World Journal of Computer Application and Technology 2(5): 104-113, 2014 http://www.hrpub.org
DOI: 10.13189/wjcat.2014.020502

Knowledge Reduction Information Retrieval Model in Pathology Medical Domain

Changwoo Yoon

Electronics and Telecommunication Research Institute, Daejeon, Korea
*Corresponding Author: cwyoon@etri.re.kr

Abstract We present an efficient intelligent information retrieval model using reduction of domain-specific expert knowledge, demonstrating its use in the pathology medical domain.

We created an information retrieval model that incorporates domain-specific knowledge to provide knowledgeable answers to users. This model converts domain-specific knowledge to a relationship of terms represented as quantitative values, which gives improved efficiency. The conversion technology, called "knowledge reduction," enables the off-line calculation of knowledge separate from the information retrieval (IR) process. This results in the real-time processing of retrieval results.

We performed a simulation of the developed Intelligent IR model in the Pathology medical domain. Our approach resulted in an approximately 30% performance gain measured by average precision at 11 standard recall levels metrics when compared with the vector space model based IR method.

Keywords Information Retrieval, Intelligent Information Retrieval, Knowledge Representation, Vector Space Model, Bayesian Network

1. Introduction

The advent of the Internet has led to vast quantities of publicly available information (i.e., open-domain data [1]) that continue to grow at an exponential rate. Nowadays the data amounts are so large and complex, it becomes difficult to process those properly [2,3]. To assist in the location and retrieval of this information, several information retrieval (IR) engines using well-known IR methods (e.g., Boolean, probabilistic, or vector space model) [4,5] were developed and have earned a commercial success.

Contrary to open-domain data, many institutions maintain their own secretly managed data, i.e., closed-domain data [1]. For example, in a medical field, each hospital maintains its own patient information containing patient reports, personal disease history, and billing information.

Closed-domain data have their own descriptive language consisting of a term dictionary and relations that exist between terms. Examples of such languages are the medical field's Unified Medical Language System (UMLS) and Systematized Nomenclature of Medicine (SNOMED) [6,7], which we call domain specific knowledge.

Except the conventional information retrieval methods such as Boolean model, and classical vector space IR model, if we can use domain specific knowledge, we can use Intelligent IR models. The examples are query expansion by using a thesaurus [8-11], a term relationship measurement like Latent Semantic Indexing (LSI) [12], and a probabilistic inference engine using Bayesian Network [13,14]. Since a figure is often worth a thousand words, other efforts examined the visualization of information [15].

The nature of closed-domain data allows us to use better semantics than that of general-domain data [14]. It is a better approach in terms of knowledge management if we can use the closed-domain knowledge in developing an Intelligent IR model. Applying knowledge in the information retrieval process normally requires significant computation. This computation occurs when the intelligent information retrieval system tries to search the knowledge space during the retrieval process.

The objective of this research is to create an intelligent information retrieval model that uses a computationally efficient method to produce effective results reflecting knowledge. We used a semantic network and a Bayesian network to express the closed-domain specific knowledge in the Pathology medical domain's SNOMED.

We used the Pathology domain as our target closed-domain IR area. The data are Pathology patient reports describing specimens, their diagnoses, and retrieval and charge specification codes.

Medical language is extremely rich, varied, and difficult to comprehend and standardize, with vagueness and imprecision. This has resulted in the development of the Unified Medical Language System (UMLS) and Systematized Nomenclature of Medicine (SNOMED) [10,11]. In this research, we used SNOMED II as our knowledge base. SNOMED II consists of five main axes beginning with a hierarchical listing of anatomical systems

(Topography). Any change in form of those structures throughout life is characterized in the (Morphology) axis. Causes or etiologies for those changes are listed in the (Etiology) axis. All human functions, normal and abnormal, are listed in the (Function) axis. Combinations of Topography, Morphology, Etiology, and Function may constitute a disease entity or syndrome and are classified in the (Disease) axis. Using the T, M, E, F, and D axes it is possible to code nearly all-anatomic and physiologic features of a disease process. See Figure 1.

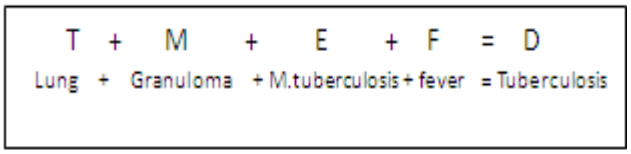


Figure 1. "Equation" of SNOMED disease axes

In this paper, I summarize the result of previous paper [25], and suggest future possible usage of knowledge reduction.

2. Overview of Intelligent IR Model

Figure 2 shows the architecture of our knowledge-based information retrieval model. The edges of this diagram represent procedures or actions taken in processing the information represented by the nodes, which represent data or subsystems. The overall operation of the Intelligent IR model is as follows.

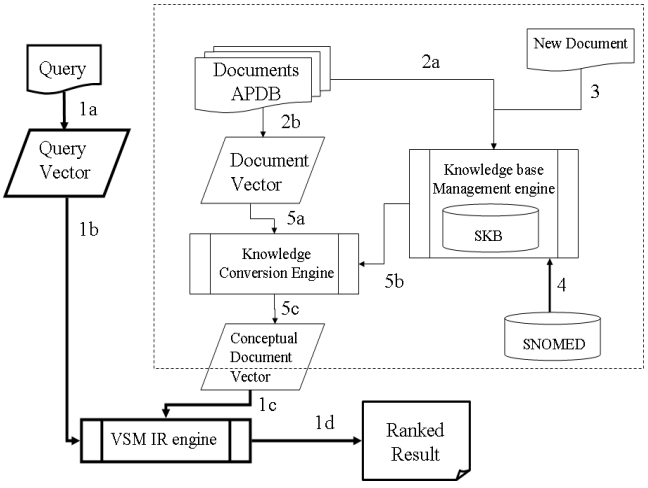


Figure 2. Knowledge-based information retrieval model

The bold edge processes (1a, 1b, 1c, 1d) are on-line processes, while the other edges (2a, 2b, 3, 4, 5a, 5b, 5c) are off-line processes completed before processing any user's query shown inside the dashed line box. We compute the conceptual document vector in off-line processes from documents (Pathology patient report) and expert knowledge (SNOMED). We can say that the conceptual document vector contains expert knowledge. The off-line process is a one-time calculation or a regularly calculated process introducing new documents to the knowledge base. After we obtain the conceptual document vector, the IR user can use the Intelligent IR machine using keywords as a query. We call these processes (1a, 1b, 1c, 1d) on-line processes because these are the only processes used during the retrieval.

