Multilinugal Patent Knowledge Analysis

Aviv Segev Jussi Kantola
Department of Knowledge Service Engineering
KAIST
Daejeon, Korea
{aviv, jussi} @kaist.edu

***Abstract*—Knowledge extraction for patent requests depends on analyzing current state of the art in multiple languages. Currently the process is usually limited to the languages the patent seeker knows. The paper describes a model for representing the patent request by a set of concepts related to existing multilingual knowledge ontology. The search for patent information is based on Fuzzy Logic decision support, allowing a multilingual search. The model was analyzed in assisting the decision process in the Korean Patent Office based on patents in the Korean, English, and Chinese languages.**

Keywords- Multilingual Knowledge, Decision Support, Patent, Fuzzy Logic, Ontology

I. INTRODUCTION

Patent knowledge analysis is a challenging task due to the language barrier in the growing number of open markets. Difficulties in patent knowledge analysis arise since patents in different countries are in multiple languages and are not classified under one classification system. Patent support service is required to assist in identifying similar domains and patterns that would facilitate the decision whether to grant the patent request [4].

Knowledge analysis for patent request usually involves identifying the main concepts of the invention and searching for existing documents relating to the innovation. The process of knowledge analysis is usually limited to the languages of the patent seeker.

The use of automatic tools for language translation has been suggested as a solution for multilingual applications [21]. However, this solution is not viable, since automatic machine translation (MT) today has yet to achieve a level of proficiency comparable to human translation [8]. Furthermore, while human translation can identify errors and deficiencies that can be corrected or improved, MT has yet to acquire this ability. A person who makes a mistake once can learn for the future, but MT still cannot. Currently, any prospect of a fully automatic general-purpose system capable of good quality translation without human intervention is beyond the scope of MT.

The patent knowledge extraction method described in this paper presents a model for designing a service based on multilingual ontology for the domain representation of the patent request combined with Fuzzy Logic for the decision support. The main advantage is both that the knowledge representation supplied by the multilingual ontology modeling technique is utilized and that the user is presented with powerful reasoning of knowledge extraction using the Fuzzy Logic methods. The model is based on two types of inputs. The first type is the patent request document, which is written in free text. The second type is the queries, which can be either structured or free text, asked by the service user, the patent officer. The service assists in extracting relevant knowledge for determining the likelihood that the patent request is covered by previous patents or existing knowledge. The service allows the decision maker an option to drill down and identify the reasoning and to modify the requirements or the decision qualifications for each patent request.

The Knowledge Extraction Patent Service model is described in Figure 1 and includes the following main modules: The *Patent Knowledge Extraction* process is based on extracting information from the free text based documents. The extraction process includes the identification, in multiple languages, of keywords that describe the context of the patent request and the association of relevant weights to each descriptor. The *Patent Domain Representation* is based on using a multilingual ontology that allows all existing patents to be mapped according to the predefined concepts. The process allows the patent officer to create new concepts according to which existing patents can be automatically classified. The process can also be used to cluster the patents in order to seek new patent classifications. The *Multilingual Domain Representation* involves a process directed by the patent officer of classifying the patent domain according to user perspective of knowledge. The patents are represented in multiple languages by general concepts and by an existing structure according to which the patent office workers define the patent. The problem of patent search is that the inquirer cannot always find those documents having the maximum relevance. The reason for this is the crisp approach of searching for the relevance. Fuzzy Set theory [23] and Fuzzy Logic [24] provide a robust and tractable way to move away from a precise search approach. An imprecise fuzzy patent search can find related documents that otherwise cannot be found. This is possible when we introduce the degree of relevance to the patent search. Thus, the knowledge interface becomes fuzzy - like it is in the real

