Expert Systems with Applications 40 (2013) 7010–7023

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Analyzing multilingual knowledge innovation in patents

Aviv Segev [a,*], Jussi Kantola [b], Chihoon Jung [a], Jaehwa Lee [a]

[a] Department of Knowledge Service Engineering, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 305-701, South Korea
[b] Department of Production, University of Vaasa, P.O. Box 700, Vaasa FI-65101, Finland

ARTICLE INFO

Keywords:
Conceptual modeling
Ontologies
Knowledge management applications
Database semantics

ABSTRACT

In the process of analyzing knowledge innovation, it is necessary to identify the existing boundaries of knowledge so as to determine whether knowledge is new – outside these boundaries. For a patent to be granted, all aspects of the patent request must be studied to determine the patent innovation. Knowledge innovation for patent requests depends on analyzing current state of the art in multiple languages. Currently the process is usually limited to the languages and search terms the patent seeker knows. The paper describes a model for representing the patent request by a set of concepts related to a multilingual knowledge ontology. The search for patent knowledge is based on Fuzzy Logic Decision Support and allows a multilingual search. The model was analyzed using a twofold approach: a total of 104,296 patents from the United States Patent and Trademark Office were used to analyze the patent extraction process, and patents from the Korean, US, and Chinese patent offices were used in the analysis of the multilingual decision process. The results display high recall and precision and suggest that increasing the number of languages used only has minor effects on the model results.

1. Introduction

In the analysis of the boundary of knowledge, such as in the process of granting patents, there is a difference between the need to locate knowledge and the need to identify whether similar knowledge exists. The search of the boundary of knowledge examines whether given concepts exist, while regular knowledge search looks for instances of existing concepts. Contemporary knowledge-based services depend on using existing knowledge, while Patent Knowledge Extraction is required to assist in identifying similar domains and patterns that will facilitate the decision whether to grant the patent request (Cong & Tong, 2008). Furthermore, another difficulty is that patents in different countries are not classified under one classification system and employ multiple languages.

Conversely, to invalidate a patent, relevant documents must be identified as "prior art", open to the public before the patent was filed. Analysis of patents involves searching for relevant patents and documents that could invalidate a claim within the patent or for a set of patents that could invalidate a claim when used together.

The main problem encountered when searching for existing patents is verifying that all relevant documents related to the current invention were retrieved. If a relevant document is missed, low recall, then a patent could be granted to an already existing work. Conversely, retrieving an irrelevant document, low precision, would only lead to minor additional work from the patent inquirer or decision maker. The current decision process for granting patents averages 3–4 years depending on the specific field of technology. The main advantage of the model presented here is that it decreases the time required to review a patent request by supplying a semi-automatic guided search. The model aims at benefitting both the patent office decision maker who needs to decide whether to grant a patent for each request and inventors and companies that would like to inquire about existing patented technology.

In the growing number of open markets, the identification of patent knowledge is a challenging task due to the language barrier. Analyzing knowledge innovation for a 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 Patent Knowledge Extraction method described in this paper presents a model based on ontology for the domain representation of the patent request combined with Fuzzy Logic for the decision support. The Patent Knowledge Extraction method has two main advantages: the knowledge is represented using the ontology modeling technique and the user is presented with powerful reasoning in knowledge extraction using the Fuzzy Logic methods.

The Patent Knowledge Extraction method is based on free text input in the language of the patent. An example of a sample patent

* Corresponding author. Tel.: +82 10 3402 1881.
E-mail addresses: aviv@kaist.edu (A. Segev), jussi.kantola@uwasa.fi (J. Kantola), chihoon.jung@kaist.ac.kr (C. Jung), tkod119@kaist.ac.kr (J. Lee).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.06.013

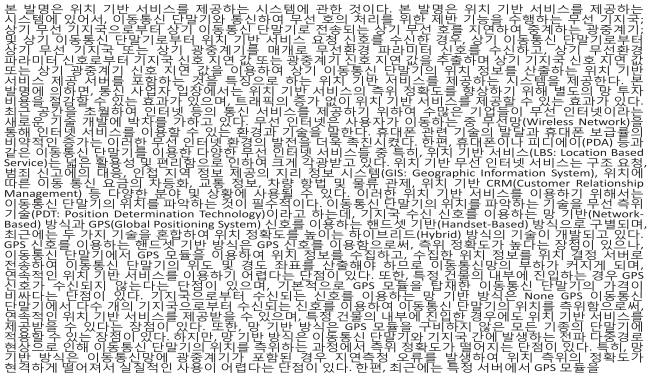
본 발명은 위치 기반 서비스를 제공하는 시스템에 관한 것이다. 본 발명은 위치 기반 서비스를 제공하는 시스템에 있어서, 이동통신 단말기의 통신하여 무선 호의 처리를 위한 기능을 수행하는 무선 기지국; 상기 무선 기지국으로부터 상기 이동통신 단말기로 전송되는 상기 무선 호를 지원하여 중계하는 광중계기; 및 상기 이동통신 단말기로부터 위치 기반 서비스 요청 신호를 수신한 경우, 상기 이동통신 단말기로부터 상기 무선 기지국 또는 상기 광중계기를 매개로 무선환경 파라미터 신호를 수신하고, 상기 무선환경 파라미터 신호로부터 기지국 신호 지연 값 또는 광중계기 신호 지연 값을 추출하며 상기 기지국 신호 지연 값 또는 상기 광중계기 신호 지연 값을 이용하여 상기 이동통신 단말기의 위치 정보를 산출하는 위치 기반 서비스 제공 서버를 포함하는 것을 특징으로 하는 위치 기반 서비스를 제공하는 시스템을 제공한다. 본 발명에 의하면, 통신 사업자 입장에서는 위치 기반 서비스의 즉위 정확도를 향상하기 위해 별도의 망 투자 비용을 절감할 수 있는 효과가 있으며, 트래픽의 증가 없이 위치 기반 서비스 제공 수 있는 효과가 있다. 최근 광대역을 초월하여 인터넷 등의 통신 서비스를 제공하기 위하여 수많은 기업들이 무선 인터넷이라는 새로운 기술 개발에 박차를 가하고 있다. 무선 인터넷은 사용자가 이동하는 중 무선망(Wireless Network)을 통해 인터넷 서비스를 이용할 수 있는 환경과 기술을 말한다. 휴대폰 관련 기술의 발달과 휴대폰 보급률의 비약적인 증가는 이러한 무선 인터넷 환경의 발전을 더욱 촉진시켰다. 하퍼, 휴대폰이나 피디에이(PDA) 등과 같은 이동통신 단말기를 이용한 다양한 무선 인터넷 서비스들 중 특히, 위치 기반 서비스(LBS: Location Based Service)는 넓은 활용성 및 편리함으로 인하여 크게 각광받고 있다. 위치 기반 무선 인터넷 서비스는 구조 요청, 범죄 신고에의 대응, 이정 지역 정보 제공의 지리 정보 시스템(GIS: Geographic Information System), 위치에 따른 이동 통신 요금의 자등화, 교통 정보, 차량 항법 및 물류 관제, 위치 기반 CRM(Customer Relationship Management) 등 다양한 분야 및 상황에 사용될 수 있다. 이러한 위치 기반 서비스를 이용하기 위해서는 이동통신 단말기의 위치를 파악하는 것이 필수적이다. 이동통신 단말기의 위치를 파악하는 기술로 무선 측위 기술(POT: Position Determination Technology)이라고 하는데, 기지국 수신 신호를 이용하는 망 기반(Network-Based) 방식과 GPS(Global Positioning System) 신호를 이용하는 핸드셋 기반(Handset-Based) 방식으로 구별되며, 최근에는 두 가지 기술을 혼합하여 위치 정확도를 높이는 하이브리드(Hybrid) 방식의 기술이 개발되고 있다. GPS 신호를 이용하는 핸드셋 기반 방식은 GPS 신호를 이용함으로써, 측위 정확도가 높다는 장점이 있으나 이동통신 단말기에서 GPS 모듈을 이용하여 위치 정보를 수집하고, 수집한 위치 정보를 위치 결정 서버로 전송하여 이동통신 단말기의 위도 및 경도 좌표를 산출해야 하므로 이동통신망의 부하가 커지게 되며, 연속적인 위치 기반 서비스를 이용하기 어렵다는 단점이 있다. 또한, 특정 건물의 내부에 진입하는 경우 GPS 신호가 수신되지 않는다는 단점이 있으며, 기본적으로 GPS 모듈을 탑재한 이동통신 단말기의 가격이 비싸다는 단점이 있다. 기지국으로부터 수신되는 신호를 이용하는 망 기반 방식은 None GPS 이동통신 단말기에서 다수 개의 기지국으로부터 수신되는 신호를 이용하여 이동통신 단말기의 위치를 측위함으로써, 연속적인 위치 기반 서비스를 제공받을 수 있으며, 특정 건물의 내부에 진입한 경우에도 위치 기반 서비스를 제공받을 수 있다는 장점이 있다. 또한, 망 기반 방식은 GPS 모듈을 구비하지 않은 모든 기종의 단말기에 적용할 수 있는 장점이 있다. 하지만, 망 기반 방식은 이동통신 단말기와 기지국 간에 발생하는 전파 다중경로 현상으로 인해 이동통신 단말기의 위치를 측위하는 과정에서 측위 정확도가 떨어지는 단점이 있다. 특히, 망 기반 방식은 이동통신망에 광중계기가 포함된 경우 지연측정 오류를 발생하여 위치 측위의 정확도가 현격하게 떨어져서 실질적인 사용이 어렵다는 단점이 있다. 한편, 최근에는 특정 서버에서 GPS 모듈을

Fig. 1. Sample free text input – patent in Korean.

input in Korean is displayed in Fig. 1. Current methods require translation of the patent or identification of the main related issues manually before searching for similar patents in multiple languages. The proposed solution is based on the automatic identification of related concepts represented in multiple languages and on the automatic extraction of relevant documents in different languages.