The off-line processes consist of two subsystems: the Knowledge-based Management Engine (KME) and the Knowledge Conversion Engine (KCE).

The KME manages knowledge representation converted from the knowledge sources: documents (2a, 3) and SNOMED (4). For this domain the knowledge base is named the SNOMED Knowledge Base (SKB). There are two types of knowledge about the Pathology domain: the first is known knowledge which is written in SNOMED and used daily by Pathologist (4) while the second is an unknown knowledge that can be found by investigating patient reports (2a, 3). Section 3 discusses the details of KME including knowledge conversion methods from SNOMED and documents to SKB.

The KCE makes a conceptual document vector (5c) from the document vector (5a) and KME's SKB (5b). The documents used in the pathology domain are pathology reports called Anatomic Pathology (AP), which we preprocessed into a database, the Anatomic Pathology Database (APDB). The Document Vector is produced (2b) from the APDB. Section 4 explains the details of KCE.

Periodically, the KCE updates the Conceptual Document Vector to reduce the computational needs rather than updating the Conceptual Document Vector every time a new document is added.

Two key features distinguish this knowledge-based IR model from conventional models. First, while other models perform knowledge level information retrieval tasks such as ontology comparison and ontological query expansion [8,15], this model reduces the knowledge level represented by the knowledge base to a statistical model such as the vector space model's document vector (5a, 5b, 5c). We used semantic networks and naïve Bayes model for knowledge representation. These graphical knowledge representations are human friendly and easily understandable by human, but they are computationally complex. The reduced statistical form of knowledge, such as a conceptual document vector, is not human friendly but is computer friendly and computationally efficient.

And second, unlike other knowledge-based IR models, which have a heavy computation requirement because they compare concepts between the IR model and the query when the user requests information [14,15], this model uses the off-line application of knowledge to the document vector leaving only a similarity measurement calculation between the query and the documents shown in dotted box Figure 2. Only the conceptual document vector, which is obtained from the document vector and the knowledge base, is involved in the on-line process of producing ranked results by comparing a user's query and the documents.

3. Knowledge Base Management Engine (KME)

The knowledge base for this Intelligent IR system is based on the Systematized Nomenclature of Medicine (SNOMED) and consists of pre-coordinated knowledge and post-coordinated knowledge. The pre-coordinated knowledge is the expert knowledge described in SNOMED that the Pathologist uses in writing and understanding a patient's report. The post-coordinated knowledge is a special form of knowledge that can be obtained from a patient's report. This is augmentable knowledge that can be found from parsing the document and introducing new data.

The main objective of KME is expressing and storing closed-domain expert knowledge (Pathologist's pre- and post-coordinated knowledge) into an ontological representation that forms the SNOMED Knowledge-base (SKB).

3.1. Ontology of Pre-Coordinated SNOMED Knowledge

SNOMED is a detailed and specific coded vocabulary of names and descriptions designed for use in computerized patient records. We can classify the term-to-term relationships, which are called the "pre-coordinated relationship" in SNOMED, as one of three types. See Figure 3.

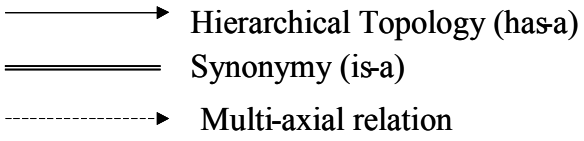


Figure 3. The three types of SNOMED term relation

The first type is a hierarchical topology. The SNOMED terms are all arranged in a hierarchy, represented by an alphanumeric code where each digit represents a specific location in the hierarchy. Figure 4 illustrates the hierarchical structure of this knowledge modeled as a semantic network. Arcs expressing the "part of" or "has-a" relation connect the nodes of this network

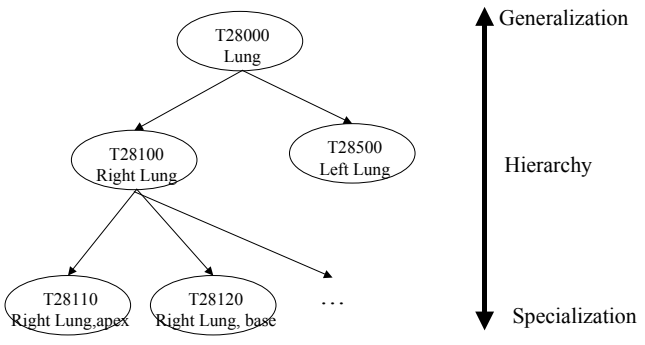


Figure 4. SNOMED hierarchical term relationship

SNOMED has controlled vocabulary characteristics, which allows individuals to record data in a patient's record using a variety of synonyms, where each references a primary concept. For example, "bacterial infectious disease", "bacterial sepsis", and "bacterial infection" are classified as symptoms of "disease caused by bacteria" with each carrying the same term code as shown in the semantic network of Figure 5. This relationship of synonyms is an "is-a" relationship where the synonym relation is explicit to each node with no propagation among the nodes.