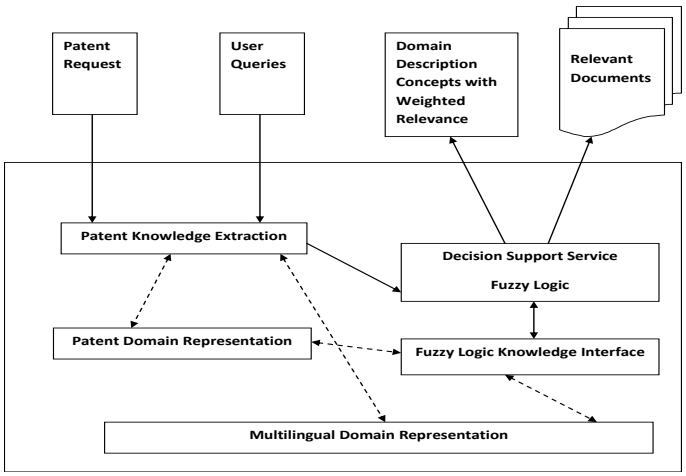


Figure 1 – Knowledge Extraction Patent service model

world. The *Fuzzy Logic Knowledge Interface* presents the weighted concepts that were automatically extracted to describe both the patent domain and the multilingual domain representation. The *Fuzzy Logic Decision Support Service* allows the user to modify the result by adjusting the fuzziness level and marking more relevant results to optimize the recall and satisfy the precision performance.

The rest of the paper is organized as follows. The next section describes the related work. Section III presents the patents service model. Section IV presents the implementation performed on real patents and analyzed with Korean Intellectual Property officers using Korean, English, and Chinese patents. Section V describes experiments and results. Section VI presents discussion and conclusions.

II. RELATED WORK

Ontologies have been defined and used in various research areas, including philosophy (where it was coined), artificial intelligence, information sciences, knowledge representation, object modeling, and most recently, eCommerce applications. In his seminal work, Bunge defines Ontology as a world of systems and provides a basic formalism for ontologies [3]. Typically, ontologies are represented using Description Logic [2], where subsumption typifies the semantic relationship between terms, or Frame Logic [9], where a deductive inference system provides access to semi-structured data.

Recent work has focused on ontology creation and evolution and in particular on schema matching. Many heuristics were proposed for the automatic matching of schemata (e.g., Cupid [13] and OntoBuilder [7]), and several theoretical models were proposed to represent various aspects of the matching process, e.g., [15].

The realm of information science has produced an extensive body of literature and practice in ontology construction, e.g., [20]. Other undertakings, such as the DOGMA project [19], provide an engineering approach to ontology management. Work has been done in ontology learning, such as Text-To-Onto [14] and Mapping Context to Ontology [17], to name a few. Finally, researchers in the field of knowledge representation have studied ontology interoperability, resulting in systems such as Protégé [16].

Fuzzy Logic is reasoning with imprecise things. Fuzzy Logic has two principle components. The first is a translation system for representing the meaning of propositions and other semantic entities. The second component is an inferential system for arriving at an answer to a question that relates to the information resident in a knowledge base [25]. Fuzzy logic provides decision support systems with powerful reasoning capabilities. Vagueness in linguistics can be captured mathematically by applying fuzzy sets [11]. Fuzzy sets represent objects and concepts better than do crisp sets. There are two reasons for this. First, the predicates in propositions representing a system do not have crisp denotations. Second, explicit and implicit quantifiers are fuzzy [25]. A fuzzy set can be defined mathematically by assigning to each possible individual in the universe of discourse a value representing its grade of membership in the fuzzy set. This grade corresponds to the degree to which that individual is similar to or compatible with the concept represented by the fuzzy set [10].

An ongoing work in the European Union called PATexpert [22] targets several areas of patent services. The goal of the project is to bring patent services to a new level by applying several new approaches and methods to various areas in patent services. The search method proposed in this paper is different from the approach described in PATexpert. First, in PATexpert the classification process is manual. In our method the classification/search is a semi-automatic process. Second, the meaning of fuzzy in PATexpert is in the morphological and spelling sense. In the method proposed in this paper, the fuzzy refers to Fuzzy Sets and Fuzzy Logic for the reasoning and decision making process.

There have been many publications about fuzzy information or document retrieval from the early 1970s till today, see for example [1],[5], and [12], but we could not find any work about fuzzy concept search, as described in this paper. We believe that the value of this research in comparison to existing research lies in the joint application of ontology matching and Fuzzy Sets that enables a searcher-friendly service which considerably decreases the search time period and expands the relevant results.

III. PATENT SERVICE MODEL

The implementation of the model begins when the patent office user begins the process of evaluating the patent request. A simple syntactic search might look for documents relating to a term, such as *Length*, which appears in the text. However, the described model expands the search results to include documents related to additional concepts not mentioned in the text.

A. Patent Knowledge Extraction

Each claim is analyzed separately through the Domain Representation process. To analyze the claims, a context

extraction algorithm can be used. To handle the different vocabularies used by different information sources, a comparison based on context is used in addition to simple string matching. For each document the context is extracted by the Patent Knowledge Extraction and then compared with the ontology concept by the Patent Domain Representation.

