An ontology-based approach to learnable focused crawling

Hai-Tao Zheng, Bo-Yeong Kang, Hong-Gee Kim

Biomedical Knowledge Engineering Laboratory, Seoul National University, 28 Yeongeon-dong, Jongro-gu, Seoul, Republic of Korea

ABSTRACT

Focused crawling is aimed at selectively seeking out pages that are relevant to a predefined set of topics. Since an ontology is a well-formed knowledge representation, ontology-based focused crawling approaches have come into research. However, since these approaches utilize manually predefined concept weights to calculate the relevance scores of web pages, it is difficult to acquire the optimal concept weights to maintain a stable harvest rate during the crawling process. To address this issue, we proposed a learnable focused crawling framework based on ontology. An ANN (artificial neural network) was constructed using a domain-specific ontology and applied to the classification of web pages. Experimental results show that our approach outperforms the breadth-first search crawling approach, the simple keyword-based crawling approach, the ANN-based focused crawling approach, and the focused crawling approach that uses only a domain-specific ontology.

1. Introduction

Due to the tremendous amount of information on the web, it is increasingly difficult to search for useful information. For this reason, it is important to develop document discovery mechanisms based on intelligent techniques such as focused crawling [6,8–10,12,14,17,18,21–23,28,30,32] and personalized search [2,3,11,13,19]. In a classical sense, crawling is one of the basic techniques for building data storages. Focused crawling goes a step further than the classical approach. It is proposed to seek out selectively pages relevant to a predefined set of crawling topics.

One of the main issues of focused crawling is how to traverse effectively off-topic areas and maintain a high harvest rate during the crawling process. To address this issue, an effective strategy is to apply background knowledge of crawling topics. Since an ontology [25] is defined as a well-organized knowledge scheme that represents high-level background knowledge with concepts and relations, ontology-based focused crawling approaches have come into research [10,18,32]. Such approaches use not only keywords, but also an ontology to evaluate the relevance between web pages and crawling topics.

In the ontology-based focused crawling approaches, it is difficult to acquire the optimal concept weights to maintain a stable harvest rate during the crawling process. This is because concept weights are heuristically predefined before being applied to calculate the relevance scores of web pages. To address this issue, a learnable focused crawling approach based on ontology is proposed in this paper. Since an ANN (artificial neural network) is particularly useful for solving problems that cannot be expressed as a series of steps—such as recognizing patterns, classifying into groups, series prediction and data mining—it is suitable to apply to focused crawling. In our approach, an ANN is constructed by using a domain-specific ontology and applied to the classification of web pages.

The outline of this paper is as follows: Section 2 reviews the related work to provide an overview of focused crawling. Section 3 describes the framework of our learnable focused crawling approach and elaborates the mechanism of relevance.

10.1016/j.ins.2008.07.030
2. Related work

Focused crawlers aim at searching only the subset of the web related to a specific category. There are numerous methods that have been proposed for focused crawling [6,8–10,12,14,17,18,21–23,28,30,32]. A generic architecture for focused crawling was proposed by Chakrabarti et al. [8]. This architecture included a classifier that evaluates the relevance of a web page with respect to the crawling topics and a distiller that identifies web nodes with high access points to many relevant pages within their links. Diligenti et al. [9] proposed a focused crawling algorithm that constructs a “context graph” model for the contexts in which topically relevant pages occur on the web. This context graph model could capture typical link hierarchies in which valuable pages occur, and model content on documents that frequently occur with relevant pages. In [14], focused crawling using a context graph was improved by constructing a “relevancy context graph”, which can estimate the distance and the relevancy degree between the retrieved document and the given topic. Gao et al. [12] proposed a focused collaborative crawling method to realize geographically focused crawling using a group of crawling nodes. A domain search engine, MedicoPort [6], used a topical crawler, Lokman crawler, to build and update a collection of documents relevant to the health and medical field.

To help a crawler to traverse hyperlinks in an intelligent fashion, various machine learning techniques have been employed [17,21,28,30]. Rennie and McCallum [28] proposed a machine learning oriented approach to focused crawling: the Q-learning algorithm was used to guide the crawler to pass through off-topic documents to highly relevant documents. McCallum et al. [21] also used a reinforcement learning algorithm to develop an efficient crawler for building domain-specific search engines. Shokouhi et al. [30] presented an intelligent crawler called Gcrawler that used a genetic algorithm to estimate the best path for crawling and expanded its initial keywords during the crawling process. Liu et al. [17] proposed a new approach for prediction of the links leading to relevant pages based on a hidden Markov model (HMM).

