A Framework for Semantic Web of Patent Information

Yung Chang Chi[1]
Hei Chia Wang[2]
Department of Industrial and Information Management and
Institute of Information Management,
National Cheng Kung University,
Tainan City, Taiwan ROC
e-mail:charles.y.c.chi@gmail.com[1]
hcwang@mail.ncku.edu.tw[2]

Ying Maw Teng[3]
Department of International Business,
I-Shou University,
Kaohsiung City, Taiwan ROC
e-mail: morris@isu.edu.tw[3]

***Abstract* - This paper aims to propose a framework for an ontology-based semantic web of patent information by employing PATExpert, which is based on the ontological approach of constructing knowledge and technique from patent documents. This method will analyze patent infringement issues from judicial judgments in the USA and Europe. Having examined relative patent documents and the analysis of patent infringement by comparison, one can identify particular kinds of products and technologies that are interrelated with the analysis of the two different databases while enhancing the feasible construction of the semantic web for patent information.**

Keywords-patent; PATExpert; patent infringement; content analysis; Ontology; semantic web.

I. INTRODUCTION

Patents are important sources of knowledge for industrial research and product development because of their innovation and practicability. In recent years, patent analyses have increased in importance for high-technology management as the process of innovation has become more complex, the cycle of innovation shorter, and the market demand more volatile [15].

An emerging research topic, patent mining consists of patent retrieval, patent categorization, and patent clustering [1]. So far, little research has been done on the topic.

The European project PATExpert, (Advanced Patent Document Processing Techniques), coordinated by Barcelona Media (BM), has successfully accomplished the objectives after the pre-established 30 months (from February 2006 to July 2008). Thus, it has been ratified by the representative and the two external supervisors designated by the European Commission, in the final review of the project celebrated on 21st October at the BM headquarters [19].

In the frame of the Sixth Framework Program of Research and Technological Development (2002-2006), PATExpert has a global objective to change the present textual processing of patents to semantic processing (treating the patents as multimedia knowledge objects) [19].

WordNet was created in the Cognitive Science Laboratory of Princeton University under the direction of psychology professor George Armitage Miller starting in 1985 [24].

WordNet superficially resembles a thesaurus, in that it groups words together based on their meanings. However, there are some important distinctions. First, WordNet interlinks not just word forms—strings of letters—but specific senses of words. As a result, words that are found in close proximity to one another in the network are semantically disambiguated. Second, WordNet labels the semantic relations among words, whereas the groupings of words in a thesaurus do not follow any explicit pattern other than meaning similarity [24].

The applications employ the Semantic Web to make useable sense out of large, distributed information found throughout the World Wide Web. A definition for the Semantic Web begins with defining semantic. Semantic simply means meaning. Meaning enables a more effective use of the underlying data. Meaning is often absent from most information sources, requiring users or complex programming instructions to supply it. For example, web pages are filled with information associated tags. Most of the tags represent formatting instructions, such as <H1> to indicate a major heading. Semantically, we know that words surrounded by <H1> tags are more important to readers than other texts because of the meaning of H1. Some web pages add basic semantics for search engines by using the <META> tag; however, they are merely isolating keywords and lack linkages to provide a more meaningful context. These semantics are weak and limit searches to find out the exact matches. Similarly, databases will render more accurate matches if tables and columns of database are well named and well organized. Semantics give a keyword symbol useful meaning through the establishment of relationships [4].

Relational databases depend on a schema for structure. A knowledge base depends on ontological statements to establish structure. Relational databases are limited to one kind of relationship – the foreign key. The Semantic Web offers multidimensional relationships such as inheritance, part of, associated with, and many other types, including logical relationships and constraints. One important note is that the language used to form structure and the instances themselves may be the same language in a knowledge base but actually quite different in relational databases [4].

Retrieving patent documents can be done through the cluster-based approach [10]. Distributed information retrieval for patents can be done by generating ranking lists

 ISBN: 978-1-61208-486-2

for the query by CORI (The collection retrieval inference network) or KL (Kernighan–Lin) algorithms [13]. Categorizing patent documents can be done automatically by using the k-Nearest Neighbor classifiers and Bayesian classifiers [12][14], or a variety of machine learning algorithms [2], the k-Nearest Neighbor on the basis of patent's semantic structure [6], or the classifier built through back-propagation network [23]. Clustering algorithms can also be adopted to form a topic map for presenting patent analysis and summarization results [23], and to create a system interface for retrieving patent documents [5].

