TFIDF, LSI and Multi-word in Information Retrieval and Text Categorization

Wen Zhang, Taketoshi Yoshida
School of Knowledge Science,
Japan Advanced Institute of Science and Technology,
1-1, Ashahidai, Tatsunokuchi, Ishikawa 923-1292, Japan
{zhangwen, yoshida}@jaist.ac.jp

Xijin Tang
Institute of Systems Science,
Academy of Mathematics and Systems Science, Chinese
Academy of Sciences, Beijing 100080, P.R.China
xjtang@amss.ac.cn

***Abstract*—Text representation, which is a fundamental and necessary process for text-based intelligent information processing, includes the tasks of determining the index terms for documents and producing the numeric vectors corresponding to the documents. In this paper, multi-word, which is regarded as containing more contextual semantics than individual word and possessing the favorable statistical characteristics, is proposed as an alternative index terms in vector space model for text representation with theoretical support. We investigate the traditional indexing methods as TF*IDF (term frequency inverse document frequency) and LSI (latent semantic indexing) for comparative study. The performances of TF*IDF, LSI and multi-word are examined on the tasks of text classification, which includes information retrieval (IR) and text categorization (TC), in Chinese and English document collection respectively. We also attempt to tune the rescaling factor of LSI and observe its effectiveness in text classification. The experimental results demonstrate that TF*IDF and multi-word are comparable when they are used for IR and TC and LSI is the poorest one of them. Moreover, the rescaling factor of LSI has an insignificant influence on its effectiveness on text classification for both Chinese and English text classification.**

***Keywords*—text representation, TF*IDF, LSI, multi-word, text classification**

I. INTRODUCTION

Any text-based system requires some representation of documents, and the appropriate representation depends on the kind of task to be performed [1]. Different from data mining that handles the well-structured data, text mining deals with a collection of unstructured documents without any special requirements for their composition except some general grammar and lexical rules. This makes that one of the main themes supporting text mining is the transformation of text into numerical data, i.e., text representation.

In information retrieval, the stored documents and records are normally identified by sets of terms or keywords that are collectively used to represent the document content. Vector space model (VSM) [2] is one of the most widely used models for representation, mainly because of its conceptual simplicity and the appeal of the underlying metaphor of using spatial proximity for semantic proximity. Generally, there are two kinds of works involved in text representation: indexing and term weighting. Indexing is the job to assign the indexing terms for the documents. We should clarify here that in this paper, we will not discuss the effectiveness of indexing and term weighting separately but to discuss the problem as the level of text representation. Usually, the index terms can be predefined as a fixed set (controlled vocabulary indexing) or can be any additional words indexers regard them as related with the topic of the document (free indexing). As more and more texts are available, natural language indexing, the computer selection of indexing terms from texts has become increasingly used. Term weighting is the job to assign the weights of terms which measure the importance of terms in documents. Currently, there are many term weighting methods which are derived from the different assumptions of terms' characteristics or behaviors in texts. For instance, IDF (inverse document frequency) assumes that the importance of a term is inversely proportional to the frequency of occurrence of this term in all the documents and RIDF (residual inverse documents frequency) holds the assumption that the importance of a term should be measured by the difference between the frequency of actual occurrence in all the documents and the predicted frequency of occurrence by Poisson distribution (random occurrence). Essentially, in the task of text classification which includes information retrieval (IR) and text categorization (TC), we are mainly concerned with two kinds of properties of the indexing term: semantic quality and statistical quality [3]. Semantic quality is related with the meaning the index term contains, i.e., to how much extent the index term represents the text content; statistical quality is related with the discriminative (resolving) power of the index term to discriminate the document it belongs to from other texts in the collection.

The motivation of this research is to investigate the performance of text classification of different representation methods which are developed from different underlying hypotheses concerning indexing and term weighting. Based on the intuition for text representation, multi-word, which is a greater lexical unit than individual word and is anticipated to have both semantic quality and statistical quality, is proposed as a competitive index term. In order to disclose the preferred properties of terms used for representation, TFIDF, LSI and multi-word are comparatively studied to conduct the task of text classification in both Chinese document collection and English document collection respectively.

