Connotation Searching Method for Paper Retrieval System Based On Fuzzy Rules

Hailiang Zhao [1] **Yue Ma** [2]

[1] Department of Applied Mathematics, Southwest Jiaotong University, Chengdu 610031, China
[2] Scientific Research Office of Southwest Jiaotong University, Chengdu 610031, China

Abstract

In this paper, an overall framework on paper retrieval system based on paper's connotation is proposed. Paper database is sorted into four ranks. Each of them is mainly described by an extended keyword set, which is served as the carrier of the precise connotation of a paper. Based on the matching degree between the paper introduction's vocabulary and the extended keywords set of the topics, papers in those topics is selected by using fuzzy rules. So the papers whose connotation approximates to the user's interest can be obtained. Furthermore, an automatic method to identify new and hot topics is presented.

Keywords: Document Searching, Latent semantic indexing, Fuzzy information retrieval, Document Clustering, Web searching

1. Introduction

Reference retrieval is a very important part of the research works. Computer techniques make the available reference data increase everyday. How to find the interest references precisely with an easy and fast way become a hot point in document retrieval topics. At present, the ordinary search methods are almost based upon keyword matching mode. There are several problems with their outputs. The first is the methods can not distinguish the papers from the same topic but different precise topic. When several keywords are submitted, the outputs may be presented with lots of papers. To find the interest ones that usually may be only few of them, the user have to identify them one by one. That usually is a boring work. Second, the methods can not identify the papers from the same topic but with different keyword. Authors with different writing manner may describe the same problem with different vocabularies, so the worst thing with these methods is that some important papers may be left out. Third, these methods can not distinguish the references from different subject but sharing some common keywords. The techniques borrowing from other research areas often employ the same terms, but their meanings are very different. So papers on unrelated topics may have some common keywords. Hence, there are often lots of uninterested papers in the outputs of search methods based on keywords. That is unpleasant to users. Therefore, to improve and develop the techniques on document retrieval, and find a way that can let users to effectively obtain what they really want to are urgent problems.

Focusing on these problems, there are many works, say some [1-9], present some search methods. All the suggested methods, for example, the similarity search methods, latent semantic indexing and conceptual word-chains, are make much progress relative to the original keyword matching mode. But their efficiency is also needed to improve.

Motivating from the ideas in works mentioned above, in this paper, we provide an overall framework on paper retrieval system based on paper's connotation. The retrieval system includes papers sorting way, paper edit format, paper connotation searching methods and hot research topics predicting. Obviously, one of the most efficiency search approaches is to institute a paper database with a sorting way which is appropriate to paper searching. For this reason, a new standard paragraph format of paper is proposed, in which the paper's sort code of itself and a paragraph consisting of extended keywords are included in a special field of the paper. Extended keywords are served as the carrier of the precise connotation of a paper. If a paper database is composed of papers in this format, the suggested search process will be easy, fast and more accurate with less computing. In this paper, paper database is sorted into four ranks, i.e. subject, sub-subject, topic and specific topic. Among them, the later one is a subset of the former one in sequences. Each of them is described by a general keyword set and an extended keyword set. The general keyword set consists of only several words, which is used to identify the sub-subject or topics the user interested in. The extended keywords set consist of

words with several decades, which is employed to identify the topic and specific topic. Since searching interface can make the retrieval system comprehend and grasp the user's interests, so our search process includes two phases. In the first phase, the retrieval system roughly identify the user's interest sub-subject or topics according to the user's query by some fuzzy rules based upon the relation between the general keywords and their implication paper class. Then, the embedded extended keywords relating to the sub-subject or topics are shown to the user. The user must select his interest words among them and submitted his select to the search system. In the second one, considering that user's searching aims usually are some papers on some narrow topics, search is limited in the user's interest sub-subject or topics as follows. Based on the matching degree between the paper introduction's vocabulary and the user select word set, papers in that topics are selected by using fuzzy rules again. So the papers which have the approximate connotation with the user's interest can be obtained. The final output to the user is arranged in the matching degree. Furthermore, an attribute vector of a keyword is suggested, which components consist of some statistical searching information about the keyword. The attributes include the original time as a keyword, the total number of papers which take the word as a keyword of themselves, the stage number of papers which take the word as a keyword of themselves in the evaluation period, and the times to be searched in the evaluation period. An automatic method to identify new topics is presented, which can also be used to predict hot research topics and issues.

