Cosema: Content-based Semantic Annotator

Angela Fogarolli
University of Trento
Via Sommarive 14, 38100 Trento, Italy
afogarol@disi.unitn.it

***Abstract*—In this paper, we present a library for creating automatic annotations for entities and concepts inside any textual content. The tool is based on DBpedia. In particular, the annotations are generated using the DBpedia link structure as a source of knowledge for Word Sense Disambiguation. DBpedia is used as a reference to obtain information on lexicographic relationships. By using such information in combination with statistical information extraction techniques, it is possible to deduce concepts related to the terms extracted from a corpus. Moreover, by combining statistical information extraction with named entity recognition and the use of the OKKAM ENS infrastructure, it is also possible to obtain unique annotations for entities in the content. The advantage of this approach, in addition of improving information retrieval and categorization capabilities, consists in the fact that the generate concept and entity annotations can be referred to with unique identifiers around the Web. For this reason different description for the same entity or concept can be semantically aggregated from the Web.**

I. Introduction

A common practice to avoid information overloading is to enable efficient access to a resource by associating to documents a set of metadata which describe their content. These metadata usually provide additional information about the content of the resources they are describing, such as author, main topic, language, etc. Descriptions should have a high level of semantics in order to be used for answering human needs of classification and retrieval. Various standardized metadata descriptors, which fulfill these requirements to different extents, are available today.

Metadata can be manually generated this is costly, time consuming and can be error-prone. Also the agreement between annotators can notably differ and usually requires domain expertise and controlled vocabularies. Since the amount of documents people are dealing with is constantly increasing, manual annotation faces increasing challenges in terms of sustainability. However, knowing what a document is about is of fundamental importance for effective knowledge management. Automatic or semi-automatic techniques can be employed instead as an alternative to human annotation. The limitation of automatic annotation is usually low recall when annotations are missing, low precision when the annotations are inaccurate, or the extraction of relationships [1] among them. Additionally, annotations alone do not establish the semantics of the vocabulary used in the annotations.

A solution to this problem can be inspired by the Semantic Web. The Semantic Web as envisioned in [2] allows semantic interoperability between machines and users. It provides a stack of languages for supporting the representation of knowledge, in the form of ontologies and metadata. Semantic Web technologies aim to annotate documents based on domain Ontologies. In this way the semantics of the produced annotations are well defined. An ontology [3] is a conceptualization of a domain with a controlled vocabulary and grammar for describing objects and the relations between them in a formal way. Ontologies are populated with individuals, often referred to as (named) entities. Typical entities are specific (individual) people, organizations, events, artifacts ("Mona Lisa"), places, products, etc.

The vision of the Semantic Web involves re-use mainly of the schematic parts of ontologies, i.e. concepts and their definition. Uniform Resource Identifiers (URIs) are used for referring to any resource, relations between resources can be stated in RDF [4] statements, and the vocabulary used for describing these relations is specified using RDF Schema [5] or OWL citeowl ontologies. The benefit of using this kind of formalization is that information can rather easily be aggregated (by detecting identical URIs in datasets), and that they enable certain kinds of reasoning (e.g., about class hierarchies) that can produce query results beyond what is currently possible using relational databases or information retrieval systems.

The environment described in this paper aims to provide a way for automatically generating semantic annotations for a given text compliant with best practices of the Semantic Web, being easy interlinking and distributedness. We thus enable extraction and sharing of the knowledge implicit in content, on the Web. For guaranteeing domain independence, the tool is based on the DBpedia [7] knowledge. DBpedia can be considered a light weight ontology which spans different domains. In this way, any type of content can be annotated by Cosema library.

The rest of the article is organized as follows. In the next section we will give a brief overview of the related work.

 ISBN: 978-1-61208-104-5

In Section III we describe a novel approach for automatic generation of concepts and entities semantic annotations. In Section IV we will evaluate the quality of the automatic generated annotations. The conclusions summarize our contribution to the Semantic Annotation field.

II. Related Work

Semantic Web technologies aim to automatically or semi-automatically annotate documents based on domain ontologies. In this way the semantics of the produced annotations are well defined. Semantic annotations define in a formal way concepts and relationships between them. There are different approaches from manual to automatic generation of annotations. In [1] a review of the state of the art in the field is presented.

The use of annotations has been investigated in various fields. Examples include: television and radio news [8], bioinformatics [9], heritage [10] and content classification of web pages [11] Human annotation is costly, time consuming and prone to errors. Also the agreement between annotator can notably differ and requires domain expertise.