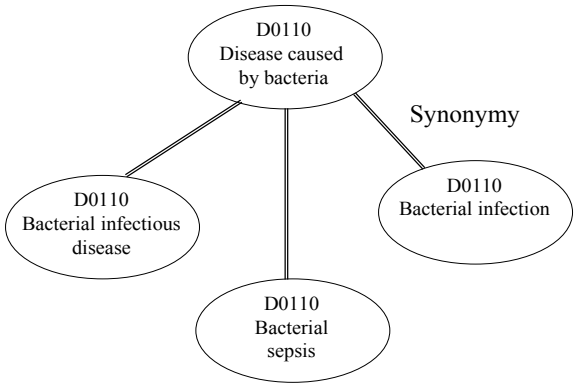


Figure 5. SNOMED synonyms relationship

The third relationship of SNOMED terms is a multi-axial relation shown in Figure 6, which refers to the ability of the ordered set of names to express the meaning of a concept across several axes. We can find examples of this relationship over all axes with it most apparent in the disease axis. Fava beans umask G6PD resulting in Favism. The SNOMED D code representing "Favism" has an information link to the E code representing "Fava bean" and F code "G-6PD". This relationship is pre-coded, mirroring the knowledge encoded at the time of SNOMED's standardization.

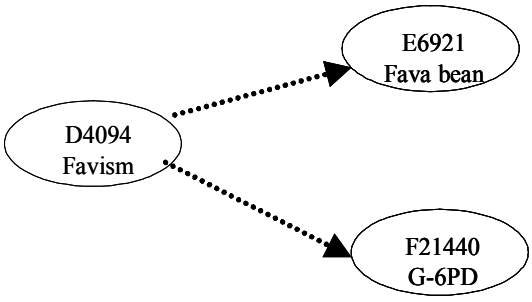


Figure 6. SNOMED Multiaxial relationship

The pre-coordinated knowledge is known and verified knowledge by a Pathologist containing no ambiguity.

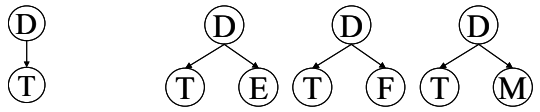
3.2. Characteristics of the Post-Coordinated SNOMED Knowledge

Contrary to the fixed or known knowledge, there is knowledge with partial truth – what is called

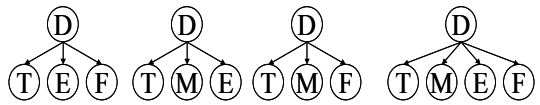
post-coordinated knowledge in pathology. Example of post-coordinated knowledge is a SNOMED equation. SNOMED consists of five categories: Topography, Morphology, Etiology, Function, and Disease. A patient report has <retrieval code> terms showing matching SNOMED categories coded as a numbers – what is called the *SNOMED equation* shown in figure 1.

We examined the Anatomic Pathology (AP) data sets from 1983 to 1994. From these documents, we extracted the SNOMED equations. The extracted SNOMED equation could be complete or not because the coding can be influenced by a physician's experience and error involved during a coding.

We can extract statistical data among SNOMED codes representing the certainty of a relationship. From the set of SNOMED codes in each document, we can extract post-coordinated knowledge. Because of the uncertainty, the pathologist does not know or describe the SNOMED equation exactly resulting in a partial description of knowledge. We only count the description of SNOMED code as post-coordinated knowledge if it contains the "D" axis. If the pathologist described SNOMED code includes "D", there is an acceptable certainty that a SNOMED equation exists. Figure 7 shows the four kinds of SNOMED equations found in the document space.



(a) Two axis relation (b) Three axis relationships



(c) Four axis relationships (d) Five axis relationship

Figure 7. Classification of post-coordinated knowledge

We can use the Bayesian classifier [16] in medical diagnostics to find the probable disease from the given symptoms. The concept that combines the Bayes theorem and the conditional independence hypothesis is proposed by several names [17,18]. The naïve Bayes (NB) approach [19] is the simplest form of classifier based on Bayesian networks.

Table 1. Statistics on post-coordinated knowledge

Post-coordinated knowledge relations	Number of unique relations
D-T	568
D-T-E	26
D-T-F	38
D-T-M	7,425
D-T-E-F	3
D-T-M-E	305
D-T-M-F	534
D-T-M-E-F	68

Table 1 shows the amount of post-coordinated knowledge found in the document space. We use this knowledge to induce possible diseases from incomplete SNOMED equations.

3.3. Naïve Bayes Model of Post-Coordinated Knowledge

It is possible to create or learn a Bayesian network from the data. We can estimate the probability density (that is, a joint probability distribution) from the data. Four different cases can result when we learn a Bayes network from the data: structure known or unknown, and all variables observable or some unobservable. For our case, the structure is known and some variables are unobservable.

To model "*post*-coordinated *knowledge*", we have several assumptions:

- We consider only the knowledge consisting of a SNOMED equation. Figure 7.(d) shows the basic architecture of a SNOMED equation expressed using a Bayesian network.
- We assume we have complete knowledge before processing a patient's report. This can be obtained from searching the complete SNOMED equations from the document's space. We call this complete knowledge "*post-coordinated knowledge"*.
- The "*post-coordinated knowledge*" consists of combinations of the five axes with the disease axis being mandatory.
- Complete knowledge is unique.
- Each disease is independent.
- The four axes (T, M, E, and F) are independent of each other.
- T, M, E, and F are conditionally dependent upon the instantiation of D.

In our case, the structure of the resulting Bayes network is fixed, having one of the forms shown in Figure 7. We can consider the knowledge is complete only if the disease axis in the SNOMED equation (i.e., in the document) exists. We use the following algorithm to extract the knowledge.

- Look through the documents to find a SNOMED equation in the document having one of the complete post-coordinated knowledge forms shown in Figure 7.
- Extract only the complete knowledge form from the documents retrieved.
- Use an expert to verify that the extracted knowledge is correct. Generally, we can consider the equation to be complete if it contains a "*D*" axis.
- Add the extracted and verified knowledge into the system's knowledge within the "Post-coordinated knowledge base" (PCKB) which is a subset of SKB.