2. Fuzzy relations between paper classes and vocabulary

In this paper, it is assumed that all the papers in the retrieval system are sorted to four level classes in term of subject, sub-subject, topic and specific topic. Among them, the later one is a subset of the former one in sequences.

As known, according to the keywords of a paper we can roughly identify the paper's class of subject or sub-subject. Since the intersection between different research areas there are no crisp edges among different subjects. So the edges among the same level classes are also inexplicit, and the vocabularies of papers in different classes distribute continuously over the total word set. Therefore, we can conclude that there are fuzzy relations between the vocabularies and the classes of the papers, and to select appropriate vocabularies or keyword sets to express the classes on each level of the four are feasible. Based upon the considerations above, we give the next analysis.

The vocabularies used in each paper class can be collected, which can be selected from the contributions of specialist on each topic. But how to get the vocabularies is not discussed in this paper.

2.1. Use word set expressing paper's connotation

Assume that the paper database $\boldsymbol{W}$ is composed of $\boldsymbol{m}$ subjects and an unknown subjects that is

$$\boldsymbol{W}=\left(\bigcup_{\boldsymbol{i}=1}^{\boldsymbol{m}}\boldsymbol{A_i}\right)\cup\boldsymbol{N} \tag{1}$$

Where, $\boldsymbol{A_i}$ denotes the ***i*–*th*** subject, and $\boldsymbol{N}$ the unknown subjects.

Let $\boldsymbol{A_{ij}}$ be the ***j*–*th*** sub-subject of $\boldsymbol{A}_i$, and $\boldsymbol{A_i}$ is composed of $\boldsymbol{n(i)}$ sub-subjects. Similarly, Let $\boldsymbol{A_{ij}}$ consists of $\boldsymbol{n(i,j)}$ topics, $\boldsymbol{A_{ijk}}$ is its ***k*-*th*** topic. Furthermore, Let $\boldsymbol{A_{ijk}}$ consists of $\boldsymbol{n(i,j,k)}$ specific topics, $\boldsymbol{A_{ijkl}}$ is the ***l*-*th*** specific topic.

Obviously, different specific topics have different keywords. A specific topic can be characterized by using of a set with keywords many enough, which is called the sufficient keyword set of the specific topic, and is also denoted by $\boldsymbol{A_{ijkl}}$. Similarly, the sufficient keyword set of subject $\boldsymbol{A_i}$, sub-subject $\boldsymbol{A_{ij}}$, topics $\boldsymbol{A_{ijk}}$ are also denoted by $\boldsymbol{A_i}$, $\boldsymbol{A_{ij}}$ and $\boldsymbol{A_{ijk}}$, respectively. It is evidently we have the conclusions:

$$\boldsymbol{A_{ijk}}=\bigcup_{\boldsymbol{l}=1}^{\boldsymbol{n(i,j,k)}}\boldsymbol{A_{ijkl}} \tag{2}$$

$$\boldsymbol{A_{ij}}=\bigcup_{\boldsymbol{k}=1}^{\boldsymbol{n(i,j)}}\boldsymbol{A_{ijk}}=\bigcup_{\boldsymbol{k}=1}^{\boldsymbol{n(i,j)}}\bigcup_{\boldsymbol{l}=1}^{\boldsymbol{n(i,j,k)}}\boldsymbol{A_{ijkl}} \tag{3}$$