The Patent Knowledge Extraction process uses the World Wide Web as a knowledge base to extract multiple contexts in multiple languages for the textual information. The algorithm input is defined as a set of textual propositions representing the claim information description. The result of the algorithm is a set of contexts - terms that are related to the propositions in multiple languages. The context recognition algorithm was adapted from [18] and consists of the following three steps:

1. Context retrieval: Submitting each token to a Web-based search engine. The contexts are extracted and clustered from the results.
2. Context ranking: Ranking the results according to the number of references to the keyword, the number of Web sites that refer to the keyword, and the ranking of the Web sites.
3. Context selection: Assembling the set of contexts for the textual proposition, defined as the outer context.

The *external* weight of each context is determined according to the number of retrieved Web references related to the concept and the number of references to the concepts in the patents. In addition, the Term Frequency/Inverse Document Frequency (TF/IDF) method analyzes the patent from an *internal* point of view, i.e., what concept in the text best describes the patent.

B. Patent Domain Representation

The Patent Domain Representation performs the ontology matching process that directs the claim to the relevant ontological concepts. One of the difficult tasks is matching each information datum with the correct concepts without the usual training process required in ontology adjustment and usually performed over a long period of time.

To process the ontology for optimal information flow, the following method is proposed. A simplified representation of an ontology is $O \equiv < C, R >$, where $C = \{c_1, c_2, \ldots, c_n\}$ is a set of concepts with their associated relation *R*. A concept can consist of multiple context descriptors and can be viewed as a meta-representation of the patent domain. The added value of having such a meta-representation is that a concept is associated with multiple contexts, each in a different language. Furthermore, each context descriptor can belong to several ontology concepts simultaneously. For example, a context descriptor <Length, 2> can be shared by many ontology concepts that have interest in length analysis, such as 거리(*Distance* in Korean) or 波 (*Wave* in Chinese), although it is not in their main role definition (and hence, low weight is assigned to it).

The weight is calculated according to the number of references to the concept in the Web combined with the number of references to the concept in the document. For example, a patent can be associated with concept 거리 (*Distance)* with weight 0.4 and concept 波 *(Wave)* with weight 0.3. To evaluate the matching of the concepts with the information and its context, a simple string-matching function is used, denoted by $\boldsymbol{match_{str}}$, which returns 1 if two strings match and 0 otherwise.

C. Multilingual Domain Representation

When a new patent request is processed, the first step involves the multilingual ontology matching process. Once the patent request is classified, the following relations with existing patents can occur:

- If the patent is related to concepts associated with existing patents, the decision process requires reviewing the existing patents and comparing them to the request.
- If the patent is not related to concepts similar to existing patents, the decision maker can extend the search according to related concepts until related patents are identified with overlapping concepts associated with the patent request.

If the second option is encountered, the decision maker faces a dilemma of whether to grant the patent based on the relation of existing patents to the current patent. To assist in the process of decision making in these instances, a fuzzy logic process is presented.

D. Fuzzy Logic Knowledge Interface

In fuzzy information retrieval the relevance of the index terms is expressed by a fuzzy relation: $R: X \times Y \rightarrow [0, 1]$ where the membership value $R(x, y)$ for each x_i and y_i represents the grade of relevance of index term x_i to document y_i [1]. The basic scheme of fuzzy information retrieval is where U1 is a fuzzy set representing a particular inquiry. When U1 is composed with Thesaurus (T), then U2 becomes an inquiry augmented by associated index terms: $U2 = U1 \circ T$. U2 can be expressed as follows: $U2(x_i)=\max \min [U2(x_i), T(x_i , x_j)]$. Then a relevant document search can be expressed by: $D = U2 \circ R$. Usually $\circ$ is understood as the max-min composition (max-min implication) [1]. The role of Fuzzy Thesaurus T can be carried out by a set of ontologies that are further linked to the lexical database Wordnet [6], [c.f. 23]. In the proposed approach the role of the fuzzy thesaurus (T) is carried out by the ontology matching process (O). The basic scheme of fuzzy information retrieval U2 becomes an inquiry augmented by associated index terms from ontology matching: $U2 = U1 \circ O$ (Figure 2).