However, most of the above approaches do not consider applying background knowledge. Background knowledge, such as the interests of users, or domain knowledge, is important for understanding situations and problems in information management systems. Since an ontology is a description of the concepts and relationships that can well represent background knowledge, it has been applied to various information management systems for improving their performance. For example, Preschner and Gauch [26] explored ontologies to represent users’ interests to improve search results. Buccafurri et al. [5] proposed an agent-based recommender system based on a new model (concept-graph) that represented user-behavior-dependent relationships among concepts. Rosaci [29] also exploited ontologies and ANNs to represent the interests and behavior of users, in order to realize mutual monitoring for the improvement of recommendation quality in autonomous multi-agent systems. Differing from these approaches’ use of ontologies to address the interests of users for personalized search, the focus of our method is to employ ontologies representing domain knowledge in order to locate relevant documents during the crawling process.

In the field of focused crawling, many approaches have applied background knowledge to improve their effectiveness [10,18,22,23,32]. In Inforspiders [22,23], a population of agents searched for pages relevant to the topic, using evolving query vectors and ANNs to decide which links to follow. The relevance feedback of users was used as background knowledge to improve the performance of ANNs. One limitation of the Inforspiders crawlers is that they require the users’ intensive labor to acquire the relevance feedback.

In contrast to Inforspiders’ use of relevance feedback, a number of focused crawling approaches have employed ontologies to represent relevant domain knowledge. Su et al. [32] presented an intelligent focused crawler algorithm in which an ontology was embedded to evaluate a page’s relevance to the topic. Another prevalent focused crawling approach based on ontologies is called ontology-focused crawling [10,18]. Fig. 1 depicts the ontology-focused crawling framework. In this approach, the crawler exploits the web’s hyperlink structure to retrieve new pages by traversing links from previously retrieved ones. To judge whether or not one page is relevant to the specific topic, a domain-specific ontology is used to compute the relevance score. After preprocessing, entities (words occurring in the ontology) are extracted from the visited pages and counted. Using the entity distance, the linking steps between an entity in the ontology and the crawling topic, concept weights are calculated by a heuristically predefined discount factor raised to the power of the entity distance. The relevance score is the summation of concept weights multiplied by the frequencies of corresponding entities in the visited web pages. With this relevance score, a candidate list of web pages is maintained in the priority queue in order of increasing priority. Based on the priority queue, the crawler can crawl the relevant web pages and store them in the topic-specific web page repository.

Most existing ontology-focused crawling approaches use an ontology as background knowledge and apply the weights of concepts in the ontology to compute the relevance score. Since the concept weights are heuristically predefined without using machine learning techniques, the relevance score can not optimally reflect the relevance of concepts to the crawling topics during the whole crawling process. Our method also applies an ontology as background knowledge for focused crawling, but additionally uses an ANN to classify the visited web pages by quantifying the concept weights based on a set of training examples. The idea of combining ontology and an ANN for focused crawling has not been researched in detail until now.
3. Ontology-based learnable focused crawling

In this section, we propose the framework for our learnable focused crawling approach. Based on this framework, we describe each step of the crawling process in detail. Then, the mechanism of relevance computation is elaborated by an example.

3.1. Framework

The proposed learnable focused crawling approach consists of three stages with distinct functions. Fig. 2 depicts these three stages: the data preparation stage, the training stage, and the crawling stage. The data preparation stage is responsible for preparing training examples that are used for the ANN construction. In the training stage, an ANN is trained by the training examples. In the crawling stage, web pages are visited and the ANN determines whether or not they will be downloaded. Finally, the downloaded web pages are stored in a topic-specific web page repository.

3.1.1. Data preparation stage

In the data preparation stage, we construct a set of training examples by crawling web pages. The WebLech [33] crawler is employed to crawl web pages. The WebLech crawler is a breadth-first crawler that supports many features required to download web pages and emulates standard web-browser behavior as much as possible. Given a seed URL, the WebLech crawler will crawl all the web pages near this URL without bias. After crawling a collection of web pages, we manually determine the relevance of the web pages to the crawling topics. As a result, a set of training examples is composed of positive training examples and negative training examples.
3.1.2. Training stage

All the training examples constructed in the first stage are fed to the training stage for training an ANN. In the training stage, the crawler selects a set of concepts from a given domain-specific ontology to represent the background knowledge of crawling topics. These concepts are selected because they are considered to have close relationships with the crawling topics. These selected concepts are called relevant concepts. Each web page in the training examples is preprocessed to get a list of entities (words occurring in the set of relevant concepts).