$$\boldsymbol{A_i}=\bigcup_{\boldsymbol{j}=1}^{\boldsymbol{n(i)}}\boldsymbol{A_{ij}}=\bigcup_{\boldsymbol{j}=1}^{\boldsymbol{n(i)}}\bigcup_{\boldsymbol{k}=1}^{\boldsymbol{n(i,j)}}\bigcup_{\boldsymbol{l}=1}^{\boldsymbol{n(i,j,k)}}\boldsymbol{A_{ijkl}} \tag{4}$$

$$\boldsymbol{W}=\left(\bigcup_{\boldsymbol{i}=1}^{\boldsymbol{m}}\boldsymbol{A_i}\right)\cup\boldsymbol{N}=\left(\bigcup_{\boldsymbol{i}=1}^{\boldsymbol{m}}\bigcup_{\boldsymbol{j}=1}^{\boldsymbol{n(i)}}\bigcup_{\boldsymbol{k}=1}^{\boldsymbol{n(i,j)}}\bigcup_{\boldsymbol{l}=1}^{\boldsymbol{n(i,j,k)}}\boldsymbol{A_{ijkl}}\right)\cup\boldsymbol{N} \tag{5}$$

Especially, the sufficient keyword set of the unknown subject $\boldsymbol{N}$ is employed as an orphanage. All new keywords that are not included in $\bigcup_{\boldsymbol{i}=1}^{\boldsymbol{m}}\boldsymbol{A_i}$ are registered in $\boldsymbol{N}$. Call $\boldsymbol{W}$ the universe of discourse of the sufficient keyword for the paper database of the retrieval system. Note that $\boldsymbol{W}$, $\boldsymbol{N}$, $\boldsymbol{A_i}$, $\boldsymbol{A_{ij}}$, $\boldsymbol{A_{ijk}}$, $\boldsymbol{A_{ijkl}}$ both denote the paper classes in each level and their sufficient keyword sets. That makes no confusions in symbols, since the membership relation between a paper and a class depends on the intersection of paper's vocabulary and the sufficient keyword set of

the class. Whether they are regarded as paper sets or a keyword sets, their latent class attributes are the same.

2.2. Sufficient keyword set and paper's ID for specific topics

To classify the paper database in the four levels as clearly as possible, it is demanded that every paper to be published add a sort code as paper's ID, with which the paper's specific topic is indicated. Furthermore, each level's sufficient keyword set is divided into two parts. One is the general keyword set, another one is the extended keyword set. For the later use, the two parts in each level are denoted by capital letter $\boldsymbol{B}$ and $\boldsymbol{E}$ with some subscripts, respectively. That is

$$\boldsymbol{A}_{ijkl}=\boldsymbol{B}_{ijkl}\cup\boldsymbol{E}_{ijkl} \quad (6)$$

$$\boldsymbol{A}_{ijk}=\boldsymbol{B}_{ijk}\cup\boldsymbol{E}_{ijk} \quad (7)$$

$$\boldsymbol{A}_{ij}=\boldsymbol{B}_{ij}\cup\boldsymbol{E}_{ij} \quad (8)$$

$$\boldsymbol{A}_{i}=\boldsymbol{B}_{i}\cup\boldsymbol{E}_{i} \quad (9)$$

$$\boldsymbol{W}=\boldsymbol{B}\cup\boldsymbol{E}\cup\boldsymbol{N} \quad (10)$$

Corresponding to equalities (2) to (5), between different level's sufficient keyword set we have the similar relations for $\boldsymbol{B}$ and $\boldsymbol{E}$ as the sufficient keyword set. For instance

$$\boldsymbol{B}=\bigcup_{i=1}^{m}\boldsymbol{B}_{i} \quad (11)$$

$$\boldsymbol{B}_{i}=\bigcup_{j=1}^{n(i)}\boldsymbol{B}_{ij} \quad (12)$$

$$\boldsymbol{B}_{ij}=\bigcup_{k=1}^{n(i,j)}\boldsymbol{B}_{ijk}=\bigcup_{k=1}^{n(i,j)}\bigcup_{l=1}^{n(i,j,k)}\boldsymbol{B}_{ijkl} \quad (13)$$

$$\boldsymbol{E}_{ij}=\bigcup_{k=1}^{n(i,j)}\boldsymbol{E}_{ijk}=\bigcup_{k=1}^{n(i,j)}\bigcup_{l=1}^{n(i,j,k)}\boldsymbol{E}_{ijkl} \quad (14)$$

Remarks: A word in the general keyword set must be the word which is the most common used in that topic. A word in the extended keyword set must be the word that is especially used in the papers of that topic, including ones even may be rarely used. The extended keyword set must be large enough, and include enough terms in different names or expressions. For instance, if an extended keyword set for some topics includes the words "inner product", then it should include the words "dot product" and "Scalar product".