The inquirer can inspect all the documents that have support D, or she can filter the inspection to those supported by some α-cuts [1]. The search index must have full relevance to the document index. The inquirer can "expand" the patent inquiry by setting α-cut to a lower level. For example, α-cut level 0.5 would also bring up those

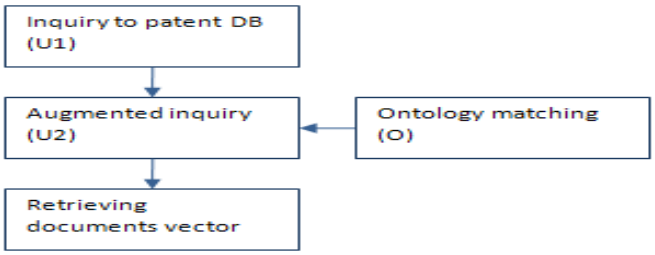


Figure 2 - Fuzzy information retrieval and ontology matching scheme

documents that are meaningful to a specific search but not to a full degree. Setting α-cut to a very low level would bring up those documents that are vaguely related to a given inquiry. A person finds it difficult or impossible to think of the concepts that are vaguely related to a given inquiry. That is the justification for using ontology matching to augment the original inquiry.

E. Decision Support Service Fuzzy Logic

In the proposed approach the user can expand the search to other possibly related concepts by selecting a mode for extended search by choosing *Strict* mode or *Vague* mode. In the *Strict* search mode the system is tuned to find those patent documents that are closely related to the original document, and in the *Vague* search mode the system is set up to find documents that are loosely related to the original document. The user enters a document into the Web based ontology matching process. A list of related concepts, together with the degrees of relevance, is presented. The degree of relevance (μ) is calculated based on the concept weight in searched documents provided by the ontology matching algorithm and fuzzy membership functions. The fuzzy set defined by the membership function is different for the "Strict" and for the "Vague" search modes.

The Strict and Vague membership functions result in different degrees of relevance with the same weight from the ontology matching algorithm. For example, the weight 0.28 for the 波 *(Wave)* concept from the ontology matching algorithm results in 0.5 (degree of relevance) according to the Vague membership function but only in 0.23 according to the Strict membership function. Concept weight 0.06 for the 거리 (*Distance)* concept returns 0.32 in Vague mode and 0 in Strict mode. The parameters for the membership functions were adjusted according to tests performed during the model implementation.

Figure 3 illustrates how the α-cuts are used to filter the new expanded set of results. For example, in Strict mode the 波 concept is part of the new expanded index set if the α-cut is set to a level of 0.15. However, the 거리 concept is not part of the result set if the α-cut level is 0.48.

According to this proposed method, the patent officer can carry out expanded searches by using her own language. Therefore, the user does not need to convert meanings to some numerical scale, index, or variable. The method offers more meaningful results and at the same time provides a more human-like search approach for the users.

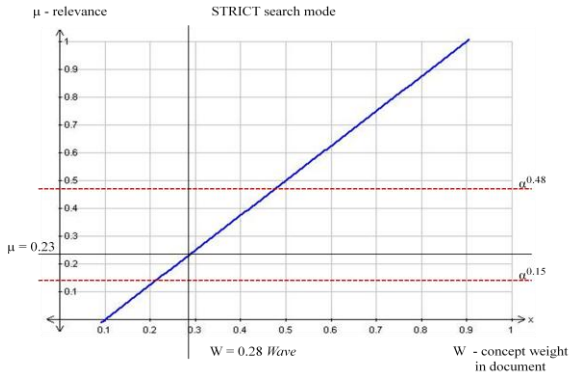


Figure 3 - The relevance of concepts

IV. PATENT SERVICE MODEL IMPLEMENTATION

The implementation of the model is currently being tested at the Korean Intellectual Property Office (KIPO). KIPO seeks to improve the ability to identify and classify new patents. KIPO's goal is to optimize the examination infrastructure, improve the quality of examinations, and enhance the effectiveness of quality management.

There is a new international grouping of major Intellectual Property (IP) offices. The trilateral cooperation among the US, Japan, and Europe has been expanded to include Korea and China. These five major offices, known as IP5, are undertaking ten foundation projects designed to improve the quality of examinations and promote the creation of high-quality patents. The IP5 offices handle an aggregate of approximately 1.35 million patent applications, which represent 76 percent of all the patent applications filed throughout the world. There are about 173 million pieces of patent information on the database as of 2008 and the quantity of information is increasing, up by 14 million pieces from 2007 to 2008.

