2004－ICS－138　(1)
2004／12／4

Proposal of M2VSM and Its Comparison with Conventional VSM

Toru Ishibashi† and Yasufumi Takama†

Information retrieval based on Vector Space Model (VSM) only employs typical indexing terms contained in documents. For that reason, when we apply it to a specific field such as medicine, it can crowd the documents in the vector space, which makes it difficult to retrieve and cluster them. In this paper, modified VSM based on meta keywords such as adjectives and adverbs, which is called M2VSM (Meta keyword-based Modified VSM), is proposed for separating the crowded documents using meta keywords as additional value of indexing terms. Experimental results by applying M2VSM to Medline (medical literature database) show that it can separate documents crowded in the vector space.

1. Introduction

We can find huge databases easily on the Web in recent years because of breakthroughs in technique for information acquisition and dramatically low-pricing of the mass storage devices. Especially in the medical field, there is an increasing trend of storing medical documents and literature in the databases, for example, documents such as patient reports in hospital database, medical literature in medical libraries, and general medical information on the Web. As a result, the efficient retrieval of these documents from databases is becoming increasingly important.

Data mining is the process of detecting the new knowledge and rules on these huge databases and utilizing them. Various kinds of data mining and information retrieval techniques have been developed based on VSM which has several advantages. One of them is the ability to rank the documents in order of the expectation that documents are appropriate to a user's query. Another advantage is the ability to implement the relevance feedback [1]. However, when they are applied to a specific field such as medicine, the documents tend to form dense clusters in the vector space because of high similarity between them, therefore their performance decreases. Increasing the number of dimension by increasing the number of unique terms can make the vector space sparse, however this sometimes leads to problems such as curse of dimensionality, which prevents the expression of the accurate relationship between documents.

In this paper, we propose modified VSM based on meta keywords such as adjectives and adverbs, called M2VSM (Meta keyword-based Modified VSM), which divides the dense clusters into small semantically similar clusters.

Since medical data is often used as common data in the data mining field, and the concept of "EBM (Evidence-Based Medicine) based on Statistical Evidence" [2], which has had much attention recently, is related to document retrieval, we apply M2VSM to Medline, a famous online medical literature database of national library of medicine in USA, and present that effectiveness. Adjectives and adverbs are used as not additional indexing terms but additional value for indexing terms in M2VSM.

We review the conventional VSM in Section 2, and then propose M2VSM in Section 3, followed by comparative experiments by applying VSM and M2VSM to Medline in Section 4.

2. Vector Space Model

The VSM has been widely used in the traditional information retrieval field. Most search engines also use similarity measures based on VSM to rank documents on the Web. The model creates a multi-dimensional space, in which both documents and queries are represented by vectors. For a fixed collections of documents, a N_w-dimensional vector is generated for each document and query from sets of terms associated weights, where N_w is the number of indexing terms in the document collection. Then the similarity between a document and a query, and documents is calculated by cosine measure. In VSM, weight w_{ij} associated with

† Tokyo Metropolitan Institue of Technology

the term t_i in document D_j is often calculated by TFIDF (Term Frequency Inverse Document Frequency) measure [3]. The important characteristic of TFIDF measure is that the more often t_i appears in D_j, the more important t_i is in D_j, on the other hand, the more documents t_i belongs to, the less discriminative power it has, and thus the less important it is. How to calculate TFIDF value for t_i in D_j, $TFIDF(t_i, D_j)$, is defined as Eq.(1),

$$TFIDF(t_i, D_j) = \frac{m_{ij}}{M_j} \times log\frac{N_D}{DF(t_i)}, \quad (1)$$

where m_{ij} represents the number of occurrences of t_i in D_j, M_j represents the total number of indexing terms in D_j, N_D is the total number of documents and $DF(t_i)$ is the number of documents containing t_i.

The similarity $sim(q, D_j)$ between a query q and a document D_j is defined as Eq.(2), i.e., inner product of the query vector $\vec{q} = (\mathrm{q_1}, \mathrm{q_2}, ..., \mathrm{q_{Nw}})^{\mathrm{t}}$ and the document vector $\vec{D_j} = (\mathrm{w_{1j}}, \mathrm{w_{2j}}, ..., \mathrm{w_{Nwj}})^{\mathrm{t}}$.

$$sim(q, D_j) = \frac{\sum_{n=1}^{Nw} q_n w_{nj}}{\sqrt{\sum_{k=1}^{Nw} q_k^2}\sqrt{\sum_{l=1}^{Nw} w_{lj}^2}}. \quad (2)$$

3. M2VSM

As we mentioned this in Section 1, when the conventional VSM is applied to a database in a specific field such as medicine, it can crowd the documents in the vector space. That makes it hard to retrieve and cluster the documents. One of the reasons causing this problem is the existence of the indexing terms appearing in many documents, because they have the general meanings in the field. Therefore increasing the number of the indexing terms dose not only resolve the problem but also causes the curse of dimensionality at worst. In order to deal with this matter, we propose to expand the Vector Space Model based on Meta keywords (adjectives and ad-verbs).

