abd El-Metaal
Statistics department
Faculty of economics and political science, Cairo University
Giza, Egypt
e-mail: mismail@feps.edu.eg azahran@feps.edu.eg eman_mahmoud@feps.edu.eg

***Abstract*—Egypt has faced a major problem in balancing electricity produced and electricity consumed at any time in the day. Therefore, short-term forecasts are required for controlling and scheduling of electric power system. Electricity demand series has more than one seasonal pattern. Double seasonality of the electricity demand series in many countries have considered. Double seasonality pattern of Egyptian electricity demand has not been investigated before. For the first time, different double seasonal autoregressive integrated moving average (DSARIMA) models are estimated for forecasting Egyptian electricity demand using maximum likelihood method. $DSARIMA\ (3,0,1)\ (1,\ 1,\ 1)_{24}\ (2,\ 1,3)_{168}$ model is selected based on Schwartz Bayesian Criterion (SBC). In addition, empirical results indicated the accuracy of the forecasts produced by this model for different time horizon.**

Keywords-multiple seasonality pattern; post-sample forecasts; Double Seasonal ARIMA models.

I. INTRODUCTION

Electricity is one of the ordinary life necessities, and a major driving force for economic growth and development. The unstorable nature of electricity means that the supply of electricity must be always available to satisfy the growing demand. Therefore, electricity utilities throughout the world have given a remarkable interest for forecasting electricity demand. Decision makers around the world widely use energy demand forecasting as one of the most important policy making tools. An accurate hourly demand forecasting up to one day ahead is a vital process in electricity industry planning. It is critical for nations in order to balance electricity produced and electricity consumed at any time in the day, to increase the reliability of power supply, to minimize costs and to provide correct decisions for future development [1][2] .

Electricity demand is mainly influenced by seasonal effects (daily and weekly cycles, calendar holidays). A within-day seasonal cycle is apparent if similarity of the hourly demand from one day to the next exists, while a within-week seasonal cycle is apparent if similarity of the daily demand exists week after week. Therefore, using a forecasting method that is able to capture both seasonal patterns (daily and weekly) is mandatory.

Seasonal Autoregressive integrated moving average (SARIMA) model is used for time series data with single seasonal pattern. However, SARIMA model can be extended to cope with multiple seasonal cycles [3]. SARIMA model that includes two cycles is known by DSARIMA model.

DSARIMA was used by many authors for forecasting electricity demand. In [4], DSARIMA model was used for forecasting hourly electricity load in England and Wales and was compared with single seasonal Holt-Winters exponential smoothing method and with a double seasonal Holt-Winters exponential smoothing method. The forecasts produced by the DSARIMA model were well and outperformed those from Holt-Winters exponential smoothing method that considered only single seasonal pattern but were outperformed by those from double seasonal Holt-Winters method. In [5], six forecasting methods including DSARIMA, double seasonal exponential smoothing, a method based on the principal component analysis (PCA), artificial neural network (ANN), a random walk model and a seasonal version of the random walk were considered for forecasting hourly electricity demand for the state of Rio de Janeiro in Brazil and half-hourly electricity demand for England and Wales. Among those forecasting methods, DSARIMA model was competitive and performed well for Rio data and England and Wales data. The same pervious methods were also applied on ten European countries and the same conclusion was reached [6].

In a recent study [7], the DSARIMA model was investigated for forecasting the double seasonal (daily and weekly) Malaysian electricity demand series. In (2011), it was compared with SARIMA model and concluded that DSARIMA model outperformed the SARIMA model [8]. Therefore, our target is to investigate DSARIMA model in forecasting Egyptian electricity demand series.

The rest of this paper is organized as follows. Section II describes the Egyptian electricity demand series. Section III describes DSARIMA model. Section IV discusses the results. The conclusion and future work close the article.

 ISBN: 978-1-61208-445-9

II. EGYPTIAN ELECTRICITY DEMAND SERIES

The Egyptian electricity demand series consists of hourly time series data of Egyptian electricity demand measured in Megawatt (MW) for a one year starting on Saturday 7 January 2012 and ending on Friday 28 December 2012. All the data is used to estimate parameters except for the last 4 weeks that are put aside to evaluate post-sample accuracy of forecasts.

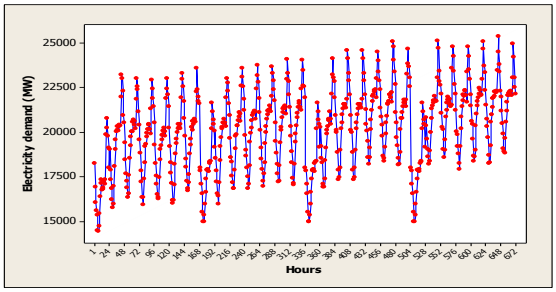


Figure 1. Time plot for the Egyptian electricity demand series from Friday 1 June 2012 to Thursday 28 June 2012

Figure 1 shows a time series plot covering the period from Friday 1 June 2012 to Thursday 28 June 2012. In the figure, the first day is represented by hours from hour 1 till hour 24, while from hour 24 till hour 48 represents the second day and so on. Figure 1 shows a within-day seasonal cycle and a within-week seasonal cycle. A within-day seasonal cycle is apparent from the similarity of the demand from one day to the next. A within-week seasonal cycle is also apparent from comparing the demand on a certain day of different weeks. It is clear that the weekdays show similar patterns of demand, while the weekend days, which have the lowest peak of electricity demand, have a different electricity demand pattern.

III. DSARIMA MODELS

A multiplicative SARIMA model has introduced by [9] to analyze single seasonal pattern time series data. SARIMA model is denoted as ARIMA $(p, d, q)\,(P, D, Q)_s$ where p and P are the orders of nonseasonal and seasonal autoregressive terms, respectively, d and D are the orders of nonseasonal and seasonal differencing, respectively, while q and Q are the orders of nonseasonal and seasonal moving average terms and s is the seasonal period. SARIMA can be expressed as

$$\emptyset_p(B)\Phi_P(B^s)\,\nabla^d\,\nabla_s^D\,y_t = \theta_q(B)\,\Theta_Q(B^s)\,\varepsilon_t, \qquad (1)$$

where ∇^d and ∇_s^D are the nonseasonal and seasonal differencing operators, respectively ; B is the backward shift operator; $\{\varepsilon_t\}$ is a white noise process with mean zero and a constant variance; $\emptyset_p(B)$ *and* $\Phi_P(B^s)$, are polynomials of order p and P, represent the nonseasonal and seasonal autoregressive terms, respectively; $\theta_q(B)$ and $\Theta_Q(B^s)$ are polynomials of order q and Q, represent the nonseasonal and seasonal moving average terms, respectively.

SARIMA model can be extended for DSARIMA model [3]. DSARIMA model has been expressed by [4] to capture two seasonality cycles (within-day and the within-week seasonal cycles). The multiplicative DSARIMA model, which is denoted as ARIMA $(p, d, q)\,(P_1, D_1, Q_1)_{s_1}\,(P_2, D_2, Q_2)_{s_2}$, can be written as

$$\emptyset_p(B)\,\Phi_{P_1}(B^{s_1})\,\Omega_{P_2}(B^{s_2})\,\nabla^d\,\nabla_{s_1}^{D_1}\,\nabla_{s_2}^{D_2}\,y_t = \theta_q(B)\,\Theta_{Q_1}(B^{s_1})\,\Psi_{Q_2}(B^{s_2})\,\varepsilon_t, \qquad (2)$$

where $\nabla_{s_1}^{D_1}$ is the daily seasonal differencing operator; $\nabla_{s_2}^{D_2}$ is the weekly seasonal differencing operator; $s_1 and\ s_2$ are the two seasonal periods which are 24 and 168, respectively in our Egyptian electricity demand data set series; $\Phi_{P_1}(B^{s_1})$ and $\Omega_{P_2}(B^{s_2})$ are polynomials of orders P_1 and P_2, respectively; and $\Theta_{Q_1}(B^{s_1})$ and $\Psi_{Q_2}(B^{s_2})$ are moving average polynomials of orders Q_1 and Q_2, respectively.

Stationarity of the Egyptian electricity demand series is investigated in the next section. If the data series is nonstationary, suitable differences and/or transformations should be made to render stationarity.

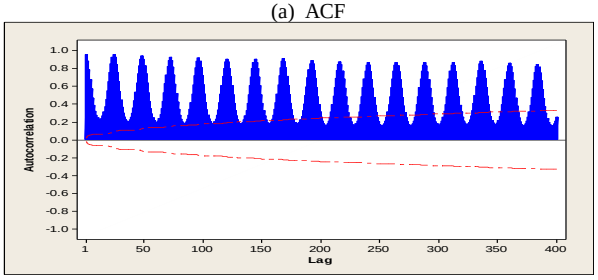
IV. EMPIRICAL RESULTS

Different DSARIMA models are used for forecasting Egyptian electricity demand. At first, in order to identify a suitable DSARIMA model and check whether the series is stationary, we plotted the autocorrelation function (ACF) and the partial autocorrelation function (PACF) of the Egyptian electricity demand series. Figure 2 shows the ACF and PACF of the hourly Egyptian electricity demand series. It is clear from the ACF the presence of daily seasonal pattern. A daily seasonal differencing $(D_1 = 1, s_1 = 24)$ is considered to convert the nonstationary series that results from the daily pattern into a stationary series. Plotting the ACF and PACF after the daily seasonal differencing, Figure 3 shows another seasonal pattern which is the weekly seasonal pattern; therefore the weekly seasonal differencing $(D_2 = 1, s_2 = 168)$ is also considered.

The ACF and PACF after daily and weekly seasonal differencing, as shown in Figure 4, indicate that the series becomes stationary after eliminating the daily and weekly patterns. Lag polynomials up to order three was considered for the seasonal autoregressive polynomials and seasonal moving average polynomials. Different double seasonal ARIMA models have been estimated by maximum likelihood method using SAS software. All the data is used to estimate parameters except for the last 4 weeks that are put aside to evaluate post-sample accuracy of forecasts. The Schwartz Bayesian Criterion (SBC) for the different models was calculated and compared. By choosing the model corresponding to the minimum value of SBC, one is attempting to select the model corresponding to the highest Bayesian posterior probability. $DSARIMA\,(3,0,1)\,(1,\,1,\,1)_{24}\,(2,\,1,3)_{168}$ model was selected with the lowest SBC.

 ISBN: 978-1-61208-445-9

(a) ACF



(b) PACF

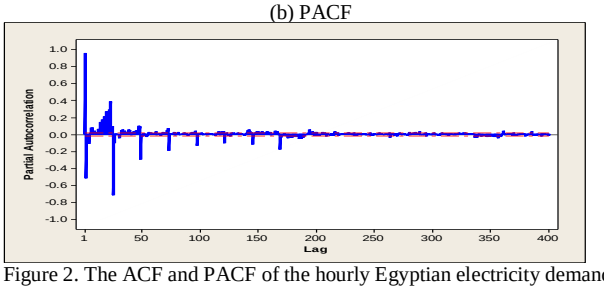
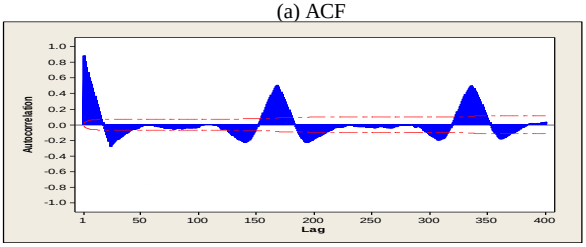


Figure 2. The ACF and PACF of the hourly Egyptian electricity demand

(a) ACF



(b) PACF

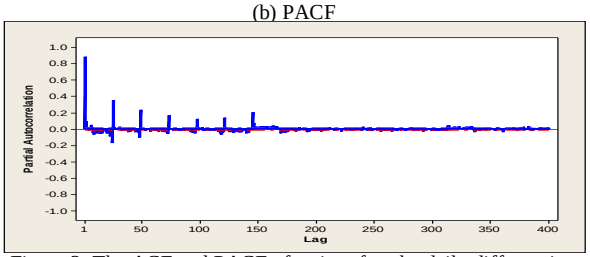


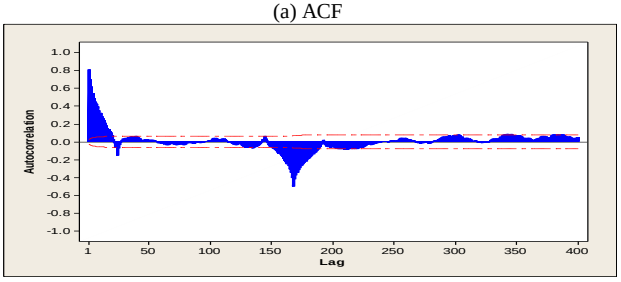
Figure 3. The ACF and PACF of series after the daily differencing

The selected model is estimated using the maximum likelihood method. The fitted model is given by:

$$(1 - 1.68\,B + 0.63\,B^2 + 0.06\,B^3)\,(1 - 0.12\,B^{24})\,(1 + 0.30\,B^{168} - 0.62\,B^{336})\,\nabla^0\,\nabla_{24}^1\,\nabla_{168}^1\,y_t = (1 - 0.92B)(1 - 0.77B^{24})(1 - 0.56B^{168} - 0.93B^{336} + 0.49\,B^{504})\varepsilon_t\,, \quad (3)$$

Forecasts are obtained for the last 4 weeks of our data set from the above fitted model.

(a) ACF



(b) PACF

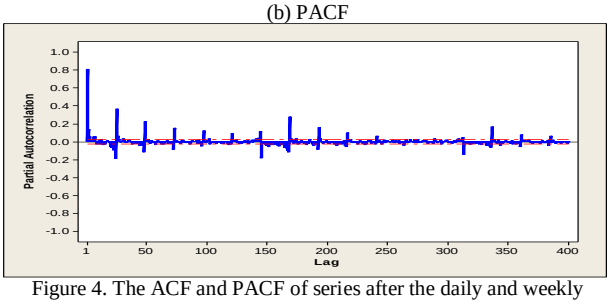


Figure 4. The ACF and PACF of series after the daily and weekly differencing

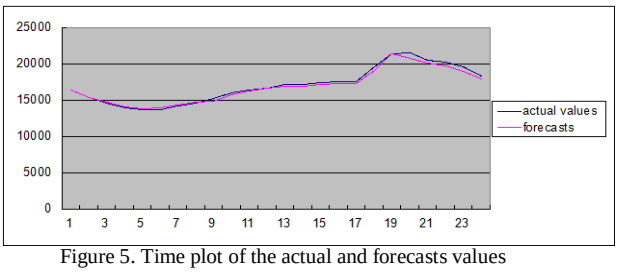


Figure 5. Time plot of the actual and forecasts values

The actual values of the Egyptian electricity demand series and its forecasts up to a day ahead are represented in Figure 5. It is observed that the forecasts are close to the actual values. In addition, the mean absolute percentage error (MAPE) is calculated for different time horizons to evaluate the accuracy of the selected model. The MAPE is the average of the absolute percentage prediction error. Low values of this statistic are preferred. The MAPE of the forecasts produced by the selected DSARIMA model up to one week horizon, two weeks horizon, three weeks horizon and a month horizon are 1.32%, 1.79%, 2.58% and 3.73%, respectively. Although, forecasting accuracy is less accurate for longer horizons, the selected model provides accurate forecasts for the Egyptian electricity demand.

 ISBN: 978-1-61208-445-9

V. CONCLUSION AND FUTURE WORK

In this paper, the DSARIMA model was investigated for forecasting Egyptian electricity demand. Different DSARIMA models were estimated. Forecasts produced by the selected model were accurate for different time horizons. The results agree with those reported in the literature for other countries. Different techniques and methods, such as exponential smoothing method and artificial neural networks, may be used and compared with DSARIMA model in forecasting the Egyptian electricity demand series. Obtained results would be of a great importance for policy makers.

REFERENCES

[1] D. W. Bunn, "Forecasting loads and prices in competitive power markets, " Proc. of the IEEE, vol. 88, pp. 163-169, Feb. 2000.

[2] M. P. Garcia and D. S. Kirschen,"Forecasting system imbalance volumes in competitive electricity markets, " Power Systems, IEEE Transactions, vol. 21, pp. 240-248, Feb. 2006.

[3] G. Box, G. M. Jenkins, and G. Reinsel , Time series analysis: Forecasting and control, 3rd ed. Englewood Cliffs, New Jersey: Prentice Hall, pp. 333,1994.

[4] J. W. Taylor, "Short-term electricity demand forecasting using double seasonal exponential smoothing, " Journal of the Operational Research Society, vol.54, pp.799– 805, 2003.

[5] J. W. Taylor, L. M. de Menezes, and P. E. McSharry, "A comparison of univariate methods for forecasting electricity demand up to a day a head, " International Journal of Forecasting, vol. 22, pp. 1–16, 2006.

[6] J. W. Taylor and P. E. McSharry, "Short-Term Load Forecasting Methods: An Evaluation Based on European Data, " IEEE Transactions on Power Systems, vol.22, pp.2213-2219, 2008.

[7] N. Mohamed , M. H. Ahmad, Z. Ismail, and Suhartono, "Double Seasonal ARIMA Model for Forecasting Load Demand," MATEMATIKA, vol. 26, pp.217–231, 2010.

[8] N. Mohamed, M. H. Ahmad, Suhartono, and Z. Ismail, "Improving Short Term Load Forecasting Using Double Seasonal Arima Model, " World Applied Sciences Journal, vol. 15, pp. 223-231, 2011.

[9] G. E. P. Box and G. M. Jenkins, Time Series Analysis: Forecasting and Control. Revised edition, San Francisco: Holden-day, 1976.

 ISBN: 978-1-61208-445-9

A Comparison of Classification Systems for Rule Sets Induced from Incomplete Data by Probabilistic Approximations

Patrick G. Clark

Department of Electrical Eng. and Computer Sci.
University of Kansas, Lawrence, KS, USA
e-mail: patrick.g.clark@gmail.com

Jerzy W. Grzymala-Busse

Department of Electrical Eng. and Computer Sci.
University of Kansas, Lawrence, KS, USA
Department of Expert Systems and Artificial Intelligence
University of Information Technology and Management,
Rzeszow, Poland
e-mail: jerzy@ku.edu

***Abstract*—In this paper, we compare four strategies used in classification systems. A classification system applies a rule set, induced from the training data set in order to classify each testing case as a member of one of the concepts. We assume that both training and testing data sets are incomplete, i.e., some attribute values are missing. In this paper, we discuss two interpretations of missing attribute values: lost values and "do not care" conditions. In our experiments rule sets were induced using probabilistic approximations. Our main results are that for lost value data sets the strength only strategy is better than conditional probability without support and that for "do not care" data sets the conditional probability with support strategy is better than strength only.**

***Index Terms*—Data mining; rough set theory; probabilistic approximations; MLEM2 rule induction algorithm; lost values and "do not care" conditions.**

I. Introduction

In this paper, we investigated the correctness of rule sets evaluated by the error rate, a result of ten-fold cross validation, with a focus on the choice of classification strategy. For a given rule set and testing data set the question is what is the best strategy for the classification system. In our experiments we used the Learning from Examples using Rough Sets (LERS) data mining system [1]–[3] with which we may use four strategies: strength of a rule combined with support, strength only, a conditional probability of the concept given the set of all training cases the rule matches combined with support, and the conditional probability, without any support.

In Sections 2 and 3, background material on incomplete data and probabilistic approximations are covered. Section 4 introduces and explains the four classification strategies used during the experiments described in Section 5. In Section 6, conclusions are discussed with the main results being that for the data sets with lost values the strategy based on strength only is better than conditional probability without support. For data sets with "do not care" conditions the strategy based on conditional probability with support is better than the strategy based on strength only.

TABLE I
TRAINING DATA SET

	Attributes			Decision
Case	Temperature	Headache	Cough	Flu
1	high	no	no	no
2	very-high	yes	*	no
3	normal	*	no	no
4	normal	no	*	no
5	?	?	yes	yes
6	very-high	yes	no	yes
7	*	yes	?	yes
8	high	yes	*	yes

II. Incomplete Data

We assume that the input data sets are presented in the form of a decision table. An example of a decision table is shown in Table I. Rows of the decision table represent cases, while columns are labeled by variables. The set of all cases will be denoted by U. In Table I, U = {1, 2, 3, 4, 5, 6, 7, 8}. Independent variables are called attributes and a dependent variable is called a decision and is denoted by d. The set of all attributes will be denoted by A. In Table I, A = {*Temperature*, *Headache*, *Cough*}. The value for a case x and an attribute a will be denoted by $a(x)$.

In this paper, we distinguish between two interpretations of missing attribute values: lost values and attribute-concept values. Lost values, denoted by "?", mean that the original attribute value is no longer accessible and that during rule induction we will only use existing attribute values [4][5]. "Do not care" conditions (denoted by $*$) correspond to a refusal to answer a question. With a "do not care" condition interpretation we will replace the missing attribute value by all possible attribute values. The error rate does not differ significantly for both interpretations of missing attribute values [6].

One of the most important ideas of rough set theory [7] is

 ISBN: 978-1-61208-445-9

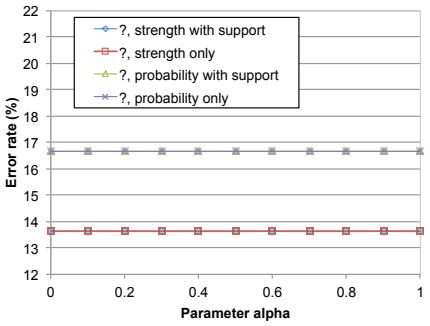


Fig. 1. Error rate for the *Bankruptcy* data set with lost values with lost values

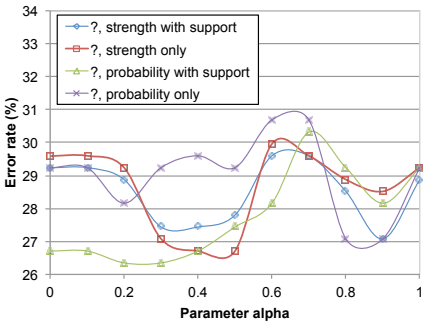


Fig. 2. Error rate for the rule set for the *Breast cancer* data set with lost values

an indiscernibility relation, defined for complete data sets. Let B be a nonempty subset of A. The indiscernibility relation $R(B)$ is a relation on U defined for $x, y \in U$ as defined in equation 1.

$$(x, y) \in R(B) \text{ if and only if } \forall a \in B \ (a(x) = a(y)) \quad (1)$$

The indiscernibility relation $R(B)$ is an equivalence relation. Equivalence classes of $R(B)$ are called *elementary sets* of B and are denoted by $[x]_B$. A subset of U is called *B-definable* if it is a union of elementary sets of B.

The set X of all cases defined by the same value of the decision d is called a *concept*. For example, a concept associated with the value *yes* of the decision *Flu* is the set {5, 6, 7, 8}. The largest B-definable set contained in X is called the *B-lower approximation* of X, denoted by $\underline{appr}_B(X)$, and defined in equation 2.

$$\cup\{[x]_B \mid [x]_B \subseteq X\} \quad (2)$$

The smallest B-definable set containing X, denoted by $\overline{appr}_B(X)$ is called the *B-upper approximation* of X, and is defined in equation 3.

$$\cup\{[x]_B \mid [x]_B \cap X \neq \emptyset\} \quad (3)$$

For a variable a and its value v, (a, v) is called a variable-value pair. A *block* of (a, v), denoted by $[(a, v)]$, is the set $\{x \in U \mid a(x) = v\}$ [8].

For incomplete decision tables the definition of a block of an attribute-value pair is modified in the following way.

- If for an attribute a there exists a case x such that $a(x) =$?, i.e., the corresponding value is lost, then the case x should not be included in any blocks $[(a, v)]$ for all values v of attribute a,
- If for an attribute a there exists a case x such that the corresponding value is a "do not care" condition, i.e., $a(x) = *$, then the case x should be included in blocks $[(a, v)]$ for all specified values v of attribute a.

For the data set from Table I the blocks of attribute-value pairs are:

[(Temperature, normal)] = {3, 4, 7},

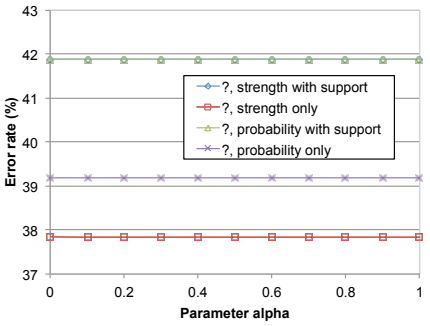


Fig. 3. Error rate for the rule set for the *Echocardiogram* data set with lost values

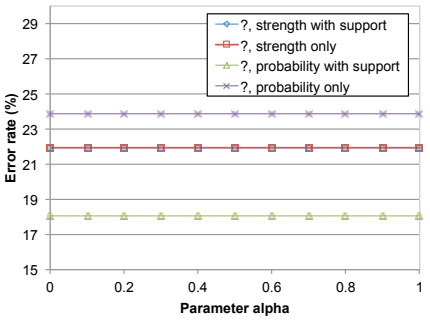


Fig. 4. Error rate for the rule set for the *Hepatitis* data set with lost values

[(Temperature, high)] = {1, 7, 8},
[(Temperature, very-high)] = {2, 6, 7},
[(Headache, no)] = {1, 3, 4},
[(Headache, yes)] = {2, 3, 6, 7, 8},
[(Cough, no)] = {1, 2, 3, 4, 6, 8}, and
[(Cough, yes)] = {2, 4, 5, 8}.