Content analysis is indigenous to communication research and is potentially one of the most important research techniques in social sciences. It seeks to analyze data within a specific context in view of the meaning someone – a group or a culture – attributes to them. Communications, messages, and symbols differ from observable events, things, properties, or people in that they inform about something other than themselves; they reveal some properties of their distant producers or carriers, and they have cognitive consequences for their senders, their receivers, and the institutions in which their exchange is embedded [7].

Content analysis is a research technique for making replicable and valid inferences from texts to the contexts of their use. As a technique, content analysis involves specialized procedures. It provides new insight, increases a researcher's understanding of particular phenomena, or informs practical actions. Content analysis is a scientific tool [8].

The judgements of patent infringement, unlike the patent documents, can be mined by using text mining techniques, since the judgments are legal documents. The judgments can be transformed into patterns by content analysis, and readers can easily access them the same way as reading newspapers to understand the key points and issues in dispute.

This study proposes to enhance a semantic web of patent information and patent infringement framework, and the rest of the paper is structured as follows. Section II presents the research background and states the objective. In Section III, we describe our proposed research method. The paper concludes with expected results and future work considerations.

II. RESEARCH BACKGROUND AND OBJECTIVE

So far, patent analysis technologies include patent bibliometric data analysis [3], patent citation analysis [16], patent statistical analysis [22], and patent classification. Patent mining consists of patent retrieval, patent categorization and patent clustering [22]. However, patent infringement constitutes the biggest threat in patent use.

This framework proposes that through patent document mining and judgements of patent infringement analysis, one can discover newly developed products and their similarity with the claims of other patents, and foresee where potential patent threats are and the likelihood of patent infringement. While constructing the patent Semantic Web, the ontology and rules engines must include it.

This framework of patent database is based on the databases of United States Patent and Trademark Office (USPTO) and the European Patent Office (EPO). The patent infringement judgements are based on the judicial judgments in United States and European Union.

The purposes of this study is to construct the Semantic Web of patent information, and reference regarding patent infringement, technology trends for new product designers and technology research engineers at the stage before and after developing a new product or technology. With the analyzed information, managerial team can make sound strategic decisions.

It is difficult for laymen or ordinary readers who are short of legal background to fully grasp the gist of judicial judgments rendered by judges. With the implementation of content analysis and the big data concept, ordinary people can use content analysis technology to analyze patent infringements with the help of a semantic web.

III. RESEARCH METHOD

Patent documents can be collected from United States Patent and Trademark Office (USPTO) patent database and the European Patent Office (EPO) patent database. The patent infringement content can be collected from the "West Law" database.

A. Patent documents analysis

Based on the collected patent documents and the subject-action-object (SAO) structures extracted by using Natural Language Processing (NLP), the study uses a content analysis approach to generate the concepts and relationships of related patent documents.

NLP is a text mining technique that can conduct syntactic analysis of natural language; NLP tools include the Stanford parser (Stanford 2013) [21], Minipar (Lin 2003) [17], and Knowledgist TM2.5 [11].

NLP tools will be used for building a set of SAO structures from the collected patents.

Multidimensional scaling (MDS) is a statistical technique used to visualize similarities in data [9][20]. Patent documents in different fields have different key issues that trigger different multidimensional scaling, so the paper will design a new algorithm to identify which particular patent field shall correspond to what extent of scaling.

The analysis patent documents for specified keywords and returns a list of the documents where the keywords were found. Then, the framework will use data and text mining technology to design a specified algorithm (still in progress) in order to analyze the legal documents and try to find out the most similar patents or patent group.

B. Patent infringement content analysis

The most obvious source of data appropriate for content analysis is the text to which meanings are conventionally attributed: verbal discourse, written documents, and visual representations. The text in the patent infringement judgements is important because that is where the meanings are. For this reason, it is essential for the content analysis technology to analyze the patent infringement text in order to

 ISBN: 978-1-61208-486-2

develop strategies and preventive measures in patent litigation.