Document retrieval systems usually employ nouns as indexing terms, which refer to a person or place or thing, expressing the topic of a document. The noun which shows the topic of a document is considered most suitable in information retrieval, but relatively high frequency terms tend to have the only general meanings, and when those words are modified by adjectives and adverbs that define the concept of terms, they can have specific meanings. Therefore we consider if same indexing terms have the different meta keywords in a different document, each document refers to the different topics. The research of using adjectives and adverbs as seed words in opinion extraction on the Internet [4] has been proposed already. To determine which words are semantically oriented, in what direction and the strength of their orientation, these kinds of research measure their co-occurrence with words from the seed set of semantically oriented words. We, however, employ adjective and adverb as meta keyword of indexing terms, so these researches and ours differ in this point. Meta often means "above" or "upper level" in artificial intelligence and knowledge engineering fields, but in M2VSM, meta means rather "with" or "metaphysical". We employ meta keywords as additional value of indexing terms, not additional indexing terms.

Given the collection of meta keywords S_M, we difine the similarity as Eq.(3),

$$sim(q, D_j) = \frac{\sum_{n=1}^{Nw} q_n w_{nj} F_{nj}}{\sqrt{\sum_{k=1}^{Nw} q_k^2}\sqrt{\sum_{l=1}^{Nw} w_{lj}^2}},$$

$$F_{ij} = \left\{ \begin{array}{l} 1....|ME_{ij}| = |ME_{iq}| = 0 \\ \frac{|ME_{ij} \bigcap_{M} ME_{iq}|}{}....otherwise \end{array} \right\}, \quad (3)$$

$$ME_{ij}, ME_{iq} \subseteq S_M, |ME_{ij}|, |ME_{iq}| \leq M,$$

where ME_{ij} represents the set of meta keywords of indexing term t_i in a document D_j, ME_{iq} represents the set of meta keywords of t_i in a query q, and F_{ij} is defined based on the intersection of ME_{ij} and ME_{iq}. In this paper, we employ two modifiers at maximum as meta keyword for t_i (i.e. $M = 2$), which is based on the observation result that even though lots of adjectives and adverbs modify an indexing term, only 2 modifiers immediately before the term are important.

Example sentences are given in Table 1. Both document D_1 and D_2, which are assumed to contain a single sentence, have "immunity" in common as an indexing term. When we suppose t_i=immunity, the sets of meta keywords are ME_{i1}=(simultaneously, virus-specific) and ME_{i2}=(cellular), and there is no common meta keyword for t_i, so these two documents are going to be divided.

Table 1 Example sentences in document D_1 and D_2

document	sentence
D_1	Simultaneously, virus-specific immunity is induced by antigen.
D_2	Hcv subverts cellular immunity by IL-10.

4. Comparative Experiments

4.1 Method

We prepare two fixed collections of documents from Medline for experiments. One consists of 100 documents about hepatitis c virus, and the other consists of 240 documents about stem cell. We extract adjectives and adverbs from these documents manually in advance, and make the collection of meta keywords S_M which consists of 760 words. As for indexing terms, we use Medical Subject Headings (MeSH), which is the National Library of Medicine's controlled vocabulary thesaurus.

We apply VSM and M2VSM to prepared two fixed collections and calculate the similarity between the documents, and visualize the similarity to see the dispersion of the documents by Keyword Map (KM) [5]. As the function of KM, when the similarity between a query and a document is more than 0, the query links to the document, and when a document is at close range to a query on a map, the similarity between them is comparatively high.