It is possible that an individual document can contain incomplete knowledge due to a lack of expert knowledge or an error. This means some variables of the *"Post-coordinated knowledge base" (PCKB)* are not observable in some documents. In this case, we must induce the value of the unobserved variables in the complete PCKB.

To do this, we need to estimate the probability values of the PCKB structure's variables.

It is easier to start by estimating *P(D)*. This is computed by counting how many times *D* is true (i.e., found positive) in the data set (documents) and dividing by n, the total numbers of documents. To obtain an estimate of the probability that *T* is true given that *D* is true, we just count the number of cases in which *T* and *D* are both true, and divide by the number of cases in which *D* is true. The probability of *T* given not *D* is similarly shown below.

$$P(D) \approx \frac{\#(D = true)}{n}$$

$$P(\sim D) \approx 1 - P(D)$$

$$P(T \mid D) \approx \frac{\#(T = true \cap D = true)}{\#(D = true)}$$

$$P(T \mid \sim D) \approx \frac{\#(T = true \cap D = false)}{\#(D = false)}$$

Each PCKB has the probability estimations shown in Figure 8.

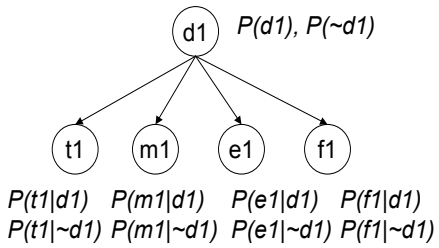


Figure 8. PCKB component structure and probability estimation

4. Knowledge Conversion Engine

The Knowledge Conversion Engine (KCE) converts the Vector Space Model (VSM) document vector to a conceptual document vector reflecting the knowledge of the SNOMED Knowledge Base (SKB).

The best-known model in information retrieval is the Vector Space Model (VSM) [20]. In the VSM, documents and queries reside in a vector space. In this space, each document can be represented as a linear combination of term vectors. A document vector is defined as equation (1) where *w* is weighting scheme.

$$\overrightarrow{d_j} = (w_{1,j}, w_{2,j}, ..., w_{t,j})^T . \quad (1)$$

The VSM document vector uses the term frequency $tf_{i,j}$ and inverse document frequency $idf_{i,j}$ as conceptual imbuement to the information retrieval model

In the Vector Space Model, term vectors are pair-wise orthogonal, implies that terms are assumed to be independent. There was an attempt to incorporate term dependencies, which gives semantically rich retrieval results [21, p. 239]. They used a term context vector to reflect the influence of terms in the conceptual description of other terms. The term context vector is defined as (2).

$$T = \begin{pmatrix} c_{11} & c_{21} & \cdots & c_{n1} \\ c_{12} & c_{22} & \cdots & c_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ c_{1n} & c_{2n} & \cdots & c_{nn} \end{pmatrix} \quad (2)$$

The c_{ik} represents the influence of term t_k on term t_i [21, p239].

In this research we used the definition of term context vector above to convert document vector into conceptual document vector.

The Knowledge Conversion Engine (KCE) converts relationships among index terms within the SKB into a term context vector. In the following, we discuss how the elements of matrix T can be obtained from the domain-specific knowledge base representation, the SKB.

4.1. KCE: Knowledge Reduction

There are two types of knowledge to convert: *pre-coordinated* and *post-coordinated knowledge.* We reduce the dimension of the knowledge of the *pre- and post-coordinated* knowledge to a conceptual document vector. The form of knowledge expressed by a graph (in our case, a semantic network) is a human friendly form, but it is computationally complex. We convert that knowledge into a computer friendly and efficient statistical form. The concept of knowledge reduction is shown in Figure 9.

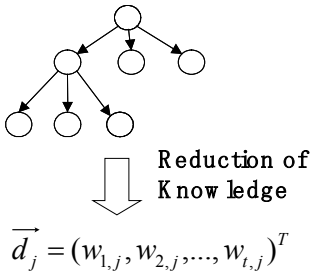


Figure 9. Knowledge reductions

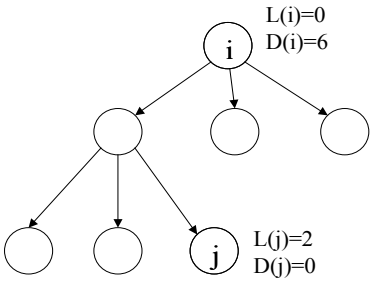


Figure 10. Attributes of the SNN-KB hierarchical topology relation

4.2. KCE: Conversion of Pre-Coordinated Knowledge

Three types of relationships exist within the SKB model representing SNOMED. In the first type, the hierarchical topology relationship, each node has attributes denoting its characteristics on the hierarchical tree. See Figure 10.

L(i) is the level of term i in a knowledge tree. D(i) is the number of descendents of term node i in the tree. The term influence between i and j is inversely proportional to the distance, which is the difference of the levels. Having many descendents means that a node is a more general term than some node having a smaller number of descendents, so term influence is inversely proportional to the number of descendents. Thus, we can calculate the SNN-KB hierarchical topology relationship between the two terms i and j as:

$$c_{ij} = C(Sht) \times \frac{1}{d(i,j)} \times \log \frac{1}{D(i) + D(j)} \quad (3)$$

where *C(Sht)* is the coefficient for the SNOMED hierarchical topology relation, *d(i,j)* is a difference of level between node i and j, *D(i)* is the number of descendents of node i and *D(j)* is the number of descents of node j.

For the synonym relations: the SNN-KB synonym relationship between the two terms i and j is defined as

$$c_{ij} = C(Ss) \quad (4)$$

where *C(Ss)* is the coefficient for the SNOMED synonym relationship.

For the multi-axial relations: the SNN-KB multi-axial relationship between the two terms i and j is defined as

$$c_{ij} = C(Sm) \quad (5)$$

where *C(Sm)* is the coefficient for the SNOMED multi-axial relationship.