For a case $x \in U$ and $B \subseteq A$, the *characteristic set* $K_B(x)$ is defined as the intersection of the sets $K(x, a)$, for all $a \in B$, where the set $K(x, a)$ is defined in the following way:

- If $a(x)$ is specified, then $K(x, a)$ is the block $[(a, a(x))]$ of attribute a and its value $a(x)$,
- If $a(x) =$? or $a(x) = *$ then the set $K(x, a) = U$.

For Table I and $B = A$,

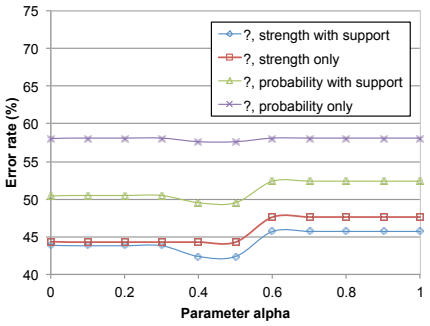


Fig. 5. Error rate for the rule set for the *Image segmentation* data set with lost values

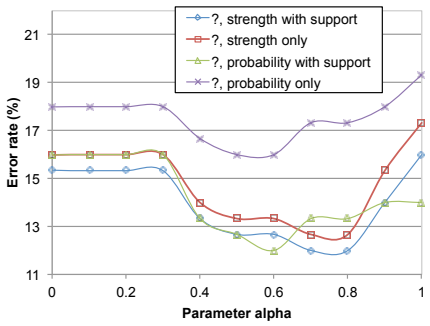


Fig. 6. Error rate for the rule set for the *Iris* data set with lost values

$K_A(1) = \{1\}$,
$K_A(2) = \{2, 6, 7\}$,
$K_A(3) = \{3, 4\}$,
$K_A(4) = \{3, 4\}$,
$K_A(5) = \{2, 4, 5, 8\}$,
$K_A(6) = \{2, 6\}$,
$K_A(7) = \{2, 3, 6, 7, 8\}$, and
$K_A(8) = \{7, 8\}$.

Note that for incomplete data there are a few possible ways to define approximations [9], we used *concept* approximations since our previous experiments indicated that such approximations are most efficient [10]. A B-*concept lower approximation* of the concept X is defined in equation 4.

$$\underline{B}X = \cup\{K_B(x) \mid x \in X, K_B(x) \subseteq X\} \tag{4}$$

The *B-concept upper approximation* of the concept X is defined by the equation 5.

$$\begin{aligned} \overline{B}X &= \cup\{K_B(x) \mid x \in X, K_B(x) \cap X \neq \emptyset\} \\ &= \cup\{K_B(x) \mid x \in X\} \end{aligned} \tag{5}$$

For Table I, A-concept lower and A-concept upper approximations of the concept {5, 6, 7, 8} are
$\underline{A}\{5, 6, 7, 8\} = \{7, 8\}$ and
$\overline{A}\{5, 6, 7, 8\} = \{2, 3, 4, 5, 6, 7, 8\}$, respectively.

III. Probabilistic Approximations

For completely specified data sets a *probabilistic approximation* is defined by equation 6, where α is a parameter, $0 < \alpha \leq 1$, see [10]–[15]. Additionally, for simplicity, the elementary sets $[x]_A$ are denoted by $[x]$. For discussion on how this definition is related to the variable precision asymmetric rough sets see [1][10].

$$appr_\alpha(X) = \cup\{[x] \mid x \in U, P(X \mid [x]) \geq \alpha\}. \tag{6}$$

For incomplete data sets, a *B-concept probabilistic approximation* is defined by equation 7 [10].

$$\cup\{K_B(x) \mid x \in X,\ Pr(X|K_B(x)) \geq \alpha\} \tag{7}$$

Where $Pr(X \mid K_B(x)) = \frac{|X \cap K_B(x)|}{|K_B(x)|}$ is the conditional probability of X given $K_B(x)$ and $|Y|$ denotes the cardinality

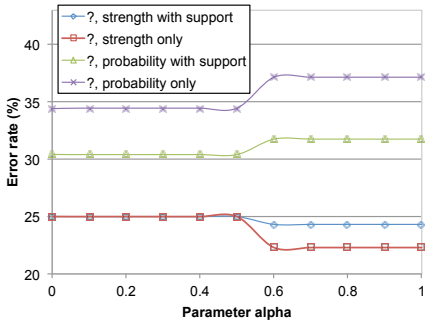


Fig. 7. Error rate for the rule set for the *Lymphography* data set with lost values

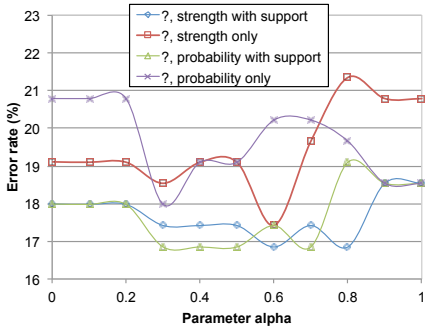


Fig. 8. Error rate for the rule set for the *Wine recognition* data set with lost values

of set Y. Note that if $\alpha = 1$, the probabilistic approximation becomes the standard lower approximation and if α is small, close to 0, in our experiments it was 0.001, the same definition describes the standard upper approximation.

For simplicity, we will denote $K_A(x)$ by $K(x)$ and the *A-concept probabilistic approximation* will be called a *probabilistic approximation*.

For Table I and the concept X = {5, 6, 7, 8}, there exist three distinct probabilistic approximations:
$appr_{1.0}(\{5, 6, 7, 8\}) = \{7, 8\}$
$appr_{0.6}(\{5, 6, 7, 8\}) = \{2, 3, 6, 7, 8\}$ and
$appr_{0.001}(\{5, 6, 7, 8\}) = \{2, 3, 4, 5, 6, 7, 8\}$.

 ISBN: 978-1-61208-445-9

IV. Classification

Rule sets, induced from data sets, are used most frequently to classify new, unseen cases. A *classification system* has two inputs: a rule set and a data set containing unseen cases. The classification system classifies every case as being member of some concept. A classification system used in LERS is a modification of the well-known bucket brigade algorithm [16]–[18].

The decision to which concept a case belongs is made on the basis of two factors: *strength* and *support*. *Strength* is the total number of cases correctly classified by the rule during training. The second factor, *support*, is defined as the sum of strengths for all matching rules indicating the same concept. The concept C for which the support, i.e., the following expression

$$\sum_{matching\ rules\ r\ describing\ C} Strength(r) \tag{8}$$

is the largest is the winner and the case is classified as being a member of C. This strategy is called *strength with support*. There exist three additional strategies. We may decide to which concept a case belongs on the basis of the strongest rule matching the case. This strategy will be called *strength only*. In the next strategy for every rule we compute ratios of the strength to the rule domain equal to the total number of cases matching the left-hand side of the rule. Such a ratio is a conditional probability of the concept given rule domain. A rule with the largest probability decides to which concept a case belongs. This strategy is called *probability only*. The fourth strategy, highly heuristic, in which all probabilities for rules indicating the same concept are added up is called *probability with support*.

In the classification system of LERS, if complete matching is impossible, all partially matching rules are identified. These are rules with at least one attribute-value pair matching the corresponding attribute-value pair of a case. For any partially matching rule r, the additional factor, called *Matching_factor* (r), is computed. Matching_factor (r) is defined as the ratio of the number of matched attribute-value pairs of r with a case to the total number of attribute-value pairs of r. In partial matching, the concept C for which the following expression

$$\sum_{\substack{partially\ matching \\ rules\ r\ describing\ C}} Strength(r) * Matching\_factor(r) \tag{9}$$

is the largest is the winner and the case is classified as being a member of C.

The problem is how to classify unseen cases with missing attribute values. In the LERS classification system, when an unseen case x is classified by a rule r, case x is considered to be not matched by r if for an attribute a, $a(x)$ = ? and the rule r contained a condition of the type (a, v), where v was a value of a. If for an attribute a, $a(x)$ = * and if the rule r contained a condition of the type (a, v), then case x is considered to be matched by r, does not matter what v is. In both cases interpretation of lost values and "do not care" conditions were strictly adhered to.

TABLE II
THE BEST RESULTS FOR ERROR RATES (%)—EXPERIMENTS ON DATA WITH *lost values*

Data set	Error rate (%) for			
	strength with support	strength only	probability with support	probability only
Bankruptcy	13.64	13.64	16.67	16.67
Breast cancer	27.08	26.71	26.35	27.08
Echocardiogram	41.89	37.84	41.89	39.19
Hepatitis	21.94	21.94	18.06	23.87
Image segmentation	42.38	44.29	49.52	57.62
Iris	12.00	12.67	13.33	16.00
Lymphography	24.32	22.30	30.41	34.46
Wine recognition	16.85	17.42	16.85	17.98

Using α = 0.333, the following rule set was induced by LERS from the data set from Table I

R1. (Headache, no) $\rightarrow$ (Flu, no), with strength = 3 and domain rule size = 3,

R2. (Temperature, very-high) $\rightarrow$ (Flu, no), with strength = 1 and domain rule size = 3,

R3. (Headache, yes) $\rightarrow$ (Flu, yes), with strength = 3 and domain rule size = 5, and

R4. (Cough, yes) $\rightarrow$ (Flu, yes), with strength = 2 and domain rule size = 4.

V. Experiments

Eight real-life data sets taken from the University of California at Irvine *Machine learning Repository* were used for experiments. Three of our data sets: *Bankruptcy*, *Echocardiogram* and *Iris* were numerical. All eight data sets were enhanced by replacing 35% of existing attribute values by missing attribute values, separately by lost values and by "do not care" conditions.

For all data sets there was a maximum value for the percentage of missing attribute values successfully replaced. In our experiments we chose the largest percentage common to all datasets, 35%, as it is the maximum percentage for the *bankruptcy* and *iris* data sets. As a result, 16 data sets were used, eight with 35% *lost values* and eight with 35% *"do not care" conditions*. Using the 16 data sets, experiments with 11 alpha values and four classification strategies were conducted, resulting in 704 ten-fold cross validation experiments.

Results of our experiments are presented as Figures 1 - 16 and Tables II and III. Results of experiments are presented in

 ISBN: 978-1-61208-445-9

TABLE III
THE BEST RESULTS FOR ERROR RATES (%)–EXPERIMENTS ON DATA WITH *"do not care" conditions*

Data set	Error rate (%) for			
	strength with support	strength only	probability with support	probability only
Bankruptcy	16.67	22.73	15.15	19.70
Breast cancer	28.16	28.88	27.08	27.80
Echocardiogram	24.32	27.03	27.03	28.38
Hepatitis	19.35	18.71	18.71	19.35
Image segmentation	47.14	51.43	46.19	49.52
Iris	36.00	38.67	28.67	25.33
Lymphography	24.32	26.35	25.00	31.76
Wine recognition	14.04	17.42	14.04	15.73

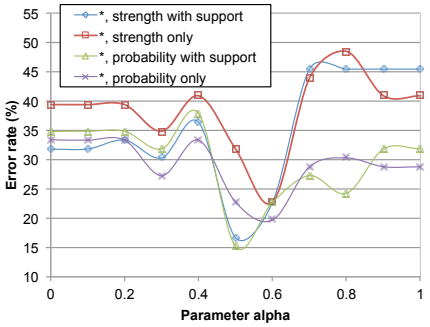


Fig. 9. Error rate for the *Bankruptcy* data set with "do not care" conditions

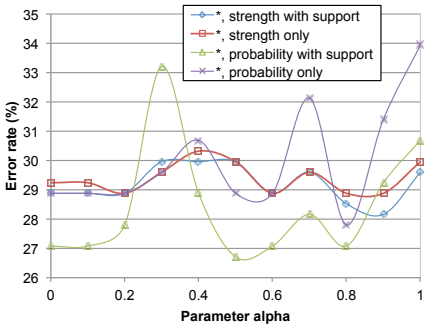


Fig. 10. Error rate for the rule set for the *Breast cancer* data set with "do not care" conditions

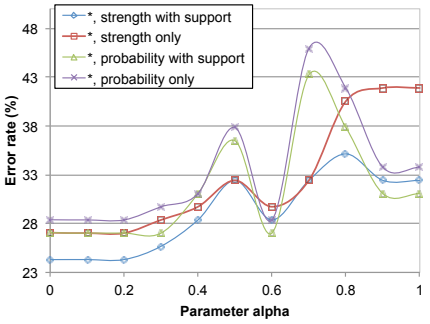


Fig. 11. Error rate for the rule set for the *Echocardiogram* data set with "do not care" conditions

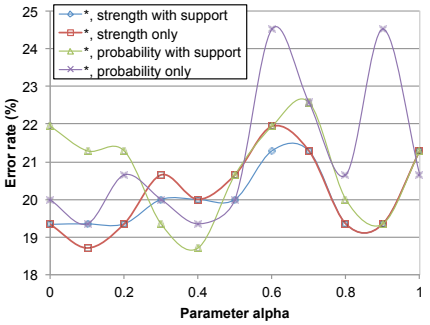


Fig. 12. Error rate for the rule set for the *Hepatitis* data set with "do not care" conditions

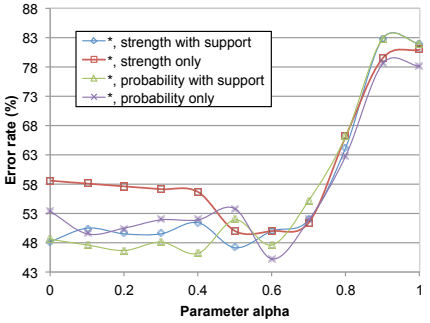


Fig. 13. Error rate for the rule set for the *Image segmentation* data set with "do not care" conditions

terms of error rate, a percentage of incorrectly classified cases when run in a 10-fold cross validation system.

In Tables II and III, the best results for all four strategies are shown. For each data set, strategy and interpretation of missing attribute values, we selected the smallest error rate from Figures 1 - 16. It is justified by practice of data mining, we always pick the value of the parameter α that corresponds to the smallest error rate. For example, for the *bankruptcy* data set, for two strategies, *strength with support* and *strength only*, for lost values, the error rate is 13.64%, so the corresponding entries in Table II are 13.64 (in this specific situation, the error rate does not depend on α).

Surprisingly, the strategy *strength only* seems to be the best strategy for data with lost values while the same strategy looks like the worst strategy for data with "do not care" conditions.

The Friedman test (5% level of significance), ties were taken into account shows that for both Tables II and III the null hypothesis that all four strategies do not differ significantly with respect to error rate must be rejected. For post hoc analysis we used the distribution-free pairwise comparisons based on Friedman rank sums (5% level of significance). The only results are: for data sets with lost values, the strategy based on *strength only* is better than the strategy

 ISBN: 978-1-61208-445-9

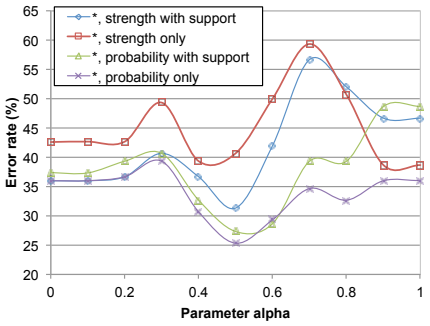


Fig. 14. Error rate for the rule set for the *Iris* data set with "do not care" conditions

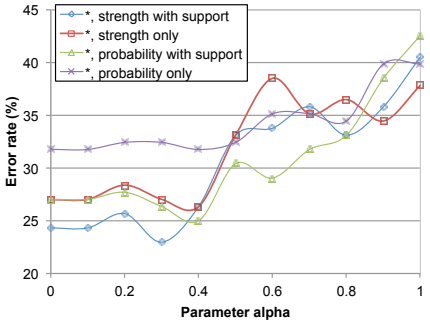


Fig. 15. Error rate for the rule set for the *Lymphography* data set with "do not care" conditions

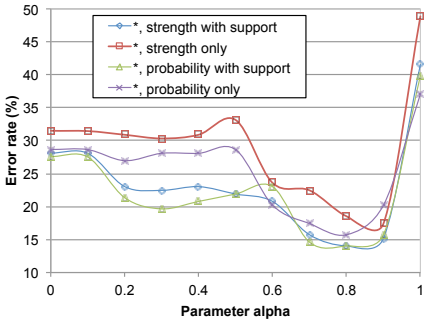


Fig. 16. Error rate for the rule set for the *Wine recognition* data set with "do not care" conditions

based on *probability only*, for data sets with "do not care" conditions, the strategy based on *probability and support* is significantly better than the strategy based on *strength only*. For other strategies differences are not statistically significant. For example, as follows from Table II, for data with lost values, the strategy based on *strength with support* is in most cases better than the strategy based on *probability only*, but that difference is not statistically significant.

VI. Conclusions

In this paper we report results of experiments on four different strategies of classification: *strength with support*, *strength only*, *probability with support* and *probability only* used for classification incomplete data by rule sets induced from incomplete data using probabilistic approximations.

Our main result is that for the data sets with lost values the strategy based on strength only is better than conditional probability without support. For data sets with "do not care" conditions the strategy based on conditional probability with support is better than the strategy based on strength only.

Additionally, results of our experiments show that for any given incomplete data set all four strategies should be applied and the best strategy should be selected as a result of ten-fold cross validation.

References

[1] P. G. Clark and J. W. Grzymala-Busse, "Experiments on probabilistic approximations," in *Proceedings of the 2011 IEEE International Conference on Granular Computing*, 2011, pp. 144–149.

[2] J. W. Grzymala-Busse, "A new version of the rule induction system LERS," *Fundamenta Informaticae*, vol. 31, pp. 27–39, 1997.

[3] ——, "MLEM2: A new algorithm for rule induction from imperfect data," in *Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 2002, pp. 243–250.

[4] J. W. Grzymala-Busse and A. Y. Wang, "Modified algorithms LEM1 and LEM2 for rule induction from data with missing attribute values," in *Proceedings of the Fifth International Workshop on Rough Sets and Soft Computing (RSSC'97) at the Third Joint Conference on Information Sciences (JCIS'97)*, 1997, pp. 69–72.

[5] J. Stefanowski and A. Tsoukias, "Incomplete information tables and rough classification," *Computational Intelligence*, vol. 17, no. 3, pp. 545–566, 2001.

[6] P. G. Clark, J. W. Grzymala-Busse, and W. Rzasa, "Mining incomplete data with singleton, subset and concept approximations," *Information Sciences*, vol. 280, pp. 368–384, 2014.

[7] Z. Pawlak, "Rough sets," *International Journal of Computer and Information Sciences*, vol. 11, pp. 341–356, 1982.

[8] J. W. Grzymala-Busse, "LERS—a system for learning from examples based on rough sets," in *Intelligent Decision Support. Handbook of Applications and Advances of the Rough Set Theory*, R. Slowinski, Ed. Dordrecht, Boston, London: Kluwer Academic Publishers, 1992, pp. 3–18.

[9] ——, "Rough set strategies to data with missing attribute values," in *Workshop Notes, Foundations and New Directions of Data Mining, in conjunction with the 3-rd International Conference on Data Mining*, 2003, pp. 56–63.

[10] ——, "Generalized parameterized approximations," in *Proceedings of the RSKT 2011, the 6-th International Conference on Rough Sets and Knowledge Technology*, 2011, pp. 136–145.

[11] J. W. Grzymala-Busse and W. Ziarko, "Data mining based on rough sets," in *Data Mining: Opportunities and Challenges*, J. Wang, Ed. Hershey, PA: Idea Group Publ., 2003, pp. 142–173.

[12] Z. Pawlak, S. K. M. Wong, and W. Ziarko, "Rough sets: probabilistic versus deterministic approach," *International Journal of Man-Machine Studies*, vol. 29, pp. 81–95, 1988.

[13] S. K. M. Wong and W. Ziarko, "INFER—an adaptive decision support system based on the probabilistic approximate classification," in *Proceedings of the 6-th International Workshop on Expert Systems and their Applications*, 1986, pp. 713–726.

[14] Y. Y. Yao, "Probabilistic rough set approximations," *International Journal of Approximate Reasoning*, vol. 49, pp. 255–271, 2008.

[15] W. Ziarko, "Variable precision rough set model," *Journal of Computer and System Sciences*, vol. 46, no. 1, pp. 39–59, 1993.

[16] L. B. Booker, D. E. Goldberg, and J. F. Holland, "Classifier systems and genetic algorithms," in *Machine Learning. Paradigms and Methods*, J. G. Carbonell, Ed. Boston: MIT Press, 1990, pp. 235–282.

[17] J. H. Holland, K. J. Holyoak, and R. E. Nisbett, *Induction. Processes of Inference, Learning, and Discovery*. Boston: MIT Press, 1986.

[18] J. Stefanowski, *Algorithms of Decision Rule Induction in Data Mining*. Poznan, Poland: Poznan University of Technology Press, 2001.

 ISBN: 978-1-61208-445-9

A Novel Framework to Describe Technical Accessibility of Open Data

Jolon Faichney and Bela Stantic
School of Information and Communication Technology
Griffith University
Gold Coast, Australia
email: {j.faichney,b.stantic}@griffith.edu.au

***Abstract*—Open Data is a recent and important movement that has economic, social, and political benefits. Despite a lot of attention in literature there are still limitations with the existing Open Data frameworks in describing technical accessibility of Open Data. In this paper, at first, we review the emergence of Open Data and the current state of frameworks and standards. We also describe our progress and findings working with Open Data at the local, state, and federal level in Australia. We then present a new Open Data Accessibility Framework (ODAF), which more completely defines levels of Open Data accessibility, guiding data custodians to make data more accessible for Open Data consumers.**

Keywords-Open Data; Open Government; Case Study; Framework.

I. INTRODUCTION

Open Data is a relatively recent movement, with the Unites States launching its Open Data portal in 2009 and the United Kingdom in early 2010 [1]. Open Data is a broad term, which has been defined as "accessible at marginal cost and without discrimination, available in digital and machine-readable format, and provided free of restrictions on use or redistribution" [1]. Open Government Data is a subset of Open Data, however Kloiber [2] states that the majority of uses of the term "Open Data" is used synonymously for "Open Government Data".

The United Kingdom has led the way in implementing and utilising Open Data being ranked number 1 in the world in both the Open Data Barometer [3] and the Global Open Data Index [4].

Due to the relative recentness of Open Data there are only several attempts to define frameworks for Open Data including the Open Definition (2005) [5], Sunlight Principles (2010) [6], Tim Berners-Lee's 5-star Linked Open Data (2010) [7], and Open Data Certificates (2013) [9]. Open Data Certificates [9] currently represents the most comprehensive framework combining three previous frameworks into four levels of Open Data publishing quality.

In Section 2, we present the widely accepted existing Open Data frameworks. In Section 3, we provide an overview of the current level of Open Data support and collaboration at the Local, State, and Federal levels in the City of Gold Coast region. In Section 4, we describe our experiences working with the City of Gold Coast outlining issues and challenges with the current frameworks. In Section 5, we propose and present a new framework describing the technical accessibility of Open Data. In Section 6, we discuss the challenges to adopting the proposed Open Data Accessibility framework. In Section 7 we present our conclusions and proposed future work.

II. BACKGROUND

A. Open Knowledge Definition

Underpinning the majority of Open Data definitions is the Open Definition provided by the Open Knowledge Foundation, now at version 2.0, which states: "*Knowledge is open if anyone is free to access, use, modify, and share it — subject, at most, to measures that preserve provenance and openness*" [10].

The Open Definition is broad, primarily focussing on the licensing of Open Data rather than the technical aspects.

B. Sunlight Foundation Open Data Principles

In 2010, the Sunlight Foundation defined 10 principles of Open Data (extending the previous 8 Sebastopol Principles): Completeness, Primacy, Timeliness, Ease of Physical and Electronic Access, Machine readability, Non-discrimination, Use of commonly Owned Standards, Licensing, Permanence, and Usage costs [6].

Many of the Sunlight Foundation principles are now covered in the Open Definition 2.0, specifically the last five principles listed above.

It's worth noting that the first two principles of Completeness and Primacy show the prioritising that the complete, raw, original data is made available. This is an important priority for Open Data as it means that the public has access to the original data. However, our experience discussed in the next section shows that non-raw, processed data can be of benefit for Open Data adoption. The Sunlight Foundation principles do not promote processed data other than making the data available in open, machine-readable formats.

C. 5-star Linked Open Data

Based on our experience, raw, unprocessed data can make Open Data less accessible. Tim Berners-Lee introduced the 5-star Linked Open Data framework with an emphasis on technical accessibility [7]. Each level makes the

 ISBN: 978-1-61208-445-9

data more accessible to applications. The five levels of the Linked Open Data framework are shown below:

1. Make the data available on the web in any format with an open license
2. Make it available as structured, computer-readable data (not in image or PDF formats)
3. Use non-proprietary formats such as CSV and XML
4. Use URIs within data so that other websites can point to resources
5. Link data to other data to provide context

The 5-star framework puts an important focus on technical accessibility. Open Data which does not have inherent links, only has to satisfy the first 3 levels. The 3rd level stipulates that the data must use non-proprietary (or open) formats. This is already covered in the Sunlight Foundation and Open Definition 2.0. However, it is important to note that the 5-star framework has emerged from Berners-Lee's work on linked data, the influence can be seen in the 4th and 5th levels which centre around linked data . Therefore the 5-star framework doesn't provide a greater level of detail in technical accessibility apart from adding levels for linked data.