2.3. New standard format of paper

Papers are written in the following paragraph format is called a paper with standard format, or simply called a standard paper. Its paragraph sequence is as follows.

Sort code;
Title;
Authors name;
Author affiliation;
Abstract;
Keywords (i.e. the General keyword set);
Extended Keywords (i.e. the Extended keyword set);
Introduction;
Main content;
Conclusions;
References;

Let $\boldsymbol{x}$ denote a paper, for simplicity, the paragraphs are denoted by $\boldsymbol{Code}(\boldsymbol{x})$, $\boldsymbol{Tilte}(\boldsymbol{x})$, $\boldsymbol{Author}(\boldsymbol{x})$, $\boldsymbol{Abstract}(\boldsymbol{x})$, $\boldsymbol{Keywords}(\boldsymbol{x})$, $\boldsymbol{Ext\_Keywords}(\boldsymbol{x})$, $\boldsymbol{Introduction}(\boldsymbol{x})$, $\boldsymbol{Main}(\boldsymbol{x})$, $\boldsymbol{Conclusions}(\boldsymbol{x})$, $\boldsymbol{References}(\boldsymbol{x})$.

Compromised the information capacity with the amount of word, the suitable size of the extended keyword paragraph is considered as about 50 words. We add that paragraph is just to provide enough information of vocabulary for paper's connotation retrieval. As to the general keywords paragraph, i.e. the paragraph $\boldsymbol{Keywords}(\boldsymbol{x})$ in paper **x,** we maintains five words in it. It should be emphasized that every word in $\boldsymbol{Keywords}(\boldsymbol{x})$, $\boldsymbol{Ext\_Keywords}(\boldsymbol{x})$ must be complied with the remarks on subsection B.

Comparing with the usual paper format, we can find that in the standard format there are two paragraphs more than the usual one. The two paragraphs are the sort code and extended keywords. Although the new standard is only a few words more than the usual format, its latent search information has been increased on a large extent. Moreover, the new standard format is compatible with the usual one, so the existed document retrieval method can work simultaneously.

It should be noted that to paper authors, there are no too much works have to be done, but their work can be found easily, so the new standard can be willingly accepted.

Remarks: For a paper $\boldsymbol{x}$ to be published, it is allowed to present one or two new keywords in $\boldsymbol{Keywords}(\boldsymbol{x})$, and several extended keywords in $\boldsymbol{Ext\_Keywords}(\boldsymbol{x})$, All the new keywords that are found in the two paragraphs but are not included in $\bigcup_{i=1}^{m}\boldsymbol{A}_{i}$ are registered in $\boldsymbol{N}$.

3. Clustering methods for database of the retrieval system

Papers in new standard format can be classified by its sort code. So we only discuss the clustering methods

for the papers in usual format. For simplicity, at first, we suggest some symbols and terms. For a given finite set A, we use $P(A)$ to express the cardinality of A.

Definition 1. Let X be the universe of discourse, A and B are subsets of X, and $P(A)$ is a finite. Let

$$M(A \subseteq B) = \frac{P(A \cap B)}{P(A)} \quad (15)$$

Call $M(A \subseteq B)$ the matching degree of A with respect to B.

Given a paper, to reduce computing complexity, clustering process includes two steps, i.e. the rough level and the fine level. The first one is to identify the paper's class in the levels of sub-subject or subject. The second one is to determine the paper's class in the levels of precise topic or topic.