Figure 4 shows the Fuzzy Logic Ontology Context Knowledge (FLOCK) demonstrator application that was used to test the model described in this paper. The basic steps in the use of the demonstrator are as follows:

Load a "New patent" document.

1. Select Vague or Strict search mode from the radio button list.
2. Set the filter (alpha cut level) to a suitable level. The top filter filters the Internal (I) concepts based on the TF/IDF algorithm. The bottom filter filters the External (E) concepts based on the Web context retrieval.
3. Discard some general search terms, such as *map*, *design*, and *music*, by selecting those search terms. The result of steps 2 and 3 can be seen in the "Search terms" list automatically.
4. Approve the search terms (Approved search terms list) by clicking either A) Search patents (string) button or B) Search patents (degree) button to locate the target folder for patent documents and to search for relevant documents. The String search is traditional string matching search, whereas the Degree search compares the context matching index of the new patent application to the context matching indexes of the existing patents.

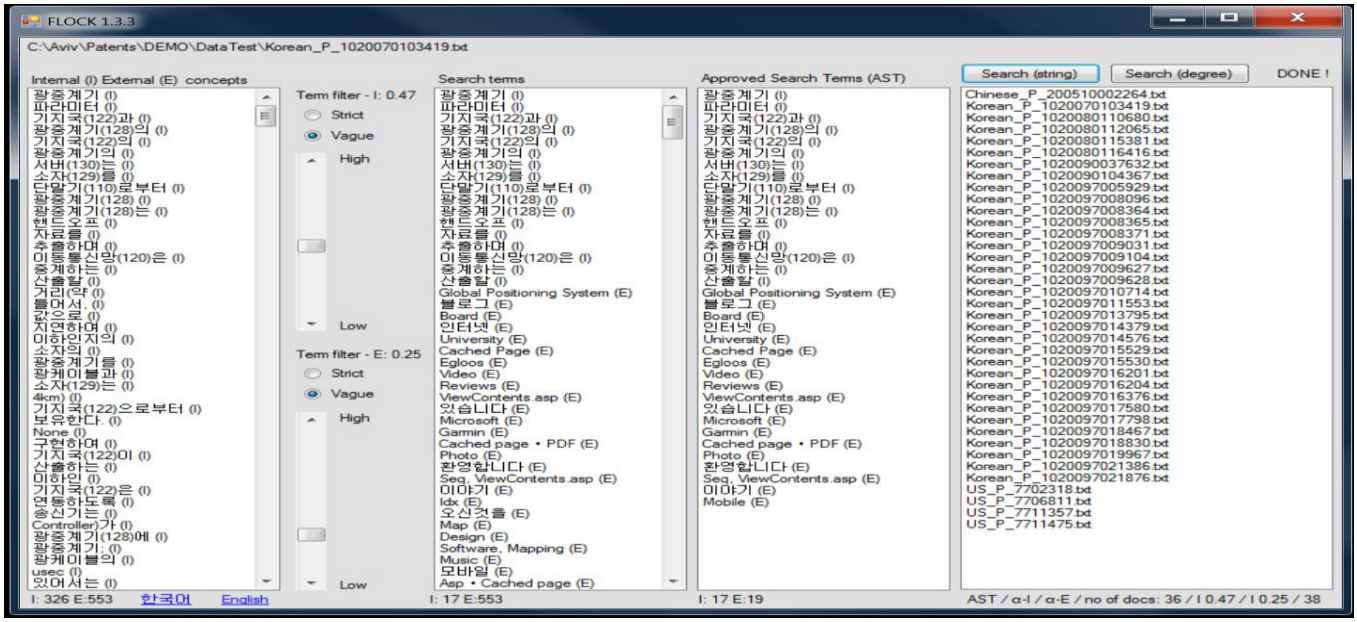


Figure 4 – The FLOCK demonstrator tested at KIPO

5. See the documents found by the application on the list on the right. The patent officer can now look into those existing patents.

The proposed method was tested in Korean, English and Chinese. The context matching algorithm searches the Internet using the language in the new patent application and the results are extracted in multiple languages, allowing the patent database to be searched in multiple languages. For example, a new patent application written in Korean is matched against Internet content written in Korean, English, and Chinese and patents written in all these languages can be searched.