The Patent Knowledge Analysis model is described in Fig. 2. The model is based on two types of inputs. The first type is the patent submission request document, which is written in free text (Fig. 1). The second type is the queries performed by the service user, the patent officer, on either structured text or free text. Queries on structured text can be performed by adjusting relevant concepts weights. Queries on free text can be performed by modifying proposed concepts descriptors. The model assists in extracting relevant knowledge for determining the likelihood that the patent request is covered by previous patents or existing knowledge. The model 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 Patent Knowledge Analysis model includes the following main modules: Patent Knowledge Extraction, Patent Domain Representation, Multilingual Domain Representation, Fuzzy Logic Knowledge Interface, and Fuzzy Logic Decision Support. The arrows represent the process flow, and the dotted arrows represent data extraction from the Patent Domain Representation, the Multilingual Domain Representation, and the storage of the Patent Ontology and the Patent Corpus.

The Patent Knowledge Extraction process is based on extracting knowledge from the free text based documents. The extraction process includes the identification of keywords that describe the context of the patent request and the association of relevant weights to each descriptor. The Patent Knowledge Extraction process forwards the knowledge to the Patent Domain Representation and Multilingual Domain Representation modules.

The Patent Domain Representation is based on using a multilingual ontology that allows all existing patents to be mapped according to the predefined concepts. Each concept is represented in multiple languages. 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 process is directed by the patent officer who classifies the patent domain according to the user perspective of the knowledge. The knowledge is usually defined according to the domain of expertise and languages of the patent officer. Consequently, a specific patent can be classified both by the general concepts and by an existing structure that defines the patent office workers' expertise. The multilingual representation allows the user to classify the patent in one language and match it with similar patents according to the multilingual ontology.

The problem of patent search is that the inquirer cannot always find those documents that have the maximum relevance, because of the crisp approach which is defined as the exact approach of searching for relevance in database systems. Fuzzy Set theory

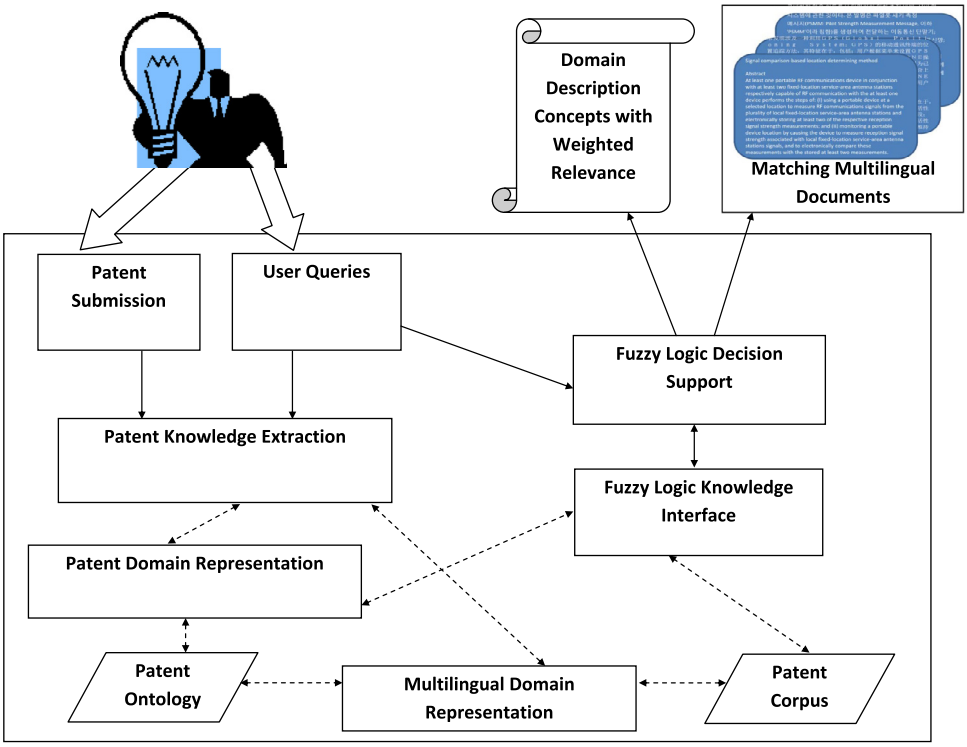

Fig. 2. Patent Knowledge Analysis model outline.

(Zadeh, 1965) and Fuzzy Logic (Zadeh, 1973) 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 world.

The Fuzzy Logic Knowledge Interface presents the weighted concepts that were automatically extracted to describe both the patent domain and the multilingual domain. The Fuzzy Logic Decision Support allows the user to modify the result by adjusting the relevance level and marking more relevant concepts 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 3 presents the Patent Knowledge Analysis model. Section 4 presents the implementation performed on real patents from the Korean Intellectual Property Office. Section 5 describes the experiments and results. Section 6 discusses the model and analyzes the implementation with officers in the Korean Intellectual Property Office and the Israeli Patent and Trademark Office, and Section 7 presents the conclusions.

2. Related work

2.1. Ontology

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 (Bunge, 1979). Typically, ontologies are represented using Description Logic (Borgida & Brachman, 1993; Donini, Lenzerini, Nardi, & Schaerf, 1996), where subsumption typifies the semantic relationship between terms, or Frame Logic (Kifer, Lausen, & Wu, 1995), where a deductive inference system provides access to semi-structured data. Ontologies are used widely used in the Semantic Web with ontology languages OWL (Bechhofer et al., 2004) and OWL 2 (W3C OWL Working Group, 2009).

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 (Madhavan, Bernstein, & Rahm, 2001), GLUE (Doan, Madhavan, Domingos, & Halevy, 2002), and OntoBuilder (Gal, Modica, Jamil, & Eyal, 2005)), and several theoretical models were proposed to represent various aspects of the matching process (Melnik, 2004; Madhavan, Bernstein, Domingos, & Halevy, 2002). The ontology matching workshop is dedicated to research on schema matching in areas such as learning of link specifications (Ngomo, Lehmann, Auer, & Höffner, 2011) and data interlinking evaluation (Euzenat, 2012).

The realm of information science has produced an extensive body of literature and practice in ontology construction, e.g. (Vickery, 1966). Other undertakings, such as the DOGMA project (Spyns, Meersman, & Jarrar, 2002), provide an engineering approach to ontology management. Work has been done in ontology learning, such as Text-To-Onto (Maedche & Staab, 2001), Mapping Context to Ontology (Segev & Gal, 2007), and OntoMiner (Davulcu, Vadrevu, Nagarajan, & Ramakrishnan, 2003), to name a few. Finally, researchers in the field of knowledge representation have studied ontology interoperability, resulting in systems such as Chimaera (McGuinness, Fikes, Rice, & Wilder, 2000) and PROMPT (Noy & Musen, 2000).

2.2. Translation and multilingual information retrieval

The use of automatic tools for language translation has been suggested as a solution for multilingual applications (Vossen, 1999). However, this solution is not viable, since automatic machine translation (MT) today has yet to achieve a level of proficiency comparable to that of human translation (Hutchins, 2005). 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. One of the factors influencing MT performance is the dependence on incorporation of the "life-meaning" of texts, drawing on the knowledge and common sense used in the lives of the speaker (Basden & Klein, 2008). Previous work used ontological concepts specified in multiple languages to assist in resolving cross-language and local variation language ambiguities (Segev & Gal, 2008). Other work developed the Latent Semantic Indexing (LSI)-based multilingual document clustering technique, which generated knowledge maps (i.e., document clusters) from multilingual documents (Wei, Yang, & Lin, 2008). However, previous work using ontological concepts analyzed the classification of existing information, while this paper deals with the identification, in multiple languages, of whether current knowledge is new.

2.3. Fuzzy Logic

Vagueness in linguistics can be captured mathematically by applying Fuzzy Sets (Lin & Lee, 1996). Fuzzy Sets represent objects and real world 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 (Zadeh, 1983). 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. A *fuzzy set* is a pair (U,m) where U is a set and $m{:}U \rightarrow [0,1]$. This grade corresponds to the degree to which that individual is similar to or compatible with the concept represented by the fuzzy set (Klir & Yuan, 1995).

Fuzzy Logic is reasoning with imprecise things. Fuzzy Logic has two principal components. The first is a translation system for representing the meaning of propositions and other semantic entities. *Fuzzy Logic* is an extension of the case of multi-valued logic, valuations ($\mu{:}V_0 \rightarrow W$) of propositional variables (V_0) into a set of membership degrees (W) can be thought of as membership functions mapping predicates into Fuzzy Sets. 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 (Zadeh, 1983). Fuzzy Logic provides decision support systems with powerful reasoning capabilities.

In an ongoing work in the European Union called PATexpert (Wanner et al., 2008), several areas of patent services are targeted. 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. An initial outline of a possible solution was presented in Segev and Kantola (2010). However, the description did not include the model, implementation, and validation presented in this work.

Research in the field of fuzzy information from the early 1970s till today has focused on document retrieval, see for example (Aliev & Aliev, 2001; Cross, 2008; Lucarella & Morara, 1991; Melnik, 2004; Miyamoto, 1990). Recent publications have focused on ontology and fuzzy theory, see for example (Kang, Kim, & Kim, 2005; De Maio, Fenza, Loia, & Senatore, 2012). However, the approach presented in this paper is different from the fuzzy concept

search in existing work. We believe that the value of this research in comparison to existing research lies in the joint application of ontology matching and Fuzzy Sets, a combination that enables a searcher-friendly service that considerably decreases the search time period and expands the relevant results.