The rest of this paper is organized as follows. Section 2 describes the traditional representation methods as TFIDF and LSI. Section 3 proposes multi-word as index term and its properties on text classification are discussed. Section 4 is the

experiments and evaluation for TFIDF, LSI and multi-word in IR and TC on Chinese and English corpus. Section 5 is the discussion for our experimental results and concluding remarks.

II. TF*IDF AND LSI

This section describes the usually adopted representation methods as TFIDF and LSI in IR field. The basic ideas behind these two methods are discussed. Their benefits and shortcomings in text classification are presented.

A. TF*IDF

TF*IDF is evolved from IDF which is proposed by Sparck Jones [4, 5] with the heuristic intuition that a query term which occurs in many documents is not a good discriminator, and should be given less weight than one which occurs in few documents. Eq.1 is the classical formula of TF*IDF used for term weighting.

$$w_{i,j} = tf_{i,j} \times \log(\frac{N}{df_i}) \quad (1)$$

where $w_{i,j}$ is the weight for *i*th term in *j*th document, N is the number of documents in the collection, $tf_{i,j}$ is the term frequency of *i*th term in *j*th document and df_i is the document frequency of *i*th term in the collection.

The basis of TF*IDF is from the theory of language modeling that the terms in a given document can be divided into with and without the property of eliteness [6], i.e., the term is about the topic of the given document or not. The eliteness of a term for a given document can be evaluated by TF and IDF is used for the measure of importance of this term in the collection.

However, there are some deficiencies of TF*IDF method. The first one is that it is sometimes criticized as 'ad-hoc' because it is not directly derived from a mathematical model of term distribution or relevancy analysis although usually it is explained by Shannon's information theory [6]. The second one is the dimensionality of text data is the size of the vocabulary across the entire data-set. And it brings out a huge computation on the weight of each terms occurring in each document [7].

B. LSI

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

LSI (Latent Semantic Indexing) [8] is one of the most popular linear document indexing methods which produce low dimensional representations using word co-occurrence which could be regarded as semantic relationship between terms. LSI aims to find the best subspace approximation to the original document space in the sense of minimizing the global reconstruction error (the Euclidean distance between the original matrix and its approximation matrix). It is fundamentally based on SVD (Singular Value Decomposition) and projects the document vectors into the subspace so that cosine similarity can accurately represent semantic similarity. Given a term-document matrix $X = [x_1, x_2, ..., x_n] \in R^m$ and suppose the rank of X is r, LSI decomposes X using SVD as follows:

$$X = U\Sigma V^T \quad (2)$$

where $\Sigma = diag(\sigma_1, ..., \sigma_r)$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ are the singular values of X. $U = [u_1, .., u_r]$ and u_i is called the left singular vector. $V = [v_1, ..., v_r]$ and v_i is called the right singular vector. LSI uses the first k vectors in U as the transformation matrix to embed the original documents into a k-dimensional space.

There are also some deficiencies of LSI method. The first one is that there are some negative values in the reconstruction matrix we can not give a plausible explanation. It also has a huge computation as $O(n^2r^3)$, where n is the smaller of the number of documents and the number of terms, r is the rank of X [7].

III. MULTI-WORD

This section will introduce the basic ideas and assumptions of multi-word for document indexing. The method for multi-word extraction used in this paper is also discussed.

A. Definition of multi-word and motivation

A word is characterized by the company it keeps [9]. That means not only the individual word but also the context of the individual word should be laid on great emphasis for further processing. This simple and direct idea motivates the researches on multi-word which is expected to capture the context information of the individual words of its own. Although multi-word has no satisfactory formal definition, it can be defined as a sequence of two or more consecutive individual words, which is a semantic unit, including steady collocations (e.g. proper nouns, terminologies, etc.) and compound words [10, 11]. Usually, it is made up of a group of individual words and its meaning is either changed to be entirely different from (e.g. collocation) or derived by the straight-forward composition of the meanings of its parts (e.g. compound phrase).