The FLOCK system for extracting concepts and relevant patent documents was evaluated by six KIPO Patent Officers who routinely process patent requests. A patent officer regularly analyzes each patent claim in relation to all existing patents worldwide.

V. EXPERIMENTS

The paper describes a model for representing the patent request by a set of concepts related to existing knowledge in multiple languages. The search for patent information is based on applications of Fuzzy Sets and Fuzzy Logic decision support to allow the query expansion for relevant documents. The model was analyzed to evaluate the relevance of the patents extracted in multiple languages.

A. Data Set and Methods

The data consists of a total of 169 patents extracted from the Korean Intellectual Property Office, United States Patent and Trademark Office, and China Patent and Trademark Office. The patent documents included free text description of the patents from classifications such as: location based systems, organic, and food. The patents collected were processed through the Patent Knowledge Extraction process as described in Section III.A. The patents were analyzed using the Fuzzy Logic module as described in III.D. The interface is based on the FLOCK system as described in section IV.

The experiments analyzed precision and recall of the patent extraction process. The precision is calculated as the fraction of retrieved patents relevant to the search divided by all the retrieved patents. The recall is calculated as the fraction of retrieved patents relevant to the search divided by all the relevant patents.

B. Experiments Results

The first set of tests analyzed precision versus recall for the patents. A randomly selected set of 10 patents was used and the precision and recall were calculated for each one at predefined alpha cut values. An ideal result for a recall versus precision graph would be a horizontal curve with high precision value; a poor result has a horizontal curve with a low precision value. The recall-precision curve is widely considered by the Information Retrieval community and patent officers to be the most informative graph showing the effectiveness of the methods. The average precision versus recall is displayed in Figure 5. The results present high relevance and accuracy with precision falling below 80% only when recall reaches 65.56%.

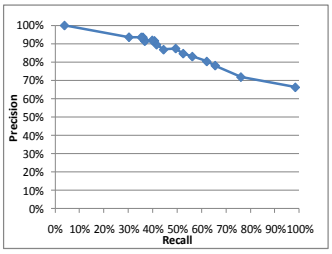
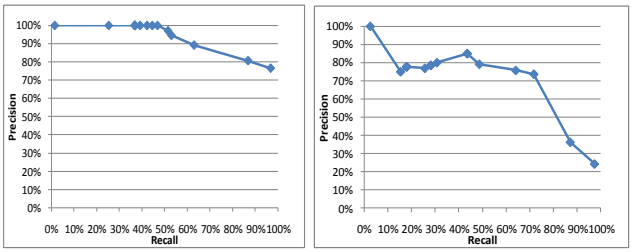


Figure 5 – Precision vs. recall average for 10 results

Figure 6.A presents the worst sampled patent results where the precision drastically declines after the recall increases over 73.68%. The sharp decline can be explained by an increasing amount of irrelevant concepts that are added to the concept collection at this stage.Manual filtering



A B

Figure 6 – Precision vs. recall - worst and best sample cases

by the patent user can decrease the decline. Figure 6.B presents the best sampled patent results. The results achieve 100% precision until the recall drops below 46.92%.

The second set of experiments analyzes how the increase in the number of languages used in the data set influences the recall and precision. Figure 7 presents two data sets. The first data set includes only the Korean patents. The second data set includes the Korean, US, and Chinese patents. The recall versus precision results displays a minimal difference between the two graphs at any specific point. Furthermore, the increase in the number of languages did not decrease all the values to create a similar graph shifted downward as expected. The results suggest that increasing the number of languages used can have minor effects on the model.

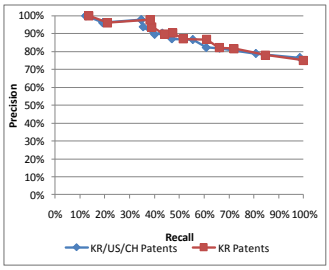


Figure 7 – Korean versus multiple languages (Korean, English, & Chinese)

VI. Discussion and Conclusion

The patent search model described in the paper allows queries to be performed in multiple languages. The model shows promise in extending the field of patent search where the patent inquirer or decision maker can automatically classify the concepts related to the patent, unlike manual patent classification used in the past [22]. The results show the advantage of query expansion in the search process based on extracting relevant information from the Web instead of limiting the search to concepts that appear in the patent itself. The results show high precision versus recall results. The method allows the user to perform a gradual expansion of the related work using Fuzzy Sets and assists in minimizing the time required to make a patent-related decision. Future work includes analyzing the model in relation to the strict versus vague fuzzy search modes, as well as analyzing additional fuzzy reasoning methods.