The judgements of patent infringement, unlike the patent documents, can be mined using text mining techniques, since the judgements are legal documents. The judgements can be transformed into patterns by content analysis, and readers can easily access them the same way as reading newspapers to understand the key points and issues in dispute.

Content analyses commonly contain six steps that define the technique procedurally, as follows:

Design. Design is a conceptual phase during which analysts (i) define their context, what they wish to know and are unable to observe directly; (ii)explore the source of relevant data that either are or may become available; and (iii)adopt an analytical construct that formalizes the knowledge available about the data-context relationship thereby justifying the inferential step involved in going from one to the other.

Unitizing. Unitizing is the phase of defining and ultimately identifying units of analysis in the volume of available data. Sampling units make possible the drawing of a statistically representative sample from a population of potentially available data, such as issues of a newspaper, whole books, television episodes, fictional characters, essays, advertisements.

Sampling. While the process of drawing representative samples is not indigenous to content analysis, there is the need to (1) undo the statistical biases inherent in much of the symbolic material analyzed and (2) ensure that the often conditional hierarchy of chosen sampling units become representative of the organization of the symbolic phenomena under investigation.

Coding. Coding is the step of describing the recording units or classifying them in terms of the categories of the analytical constructs chosen. This step replicates an elementary notion of meaning and can be accomplished either by explicit instructions to trained human coders or by computer coding. The two evaluative criteria, reliability as measured by inter coder agreement and relevance or meaningfulness, are often at odds.

Drawing inferences. Drawing inferences is the most important phase in a content analysis. It applies the stable knowledge about how the variable accounts of coded data are related to the phenomena the researcher wants to know about.

Validation. Validation is the desideratum of any research effort. However, validation of content analysis results is limited by the intention of the technique to infer what cannot be observed directly and for which validation evidence is not readily available.

The patent infringement content analysis searches the patent infringement judgements for specific keywords and returns a list of the documents as above patent documents by introducing the content analysis technology into specified design algorithm (still in progress) in order to analyze the infringement cases/precedents. It also finds the nearest infringement judgements/precedents.

C. Constructing Semantic Web

Our proposed semantic web, which has a structure based on Figure 4, is a program that has the following parts:

(1) The first part searches patent documents for specified keywords and returns a list of the documents where the keywords were found. Then, the program will use data and text mining technology to design a specified algorithm in order to analyze the legal documents and try to find out the most similar patents or patent group concepts and relationships. The results from part (1) constructing the knowledge repository for the Reasoners are shown in Figure 1.

(2) Next, it searches the patent infringement judgments for specific keywords and returns a list of the documents as above patent documents by introducing the content analysis technology into a specified design algorithm in order to analyze the infringement cases/precedents. It also finds the nearest infringement judgements/precedents and description logic. The results from part (2) constructing the knowledge repository for the Rules engines are shown in Figure 3.

(3) Reasoners: Reasoners add inference to the Semantic Web. Inference creates logical additions that offer classification and realization. Classification populates the class structure, allowing concepts and relationships to relate properly to others, such as a person is a living thing, father is a parent, married is a type of relationship, or married is a symmetric relationship. Realization offers the same, for example, John H is the same as JH, for instance. There are several types of reasoners offering various levels of reasoning. Reasoners often plug in other tools and frameworks. Reasoners leverage asserted statements to create logically valid ancillary statements [4].

(4)The semantic web components of Rules engines support inference typically beyond what can be deduced from description logic. The engines add a powerful dimension to the knowledge constructs. Rules enable the merging of ontologies and other larger logical tasks, including programming methods such as count and string searches [4].

So far, the whole approach is purely theoretical at the moment. But in the patent document analysis, we have successful employed WordNet by the word similarity matrix clustering of words and merged with the similar semantic terms from a lower term dimensional approach. In the related data-mining fields, the semantic web system as WordNet can be employed to identify keywords, connect similar words, features, and sparse matrix to prevent the miscarriage of patent retrievals, waste of time and risks of patent infringement as well. The WordNet can also save storage memory while advancing the accuracy of text-mining in regards to ordinary, literal, and professional meaning of keywords and promoting the retrieval speed of patent research.