D. Open Data Certificates

The Open Data Institute (ODI) has developed the Open Data Certificates [9] which combine the three previously discussed Open Data frameworks into four levels of Open Data access, which are:

Raw – A great start at the basics of publishing open data

Pilot – Data users receive extra support from, and can provide feedback to the publisher

Standard – Regularly published open data with robust support that people can rely on

Expert – An exceptional example of information infrastructure

The Expert level technical requirements can be summarised as follows:

- Provide database dumps at dated URLs,
- provide a list of the available database dumps in a machine readable feed,
- statistical data must be published in a statistical data format,
- geographical data must be published in a geographical data format,
- URLs as identifiers must be used within data,
- a machine-readable provenance trail must be provided that describes how the data was created and processed.

The expert level provides greater technical details than the preceding frameworks. However, in the course of working with Open Data, in our case we have found that the above frameworks do not adequately describe the requirements of software applications, which require technical access to the Open Data and we have developed an Open Data Accessibility Framework.

However, before we describe the Open Data Accessibility framework (ODAF) we will discuss our experiences working with Open Data.

III. Open Data in the City of Gold Coast

The City of Gold Coast, located in the State of Queensland, Australia, is unique in that there is strong support for Open Data at the local, state, and federal levels.

This section describes Open Data adoption at the federal, state, and local levels, and Griffith University's participation.

A. Federal Government

Australia is ranked at number 7 in the world in the Open Data Barometer [3] and is currently ranked number 5 in the world alongside New Zealand in the Open Data Index [4].

The Australian Federal Government launched its open data portal *data.gov.au* in 2012 and appointed the role of Director of Co-ordination and Gov 2.0. The open data portal can be used by any individual or organisation within Australia to host open data including federal, state, and local governments. The portal was migrated to the Open Knowledge Foundation's CKAN [11] platform in 2013 and currently hosts 5,200 data sets from 159 organisations. Since 2012 the federal government has run a national hackathon called *GovHack* where participants from around Australia compete for a pool of prizes. In 2014 GovHack was run in 11 cities with 1300 participants and observers competing for $256,000 in prizes. The federal Minister for Communications gave the keynote speech at the 2014 GovHack awards.

B. State Government

The City of Gold Coast resides in the state of Queensland, the second largest state in Australia, but the third-most populous. In 2013, the Premier of Queensland launched the state's open data initiatives, which included a competition titled the *Premier's Open Data Awards*. Unlike the GovHack hackathon, the Premier's Open Data Awards runs for several months providing participants time to work on larger projects. The Premier presented the awards to participants at both the 2013 and 2014 award ceremonies.

Despite the federal government providing the *data.gov.au* portal for all levels of government to use, the Queensland state government launched its own portal *data.qld.gov.au*, which currently hosts 1577 data sets.

C. Local Government

In 2013, the City of Gold Coast began to spearhead its open data initiative through a collaborative effort between the Economic Development and Information Services Departments. This involved establishing a data portal,

 ISBN: 978-1-61208-445-9

engaging with departments to identify and release data, running community forums to educate the public on open data, and supporting other open data initiatives.

The City of Gold Coast supported Griffith University in running a Premier's Open Data Awards information event in 2013 and also sponsored and helped organise the 2013 and 2014 local GovHack events.

The City of Gold Coast has decided to use the federal government's *data.gov.au* portal to host its data.

The City of Gold Coast has also been active in sponsoring development of apps which utilise Open Data including apps developed by Griffith University.

D. Federal, State, and Local Government Interaction

In 2013 the Director of Co-ordination and Gov 2.0 stated that the City of Gold Coast region was unique within Australia in having strong support from local, state, and federal government levels. In 2013 the Economic Development office of the City of Gold Coast along with state and federal departments arranged for Tim Berners-Lee to speak at Griffith University.

The City of Gold Coast has been very supportive of events and initiatives run by both state and federal governments. The federal government has also been very supportive of the Gold Coast region.

IV. Open Data Case Study

This section describes our experiences working with Open Data for three mobile apps and identifies issues during the process.

A. Cultural Challenges

Our first experience with Open Data in the City of Gold Coast began in 2013 with a smartphone app for disability car parks initiated by Regional Development Australia Gold Coast. Having no knowledge of the City of Gold Coast's Open Data support, the committee developing the app first asked the question, can we access the data? Fortunately, the City of Gold Coast had just started their open data initiative with the ultimate goal of "open by default". Despite the new open data initiative it took Council's enterprise architect three weeks to get the data due to traditional mindsets, policies, and procedures towards data protection.

The disability app is shown in Figure 1(a) and has more recently been expanded to show disability toilets and access ramps. The data required for the app is simply a list of latitude/longitude points for disability car parks on the Gold Coast in addition to polygon outlines of the carparks. The data is not sensitive, nor should it require a license, as the carparks can be seen simply by driving around the city. However, the traditional policies of the local council would've made it difficult to acquire and utilise the data. However, due to the council's Open Data initiative, which aims to not only release data publicly, if possible, but also under a license that allows the data to be freely used, re-used, and re-distributed, we were able to easily utilise the data once made available. Additionally the data was then made available for the general public on the federal Open Data portal *data.gov.au*.

We have since worked on two further Open Data-based apps including GC Heritage shown in Figure 1(b), for displaying heritage sites, and GC Dog Parks shown in Figure 1(c) and (d). Despite cultural and policy changes within the City of Gold Coast, acquiring data can still be a time consuming process as data is prepared in formats not previously required. However, the benefits of releasing this data are that the community now has easy access to disability, heritage, and dog park site information.

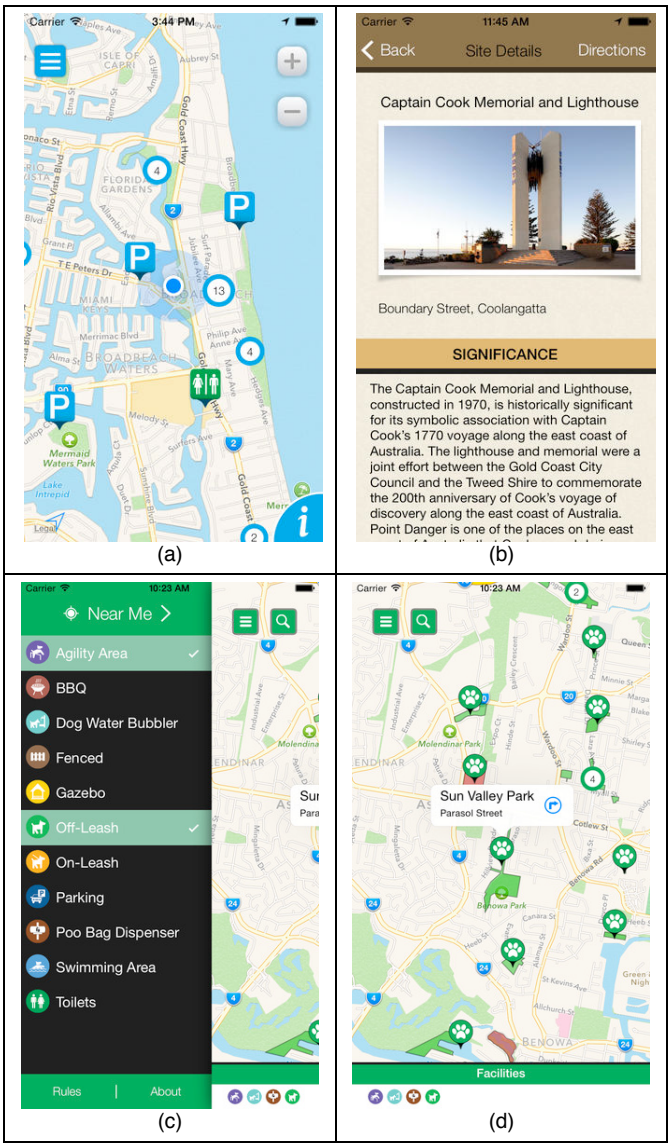


Figure 1. Apps developed by Griffith University using Open Data provided by City of Gold Coast: (a) Access GC, (b) GC Heritage, (c) and (d) GC Dog Parks

B. Data Cleaning Challenges

While working on Open Data for disability car parks there were a number of data cleaning steps required. The data cleaning challenges were as follows:

1. The data came in two files: on-street and off-street car parks, even though the app required no distinction between on-street and off-street car parks.
2. The files weren't named in a way to be able to identify which were on-street and off-street car parks.
3. The files contained *all* car parks in the City of Gold Coast not just disability car parks.
4. The two different files used different notations to identify a disability car park.
5. There were some minor formatting errors in the files.
6. The files were in a large XML file format with a lot of unnecessary data, the files were converted to CSV files suitable for mobile applications reducing the file size by about 100 times.

As it can be seen above, there were 6 data cleaning challenges. We faced similar issues with the additional data sets for other applications, which extended Access GC and for the GC Heritage and GC Dog Parks data. We proposed that the City of Gold Coast adopt our data cleaning process so that future updates to the data would be ready to use. However, the City of Gold Coast was not able at this stage to adopt a data cleaning process for the following reasons:

1. There is a lot of data to be made available and the highest priority is to release the data in the most accessible form,
2. The data custodians responsible for maintaining the data sets do not currently have the responsibility of cleaning it once it is exported,
3. The data custodians don't have the resources to facilitate regular data cleaning for Open Data purposes.

In addition, some of the cleaning processes require programming skills which the data custodians may not have.

Our experiences with using Open Data to date have formed the motivation to develop an Open Data Accessibility Framework. By having a technical accessibility framework, Open Data providers will be able to allocate sufficient resources to ensure Open Data is more accessible and more broadly adopted.

V. OPEN DATA ACCESSIBILITY FRAMEWORK

The 5 levels of the Linked Open Data Framework are aimed to address technical accessibility of Open Data. However, it is possible to achieve level 5 in this framework whilst still presenting many technical challenges to users of the data.

Our aim is not so much to replace the 5 levels but rather expand the 3rd level (use non-proprietary formats) to provide greater detail on technical accessibility.

We have identified six technical aspects that affect Open Data accessibility. These are not so much levels but rather checkboxes. Not all will be attainable by Open Data providers but provides a measure to evaluate the technical accessibility for both Open Data producers and consumers.

The Open Data Accessibility Framework is summarized in Table 1 and described in the following subsections using specific examples from our experiences working with Open Data.

TABLE I. OPEN DATA ACCESSIBILITY FRAMEWORK

Open Data Accessibility Framework (ODAF)
Resource Naming
Data Coalescing
Data Filtering
Data Consistency
Data Formats
API Accessibility

A. Resource Naming

When working with disability carparks we were provided with two files: *carparks.kmz* and *parking.kmz*. One represented on-street parking and the other represented off-street parking. It wasn't clear which file was which. The files or URLs should clearly indicate the contents of the file. In this case a name such as *onstreet_parking.kmz* and *offstreet_parking.kmz* should be used.

Resource names may also benefit from additional information such as the date of release of the data and the region they are from.

There currently is no standard for naming Open Data resources however the Expert level Open Data Certificates do stipulate that URLs should contain dates [9].

B. Data Coalescing

In the disability carpark example the data came in two files: on-street and off-street. There is little need for a distinction between the two types of carparks in most usage scenarios. In addition, the distinction would be more appropriately indicated as an attribute of a carpark record rather than being provided in separate files.

Open Data providers should aim to provide data as single files where there is no need for separate files.

Another example would be providing data separated into files by zip code. Most software applications will find it easier to deal with a single file and have the zip code as an attribute of the data rather than separated into individual files.

The Open Definition 2.0 states that "*the work shall be available as a whole*" [10] and the Sunlight Foundation principles state that "*Datasets released by the government should be as complete as possible, reflecting the entirety of what is recorded about a particular subject*" [6]. However, neither definitions stipulate whether this refers to a single file or multiple files, additionally the focus is on the primacy or

 ISBN: 978-1-61208-445-9

the original raw data, rather than data processed to be more accessible.

C. Data Filtering

In contrast with the previous requirement of Data Coalescing, there are often requirements for data to be filtered. For example the on-street and off-street parking data for the Gold Coast region consists of 22.7MB of uncompressed KML files. In contrast, the extracted disability carparks represented in CSV format were less than 200KB. Mobile apps are an important use of Open Data and a 22.7MB XML file would place a heavy resource burden on a mobile app.

It would be useful for datasets to be filtered for particular domains, in this case disability. Note that this requirement is not in conflict with the previous requirement of Data Coalescing. Data Coalescing should remove unnecessary separations of data whereas Data Filtering should provide useful application-oriented data separation.

A key point that we will address in the next section, is that Data Filtering and Data Coalescing must be driven by the Open Data consumer, as the Open Data producer may not be aware of the needs of the consumer.

D. Data Consistency

Open Data frameworks have identified the need for data cleanness. However, equally important is the need for notations to be consistent between files. As an example our work with carpark data used two different notations to represent disability carparks between the onstreet and offstreet files. One file used an identifier NUM_DISABLED_SPACES followed by a number, whereas the other file used simply the keyword "Disabled Parking".

The Expert level Open Data Certificate stipulates that URLs must be used consistently; however, there is no mention of consistency of other data types [9].

E. Data Formats

The carpark geospatial data provided was in an XML format (KML – Keyhole Markup Language). XML is a nested, structured data format, which is more challenging to process for mobile and web apps than the record-based CSV file format. XML however allows for complex data relationships to be represented and may in fact better represent the original data.

To minimize resource usage, CSV and other similar formats are more suitable for mobile and web applications due to reduced file size and simplicity in processing data.

The disability app data requirements were simply GPS co-ordinates of the disability resource. No additional information was required. Removing unnecessary data resulted in a file size reduction factor of 100.

Note that the recommendation here is not to replace the original data with a filtered CSV-like format, but to provide data in forms that are most suitable for mobile and web applications *in addition* to the original raw data.

The Expert level Open Data Certificate [9] recommends that geographical data be made available in geographical formats such as KML, however, our experience is that these formats are not the most suitable for mobile apps and preprocessing is often required.

F. API Accessibility

Most of the Open Data provided to us has been through files. However there are advantages to providing an API. One advantage is that the entire file does not need to be transferred. One dataset that we had access to was almost 1TB and had to be transferred on a hard drive.

APIs also provide additional benefits such as allowing the data to be filtered and destination formats to be determined at the time of the request.

Open Data portals, such as CKAN (Comprehensive Knowledge Archive Network) [11], utilized by the UK, EU, and Australia, allow for data uploaded in one format to be accessible through an API. However, the API doesn't allow searching and filtering of the data.

Emerging Open Data portals such as Open Data Architectures and Infrastructures (Open-DAI) [12] are beginning to provide support for data filtering. Alternatively technologies such as Elasticsearch [13] can be used to provide comprehensive RESTful API functionality however this would be beyond the skillset of most data custodians.

VI. DISCUSSION

Technical accessibility is an important factor in Open Data adoption. The Open Data Accessibility Framework (ODAF) we proposed identifies six factors that improve Open Data technical accessibility. We will now discuss some of the considerations and consequences of ODAF.

Technical accessibility aims to make it easier for Open Data consumers and software to process Open Data. ODAF identifies characteristics that improve technical accessibility that will require changes to the data and the processes that produce them. We will now discuss these implications.

Firstly, the most important aspect of Open Data is making the raw data available. Even though ODAF promotes changing the data and often removing data, it is important that the raw, original data is still made available. ODAF does not promote reducing the availability of data, but instead providing *additional modes* of the data.

Secondly, ODAF does not prescribe specifically what changes should be made. ODAF does not specify how data should be coalesced, filtered, made consistent, or which formats or APIs to provide it with. Ultimately it is up to the data custodians and consumers to determine these. ODAF is therefore a checklist that describes how successful the Open Data Consumer is in responding to the Open Data Producer.

It would be unrealistic for the Open Data Producer to provide data in every possible combination that could be

 ISBN: 978-1-61208-445-9

required. However, by keeping the ODAF factors in mind, it should result in better quality data sets at the outset.

By adopting an API-based approach the Open Data Producer can satisfy many of the ODAF requirements. An API can often coalesce many data sets into one API resource. Naming is likely improved, and API queries aid with specific queries. APIs generally allow multiple data format responses such as XML, JSON, and CSV and customized fields.

The onus however is still on the Open Data Producer to adopt processes that make the data more accessible. This may be beyond the resources that have been allocated to make Open Data available.

The most important step of Open Data is to make the original raw data available. However to allow Open Data to be useful and widely adopted it must also satisfy the ODAF requirements. This may require adopting an API-based Open Data Portal. However, existing Open Data Portals are limited in their ability to clean, filter, and coalesce structured data. Open Data Portals must be extended to provide querying abilities within structured data to satisfy the requirements of ODAF.

VII. CONCLUSION

In this paper we have explored existing Open Data frameworks and highlighted their weaknesses in describing requirements for technical accessibility. Based on our own experiences working on three Open Data projects and also being involved with Open Data initiatives at the local, state, and federal government levels in this work we propose the Open Data Accessibility Framework which presents factors which improve the technical accessibility of Open Data.

Adopting the ODAF factors will require a commitment from Open Data Producers to listen to their consumer's needs and make appropriate changes. It will require more resources to make the Open Data more technically accessible. Ultimately it should result in the data being available through an API. APIs can open up other opportunities such as crowdsourcing data, transitioning from e-Government to "we-Government" [14][15], progressing to what O'Reilly defines as "Government as a Platform" [16].

Open Data is an emerging initiative. Great progress has already been made in adoption at all levels of government throughout the world. Much of the progress has been at the policy and cultural level. There has been a focus on releasing data in a timely manner including the proposal of the timeliness measure *tau* [17]. However, much more work needs to be done at the technical level and this ODAF is a framework that defines attributes of technically accessible Open Data.

REFERENCES

[1] M. Heimstadt, F. Saunderson, and T. Heath, "From Toddler to Teen: Growth of an Open Data Ecosystem," JeDEM, vol. 6, no. 2, 2014, pp. 123-135.

[2] J. Kloiber, "Open Government Data – Between Political Transparency and Economic Development", Masters Thesis, Utrecht University, 2012.

[3] T. Davies, "Open Data Barometer 2013 Global Report", http://www.opendataresearch.org/content/2014/666/open-data-barometer-2013-global-report, 2013 [retrieved: March, 2015].

[4] Open Knowledge Foundation (OKFN) Global Open Data Index, http://index.okfn.org/place/, December, 2014 [retrieved: March, 2015].

[5] Open Knowledge Foundation (OKFN), Open Definition 1.0, http://opendefinition.org/history, August, 2005 [retrieved: March, 2015].

[6] Sunlight Foundation, "Ten principles for opening up government", http://sunlightfoundation.com/policy/documents/ten-open-data-principles/, August, 2010 [retrieved: March, 2015].

[7] T. Berners-Lee, "Is your Linked Open Data 5 star?", http://www.w3.org/DesignIssues/LinkedData.html, 2010 [retrieved March, 2015].

[8] C. Bizer, T. Heath, and T. Berners-Lee, "Linked Data – The story so far", Special Issue on Linked Data, International Journal on Semantic Web and Information Systems, 2009, pp. 1-22.

[9] Open Data Certificates, https://certificates.theodi.org, 2013 [retrieved: March, 2015].

[10] Open Knowledge Foundation (OKFN), Open Definition 2.0, http://opendefinition.org/od/, December, 2014 [retrieved: March, 2015].

[11] Comprehensive Knowledge Archive Network (CKAN), http://ckan.org [retrieved: March, 2105].

[12] R. Iemma, F. Morando, and M. Osella, "Breaking Public Administrations' Data Silos", JeDEM, vol. 6, no. 2, 2014, pp. 112-122.

[13] O. Kononenko, O. Baysal, R. Holmes, M. Godfrey, and D. Cheriton, "Mining Modern Repositories with Elasticsearch", MSR 2014, Hyderabad, India, May, 2014, pp. 328-331.

[14] D. Linders, "From e-government to we-government: Defining a typology for citizen coproduction in the age of social media", Government Information Quarterly, vol. 29, no. 4, 2012, pp. 446-454.

[15] T. Nam, "Suggesting frameworks of citizen-sourcing via Government 2.0", Government Information Quarterly, vol. 29, no. 1, 2012, pp. 12-20.

[16] T. O'Reilly, "Government as a Platform", Innovations, vol. 6, no. 2, 2011, pp. 13-40.

[17] U. Atz, "The Tau of Data: A new metric to assess the timeliness of data in catalogues", Proceedings of the International Conference for E-Democracy and Open Government, Krems, Austria, May, 2014, pp. 258-268.

 ISBN: 978-1-61208-445-9

RDF based Linked Open Data Management as a DaaS Platform

LODaaS (Linked Open Data as a Service)

Seonho Kim, Ivan Berlocher, Tony Lee
Saltlux, Inc.
Seoul, South Korea
e-mail: {shkim, ivan, tony}@saltlux.com

***Abstract*—In this paper we discuss the architecture and the processes for Resource Description Framework (RDF) based Linked Open Data as a Service (LODaaS), considering practical use cases. LODaaS is different from the usual Linked Data Platform (LDP) or Data Warehouse (DW), and as such, has to consider its own stakeholders and the processes of data publishing and consuming based on agreed ontology schema. The datasets should be transformed, published and consumed following the schema so that datasets could be shared, linked together and queried by 3rd party services without difficulty regarding data structures and queries. We implemented the Personalized and Localized Urban Quality Index (PLUQI) application for a data consumption use case, utilizing DaPaaS as a LODaaS platform. To implement this, s we designed an ontology schema, collected and published the datasets into the DaPaaS platform and reused these via the endpoint for PLUQI web service.**

Keywords—Linked Data Platform, RDF, Data as a Service, Data Integration, Open Data.

I. INTRODUCTION

W3C proposed the recommendation of Linked Data Platform (LDP) [1], which is a Linked Data specification defining a set of application integration patterns for building RESTful HTTP services that handle RDF documents. It provides a set of best practices and a simple approach for a read-write Linked Data architecture based on HTTP access to web resources that describe their state using the RDF data model [2].

Regarding the applications that use Linked Data, the data should be integrated for the domain specific purpose for which they aim to, but the recommendation does not cover this point. The graph-based RDF data model is flexible to integrate data from multiple data sources, but for data consumers who need to write a SPARQL query to retrieve data from the LDP, it is difficult to retrieve what they exactly want, because they do not have any preliminary information on what they should expect to get from it. Therefore, data consumers have to explore and look into the data first before writing queries. In addition, the data should be inter-connected, but composed of many namespaces that may have their own data schema for each, but not published on the LDP. Also, even if they are well constructed, the data integration itself has three kinds of heterogeneity problems to be solved: syntactic heterogeneity, structural heterogeneity, and semantic heterogeneity [3].

Generally, these problems are not handled or c resolved by most of the related research or solutions, because many of them are focusing on constructing a general knowledge base like Freebase, or have a specific purpose such as Linked Closed Data [4], Linked Government Data [5], or Linked Enterprise Data [6], so that they do not have to consider multiple purposes or requirements from the third party application developers or data consumers.

In this paper we propose the Linked Open Data as a Service (LODaaS) concept and its architecture, with the processes to manage the Linked Data for which the LDP could be used as Data as a Service (DaaS).

Section II describes the concepts and the considerations of the LODaaS, relevant stakeholders and core features. Section III describes the use case – PLUQI, which is a web-based service based upon the LODaaS. Section IV summarizes the roles and the features of LODaaS considered in this paper.

II. METHODOLOGY

A. Stakeholders

LODaaS has three relevant stakeholders: data owners, data publisher, and data consumers, as represented in Figure 1. The data owner is the one who owns the datasets to be published as open data, and the data publisher transforms the datasets and imports them into the LODaaS repositories. Both roles enable the reuse of data by the data consumer.

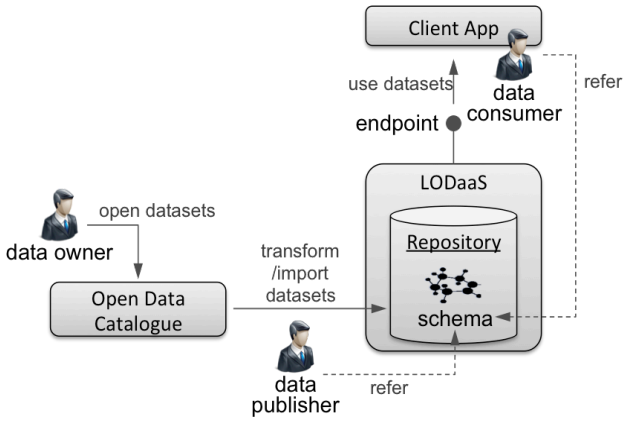


Figure 1. Relevant stakeholders of LODaaS

However, some of the open datasets are required to be transformed and imported according to the domain specific needs in the agreed format so that the data consumers can reuse them. This means that datasets should be transformed based on the agreed ontology schema so that the data consumers can simply query following the schema, rather than concern themselves with the actual datasets. So, the ontology schema should be designed accordingly, considering data consumer's perspectives and this is one of the data publisher's roles.