References

[1] Aliev, R.A., Aliev, R.R., Soft Computing and Its Applications: World Scientific, Singapore (2001)

[2] Borgida, A., Brachman, R.J., "Loading data into description reasoners," In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 217–226. ACM, (1993)

[3] Bunge, M., "Treatise on basic philosophy," vol. 4: Ontology II: A World of Systems. D. Reidel Publishing Co., Inc., New York (1979)

[4] Cong H., Tong L. H., "Grouping of TRIZ inventive principles to facilitate automatic patent classification", Expert Systems with Applications, 34, pp. 788‐795 (2008)

[5] Cross, V., "Fuzzy information retrieval," Journal of Intelligent Information Systems, 3 (1), pp. 29-56 (1994)

[6] Fellbaum C., "WordNet: An electronic lexical database," Cambridge (MA): MIT Press (1998)

[7] Gal, A., Modica, G., Jamil, H.M., Eyal, A., "Automatic ontology matching using application semantics," AI Magazine, 26(1), 2005.

[8] Hutchins J., "Current commercial machine translation systems and computer-based translation tools: System types and their uses," Int. Journal of Translation, 17(1-2), pp. 5-38, 2005.

[9] Kifer, M., Lausen, G., Wu, J., "Logical foundation of object-oriented and frame-based languages," Journal of the ACM 42, 1995.

[10] Klir, J. G., Yuan, B., Fuzzy Sets and Fuzzy Logic, Theory and Applications, Prentice – Hall Inc., 1995.

[11] Lin C.-T., Lee C. S., Neural fuzzy systems: A neuro-fuzzy synergism to intelligent systems, Prentice-Hall, Inc., 1996.

[12] Lucarella, D., Morara, R., "FIRST: Fuzzy Information Retrieval SysTem," Journal of Information Science, 17 (2), pp. 81-91, 1991.

[13] Madhavan, J., Bernstein, P.A., Rahm, E., "Generic schema matching with Cupid," In: Proceedings of the International conference on Very Large Data Bases (VLDB), pp. 49–58, Rome, Italy, 2001.

[14] Maedche, A., Staab, S., "Ontology learning for the semantic web," IEEE Intelligent Systems 16, 2001.

[15] Melnik, S. (ed.), Generic Model Management: Concepts and Algorithms, Springer, Heidelberg, 2004.

[16] Noy, F.N., Musen, M.A., "PROMPT: Algorithm and tool for automated ontology merging and alignment," In: Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000), pp. 450–455, Austin, TX, 2000.

[17] Segev A., Gal A., "Putting things in context: A topological approach to mapping contexts to ontologies," Journal on Data Semantics IX, pp. 113–140, 2007.

[18] Segev A., Leshno M., Zviran M., "Context recognition using internet as a knowledge base," Journal of Intelligent Information Systems, 29(3), 2007.

[19] Spyns, P., Meersman, R., Jarrar, M., "Data modelling versus ontology engineering," ACM SIGMOD Record 31(4), 2002.

[20] Vickery, B.C., "Faceted classification schemes," Graduate School of Library Service, Rutgers, the State University, 1966.

[21] Vossen P., Eurowordnet general document, LE2-4003 LE4-8328, EuroWordNet, 1999.

[22] Wanner, L., Baeza-Yatesa, R., Brügmann, S., Codina, J., Diallo, B., Escorsa, E., Giereth, M., Kompatsiaris, Y., Papadopoulos, S., Pianta, E., Piella, G., Puhlmann, I., Raoe, G., Rotard, M., Schoester, P., Serafini, L., Zervaki, V., "Towards content-oriented patent document processing," World Patent Information, 30(1), pp. 21-33, 2008.

[23] Zadeh, L.A., "Fuzzy sets," Information and Control, Vol. 8, pp. 338-353, 1965.

[24] Zadeh, L.A., "Outline of a new approach to the analysis of complex systems and decision processes," IEEE Transactions on Systems, Man, and Cybernetics, Vol. 1, No. 1, pp. 28-44, 1973.

[25] Zadeh, L. A., "Commonsense knowledge representation based on fuzzy logic," Computer, Vol. 16, October, pp. 61-65, 1983.