Co-occurrence among the terms in a document expresses some kinds of semantic correlations of them. Further, co-occurrence of the relative positions of these terms proved that their co-occurrence were not accidental but to indicate a special content in documents. Moreover, in order to avoid repetitive writing, we often use the different words to refer to the same topic in a document, but this phenomenon less happened when we use a multi-word to describe a concept, i.e., there are less ambiguity and variation in multi-word.

From the terminologist's point of view [12], it appears that most of the terms encountered in technical fields are noun

phrases corresponding to a limited number of syntactic patterns while noun is widely accepted as the most significant part of speeches for indexing. It is more likely that content words (topic focused words) are included in multi-words than non-content words (topic un-focused words) since content words are always clustered in a document relevant to the topic of text, that is, their occurrence is topic dependent. According to the viewpoint of Katz [13], there is a phenomenon called burstiness for content words and phrases in texts but for non-content words and phrases, they have random occurrences which could be described by Poisson model. Nevertheless, in the field of IR, multi-word is more discriminative than individual word. As for a document d , its relevance ranking function is as follows [14]:

$$g'(d) = \sum_i X_i \log \frac{p_i}{1-p_i} + \sum_i X_i \log \frac{1-q_i}{q_i} \quad (3)$$

where $X_i = \begin{cases} 1, \text{term i is present in document} \\ 0, \text{term i is absent in document} \end{cases}$,

$p_i = P(X_i = 1/R)$ and $q_i = P(X_i = 1/\neg R)$. R means the document is relevant to the query term and $\neg R$ means the document is not relevant to the query term. For multi-word, p_i will increase and q_i will decrease because multi-word occurs more likely in a relevant document than irrelevant document. Thus, the value of $g'(d)$ will be amplified for a relevant incoming query term.

However, there are also some shortcomings with multi-word for indexing. On the one hand, it is not derived from a classical mathematic model. Although it's superiority in text classification (IR & TC) could be explained with respect to N-Grams. But there is also not an established theory for N-Grams, though N-Gram is validated in some practical application [15]. On other hand, the effectiveness of multi-word is restricted by the types of literature (genres). For instance, it will be effective with the documents in which the fixed expressions (terminologies, collocations, etc) are usually used such as news and academic papers, but not effective for the documents with extensive topics in which the fixed expressions are not usually used such as essays.

B. Multi-word Extraction

In order to extract the multi-word from documents, many multi-word extraction methods which mainly employ statistical methods based on mutual information and linguistic methods based on syntactical properties are proposed as [16-18].

For simplicity, we adopt the idea of Justeson and Katz [19] concerning the linguistic properties with multi-word to extract the multi-word from both Chinese and English documents. Basically in our method, we concentrate on the noun multi-word and assume the repetition property for them in a document. Our rule set for multi-word is that the length of the multi-word should be between 4 and 6 and the multi-word should meet the regular expression according to language type and also repeat at least two times in a document. The regular expression is described as Eq.4 for Chinese multi-word extraction and Equ.5 for English multi-word extraction.

$$((A|N)*)\ N \quad (4)$$

$$((A|N)^+|\ (A|N)*(N\Pi)^?\ (A|N)*)\ N \quad (5)$$

where A is an adjective, N is a noun and P is a preposition. It should be pointed out here that in our research, we generate the multi-word candidate for further linguistic verification by matching any two sentences in a document to look for the repetitive patterns between them as the multi-word candidates instead of using the traditional N-gram method which regards a word as a gram.

IV. Experiments and Evaluation

We carried out a series of experiments using TF*IDF, LSI and multi-word as indexing methods to examine their performance on text classification in our Chinese and English corpus respectively. Following is the details about our experiments and the results.

A. The Chinese and English corpus

As for the Chinese corpus, TanCorpV1.0 is used in this research which is available online: http://www.searchforum.org.cn/tansongbo/corpus.htm. On the whole, this corpus has 14,150 documents with 20 categories from Chinese academic journals concerning computer, agriculture, politics, etc. In this paper, documents from 4 categories as "agriculture", "history", "politics" and "economy" are fetched out as target Chinese document collection. For each category, 300 documents were selected randomly from original corpus so that totally 1,200 documents were used which have 219,115 sentences and 5,468,301 individual words in sum after morphological analysis[1].