The data publisher should define the ontology schema to transform the datasets and import them to the repository. This means that the data publisher has to manage the repository, including the schema, to meet the needs of the services using the published data. Therefore, the schema should not be complex and not changed often in perspective of service management to provide the datasets via the endpoint, and to keep the processes of transformation/import efficiently.

B. Features

DaaS is an emerging subset of the "as a service" (XaaS) models for Cloud Computing services, where data is accessed, queried and updated on demand through a predefined service interface (usually a RESTful service). DaaS is based on Service-Oriented Architecture (SOA), and offers data services to be consumed by the third party applications in a unified format and to import data from data publishers.

Many of DaaS platform architecture includes several common components that LODaaS also does. Therefore, this section describes the processes and the requirements for each of the components as listed below:

- Data importing / transforming
- Data publishing
- Data retrieving / querying

1) Data Importing / Transforming

Data importing and transforming features are particularly important as a DaaS that provides RDF based Linked Data, because each of the resources should be named with URIs and the names should reuse the vocabularies to have unique ones for the same concepts or resources to be interlinked between the resources. In essence, the data publishing feature has to be combined with data importing feature to transform dataset provided by data publishers into RDF format, and store them into the repository.

To differ from general data warehouse platforms, this architecture should consider the characteristics of Linked Data and Open Data as described in TABLE I, in the perspective of Extract-Transform-Load (ETL) processes.

TABLE I. COMPARING ETL PROCESSES BETWEEN PLATFORMS

Process	Platform	
	Data Warehouse	***LODaaS***
Extract	extracting data from homogenous or heterogeneous data sources	extracting data from open data sets
Transform	deriving data to be loaded into end target	deriving data to be loaded into end target + URI mapping
Load	loading data into the end target	loading data into the triple store

2) Data Publishing

Data publishing process covers the processes for data transforming and providing machine-accessible data to the public [7]. Therefore, the architecture to support those processes should be designed. There are several related tools. D2R Server [8] provides the environment to publish relational data as Linked Data, Pubby [9] provides Linked Data interface for clients with HTML/RDF browsers, not only providing SPARQL endpoints. Paget [10] is a framework for building Linked Data applications, and PublishMyData [11] supports publishing Linked Data on the cloud and having access to it.

A case study of Linked Open Government Data (LOGD), "Publications Office of the European Union" indicates RDF based store on dedicated ontology (CDM, Common Data Model) as a key resource of its business model [12]. In addition, they have achieved the integration of their content and metadata based on the CDM. It is an example of how the Linked Data can be inter-connected based on the specific needs of users, which is useful because it may be adjusted to LDP to be used for domain specific purpose. This implies that LODaaS should be available for multiple purposes based on the ontology schemas (for each of them), and data should be published to the data layer of LODaaS following the schema. The data layer could be designed in two ways: to have independent repositories for each purpose, or to have integrated repository (Figure 2). However, regardless of the design, each datasets should be mapped to its corresponding ontology schema to be transformed and imported to the repository.

The first option is a more efficient way to manage data, because there is no confusion between ontology schemas when it performs ontology structural inference, and normally the application using LODaaS uses limited datasets stored in a specific repository instead of having the entire datasets. But if it is clear that there is no contradiction or conflicts, then the latter option is the preferred way to have expanded dataset based on Linked Data.

In summary, LODaaS architecture should support data transforming and importing, and the ontology schema should be loaded in the repository providing data interfaces.

 ISBN: 978-1-61208-445-9

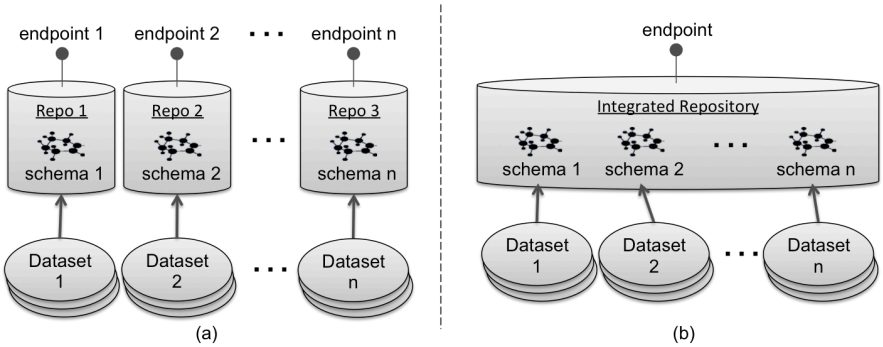


Figure 2. Two ways to locate repositories for LODaaS: (a) using independent repositories (b) using integrated repository

3) Data Retrieving / Querying

Linked Data relies on documents containing data in RDF format [13], therefore, LODaaS should provide SPARQL endpoints so that the clients can access the data through the interface. The clients need to write SPARQL query to retrieve data, which implies they need to use specific namespaces and URIs for resources. The data publishers and the clients as data consumers need to share the ontology schema information defined in the corresponding repositories or separated documentations, and this is where the ontology schema design is needed considering use cases, so that the data publishers and the consumers can be independent from each other. Details for this point is described in Section III (implementation), including the use case we have tried.

III. IMPLEMENTATION

A. Use case: PLUQI

Personalized and Localized Urban Quality Index (PLUQI) is the use case for this research defined as a customizable index model and mobile/Web application that can represent and visualize the level of well-being and sustainability for given cities based on individual preferences.

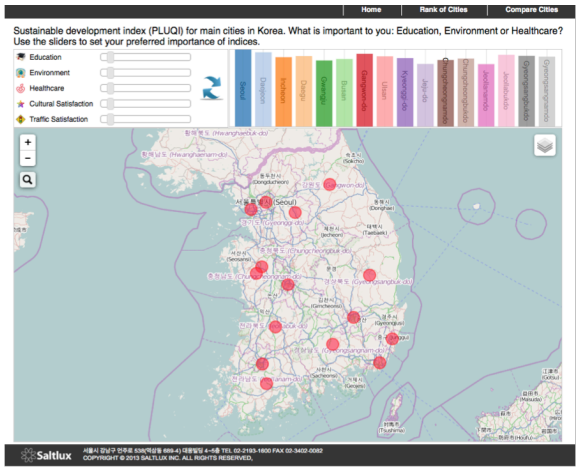


Figure 3. PLUQI Web Application

Figure 3 is a screenshot of the service webpage showing PLUQI indices. This is based on the open datasets published on DaPaaS, whose goal is to develop an integrated Data-as-a-Service (DaaS) and Platform-as-a-Service (PaaS) platform for open data applications. The proposed DaPaaS architecture will support unified accessibility to heterogeneous open datasets by using semantic technologies.

The users of PLUQI app will be provided with information about general satisfaction information from numbered index for the main regions/cities in Korea calculated based on the open datasets.

B. Processes & Architecture

To develop the PLUQI use case, we use the DaPaaS as a LODaaS to get open datasets about locational satisfaction, such as education, environment, transportation, etc. The steps for data publishing and consuming are as follows:

1. Collecting open datasets suitable for PLUQI from Korean open data catalogues
2. Transforming the open datasets into RDF format using Grafter, which is a Linked Data manufacturing tool for tabular data (Grafter [14], developed by Swirrl)
3. Importing and publishing the transformed data into DaPaaS
4. Consuming the published data from PLUQI app deployed in the DaPaaS via its SPARQL endpoint

The stakeholders can be summarized as shown in TABLE II.

TABLE II. STAKEHOLDERS FOR PLUQI USE CASE

Stakeholder	Participant
Data owner	Public sector (the person in charge of publishing open government data)
Data publisher	DaPaaS user (developer to transform and import the datasets, managing repositories)
Data consumer	PLUQI web application (developer to write queries and the user of PLUQI app)

We created a new repository of collected and transformed open datasets for PLUQI, from DaPaaS publisher portal GUI, and the SPARQL endpoint is provided. We also imported the PLUQI ontology schema to follow the architecture represented in Figure 2.

C. Ontology Schema Design

The PLUQI ontology schema helps data publishers and consumers to be independent from each other. The ontology schema can be represented as shown in Figure 4. *Quality_Index* can have *Value* to define measures described in open datasets, such as number of high schools as for educational satisfaction, which belongs to *Level_of_opportunity*. A *Value* has its *Location*, so we can describe that each location (regions/cities) has data of each categories represented as sub-classes of *Quality_Index*.

Following the design, data publishers can publish their datasets mapping for the categories, and the data consumers can query the data mapped for the categories they want to take. This means data consumers can get all the published data for *Quality_Index* without being too concerned about what kind of datasets exists for the categories. Especially, PLUQI should be available to get up-to-date indices and the data without changing queries used in the app regardless of what new datasets are published on DaPaaS.

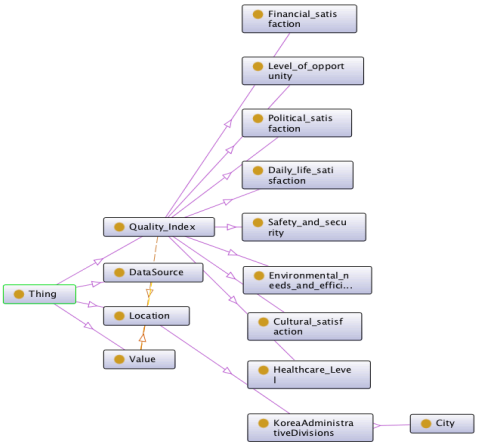


Figure 4. PLUQI Ontology Schema

IV. CONCLUSION

LODaaS has different features from the ones in LDP or DW, because it deals with open datasets, which are already open to the public and allows publishing their datasets. In addition, unlike Open Data catalogues, such as CKAN, LODaaS supports domain specific needs to provide useful data, so that the data is well transformed and managed considering the needs and creating a link between them.

To meet this need, we described the considerations for the processes of publishing open datasets to form Linked Data for specific use-case, and the architectures to manage and consume the data. It is difficult to be 'generally opened' and also 'targeting domain specific needs' at the same time, but this issue could be resolved by the agreed ontology schema between the stakeholders participating in the use case and by separating the repositories for each of them.

Our future work will be devoted to collect various data not only open datasets but also from social media which will be integrated with open datasets based on the PLUQI ontology schema to be published on DaPaaS.

ACKNOWLEDGMENT

This work was supported by the Industrial Strategic Technology Development Program (10044494, WiseKB: Big data based self-evolving knowledge base and reasoning platform) funded by the Ministry of Science, ICT & Future Planning (MSIP, Korea), and DaPaaS project which is funded by the European Commission under the 7th Framework Programme, Project No. 610988, http://dapaas.eu/.

REFERENCES

[1] W3C Linked Data Working Group, "Linked Data Platform 1.0W3C Proposed Recommendation 11 December 2014." https://dvcs.w3.org/hg/ldpwg/raw-file/default/ldp.html [retrieved: 2015.02.24]

[2] N. Mihindukulasooriya, R. García-Castro, and M. E. Gutiérrez, "Linked Data Platform as a novel approach for Enterprise Application Integration," COLD 2013.

[3] M. Gagnon, "Ontology-based integration of data sources." Proceedings of 10th International Conference on 10th International Conference on Information Fusion(FUSION2007), 2007, pp.1-8.

[4] M. Cobden, J. Black, N. Gibbins, Les Carr, and N. R. Shadbolt, "A Research Agenda for Linked Closed Dataset," COLD 2011. [Online]. Available from: http://eprints.soton.ac.uk/272711/ [retrieved: 2015.02.24]

[5] D. Wood, "Linking government data," Springer, 2011.

[6] D. Wood, "Linking Enterprise Data, 1st edition," Springer, 2010.

[7] W3C Government Linked Data Working Group, "Best Practices for Publishing Linked Data", http://www.w3.org/TR/ld-bp/ [retrieved: 2015.02.24]

[8] D2R Server. [Online]. Available from: http://www4.wiwiss.fu-berlin.de/bizer/d2r-server [retrieved: 2015.02.24]

[9] Pubby. [Online]. Available from: http://wifo5-03.informatik.uni-mannheim.de/pubby [retrieved: 2015.02.24]

[10] Paget. [Online]. Available from: https://code.google.com/p/paget [retrieved: 2015.02.24]

[11] PublishMyData [Online]. Available from: http://www.swirrl.com/publishmydata [retrieved: 2015.02.24]

[12] European Commission, "Study on business models for Linked Open Government Data", 2013. [Online]. Available from: https://joinup.ec.europa.eu/sites/default/files/85/31/25/Study_on_business_models_for_Linked_Open_Government_Data_BM4LOGD_v1.00.pdf. [retrieved: 2015.02.24]

[13] C. Bizer, T. Heath, and T. Berners-Lee, "Linked Data - the story so far," International Journal on Semantic Web and Information Systems, vol. 5, no. 3, pp. 1–22, 2009.

[14] Grafter [Online]. Available from: http://grafter.org [retrieved: 2015.02.24]

 ISBN: 978-1-61208-445-9

Ontology Learning from Text

Departing the Ontology Layer Cake

Abel Browarnik and Oded Maimon
Department of Industrial Engineering
Tel Aviv University
Tel Aviv, Israel
email: {abel, maimon}@eng.tau.ac.il

***Abstract*— We analyze the ontology learning objectives, reviewing the type of input one would expect to meet when learning ontologies - peer-reviewed scientific papers in English, papers that undergo quality control. We analyze the Ontology Learning Layer Cake model and its shortcomings, proposing alternative models for ontology learning based on linguistic knowledge and existing, wide coverage syntactical, lexical and semantic resources, using constructs such as clauses. We conclude, after showing that the Ontology Learning Layer Cake has a low maximum F measure (probably below 0.6), that the alternatives should be explored.**

Keywords - Ontology Learning from text; Ontology Learning Layer Cake Model; Language Modeling; Clauses; Subsentences.

I. INTRODUCTION

We define an *ontology* as a formal, explicit specification of a shared conceptualization. Most available ontologies are crafted and maintained with human intervention. Ontologies represent reality, and as such, require frequent updates, making them both costly and difficult to maintain. Ontology Learning (OL) has been developed in order to overcome this problem. Learning is interpreted in the literature as the process of creating the ontology and populating it. In this paper, the goal of OL is defined to be the (at least semi) automatic extraction of knowledge, possibly structured as simple or composite statements, from a given corpus of textual documents, to form an ontology.

Most, if not all, OL approaches [11][12][14][22][24] follow a model named the Ontology Learning Layer Cake (OLC), and share many features, such as statistical based information retrieval, machine learning, and data and text mining, resorting to linguistics based techniques for certain tasks.

This paper will argue that the Ontology Learning Layer Cake approach is not the best choice for Ontology Learning. The paper reviews the following issues:

Understanding Ontology Learning from text. An ontology represents, in our view, a "portion" of the world that we are looking at, for example, toxicity of engineered nanoparticles (or nanotoxicity). Every new paper published on the subject may add a new entity or relationship to the nanotoxicity model. We see OL as a tool for modeling a domain and keeping this model updated.

The input used to learn ontologies. The input to the OL process depends on the domain itself. Modeling scientific domains, such as nanotoxicity, normally draws on peer-reviewed papers or other scientific articles, magazines or books. These are well-formed and quality-checked texts. We will argue that the quality of the input is one of the parameters to be taken into account when devising an OL framework.

Analysis of the Ontology Learning Layer Cake Model. The Ontology Learning Layer Cake model aims at learning ontologies by using a multistep approach. Most OLC driven methods rely on at some stage to statistics or machine learning, the basis for unsupervised learning methods, either to extract terms, to build taxonomies or to extract relations and rules. We will argue that the sequential nature of OLC results in rather low overall recall and precision for the whole process.

Alternative models for Ontology Learning. Based on the conclusions of our analysis of OL, the input to the OL process, and the OLC model, we find that the entire subject of OL could be tackled in a different manner. Assuming that many of the target domains are defined by well-formed texts, we introduce the fundamentals of alternative OL frameworks. These fundamentals build on English language structure.

The paper includes, following this Introduction, a background section, an review of alternative models and a discussion where statistical models are compared to linguistic methods. The paper also includes a conclusion section.

II. BACKGROUND

A. Understanding Ontology Learning from text

Ontology Learning from text aims to obtain knowledge on the domain covered by the text. It is often seen as the extraction of ontologies by applying natural language analysis techniques to texts.

Cimiano [9] describes the tasks involved in OL as forming a layer cake. The cake is composed, in ascending order, by terms, sometimes synonyms, concepts, taxonomies, relations and finally axioms and rules.

This approach can be seen as a cornerstone in OL. It assumes that terms (gathered through term extraction methods) are the basic building blocks for OL. There are many term extraction methods [3][19][29] and many tools are publicly available [1][24][28]. The synonym layer is either based on sets, such as WordNet synsets [23] (after sense

 ISBN: 978-1-61208-445-9

disambiguation), on clustering techniques [3][12][22][11] or other similar methods, or on web-based knowledge acquisition.

The concept layer perception depends on the definition of concept. The consensual view is that it should include: an intensional definition of the concept, a set of concept instances, i.e., its extension, and a set of linguistic realizations, i.e., (multilingual) terms for this concept.

The concept hierarchy level (i.e., the taxonomic level) uses one of three paradigms to induce taxonomies from text:

- The application of lexico-syntactic patterns to detect hyponymy relations [16]. This approach is known to have reasonable precision but very low recall.
- The exploitation of hierarchical clustering algorithms to automatically derive term hierarchies from text, based on Harris' distributional hypothesis [15], that terms are similar in meaning to the extent in which they share syntactic contexts.
- A document-based notion of term subsumption, as proposed, for example, in Sanderson and Croft [27]. Salient words and phrases extracted from the documents are organized hierarchically using a type of co-occurrence known as subsumption.

The relation level has been addressed primarily within the biomedical field. The goal is to discover new relationships between known concepts (such as symptoms, drugs, and diseases) by analyzing large quantities of biomedical scientific articles. Relation extraction through text mining for ontology development was introduced in work on association rules in Maedche and Staab [21]. Recent efforts in relation extraction from text have been carried on under the Automatic Content Extraction (ACE) program, where entities (i.e., individuals) are distinguished from their mentions. Normalization, the process of establishing links between mentions in a document and individual entities represented in a ontology, is part of the task for certain kind of mentions (e.g., temporal expressions).

The rule level is at an early stage [20]. The European Union-funded project Pascal [10] on textual entailment challenge has drawn attention to this problem.

Our analysis of OL takes Wong [30] and Wong, Liu and Bennamoun [31] as its starting point.

The following remarks represent the consensus among OL reviews:

- The fully automatic learning of ontologies may not be possible.
- A common evaluation platform for ontologies is needed.
- The results for discovery of relations between concepts are less than satisfactory.
- The more recent literature points to an increase in interest in using the Web to address the knowledge acquisition bottleneck and to make OL operational on a Web scale.

Ontology Learning starts at a given point in time. It collects the existing knowledge by using the methods available and builds a representation of this knowledge. There are many schemes for knowledge representation, such as Extensible Markup Language (XML), Resource Description Framework (RDF)/RDF Schema (RDFS), Web Ontology Language (OWL)/OWL2 and Entity-Relationship Diagrams (ERD).

The representation scheme chosen affects the extent of reasoning that the Ontology will allow. An XML represented Ontology will allow less reasoning than a First Order Logic scheme.

As knowledge is added, the representation absorbs the new knowledge incrementally. The scheme should not permit contradictory knowledge. Therefore, if new knowledge contradicts existing knowledge, a protocol is needed to resolve the contradictions.

New knowledge is created by scientific work published (e.g., books, papers, proceedings). The input is processed and incorporated into the knowledge representation.

B. The Input Used to Learn Ontologies

There are a few types of ontologies. Upper or foundation ontologies are general purpose ontologies and define reality. Domain ontologies, on the other hand, are used to depict a domain. A domain ontology plays a role similar to that of the conceptual layer of an ERD in the area of system analysis. In both cases the relevant concepts are entities, attributes, relationships and more.

System analysis is performed by humans – system analysts – that gather information from humans involved in the domain, together with environmental details, to create the conceptual layer of an ERD for that domain. An ERD has two additional layers, the logical layer and the physical layer. These two layers deal with implementation details and therefore are not relevant to our discussion. From now on, when we refer to ERD we mean the conceptual layer of an ERD. We argue that an ERD is equivalent to an ontology, because an ERD of the domain represents conceptually the entities involved and the relations between the entities. OL is the task of gathering the information necessary to build the ontology of the domain (and perhaps to populate it). This is similar to building an ERD, even though the means to build an ERD are not necessarily the same means required to learn an ontology.

The main difference between creating an ERD for a business and learning an Ontology for a domain is the fact that the domain builds on a body of scientific books or papers that are a strong basis for a learning process without human intervention (except for paper writing), while building an ontology for an information domain such as an Enterprise Requirements Planning (ERP) system relies on knowledge that is seldom written, let aside formalized. Yet, in both cases we target a model of the domain. Thus, we see OL as a modeling technique.

We should consider the sources of text used towards learning ontologies, and the quality of these texts. To this end we could think of a Martian visiting Earth. The visitor could find him/herself browsing the New York Times website on November 21st, 2013. He/She could see there that "Applicants Find Health Website Is Improving, but Not Fast Enough ". Having no worldly knowledge he/she would not understand that this issue is related to the United States (US)

health reform commonly referred to as Obamacare. This is where an ontology comes of use. An ontology of US politics would provide the visitor with the background knowledge he would need to understand the newspaper. The source for this OL task would be newspapers and books. As we deal with learning ontologies from text we do not consider video or audio sources. We do not consider new media, such as Tweeter, email or Facebook either, because language quality in new media cannot be taken for granted. Newspapers, magazines and books undergo editing which is a sort of quality control. This is not to say that there is no use for new media. It can be used for less formal tasks, as is the case with sentiment analysis.

Most existing methods for OL from text rely on well-formed text. There is no clear guidance on this issue. Our Literature review reveals that existing tools such as ASIUM [12], OntoLearn [24] and CRCTOL [17] perform term extraction using sentence parsing. Text-to-Onto [22], TextStorm/Clouds [25] and Syndikate [14] perform term extraction using syntactic structure analysis and other techniques. OntoGain [11] uses shallow parsing for term extraction. If the text is not well-formed, these tasks would not be feasible. Thus, we assume that the input for OL from text consists of well-formed text.

C. Analysis of the Ontology Learning Layer Cake Model

Methods using the Ontology Learning Layer Cake model divide the OL task into four or five sequential steps. These steps result in the following outputs:

- Terms
- Concepts
- Taxonomic relations
- Non-taxonomic relations
- Axioms

Some methods perform all the steps, while some perform only part of them. Recall and precision obtained by the methods vary. ASIUM [12], Text-to-Onto [22], Ontolearn [24] and Ontogain [11] do not provide an overall figure of precision and recall for the whole OL process. TextStorm/Clouds [25] cites an average result of 52%. Syndikate [14] mentions high precision (94 to 97 % for different domains) and low recall (31 to 57% correspondingly, for the same domains). CRCTOL [17] reports a figure of 90.3% for simple sentences and 68.6% for complex sentences (we assume that these figures represent the F measure of the method). The main characteristics shared by the methods based on the OLC model are:

- The method is split into sequential steps. The output of step i is the input for step $i + 1$ (though there may be additional inputs from other sources).
- Individual steps may produce, in addition to the main output expected, other results. As an example, ASIUM, OntoLearn and CRCTOL perform term extraction using sentence parsing. This can be considered a secondary output. Secondary output from step i is not passed to step $i + 1$.
- If a method has four or five sequential steps, each step depends on the previous one. If every step has precision and recall (and therefore their harmonic mean, the F measure) bound by p ($p < 1$), then the method cannot obtain recall and precision better than p^n (n is the number of steps). As an example, if we assume the F measure of each step to be 0.8, the F measure of the whole OL method with 4 steps will be 0.41. With 0.9 (a result seldom attained) per step the F measure is 0.59!
- A step which uses statistical or machine learning methods requires considerable amounts of data to give significant results. In general, it also requires the data to be split into a training set and a test set.
- Statistical and machine learning methods have to beware of the danger of over-fitting and wrong choices of training and test sets. These may result in output distortion.
- OLC methods require statistical evidence regarding knowledge of the area being studied. Thus, features such as co-occurrence of terms or words may induce conclusions that are nonsensical to subject experts.
- The statistical nature of some steps makes it impossible to trace back specific results. As an example, a method may find a relation between two concepts following the co-occurrence of the two concepts in the same sentence or paragraph in different portions of text, or even in different documents.

Often the unsupervised nature of statistical or machine learning methods is an incentive to choose such methods, as less human effort is required to understand the subject matter. Such understanding is critical for the success of non-statistical, non-machine learning methods. The human effort and the fact that results are sometimes similar for both supervised and unsupervised methods tip the scales, leading the practitioner to choose unsupervised methods. In this case, however, we see that OLC methods sometimes use supervised techniques. Such may be the case, for example, in TextStorm/Clouds. This method uses part of speech tagging (using WordNet), syntactic structure analysis, and anaphora resolution for any of the steps of the OLC process, for example, term extraction, and taxonomic and non-taxonomic relation learning. Yet, this is an OLC method with its "cascading" nature.