A preprocessing module includes functions for parsing web pages from html to text, performing POS (part-of-speech) tagging, stop word removal, and word stemming. HTML Parser [15] is employed to process the html by removing all the html tags. QTagger [27] is used to do the POS tagging. A lexical tool called LexAccess [16] is used for the word stemming. LexAccess is designed to help users access information from the SPECIALIST Lexicon [31], which is a database of lexicographic information for use in natural language processing. After preprocessing the visited web pages, the entities (words occurring in the set of relevant concepts) are extracted, and from them, the term frequencies of relevant concepts are calculated.

After calculating the term frequencies of relevant concepts, we use them to train an ANN. The ANN is initialized as a feedforward neural network with three layers (Fig. 3). The first layer is composed of a linear layer with transfer function \( y = \frac{1}{1 + e^{-x}} \). There are four nodes in the hidden layer. Considering the ANN will be used for binary classification, the output layer is also composed of a sigmoid layer. The ANN outputs are the relevance scores of corresponding web pages. We apply the well-known backpropagation algorithm to train the ANN. The training process will not stop until the root mean squared error (RMSE) is smaller than 0.01. A more detailed description of the backpropagation algorithm can be found in [24].

3.1.3. Crawling stage

The crawling stage is concerned with the actual crawling and the use of the constructed ANN for classification. First, our crawler retrieves the robots.txt information for the host either from the metadata store or directly from the host, and determines whether or not it is allowed to crawl the page based on this specification. If the URL passes the check, our crawler begins to crawl web pages and judge their relevance to the crawling topics. Given a domain-specific ontology, a set of concepts, which are called relevant concepts, is selected to represent the background knowledge of crawling topics. When the crawler visits web pages, the term frequencies of relevant concepts are calculated after preprocessing. Then the term frequencies of relevant concepts are acquired as inputs for the ANN, which computes relevance scores for the web pages based on the crawling topics. The mechanism of relevance computation will be described in detail in the next section. If the web pages are classified as relevant to the crawling topics, they are downloaded and saved into the topic-specific web page repository.

3.2. Mechanism of relevance computation

The most important component within the framework is the relevance computation component in the crawling stage. First, a domain-specific ontology is given as a background knowledge base.

**Definition 3.1 (Domain-Specific Ontology).** Ontology \( O = \{ C, P, I, \text{is-a, inst, prop} \} \), with a set of classes \( C \), a set of properties \( P \), a set of instances \( I \), a class function is-a: \( C \rightarrow C \) (is-a(\( C_1 \)) = \( C_2 \) means that \( C_1 \) is a sub-class of \( C_2 \)), a class instantiation function...
inst: $C \rightarrow 2^I$ (inst($C$) = $I$ means $I$ is an instance of $C$), a relation function prop: $P \rightarrow C \times C$ (prop($P$) = ($C_1$, $C_2$) means property $P$ has domain $C_1$ and range $C_2$).

Given the domain-specific ontology, relevant concepts are selected from the knowledge base to calculate the relevance of the web pages. The selection of relevant concepts is based on the distance between those concepts and the crawling topics.

**Definition 3.2 (Distance between Concepts).** Distance between concepts $d(t, c_i) = k$, where $k \in \mathbb{N}$ being the number of links between $t$ and $c_i$ in ontology $O$; $t$ being the crawling topic contained in ontology $O$; $c_i$ being concepts in ontology $O$, $i \in \{1, \ldots, n\}$, $n$ being the number of concepts of ontology $O$.

The distance of concepts reflects the pertinence between concepts in the ontology and the crawling topic. If $d(t, c_i)$ is too large, the concept $c_i$ will have low relativity with the crawling topic $t$. Moreover, the number of concepts that need to be selected from the ontology will increase exponentially as the distance between concepts grows. Because of that, the time complexity of crawling will increase. To maintain a high crawling efficiency, we need to perform experiments to choose a suitable distance $d(t, c_i)$.