A paper's paragraphs in usual format are as follows: **Title, Authors name, Author affiliation, Abstract, Keywords, Introduction, Main content, Conclusions, references.** By means of the general keyword set and the extended keyword set of sub-subjects, and the threshold principle in fuzzy recognition, we give the following fuzzy rules to identify the class of a given paper.

$$\text{If } M(Keywords(x) \subseteq A_{ij}) \geq r \text{ then } x \in A_{ij} \quad (16)$$

Where, $i=1,\ldots,m$, $j=1,\ldots,n(i)$, x denotes a paper, $Keywords(x)$ the keyword set of paper x, and $r \in [0,1]$ is a threshold.

For a given paper x, in the sense of the rule (16), assume that $\exists D_k \in \{ A_{ij} | \ i=1,\ldots,m, \ j=1,\ldots,n(i)\}$, $k=1,2,\ldots,s$. such that $x \in D_k$ then $x \in \bigcap_{k=1}^{s} D_k$, i.e. x is assigned to sub-subject $\bigcap_{k=1}^{s} D_k$.

To find the specific sub-subject of paper x, Let $Ext(D_k)$ be the extended keyword set of D_k, $k=1,2,\ldots,s$. Let

$$\lambda_k = M(Introduction(x) \subseteq Ext(D_k)) \quad (17)$$

$$max(\lambda)=\{k | \lambda_k \geq \lambda_j, \ j=1\ldots s\} \quad (18)$$

then x is assigned to $\bigcap_{j \in \max(\lambda)} D_j$, i.e. $x \in \bigcap_{j \in \max(\lambda)} D_j$.

The number of the extended keywords of a sub-subject that appear in the paper's introduction plays an important role in the method above. Because a paper's introduction usually includes topic background, the existed conclusions and the current research state about the topic, the concepts and the special terms used in the specific topic have more opportunities to appear in it. Therefore, to classify the papers by its connotation, the presented clustering method is reasonable.

If there does not exist a $t \in \{1,\ldots s\}$, such that $D_t = \bigcap_{j \in \max(\lambda)} D_k$, then x can be regarded as belonging to several sub-subjects. That is also reasonable because of the boundary vagueness of sub-subjects.

In the similar way, replacing A_{ij} with A_{ijk} and A_{ijkl} in rule (16), respectively, we can assign any paper to its fuzzy class in the level of topic and the level of specific topic. Thus the paper database is arranged in the form showing in Fig.1.

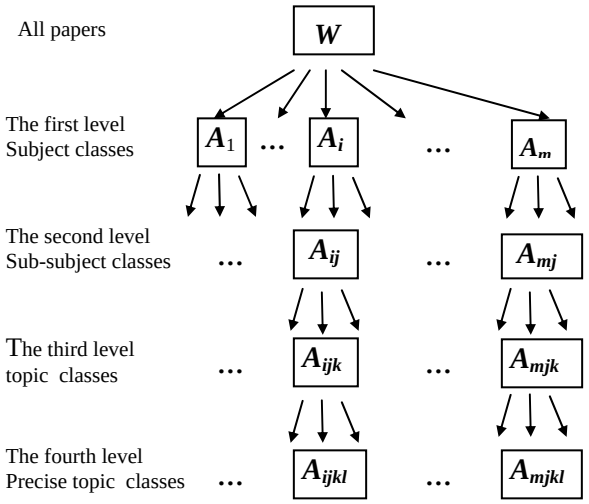

Fig 1 Architecture of the paper database

4. Interface connotation searching method with fuzzy rules

4.1. Sub-subject Identification For The User's Interest

Let $\{B_{ij} | \ i=1,\ldots,m, \ j=1,\ldots,n(i)\}$ is a complete partition of B, which is the universe of discourse of the general keyword, Here, B has the same meaning as in (11), i.e. if $W = B \cup E \cup N$, then $B = \bigcup_{i=1}^{m} \bigcup_{j=1}^{n(i)} B_{ij}$.