It is possible that the OLC approach was inspired by the "divide and conquer" algorithm design paradigm. A divide and conquer algorithm works by recursively breaking down a problem into smaller sub-problems of the same (or related) type, until these become simple enough to be solved directly. The solutions to the sub-problems are then combined to give a solution to the original problem. Problems in data mining are often solved using "cascading" algorithms built on the divide and conquer paradigm. The fact that data mining was followed by textual data mining which, in turn, inspired OL may be one of the reasons for choosing OLC.

III. ALTERNATIVE MODELS FOR ONTOLOGY LEARNING

The approaches and methods reviewed above stem mainly from Cimiano's ontology layer cake. That is, there is

consensus that first one has to gather terms (and probably also synonyms), then concepts, and finally extract relations (taxonomic for all the systems, with some of the systems and approaches aiming also at non-taxonomic relations, with a variable degree of success). In addition, few systems cross the reasoning threshold. Some of the methods are purely statistic; most use a mixture of statistical based and linguistic and/or Natural Language Processing (NLP) based methods, with statistics based methods taking an important role. The reason for this may be based on Brants conjecture [4]. Brants argues that NLP contribution to Information Retrieval related tasks is rather ineffective. Is this the only way to proceed? We would initially ask two questions:

- Would it be possible to start, for example, by gathering relations (any relation, not necessarily taxonomic) and then proceed to the other layers mentioned in the OLC?
- If we want to store knowledge in RDF or RDFS is there any requirement that the order should respect the OLC order?

It is worth mentioning that even linguistic or NLP based methods may rely on corpora. It is said that the most promising trend exploring the web as the corpus of choice due to its extent and coverage.

There is a third question:

- We are dealing with a specific and bound subject – ontology learning. Would it be appropriate to deal with text in a purely linguistic, even linguistic-theoretic manner? In other words, do we have to rely on corpora, or can we use language modeling to obtain results?

A. *Extracting semantic representation from text*

Research on Psycholinguistics and Neurolinguistics looks at how humans gather information from text (see for example, Caplan, [6]). It is generally agreed that humans gather information from text at the sentence level or even at the clause level, and not at the document (or corpus) level. Thus, extracting the semantic payload of text would ideally include deep parsing, semantic labeling of the text and a process of knowledge accumulation. From a practical point of view, the above may not be feasible. To overcome these limitations, researchers apply practical approaches based on heuristics and partial methods.

The literature shows several attempts to gather information from text at the sentence level. The model proposed by Chen and Wu [8] makes extensive use of FrameNet [13]. A semantic frame can be thought of as a concept with a script. It is used to describe an object, state or event. Chen and Wu avoid the need to deep-parse the sentences that constitute the text by using Framenet.

Chaumartin [7] presents another attempt to tackle the semantic representation issues. Instead of using Framenet, Antelope, the implementation of Chaumartin's work uses VerbNet [18] for the lexical-semantic transition.

Both methods (Chaumartin [7] and Chen and Wu [8]) deal with text at the sentence level, without taking into account sub-sentence components. Chen does not provide a tool to showcase the capabilities of his approach, except for an example in the paper: "*They had to journey from Heathrow to Edinburgh by overnight coach.* ". The example is assigned Framenet's frame *Travel* with all its elements (traveler, source and goal). Chaumartin released a full-fledged toolbox to test the capabilities of his approach. The system includes an example with its result, a clear semantic representation of the sentence in terms of VerbNet classes and all the resulting constraints. The representation includes all the semantic details necessary to assess the situation and allow for higher order activities such as question answering, reasoning and maybe automatic translation. Yet, for other sentences, results are not satisfactory, as in:

"Most of these therapeutic agents require intracellular uptake for their therapeutic effect because their site of action is within the cell"

The example above yields no result (i.e., no VerbNet class is recognized and therefore no semantic representation is extracted). One of the reasons for failing to discover the semantic contents of complex or compound sentences may be that such a sentence structure requires more than one frame or verb class to be found.

B. *From clauses or subsentences to RDF triples and RDFS*

The Resource Description Framework (RDF) data model, defined in http://www.w3.org/RDF/, makes statements about resources in the form of subject-predicate-object expressions known as triples. The subject denotes the resource, and the predicate denotes traits or aspects of the resource and expresses a relationship between the subject and the object.

The Longman Grammar of Spoken and Written English (LGSWE) [2] defines a clause as a unit structured around a verb phrase. The lexical verb in the verb phrase denotes an action (drive, run, shout, etc.) or a state (know, seem, resemble, etc.). The verb phrase appears with one or more elements denoting the participants involved in the action, state, etc. (agent, affected, recipient, etc.), the attendant circumstances (time, place, manner, etc.), the relationship of the clause to the surrounding structures, etc. Together with the verb phrase, these are the clause elements. The clause elements are realized by phrases or by embedded clauses. A clause may be divided into two main parts: a subject and a predicate. The subject is generally a nominal part, while the predicate is mainly a verbal nucleus. Preisler [26] states that a clause contains the constituents Subject, Verbal, Complement and Adverbial, all or some of them. Rank shifting adds complexity to the subject. In this context, rankshifted clauses are called subclauses, while non-rankshifted clauses are called main clauses.

Clauses appearing together in a larger unit (generally sentences, but possibly phrases in the event of a rank-shifted clause) are linked by structural links, the principal types being coordinators, subordinators and wh-words. Coordinators create coordinated clauses. On the other hand, subordinators and wh-words create embedded clauses.

Subsentences, a concept introduced by Browarnik and Maimon [5] sometimes overlap with clauses. Yet, subsentences keep the construct simpler because of the restriction to the number of Verbal Constructs (VC) per subsentence.

 ISBN: 978-1-61208-445-9

The above definitions give a clue on to how to represent knowledge extracted by linguistic modeling by using RDF constructs, e.g., an RDF triple and a clause or a subsentence seem to represent entities and relationships. In other words, knowledge extracted from a clause or a subsentence can be represented by an RDF triple. Generally, an RDF triple is defined by an RDF scheme. In our case, as we start from knowledge extracted from a clause or a subsentence to obtain an RDF triple, the RDF scheme (or RDFS) should be obtained from a generalization of the RDF triples obtained, in a bottom-up fashion.

C. Advantages of language modeling approaches

Traceability. No matter whether one picks deep parsing or one of the heuristic methods (using sentences, clauses or subsentences), the result is that one sentence is derived into a set of RDF triples. It is possible that one RDF triple is derived from more than one sentence. This mechanism creates a clear relationship between the input sentence and the output RDF triple. A human reviewer may decide to check the Ontology Learning results. Such a review is feasible. Moreover, if the result turns out to be erroneous, it may be fixed. Such fixing may have an impact on the method and change other results.

Contradictory facts. If two scientific papers contain contradictory statements and both statements are used towards learning an Ontology we face a problem. While it is possible that the issue remains unresolved in the scientific community, we cannot assert both results in the resulting Knowledge Base. This situation is better than what a Machine Learning approach would provide, i.e., a statistical procedure that would add the most statistically significant result into the Knowledge Base, making it hard to clarify afterwards whether the result was appropriate.

Big Corpora. Most OLC methods are based on statistical processing. The corpus has to be split into a training set and a test set. The outcome is measured by recall and precision. As mentioned before, the performance of each step in an OLC method is bounded and therefore the result of a 4-step cascading method is theoretically bounded. On the other hand, Language Modeling does not require the use of big corpora, at least explicitly.

Recall and Precision. Recall measures the percentage of results that should have been returned by the method used. Methods based on Language Modeling make this measure less relevant. The methods process every sentence on the input text and return a result. Therefore one could argue that recall would always be 100%. Precision, while still relevant to the Language Modeling methods, may be interpreted differently. The methods do return a result, yet the result may be wrong, therefore reducing precision. If and when the mistake is discovered, the traceability mentioned above can be used to correct it, thus improving the model's precision.

IV. DISCUSSION: STATISTICAL VS. LINGUISTIC-BASED METHODS

The attempt to construct a model follows one of two possible approaches, and sometimes a mixture of the two approaches.

The Language Modeling approach aims at understanding the subject matter. Such approaches generally rely on a thorough knowledge of the subject matter. The results are generally accurate. When results accumulate they either confirm the adequacy of the model, making it widely accepted, or undermine it, leading it ultimately to be discarded. Physics shows plenty of theoretical models that were accepted after obtaining more and more experimental confirmations. Even more models were rejected after experimental evidence showed they were wrong.

The other approach aims at creating models by gathering facts and statistics that give us a hint about the "internals" of the subject matter, without obtaining a detailed understanding of these internals. Engineering and Medicine are areas where such methods flourish. A good example is the area of queuing theory. To forecast the arrivals of requests, one often uses heuristics to decide on a given probability distribution. Such decisions give a good approximation to the real conditions of the problem, but not necessarily the best theoretical fit.

Most methods for Natural Language Processing (NLP), and especially the methods used for OL, draw on the second approach. To mention only the most prominent OL systems, we see that:

- ASIUM [12] uses agglomerative clustering for taxonomy relations discovery.
- Text-To-Onto [22] uses agglomerative clustering, hypernyms from WordNet and lexico-syntactic patterns for taxonomic relation extraction. Non-taxonomic relations are extracted using association rule mining.
- In TextStorm/Clouds [25], both taxonomic and non-taxonomic relations are obtained using part of speech tagging, WordNet, syntactic structure analysis and anaphora resolution.
- Syndikate [14] implements semantic templates and domain knowledge for relations extraction.
- OntoLearn [24] relation extraction relies on hypernyms from WordNet (relations extracted are only taxonomic).
- CRCTOL [17] achieves relation extraction (taxonomic and non-taxonomic) using lexico-syntactic patterns and syntactic structure analysis.
- OntoGain [11] applies aglomerative clustering and formal concept analysis to extract taxonomic relations and Association rule mining for non-taxonomic relations.

Moreover, for most of the reviewed OL methods and systems, even the term and concept layers (stemming from Cimiano's ontology layer cake) are extracted using statistical methods.

The Language Modeling approaches for OL from English texts are based on the following facts:

- An ontology can be represented by RDF triples.
- RDF triples are subject-predicate-object expressions.
- Clauses are components of sentences and include a subject, a verbal part, a complement and an adverbial

part, all or some of them. Subsentences are a textual passage built around one verbal construct.

- RDF triples are equivalent to clauses or subsentences.

Therefore, an RDF triple can be constructed from a clause or a subsentence. But, how does one extract the triple from a clause? And how does one find a clause from a sentence? Our preferred solution is to use Clause Boundary Detection or Subsentence Detection. Both are characterized by near linear time complexity.

Chaumartin shows how to extract a kind of role based frame from a sentence, although, as we indicated above, working at the sentence level has succeeded only partially.

Yet, statistical methods are very useful and should by no means be neglected. Constructing resources, such, as part of speech taggers, WordNet, VerbNet or FrameNet, do profit from statistical methods. Based on these resources, one can create a somehow theoretical linguistic model that would not rely on corpora in order to extract clauses or subsentences from sentences, and in turn convert it into RDF triples, thus learning an ontology with no – direct – use of corpora.

V. CONCLUSION

We have shown that OLC methods have generally low recall and precision (less than 0.6). We have shown that OL is generally based on well-formed text, and that text can be decomposed into clauses (or subsentences), and then translated into RDF statements. The Language Modeling approaches make backtracking results possible, therefore allowing for a correction option. The approaches do not rely on big corpora, making these approaches potentially more efficient than OLC. The Language Modeling approaches elude OLC problems such as bounded performance due to OLC cascading nature, limitations of its statistical basis, the need for big corpora, and the problems associated with such corpora.

REFERENCES

[1] M. Baroni and S. Bernardini, (2004), "BootCaT: Bootstrapping corpora and terms from the web", In Proceedings of LREC (vol. 4), L04-1306, Lisbon, Portugal.

[2] D. Biber, S. Johansson, G. Leech, S. Conrad, and E. Finegan, (1999), Longman Grammar of Spoken and Written English, Longman Publications Group, ISBN 0-582-23725-4.

[3] D. Bourigault and C. Jacquemin, (1999), "Term extraction+ term clustering: An integrated platform for computer-aided terminology", In Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics, (pp. 15-22), Association for Computational Linguistics.

[4] T. Brants, (2003), "Natural Language Processing in Information Retrieval", In Bart Decadt, Véronique Hoste, Guy De Pauw (Eds.): Computational Linguistics in the Netherlands 2003, December 19, Centre for Dutch Language and Speech, University of Antwerp. Antwerp papers in Linguistics University of Antwerp 2003.

[5] A. Browarnik and O. Maimon, (2012), "Subsentence Detection with Regular Expressions", Presented at the XXX AESLA International Conference, Lleida, April 2012.

[6] D. Caplan, (2003), "Neurolinguistics", In The Handbook of Linguistics, UK: Blackwell Publishing.

[7] F. R. Chaumartin, (2005), "Conception and Implementation of a Syntactic/Semantic Interface Using English Large Coverage Resources". Master Thesis in Linguistics and Information Technology (in French), Universite Paris7 – Denis Diderot, 2005.

[8] E. Chen and G. Wu, (2005), "An Ontology Learning Method Enhanced by Frame Semantics", ISM 2005: 374-382.

[9] P. Cimiano, (2006), "Ontology Learning and Population from Text.Algorithms, Evaluation and Applications", ISBN: 978-0-387-30632-2, Springer, 2006

[10] I. Dagan, O. Glickman, and B. Magnini, (2006), "The PASCAL Recognising Textual Entailment Challenge". Lecture Notes in Computer Science, Volume 3944, Jan 2006, Pages 177 - 190.

[11] E. Drymonas, K. Zervanou, and E. Petrakis, (2010), "Unsupervised ontology acquisition from plain texts: The OntoGain system", Natural Language Processing and Information Systems, 277-287.

[12] D. Faure and C. Nedellec, (1999), "Knowledge acquisition of predicate argument structures from technical texts using machine learning: The system ASIUM". Knowledge Acquisition, Modeling and Management, 329-334.

[13] C. J. Fillmore, (1976), "Frame semantics and the nature of language", Annals of the New York Academy of Sciences, (280):20–32.

[14] U. Hahn, M. Romacker, and S. Schulz, (2000), "MedSynDiKATe--design considerations for an ontology-based medical text understanding system", In Proceedings of the AMIA Symposium (p. 330), American Medical Informatics Association.

[15] Z. Harris, (1968), Mathematical Structures of Language, John Wiley & Sons, 1968.

[16] M. A. Hearst, (1992), "Automatic acquisition of hyponyms from large text corpora", In Proceedings of the 14th conference on Computational linguistics-Volume 2 (pp. 539-545), Association for Computational Linguistics.

[17] X. Jiang and A. H. Tan, (2009), "CRCTOL: A semantic‐based domain ontology learning system", Journal of the American Society for Information Science and Technology, 61(1), 150-168.

[18] K. Kipper Schuler, (2005), "VerbNet - a broad-coverage, comprehensive verb lexicon", PhD thesis, University of Pennsylvania. http://verbs.colorado.edu/~kipper/Papers/dissertation.pdf

[19] L. Kozakov, Y. Park, T. Fin, Y. Drissi, Y. Doganata, and T. Cofino, (2004), "Glossary extraction and utilization in the information search and delivery system for IBM Technical Support", IBM Systems Journal, 43(3), 546-563.

[20] D. Lin and P. Pantel, (2001), "DIRT - Discovery of Inference Rules from Text", In Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2001, pp. 323-328.

[21] A. Maedche and S. Staab, (2000), "Discovering conceptual relations from text". In W. Horn, editor, Proceedings of the 14th European Conference on Artificial Intelligence (ECAI'2000), 2000.

[22] A. Maedche and S. Staab, (2000, August), "The text-to-onto ontology learning environment", In Software Demonstration at ICCS-2000-Eight International Conference on Conceptual Structures.

[23] G. A. Miller , R. Beckwith , C. Fellbaum , D. Gross , and K. J. Miller, "Introduction to WordNet: An On-line Lexical Database", Int J Lexicography 3: 235-244.

[24] R. Navigli and P. Velardi, (2004), "Learning Domain Ontologies from Document Warehouses and Dedicated Websites", Computational Linguistics, 30(2), MIT Press, 2004, pp. 151-179.

 ISBN: 978-1-61208-445-9

[25] A. Oliveira, F. C. Pereira, and A. Cardoso, (2001), "Automatic reading and learning from text", In Proceedings of the International Symposium on Artificial Intelligence (ISAI).

[26] B. Preisler, (1997), A Handbook of English Grammar on Functional Principles, Aarhus University Press, Denmark, ISBN 87 7288 405 3.

[27] M. Sanderson and B. Croft, (1999), "Deriving concept hierarchies from text", In Research and Development in Information Retrieval, pages 206–213. 1999.

[28] F. Sclano and P. Velardi, (2007), "TermExtractor: a Web Application to Learn the Shared Terminology of Emergent Web Communities". To appear in Proc. of the 3rd International Conference on Interoperability for Enterprise Software and Applications (I-ESA 2007), Funchal (Madeira Island), Portugal, March 28–30th, 2007.

[29] J. Wermter and U. Hahn, (2005), "Finding new terminology in very large corpora", In Proceedings of the 3rd international conference on Knowledge capture (pp. 137-144), ACM.

[30] W. Y. Wong, (2009), "Learning lightweight ontologies from text across different domains using the web as background knowledge", (Doctoral dissertation, University of Western Australia).

[31] W. Wong, W. Liu, and M. Bennamoun, (2012), "Ontology Learning from Text: A Look back and into the Future", ACM Computing Surveys (CSUR), 44(4), 20.

 ISBN: 978-1-61208-445-9

Applying Semantic Reasoning in Image Retrieval

Maaike de Boer, Laura Daniele,
Paul Brandt, Maya Sappelli
TNO
Delft, The Netherlands
email: {maaike.deboer,
laura.daniele, paul.brandt,
maya.sappelli}@tno.nl

Maaike de Boer, Maya Sappelli

Radboud University
Nijmegen, The Netherlands,
email: {m.deboer,
m.sappelli}@cs.ru.nl

Paul Brandt
Eindhoven University of
Technology (TU/e)
Eindhoven, The Netherlands
email: p.brandt@tue.nl

***Abstract*—With the growth of open sensor networks, multiple applications in different domains make use of a large amount of sensor data, resulting in an emerging need to search semantically over heterogeneous datasets. In semantic search, an important challenge consists of bridging the semantic gap between the high-level natural language query posed by the users and the low-level sensor data. In this paper, we show that state-of-the-art techniques in Semantic Modelling, Computer Vision and Human Media Interaction can be combined to apply semantic reasoning in the field of image retrieval. We propose a system, GOOSE, which is a general-purpose search engine that allows users to pose natural language queries to retrieve corresponding images. User queries are interpreted using the Stanford Parser, semantic rules and the Linked Open Data source ConceptNet. Interpreted queries are presented to the user as an intuitive and insightful graph in order to collect feedback that is used for further reasoning and system learning. A smart results ranking and retrieval algorithm allows for fast and effective retrieval of images.**

Keywords-semantics; natural language queries; semantic reasoning; image retrieval; ranking.

I. INTRODUCTION

More and more sensors connected through the Internet are becoming essential to give us support in our daily life. In such a global sensor environment, it is important to provide smart access to sensor data, enabling users to search semantically in this data in a meaningful and, at the same time, easy and intuitive manner. Towards this aim, this paper presents the GOOgle™ for Sensors (GOOSE) system, which is a general-purpose search engine conceived to enable any type of user to retrieve images and videos in real-time from multiple and heterogeneous sources and sensors [1]. The proposed system especially focuses on cameras as sensors, and aims at bridging the semantic gap between natural language queries that can be posed by a user and concepts that can be actually recognized by detectors. These detectors are built using computer vision techniques, and the number of detectors is limited compared to all possible concepts that may be in the user's mind.

This work addresses the semantic interpretation of user queries to support the task of image retrieval. Our approach is general-purpose, i.e., not restricted to a specific domain, since it gives the flexibility to search for images that can contain any kind of concepts. Users can pose queries in natural language, which are parsed and interpreted in terms of objects, attributes, scenes and actions, but also semantic and spatial relations that relate different objects to each other. The system uses semantic graphs to visually explain to its users, in an intuitive manner, how a query has been parsed and semantically interpreted, and which of the query concepts have been matched to the available image detectors. For unknown concepts, the semantic graph suggests possible interpretations to the user, who can interactively provide feedback and request to train an additional concept detector. In this way, the system can learn new concepts and improve the semantic reasoning by augmenting its knowledge with concepts acknowledged by the user. Images corresponding to the recognized concepts are retrieved from video streams, ranked and presented to the user as result. The reasoning to build the semantic graphs is fully automated and uses ConceptNet [2], an external large knowledge base with concepts and semantic relations, constructed by combining multiple sources on the Web.

The main challenge in this work is the integration of several research areas in which semantics is intended in different ways. In Computer Vision, applying semantics is the process of converting elementary visual entities, e.g., pixels, to symbolic forms of knowledge, such as textual tags and predicates. In Human Media Interaction, semantics is mainly used in terms of tags used to annotate images by users. In Semantic Modelling, semantics is intended in terms of semantic models used to describe domain knowledge, such as ontologies, and inference using rules.

The goal of this paper is to show how state-of-the-art techniques in these three research areas can be combined in one single application able to semantically reason and learn, allowing its users to pose natural language queries about any topic and interact with the system to retrieve images corresponding to their queries. An overview paper about the whole GOOSE application is given in [3], where this paper is only focused on the semantic interpretation. The working of the image classification and quick image concept learning is given in [4] and fast re-ranking of visual search results is presented in [5].

This paper is structured as follows: Section II describes related work, Section III presents a short overview of the application, Section IV explains the semantic reasoning in the application, Section V contains the discussion and Section VI consists of the conclusion and future work.

 ISBN: 978-1-61208-445-9

II. RELATED WORK

Most of the effort in applying semantics in Computer Vision is aimed at training detectors and classifiers using large sources of visual knowledge, such as ImageNet [6] and Visipedia [7]. ImageNet is based on the WordNet [8] hierarchy of nouns, which allows to reason about *objects* in the images, but not about *actions*. Moreover, only a part of the ImageNet images is manually annotated with bounding boxes, which limits the results of the classifiers and detectors training process. Visipedia is an augmented version of Wikipedia with annotated images. Annotation is a time consuming and error-prone activity that is usually delegated to motivated crowds, who need to be trained to reduce the subjective noise in the process of image labelling. Concerning annotation, considerable effort has been spent in Human Media Interaction in labelling images for the purpose of retrieving video events. Usually, domain-specific ontologies are used as basis for annotation, such as the ontologies in [9] [10] that are used to annotate soccer games. Another example of domain-specific ontology is presented in [11] for the purpose of action recognition in a video surveillance scenario. In general, the efforts mentioned above focus on the specific algorithms for image processing and/or on the annotation of images, rather than on the semantic interpretation that should facilitate users in understanding the reasoning behind the system. Therefore, more attention should be given at integrating computer vision and semantic reasoning techniques with human interaction aspects. In this paper, three systems that integrate these aspects are discussed [12] [13] [14].

The first of these systems facilitates natural language querying of video archive databases [12]. The underlying video data model allows identification of regions (bounding boxes), spatial relations between two bounding boxes, temporal relations in terms of intervals, and trajectories. Queries are processed in terms of *objects*, *attributes*, *activities* and *events* using information extraction techniques. This is especially relevant to structure the initial user query in semantic categories that facilitate the matching with available video detectors. The query processing is realized using a link parser [15] based on a light-parsing algorithm that builds relations between pairs of words, rather than constructing constituents in a tree-like hierarchy. This is sufficient for the specific kind of word groups considered in the system [12], but is limitative for more complex queries. In contrast, a typed dependencies parser, such as the Stanford Parser [16], facilitates the processing of complex queries and allows sentences to be mapped onto a directed graph representation. In this representation, the nodes represent words in the sentence and the edges represent the grammatical relations. Moreover, the query expansion in this system [12] could benefit from a semantically richer knowledge base than WordNet [8], such as ConceptNet [2] , which is a large knowledge base constructed by combining multiple web sources, such as DBpedia [17], Wiktionary [18] and WordNet [8].

The Never Ending Image Learner (NEIL) proposed in [13] is a massive visual knowledge base that runs 24 hour a day to extract semantic content from images on the Web in terms of *objects*, *scenes*, *attributes* and their *relations*. The longer NEIL runs, the more relations between concepts detected in the images it learns. NEIL is a general-purpose system and is based on learning new concepts and relations that are then used to augment the knowledge of the system. In this way, it continuously builds better detectors and consequently improves the semantic understanding of the images. NEIL aims at developing visual structured knowledge fully automatically without human effort. However, especially in semantic reasoning, lots of knowledge stays implicit in the user's mind. Therefore, it is desirable to provide the user with mechanisms to generate feedback to improve the semantic understanding of the system. Besides the lack of a user interface for collecting feedback, NEIL does not detect *actions*. Moreover, although NEIL considers an interesting set of semantic relations, such as taxonomy (*IsA*), partonomy (*Wheel is part of Car)*, attribute associations (*Round_shape is attribute of Apple* and *Sheep is White)*, and location relations (*Bus is found in Bus_depot*), most of the relations learned so far are of the basic type *IsA* or *LooksSimilarTo*.