Cucerzan in [12] described an interesting approach for associating Name Entities in a corpus with Wikipedia definitions. The goal of this approach is similar to ours, the main difference is that we do not limit the corpus analysis to Name Entities and we considered also multilanguage material. They explored various strategies to decrease the numbers of attributes to consider. They reduce the context information by extracting entities with a certain number of mentions in the article or using some TF-IDF threshold. For learning about topic dependencies for annotation, in this paper we consider only strong links [13] among articles.

Synarcher [14] is another work based on Wikipedia knowledge, which searches for synonyms and related terms in the Wikipedia category structure and analyzing hyperlinks between pages. The algorithm could be used to extend queries in a search engine, or as an assistant for forming a dictionary of synonyms. Another work which explores categories in Wikipedia is the one of Chernov et al. [15]. The authors suggest that semantic information can be extracted from Wikipedia by analyzing the category structure and they propose a way to calculate a connectivity ratio which correlates with the strength of the semantic connection among them. Wikipedia categories are also used for document classification by Schonhofen [16] and by Thom et al. [17] for improving entity ranking effectiveness. Watanabe et al. present another work on Name Entity categorization [18] based on category information extracted from the linked HTML text in the articles. Syed et al. in [19] describe an approach for identifying topics and concepts associated with a set of documents. The approach is based on the Wikipedia category graph for predicting generalized concepts and uses article links to help predict concept when an article is not associated with a specific category.

Adafre and de Rijke [20] firstly analyzed the link structure in Wikipedia, in 2005. They tackle the problem of missing links between articles. For doing this they cluster similar pages based on similar link structure and then they examined these cluster to find missing links between them. Voss [21] described the Wikipedia link structure as a power law function in which there is an exponential growth of links. Whenever a non-existing article is linked is more likely someone will create it. Kamps and Koolen [22] examined Wikipedia link structure and stated that link structure is an indicator of relevance especially if considering links between pages retrieved in response to a search request. In other words links can help defining a context and can improve performance in information retrieval. Hyperlinks structure in Wikipedia is also used for calculating related pages to an article. Ollivier and Senellart [23] process these relationships using Green Measures which is a function introduced in electrostatic theory for computing the potential created by a charge distribution. Green measures are applied as a finite Markov chain to a graph modeled by hyperlinks among Wikipedia articles.

Mihalcea in [24] and [25] discuss the use of Wikipedia for Word Sense Disambiguation (WSD). In [24], the author reports about the use of Wikipedia content for avoiding the bottleneck in WSD of not having enough examples of a term usage. In her approach, she selects all paragraphs in Wikipedia which contain a contextualized reference to an ambiguous term in the link label and then maps the different Wikipedia annotations to word senses instead of relying on the Wikipedia disambiguation pages. This is due to the face that sometimes not all meanings are elicited in the disambiguation page. Finally, the labels which describe the possible senses for a word are manually mapped to WordNet senses. In this way the number of example for each word can increase improving the performance of a classifier. In her second work [25], Mihalcea describes an use case of her WSD algorithm to an application which associate terms in an input text to Wikipedia definitions. The keyword extraction from the text is done using a controlled vocabulary. WSD is done in three different ways. Using a Knowledge-Based calculating the overlap of the Wikipedia definition with the paragraph where the text occurs (similar to Lesk algorithm). A second approach that has also been tested in [25] is a data-driven method which use a machine learning classifier, giving as a training all the occurrences where the word is found in the link plus all the possible Wikipedia definition articles, which represents the possible meanings. Additionally they experimented also a combination of the first two approaches.

The OKKAM research project [26] is an attempt to solve the identity problem on the Semantic Web. OKKAM aims to enable and bootstrap the Web of Entities, a global decentralized information space in which every entity is identified by a global identifier, and in which global identifiers are

 ISBN: 978-1-61208-104-5

consistently used for specifying relations between entities, across system boundaries. As the World Wide Web (WWW) was the result of integrating local Webs of documents into a global (universal) space of resources addressable through global identifiers (the well-known URLs), so the Web of Entities will be the result of integrating local webs of entities (i.e. any local space of information about a collection of entities, like a directory, a catalogue, an information system, a knowledge base, a database, a data intensive web site, and so on) into a global information space where every entity is identified through a global (universal) identifier. However, with respect to the WWW, the domain of entities is extended beyond the realm of digital resources, and links between entities are extended beyond hyperlinks to include virtually any type of relation. As a result, the vast amount of information, which today is not integrated, could be aligned and become part of a global information space that has entities as pivot objects, instead of documents.