2.4. Patent retrieval

Previous workshops in Information Retrieval (IR) have targeted patent documents. During the NTCIR Workshops (Iwayama, Fujii, Kando, & Marukawa, 2006; Fujii, Iwayama, & Kando, 2004) a patent retrieval task was organized in which a test collection of patent documents was produced and used to evaluate a number of participating IR systems. In the NTCIR-3 patent retrieval task, participant groups were required to submit a list of relevant patent documents in response to a search topic consisting of a newspaper article and a supplementary description. Search topics were in four languages. All topics were initially written in Japanese and were manually translated into English, Korean, and traditional or simplified Chinese. In NTCIR-4 the search topic files were Japanese patent applications that were rejected by the Japanese patent office. The English patent abstracts were human translations of the Japanese patent abstracts. Currently, the NTCIR tasks aim at machine translation of sentences and claims from Japanese to English. Other work analyzed Japanese–English cross-language patent retrieval using Kernel Canonical Correlation Analysis (KCCA), a method of correlating linear relationships between two variables in the kernel defined by feature spaces (Li & Shawe-Taylor, 2007). Additional approach of patent classification dealt with identification of trends from patents using self-organizing maps (Segev & Kantola, 2012).

The Workshop of Cross-Language Evaluation Forum (CLEF 2009) (Roda, Tait, Piroi, & Zenz, 2010) gave separate topic sets for the language tasks, when the document language of the topics was English, German, and French. CLEF-IP included Prior Art Candidate Search task (PAC) and Classification task (CLS). Participants in the PAC task were asked to return documents in the corpus that could constitute prior art for a given topic of patents. Participants in the CLS task were given patent documents that had to be classified using the International Patent Classification codes. In addition, evaluations were performed on chemical datasets in chemical IR in general and chemical patent IR in particular. A chemical IR track in TREC (TREC-CHEM) (Lupu, Huang, Zhu, & Tait, 2009) addressed the challenges in chemical and patent IR.

Retrieval methods included language models, vector-space and probabilistic approaches, and translation resources ranged from bilingual dictionaries, parallel and comparable corpora to online MT systems and Wikipedia. Groups often used a combination of more than one resource. Although different implementations took part in the PAC and CLS tasks, the retrieval models presented a uniform approach to the translation problem. There was a very strong indication of the validity of the Google Translate function (Ferro & Peters, 2010).

Previous work analyzed automatic patent retrieval, while we describe a method that involves a manual decision process assisted by an automatic suggestion of relevant concepts related to patents. In addition, the proposed method allows concept generation and patent extraction in multiple languages without the need to translate the patent or the query.

3. Patent Knowledge Analysis model

The implementation of the model begins when the patent office user initializes the process of evaluating the patent request in his native language (Fig. 1). The model identifies the main context of the patent, a set of descriptors which are semantically related to the patent. A simple syntactic search might look for documents relating to a descriptor, such as *Length*, which appears in the text. However, the described model expands the search results to include documents related to additional descriptors, such as *Wave* in Chinese or *Distance* in Korean, that are not mentioned in the text. The patent officer can perform a query regarding a patent request. The query is the patent document itself. The query is represented by a context, a set of textual descriptors. The context of the patent is matched with ontology concepts which are also represented by sets of descriptors. Each patent is matched with concepts in the ontology based on overlap between descriptors.

3.1. Patent Knowledge Extraction

Each patent claim is analyzed separately through the Domain Representation process. To analyze the claims, a context extraction algorithm and a term frequency/inverse document frequency 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. A context comparison involves comparing the set of descriptors which represent the patent but are not limited to words appearing in the document. For each document the context is extracted by the Patent Knowledge Extraction and then compared with the ontology concept by the Patent Domain Representation.

3.1.1. Context extraction

We define a context descriptor c_i from domain $\mathcal{DOM}$ as an index term used to identify a record of information (Mooers, 1972), which in our case is a patent claim. It can consist of a word, phrase, or alphanumerical term. A weight $w_i \in \Re$ identifies the importance of descriptor c_i in relation to the patent. For example, we can have a descriptor $c_1 = Length$ and weight $w_1 = 2$. A *descriptor set* $\{\langle c_i, w_i\rangle\}_i$ is defined by a set of pairs, descriptors and weights. Each descriptor can define a different point of view of a concept. The descriptor set eventually defines all the different perspectives and their relevant weights, which identify the importance of each perspective.

By collecting all the different view points delineated by the different descriptors, we obtain the *context*. A context $\mathcal{C} = \{\{\langle c_{ij}, w_{ij}\rangle\}_i\}_j$ is a set of finite sets of descriptors, where *i* indexes each context descriptor and *j* represents the index of each set. For example, a context $\mathcal{C}$ may be sets of words (hence $\mathcal{DOM}$ is a set of all possible character combinations) defining a patent and the weights can represent the relevance of a word in a descriptor set to the patent. In classic Information Retrieval, $\langle c_{ij}, w_{ij}\rangle$ may represent the fact that the word c_{ij} is repeated w_{ij} times in the patent.

The Patent Knowledge Extraction process uses the World Wide Web as a knowledge base to extract multiple context descriptors for the textual information. This use of the World Wide Web has the following three advantages. First, use of the Internet takes advantage of an existing database that is not limited to a predefined knowledge domain. Second, the Internet can serve as an unlimited knowledge domain that is constantly updated and maintained. The noise introduced when querying the Web for specific knowledge can be overcome by analyzing large amounts of data extracted by multiple queries. Last but not least, the Web provides a perfect infrastructure for the proposed method because of its multilingual nature. The Web allows queries to be performed in one language and the results to be received in multiple languages automatically, without the need to translate.

The algorithm input is defined as a set of textual propositions representing the patent claim description. The patent claim is separated into sentences, when each sentence forms a textual proposition. The algorithm produces for each textual proposition a set of descriptors. The result of the algorithm is a context – sets of descriptor terms that are related to the propositions. The context

recognition algorithm was adapted from Segev, Leshno, and Zviran (2007) and consists of the following three steps:

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

The Web pages clustering algorithm is based on the *concise all pairs profiling* (CAPP) clustering method (Valdes-Perez & Pereira, 2000). This method approximates profiling of large classifications. It compares all classes pairwise and then minimizes the total number of features required to guarantee that each pair of classes is contrasted by at least one feature. Then each class profile is assigned its own minimized list of features, characterized by how these features differentiate the class from the other features.

The algorithm then calculates the total number of Web pages that contain the same descriptor and the sum of number of references to the descriptor in the patent. A high ranking in only one of the weights does not necessarily indicate the importance of the context descriptor. For example, a high ranking in only Web references may mean that the descriptor is important since the descriptor widely appears on the Web, but it might not be relevant to the topic of the patent.

The weights can be calculated as follows. For each descriptor, c_i, we measure how many Web pages refer to it, defined by weight w_{i1}, and how many times it is referred to in the patent, defined by weight w_{i2}. For example, *Distance* might not appear at all in the patent, but the descriptor based on clustered Web pages could refer to it twice in the patent, and a total of 235 Web pages might be referring to it. The algorithm allows having an external source, the Web, supplying additional descriptors. The descriptor's weight, w_i, can be calculated according to the following methods:

- Set all n descriptors in descending weight order according to the number of Web page references:
  $\{\langle c_i,w_{i1}\rangle_{1\leqslant i1\leqslant n-1}|w_{i1}\leqslant w_{i1+1}\}$ Current References Difference Value, $\mathcal{D}(\mathcal{R})_i=\{w_{i1+1}-w_{i1,1\leqslant i1\leqslant n-1}\}$.
- Set all n descriptors in descending weight order according to the number of appearances in the patent:
  $\{\langle c_i,w_{i2}\rangle_{1\leqslant i2\leqslant n-1}|w_{i2}\leqslant w_{i2+1}\}$ Current Appearances Difference Value, $\mathcal{D}(\mathcal{A})_i=\{w_{i2+1}-w_{i2,1\leqslant i2\leqslant n-1}\}$.
- Let $\mathcal{M}_r$ be the Maximum Value of References and $\mathcal{M}_a$ be the Maximum Value of Appearances:
  $\mathcal{M}_r=\max_i\{\mathcal{D}(\mathcal{R})_i\}$,
  $\mathcal{M}_a=\max_i\{\mathcal{D}(\mathcal{A})_i\}$.
- The combined weight which can be used for the α-cut, w_i of the number of appearances in the patent and the number of references in the Web, is calculated according to the following formula, which is based on distance between the weights:

$$w_i=\sqrt{\left(\frac{2*\mathcal{D}(\mathcal{A})_i*\mathcal{M}_r}{3*\mathcal{M}_a}\right)^2+(\mathcal{D}(\mathcal{R})_i)^2} \tag{1}$$

The weight of each context can be determined according to the number of retrieved Web references related to the concept or the number of references to the concepts in the patents. Alternatively, the weight can contribute equally to both the number of Web references and number of patent references to the concept. Another option is setting the weight as the square root of the sum of the number of Web references squared and the number of patent references squared. All four methods described above are evaluated in the experiments section.

3.1.2. Term Frequency/Inverse Document Frequency

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.