The work in [14] presents a video search system for retrieving videos using complex natural language queries. This system uses the Stanford Parser [16] to process the user sentences in terms of *entities*, *actions*, *cardinalities*, *colors* and *action modifiers*, which are captured into a semantic graph that is then matched to available visual concepts. Spatial and semantic relations between concepts are also considered. The system [14] is not general-purpose, but tailored for a use case of autonomous driving, which provides sufficient complexity and challenges for the video detection. This includes dynamic scenes and several types of objects. This use case limits the semantic search capability to the set of concepts that are relevant, resulting in five entity classes, i.e., *cars*, *vans*, *trucks*, *pedestrians* and *cyclists*, and fifteen action classes, such as *move*, *turn*, *park* and *walk*. The semantic graph provides an intuitive and insightful way to present the underlying reasoning to the users.

III. APPLICATION

Our application is a general-purpose search engine that allows users to pose natural language queries in order to retrieve corresponding images. In this paper, we show two visual environments in which the application has been used. As a first environment, a camera is pointed at a table top on which toy sized objects can be placed to resemble real objects. Images of (combinations of) these objects can be manually taken and sent in real time to a database. In this environment, 42 concepts and 11 attributes, which are colors, are trained using sample images in the same environment. This number can grow, because of the ability to learn new concepts.

As a second environment, we tap into highway cameras. From these cameras, images are taken continuously and are automatically processed and stored by the image classification system. At this moment, only one concept (*car*) can be detected. Up to 12 colors are available for these cars. This environment can for example be used in

 ISBN: 978-1-61208-445-9

applications for police or defense organizations, such as following suspect cars or searching for specific accidents.

Two main use cases are supported. Firstly, users can search in historical data. Secondly, real-time images can be retrieved using notifications on outstanding queries. In the next section, we will focus on the semantic reasoning in this application.

IV. SEMANTIC REASONING

Figure 1 shows an overview of the system in which green and blue parts represent the components that realize the semantic reasoning, yellow parts represent the components dedicated to the image classification task and the white parts represent external components. Information about the image classification task is out of the scope of this paper, but elaborated in [4]. In the image classification, the semantics of Computer Vision is captured when the pixels are translated into annotated images.

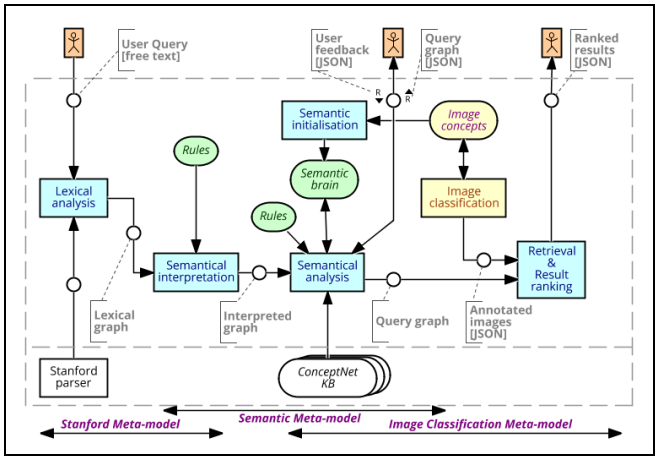


Figure 1.System overview

The input for the GOOSE system is a user query in natural language. The query is passed through four modules, while a fifth module takes care of initializing the system and learning new concepts. In the first stage, the query is sent to the Lexical Analysis module that parses it using the Stanford Parser [16], as opposed to the light link parser in [12]. The Stanford Parser returns a lexical graph, which is used as input to the Semantic Interpretation module. In this module, a set of rules is used to transform the lexical elements of the Stanford meta-model into semantic elements *objects*, *attributes*, *actions*, *scenes* and *relations*. The interpreted graph is sent to the Semantic Analysis module that matches the graph nodes against the available image concepts. If there is no exact match, the query is expanded using an external knowledge base, i.e., ConceptNet, to find a close match. The interpretation resulting from the Semantic Analysis is presented as a query graph to the user, who can interactively provide feedback used to gradually augment the Semantic Brain of the system, as inspired by NEIL [13]. The interactive part reflects the semantics of the Human Media Interaction. The query graph is inspired by the system in [14], which is, in contrast to our system, a domain-specific system. The query graph is also used as input for the Retrieval & Result ranking module, which provides the final result to the user. In the following subsections the complete process is described in detail using the example query *find an animal that is standing in front of the yellow car*.

A. Semantic Initialisation

This module provides an initial semantic capability by populating the Semantic Brain with image concepts (*objects*, *actions*, *scenes* and *attributes*) that the image classification part is capable of detecting. It also handles updates to the Semantic Brain following from new or modified image classification capabilities.

B. Lexical Analysis

In the Lexical Analysis module, the user query is lexically analyzed using the Typed Dependency parser (englishPCFG) of Stanford [16]. Before parsing the query, all tokens in the query are converted to lower case. In the example of *find an animal that is standing in front of the yellow car*, the resulting directed graph from the Lexical Analysis is shown in Figure 2.

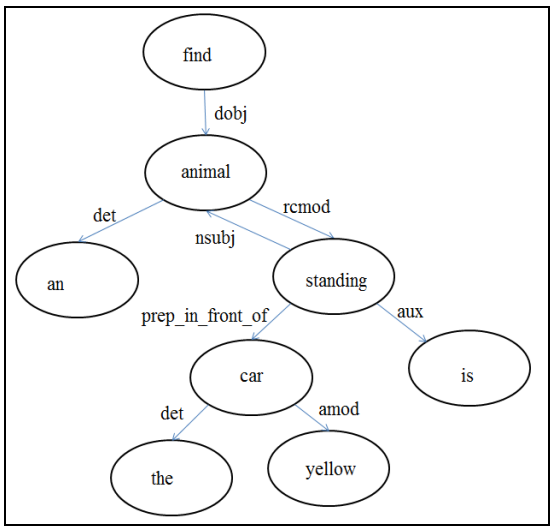


Figure 2. Lexical Graph

C. Semantic Interpretation

Since GOOSE is positioned as a generic platform, its semantics should not depend on, or be optimized for, the specifics of one single domain. Instead, we apply a generic ontological commitment by defining a semantic meta-model, shown in Figure 3, which distinguishes objects that might (i) bear attributes (*a car having a yellow color*), (ii) take part in actions (*running*), (iii) occur in a scene (*outside*), and (iv) have relations with other objects, in particular ontological relations (*a vehicle subsumes a car*), spatial relations (*an animal in front of a bus*), and temporal relations (*a bus halts after driving*). This meta-model is inspired by [12] [13] [14].

 ISBN: 978-1-61208-445-9

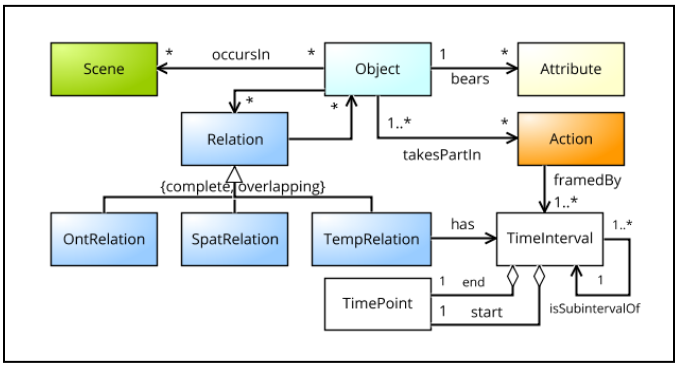


Figure 3. Semantic Meta-model

In the Semantic Interpretation module, a set of rules is used to transform the elements from the lexical graph into *objects*, *attributes*, *actions*, *scenes* and *relations*, according to the semantic meta-model in Figure 3. These rules include the following examples:

- Derive *cardinality* from a *determiner* (*det* in Figure 2), e.g., *the* in a noun in the singular form indicates a cardinality of 1, while *a/an* indicates at least 1;
- Derive *attributes* from *adjectival modifiers* (*amod* in Figure 2), i.e., adjectival phrases that modify the meaning of a noun;
- Derive *actions* from *nominal subjects* and *direct objects* (*nsubj* and *dobj* in Figure 2), i.e., the subject and object of a verb, respectively;
- Actions that represent the query command, such as *find*, *is*, *show* and *have*, are replaced on top of the tree by the subject of the sentence.

The output of the Semantic Interpretation for *find an animal that is standing in front of the yellow car* is shown in Figure 4. This is the basis of the query graph.

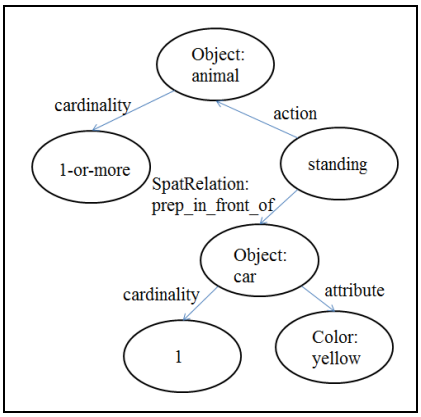


Figure 4. Interpreted Graph

D. Semantic Analysis

In the Semantic Analysis module, the elements from the interpreted graph, which are the query concepts, need to be matched against the concepts that can be detected by the image analysis component. The concepts that can be detected are represented by a label and stored in the system as image concepts. During the semantic analysis, the query concepts are matched against the image concepts in the Semantic Brain. If none of the objects or attributes can be detected by the image analysis module, the query concepts are expanded using ConceptNet. ConceptNet is used as opposed to WordNet in [12], because it has a more extensive knowledge base. Concept expansion is performed as follows: ConceptNet 5.2 [2] is accessed using the REST API and, among all the possible relations, we select the *IsA* relations (*OntRelation* in Figure 4) for objects, scenes or attributes, and the *Causes* relations (*TempRelation* in Figure 4) for actions. If one of the expanded objects, scenes or attributes has an exact match to one of the image concepts, that concept is added to the query graph with its corresponding relation. If there is still no match, the expansion cycle is repeated a second time. In this way, if there is, for example, no corresponding image concept for *Volvo*, this can anyway be expanded to the *car* image concept. However, when a *car* image concept is not available, the query will be further expanded in the second stage to the *vehicle* image concept. At this moment, we do not expand further than 2 iterations due to its combinatorial explosion, any potential cyclic concepts and its increasing semantic inaccuracy. The expansions, notably those originating from *IsA*, can be directed into both generalizing and specializing fashion, such that expansion of *animal* results both in *cow* as well as *creature*. Therefore, in the expanded query graph as visualized for the user, the *IsA* arrow stands for *expanded to* as opposed to a direction of subsumption. An example of the Query Graph for *find an animal that is standing in front of the yellow car* is shown in Figure 5.

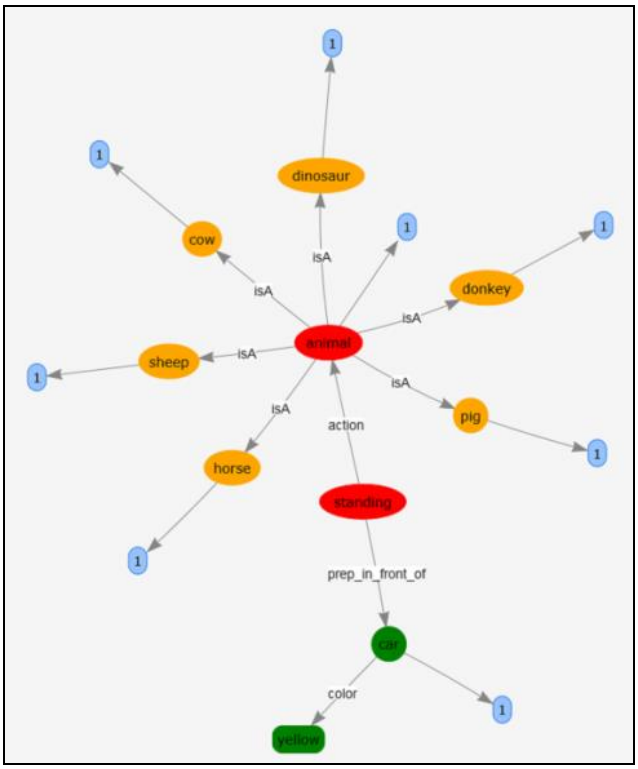


Figure 5.Example of an Expanded Query Graph

In the visualization, green colored nodes are query concepts that have a direct match with an available image concept. Red colored nodes represent query concepts that cannot be matched against an available image concept. Orange colored nodes represent suggested interpretations of query concepts using ConceptNet. For these concepts, it is uncertain whether they convey the user's intent and, therefore, require feedback from the user. Blue colored nodes represent the cardinality of the concept, e.g., the number of instances of this concept that is requested in the query. Relations between the concepts are depicted using labeled connections between the nodes corresponding to the concepts.

E. Retrieval and Ranking

The retrieval and ranking function need to be able to take into account the interpreted cardinality and attributes of the concepts in the query. This module retrieves those images that contain one or more concepts required by the query graph and excludes those that contain concepts that are explicitly not required (*a bicycle and no car*). The retrieval function is non-strict to ensure that there are a sufficient number of images that can be returned to the user.

The ranking on the images is based on concept, cardinality and attribute match. The motivation is that the most important elements in image search are the concepts that need to be found. For example, if a person searches for *a red car*, then it is more important that the *car* is visible in the image, and to a lesser extent whether this car is indeed *red*. Of course, this is also dependent on the context of the application.

The ranking function is penalty based. The image is included in the results if all requested concepts are present. The inverse of the confidence of the classifier for the concept, which is a value between zero and one, is taken as the basic penalty. This means that a high confidence gives a low penalty. For each concept requested in the query, a penalty of 0.5 is added to the basic penalty if 1) an attribute of a concept in the image does not match the attribute requested in the query and 2) when the image contains too few instances of the requested concept. The image is, however, not penalized if it contains too many instances. This is a choice that is also dependent on the application area. The lowest penalized images are displayed first in the results list.

Figure 6 shows the ranked result list for the query *find an animal that is standing in front of the yellow car*. In the results, we see that all images that contain a yellow car are ranked higher than those images that contain a car, but of which the attribute color is wrong, i.e., red. Even images with multiple cars of which one car is yellow are ranked higher than the images with a single car that is red. This is coherent with the interpretation of *a yellow car*, since the query states nothing explicitly about the exclusion of cars of different colors. Images with multiple cars and an animal are ranked lower than most of the images that contain a single yellow car and an animal, because the classifier is less confident that it has indeed observed a yellow car among the five vehicles in the picture.

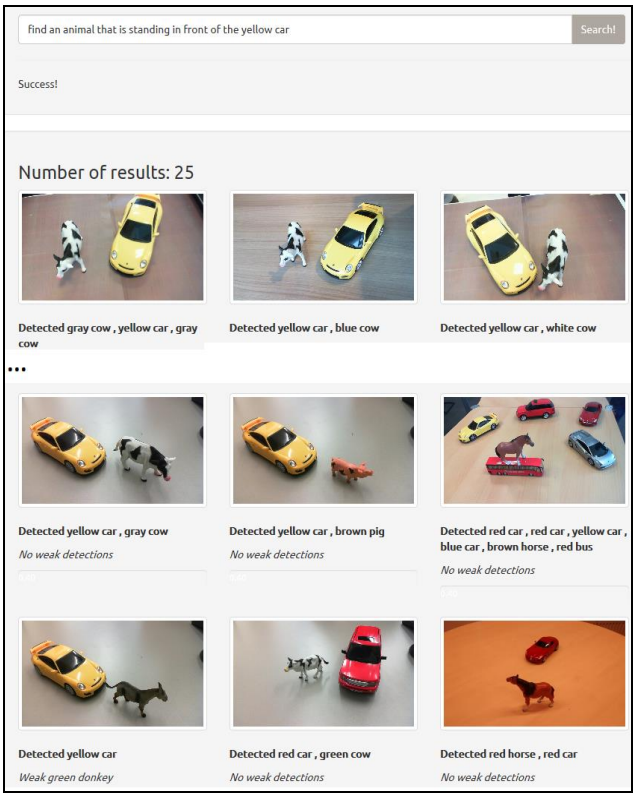


Figure 6. Result for Example Query

V. DISCUSSION

During the implementation of the system we encountered various obstacles. When using semantic analysis as a method to understand which components need to be present in an image, it is important to keep in mind the limitations of the image classification.

The use of spatial relations in image search is a challenge. In terms of image analysis, objects can be positioned relative to each other using bounding boxes. An object can be left of, right of, above, under or overlapping. But, how should we interpret spatial relations such as *in front of* in the user query? With the query *animal in front of the yellow car*, is the interpretation of *in front of* based on the perspective of someone in the car, or based on the perspective of someone looking at the picture? In the former case, this would mean that an image with an animal that is to the side of the front of the car needs to be ranked higher, while in the latter case an animal that is closer to the observer (whether it is at the front, the back or the side) needs to be ranked higher. Depending on which interpretation the user wishes, the image classifier may have a higher burden, because it would need to analyze the orientation of the car, and detect on which side the front of the car is.

An additional complication concerns prepositions, such as *with*, that have ambiguous meaning. For example, the query *the woman with the red dress* is most likely interpreted as *the woman wearing a red dress*. From an image detection point of view this can be seen as *a woman who is for a large part red*. On the other hand, in the case of *the woman with the*

 ISBN: 978-1-61208-445-9

dog, the interpretation of the two concepts cannot be merged. One possible solution would be to take the type of object into account (*a dress is clothing*).

Query expansion can also be a complicating factor. ConceptNet sometimes does not provide the expected relations. For example, no *IsA* relation between *animal* and *cow* exists (but a *RelatedTo*). On the other hand, a relation between *Mercedes* and *person* and *animal* is available, which should be filtered if one is looking for a car. The specific dataset that is used plays a role here. Manual additions that are specific to the dataset under consideration can be meaningful to ensure that all relevant concepts can be found during query expansion.

The combination of attributes is another point of discussion. Again, this is difficult both from the point of view of the user as well as the capabilities of the image classification. For example, the query *blue and red car* can mean that someone is searching for an image with a blue car and a red car, or that one is searching for an image with a car that is partly blue and partly red. In order to provide the required results, these kinds of ambiguities can be detected and resolved before the image classification by requesting the user to identify the correct interpretations out of the possible ones. The image classifier that we used was capable of attributing only one color to each concept, making the second interpretation impossible to detect in images.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, a prototype of the GOOSE system is presented. We have shown that state-of-the-art techniques from Computer Vision, Human Media Interaction and Semantic Modelling can be combined and used in an application, while at the same time pinpointing several challenges. In the semantic part of the system, the user query is transformed into a query graph through semantic reconciliation using the Stanford Parser and ConceptNet, their meta-models and a semantic meta-model. This query graph is presented to the user in an intuitive and insightful way in order to collect feedback that is used for further reasoning and system learning.

In the future, the user should be able to have user-specific entries in the Semantic Brain. Good and bad query expansions and results are subjective and, therefore, need user-specific care.

An additional point to be further investigated concerns the semantic interpretation of the image classifiers. Here, potentially ambiguous names were used to identify these classifiers. This is in particular an issue when dealing with user-generated classifiers.

Finally, an evaluation of the semantic interpretation and result ranking modules of the GOOSE system should be performed in the future in order to validate our approach and show that our implementation can handle simple and complex queries in different domains.

ACKNOWLEDGMENT

This research was performed in the context of the GOOSE project, which is jointly funded by the enabling technology program Adaptive Multi Sensor Networks (AMSN) and the MIST research program of the Dutch Ministry of Defense.

REFERENCES

[1] R. Speer and C. Havasi "Representing general relational knowledge in conceptnet 5" LREC, 2012, pp. 3679-3686.

[2] K. Schutte et al. "GOOSE: semantic search on internet connected sensors", Proc. SPIE, vol. 8758, 2013, pp. 875806.

[3] K. Schutte et al. "Interactive detection of incrementally learned concepts in images with ranking and semantic query interpretation", unpublished.

[4] H. Bouma, P. Eendebak, K. Schutte, G. Azzopardi, G. Burghouts "Incremental concept learning with few training examples and hierarchical classification", unpublished.

[5] J. Schavemaker, M. Spitters, G. Koot, M. de Boer "Fast re-ranking of visual search results by example selection", unpublished.

[6] J. Deng, W. Dong, R. Socher, L. Li, K. Li, L. Fei-Fei "Imagenet: a large-scale hierarchical image database," in IEEE CVPR, 2009, pp. 248–255.

[7] P. Perona "Visions of a Visipedia", Proc. of IEEE, 98.8, 2010, pp. 1526-1534.

[8] C. Fellbaum "WordNet", Blackwell Publishing Ltd, 1998.

[9] L. Bai1, S. Lao, G.J.F. Jones, and A.F. Smeaton "Video Semantic Content Analysis based on Ontology". In Int. Machine Vision and Image Processing Conf., 2007, pp. 117-124.

[10] A.D. Bagdanov, M. Bertini, A. Del Bimbo, G. Serra, C. Torniai "Semantic annotation and retrieval of video events using multimedia ontologies". In Int. Conf. on Semantic Computing, 2007, pp. 713-720.

[11] A. Oltramari and C. Lebiere "Using Ontologies in a Cognitive-Grounded System: Automatic Recognition in video Surveillance", In Proc. 7 Int. Conf. on Semantic Technology for Intelligence, Defense, and Security, 2012.

[12] G. Erozel, N. K. Cicekli, I. Cicekli "Natural language querying for video databases," Information Sciences, 178.12, 2008, pp. 2534–2552.

[13] X. Chen, A. Shrivastava, and A. Gupta "NEIL: Extracting Visual Knowledge from Web Data", IEEE Int. Conf. on Computer Vision, 2013, pp. 1409-1416.

[14] D. Lin, S. Fiedler, C. Kong, R. Urtasun "Visual Semantic Search: Retrieving Videos via Complex Textual Queries". In IEEE Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2657-2664.

[15] D. Sleator and D. Temperley, "Parsing English with a Link Grammar", In Third Int. Workshop on Parsing Technologies, 1993.

[16] M.-C. de Marneffe, B. MacCartney, C. D. Manning. "Generating Typed Dependency Parses from Phrase Structure Parses", LREC, 2006, pp. 449-454.

[17] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives "Dbpedia: A nucleus for a web of open data" Springer Berlin Heidelberg, 2007, pp. 722-735.

[18] E. Navarro et al, "Wiktionary and NLP: Improving synonymy networks". In Proc. of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources, 2009, pp. 19-27, Association for Computational Linguistics.

 ISBN: 978-1-61208-445-9

Plant Leaves Classification

Mohamed Elhadi Rahmani, Abdelmalek Amine, Mohamed Reda Hamou

GeCoDe Laboratory
Department of computer Science
University of Dr. Tahar Moulay
Saida, Algeria
r_m_elhadi@yahoo.fr; amine_abd1@yahoo.fr; hamoureda@yahoo.fr

Abstract:- **A leaf is an organ of vascular plant and is the principal lateral appendage of the stem. Each leaf has a set of features that differentiate it from the other leaves, such as margin and shape. This paper proposes a comparison of supervised plant leaves classification using different approaches, based on different representations of these leaves, and the chosen algorithm. Beginning with the representation of leaves, we presented leaves by a fine-scale margin feature histogram, by a Centroid Contour Distance Curve shape signature, or by an interior texture feature histogram in 64 element vector for each one, after we tried different combination among these features to optimize results. We classified the obtained vectors. Then we evaluate the classification using cross validation. The obtained results are very interesting and show the importance of each feature.**

Keywords:- Plants leaves classificatin; supervised classification; KNN; Decision tree; Naïve Bayes.

I. INTRODUCTION

For all forms of life, plants form the basic food staples, and this is just one reason why plants are important. They are the major source of oxygen and food on earth since no animal is able to supply the components necessary without plants. The fish we eat consume algae and the cattle we eat as beef feed on grass, so even if you are not a fan of salads, your food source relies on plants. Plants also provide animals with shelter, produce clothing material, medicines, paper products, reduce noise levels and wind speed, reduce water runoff and soil erosion. Coal is also produced from plant materials that were once alive. All that gives plants its important role in life on earth. For example, as natural resource managers, they must understand what they manage, and plant identification is a key component of that understanding. The ability to know, or identify plants allows them to assess many important rangeland or pasture variables that are critical to proper management: range condition, proper stocking rates, forage production, wildlife habitat quality, and rangeland trend, either upward or downward. Natural resource managers, especially those interested in grazing and wildlife management must be able to evaluate the presence or absence of many plant species in order to assess these variables.

In nature, plant leaves are two dimensional containing important features that can be useful for classification of various plant species, such as shapes, colours, textures and structures of their leaf, bark, flower, seedling and morph. According to Bhardwaj et al. [8], if the plant classification is based on only two dimensional images, it is very difficult to study the shapes of flowers, seedling and morph of plants because of their complex three dimensional structures.

The present paper proposes a comparison of the classification of different representation of plant leaves based on its margin, shape and textures; we used for each representation different classical supervised data mining algorithms. The organization of this paper is given as follows: Section 2 provides an overview of the related works; Section 3 gives details about dataset used in our experiment, Section 4 presents used approaches, discussion of the results will show in Section 5, and finally Section 6 gives the overall conclusion and the scope for future research.