We used 1 as a value of all coefficients in simulation.

4.3. KCE: Conversion of the Naïve Bayes Model of Post-Coordinated Knowledge

We defined the naïve Bayes model of the post-coordinated knowledge in Section 3. After processing documents for post-coordinated knowledge (PCK), we have n documents and m PCKs.

Each PCK has a specific form shown in Figure 7. The objective of inference in the knowledge-based information retrieval model is to find a disease from the given findings (combinations of T, M, E, and F). Each document does not contain a complete PCK normally. Because of the lack of expert knowledge, it is impossible to write a complete form of the PCK in a patient's report, so we must estimate what kind of disease is most likely from the given findings in the document. This is the key to improving the knowledge enhancement of the retrieval process.

We modeled the PCKs using naïve Bayes as shown in Section 3. We can define the posterior probability that we are attempting to calculate as:

$$P(D \mid t, m, e, f),$$

where *D* is the set of diseases that has a relationship with the given findings (t, m, e, and f) found by searching the PCKs. The posterior probability can be solved by Bayes theorem:

$$P(D \mid t, m, e, f) = \frac{P(D)P(t, m, e, f \mid D)}{p(t, m, e, f)}.$$

In practice, we are only interested in the numerator of above fraction, since the denominator does not depend on D and the values of the *t, m, e, and f* that are given, so the denominator is constant.

By the independence assumption, we can rewrite the fraction as:

$$P(D \mid t, m, e, f) = \frac{1}{Z} P(D) \prod_{i=1}^{n} P(F_i \mid D)$$

where F_i is the set of findings.

The post-coordinated knowledge has specific relations with the individual documents. Actually, the individual knowledge is defined from the specific contents of each document, so we cannot use knowledge reduction in this case. Knowledge reduction handles the general knowledge conversion cases. So, we have to apply the post-coordinated knowledge to each document: more specifically to each individual document vector.

We can classify several cases for conversion of post-coordinated knowledge. Refer to Figure 7 for the classification of post-coordinated knowledge. We use PCKB-a for a *one axis relation*, PCKB-b for a *two axes relations,* PCKB-c for a *three axes relation*, and PCKB-d for a *four axes relationship*.

Case **1:** The *document contains all four axes: t, m, e, and f.*

We must find the probability of *d* based upon the existence of *(t,m,e,f).* This is performed by searching PCKB-d. Searching PCKB-a, PCKB-b, or PCKB-c is not necessary because they have less information. We can obtain only one component of knowledge from PCKB because with the five axes of information, the knowledge is complete and unique. If we find different diseases d1 and d2 from same combination of symptom axis, we have to calculate the probability of d1 and d2.

Case 2: *The* document *contains three axes: all except d.*

Figure 11 shows an example of this case. Here, we must compute the probability of each possible diseases, then another axis's, i.e.,

$$P(d1 \mid t, m, e) \quad (6)$$

$$P(d2 \mid t, m, e) \quad (7)$$

after finding the possible post-coordinate knowledge from PCKB-c and PCKB-d.

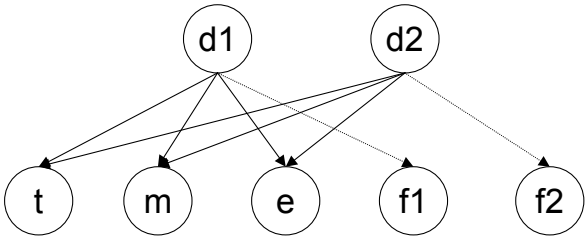


Figure 11. Examples of case 2

We already know P(d1), P(d2), P(t|d1), P(m|d1), P(e|d1),

P(f1|d1), P(t|d2), P(m|d2), P(e|d2), and P(f2|d2). By the naïve Bayes theorem, the posterior probability (6) and 7) can be calculated and compared by:

$$P(d1 \mid t,m,e) = \frac{1}{Z} P(d1) \prod_{i=1}^{n} P(F_i \mid d1) = \alpha P(d1) P(t \mid d1) P(m \mid d1) P(e \mid d1)$$

$$P(d2 \mid t,m,e) = \frac{1}{Z} P(d2) \prod_{i=1}^{n} P(F_i \mid d2) = \alpha P(d2) P(t \mid d2) P(m \mid d2) P(e \mid d2)$$

.

Then, we can augment the document vector according to the relative normalized value of *P(d1|t, m, e)* and *P(d2|t, m, e)* with some coefficient. The complexity of this algorithm, is O(mn) where n is the number of documents and m is a count of the post-coordinate knowledge.

Case 3 is the case when two axes relations are found in a document and case 4 is the case when one axis relation is found in a document. These calculations are as straightforward as case 2's. Figure 12 shows an algorithm to calculate the naïve Bayes form of the SNOMED post-coordinated knowledge.

- For each document di
  - If it contains partial knowledge (without d axis), extract partial axes EP_{ax}, where ax is the number of axes
  - Repeat from the lower class of the post-coordinated knowledge PK_{ix}, where j is class & k is the k^{th} pre-defined knowledge
    - If (ax <= j) search pattern of EP_{ax} from PK_j
    - Calculate $P(d_i \mid EP_{ax})$ using the naïve Bayes equation for all matched PK's, store to the candidate knowledge sets CK.
  - Apply relative values of CK's disease probability and unseen axes into the document vector.

Figure 12. Algorithm of naïve Bayes knowledge calculation

4.4. KCE: Generating the Conceptual Document Vector

By converting the SNOMED knowledge and domain-specific knowledge to the term-relation matrix T defined in (2), we can transform each initial document vector into a conceptual document vector using the equation defined in [21, p. 240]. We are applying the methodology proposed in the paper [21] to convert the term-relation matrix T into conceptual document vector.

The division of the elements in the term context vectors by the length of the vector is a normalization step.