TF/IDF is a common mechanism in IR for generating a robust set of representative keyword/term descriptors from a corpus of documents, although other methods can be used for classifying text streams by keyword descriptors (Yang, Zhang, & Li, 2011). The TF/IDF method is applied here to the patent documents. By using a large enough corpus of documents, irrelevant terms are more distinct and can be thrown away with a higher confidence. To formally define TF/IDF, we start by defining $freq(t_i,\mathcal{D}_i)$ as the number of appearances of the term t_i within the document $\mathcal{D}_i$. We define the term frequency of each term t_i as:

$$tf(t_i)=\frac{freq(t_i,\mathcal{D}_i)}{|\mathcal{D}_i|} \tag{2}$$

We define $\mathcal{D}_{patent}$ to be the corpus of patent documents. The inverse document frequency is calculated as the ratio between the total number of documents and the number of documents that contain the term:

$$idf(t_i)=\log\frac{|\mathcal{D}_{patent}|}{|\{\mathcal{D}_i:t_i\in\mathcal{D}_i\}|} \tag{3}$$

The TF/IDF weight of a term, annotated as $w(t_i)$, is calculated as:

$$w(t_i)=tf(t_i)\times idf^2(t_i) \tag{4}$$

While the common implementation of TF/IDF gives equal weights to the term frequency and inverse document frequency (i.e., $w=tf\times idf$), we chose to give higher weight to the *idf* value. The reason behind this modification is to normalize the inherent bias of the *tf* measure in short documents (Robertson, 2004). Stop word filtering before the TF/IDF was found to be unnecessary in the experiments since the algorithm applies low weights to the stop words. However, additional stop word filtering can be added in the Fuzzy Logic Decision Support module for each relevant language.

3.2. Multilingual ontology domain representation

An ontology $O\equiv\langle C,R\rangle$ is a directed graph, with nodes representing a set of concepts $C=\{c_1,c_2,\ldots,c_n\}$ (things in Bunge's terminology (Bunge, 1977; Bunge, 1979)) and edges representing relationships R. We define a single concept as represented by a name and a context. 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. Each context descriptor can belong to several ontology concepts simultaneously, thus defining the relation between them according to the shared context descriptors. For example, a context descriptor $\langle Length,2\rangle$ can be shared by many ontology concepts that have length analysis as a relation, 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 relevance of the patent information to each concept is evaluated according to the weight attributed to each concept. 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 (Section 3.1.1). For example, a patent can be associated with concept 거리 (Distance) with weight 0.4 and with concept 波 (Wave) with weight 0.3 (Fig. 3).

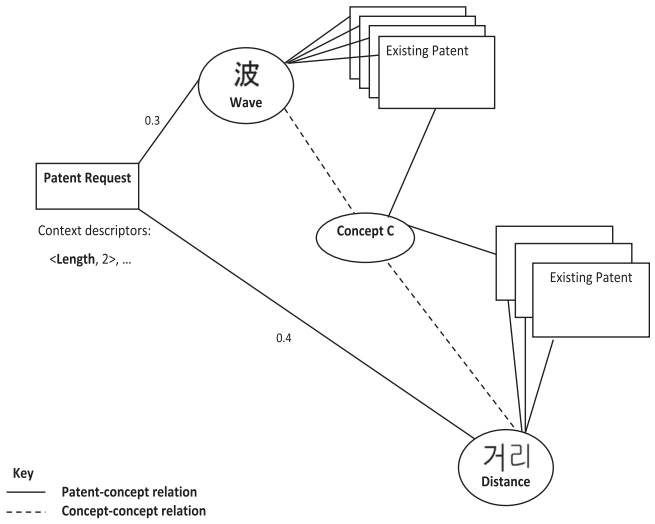

Fig. 3. Multilingual ontology domain representation.

To compute the relevance to each concept, we first define distance between two descriptors c_i and c_j with their associated weights w_i and w_j to be:

$$d(c_i, c_j) = \begin{cases} |w_i - w_j| & i = j \\ \max(w_i, w_j) & i \neq j \end{cases}$$

This distance function assigns greater importance to descriptors with larger weights, assuming that weights reflect the importance of a descriptor within a context. To define the best ranking concept in comparison with a given context we use the Hausdorff metric. Let A and B be two contexts and a and b be descriptors in A and B, respectively. Then,

$$d(a, B) = inf\{d(a, b) | b \in B\}$$
$$d(A, B) = \max\{\sup\{d(a, B) | a \in A\}, \sup\{d(b, A) | b \in B\}\}$$

The first equation provides the value of minimal distance of an element from all elements in a set. The second equation identifies the furthest elements when comparing both sets.

To expand an existing ontology with concepts represented in multiple languages, a set of documents is used for each concept to generate the context descriptor set. The documents can be in each one of the languages defined by the same concept. Another option is to use the context extraction (described in Section 3.1.1) in one language and to extract, using the Web, the related context descriptors of the concepts in multiple languages. It should be noted that the result of using the Web would include not only a direct translation of the concept but also relevant descriptors in other languages. In the analysis performed, both multiple documents in different languages and Web context extraction techniques were used to create the multilingual ontology.

3.3. Matching contexts to ontologies

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, a patent claim, with the correct concepts without the usual training process required in ontology adjustment and usually performed over a long period of time.

An ontology can be based on existing patent office classification of patent topics and relations. Alternatively, existing ontologies on specific domains can be integrated. Since each concept can be associated with multiple context descriptors, it is easy to merge existing ontologies by integrating the context descriptors. Although alternative methods of ontology merging exist (Euzenat & Shvaiko, 2007), a method based on multilingual ontology-based knowledge management (Segev & Gal, 2008), which performed well in European languages, was adopted.

To process the patent claims by mapping the contexts to existing ontologies, the following method is proposed. Let $O_1, O_2, \ldots, O_n$ be a set of ontologies, each representing different domain knowledge.

To evaluate the matching of the concepts with the patent claims context, a simple string-matching function is used, denoted by $match_{str}$, which returns 1 if two strings match and 0 otherwise. Misspelled words would have already been filtered out by the Web search engine or low TF/IDF ranking. P is defined as the patent claims, and $\mathcal{C}^P$ is the patent context descriptor set. Also, n is defined as the size of C^P.

The match between the concept c_j and the patent context descriptor set is defined as the sum of the descriptor matching values:

$$match(P, c_j) = \sum_{t_i \in C^P} match_{str}(t_i, c_j)$$

The overall match between the ontology and the patent is defined as a normalized sum of the concept matching values:

$$match(P, O_i) = \frac{1}{n} \sum_{c_j \in O_i} \sum_{t_i \in C^P} match_{str}(t_i, c_j)$$

A similar process is performed for all patents in the corpus. When a new patent request is processed, the first step involves the ontology matching process. Once the patent request is classified, the following relations with existing patents can occur:

- If the patent is related to concepts that are 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 that are 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 (Fig. 3).

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.

3.4. 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 (Aliev & Aliev, 2001). The basic scheme of fuzzy information retrieval is shown in Fig. 4 where $U1$ is a fuzzy set representing a particular query. When $U1$ is composed with Thesaurus (T), then $U2$ becomes a query augmented by associated index terms: $U2 = U1 \circ T$. $U2$ can be expressed as follows: $U2(x_i) = \max_i \min_j [U1(x_i), T(x_i, x_j)]$. 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) (Aliev & Aliev, 2001). Other implication relations can be used, but in this work we use max–min.

The role of Fuzzy Thesaurus T can be carried out by a set of ontologies that are further linked to the lexical database Wordnet (Fellbaum, 1998), [c.f. (Segev & Gal, 2007)]. In the proposed approach, the role of the fuzzy thesaurus (T) is carried out by the ontology matching process (O). The relevance of the set of concepts

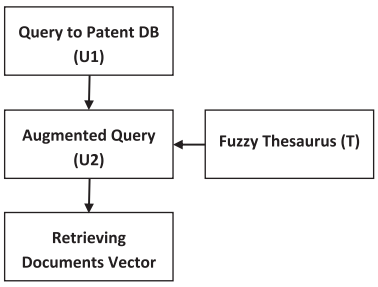

Fig. 4. Fuzzy information retrieval scheme (c.f. (Aliev & Aliev, 2001)).

and their weights to each patent supplies the fuzziness of the system. The basic scheme of fuzzy information retrieval $U2$ becomes a query augmented by associated index terms from ontology matching: $U2 = U1 \circ O$ (Fig. 5). Term operands are Fuzzy Sets as described in Section 2.3.

For the Fuzzy Logic ontology matching function, the search by string uses binary string matching (full match) and the search by degree uses the mathematical functions of the specified vague or strict Fuzzy Sets (degree of match from perfect match to no match). Vague and strict functions are displayed in Fig. 6.

Max–Min composition is used for association between concepts. The equation for matching the patent context descriptor set, $\mathcal{C}^{\mathcal{P}}$, representing the patent claims, and the concept $\mathcal{C}$ matching function is

$$match(\mathcal{C}^{\mathcal{P}}, \mathcal{C}) = min[\{\langle c_i, w_i\rangle\}_i | w_i \geqslant \mu_{strict|vague}]$$

The inquirer can inspect all the documents that have support D, or she can filter the inspection to those supported by some α-cuts (Aliev & Aliev, 2001). The search index must have full relevance to the document index. The membership functions of the Fuzzy Sets allow us to set what the response to the index is. With this we can determine the strength of the "matching response" depending on different situations. The inquirer can manually augment the patent query by setting α-cut to a lower level, which can expand the number of documents retrieved from the existing data set. For example, α-cut level 0.5 would also bring up those 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 query. Since a person finds it difficult or impossible to think of the concepts that are vaguely related to a given query, using ontology matching to augment the original query is justified.

3.5. Fuzzy Logic Decision Support

Fig. 6 shows an example of the proposed approach. Say the patent officer is examining patent claims. The user can expand the search to other possibly related concepts as well 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.