 ISBN: 978-1-61208-486-2

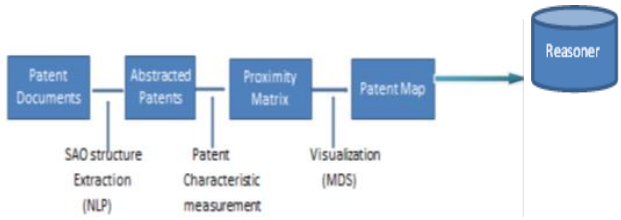

Figure 1. Part (1) the patent map result construct knowledge repository for the Reasoners modul

D. *The framework for semantic web of patent information*

The top part of framework in Figure 2 is the PATExpert, Ontology Modules and use of the W3C standard RDF. We implement the patent information in this process. The framework will construct the ontology base for Semantic Web shown in Figure 3.

As Reasoners module, the patent documents analysis process includes SAO structure extraction (NLP) and patent characteristic measurement and visualization (MDS). Here, we attempt to generate the patent concepts and relationships. In this phase, the study has generated some results based on the past research.

The other part in Figure 3 is the Rules engine modules that represent the content analysis research process [7]. The process is to implement the patent infringement judgments. The framework of the process is designed as an algorithm. Then, the study will construct the knowledge and technology logic in order to support the semantic web.

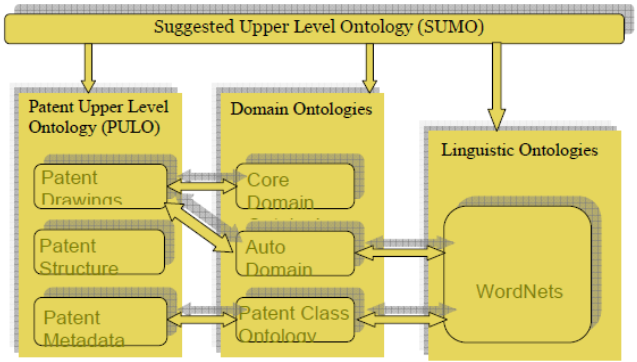

Figure 2. PATExpert Ontology Modules[18]

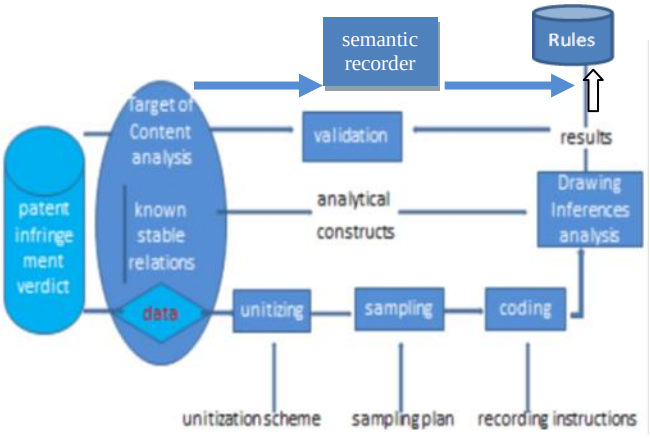

Figure 3. Part (2)content analysis result construct knowledge repository for the Rules engines

IV. EXPECTED RESULT AND FUTURE WORK

This study proposes to enhance a semantic web of patent information and patent infringement. When the user enters a keyword, the semantic web will automatically analyze the text or phrases in correspondence to related patents and potential patent threats. It can also provide the knowledge, technology, knowhow trend, and analysis. The challenge is in developing a semantic web that has the AI function.

So far the study experiment has explored WordNet to enhance the accuracy of patent searches. The evaluation index for this experiment is "Precision", "Recall" and "F-measure", and "Precision" implies that how much documents are retrieved by the system and how much are necessarily retrieved. "Recall" means how many documents need to be retrieved and how many need to be retrieved by the system. "F-measure" is compared to the harmonic mean of Precision and Recall.