After converting the document vector to the conceptual document vector, the similarity measure between the query vector ***q*** and the conceptual document vector $\boldsymbol{cd_i}$ produces a ranked list of relevant documents related to the query.

5. Performance Evaluation

In our experiment, we used differences between our algorithm and vector space model of average precision at 11 standard recall levels to evaluate the performance.

To calculate precision and recall, we must know the exact relationship between each document and the query. We selected 2000 case documents signed by a top expert, because these documents should have a low error rate in describing post-coordinate knowledge. Then, we selected 261 cases randomly among the 2000 cases to reduce the size of set to be able to examine relevancy by a human expert. The selected 261 cases were examined for their relevancy with queries "membranous nephropathy lupus" and "nephrotic syndrome". Our expert rated the relevancy between each document and the query as "Positive", "Neutral", or "Negative".

In this section, we call the query "membranous nephropathy lupus" as Q1 and "nephrotic syndrome" as Q2. Table 2 shows the result of evaluation for the 261 documents.

Table 2. Relevancy check result of 261 simulation documents

Query	# of positive	# of neutral	# of negative	Total relevant (pos.+neut.)
Q1	24	95	142	119
Q2	23	90	148	113

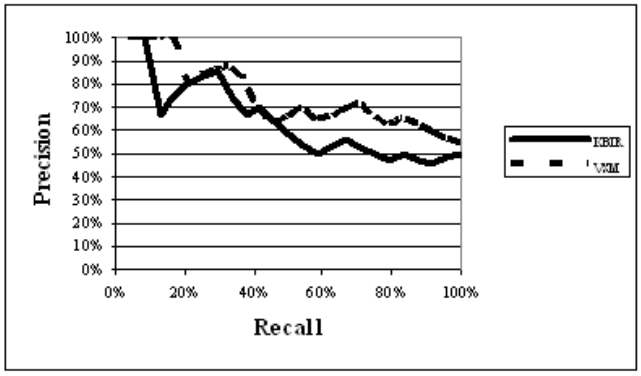


Figure 13. Comparison of performance for query1 on positive cases

5.1. Performance Evaluation with Pre-Coordinated Knowledge

Figure 13 shows the result of the query "membranous nephropathy lupus" on the positive cases. This graph shows some degradation of performance for the knowledge based information retrieval (KBIR) model compared with the vector space model (VSM). We can think of the KBIR having the same effect as query expansion. The KBIR expands the document vector instead of the query vector. If the knowledge has synonyms, the KBIR expands the document vector to include synonyms of the query "membranous nephropathy lupus". This causes an expansion to a somewhat broader range of knowledge. For example, "membranous" can be expanded to a more general term, so the degradation on the positive case may be caused by a general expansion of the knowledge of KBIR. This can be explained more by looking at the results of query 1 if we

included the neutral cases in the performance evaluation as shown in Figure 14.

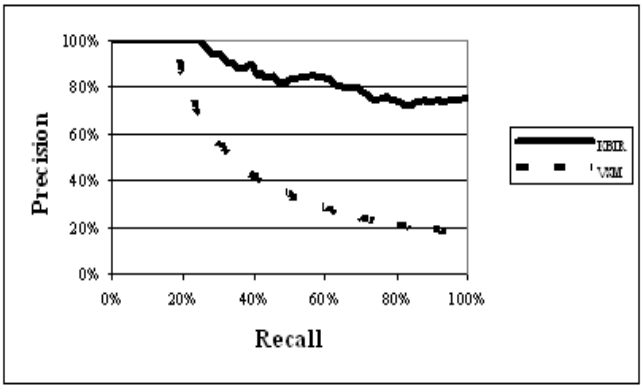


Figure 14. Evaluation results of query 1 including the neutral cases

These results show a 39.6% gain in performance when compared with the degradation that occurs with only the positive cases. If we look at the result more generally, meaning there is an importance to the neutral cases, the performance evaluation shows promising results. The gain can be explained by the expansion of knowledge in the document vector. If we look at the result of VSM, the resulting documents only have to contain one of the query terms: membranous, nephropathy, or lupus. But KBIR retrieves some documents that do not contain any of these query words because the document vector was extended to contain terms related to the existing terms in these documents. This increases the recall rate. If we look at precision, this starts to make sense when we consider the results more generally.

Figure 15 is the result of query 2, "nephrotic syndrome" on just the positive cases. When this is contrasted with the evaluation of query 1 on positive cases, the results show a performance gain. This can be explained by the characteristics of the KBIR's knowledge management. Because the number of terms in query 2 is smaller than in query 1, the amount of expanded knowledge for query 2 is less than for query 1. This means that knowledge expansion for queries having fewer query terms tends to have smaller error rates compared to queries having many terms.

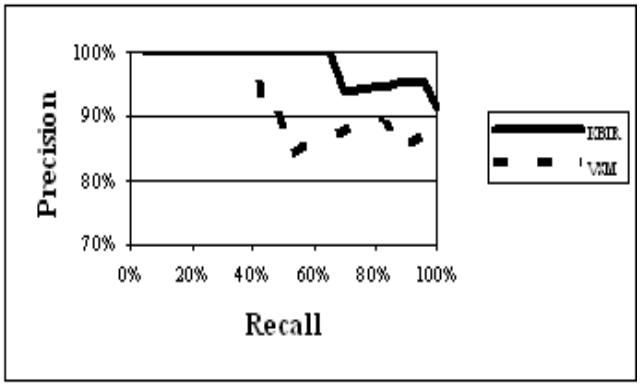


Figure 15. Evaluation results for query 2 for the positive cases

If we look at the performance evaluation results of query 2 including the neutral cases shown in Figure 16, they show a lower performance gain when compared to the results of query 1. This can be explained also by the small expansion of knowledge caused by the lower number of terms in the query.