Let U be the keyword set of user's input in his query. The retrieval system follows to identify the users interest sub-subject according to the following fuzzy identification rules:

$$\text{If } M(U \subseteq B_{ij}) \geq r \text{ then user's interested subject is } A_{ij} \quad (19)$$

Where, $i=1,\ldots,m$, $j=1,\ldots,n(i)$.We call $M(U \subseteq B_{ij})$ the interest intensity of the user to the sub-subject A_{ij}, and r is the threshold of interest intensity.

If $\forall i,j$, $i=1,\ldots,m$, $j=1,\ldots,n(i)$, $M(U \cap B_{ij}) < r$, then the search fails. When the failure cases occur, the retrieval system can reduce r or ask user to input other keyword again.

Let

$$A(r)=\{A_{ij}|M(U\subseteq B_{ij})\geq r,\ i,j,\ i=1,\ldots,m,\ j=1,\ldots,n(i)\} \quad (20)$$

$$B(r)=\{B_{ij}|M(U\subseteq B_{ij})\geq r,\ i,j,\ i=1,\ldots,m,\ j=1,\ldots,n(i)\} \quad (21)$$

4.2. Connotation Identification and Search Output

Base on the user's interest intensity to each sub-subject A_{ij} in $A(r)$, the retrieval system shows B_{ij} in $B(r)$ to user in the sequence from large interest intensity to small one. User must select some interest words form these B_{ij} as the new input to the search system. Let V denote the user's new input set. Since the words in V are some normal or special keywords, so that the user's interest can be expressed more precisely than the first input. Let $\beta_{ij}=M(V\subseteq Ext(A_{ij}))$, that is

$$\beta_{ij}=M(V\subseteq E_{ij}) \quad (22)$$

and

$$\beta_{st}=max\{\beta_{ij}|i=1\ldots m,\ j=1,\ldots,n(i)\} \quad (23)$$

then the user's interest can regarded as the sub-subject A_{st}, $\forall x\in A_{st}$, Let

$M(x)=$

$$\begin{cases} M(Ext\_keywords(x)\subseteq V) & \text{if } x \text{ is a standard paper} \\ M(introduction(x)\subseteq V) & \text{if } x \text{ is a non standard paper} \end{cases} \quad (24)$$

Call $M(x)$ the matching degree of paper x to the user's interest.

Finally, the papers in A_{st} is ranked according to the $M(x)$.

Let $\lambda\in[0,1]$ be the acceptable level of matching degree, and out(λ) denote the papers with a matching degree which is large or equal to λ, then

$$out(\lambda)=\{x\in A_{st}\mid M(x))\geq\lambda\} \quad (25)$$

out(λ) can be regarded as the searching result of the retrieval system with matching degree λ.

Note that, in the process of interactive connotation searching, the aim of the first step is to understand user's interest. The communication between user and retrieval system can gives much information about users interest topics, it is more helpful to researchers, especially for new researchers. In the second step we use lots of extended keywords to matching the vocabulary of the paper's introduction, so the method can give connotation searching in a great extent.

5. Method to discover new topics

Considering that a new topic birth always companies with a lot of searching times for some new word in a period time. So we suggest the next definition.

5.1. Attribute vector of word

In the time interval $[t_0,t]$, For any given word $w\in W$, Let

$$Attribute(w,t)=(a_1(w,t),a_2(w,t)) \quad (26)$$

call it attribute vector of the word w. Where, $a_1(w,t)$ is the total number of papers that take w as its keyword in the time interval $[t_0,t]$, simply call absolute attribute of w; $a_2(w,t)$ is the total searching times as a user's searching word in the time interval $[t_0,t]$, simply call searching attribute of w; since t_0 is the starting point of time for statistic , we can take t_0=1980 or another year if it is early enough for papers can be found in computers.

5.2. Method to discover new topics

By considering lots of cases, we suggest an approach to discover new specific topic, or new topic, even a new sub-subject and subject. The idea is based on the fact, that there must be several non general keywords that are employed as keywords in lots of publications in a period of time, and each of the several words be searched a lot of times in the same period.