III. Generation of Semantic Annotations

In this section, we describe how we extract semantic annotations from textual content. Those annotations express the most important concepts and entities in text content. There are two interfaces for accessing the functionality of the Cosema library, a web interface and a Web service interface. As input the system receives a text passage and it returns semantic annotations for contained entities and concepts. The annotations are represented by using Semantic Web URI. In this way by resolving the URI is possible to gather a detailed description of the meaning of the annotation.

A. Disambiguation Process

This section describes the WSD process we used for discriminating the correct meaning of a term based on the context where the term was found. The approach is based on the DBpedia link structure which can be assimilated to the Wikipedia link structure. The link structure in Wikipedia draws a huge network between pages which facilitates the navigation and the understanding of concepts.

The type of link we are interested in for WSD are what we called "strong links". We define a strong link as a bidirectional connection between two pages. A page P_o has a strong link with page P_d if in P_o exists a link to P_d and in P_d there is a link back to P_o.

$$P_o \longleftrightarrow P_d \tag{1}$$

A link in Wikipedia is considered to be strong if the page it points to has a link back to the starting page.

The WSD approach included in the Cosema library uses DBpedia as a source of Knowledge.

The first step for calculating semantic annotations is related to information extraction (IE). Cosema uses two IE methods: a statistical one based on TF-IDF measure and a name entity recognizer (NER) (two commercial and one opensource NER has been evaluated).

A term vector containing the most important terms on a document is extracted based on the TF-IDF measure and combined with the results of the NER. The disambiguation for an ambiguous term or entity is calculating by matching the term with a DBpedia or OKKAM identifier. The process takes into account the document domain which is defined by the terms in the same document.

In Wikipedia and so in DBpedia, different word senses are represented through a so-called disambiguation page. Each article in Wikipedia is identified by its title. The title consists of a sequence of words separated by underscores. When the same concept exists in different domains that name is concatenated with a string composed by a parenthetical expression which denotes the domain where the word has a specific sense. If a query ambiguously identifies more senses, a disambiguation page is called.

The algorithm for creating a semantic annotation uses two different resources for annotating entities or concepts. For entity annotations Cosema relies on the knowledge of OKKAM ENS which already includes all the DBPedia entities. While it uses DBpedia directly for disambiguating concepts and in the case of entities that are not present in the OKKAM ENS. It follows a separate description for the two methodologies. The results of the IE phase are two lists, one with the extracted entities derived from the NER and the second is a term vector coming from the statistical IE. Each of the extracted entities is looked up in OKKAM. In case of entity type "'Person"' there is the need of minimum two words (since the ambiguity of using just a last name or a name as discrimination will be too high) to be passed to OKKAM otherwise the entity will be resolved with the procedure used for the concepts which deals with ambiguity by taking into account the context where the word was located. If the entity is present in OKKAM then OKKAM identifiers and the entity alternative identifiers will be returned (i.e., the DBpedia identifier can be an alternative identifier).

For generating annotations for concepts or for entities in case of failure of the OKKAM lookup, the procedure will analyze every term present in the term vector created out of the text given as input. The term vector is defined as:

$$T_{i=1..N} = \{w_{ij}\}_{j=\{1..25\}}$$

where i identify a specific document, and j a term in the term vector. For each candidate definition p_{ijk}, where k is the k-th possible definition, we consider only its strong links (the concept and its links are searched in DBpedia through the SPARQL endpoint).

$$S_{zijk} = S_z\left(p_{ijk}\right)_{z=\{1..M\},k=\{1..Q\}} \tag{2}$$

 ISBN: 978-1-61208-104-5

where Q is the number of senses for pij and M is the number of strong links for the k-th sense.

Therefore S_{zijk} is the z-th strong link for the k-th sense of the j-th term of the i-th document. Hence, a strong link represents a bidirectional relation between two DBpedia pages. All strong links S_{zijk} for every term w_{ij} are taken into account for computing the disambiguation process and to be used in the query suggestion and summarization task. The best definition among the candidates is the one having the majority of words w_{ij} in the presentation material T_i in common with the target article name anchored from a strong link.