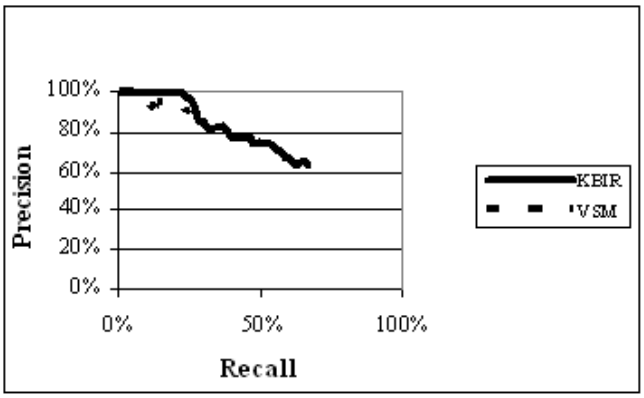


Figure 16. Evaluation results for query 2 including the neutral cases

Table 3 shows quantitative values of performance gain for the pre-coordinated knowledge addition compared to the VSM method.

Table 3. Performance gain of pre-coordinated knowledge compared to VSM

Query	Performance gain (%)
Query 1	39.6
Query 2	20.6
Average	30.1

5.2 Performance Evaluation with Naïve Bayes Post-Coordinated Knowledge

Figure 17 shows the performance gain when we use the naïve Bayes post-coordinated knowledge for query1 and Figure 18 for query 2. Table 4 shows the quantitative value of performance gain compared to VSM and pre-coordinated knowledge.

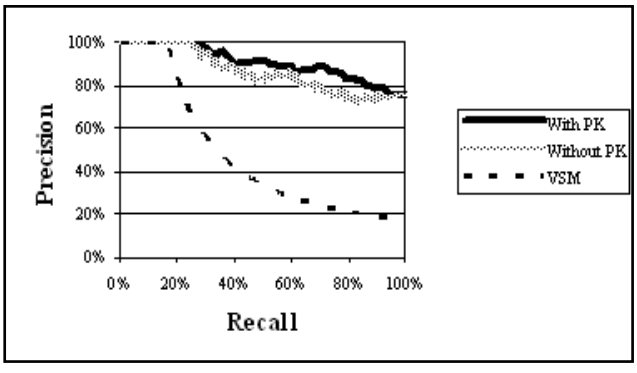


Figure 17. Evaluation results of query 1 including post-coordinated knowledge

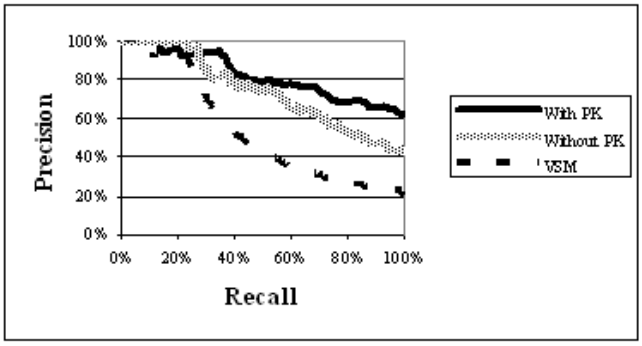


Figure 18. Evaluation results of query 2 including post-coordinated knowledge

Table 4. Value of performance gain of post-coordinated knowledge

Query	Performance gain (%) Compared to pre-coordinated knowledge	Performance gain (%) Compared to VSM
Query 1	7.0	47.0
Query 2	8.2	28.8
Average	7.6	37.9

The results show nearly the same percentage of improvement compared to the pre-coordinated knowledge case and different gain compared to the VSM case. The reason is straightforward and based on the effects of the knowledge application of our model explained in the previous section.

6. Conclusions

In this paper, we have shown significant progress towards developing an information retrieval model augmented by a knowledge base. We created a knowledge based information retrieval (KBIR) model showing meaningful performance gain while providing same speed performance in the retrieval process.

The result of our knowledge-based information retrieval model is very similar to that of query expansion or latent semantic models. Unlike those, which calculate part of the knowledge during the retrieving process, our model does its processing offline, giving the same effect with a lower computational burden. So we named it inverse query expansion.

Even if the proposed model uses domain-specific knowledge, this model can be used in an open-domain application if some types of knowledge bases are supported. One possible candidate for the open domain knowledge base is WordNet, which has a thesaurus and relations from the natural language domain.

We defined some examples of knowledge reduction methods using a semantic network. Our model has flexibility on the type of knowledge representation if we can define the knowledge reduction scheme of the selected knowledge representation model. In our model, we used a naïve Bayes network for representing post-coordinated knowledge. It has classification ability with less computational complexity and a reasonable approximation of conditional independence.

One task that needs completing is applying our model to the open domain information retrieval process. Using WordNet as a knowledge source, we can see if there is a performance gain in general domain information retrieval. Extracting knowledge automatically from given documents to use as a knowledge source for the information retrieval process is a possible approach towards applying our model to the general open domain.

Nowadays, the amount of data sets are so large and complex, it becomes difficult to process using on-hand database management tools or traditional data processing applications. Various knowledge reduction methods [22- 24] will have an important role on management of big data, extracting meaningful information and intelligence from big-data sets. One of the main objectives of knowledge reduction is attribute reduction, or dimensionality reduction that is mentioned in this paper.

The traditional von Neumann architecture based computer science faces bottleneck [26, 27] and it has totally different architecture compared with brain processing. Cognitive computing are getting spotlight to overcome the bottleneck of the von Neumann machine. Cognitive computing is a new type of computing with the goal of more accurate models of how the human brain/mind senses, reasons, and responds to stimulus. Like a human, a cognitive computing application learns by experience and/or instruction.

One of the problems of comprehending environment in cognitive computing is understanding big amounts of input data. DARPA and IBM started SyNAPSE project [28] attempt to build a new kind of cognitive computer with similar form, function, and architecture to the mammalian brain. The Human Brain Project is started by EU to create a better understanding of the human brain and its functions, as well as facilitate medical research related to healing and brain development [29]. These kinds of brain related projects attempts to resolve the problems of traditional Von Neumann architecture by mimicking human brain function.