According to the distance between concepts, given a crawling topic $t$, a set of relevant concepts $c_i$ will be selected from the domain-specific ontology. When visiting web pages, the preprocessing module parses the web pages from html into text, performs POS (part-of-speech) tagging, removes stop words, and implements word stemming. After getting a list of entities (words occurring in the relevant concepts) from the visited web pages, the term frequencies of relevant concepts are calculated based on their occurrence in the visited web pages. The term frequencies are used as inputs for the ANN, which calculates the relevance score of each visited web page and decide whether or not to download the web page.

**Fig. 4** shows an example of the relevance computation process. The relevance is calculated in several steps, starting with preprocessing such as word stemming and word counting, followed by ANN classification and concluding with a relevance judgment.

The crawling topic in the example is “Cell”. The ontology used here is the UMLS (unified medical language system) ontology (**Fig. 5**) [4]. UMLS integrates over 2 million names for some 900 000 concepts from more than 60 families of biomedical vocabularies, as well as 12 million relations among these concepts. The concepts relevant to the crawling topics are measured in terms of the distance of the concepts, i.e., $d(t, c_i) \leq 3$. Based on the distance of concepts, our crawler selects 38 concepts in the UMLS ontology that are relevant to the crawling topic—“Cell” and apply them to construct the ANN with training examples.

Note that some of the class concepts in the UMLS ontology are multi-word concepts such as “Cellular metabolism”. The concepts are firstly decomposed into single words. Then the lexical tool, LexAccess [16], is used to transform the non-noun words into their noun forms, which are all counted in the visited web pages. Finally, the term frequencies of the multi-word concepts are the sum of the term frequencies of all of the corresponding single words. For example, the multi-word concept “Cellular metabolism” is decomposed into the term “cellular” and the term “metabolism”. Using LexAccess, these two words are transformed into their noun forms such as “cell” and “metabolism”. The term frequencies of both “cell” and “metabolism” are added together as the term frequency for the concept “Cellular metabolism”.

After preprocessing the visited web pages, entities contained in relevant concepts such as “Cell” and “Tissue”, are searched for in the visited web pages. For each entity found, the corresponding relevant concept term frequency is counted. In this example, the concept “Gene” has term frequency 6, the concept “Tissue” has term frequency 5, the concept “Cell” has term frequency 13, the concept “Anatomical” has term frequency 1 and so on. The concept “Component” has term frequency

![Fig. 4. An example of relevance computation.](image-url)
0, which means there is no such entity in the visited web page. All these calculated term frequencies are used as input for the ANN, which calculates the relevance score of the corresponding web pages. Since our example result is a positive relevance judgement, with the relevance score being 1, the crawler decides to download the page.

4. Experimental results

In the experiment, we compared our approach with the standard breadth-first search crawling approach, the focused crawling approach with simple keyword spotting, the ANN-based focused crawling approach, and the ontology-focused crawling approach [10].

The ANN-based focused crawling approach is a crawling approach that utilizes an ANN to classify the visited web pages without considering domain-specific ontologies. The training examples constructed in our approach are also used to train an ANN. All the terms in the visited web pages are extracted and counted. The term frequencies of these terms are used as inputs for the ANN, which calculates the relevance score in order to crawl web pages.

Our experiment focuses on two crawling topics: “Cell” and “Mammal”. Because the UMLS ontology [4] (Fig. 5) contains abundant medical information relevant to the topics “Cell” and “Mammal”, we use it as the background knowledge for this focused crawling. For evaluating our approach, we refer to the biological definition of “Cell” [7] and “Mammal” [20] in Wikipedia to judge whether or not the web pages are relevant to the crawling topics.

The rates at which relevant pages are effectively filtered are the most crucial evaluation criteria for focused crawling. Aggarwal et al. [1] and Chakrabarti et al. [8] provides a well-established harvest rate metric for our evaluation

\[
hr = \frac{sr}{sp}, \quad hr \in [0, 1].
\]

The harvest rate represents the fraction of web pages crawled that satisfy the crawling target \(sr\) among the crawled pages \(sp\). If the harvest ratio is high, it means the focused crawler can crawl the relevant web pages effectively; otherwise, it means the focused crawler spends a lot of time eliminating irrelevant pages, and it may be better to use another crawler instead. Hence, a high harvest rate is a sign of a good crawling run.