II. RELATED WORK

Recently, plant classification became one of major researches. Shanwen et al. [2] used a combination between semi-Supervised locally linear embedding (semi-SLLE) and KNN algorithms for plant classification based on leaf images and showed its performance. James Cope et al. [6] presented plant texture classification using Gabor co-occurrences; where joint distributions for the responses from applying different scales of the Gabor filter are calculated. The difference among leaf textures is calculated by the Jeffrey divergence measure of corresponding distributions. Also Kadir et al. in [3] incorporates shape and vein, colour, and texture features to classify leaves using probabilistic neural network and proves that it gives better result with average accuracy of 93.75%. Plant leaf images corresponding to three plant types, are analysed using two different shape modelling techniques in Chaki et al. [5], authors proposed an automated system for recognizing plant species based on leaf images. One of the last works released by Bhardwaj in [8], that presented a simple computational method in computer vision to recognize plant leaves and to classify it using K-nearest neighbours. Anang Hudaya also worked on plant classification in his paper [18], presenting a scalable approach for classifying plant leaves using the 2-dimensional shape feature, using distributed hierarchical graph neuron (DHGN) for pattern recognition and k-nearest neighbours (k-NN) for pattern classification.

 ISBN: 978-1-61208-445-9

III. DATASET

The 'Leaves' dataset contains one-hundred species of leaves [7], each species represented by three 64 element vector for each of three distinct features extracted from images: a fine-scale margin feature histogram, a Centroid Contour Distance Curve shape signature, and an interior texture feature histogram. This dataset contains 1600 samples, whereas there are sixteen distinct specimens for each species, photographed as a colour image on a white background. Figure 1 shows the first 27 species from the dataset.

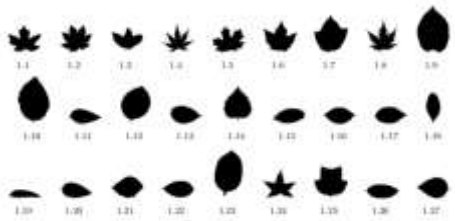


Figure 1. A silhouette image of one plant specimen each from the challenging one-hundred species leaves data set.

The data set inherently consists of having a wide set of classes with a low number of samples. Additionally, many sub species resemble the appearance of other major species, as well as many sub species with a major species can resemble a radically different appearance [7].

IV. PROPOSED APPROACHES

The present work shows a comparison of classification of seven different representations of plant leaves using three features extracted from the images; Figure 2 shows the architecture of proposed approaches:

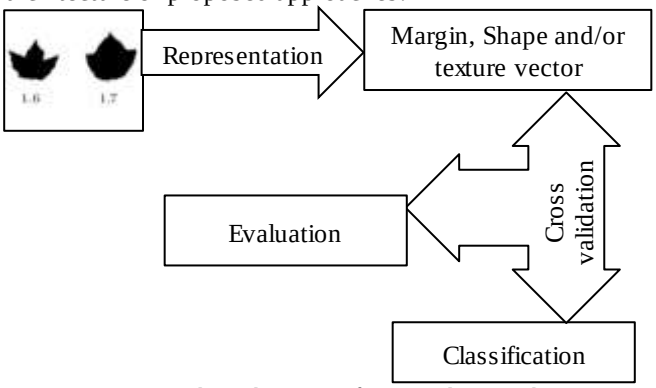


Figure 2. The architecture of proposed approaches

A. Representation of samples

Beginning with representation of species by three features extracted from images: a fine-scale margin feature histogram, then a Centroid Contour Distance Curve shape signature, and finally an interior texture feature histogram. We put values of each feature in 64 elements vector, then we tried to combine these vectors two by two in one 128 elements vector, and finally we presented species combining the three vectors together in one 192 elements vector.

B. Classification

In each case, we used three different approaches for classification: probabilistic approach using Naïve Bayes algorithm, hierarchical approach using Decision Tree C4.5 algorithm, and finally, approach based on distance calculation using K-nearest neighbours (K-NN) algorithm with k = 3, 4, 5, 6, or 7 and using Euclidian distance.

C. Evaluation

To evaluate classification, we used 10-folds cross validation. Training and testing sets are performed 10 times by partitioning randomly the dataset into 10 mutually in iteration exclusive subsets or "folds"; i, a subset Di is reserved as the test set, and the remaining partitions are collectively used to train the model.

D. Calculated measures for the evaluation

To calculate different metrics used for evaluation of classification, we have to introduce other measures:

1) True Positive (TP) present the average of the vectors that are correctly predicted relevant obtained in each iteration

2) True Negative (TN) present the average of the vectors that are correctly predicted as not relevant obtained in each iteration

3) False Positive (FP) present the average of the vectors that are predicted relevant but they are not relevant obtained in each iteration

4) False Negative (FN) present the average of the vectors that are correctly predicted not relevant but they are relevant obtained in each iteration

Using these four measures, we calculated the most famous measures that are used to evaluate classification algorithms:

1) *For classification, the accuracy estimate is the is the overall number of correct classifications from the 10 iterations, divided by the total number of tuples in the initial data.[16]*

- Accuracy = (TP + TN) / (TP + FP + TN + FN)

2) Precision and recall are the measures used in the information retrieval domain to measure how well an information retrieval system retrieves the relevant elements requested by a user. The measures are defined as follows[17]

- Precision = (TP) / (TP + FP)
- Recall = (TP) / (TP + FN).

3) Instead of two measures, they are often combined to provide a single measure of retrieval performance called the F-measure as follows[17]

- F-measure = (2 * Recall * Precision) /(Recall + Precision)

 ISBN: 978-1-61208-445-9

V. RESULTS AND DISCUSSION

The following section shows the different results obtained for each representation with each algorithm.

TABLE I. RESULTS OBTAINED BY CLASSIFICATION OF SPECIES REPRESENTED BY THE MARGIN EXTRACTED FROM IMAGES

Algorithms	Evaluation %			
	Accuracy	Precision	Recall	Fmeasure
Naïve Bayes	**85.125**	**85.9**	**85.1**	**85.5**
Decision Tree	47.437	48.2	47.4	47.7
K-NN k=3	75.5	77.1	75.5	76.2
K-NN k=4	76.5	77.9	76.5	77.2
K-NN k=5	77.062	78.3	77.1	77.7
K-NN k=6	75.75	77.3	75.8	76.5
K-NN k=7	77.312	78.4	77.3	77.8

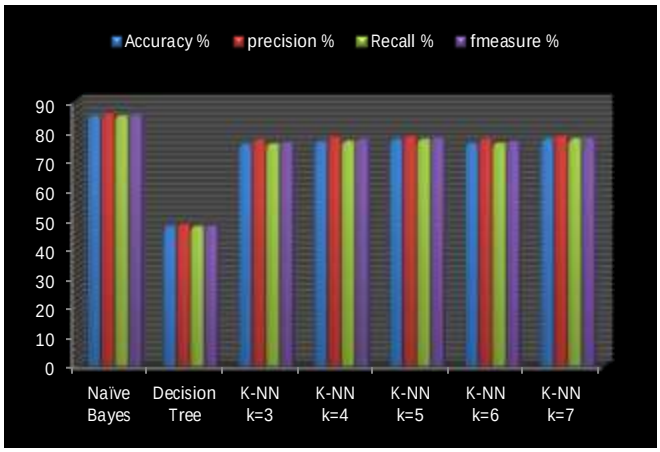


Figure 3. Results obtained by classification of species represented by the margin extracted from images

Table 1 and Figure 3 show the obtained results of the classification of species represented by the margins extracted from images where samples are 64 elements vectors. As we see, the probabilistic approach using Naïve Bayes gives best result compared with the approach based on distance calculation using K-Nearest Neighbours where the initial K (=3, 4, 5, 6, or 7) value can affect the result even a little, otherwise, hierarchical classification approach using Decision Tree gives the worst results.

TABLE II. RESULTS OBTAINED BY CLASSIFICATION OF SPECIES REPRESENTED BY THE SHAPE OF LEAVES

Algorithms	Evaluation %			
	Accuracy	Precision	Recall	Fmeasure
Naïve Bayes	52.625	53.8	52.6	53.2
Decision Tree	42.125	42	42.1	42
K-NN k=3	60.437	61.9	60.4	61.1
K-NN k=4	**61.187**	**62.5**	**61.2**	**61.8**
K-NN k=5	59	60.9	59	59.9
K-NN k=6	57.562	59.6	57.6	58.6
K-NN k=7	56.937	58.4	56.9	57.6

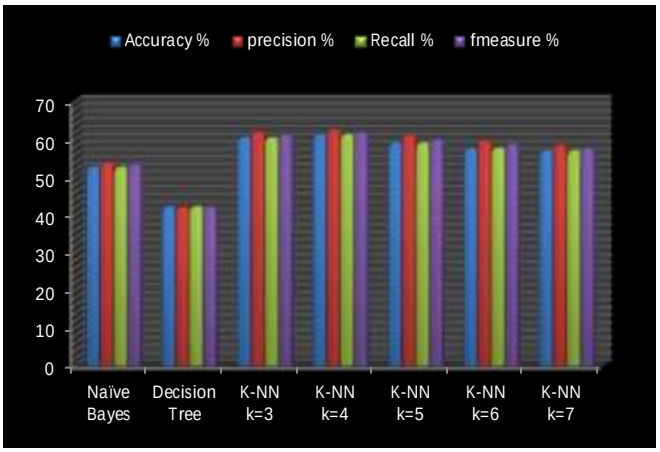


Figure 4. Results obtained by classification of species represented by the shape of leaves

In Table 2, unlike the margin representation, representation of leaves by its shape gives totally different results, in which approach based on distance calculation gives better result than probabilistic approach; even initial K value effect is more visible as we see in Figure 4.

TABLE III. RESULTS OBTAINED BY CLASSIFICATION OF SPECIES REPRESENTED BY THE TEXTURE EXTRACTED FROM IMAGES

Algorithms	Evaluation %			
	Accuracy	Precision	Recall	Fmeasure
Naïve Bayes	74.359	77.5	74.4	75.9
Decision Tree	51.97	52.9	52	52.4
K-NN k=3	76.923	78.2	76.9	77.5
K-NN k=4	76.673	78.1	76.7	77.3
K-NN k=5	**76.923**	**78.4**	**76.9**	**77.6**
K-NN k=6	76.548	78	76.5	77.2
K-NN k=7	76.86	78.4	76.9	77.6

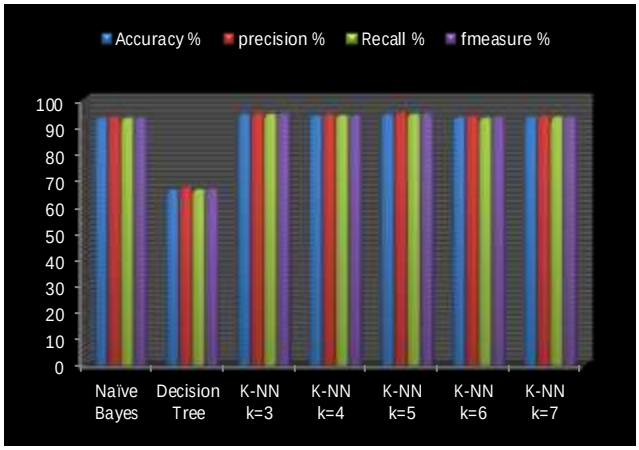


Figure 5. Results obtained by classification of species represented by the Texture extracted from images

In Figure 5 and Table 3, representation by texture vectors extracted from images gives convergent results, either between K-nearest neighbours with different initial K value, or between K-nearest neighbour and Naïve Bayes algorithm.

 ISBN: 978-1-61208-445-9

For hierarchical algorithm, it gives better results than the representation of leaves by margins or shape.

In order to optimize obtained results, we used to combine these features, where we get more efficiency in classification; the following Tables and Figures prove this idea. So, the supposed question here is, which combination can give the best result?

Beginning with combination between margin vector and shape vector in one 128 elements vector to represent each leaf, the obtained results are shown in Table 4.

TABLE IV. RESULTS OBTAINED BY CLASSIFICATION OF SPECIES REPRESENTED BY THE COMBINATION OF MARGIN AND SHAPE

Algorithms	Evaluation %			
	Accuracy	Precision	Recall	Fmeasure
Naïve Bayes	93.187	93.7	93.2	93.4
Decision Tree	66.187	67.1	66.2	66.6
K-NN k=3	**94.687**	**95.1**	**94.7**	**94.9**
K-NN k=4	94.187	94.7	94.2	94.4
K-NN k=5	94.687	95.2	94.7	95
K-NN k=6	93.5	94.2	93.2	93.8
K-NN k=7	93.562	94.2	93.6	93.8

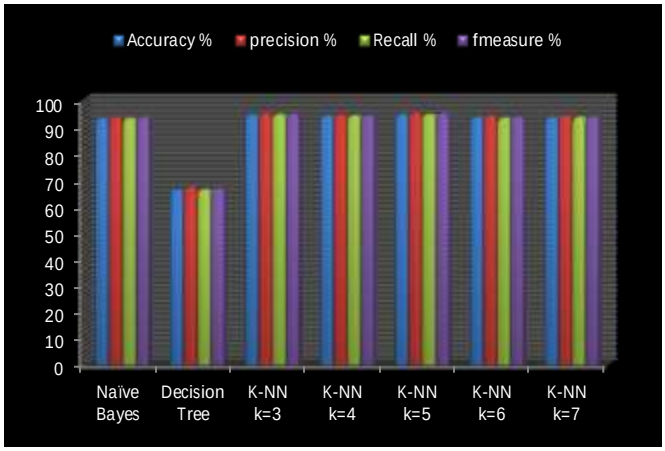


Figure 6. Results obtained by classification of species represented by the combination of margin and shape

In Figure 6 and Table 4, results given by Naïve Bayes algorithm and K-Nearest Neighbour are converged, and better than previous results (+90% of accuracy); even Decision Tree algorithm gives better result (almost 67% of accuracy).

TABLE V. RESULTS OBTAINED BY CLASSIFICATION OF SPECIES REPRESENTED BY THE COMBINATION OF MARGIN AND TEXTURE

Algorithms	Evaluation %			
	Accuracy	Precision	Recall	Fmeasure
Naïve Bayes	86.687	88.4	86.7	87.5
Decision Tree	59.375	59.9	59.1	59.5
K-NN k=3	**92**	**92.5**	**92**	**92.2**
K-NN k=4	91.562	92.1	91.6	91.8
K-NN k=5	91.312	92	91.3	91.6
K-NN k=6	91.25	91.9	91.3	91.6
K-NN k=7	91	91.8	91	91.4

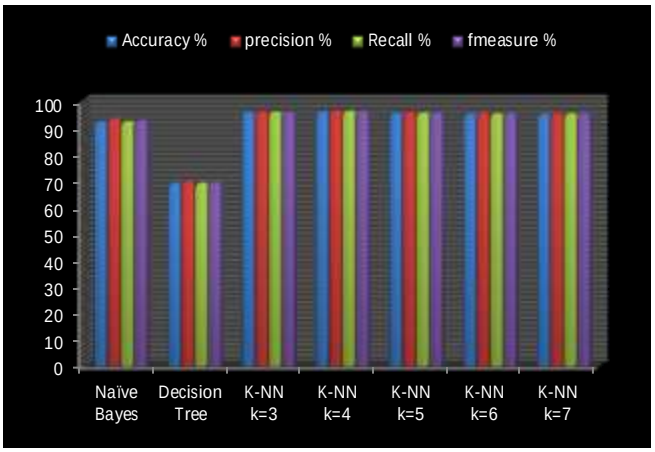


Figure 7. Results obtained by classification of species represented by the combination of margin and texture

Table 5 and Figure 7 show that obtained accuracy decreases compared with the combination between margin and shape, especially Naïve Bayes (from 93% to 87% of accuracy) and Decision Tree (from 66% to 59%).

TABLE VI. RESULTS OBTAINED BY CLASSIFICATION OF SPECIES REPRESENTED BY THE COMBINATION OF SHAPE AND TEXTURE

Algorithms	Evaluation %			
	Accuracy	Precision	Recall	Fmeasure
Naïve Bayes	84.25	86.5	84.3	85.3
Decision Tree	61.5	62.2	61.5	61.8
K-NN k=3	**87.687**	**88.6**	**87.7**	**88.1**
K-NN k=4	87.187	88	87.2	87.6
K-NN k=5	87	87.7	87	87.3
K-NN k=6	86.25	87.4	86.3	86.8
K-NN k=7	85.875	87	85.9	86.4

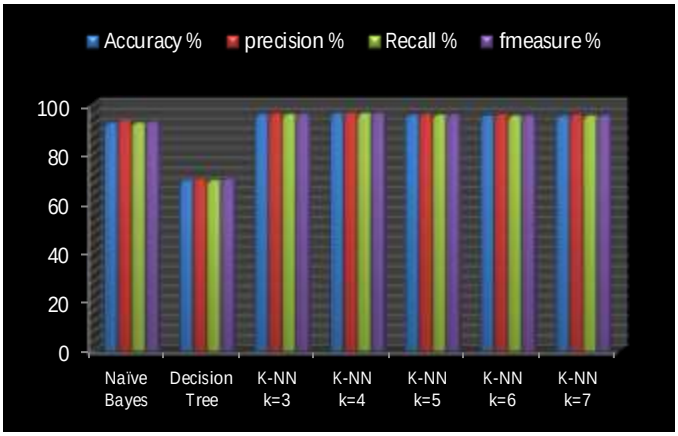


Figure 8. Results obtained by classification of species represented by the combination of margin and shape

In Table 6 and Figure 8, we see clearly that all algorithms had less performance (-90% of accuracy), where K-Nearest Neighbour gives the best result in this case.

We tried to improve results shown in Table 4 by combining the three features in one 192 elements vector in

order to represent each sample, and we got better performance as demonstrated in Table 7 and Figure 9:

TABLE VII. RESULTS OBTAINED BY CLASSIFICATION OF SPECIES REPRESENTED BY THE COMBINATION OF THE THREE FEATURES

Algorithms	Evaluation %			
	Accuracy	Precision	Recall	Fmeasure
Naïve Bayes	92.437	93.5	92.4	92.9
Decision Tree	69.125	69.8	69.1	69.4
K-NN k=3	95.937	96.2	95.9	96
K-NN k=4	**96.25**	**96.5**	**96.3**	**96.4**
K-NN k=5	95.625	96	95.6	95.8
K-NN k=6	95.312	95.8	95.3	95.5
K-NN k=7	95.25	95.7	95.3	95.4

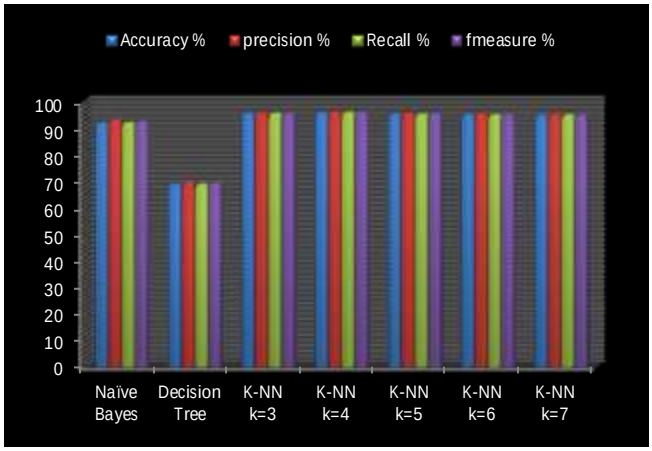


Figure 9. Results obtained by classification of species represented by the combination of the three features

Table 7 and Figure 9 prove that the combination of the three features extracted from images gives the best result in all tested representations; except Naïve Bayes that gives lower accuracy than the classification of vectors containing margin and shape values.

VI. CONCLUSION

Plants play an important role in our lives, without plants there will not be the existence of the ecology of the earth. The large amount of leaf types now makes the human being in a front of some problems in the specification of the use of plants, the first need to know the use of a plant is the identification of the plant leaf.

This work proposed a comparison of supervised classification of plant leaves, where we used to represent species in seven different representations, using three features extracted from binary masks of these leaves: a fine-scale margin feature histogram, by a Centroid Contour Distance Curve shape signature, and by an interior texture feature histogram. Results were very interesting in a way that gives as clear ideas:

- In term of representation: we can differentiate leaves by its margin better than shape or texture, but, experiments shown in this study prove our idea: the more we combine these features, the more precise the difference between samples is and that is what gives better results in classification.
- In term of classification: distance based algorithms give the best result for plant leaves classification. So, we can conclude that these algorithms are the most suitable for that task. On the other hand, the approach based on decision tree gives the worst results because of the overfitting problem. In general, a learning algorithm is said to overfit relative to a simpler one if it is more accurate in fitting known data but less accurate in predicting new data.

Use of the three features proved that there is some information more important than other. We discovered that margin representation can affect results more than the shape of the leaf. However, the combination of the three features gives the best result. To solve this problem, we plan, as future work, use of feature extraction algorithms, like PSO, to clean dataset and keep the important information in order to optimize the obtained results and avoid overfitting problem posed by decision tree algorithm. We plan also to use bio-inspired algorithms. They are part of a new research domain that is becoming more important due to its results in different areas.

REFERENCES

[1] C. Mallah, J. Cope, and J. Orwell ,"Plant leaf classification using probabilistic integration of shape, texture and margin features," Signal Processing, Pattern Recognition and Applications, 2013

[2] S. Zhang, and K. W. Chau, "Dimension reduction using semi-supervised locally linear embedding for plant leaf classification," Emerging Intelligent Computing Technology and Applications, Springer Berlin Heidelberg, 2009, pp. 948-955

[3] A. Kadir, L. E. Nugroho, A. Susanto, and P. I. Santosa, "Leaf classification using shape, color, and texture features", arXiv preprint arXiv:1401.4447, 2013

[4] A. H. M. Amin, and A. I. Khan, "One-shot Classification of 2-D Leaf Shapes Using Distributed Hierarchical Graph Neuron (DHGN) Scheme with k-NN Classifier," Procedia Computer Science, 24, 2013, pp.84-96

[5] J. Chaki, and R. Parekh, "Plant leaf recognition using shape based features and neural network classifiers," International Journal of Advanced Computer Science and Applications (IJACSA), 2011, pp. 26-29

[6] J. S. Cope, P. Remagnino, S. Barman, and P. Wilkin, "Plant texture classification using gabor co-occurrences," In Advances in Visual Computing, Springer Berlin Heidelberg, 2010, pp.669-677

[7] C. Mallah, "Probabilistic Classification from a K-Nearest-Neighbour Classifier," Computational Research, 1(1), 2013, pp.1-9

[8] A. Bahrdwaj, M. Kaur, and A. Kumar, "Recognition of plants by Leaf Image using Moment Invariant and Texture Analysis," International Journal of Innovation and Applied Studies, 3(1), 2013, pp.237-248

[9] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," In Advances in neural information processing systems, 2005, pp.1473-1480

[10] A. R. Backes, W. N. Gonçalves, A. S. Martinez, and O. M. Bruno, "Texture analysis and classification using deterministic tourist walk," Pattern Recognition, 43(3), 2010, pp.685-694

[11] Q. P. Wang, J. X. Du, and C. M. Zhai, "Recognition of leaf image based on ring projection wavelet fractal feature," In Advanced Intelligent Computing Theories and Applications, with Aspects of Artificial Intelligence, Springer Berlin Heidelberg, 2010, pp.240-246

 ISBN: 978-1-61208-445-9

[12] T. Beghin, J. S. Cope, P. Remagnino, and S. Barman, "Shape and texture based plant leaf classification," In Advanced Concepts for Intelligent Vision Systems, Springer Berlin Heidelberg, January 2010, pp.345-353
[13] A. Kadir, L. E. Nugroho, A. Susanto, and P. I. Santosa, "A comparative experiment of several shape methods in recognizing plants," arXiv preprint arXiv:1110.1509, 2011
[14] N. Vallimmal, and S. N. Geethalakshmi, "Hybrid image segmentation algorithm for leaf recognition and characterization," In Process Automation, Control and Computing (PACC), 2011 International Conference on IEEE, 2011, pp.1-6
[15] K. Singh, I. Gupta, S. Gupta, "Svm-bdt pnn and fourier moment technique for classification of leaf shape," International Journal of Signal Processing, Image Processing and Pattern Recognition, 3(4), 2010, pp.67-78
[16] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and techniques, Morgan Kaufmann Publishers is an imprint of Elsevier, 2011
[17] C. Sammut, and G. I. Webb, "Encyclopedia of machine learning, " Springer Science & Business Media, 2011

 ISBN: 978-1-61208-445-9

From Linguistic Resources to Medical Entity Recognition: a Supervised Morpho-syntactic Approach

Maria Pia di Buono, Alessandro Maisto
and Serena Pelosi
Department of Political, Social and Communication Sciences
University of Salerno
84084 Fisciano, Italy
{mdibuono, amaisto, spelosi}@unisa.it

***Abstract*—Due to the importance of the information it conveys, Medical Entity Recognition is one of the most investigated tasks in Natural Language Processing. Many researches have been aiming at solving the issue of Text Extraction, also in order to develop Decision Support Systems in the field of Health Care. In this paper, we propose a Lexicon-grammar method for the automatic extraction from raw texts of the semantic information referring to medical entities and, furthermore, for the identification of the semantic categories that describe the located entities. Our work is grounded on an electronic dictionary of neoclassical formative elements of the medical domain, an electronic dictionary of nouns indicating drugs, body parts and internal body parts and a grammar network composed of morphological and syntactical rules in the form of Finite-State Automata. The outcome of our research is an Extensible Markup Language (XML) annotated corpus of medical reports with information pertaining to the medical Diseases, Treatments, Tests, Symptoms and Medical Branches, which can be reused by any kind of machine learning tool in the medical domain.**

***Index Terms*—Medical Entity Recognition, Lexicon-Grammar, Morphosemantics, Semi-automatically Generated Lexical Resources.**

I. Introduction

The necessity to access and integrate in real time health related data that come from multiple sources, opens plenty of opportunities for the research studies in Natural Language Processing (NLP). This comes together with the need of support in the extraction and in the management of useful information and in the development of systems, which must be able to give a structure to the semantic dimension of real words and sentences.