Using the same symbols as in section II and section III, Let $W=B\cup E\cup N$, $[t_1,t_2]$ is the evaluation period of time, $\forall w\in N$, let

$$\Delta a_1(w,t_0,t_1)=a_1(w,t_1)-a_1(w,t_0) \quad (27)$$

$$\Delta a_1(w,t_1,t_2)=a_1(w,t_2)-a_1(w,t_1) \quad (28)$$

$$\Delta a_2(w,t_0,t_1)=a_2(w,t_1)-a_2(w,t_0) \quad (29)$$

$$\Delta a_2(w,t_1,t_2)=a_2(w,t_2)-a_2(w,t_1) \quad (30)$$

and σ_{11}, σ_{12} ,σ_{21}, $\sigma_{22}\in[0,+\infty)$ be proper thresholds, if the next four conditions hold:

$$\Delta a_1(w,t_0,t_1)\leq\sigma_{11} \quad (31)$$

$$\Delta a_2(w,t_0,t_1)\leq\sigma_{21} \quad (32)$$

$$\Delta a_1(w,t_1,t_2)\geq\sigma_{12} \quad (33)$$

$$\Delta a_2(w,t_1,t_2)\geq\sigma_{22} \quad (34)$$

Then Let

$$F(w)=\{x|w\in Keywords(x)\} \quad (35)$$

$$G=\bigcup_{x\in F(w)}Keywords(x) \quad (36)$$

$$B_G=\bigcap_{x\in F(w)}Keywords(x) \quad (37)$$

$$E_G=G-B_G \quad (38)$$

Then G can be regarded as a sufficient keyword set of a new paper class, which is also denoted by G. B_G and E_G can be employed as G's general keywords set and extended keywords set, respectively. Thus a new paper class is recognized and developed

5.3. Upgrading process of paper database

According to (36)-(38), G is a new paper class. To determine its class level, do the following four steps.

Step 1. Using the same symbols as in section II, we calculate the matching degrees:

$$M(G\subseteq A_i)\ ,\ i=1,\ldots,m$$

Selecting a proper threshold $\mu_1\in[0,1]$, such that for any $i,j\in\{1,\ldots,m\}$and $i\neq j$, $M(A_i\subseteq A_j)<\mu_1$. We can separate any two subject classes A_i and A_j by means of threshold μ_1. For the matching degree $M(G\subseteq A_i)$, there are all three cases.

1) $\exists\ i,j\in\{1,\ldots,m\}$ and $i\neq j$ such that $M(G\subseteq A_i)\wedge M(G\subseteq A_j)\geq\mu_1$
2) For any $i\in\{1,\ldots,m\}$, $M(G\subseteq A_i)<\mu_1$
3) Otherwise there exists a unique $i\in\{1,\ldots,m\}$, such that $M(G\subseteq A_i)\geq\mu_1$

For the cases 1) and 2) we take G as a new class in subject level. Let $W=\left(\bigcup_{i=1}^{m+1}A_i\right)\cup(N-G)$

Where $A_{m+1}=G$, and $N-G$ is the new unknown object. For case 3) ,we think that G is a sub-class of subject A_i. To determine its class level, do step 2.

Step 2. Similar as step1. We select a proper threshold $\mu_2\in[0,1]$, such that for any $j,k\in\{1,\ldots,n(i)\}$ and $j\neq k$ $M(A_{ij}\subseteq A_{ik})<\mu_2$ we can separate any two sub-subject classes A_{ij} and A_{ik} by means of threshold μ_2.

For the matching degree $M(G\subseteq A_{ij})$, there are all three cases.