Another performance metric we provide is the performance-cost ratio, which evaluates the performance of focused crawling based on the time cost. This metric can show clearly the efficiency of the crawling approach. In our experiment,
we use the number of topic hits and the computation time as the performance-cost ratio metric. The experiment environment is Windows XP Professional with Intel Pentium IV 3.00 GHz processor, 2.0 GB memory, and Java SE 5.

For the crawling topic “Cell”, the URL “http://www.biology.arizona.edu/” was used as the seed for crawling web pages for the manual assessment in the data preparation stage. Hundred training examples were filtered to train an ANN with 64 positive examples and 36 negative examples. For the crawling topic “Mammal”, the URL “http://www.enchantedlearning.com/” was used as the seed for crawling web pages for the manual assessment in the data preparation stage. Hundred and twenty training examples were filtered to train an ANN with 73 positive examples and 47 negative examples.

In the training stage, first, we need to decide the number of relevant concepts by experiment. Based on the difference value of distances $d(t,c_i)$, the relevant concepts were selected from the UMLS ontology. In the training examples, the term frequencies of those relevant concepts were counted and applied to construct the ANNs.

Figs. 6 and 7 illustrate the performance of our approach by using corresponding ANNs. For the topic “Cell” (Fig. 6), within 20 min, we can see that the learnable focused crawling process with distance $d(t,c_i) \leq 3$ outperforms the crawling processes with other distances. For the topic “Mammal” (Fig. 7), we can see that the learnable focused crawling process with the distance $d(t,c_i) \leq 4$ outperforms the crawling process with other distances. When the distance between concepts is too small, the set of relevant concepts is too small. In this case, it is difficult to crawl the relevant web pages effectively. When the distance between concepts is too large, the set of relevant concepts increases exponentially. The processing time of each web page becomes too long, which causes the number of topic hits to decrease during a crawling period. Therefore, by experiment, we choose the distance $d(t,c_i) \leq 3$ to get the relevant concepts for crawling topic “Cell” and the distance $d(t,c_i) \leq 4$ to get the relevant concepts for crawling topic “Mammal”.

For the crawling topic “Cell”, with the distance $d(t,c_i) \leq 3$, 38 relevant concepts $c_i$ were selected from the UMLS ontology, such as “Tissue” and “Gene”. In the crawling stage, we started crawling from the seed URL: “http://web.jjay.cuny.edu/acarpi/
Also, the other four crawling approaches were applied to crawl web pages with the same seed URL. After crawling 1150 web pages, we obtained the harvest rates of corresponding crawling approaches shown in Fig. 8. For the crawling topic “Mammal”, with distance \( d(t,c_i) \leq 4 \), 56 relevant concepts \( c_i \) were selected from the UMLS ontology such as “Vertebrate”, “Whale”, and so on. We applied the five crawling approaches to start crawling with the seed URL: “http://www.ucmp.berkeley.edu/mammal/mammal.html”. After crawling 980 web pages, we obtained the harvest rates shown in Fig. 9. The experimental results show our approach and the ANN-based crawling approach have higher harvest rates than the other three crawling approaches.

Table 1 shows the average harvest rate of topic “Cell” and topic “Mammal”. When the crawling topic is “Cell”, our method’s average harvest rate is 0.2044 higher than the ontology-focused crawling, 0.3237 higher than the keyword-based crawling, 0.582 higher than the breadth-first crawling, and 0.0193 higher than the ANN-based focused crawling. For the crawling topic “Mammal”, our method’s average harvest rate is 0.1323 higher than the ontology-focused crawling, 0.2574 higher than...
the keyword-based crawling, 0.4963 higher than the breadth-first crawling, and 0.0291 higher than the ANN-based focused crawling.

From the experimental results based on the harvest rate metric, we can see all the crawling processes begin with high harvest rates, but decrease as more pages are crawled. The main reason is that the seed URL has high relevance to the crawling topics. When more pages are crawled, the ratio of irrelevant web pages gets higher, and the harvest rates decrease. In the crawling process, the harvest rate of the breadth-first crawling is lowest because it does not consider the crawling topics. The keyword-based crawling also performs low because it only considers the crawling topic during the crawling process.