4.2 Results

Fig.1 and Fig.2 represent the similarity between documents on hepatitis c virus visualized by KM. We optionally chose one document as a query among the documents and inspected how the query links to the other documents. Fig.1 shows a query links to all the other documents and some documents which have relatively high similarity with the query make one cluster. The average number of links when each document is chosen as a query is 98.99. On the other hand, Fig.2 represents when we chose the same query in M2VSM, that cluster was divided into the other clusters and the average number of links is 6.93, which means documents crowded in high and low similarity with the query is divided because of meta keywords. Fig.3 and Fig.4 represents the results for the documents on stem cell retrieved with such a detailed query as "stem cell t-cell HIV" and considered much more crowded under VSM and M2VSM. In Fig.3 and Fig.4, the links between documents are not displayed. Fig.3 shows the documents are crowded on the map under the similarities calculated by VSM. On the other hand, the documents are spread on the map in Fig.4. By comparing Fig.3 and Fig.4, it is confirmed that crowded documents can be divided by meta keywords. Example pairs of indexing term with meta keyword are given in Table 2.

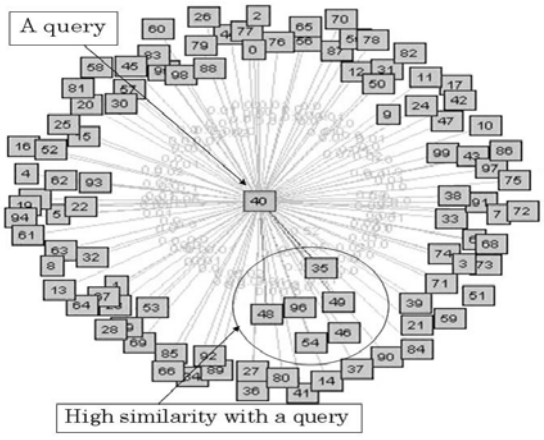

Fig. 1 visualized VSM similarity between HCV documents

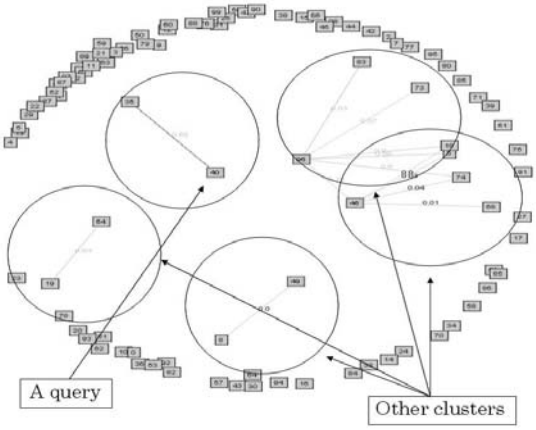

Fig. 2 M2VSM similarity between HCV documents

Having divided clusters which consist of documents that have high similarity between them into some small clusters, we confirmed whether the documents in those clusters are semantically similar or not. We chose one cluster divided into a few clusters under M2VSM and asked an expert in medicine if documents are similar in them. As a result, we found some effective meta keywords to separate the

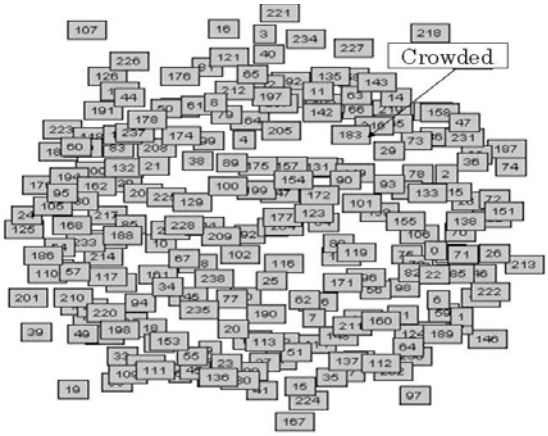

Fig. 3 Visualized similarity beween documents about Stem cell in VSM

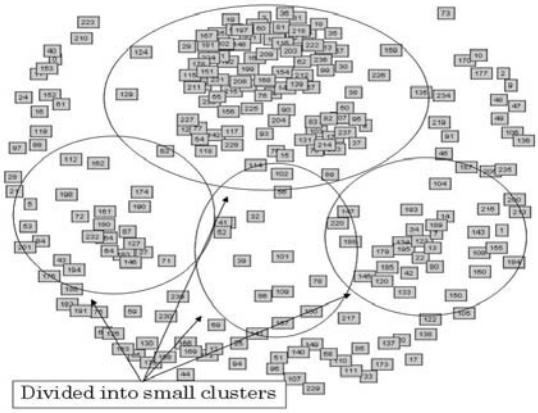

Fig. 4 Visualized similarity beween documents about Stem cell in M2VSM

documents, like "structure-base" and "non-structure-based". On the other hand, it is also found that important indexing terms were separated because of Meta keywords.