We can write this concept as function $f(i,j,k)$ where i identifies a specific document, j a term in the term vector and k a candidate definition for the term j. The function $f(i,j,k)$ will help us selecting the page p_{ij} which has the maximum number of elements in the intersection between the term vector for a presentation T_i and the target article name of the selected hard links for the candidate DBpedia definition pages, p_{ijk}. The function $f(i,j,k)$ is defined as:

$$f(i,j,k) = |T_i \cap \left\{S_{zijk,z=\{1..M\}}\right\}|$$

where z is the i-th strong link for the candidate page p_{ijk}.

The symbol | indicates the cardinality of the expression. The correct definition page p_{ij} will be identified among the p_{ijk} pages by selecting the k such that $|f(i,j,k)|$ has the largest value.

$$p_{ij} = p_{ijk}$$

which indexes are found by

$$max_k|f(i,j,k)|$$

For example if we analyze an e-Learning document (document 1) about Java Programming whose (simplified) vector is defined by:

$$T_1 = \{set, map, array, list, java, computer, collection\}$$

We consider the case of finding the right DBpedia definition for the term collection which is part of document 1. In the disambiguation page are listed the definitions for "Collection(computing)" and "Collection(museum)". For each of these pages we analyze the strong links counting the number of elements in common with the words in the term vector of the e-Lecture document in exam:

$$S_{171}Collection\,(computing) =$$

$$\{oriented, class, map, tree, set, array, list\};$$

$$S_{172}Collection\,(museum) = \{curation, curator\};$$

The group CE, contains the elements in common between the term vector and the strong link for each candidate page:

$$CE = T_1 \cap S_{171}Collection\,(computing)$$

$$CE = T_1 \cap S_{172}Collection\,(museum)$$

Since words in a term vector are stemmed, the strong links must be stemmed as well before comparing them with the keywords in the term vector. We choose the DBpedia definition page among the candidate pages to be the one which has the maximum number of elements in CE. In the example, we have $|f171| = 3$ (case of $Collection\,(computing)$) while $|f172| = 0$ (case of $Collection\,(museum)$). Therefore the disambiguated meaning of term P_17 (i.e. collection) is correctly found to be $Collection\,(computing)$. The expected result of the process is a complete disambiguated term vector Td_i composed of disambiguated words wd_{ij}.

$$Td_{i=1..N} = \{wd_{ij}\}_{j=\{1..25\}}$$

For improving the accuracy of the results we do not insert in the candidate pages only the ones with an exact match to a word in the analyzed text but all the pages which begin with that word. In this way, we are sure to include in the candidates definitions all the declinations and possible domain. More specifically, there are cases where ambiguous words are not linked to the articles mentioned by a disambiguation page, but instead they are mentioned in the related concepts section or a disambiguation page does not exist.

The disambiguation process access DBpedia online through the Web service interface, while other approaches presented also in the related work section use Wikipedia directly. The major drawback of using Wikipedia instead of DBpedia is that Wikipedia is not structured and there is not API for automatically accessing its content. For this reason for accessing Wikipedia content there is the need of using Natural Language processing techniques directly on the online version with very poor processing performance or installing and using the Wikipedia dumps. The dumps supply a complete database with the Wikipedia content; the drawbacks of this solution are in maintaining and keeping the local Wikipedia copy up to date for then calculating semantics on it. Using DBpedia instead is a very fast, lightweight and always up to date alternative for collecting information about Wikipedia content.

IV. EVALUATION

Assessing the quality of an application is very difficult and depends highly on human expertise. We evaluated the quality of the described approach in WSD. The idea behind our approach is based on a link analysis of DBpedia definition pages. In, our previous work [13], we supplied evidence that since links among Wikipedia pages connect articles that are semantically related and likely on the same context, the link structure also provides a way for identifying relationships among topics. Furthermore, we want to investigate how strong these relationships are, based on the type of link that exists between the documents. In particular, we suppose that if there is a symmetrical link relationship among two

Evaluation type	Precision
Wikipedia Based Corpus Tagging	73.4%
DBpedia Based Corpus Tagging	76.1%

Table I
WSD COMPARISON

pages, the strength of the link denotes the most important connections for describing a subject. In this section we want to evaluate how good is the approach in creating semantic annotations and for doing this we have to focus on evaluating the WSD task at the base of our approach.

The objective of the evaluation is to assess the quality of the system in recognizing different word sense. In this section we want to explore if the approach can produce good annotations for describing the content of generic text.