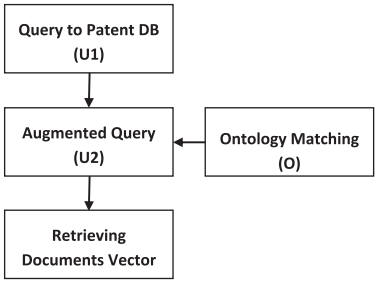

Fig. 5. Fuzzy information retrieval and ontology matching scheme.

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 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 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. Fig. 7 illustrates how the α-cuts are used to filter the new expanded set of results. For example, in Strict mode the Wave concept is part of the new expanded index set if the α-cut is set to a level of 0.15. However, the Distance concept is not part of the result set if the α-cut level is 0.48.

The patent officer can adjust the expanded search by selecting a "Strict" or "Vague" search mode and also by setting the α-cut level of the concepts (and hence the number of relevant documents retrieved) to gradually move from a Low, Medium, or High level. 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.

4. Patent 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.

The quality of a patent has two different meanings. From an economic perspective, it refers to the patent's technological value or profitability. From a legal perspective, it refers to the soundness of the decision to grant a patent and exclusion of any reasons for invalidation.

Customers have recently shown a preference for high-quality patent examinations over speedy examinations. There is also 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. KIPO has operated the IP search database since 1999 and, according to the patent technology information sharing policy, has uploaded a total of 85 patent technology databases from 21 countries and five IP offices and has continuously updated them. KIPO has also been offering them online at http://www.kipris.or.kr/since2000. 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.

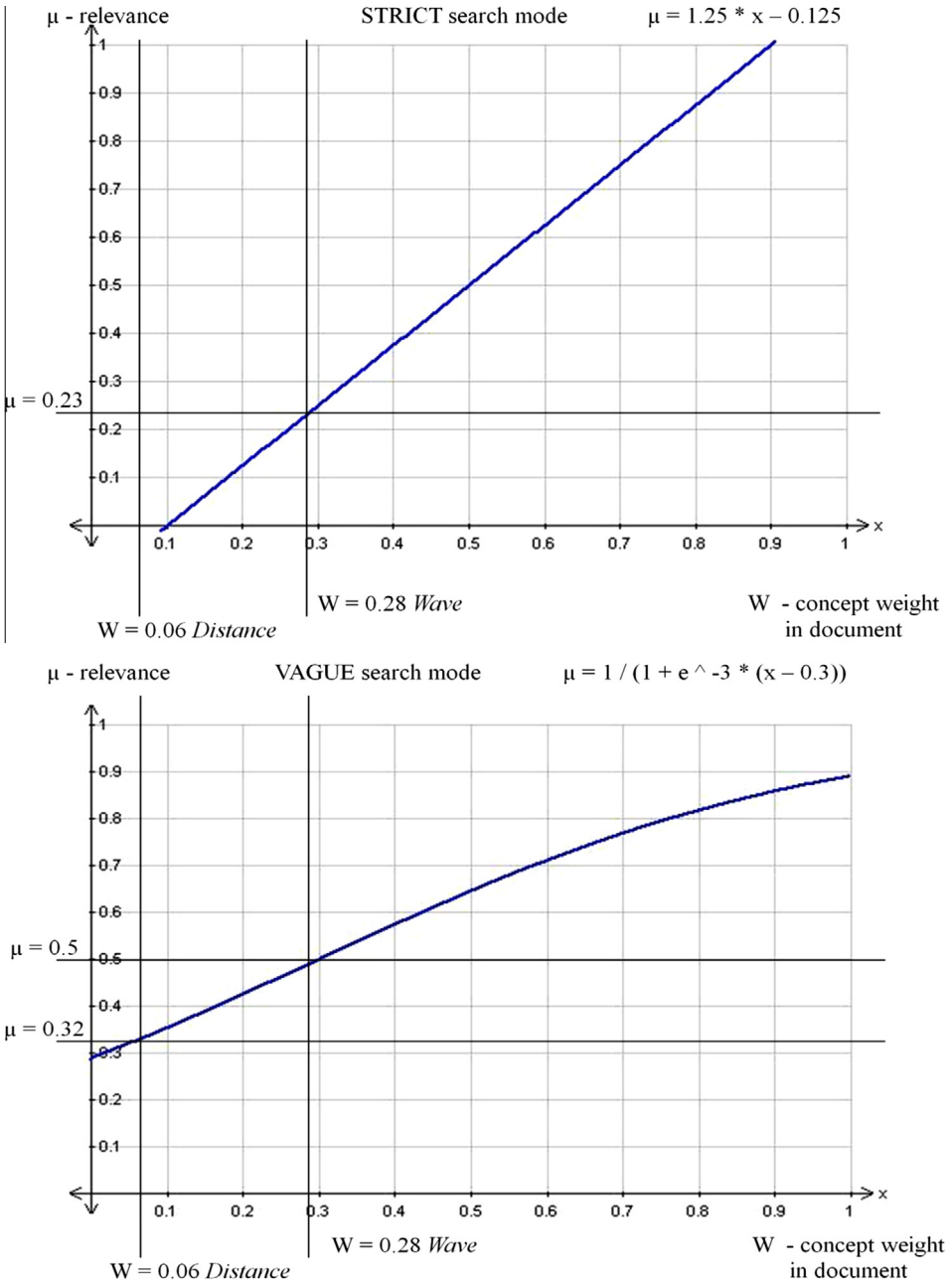

Fig. 6. The Vague and Strict membership functions.

Fig. 8 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:

Find the new patent application document by loading a "New patent" document.

1. Select Vague or Strict search mode from the radio button list.
2. Set the filter (α-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. Manually discard some general search terms, such as *map*, *design*, and *music*, by selecting those search terms and clicking them. 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. All common document types are searched. 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.
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

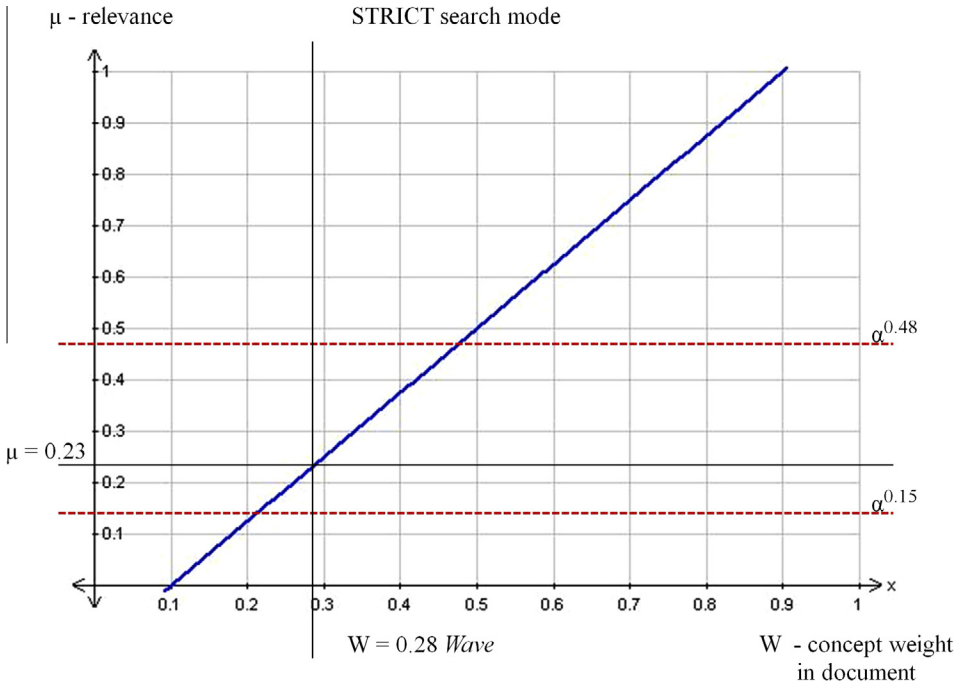

Fig. 7. The relevance of concepts.

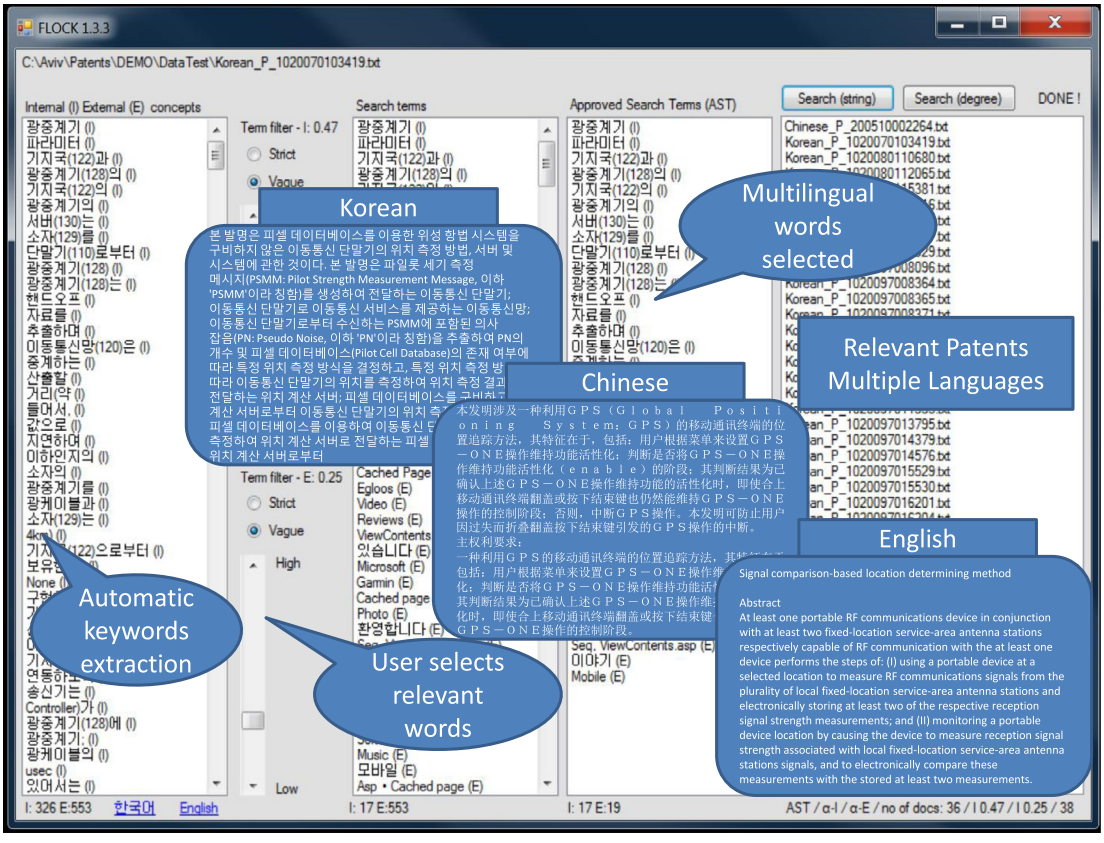

Fig. 8. The FLOCK demonstrator tested at KIPO.