4.3 Applicaton baseline of meta keyword

From experimental results, some effective meta keywords to separate the crowded documents are found, but also it is found that important indexing terms are separated because of meta keywords. For such cases as this, we need to consider the application baseline of meta keywords. Furthermore M2VSM has an interesting characteristics such as follows; if the appropriate application baseline is set up in M2VSM, dense clusters in vector space can be made sparse without breaking the structure of original vector space. That is, the dimension will be changed by increasing or decreasing the total number of indexing terms in VSM. As a result, new members will join in one cluster, or members of cluster will go out of the cluster. In M2VSM, however, the similarity between the documents is monotonically decreasing to the application baseline of meta keyword. As a result, new members will not join in the cluster, but members of cluster can come out of the cluster.

First of all, we tried to set up the application baseline based on Zipf's law [6], but we concluded that using Zipf's law to set up the application baseline of meta keyword is not adequate by preliminary experiment. In this section, we consider using the baseline related to DF value. For setting up the application baseline of meta keyword, we confirmed the change in the similarity between documents when indexing terms which have high or low DF value are deleted

Table 2 Example pairs of indexing term with meta keywords

Meta Keyword	Indexing Term
chronic	hcv
structure-based	disign
nonstructual	proteins
peptidic, intracellur	inhibitors
virus-encoded	antigens
antigen-specific	t-cell receptor
immunomodulate	autoimmune
melanoma-associated	antigens
extrachromosomal	dna
second-site	mutations
iii-transcribed	tar
iii-expressed	sirna
t-cell-based	gene therapy
lentivirus-mediated	gene transfer
intrathymic	t-cell
macrophage-tropic	hiv-1
dz-475-treated	cells
hiv-1-specific	ctl
vector-transduced	cells
achieved	clinical
hiv-related	disorders
alloreactive, allogeneic	t-cell
pathogen-specific	immune
transduced	hematopoietic stem cells
chimeric, ii-specific	immune
hiv-1-specific	t-cell
cell-associated	epstein-barr virus
nested	polymerase chain reaction
t-tropic	virus
m-tropic	virus
hiv-1-infected	pediatric
retroviral-mediated	transfer

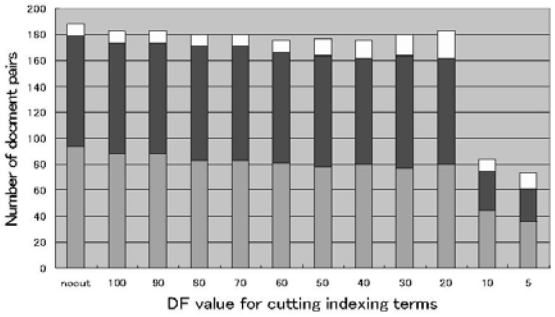

Fig. 5 Number of document pairs and DF value of cutting indexing terms that have high DF value

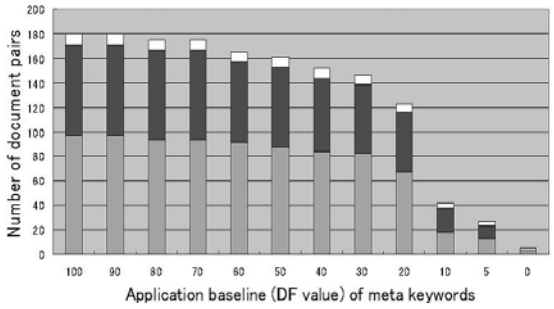

Fig. 6 Number of document pairs and application baseline (DF value) of indexing terms

in VSM. We also compared the changes in the similarity when indexing terms which have high DF value are deleted with that when meta keywords are applied to the indexing terms which have high DF value.

Fig.5 represents the number of document pairs with DF values for cutting indexing terms that have high DF values to decrease the number of dimension in VSM. Each bar is divided into 3 sections; from top, each section represents the number of document pairs which have high(over 0.6), middle(0.4 to 0.6), and low(0.3 to 0.4) similarity between them. Fig.5 dose not show the number of document pairs which have the similarity under 0.3, and this document set about stem cell is the same one used in the experiment in preceding section. In Fig.5, the number of document pairs which have high similarity between them is increasing or decreasing when the number of dimension is changing by cutting the indexing terms, and the number of document pairs are suddenly changed when indexing terms appearing in 10 or more documents are cut.