For the English corpus, Reuters-21578 distribution 1.0 is used in this paper which is also available online (http://www.research.att.com/~lewis). It collects 21,578 news from Reuters newswire in 1987. Since 1991, it appeared as Reuters-22173 and was assembled and indexed with 135 categories by the personnel from Reuters Ltd in 1996. In this research, the documents from 4 categories as "crude" (520 documents), "agriculture" (574 documents), "trade" (514 documents) and "interest" (424 documents) are assigned as the target English document collection. That is, we select totally 2,042 documents which have 50,837 sentences and 281,111 individual words in sum after stop-word elimination[2].

B. TF*IDF, LSI and multi-word for text representation

In TF*IDF method, individual words from the texts are used as index term. For each term, term Frequency, document Frequency and TF*IDF score are computed. Then, 70 percents

[1] Because Chinese is character based, we conducted the morphological analysis using the ICTCLAS tool. It is a Chinese Lexical Analysis System. Online: http://nlp.org.cn/~zhp/ICTCLAS/codes.html

[2] We obtain the stop-words from USPTO (United States Patent and Trademark Office) patent full-text and image database at http://ftp.uspto.gov/patft/help/stopword.htm. It includes about 100 usual words as stop-words. The part of speech of English word is determined by WordNet2.0 which is available online: http://wordnet.princeton.edu/obtain and Java WordNet library which is online: http://sourceforge.net/projects/jwordnet.

terms with highest TD*IDF score are retained as features for the documents in which they occur. Next, all the features from all documents in collection are aggregated to establish the vocabulary for the collection to generate the representation vector using VSM model. If a term occur in a document, the TF*IDF score of the term in that document will be used as term weight in the corresponding term-document vector. Otherwise, weight of the term in the vector will be set as 0.

In LSI method[3], initial terms for the collection are those individual words whose term frequency in this document is more than 2 and the initial term weight will be set as the corresponding term frequency in the document. Then, SVD is used to decompose the initial term-document matrix. Further, we set the rescaling factor as 1.0 and 0.7 to project the document vectors into the respective lower dimension subspace in order to compare the performance of LSI with different parameter settings.

The multi-words are extracted from the Chinese and English collection using the syntactic structure based method mentioned in Section 3.2. Then, we use the extracted multi-word as the index terms and their term frequency as weight for a document to construct the term-document matrix.

Table 1 is the dimensionalities (the number of index terms) for each indexing method using for Chinese and English text representation.

TABLE I. NUMBER OF TERMS OF TF*IDF, LSI AND MULTI-WORD FOR CHINESE AND ENGLISH COLLECTION

Indexing method	TF*IDF	LSI (0.7)	LSI (1.0)	Multi-word
Dimensionalities for Chinese collection	21,624	840	1,200	66,949
Dimensionalities for English collection	4,889	1,429	2,042	3,112

C. *The performance on IR and TC for Chinese Corpus and English Corpus*

In order to evaluate the performance of TF*IDF, LSI and multi-word in IR, 50 queries uniformly distributed in the target 4 categories in Chinese collection and 25 queries uniformly distributed in the target 4 categories in English collection are purposely utilized to conduct the task of IR respectively. For each of these queries we have checked carefully and independently the whole Chinese and English corpus to identify the corresponding set of relevant documents. Then, the query terms are represented using the same index terms as used in the corresponding indexing method to obtain the query vectors. Especially for LSI, the query vectors are projected to the same subspace reconstructed with SVD by multiplying the left single vectors. Then, cosine similarity is computed between the query vector and document vector to retrieve the relevant documents from corpus if the similarity between them is more than 0. Table 2 and Table 3 is the average measure of precision, recall and F-measure of the three methods on IR in Chinese collection and English collection.

[3] LSI is carried out with JAMA (A Java Matrix Package) which is online and can be downloaded freely: http://math.nist.gov/javanumerics/jama/.