For this purpose we have collected sixteen text passages in English, from various sources: newspapers, encyclopedia, text books and random Web pages. We asked two annotators to manually annotate the passages using three titles of Wikipedia articles for restricting the vocabulary possibilities. Next, we compared the annotations automatically generated with the manual ones using two testers. The testers after careful reading of each text passage had to judge the correctness of the automatic annotations taking into account the difference in semantics between them and the manual ones by expressing a quality value from zero to one. A zero quality value means that the automatic annotations does not describe the text passage and they are completely unrelated with the manual annotations and one means that the automatic annotation perfectly describe the text content and can be the same as the manual annotations. We let the testers free to autonomously decide the other values in the interval by judging the semantic error of the automatic annotations.

In order to calculate the result of the experiment we consider the manual annotation to be exact and we compare the automatic ones against them. Based on the two tester judgment the precision on our test collection of the automatic generated annotations is 75%.

This result is consistent, as shown in table I, with a previous evaluation we made using Wikipedia dumps for calculating WSD. This underlines that for concept disambiguation the information included in the DBpedia representation is sufficient for gathering the same accuracy results as with Wikipedia. This result support our assumption that DBpedia knowledge can be used as Wikipedia for creating semantic annotations with the advantage of a faster processing time and easier accessibility.

During the word sense evaluation we were also considering the correctness of the meaning of the annotation by pointing it to a Wikipedia article, for this reason the precision value is lower since some errors can occur in the sense disambiguation while the annotation word is still correct. For example for a text about the "9/11" an annotation Attack could be consider correct but the meaning of the connected article given by our algorithm was "Attack (30 Seconds to Mars song)" which is wrong. In the WSD evaluation the objective was to have both correct annotation and sense, on the automatic tagging evaluation the focus was only on the correct annotation. Moreover in the test we only compared the results for concept annotations and not entities annotations since our previous Wikipedia based approach was not able to distinguish between entities and concepts. Even though we do not present this type of comparison, the persons who took the test admit that the entity annotations where able to give either an higher level categorization of the text in case of events or a more specific definition in case of person entities.

V. CONCLUSIONS AND FUTURE WORK

We have presented a library tool for automatic generation of concepts and entities annotations about content. The library can be accessed through a Web interface or Web Services. In the paper the WSD approach behind the tool is described and evaluated. DBpedia has been used as a knowledge resource for WSD. The cross-links between DBpedia entries allows us to discover important relations between concepts. We applied the presented work in a digital library environment for automatically annotating and enabling searches and navigation through an unstructured multimedia and in another tool for creating multimedia presentation. The good results of the evaluation suggest that our approach might be applied in different scenarios such as text categorization and document classification, where it is crucial to automatically extract semantic information from content. This underlines the genericity and usefulness of the work presented in this paper. In the future we plan to add the functionality of generating an RDFa description of the annotations to be included where the content will be published. In this way semantic search engine will easily discover the annotated content.

REFERENCES

[1] V. S. Uren, P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta, and F. Ciravegna, "Semantic annotation for knowledge management: Requirements and a survey of the state of the art," *J. Web Sem.*, vol. 4, no. 1, pp. 14–28, 2006.

[2] T. Berners-Lee, J. A. Hendler, and O. Lassila, "The Semantic Web." *Scientific American*, vol. May, 2001.

[3] D. Fensel, *Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2003.

[4] P. Hayes, *RDF Semantics*, February 2004, http://www.w3.org/TR/rdf-mt/. [Online]. Available: http://www.w3.org/TR/rdf-mt/

[5] D. Brickley and R. G. (Eds.), *RDF Vocabulary Description Language 1.0: RDF Schema*, W3C, February 2004.

[6] P. Patel-Schneider, P. Hayes, and I. Horrocks, "Web Ontology Language (OWL) Abstract Syntax and Semantics," W3C, Tech. Rep., February 2003, http://www.w3.org/TR/owl-semantics/. [Online]. Available: http://www.w3.org/TR/owl-semantics/

[7] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, "Dbpedia : a crystallization point for the web of data," 2009. [Online]. Available: http://jens-lehmann.org/files/2009_dbpedia_jws.pdf

[8] M. Dowman, V. Tablan, H. Cunningham, and B. Popov, "Web-assisted annotation, semantic indexing and search of television and radio news," in *WWW '05: Proceedings of the 14th international conference on World Wide Web*. New York, NY, USA: ACM, 2005, pp. 225–234.