Fig.6 represents the number of document

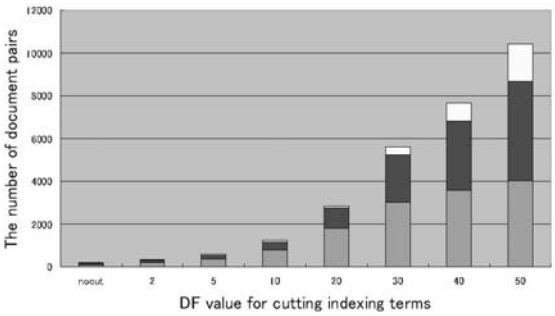

Fig. 7 Number of document pairs and DF value of cutting indexing terms that have low DF value

pairs with application baseline (DF value) of meta keywords in M2VSM. In Fig.6, the number of document pairs which have high similarity between them is getting decreasing steadily when meta keywords are applied to the indexing terms over the application baseline. Compared with Fig.5, in Fig.6 relatively high similarity document pairs are decreasing gradually, which means the vector space is getting sparse gradually because of meta keywords. As well as Fig.5, the number of document pairs are also suddenly divided when meta keywords are applied to the indexing terms which appear in 10 or more documents.

Next we confirmed how indexing terms that have low DF values affect in vector space.

Fig.7 shows the number of document paris with DF values for cutting indexing terms that have low DF values in VSM. Cutting the indexing terms that have low DF values extremely increases the number of document pairs which have high similarity between them. Only cutting the indexing term that have low DF values just makes the vector space dense, but the phenomenon of vector space being suddenly made

Table 3 DF value and the number of indexing terms(total:1329)

DF value	number of indexing terms
under 10	1134
~ 20	102
~ 30	40
~ 40	20
~ 50	10
~ 60	5
~ 70	7
~ 80	2
~ 90	2
~ 100	0
over100	7

sparse in M2VSM (in this case, the application baseline is about 10 or 20) could be compensated with this effect.

From the results, as far as this document set, indexing terms which appear in about 10 to 20 documents are important to set up the application baseline of metakeyword, though we need to investigate other document sets. The number of indexing terms per DF value is shown in Table 3. It turns out that if we suppose indexing terms which appear in under 10 documents are not important, about 85.3 percent of indexing terms are not important and only 14.3 percet of them are important, especially about 7.7 percent of them (DF values are about 10 to 20) are potent with the similarity between documents.

5. Conclusions

We propose M2VSM, a modified VSM based on meta keywords such as adjectives and adverbs. When it is applied to documents collected form Medline, it is confirmed that the documents crowded in the vector space which consist of dense clusters are divided into small clusters under M2VSM, and there are effective meta keywords to separate crowded documents. It is also found that the choice of indexing terms, to which meta keywords should be applied, is important. From this point of view, we try to set up the application baseline of meta keyword related to DF value. Regarding the qualitative evaluation, we found some document pairs, in which meta keywords supposed to work, in the process of setting up the application baseline of meta keyword. Further analysis is still in progress.

This paper only employs adjectives and adverbs as meta keyword, but we think nouns also can be meta keywords. For example, if there is an expression "involved in A" (A can be a noun or noun phrase) in a sentence, the relationship between a noun as the subjective of the sentence and, A, is important. Therefore, we need to treat nouns immediately after specific verbs and prepositions (namely, "of", "at", "with", etc.) as meta keywords.

We have not used qualifiers as meta keywords because most of the adjectives and adverbs are used as modifiers, but due to several verbs are used as adjectives in sentences, we need to consider whether verbs are used as meta keywords or not.

It is expected that M2VSM can be applied not only to medical databases but also to various databases in the specific field or specific topics. Furthermore, it will be suitable for the fields, in which meta keywords have important roles, such as opinion extraction on the Web.

References

1) Takenobu, T.: "Sec 5.2. Relevance feedback", in *Computation and Language Volume5; Information Retrieval and Natural Language Processing*, pp. 154–159 (1999)
2) TuanNam, T., Masayuki, N.: Biomedical Literature Database Retrieval System Based on Genetic Function, in proceedings of the 14th Japanese Society for AI, pp. 392–395 (2000).
3) Thorsten, J.: A probablilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization, in proceeding of the 14th International Conference on Machine Learning, pp. 143–151 (1997).
4) Hong, Yu., Vasileios H.: Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences , in proceeding of the Conference on Empirical Methods in NLP(EMNLP), pp. 129–136 (2003).
5) Yasufumi, T., Tomoki, K.,: Keyword Map-based Relevance Feedback for Web Information Retrieval, in proceeding of the 2nd Pan-Pacific Symposium on IT, pp. 41–44 (2004).
6) R.Baeza-Yates, Berthier R.: Acm Press "Sec.6.3.3. Modeling Natural Language", in *Modern Information Retrieval*, pp. 145–148 (1999)