A patent engineer will mark a most familiar patent document from the experiment in comparison with ten other similar articles of different patents, as the first row shows in Table I. Given the similarity threshold of hypothesis, the mark over "High" refers to Doca, Docb, Docc and Docd which is shown by implementing WordNet.

Given the same similarity threshold of hypothesis above, failure to use the similarity WordNet shown in the second row in Table I, Doca、Docb、Docf will be indicated as retrieved files. Without the implementation of WordNet, the value of "precision" appears to be "High", "Recall" is "Mid", and "F-measure" is "High".

By incorporating the similarity of the WordNet as the third row shows in Table I, given the similarity threshold is "High" or equivalent to "High", Doca、Docb、Docc、Docf and Docg are indicated as the retrieved files. Thus, by incorporating WordNet, the value of "Precision" tends to be is "High", "Recall" is becoming "More High", and "F-measure" also indicates "More High". Finally, the value of "F-measure" with WordNet is higher than the value

 ISBN: 978-1-61208-486-2

of "F-measure" without WordNet. Thus, the integration of WordNet is more likely to generate more precise meanings for patent searches.

The finding of the experiment is that WordNet can generate more wording accuracy of text-mining as to ordinary, literal and professional meanings of keywords, while promoting the retrieval speed of patent research and mitigating waste of time.

TABLE I. WORDNET SIMILARITY COMPARISON TABLE

	Patent Engineer marked similarity	No include WordNet income of similarity	Include WordNet income of similarity
Similarity(Doc_x,Doc_a)	**Most High**	**High**	**More High**
Similarity(Docx,Doc_b)	**More High**	**More High**	**More High**
Similarity(Docx,Doc_c)	**More High**	Mid	**High**
Similarity(Docx,Doc_d)	**High**	Low	Mid
Similarity(Docx,Doc_e)	Mid	Low	Mid
Similarity(Docx,Doc_f)	Low	**High**	**More High**
Similarity(Docx,Doc_g)	More Low	Mid	**High**
Similarity(Docx,Doc_h)	More Low	Low	Mid
Similarity(Docx,Doc_i)	Most Low	Most Low	More Low
Similarity(Docx,Doc_j)	None	Most Low	More Low

The study difficulties are in employing different analysis methods to analyze different databases and further, integrating these analysis results with the semantic web. An accurate algorithm in different fields must be constructed and achieved with the semantic web.

The obstacles are in integrating much more research math and different databases. Results need to be standardized and communicated, compared, and exchanged with each other.

Furthermore, the study aims to employ different analysis methods to analyze various databases with the analysis results in Semantic Web. An accurate algorithm in different fields can be feasibly constructed and achieved in the analysis of patent information and patent infringement.

The next steps will be to employ the Semantic Web impacts functions to increase the new data automatically. Figure 4 indicates major Semantic Web components: the right side is Rules engine, the left side is the Reasoner, over the center side is base ontology from the PATExpert ontology modules, and under the center side is language. The next steps of Semantic Web will be developed to use different languages, construct a multi-language Semantic Web, in order to retrieve from different country's patent databases.

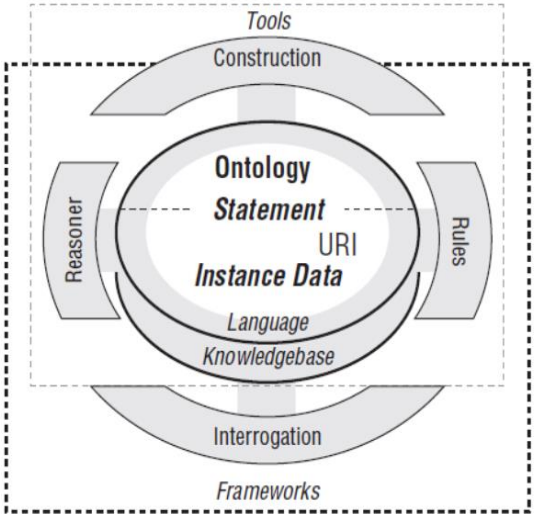

Figure 4. Major Semantic Web components [4]

The Semantic Web can employ the image recognition functions to identify drawings and pictures. If the Semantic Web has the capability of analyzing drawings and pictures, the accuracy of the results will be increased in the future.