[9] E. Camon, M. Magrane, D. Barrell, D. Binns, W. Fleischmann, P. Kersey, N. Mulder, T. Oinn, J. Maslen, A. Cox, and R. Apweiler, "The gene ontology annotation (goa) project: implementation of go in swiss-prot, trembl, and interpro." *Genome Res*, vol. 13, no. 4, pp. 662–672, April 2003. [Online]. Available: http://dx.doi.org/10.1101/gr.461403

[10] V. M. Hennie Brugman and L. Hollink, "A common multimedia annotation framework for cross linking cultural heritage digital collections," in *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, E. L. R. A. (ELRA), Ed., Marrakech, Morocco, may 2008.

[11] S. Handschuh and S. Staab, "Authoring and annotation of web pages in cream," in *WWW '02: Proceedings of the 11th international conference on World Wide Web*. New York, NY, USA: ACM, 2002, pp. 462–473.

[12] S. Cucerzan, "Large-scale named entity disambiguation based on wikipedia data," in *EMNLP 2007: Empirical Methods in Natural Language Processing,Prague, Czech Republic*, June 28-30, 2007, pp. 708–716. [Online]. Available: http://acl.ldc.upenn.edu/D/D07/D07-1074.pdf

[13] A. Fogarolli, "Word sense disambiguation based on wikipedia link structure," in *IEEE ICSC 2009*, 2009.

[14] A. Krizhanovsky, "Synonym search in wikipedia: Synarcher," *arxiv.org*, search for synomyms in Wikipedia using hyperlinks and categories. [Online]. Available: http://arxiv.org/abs/cs/0606097v1

[15] S. Chernov, T. Iofciu, W. Nejdl, and X. Zhou, "Extracting semantic relationships between wikipedia categories," in *1st Workshop on Semantic Wikis:*, June December 2006.

[16] P. Schonhofen, "Identifying document topics using the wikipedia category network," in *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*. Washington, DC, USA: IEEE Computer Society, 2006, pp. 456–462.

[17] A.-M. Vercoustre, J. A. Thom, and J. Pehcevski, "Entity ranking in wikipedia," in *SAC '08: Proceedings of the 2008 ACM symposium on Applied computing*. New York, NY, USA: ACM, 2008, pp. 1101–1106.

[18] Y. Watanabe, M. Asahara, and Y. Matsumoto, "A graph-based approach to named entity categorization in Wikipedia using conditional random fields," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 649–657. [Online]. Available: http://www.aclweb.org/anthology/D/D07/D07-1068

[19] Z. Syed, T. Finin, and A. Joshi, "Wikipedia as an ontology for describing documents," in *Proceedings of the Second International Conference on Weblogs and Social Media*. AAAI Press, March 2008.

[20] S. F. Adafre and M. de Rijke, "Discovering missing links in wikipedia," in *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*. New York, NY, USA: ACM, 2005, pp. 90–97.

[21] J. Voss, "Measuring wikipedia," in *Proceedings International Conference of the International Society for Scientometrics and Informetrics: 10 th*, 2005. [Online]. Available: http://eprints.rclis.org/archive/00003610/

[22] J. Kamps and M. Koolen, "The importance of link evidence in wikipedia." in *ECIR*, ser. Lecture Notes in Computer Science, C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, and R. W. White, Eds., vol. 4956. Springer, 2008, pp. 270–282. [Online]. Available: http://dblp.uni-trier.de/db/conf/ecir/ecir2008.html#KampsK08

[23] Y. Ollivier and P. Senellart, "Finding related pages using Green measures: An illustration with Wikipedia," in *Proc. AAAI*, Vancouver, Canada, Jul. 2007, pp. 1427–1433.

[24] R. Mihalcea, "Using wikipedia for automatic word sense disambiguation," in *Proceedings of NAACL HLT 2007*, 2007, pp. 196–203. [Online]. Available: http://www.cs.unt.edu/~rada/papers/mihalcea.naacl07.pdf

[25] R. Mihalcea and A. Csomai, "Wikify!: linking documents to encyclopedic knowledge," in *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. New York, NY, USA: ACM, 2007, pp. 233–242. [Online]. Available: http://84.11.13.37/Volumes/CIKM_07/docs/p233.pdf

[26] P. Bouquet, H. Stoermer, C. Niederee, and A. Mana, "Entity name system: The back-bone of an open and scalable web of data," *International Conference on Semantic Computing*, vol. 0, pp. 554–561, 2008.

 ISBN: 978-1-61208-104-5