Neuroscientist found that object perception depends on shape processing in the ventral visual pathway [30]. The convergent connection abstracts trivial visual patterns and then produces complex information. If we can make well-constructed parallel connections using knowledge reduction methods, we can mimic human brain's convergent connection.

REFERENCES

[1] Lamjiri, A. K., Kosseim, L., & Radhakrishnan, T., (2007). Comparing the Contribution of Syntactic and Semantic Features in Closed versus Open Domain Question Answering. International Conference on Semantic Computing,

pp.679-685

[2] White, Tom, (2012). Hadoop: The Definitive Guide, O'Reilly Media. p. 3. ISBN 978-1-4493-3877-0., 10 May 2012

[3] The Economist., Data, data everywhere., (25 February 2010) Retrieved 9 December 2012.

[4] Salton, G., (1971). The SMART Retrieval System. Experiments in Automatic Document Processing. Prentice Hall Inc., Englewood Cliffs, NJ.

[5] Gudivada, V. N., Raghavan, V. V., Grosky, W. I., & Kasanagottu, R. (1997). Information Retrieval on the World Wide Web. IEEE Internet Computing, 1(5), 58-68

[6] Systematized Nomenclature of Medicine, (Rodger A. Cote, Editor) College of American Pathologists, 1979, Skokie, IL.

[7] National Library of Medicine (1999). UMLS Knowledge Source Manual. National Library of Medicine.

[8] Qui, Y., & Frei., H. P. (1993). Concept based query expansion. In Proc. Of the 16th ACM SIGIR Conference on Research and Development in Information Retrieval, PA, USA, pp.160-169.

[9] Rila, M., Takenobu, T. & Hozumi, T. (1998). The Use of WordNet in Information Retrieval. COLING-ACL'98, pp.31-37.

[10] Sanderson, M. (2000). Retrieving with good sense. Information Retrieval, 2(1), 49-69.

[11] Kim, S. B., Seo, S. C., & Rim, H. C. (2004). Information Retrieval using Word Senses: Root Sense Tagging Approach. SIGIR'04, pp.258-265.

[12] Furnas, G. W., Deerwester, S., Dumais, S. T., Landauer, T. K., Harshman, R. A., Streeter, L. A., & Lochbaum, K. E. (1988). Information retrieval using a singular value decomposition model of latent semantic structure. In Proc. Of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.465-480.

[13] Campos, L. M., Fernandez-Luna, J. M. & Huete, J. F. (2000). Building Bayesian Network-Based Information Retrieval Systems. DEXA Workshop 2000, pp.543-552.

[14] Antal, P. B., De Moor, B., Timmerman, D., Neszaros, T., & Dobrowiecki, T. (2002). Domain Knowledge based Information Retrieval Language: an Application of Annotated Bayesian Networks in Ovarian Cancer Domain. In Proceedings of the 15th IEEE Symposium on Computer-Based Medical Systems (CBMS 2002), pp.213-218.

[15] Pfitzner, D., Hobbs, V., & Powers, D. (2003). A Unified Taxonomic Framework for Information Visualization. In Proceedings of the Australian symposium on Information visualization, 24, pp. 57-66.

[16] Miquelez, T., Bengoetxea, E., & Larranaga, P. (2004). Evolutionary Computation based on Bayesian Classifiers. Int. J. Appl. Math. Comput. Sci., 14(3). 335-349.

[17] Ohmann, C., Yang, Q., Kunneke, M., Stolzing, H., Thon, K., & Lorenz, W. (1988). Bayes theorem and conditional dependence of symptoms: Different models applied to data of upper gastrointestinal bleeding. Methods of Information in Medicine, 27(2), 73-83.

[18] Todd, B. S., & Stamper, R. (1994). The relative accuracy of a variety of medical diagnostic programs. Methods Inf Med, 33, 402-416.

[19] Domingos, P., & Pazzani, M., (1997). On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, 29(2-3), 103-130.

[20] Salton, G., Yang, C. S., & Yu, C. T. (1975). A theory of term importance in automatic text analysis. Journal of the American Society for Information Sciences, 26(1), 33-44.

[21] Billhardt, H., Borrajo, D., & Maojo, V. (2002). A Context Vector Model for Information Retrieval. Journal of the American Society for Information Science and Technology, 53(3), 236-249.

[22] He, Yuguo, (2004). Knowledge Reduction and Discovery based on Demarcation Information, eprint arXiv:cs/0405104

[23] Qihe Liu, (2005). Knowledge reduction in a new information view, Communications, Circuits and Systems, 2005. Proceedings. 2005 International Conference on (Volume:2)

[24] Miao, Duoqian, Nan Zhang ; Xiaodong Yue, (2009). Knowledge reduction in interval-valued information systems, Cognitive Informatics, 2009. ICCI '09. 8th IEEE International Conference on, pp. 320-327.

[25] Changwoo Yoon., (2005). Domain-Specific Knowledge-based Information Retrieval Model using Knowledge Reduction., Ph.D. dissertation, University of Florida

[26] Backus, John W.. Can Programming Be Liberated from the von Neumann Style? A Functional Style and Its Algebra of Programs. doi:10.1145/359576.359579.

[27] Dijkstra, Edsger W., (2008). A review of the 1977 Turing Award Lecture. Retrieved 2008-07-11.

[28] Steve Lohr, (2011). I.B.M. Announces Brainy Computer Chip, http://bits.blogs.nytimes.com/2011/08/18/ibm-announces-brainy-computer-chip/?partner=rss&emc=rss#,

[29] How to build a human brain (with a computer 1,000x faster than today's) https://www.humanbrainproject.eu/-/how-to-build-a-human-brain-with-a-computer-1-000x-faster-than-today-s-,

[30] Brincat SL, Connor CE., (2004). Underlying principles of visual shape selectivity in posterior inferotemporal cortex, Nat Neurosci 7, 880-886.