In this paper, we present a Lexicon-grammar (LG) method, that takes advantages from word combination rules and from the lexical and syntactic structures of the natural language. Our purpose is to locate and describe the meaning of phrases, sentences and even entire documents belonging to the medical domain.

In order to overcome the poor flexibility of the existing medical databases with respect to neologisms, we exploit many Morpho-semantic strategies, which can be crucial in the automatic definition of technical-scientific lexicons, in which the global meaning of the words presents strong connections with the meaning of the morphemes that compose them. In other words, we reorganize the information derived from the semantics of the word formation elements, by making the medical words derive the meaning of the morphemes with which they are formed. In this way, starting from a small number of indicators and without any dependence to limited knowledge bases, we can any time automatically build a technical-scientific dictionary of the medical text we process.

In this work, thanks to the opportunities offered by the productive morphology, we automatically locate and define the medical entities contained in a corpus of 989 medical reports. Moreover, using the theoretical insights of the LG framework, we use syntactic rules to semantically describe the categories (e.g., Disease, Treatment, Test, Symptom and Medical Branch) of the located entities. Therefore, if our starting point is a corpus of medical records in electronic format, the output of our research is a structured version of the same corpus, which can be easily reused and queried in every kind of machine learning tool, Clinical Decision Support System (CDSS) or NLP tool in the medical domain.

The paper is structured in the following way: Section II introduces the most important works on the identification and categorization of medical entities in free texts; Section III briefly describes the Lexicon-grammar framework, the Morpho-semantic approach and the tools we used to perform our tasks; in the Section IV, we introduce the automatically generated IMED dictionary and the set of syntactic rules applied to extract entity classes from Medical Records; the Section V describes the structure of the Medical Records Corpus used to test our approach and the results of the application of our method; in the end, Section VI presents the conclusions and the further developments of our research.

 ISBN: 978-1-61208-445-9

II. Related Works

The Medical Entity Recognition (MER) can be decomposed in two main tasks: the extraction of semantic information referring to medical entities from raw texts and the identification of the semantic categories that describe the located entities [1].

As regards the first task, many medical lexical databases (e.g., Medical Subject Headings (MeSH), RxNorm, Logical Observation Identifiers Names and Codes (LOINC), Systematized Nomenclature of Medicine (SNOMED), and Unified Medical Language System (UMLS), which includes all the other sources) can be used as knowledge base for the location of the medical entities.

Anyway, the quick evolution of entity naming and the slowness of the manual development and updating of the resources often make it necessary to exploit some word-formation strategies, that can be truly helpful in the automatic population of technical-scientific databases. Such strategies concern the Morpho-semantic approach and have been successfully applied to the medical domain by [2] on terminal morphemes into an English medical dictionary; by [3] on medical formative elements of Latin and Greek origin; by [4] on the suffix *-itis*; by [5] on suffixes *-ectomy* or *-stomy* and by [6] on the suffix *-osis*.

Among the most used tools for the MER, we mention MetaMap [7], a reference tool which recognizes and categorizes medical terms by matching noun phrases in free texts to the corresponding UMLS Metathesaurus and Semantic Network, and MEDSYNDIKATE [8], a natural language processor able to automatically acquire data from medical findings reports.

Examples of approaches based on the MetaMap knowledge base are the one of [9], which extracts medical entities from pathologist reports, and the one of [10], which focuses on the extraction of medical problems with an approach based on the MetaMap Transfer and the NegEx negation detection algorithm.

With reference to the second task, we can find in literature rule-based, statistical and hybrid approaches.

As regards the contributions that exploit statistical methods for the identification and classification of medical entities, we mention [11], that uses decision trees or SVMs; [12], that uses Hidden Markov Models or CRFs; [13], that presents a machine learning system which makes use of both local and syntactic features of the texts and external resources (gazetteers, web-querying, etc.); and [14], that obtains the nouns of disease, medical condition, treatment and symptom types, by using MQL queries and the Medlineplus Health Topics ontology (www.nlm.nih.gov/medlineplus/xml.html).

Rule-based methods are the ones proposed by [15], who identifies, with a set of graphical patterns, cause-effect information from medical abstracts in the Medline database, and [16], that manages to extract clinical entities disorders, symptoms and body structures from unstructured text in health records, using a rule-based algorithm.

Hybrid approaches have been proposed by [17] for the extraction of gene symbols and names; by [18] for protein-name recognition and by [19], which combines terminology resources and statistical methods with sensible improvements in terms of Precision.

III. Methodology

Our methodology is based on the LG framework, set up by the French linguist Maurice Gross during the ‘60s and subsequently applied to Italian by [20].

The LG theoretical and practical framework is one of the most consistent methods for natural language formalization, automatic textual analysis and parsing. Its main goal is to describe all mechanisms of word combinations closely related to concrete lexical units and sentence creation, and to give an exhaustive description of lexical and syntactic structures of natural language.

LG theoretical approach is prevalently based on [21], which assumes that each human language is a self-organizing system, and that the syntactic and semantic properties of a given word may be calculated on the basis of the relationships that this word has with all other co-occurring words inside given sentence contexts. The study of simple or nuclear sentences is achieved analyzing the rules of co-occurrence and selection restriction, i.e., distributional and transformational rules based on predicate syntactic-semantic properties.

As described in Section IV-B, in this work, following LG methodology, we anchored the recognition of terminological ALUs (Atomic Linguistic Units) to the sentence structures that recursively occur in medical reports. This way, on the base of co-occurrence rules, which can be characterized by different levels of variability, we could correctly annotate and classify a great part of the medical entities contained in our corpus.

As it is commonly done in literature, in our work we divided the Medical Entity Recognition into two subtasks, every one of which takes advantages from different resources.

- Semi-automatically generated lexical resources, for the extraction of semantic information from raw texts (see Section IV-A);
- Syntactic rules, for the extraction of semantic and domain information. The assumption for this step is that domain terminology is strictly interlinked with syntactic combination and co-occurrence behaviors (see Section IV-B).

Table 1 shows entity types recognized in our experiment; we also provide a description for these one.

A. NLP Tool

For our TE task we use NooJ, a software developed by Max Silberztein [22]. This system allows to formalize natural language descriptions and to apply them to corpora. NooJ is used by a large community, which developed linguistic modules,

 ISBN: 978-1-61208-445-9

TABLE I
ENTITY RECOGNITION CLASSES

Entity Type	Details
Disease	disorders and medical conditions
Treatment	therapies following diagnosis
Drug	information about prescribed drugs
Test	analysis and exams
Symptom	subjective evidences of diseases or of patient's conditions
Medical Branch	specific medical subdomains

including Finite State Automata/Transducers and Electronic Dictionaries, for more than twenty languages. The Italian Linguistic Resources have been built by the Computational Linguistic group of University of Salerno, which started its study of language formalization from 1981 [20]. Our analysis is based on the Italian module for NooJ [23], which is enriched with IMED and with grammars for Text Extraction (TE).

IV. LINGUISTIC RESOURCES

A. Italian Medical Electronic Dictionary

In order to automatically create the Italian Medical Electronic Dictionary (IMED) of the disease ALUs occurring in the corpus, we exploited morphosemantic strategies, which uses the semantics of a special kind of morphemes to identify and describe disease nouns or adjectives. Such kind of morphems are called neoclassical formative elements [24]. They come into being from Latin and Greek words and are generally used to form both technical-scientific words and ordinary words in a very productive way. They can combine themselves with other formative elements or with independent words.
In this paper we will talk about them using the word "confixes", which has been predominantly employed in literature [25]–[29].

The Medical Morphosemantic Module (M^3) we implemented is composed of the following resources:

- An Electronic Dictionary of Italian Morphemes belonging to the Medical Domain called *M3.dic*.
- Seven Morphological Grammars, denominated *M3#.nom*
- A syntactic Grammar, named *M3.nog*

The Dictionary *M3.dic* contains morphemes of the Italian medical domain which have been extracted from the electronic version of the GRADIT [30]. The morphemes has been divided into three classes: prefixes, suffixes and confixes, on the base of the positions of the morphemes in the words. Table II shows the morphemes extracted from the GRADIT. Each morpheme is described by a tag that specifies the meaning of the morpheme (i.e., *-oma* corresponds to the descriptions *tumori*, "tumours") and a tag that gives its medical subcategory (assigned with the support of a domain expert by dividing the macro class of the medicine into 25 subcategories, i.e., CARDIO, "cardiology"; ENDOCRIN, "endocrinology"; PSIC, "psychiatry" GASTRO, "gastroenterology"; PNEUMO, "pneumology"; NEURO, "neurology"; etc). We made use of a class UNKNOWN that has been used as residual category, in order to collect the words particularly difficult to classify.

TABLE II
MORPHEMES EXTRACTED FROM THE GRADIT

Manner of Use	Category	Number
Medicine	Confixes	485
Medicine	Suffixes	5
Medicine	Prefixes	7
Anatomy	Confixes	104
Anatomy	Prefixes	3

The morphemes that were not contained in the GRADIT's medical category have been manually added to our list, i.e., morphemes that are used in the formation of adjectives. The electronic dictionary of medical morphemes is classified in the following way:

- Confixes (CPX): neoclassical formative elements that appear in the initial part of the word (i.e., *pupillo-*, *mammo-*, *cefalia-*);
- Prefixes (PFX): morphemes that appear in the first part of the word and are able to connote it with a specific meaning (i.e., *-ipo*, *-iper*);
- Suffixes (SFX): morphemes that appear in the final part of the word and are able to connote it with a specific meaning (i.e., *-oma*, *-ite*);
- Suffixes for the adjectives formation: derivational morphemes that make it possible to derive and distinguish in the medical domain the adjectives (i.e., *polmonare*, "pulmonary") from the nouns that have a morpho-phonological relation with them (i.e., *polmone*, "lung").

The IMED has been completed with the addition of an electronic dictionary composed of more than 700 Concrete nouns of body/organism parts ("+Npc", i.e., *braccio*, "arm"; "+Npcorg", i.e., *cervello*, "brain") and of more than 400 Concrete nouns of drugs and medicines ("+Nfarm", i.e., *morfina*, "morphine ") developed by the Maurice Gross Laboratory of the University of Salerno.

The dictionaries works in combination with seven Morphological Grammars built with *Nooj*, which are able to find occurrences and co-occurrences of medical morphemes or nouns in medical documents' words. The seven grammars include the following combination of morphemes:

1) *confixes-confixes* or *prefixes-confixes* or *prefixes-confixes-confixes*;
2) *confixes-suffixes* or *prefixes-confixes-suffixes*;
3) *confixes-confixes-suffixes* or *prefixes-confixes-confixes-suffixes*;

 ISBN: 978-1-61208-445-9

4) *nouns-confixes*;
5) *prefixes-nouns-confixes*;
6) *confixes-nouns-confixes*;
7) *nouns-suffixes*;

In order to complete the IMED dictionary with medical multiword expressions, a syntactic grammar, built with NooJ, has been created with the goal of finding the following combination of Nouns (N), Adjectives (A) and Prepositions (P): N, NN, AN, NA, NNA, NAA, NPN.

B. Syntactic Rules

In the corpus, we noticed the presence of recursive sentence structure, in which we recognized specific and terminological ALUs. In this way, we could identify open series compounds, that are ALUs in which one or more fixed elements co-occur with one or more variable ones.
On the basis of this evidence, we developed different Finite State Automata for the TE task and for annotating treatments, tests, symptoms.

As for semantics, we observed the presence of compounds in which the head did not occur in the first position; for instance, the open series to recognize treatments *terapia di* N , "therapy of N", places the heads at the end of the compounds, being *terapia* used to explicit the notion N0 is a part of N1.

In Figure 1(a) and Figure 1(b), we recognized Treatment and Test classes by selecting a series of nouns, as fixed part, and a variable part, as head, formed by a Noun Group and/or an Adjective. We also applied a node with the option 'unknown word' (UNK); this feature allowed us to retrieve words which have not been inserted in IMED, also in order to update our dictionary.

To extract the Symptom class we developed a Finite State Automaton (Figure 1(c)) in which semantic features can be identified using grammars that are built on specific verb classes (semantic predicate sets) - i.e., *presentare, riferire, esporre, etc...*, "to present, to report, to express, etc..."; in such cases, co-occurrence restrictions can be described in terms of lexical forms and syntactic structures.

We used the grammatical information with which dictionary entries are tagged and syntactic rules as a weighting preference for the co-occurrence selection. So, we developed matrix tables in which semantic role sets, established on the basis of those constrains (properties), are matched with grammatical and syntactic rules. Matrices list a certain number of verbal entries and a specific number of distributional and syntactic properties.

During the recognition process, labeled IMED entries and FSA are the inputs. After the phase of text processing, the result is as follow:

⟨cardiology⟩ *Il Paziente* ⟨symptom⟩ *affetto da ipertensione arteriosa e BPCO* ⟨/symptom⟩. *Nel 1998 é stato sottoposto ad* ⟨treatment⟩ *intervento di sostituzione valvolare aortica mediante protesi meccanica Carbomedics* ⟨/treatment⟩ *e in quell'occasione le coronarie erano risultate prive di lesioni significative. Dopo l'intervento il Paziente ha eseguito periodici* ⟨test⟩ *controlli cardiologici* ⟨/test⟩ *presso l'Ospedale di Montichiari per* ⟨symptom⟩ *fibrillazione atriale ad elevata risposta ventricolare e scompenso cardiaco* ⟨/symptom⟩ ⟨/cardiology⟩.

"⟨cardiology⟩ The patient is ⟨symptom⟩ suffering from hypertension and BPCO ⟨/symptom⟩. In 1998 he had ⟨treatment⟩ surgery for aortic valve replacement using Carbomedics mechanical prosthesis ⟨/treatment⟩ and coronary arteries did not have significant injuries. After surgery, the patient performed periodic ⟨test⟩ cardiology checks ⟨/test⟩ at the Hospital of Montichiari for ⟨symptom⟩ atrial fibrillation with high ventricular response and heart failure⟨/symptom⟩ ⟨/cardiology⟩ ".

V. Testing and Results

The annotation process is performed on Italian clinical texts. The corpus has been built from a collection of 989 real medical records, opportunely anonymized with regards to every kind of sensitive data they contained.

Our corpus provides information about the Family History, the Physiological Anamnesis, the Past Illnesses, the Anamnesis, the Medical Diary and the Diagnosis Review for every patient. For our analysis we kept out the Family History and the Physiological Anamnesis sections, since they did not contain concepts, assertions or relation information.

The corpus, pre-processed with Nooj exploiting the traditional NLP pipeline, includes 470591 text units, 41409 different tokens and 1529774 word forms.

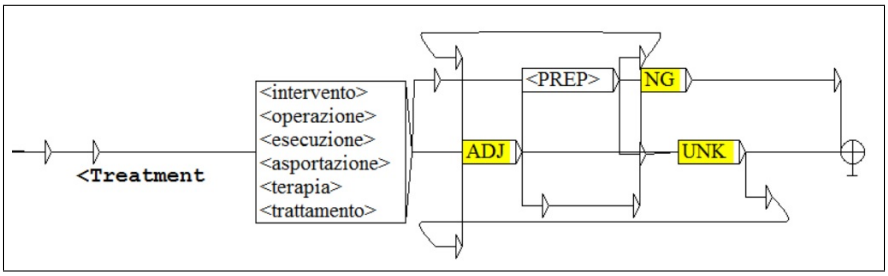
An evaluation of the results produced by our MER tool is given in Table III. We gave a measure of the validity of our method by calculating the Precision, the Recall and the F-score in the extraction of every entity class. In this phase we merged together the classes Drug/Treatment and Disease/Medical Branch because the syntactic grammars used to locate them are the same and our tool presents their results in the same list of concordances. Anyway, the tags used to annotate them are always different, so a distinction between these categories is performed any time.

As we can notice, the values present a variability with reference to the different categories, but we consider the average results very satisfying; nevertheless we are already planning to enrich our research outcomes with many other improvements.

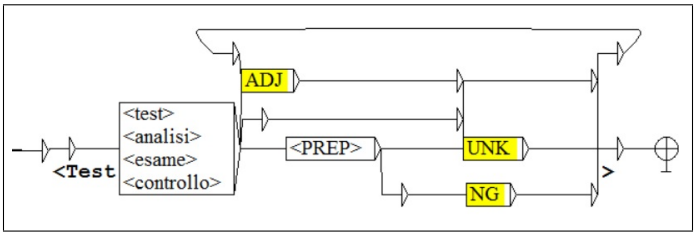
TABLE III
Evaluation

Entity Name	Precision	Recall	F-score
Symptom	0,75	0,52	0,61
Drug and Treatment	0,83	0,51	0,63
Test	0,96	0,51	0,67
Disease and Medical Branch	0,69	0,76	0,72
Average	0,80	0,58	0,66

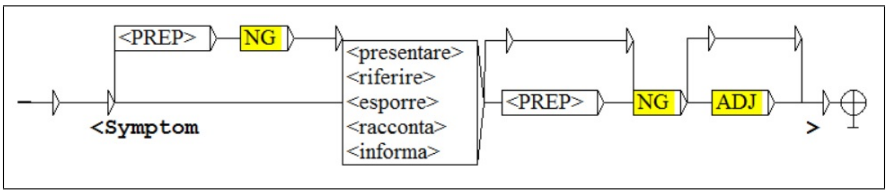
 ISBN: 978-1-61208-445-9



(a) Sample of FSA for Entity Recognition of Treatment



(b) Sample of FSA for Entity Recognition of Test



(c) Sample of FSA for Entity Recognition of Symptom

Fig. 1. Samples of FSA

All the annotations produced by the application (almost 5000 with the morpho-semantic method; more than 4000 with the syntactic strategies and about 2500 applying the preexistent dictionaries of the Italian module of Nooj) of our method and resources can be reused to enrich lexical databases or ontologies referred to the medical domain. Obviously, the size and the quality of the enrichment is strictly dependent on the largeness and on the content of the corpus on which the Nooj resources are applied. Therefore, in order to obtain widespread medical databases, it is preferable to use corpora able to cover the larger group of medical branches.

VI. Conclusion and Future Work

In this work, we presented our methodology for the extraction of entity classes from medical records, conducting different levels of linguistic analysis. Our framework is based on a robust definition language, which is used for creating and extending grammars and lexicons.

In our experiment, we considered the issue of entity boundaries carefully, but result analysis shows a request of improvement in ALUs recognizing. This comes from the specific use of medication abbreviations and word separators and from a nonstandard method to compile free-text notes. Challenging areas are the presence of ambiguous phenomena, e.g., fractions or numbers without unit or time references, and the use of brand name of drugs or a class of products, i.e., eye drops.

The combination of computational morphology and semantic distribution proposed here indicates very promising perspective: processing different corpora could instruct our system to recognize more recursive phenomena and more stop words in order to overcome partial matching issues. Future works aim at integrating our tools with the alignment of ontology constraints to syntactic relations in order to improve extraction of clinical concepts from notes, increasing the interoperability and the utility of clinical information.

References

[1] A. B. Abacha and P. Zweigenbaum, "Medical entity recognition: A comparison of semantic and statistical methods," in *Proceedings of BioNLP 2011 Workshop*. Association for Computational Linguistics, 2011, pp. 56–64.

[2] A. W. Pratt and M. Pacak, "Identification and transformation of terminal morphemes in medical english." *Methods of information in medicine*, vol. 8, no. 2, pp. 84–90, 1969.

[3] S. Wolff, "The use of morphosemantic regularities in the medical vocabulary for automatic lexical coding," *Methods of Information in Medicine*, vol. 23, no. 4, pp. 195–203, 1984.

[4] M. G. Pacak, L. Norton, and G. S. Dunham, "Morphosemantic analysis of-itis forms in medical language." *Methods of Information in Medicine*, vol. 19, no. 2, pp. 99–105, 1980.

[5] L. Norton and M. G. Pacak, "Morphosemantic analysis of compound word forms denoting surgical procedures." *Methods of Information in Medicine*, vol. 22, no. 1, pp. 29–36, 1983.

[6] P. Dujols, P. Aubas, C. Baylon, and F. Grémy, "Morpho-semantic analysis and translation of medical compound terms." *Methods of Information in Medicine*, vol. 30, no. 1, p. 30, 1991.

[7] A. R. Aronson, "Effective mapping of biomedical text to the umls metathesaurus: the metamap program." in *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2001, p. 17.

[8] U. Hahn, M. Romacker, and S. Schulz, "Medsyndikatea natural language system for the extraction of medical information from findings reports,"

 ISBN: 978-1-61208-445-9

International journal of medical informatics, vol. 67, no. 1, pp. 63–74, 2002.

[9] G. Schadow and C. J. McDonald, "Extracting structured information from free text pathology reports," in *AMIA Annual Symposium Proceedings*, vol. 2003. American Medical Informatics Association, 2003, p. 584.

[10] S. M. Meystre and P. J. Haug, "Comparing natural language processing tools to extract medical problems from narrative text," in *AMIA annual symposium proceedings*, vol. 2005. American Medical Informatics Association, 2005, p. 525.

[11] H. Isozaki and H. Kazawa, "Efficient support vector classifiers for named entity recognition," in *Proceedings of the 19th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, 2002, pp. 1–7.

[12] Y. He and M. Kayaalp, "Biological entity recognition with conditional random fields," in *AMIA Annual Symposium Proceedings*, vol. 2008. American Medical Informatics Association, 2008, p. 293.

[13] J. Finkel, S. Dingare, H. Nguyen, M. Nissim, C. Manning, and G. Sinclair, "Exploiting context for biomedical entity recognition: from syntax to the web," in *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*. Association for Computational Linguistics, 2004, pp. 88–91.

[14] M. de la Villa, F. Aparicio, M. J. Maña, and M. de Buenaga, "A learning support tool with clinical cases based on concept maps and medical entity recognition," in *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*. ACM, 2012, pp. 61–70.

[15] C. S. Khoo, S. Chan, and Y. Niu, "Extracting causal knowledge from a medical database using graphical patterns," in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2000, pp. 336–343.

[16] M. Skeppstedt, M. Kvist, and H. Dalianis, "Rule-based entity recognition and coverage of snomed ct in swedish clinical text." in *LREC*, 2012, pp. 1250–1257.

[17] D. Proux, F. Rechenmann, L. Julliard, V. Pillet, B. Jacq *et al.*, "Detecting gene symbols and names in biological texts: a first step toward pertinent information extraction," *Genome informatics series*, pp. 72–80, 1998.

[18] T. Liang and P.-K. Shih, "Empirical textual mining to protein entities recognition from pubmed corpus," in *Natural Language Processing and Information Systems*. Springer, 2005, pp. 56–66.

[19] A. Roberts, R. J. Gaizauskas, M. Hepple, and Y. Guo, "Combining terminology resources and statistical methods for entity recognition: an evaluation." in *LREC*, 2008.

[20] A. Elia, M. Martinelli, and E. D'Agostino, *Lessico e Strutture sintattiche. Introduzione alla sintassi del verbo italiano.* Napoli: Liguori, 1981.

[21] Z. S. Harris, "Notes du cours de syntaxe, traduction française par maurice gross," *Paris: Le Seuil*, 1976.

[22] M. Silberztein, "Nooj manual," *Available for download at: www. nooj4nlp.net*, 2003.

[23] S. Vietri, "The italian module for nooj." in *In Proceedings of the First Italian Conference on Computational Linguistics, CLiC-it 2014*. Pisa University Press, 2014.

[24] A. M. Thornton, *Morfologia*, R. Carocci Editore, Ed., 2005.

[25] A. Martinet, *Syntaxe generale*, A. Colin, Ed., 1985.

[26] A. Kirkness, "Aero-lexicography: Observations on the treatment of combinemes and neoclassical combinations in historical and scholarly european dictionaries," *Willy Martin ua (Hrsg.): Euralex*, pp. 530–535, 1994.

[27] S. C. Sgroi, "Per una ridenizione di "confisso": composti confissati, derivati confissati, parasintetici confissati vs etimi ibridi e incongrui," *Quaderni di semantica*, vol. 24, pp. 81–153, 2003.

[28] P. D'Achille, *L'italiano contemporaneo*, B. Il Mulino, Ed., 2003.

[29] T. De Mauro, *Nuove Parole Italiane dell'uso*, ser. GRADIT, T. UTET, Ed., 2003, vol. 7.

[30] ——, *Grande Dizionario Italiano dell'Uso*, T. UTET, Ed., 1999, vol. 8.

 ISBN: 978-1-61208-445-9