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.

The FLOCK system enables the patent officers to review each patent and remove concepts that, in their experience, would minimize the number of irrelevant documents, such as *map* or *design*.

5. Experiments

The experiments analyzed the model for representing the patent request by a set of concepts related to existing knowledge in multiple languages. The search for patent knowledge 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 concepts representing the patent. Different methods are used in combination with Fuzzy Logic in the process of identifying relevant documents. Furthermore, the model was analyzed to evaluate the relevance of the patents extracted in multiple languages.

5.1. Concept relevance analysis

5.1.1. Data set and methods

The first set of experiments analyzes different methods of evaluating the relevance of the concepts. The data consists of a total of 104,296 patents extracted from the United States Patent and Trademark Office. The patent documents included free text description of the patents with no specific classification. From the patents collected, a random set of 141 patents was processed through the Patent Knowledge Extraction process as described in Section 3.1. The patents were analyzed using the Fuzzy Logic module as described in Section 3.4. The interface is based on the FLOCK system as described in Section 4.

Four different methods were used to analyze the patents extraction process. The four methods represent different classifications for determining the weight for each concept:

- Number of Web references retrieved that are related to the concept (Web).
- Number of references to the concepts in the patent (File).
- Equal weight to both the number of Web references and patent references to the concept (Web + File).
- Square root of the sum of the number of Web references squared and the number of patent references squared $\left(\sqrt{Web^2 + File^2}\right)$.

5.1.2. Experiments results

The first set of tests analyzed the relevance of the concepts extracted in relation to the number of files retrieved according to the concept weighting techniques. The results are presented in Fig. 9. The *X*-axis represents the number of files retrieved and the *Y*-axis the number of concepts in logarithmic scale. The results of all four methods display that the top ranking concepts are the most relevant, since the number of patent files retrieved decreases as the number of concepts decreases. This is especially evident for the top 10 ranking concepts.

The next set of tests analyzed how many of the top ranking concepts are relevant. The analysis is based on evaluating all 104,296 patents against all the concepts identified. As the number of top ranking concepts is decreased, the results should show a decrease in number of relevant files retrieved. This test included simple string matching of concepts, unlike the previous weighted concept file comparison used in the previous tests. Fig. 10 displays the number of concepts versus the number of files extracted. It can be clearly seen that for only up to the top 10 ranking concepts does the simple string matching extraction of all patents have additional value. The string matching allows the patent officer to limit the number of extracted results up to 11.41% of the total amount of patents in the data set according to three of the methods and 1.41% according to the Web method. The comparison of the results in Fig. 9 shows the advantage of weighted concept comparison versus simple string matching. In the weighted value extraction the use of all four methods allows the user to consistently limit the number of extracted results up to a single file.

The last set of tests analyzed the four different methods to evaluate their effect on the concepts retrieved. Fig. 11 displays the method comparison of number of concepts in relation to the α-cut. The *X*-axis presents the α-cut value and the *Y*-axis the number of concepts in logarithmic scale used for the relevant α-cut. The Web-based method of extracting concepts according to the number of appearances on Web pages declines the fastest. This means less flexibility for the patent officer who analyzes the results. The most flexible results, which allow a gradual process of extracting new concepts, are achieved by the method that calculates the square root of the squares of both methods. Another interesting issue is that the method using the number of references to the concepts in the patents yields better results than the method of just giving equal weight to both the number of Web references and the number of patent references to the concept.

The method comparison of the number of weighted files retrieved as a result of the α-cut is displayed in Fig. 12. Unlike the previous set of results, which analyzed the concepts and which showed that the methods presented different gradual declines, these results indicate that all methods seem to decline at a similar

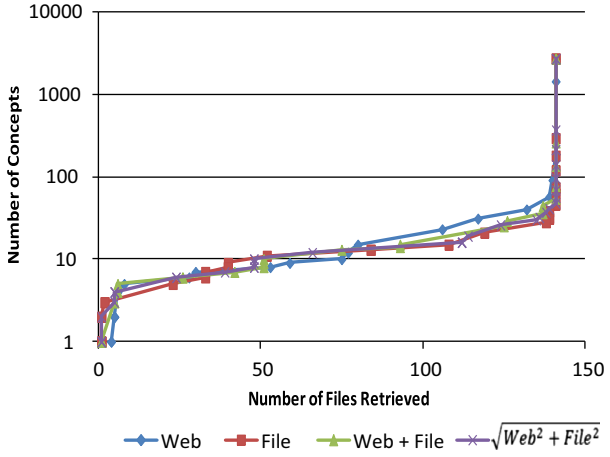

Fig. 9. Concepts relevance to number of files retrieved.

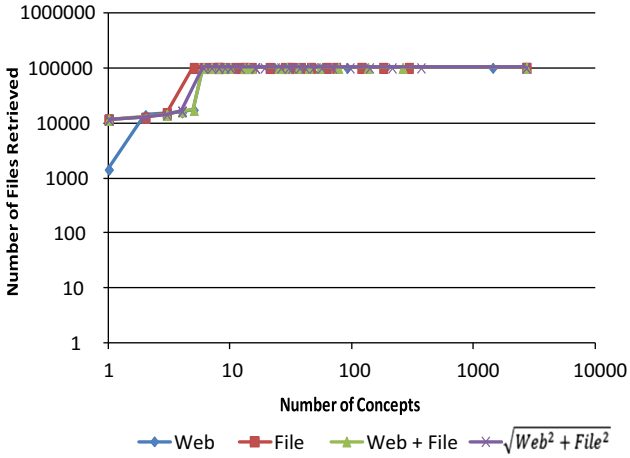

Fig. 10. Number of relevant top ranking concepts.

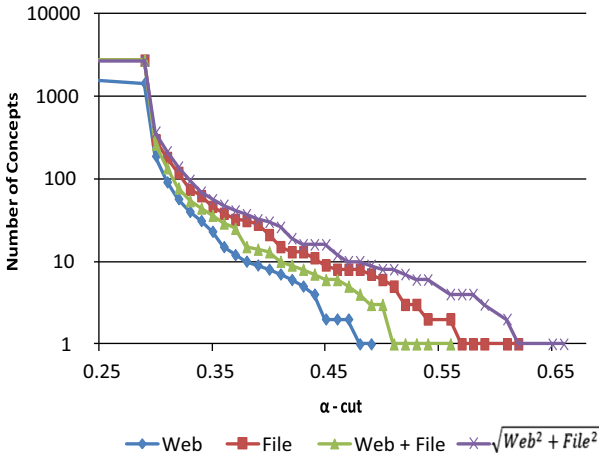

Fig. 11. Method comparison of number of concepts vs. α-cut.

slope rate. However, the beginning and ending points of decline are shifted. The decline shift appears similar to the order in Fig. 11, where the Web methods degenerate first and the method based on integrated square root of Web and File references begins and ends the decline last. The results suggest that the main emphasis in the extraction processes should be related to the number of relevant concepts.

5.2. Patent retrieval analysis

5.2.1. Data set and methods

The second set of experiments analyzed the patent retrieval performance using precision and recall. The data consist 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 Analysis model implementation. 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.

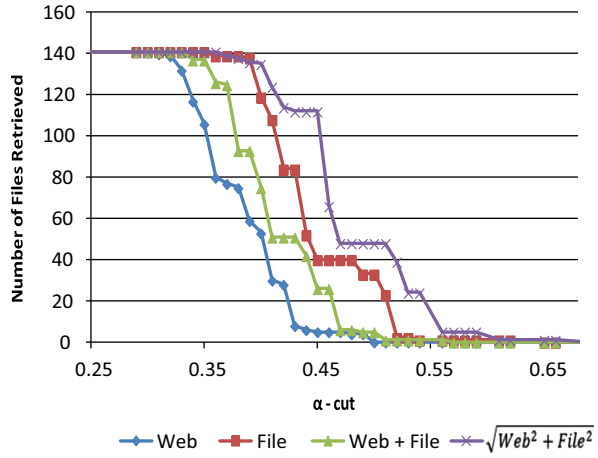

Fig. 12. Method comparison of number of weighted files vs. α-cut.

5.2.2. 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 patent according to a predefined set of 17 different α-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 Fig. 13. The results present high relevance and accuracy with precision falling below 80% only when recall reaches 65.56%.