1) $\exists\ j,k\in\{1,\ldots,n(i)\}$ and $j\neq k$ such that $M(G\subseteq A_{ij})\wedge M(G\subseteq A_{ik})\geq\mu_2$
2) For any $j\in\{1,\ldots,n(i)\}$, $M(G\subseteq A_{ij})<\mu_2$
3) Otherwise there exists a unique $j\in\{1,\ldots,n(i)\}$, such that $M(G\subseteq A_{ij})\geq\mu_2$

For the cases 1) and 2) we take G as a new class $A_{in(i)+1}$ in the sub-subject level. Let $A_i=\bigcup_{j=1}^{n(i)+1}A_{ij}$, and

$$W=\left(\bigcup_{i=1}^{m}A_i\right)\cup(N-G)$$

Where $A_{i\,n(i)+1}=G$, **and** $N-G$ is the new unknown object .

For case 3), we think that G is a sub-class of subject A_{ij}. To determine its class level, do step 3.

Step3 and **Step 4** are omitted. Their process is just like that in step 1 and step 2. We can conclude whether G is a class in topic level or a class in specific topic level. In this way, the paper class in the four levels is upgraded.

6. Simulation result

Using the presented method, we have a simple simulation with a paper set which includes 100 papers from transactions. Simulation result shows that the proposed method works well and the searching results are just we want to find. After 50 papers selected from three 'new' fields add to the paper set in two times, three hot research topics are predicted automatically.

7. Conclusions

In this paper, we suggest an idea that paper's connotation can be carried with a large enough word set. We provide an overall framework on paper retrieval system based on paper's connotation. The proposed fuzzy rules to find the papers whose connotation are user's really interested can be easily implemented by computer program. The proposed structure of the paper database guarantee that the search processes will be easy, fast and more accurate with less computing. A new paper's paragraph format is presented. Although the new standard is only a few words more than the usual paper format, its latent search information has been increased on a large extent. Moreover, the new standard format is compatible with the usual one, so the existed document retrieval method can work simultaneously. Furthermore, the attribute vector of a keyword is suggested. An automatic method to discover new topics is presented. All the method can be easily to carry out in applications.

Acknowledgment

This work was supported by Southwest Jiaotong University under Grant 2005A26.

References

[1] C.C. Aggarwal and P.S. Yu, On Effective Conceptual Indexing and Similarity Search in Text Data. *Proc. IEEE International Conference on Data Mining*, pp. 3-10, 2001.

[2] P. W. Foltz. Using latent semantic indexing for information filtering. *Proc. of the ACM SIGOIS and IEEE CS TC-OA conference on Office information systems*, Cambridge, Massachusetts, United States, pp. 40 – 47, 1990.

[3] M. Dowman, V. Tablan, H. Cunningham, B. Popov, Web-assisted annotation, semantic indexing and

search of television and radio news. *Proc. of the 14th international conference on World Wide Web*, Chiba, Japan, pp. 225 – 234, 2005.

[4] A. Dasgupta, R. Kumar, P. Raghavan, A. Tomkins, Variable latent semantic indexing. *Proc.of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, Chicago, Illinois, USA, pp. 13 – 21,2005.

[5] Sunayama Wataru, Osawa Yukio, Yachida Masahiko, Search interface for query restructuring with discovering user interest. *Proc. of the 1999 3rd International Conference on Knowledge-Based Intelligent Information Engineering Systems*, Adelaide, Aust Aug 31-Sep 1, pp. 538-541,1999.

[6] Garces, Pablo J., Olivas, Jose A.; Romero, Francisco P., Concept-matching IR systems versus word-matching information retrieval systems: Considering fuzzy interrelations for indexing web pages. *Journal of the American Society for Information Science and Technology*, 57: 564-576, 2006.

[7] B.Y. Kang, S.J. Lee, On concept based approach for determining semantic index terms. *Lecture Notes in Artificial Intelligence* (Subseries of Lecture Notes in Computer Science), v 2807, Text, Speech and Dialogue, pp. 126-131,2003.

[8] I. Witter Dian, Berry, W. Michael, Downdating the latent semantic indexing model for conceptual information retrieval. *Computer Journal*, 41: 589-601, 1998.

[9] A. Letsche Todd, W. Berry Michael, Large-scale information retrieval with latent semantic indexing. *Information Sciences*, 100: 105-137, 1997.