The ontology-focused crawling and our approach have better harvest rates than breadth-first crawling and keyword-based crawling because they relate the crawling topics to the background knowledge base in order to filter out irrelevant web pages. Compared to the ontology-focused crawling, our approach has better harvest rates. This is because the ontology-focused crawling calculates the relevance score based on heuristically predefined concept weights. These predefined concept weights are determined by humans subjectively. It is difficult to scale optimally the contributions of those relevant concepts in real-time. In our approach, concept weights are determined objectively by the constructed ANN to describe the contribution of each relevant concept. Therefore, the harvest rate of our approach is higher throughout the entire crawling process.

The ANN-based focused crawling approach also has a good performance because it applies the trained ANN to recognize the relevant web pages. Compared to the ANN-based focused crawling approach, our approach utilizes the domain-specific ontology to get the relevant concepts for crawling. Since the dimension of relevant concepts in our approach is smaller than the dimension of all the terms contained in visited web pages used by the ANN-based crawling approach, it means that our approach reduces the dimension of terms required to maintain a high crawling performance comparable to the ANN-based focused crawling approach.

For another aspect of the five crawling approaches, we carried out experiments based on performance-cost ratio. After 20 min, we obtained the number of topic hits of each of the crawling approaches (Figs. 10 and 11). The experimental results based on the performance-cost ratio metric show that our approach and the ontology-focused crawling approach have a higher number of topic hits than the other three crawling approaches throughout the first 20 min. The reason that the breadth-first crawling and keyword-based crawling perform lower than the other three approaches is that they do not consider the information related to the crawling topics. This causes these two crawling approaches to crawl more irrelevant web pages than other approaches. The ANN-based crawling approach has lower performance because it needs to gather all the terms in the visited web pages and calculate their term frequencies, which needs more time to calculate the relevant scores of visited web pages than other approaches. Throughout the crawling period, the number of topic hits for ontology-focused crawling is higher than our approach. This is because our approach employs an ANN to calculate the relevance scores and classify the web pages. The processing time of each page in our approach is longer than that of the ontology-focused crawling approach.

To conclude, although the ontology-focused crawling approach has better performance than our approach based on performance-cost ratio, our approach can maintain a higher harvest rate for the crawling topics "Cell" and "Mammal" than other approaches. It means that our approach can crawl relevant web pages more effectively than other approaches throughout the whole crawling process. Moreover, our approach has lower term dimension than the ANN-based crawling approach. Therefore, our approach can crawl more relevant web pages than the ANN-based crawling approach during a crawling period.

Fig. 10. Number of topic hits during 20 min when the crawling topic is "Cell".
5. Conclusion and future work

In order to maintain a high harvest rate during the crawling process, it is important to apply background knowledge to focused crawling. Moreover, it is also important to quantify the concept weights for focused crawling. To address these issues, we proposed an ontology-based approach to learnable focused crawling. Based on a domain-specific ontology, an ANN was trained by a set of training examples. The ANN was used to calculate the relevance score of visited web pages. The empirical evaluation shows that our approach outperforms the ontology-focused crawling, breadth-first crawling, keyword-based crawling, and the ANN-based focused crawling. Based on the experiment results, we believe that our approach improves the efficiency of focused crawling, which can be employed to build a comprehensive data collection for a given domain.

One limitation of our method is that its performance depends on which ontology is used. The comprehensiveness of the ontology, and how closely related it is to the crawling topics are significant factors in the crawling results. Specifically, the proposed system should yield good results if the ontology is well constructed to thoroughly represent the background knowledge of the crawling topics.

We will conduct further research to improve our work in the following ways. First, we will explore an effective way of utilizing the hyperlinks in the visited web pages to traverse the irrelevant web pages more efficiently. Second, more NLP (natural language processing) methods will be studied in order to help the crawler better identify the intrinsic meaning of multi-word concepts found in web pages. Finally, the crawling results for a variety of domain- and task-specific ontologies can be compared, particularly in the biomedical domain. Good examples are FMA (the foundational model of anatomy), which is known to be ontologically well designed, GO (gene ontology), which is most popularly used for biomedical research, and SNOMED CT (systematized nomenclature of medicine – clinical terms), which is a practical clinical ontology used by many hospitals.

Acknowledgements

This work was supported by ministry for health, welfare and family affairs, South Korea [A05-0909-A80405-06A2-15010A, Ontology-based EHR (electronic health record) Interoperability Technology Project]. We thank Charles Borchert for his helpful English revision. We also thank the anonymous reviewers whose comments and suggestions led us improve significantly the paper.

References