Fig. 14 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 by the patent user can decrease the decline. Fig. 15 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. Fig. 16 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 display a minimal difference between the two graphs at any specific point. Furthermore, the increase in the number of languages did not de-

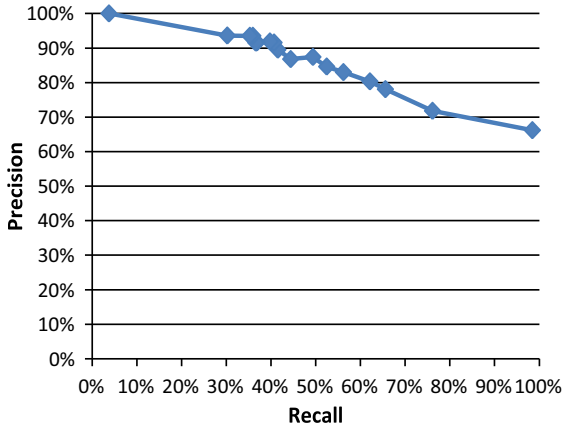

Fig. 13. Precision vs. recall average for 10 results.

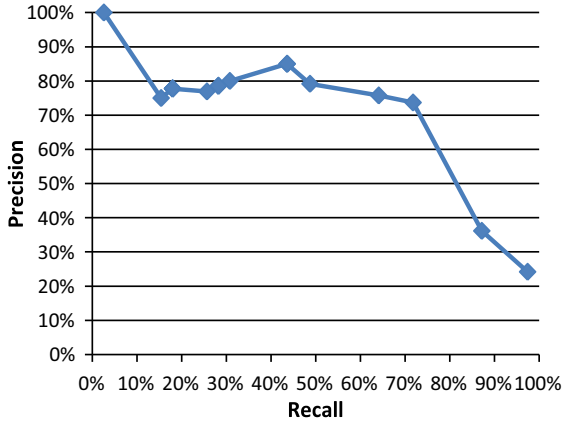

Fig. 14. Precision vs. recall – worst sample case.

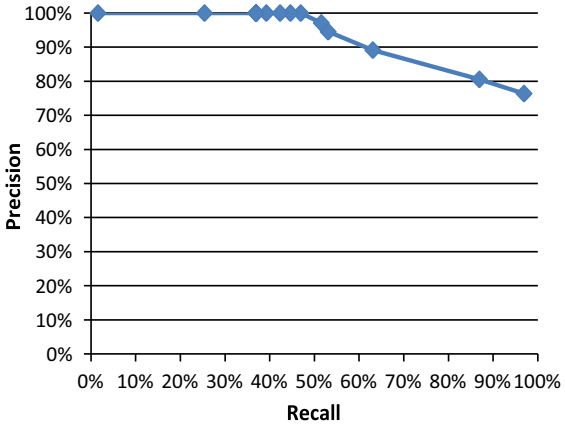

Fig. 15. Precision vs. recall – best sample case.

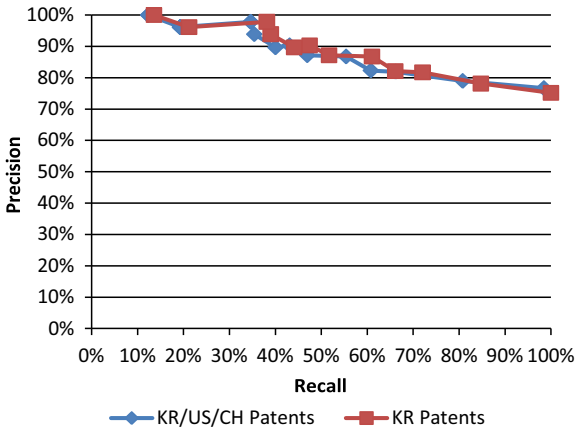

Fig. 16. Korean versus multiple languages (Korean, English, and Chinese).

crease all the values to create a similar graph shifted downward as expected. The results suggest that the increase of the number of languages used can have minor effects on the model.

5.2.3. Multilingual corpora comparison results

The second set of experiments analyzes corpora of different languages. The analysis evaluated the method dependence on different languages and how the increase in the number of languages

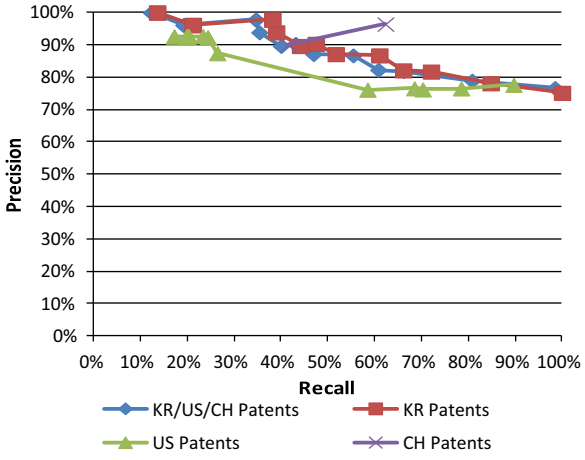

Fig. 17. Multiple languages corpus (Korean, English, and Chinese).

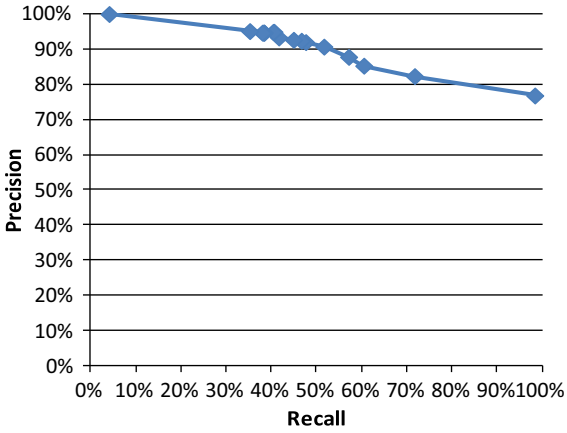

Fig. 18. Location domain.

used in the data set influences the recall and precision. Fig. 17 presents four data sets. Three data sets include patents from only one language corpus: English, Korean, and Chinese. The fourth data set includes the Korean, US, and Chinese patents as a single corpus.

The recall versus precision results display higher results for the Korean and Chinese results than for the English patents. These results point out that the English patent retrieval results actually bring down the overall average. A possible explanation could be the similarity between the languages of the Far East, which are more similar to one another than to English. Another possible explanation is that patent related documentation often includes professional vocabulary that is mostly in English. Thus, the retrieval of a document in English based on an original patent in Korean or Chinese will be more successful than will be the retrieval of a document in Korean or Chinese based on a patent in English.

The recall versus precision results display small differences between the 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.

5.2.4. Domain corpora comparison results

The experiments analyzed the precision and recall based on a different corpus based on specific domains. Patents from two different and assumed non-related domains were selected, Location and Organic Food. The average Location domain based results ap-

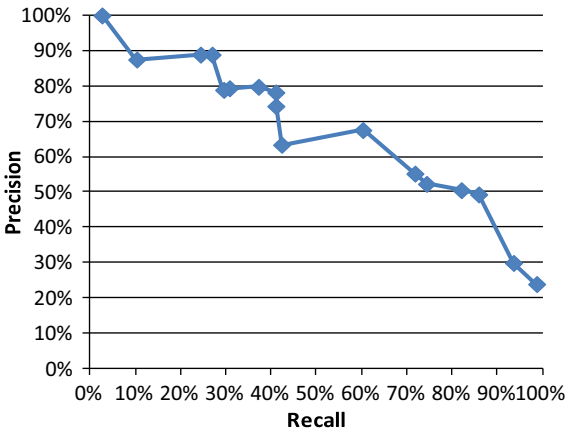

Fig. 19. Organic food domain.

pear in Fig. 18, and the average Organic Food domain based results are displayed in Fig. 19.

Although in both domains the method performs well, the comparison of both figures shows that domain dependent corpus does influence the results. Possible explanations could be dependent on the domain vocabulary where Location based patents employed a vocabulary commonly used only in this domain, while Organic Food patents can include words which are used in other domains, such as *Ruby Red*, *PureSport*, or *CakeShooters*, leading to a decrease in precision for high recall values. Manual filtering by the patent user can decrease the decline, especially if the user can identify the competing corpus that could be eliminated from the search.

6. Discussion

The technique was presented to patent examiners and managers in order to evaluate the adoption and acceptance among real users of the interactive solution offered by the system. The model, implemented as a system, was presented to both the Korean Intellectual Property Office (KIPO) and the Israeli Patent and Trademark Office. The evaluation included a presentation and collection of possible issues relating to the model.

The most critical issue was the objection of the patent examiners to their replacement by the system, although the system was presented as only a decision support system. Although this evaluation cannot be quantified, it does display the results perceived by the users. Other issues that were raised related to patent ownership, since the emphasis is on multilingual patents. The search through multilingual patents involves searching through patents in multiple countries, and this search, which requires the transfer of ownership rights between countries, is not within the scope of this paper.

The last issue raised was the complexity of the model, since the expansion of the implementation to all the patents filed at patent offices, such as KIPO, is an important issue for all the decision makers. An evaluation of processes in the model identified the response of the Vivísimo online search engine for each input as a time consuming process. To overcome this limitation, the use of parallel processing and parallel computing was analyzed. The analysis showed that parallel processing can improve the performance of processing 30 million patents to approximately 32 months on a single computer. Furthermore, the use of multiple computers in parallel will cut down time performance considerably, since there is no overhead for parsing the workload. For example, the use of 32 computers in parallel could result in processing the patents within one month.

7. Conclusion

The patent search model described in the paper allows queries to be performed on the boundaries of existing knowledge. 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 which has been used in the past (Wanner et al., 2008). The results show the advantage of query expansion in the search process, which is based on extracting relevant knowledge from the Web instead of limiting the search to concepts that appear in the patent itself. In addition, the results present the advantage of weighted concept search over the simple string search performed today. 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.