References

[1] Y. L. Chen, and Y. T. Chiu, "An IPC-based vector space model for patent retrieval" Information Processing and Management, pp.309-322.47, 2011.

[2] C. J. Fall, A. Torcsrari, K. Benzineb, and G. Karetka, "SIGIR Forum" Automated categorization in the international patent classification, pp.10-25. 37(1), 2003.

[3] V. K. Gupta, and N. B. Pangannaya, Carbon nanotubes; "Bibliometric analysis of patents", World Patent Information, pp.185-189.Vol.22,issue 3, Sep. 2000.

[4] J. Hebeler, M. Fisher, R. Blace and A. Perez-Lopez, "Semantic Web Programming "Wiley Publishing, Inc.2009.

[5] S. H. Huang, C. C. Liu, C. W. Wang, H. R. Ke, and W. P. Yang, "International Computer Symposium" Knowledge annotation and discovery for patent analysis, pp.15-20, 2004.

[6] J.H. Kim, and K.S. Choi, Patent document categorization based on semantic structural information, "Information processing & Management", pp.1200-1215.43(5), 2007.

[7] K. Krippendorff, Content analysis In E. Barnouw, G. Gerbner, W. Schramm, T. L. Worth, and L. Gross (Eds.), International encyclopedia of communication New York, NY: Oxford University Press, pp.403-407.Vol. 1, 1989.

[8] K. Krippendorff, "Content Analysis An Introduction to Its Methodology" second Edition, Sage Publications, Inc. 2004.

[9] J. B. Kruskal, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika, pp.1-27.29(1), 1964.

[10] I. S. Kang, S. H. Na, J. Kim, and J. H. Lee, "Information Processing & Management"Cluster-based patent retrieval, pp.1173-1182.43(5), 2007.

[11] Knowledgist retrieves, analyzes, and organizes information. https://invention-machine.com/ , retrieved: Apr., 2016

[12] L. S Larkey., Some issues in the automatic classification of U.S. patents. In: Working notes for the AAAI-98 workshop on learning for text categorization, pp.87-90, 1998.

[13] L. S. Larkey, Connell, M. E., and Callan, J. Collection selection and results merging with topically organized US patents and TREC data.

In Proceedings of ninth international conference on informaiton knowledge and management, pp.282-289, 2000.

[14] L. S. Larkey, A patent search and classification system. In: Proceedings of the fourth ACM conference on digital libraries, pp.79-87, 1999.

[15] Y. Liang, R. Tan, and J. Ma, "Patent Analysis with Text Mining for TRIZ" IEEE ICMIT, pp.1147-1151, 2008.

[16] J. Michel, and B. Bettels, "Patent citation analysis: a closer look at the basic input data from patent search reports", Scientometrics, pp.185-201. Vol.51. no. 1, 2001.

[17] MINIPAR http://webdocs.cs.ualberta.ca/~lindek/minipar.htm , retrieved:Dec., 2015

[18] PATExpert http://cordis.europa.eu/ist/kct/patexpert_synopsis.htm , retrieved: Feb., 2016

[19] PATExpet http://www.barcelonamedia.org/report/the-european-project-patexpert-coordinated-by-bm-finishes-with-fulfilled-objectives-and-success , retrieved: Feb., 2016

[20] U. Schmoch, "International Journal of Technology Management" Evaluation of technological strategies of companies by means of MDS maps., pp.4-5.10(4-5), 1995.

[21] The Stanford Natural Language Processing Group, The Stanford Parser: A statistical parser, http://nlp.stanford.edu/software/lex-parser.shtml

[22] Y. H. Tseng, C. J. Lin, and Y. I. Lin, "Text mining for patent mapanalysis", Information Processing & Mangement, pp.1216-1247. vol.43, issue 5, Sep. 2007.

[23] A. J.C. Trappey, F. C. Hsu, C V. Trappy,and C. I. Lin," Development of a patent document classification and search platform using a back-propagation network", Expert Systems with Applications, pp.755-765.31(4), 2006.

[24] WordNet https://wordnet.princeton.edu/, retrieved: May, 2016

 ISBN: 978-1-61208-486-2