The Mandani and Assilian (1975) type of fuzzy system models has four modules: fuzzification, rulebase, inference engine, and defuzzification modules. Further research includes adding all four of these modules to the work proposed in this paper so as to design and add enhanced humanlike capabilities to patent search. Future work also includes analyzing the model in relation to the strict versus vague fuzzy search modes, as well as analyzing additional rule-based techniques of decision making. Another direction is to continue to extend the model to other languages.

References

Aliev, R. A., & Aliev, R. R. (2001). *Soft computing and its applications*. Singapore: World Scientific.

Basden, A., & Klein, H. K. (2008). New research directions for data and knowledge engineering: A philosophy of language approach. *Data and Knowledge Engineering, 67*(2), 260–285.

Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D., Patel-Schneider, P., Stein, L. (2004). OWL web ontology language reference, W3C recommendation, W3C.

Borgida, A., Brachman, R. J. (1993). Loading data into description reasoners. In *Proceedings of the 1993 ACM SIGMOD international conference on management of data* (pp. 217–226).

Bunge, M. (1977). *Treatise on basic philosophy. Ontology I: The furniture of the world* (Vol. 3). New York, NY: D. Reidel Publishing Co., Inc..

Bunge, M. (1979). *Treatise on basic philosophy. Ontology II: A world of systems* (Vol. 4). New York, NY: D. Reidel Publishing Co., Inc..

Cong, H., & Tong, L. H. (2008). Grouping of TRIZ inventive principles to facilitate automatic patent classification. *Expert Systems with Applications, 34*, 788–795.

Cross, V. (2008). Fuzzy information retrieval. *Journal of Intelligent Information Systems, 3*(1), 29–56.

Davulcu, H., Vadrevu, S., Nagarajan, S., & Ramakrishnan, I. (2003). OntoMiner: Bootstrapping and populating ontologies from domain specific web sites. *IEEE Intelligent Systems, 18*(5), 24–33.

De Maio, C., Fenza, G., Loia, V., & Senatore, S. (2012). Hierarchical web resources retrieval by exploiting fuzzy formal concept analysis. *Information Processing and Management, 48*(3), 399–418.

Doan, A., Madhavan, J., Domingos, P., & Halevy, A. (2002). Learning to map between ontologies on the semantic web. In *Proceedings of the 11th international world wide web conference (WWW'02)* (pp. 662–673). Honolulu, HI, USA: ACM Press.

Donini, F., Lenzerini, M., Nardi, D., & Schaerf, A. (1996). Reasoning in description logic. In G. Brewka (Ed.), *Principles on knowledge representation, studies in logic, languages and information* (pp. 193–238). CSLI Publications.

Euzenat, J. (2012). A modest proposal for data interlinking evaluation. In *Proceedings of the seventh international workshop on ontology matching* (*OM-2012*).

Euzenat, J., & Shvaiko, P. (2007). *Ontology matching*. Heidelberg, DE: Springer-Verlag.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA, USA: MIT Press.

Ferro, N., & Peters, C. (2010). *CLEF 2009 ad hoc track overview: TEL and Persian tasks. Lecture notes in computer science* (Vol. 6241). Springer.

Fujii, A., Iwayama, M., Kando, N. (2004). The patent retrieval task in the fourth NTCIR workshop. In *Proceedings of the SIGIR-04* (pp. 560–561).

Gal, A., Modica, G., Jamil, H., & Eyal, A. (2005). Automatic ontology matching using application semantics. *AI Magazine, 26*(1), 21–31.

Hutchins, J. (2005). Current commercial machine translation systems and computer-based translation tools: System types and their uses. *International Journal of Translation, 17*(1-2), 5–38.

Iwayama, M., Fujii, A., Kando, N., & Marukawa, Y. (2006). Evaluating patent retrieval in the third NTCIR workshop. *Information Processing and Management, 42*, 207–221.

Kang, B., Kim, D., & Kim, H. (2005). Fuzzy information retrieval indexed by concept identification. In *Text, speech and dialogue. Lecture notes in computer science* (Vol. 3658, pp. 179–186). Springer.

Kifer, M., Lausen, G., & Wu, J. (1995). Logical foundation of object-oriented and frame-based languages. *Journal of the ACM, 42*, 741–843.

Klir, J. G., & Yuan, B. (1995). *Fuzzy sets and fuzzy logic, theory and applications*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc..

Lin, C. T., & Lee, C. S. (1996). *Neural fuzzy systems: A neuro-fuzzy synergism to intelligent systems*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc..

Li, Y., & Shawe-Taylor, J. (2007). Advanced learning algorithms for cross-language patent retrieval and classification. *Information Processing and Management, 43*(5), 1183–1199.

Lucarella, D., & Morara, R. (1991). FIRST: Fuzzy information retrieval system. *Journal of Information Science, 17*(2), 81–91.

Lupu, M., Huang, J., Zhu, J., & Tait, J. (2009). TREC-CHEM: Large scale chemical information retrieval evaluation at trec. *SIGIR Forum, 43*(2), 63–70.

Madhavan, J., Bernstein, P., Rahm, E. (2001). Generic schema matching with cupid. In *Proceedings of the international conference on very large data bases (VLDB)* (pp. 49–58). Rome, Italy.

Madhavan, J., Bernstein, P., Domingos, P., Halevy, A. (2002). Representing and reasoning about mappings between domain models. In *Proceedings of the 18th national conference on artificial intelligence and fourteenth conference on innovative applications of artificial intelligence (AAAI/IAAI)* (pp. 80–86).

Maedche, A., & Staab, S. (2001). Ontology learning for the semantic web. *IEEE Intelligent Systems, 16*(2), 72–79.

Mandani, E. H., & Assilian, S. (1975). An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Man-Machine Studies, 7*, 1–13.

McGuinness, D., Fikes, R., Rice, J., Wilder, S. (2000). An environment for merging and testing large ontologies. In *Proceedings of the seventh international conference on principles of knowledge representation and reasoning* (*KR 2000*), Breckenridge, Colorado, USA.

Melnik, S. (Ed.). (2004). *Generic model management: Concepts and algorithms*. Springer-Verlag.

Miyamoto, S. (1990). Information retrieval based on fuzzy associations. *Fuzzy Sets and Systems, 38*(2), 191–205.

Mooers, C. (1972). *Encyclopedia of library and information science* (Vol. 7). Marcel Dekker. Ch. Descriptors, pp. 31–45.

Ngomo, A. N., Lehmann, J., Auer, S., Höffner, K. (2011). RAVEN – active learning of link speciffications. In *Proceedings of the Sixth International Workshop on Ontology Matching* (*OM-2011*).

Noy, F. N., Musen, M. A. (2000). PROMPT: algorithm and tool for automated ontology merging and alignment. In Proceedings of the 17th national conference on artificial intelligence (*AAAI-2000*) (pp. 450–455). Austin, TX.

Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation, 60*(5), 503–520.

Roda, G., Tait, J., Piroi, F., Zenz, V. (2010). CLEF-IP 2009: Retrieval experiments in the intellectual property domain. In *Proceedings of the 10th workshop of the cross-language evaluation forum* (*CLEF 2009*) (pp. 385–409).

Segev, A., Kantola, J. (2010). Patent search decision support service. In *Proceedings of international conference on information technology: New generations* (*ITNG 2010*) (pp. 568–573).

Segev, A., & Gal, A. (2007). Putting things in context: A topological approach to mapping contexts to ontologies. *Journal of Data Semantics (JoDS), IX*, 113–140.

Segev, A., & Gal, A. (2008). Multilingual ontology-based knowledge management. *Decision Support Systems, 45*, 567–584.

Segev, A., & Kantola, J. (2012). Identification of trends from patents using self-organizing maps. *Expert Systems with Applications, 39*, 13235–13242.

Segev, A., Leshno, M., & Zviran, M. (2007). Internet as a knowledge base for medical diagnostic assistance. *Expert Systems with Applications, 33*(1), 251–255.

Spyns, P., Meersman, R., & Jarrar, M. (2002). Data modelling versus ontology engineering. *ACM SIGMOD Record, 31*(4), 12–17.

Valdes-Perez, R. E., & Pereira, F. (2000). Concise, intelligible, and approximate profiling of multiple classes. *International Journal of Human-Computer Studies*, 411–436.

Vickery, B. (1966). Faceted classification schemes, Graduate School of Library Service, Rutgers, The State University, New Brunswick, NJ.

Vossen, P. (1999). EuroWordNet general document, LE2-4003 LE4-8328, EuroWordNet.

W3C OWL working group. (2009). OWL 2 web ontology language: Document overview, W3C recommendation, W3C.

Wanner, L., Baeza-Yatesa, R., Brügmann, S., Codina, J., Diallo, B., Escorsa, E., et al. (2008). Towards content-oriented patent document processing. *World Patent Information, 30*(1), 21–33.

Wei, C. P., Yang, C. C., & Lin, C. M. (2008). A latent semantic indexing-based approach to multilingual document clustering. *Decision Support Systems, 45*(3), 606–620.

Yang, B., Zhang, Y., & Li, X. (2011). Classifying text streams by keywords using classifier ensemble. *Data and Knowledge Engineering, 70*(9), 775–793.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8*, 338–353.

Zadeh, L. A. (1973). Outline of a new approach to the analysis of complex systems and decision processes. *IEEE Transactions on Systems, Man, and Cybernetics, 1*(1), 28–44.

Zadeh, L. A. (1983). Commonsense knowledge representation based on fuzzy logic. *Computer, 16*, 61–65.