TABLE II. PERFORMANCES OF TF*IDF, LSI AND MULTI-WORD IN CHINESE IR

Indexing method	Av-Precision	Av-Recall	Av-F-measure
TF*IDF	0.4546	0.7510	0.4795
LSI(0.7)	0.2524	0.5104	0.2721
LSI(1.0)	0.2397	0.5130	0.2612
Multi-word	0.5374	0.5498	0.4231

TABLE III. PERFORMANCES OF TF*IDF, LSI AND MULTI-WORD IN ENGLISH IR

Indexing method	Av-Precision	Av-Recall	Av-F-measure
TF*IDF	0.4512	0.5930	0.4427
LSI(0.7)	0.0659	0.5523	0.1033
LSI(1.0)	0.0655	0.5483	0.1026
Multi-word	0.6125	0.5933	0.4726

>From Table 2 and 3, it can be seen that TF*IDF and multi-word have comparable performance in IR for both Chinese and English corpus. LSI is not sensitive to the scaling factor in Chinese and English IR. On the whole, LSI can produce a comparable recall although it can not produce a comparable precision and F-measure. We will discuss and explain these results in Section 5.

As for the evaluation on the performance of TF*IDF, LSI and multi-word with TC, SVM (Support Vector Machine)[4] is used to conduct the task of categorization for the document vectors with linear kernel because it is proved superior to non-linear kernel [20]. Here, one against all strategy is employed as we have 4 categories for both Chinese and English corpus. One of the 4 categories is assigned as the positive class and other three categories are assigned as negative. Each test is repeated 5 times so the average performance is evaluated based on total 20 tests. The 3-fold cross validation is also used here to obtain a convincing outcome. Table 4 and 5 demonstrates the performances of these three indexing methods on TC for Chinese and English collection.

TABLE IV. PERFORMANCES OF TF*IDF, LSI AND MULTI-WORD IN CHINESE TC

Indexing method	Av-Precision	Av-Recall	Av-F-measure
TF*IDF	0.8842	0.8700	0.8735
LSI(0.7)	0.3088	0.3319	0.3136
LSI(1.0)	0.3380	0.3263	0.3287
Multi-word	0.7018	0.7286	0.7041

TABLE V. PERFORMANCES OF TF*IDF, LSI AND MULTI-WORD IN ENGLISH TC

Indexing method	Av-Precision	Av-Recall	Av-F-measure
TF*IDF	0.7721	0.7627	0.7646
LSI(0.7)	0.2712	0.2380	0.2392
LSI(1.0)	0.2103	0.2626	0.2230
Multi-word	0.8156	0.8215	0.8156

It can be seen from Table 4 and 5 obviously that TF*IDF and multi-word outperform LSI on the task of TC for both Chinese and English corpus. The performance of LSI on TC is also not significantly sensitive to its scaling factor. In Chinese TC, TF*IDF have better performance than multi-word method but it is the opposite when it goes to English TC. We will discuss these results in details in Section 5.

[4] Here, libsvm is used to conduct the work of TC which is online and can be download freely: http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

V. DISCUSSION AND CONCLUDING REMARKS

In this paper, a series of experiments are carried out to examine the performance of three kinds of document indexing methods as TF*IDF, LSI and multi-word after their basic ideas and motivations are specified. Basically, two kinds of properties should be considered for the indexing term as semantic quality and statistical quality. With this motivation, multi-word is proposed because it is not only representative but also discriminative by our analysis. Most importantly, the experimental results validate our anticipation for multi-word with the conclusion that it is comparable with the benchmark method as TF*IDF for document representation in text classification.

However, there are also some points in our experiments should be clarified. The fact that LSI has a better recall than precision in IR demonstrates that LSI has captured the semantic relationships but also ignored the discrimination in some extent. This point also can be seen by its poor performance on TC. We can observe from the experimental results that TF*IDF has better performance than multi-word in Chinese document collection but when it goes to the English document collection, TF*IDF has poorer performance than multi-word method. We can regard them as comparable if without further hypothesis testing and conceive their differences in performances as the influences from types of language and genres of documents. In addition, multi-word has the greatest number of indexing terms (dimensionality) for Chinese collection, but it cannot produce the best performance in both IR and TC, because most multi-words extracted from Chinese documents are general concepts instead of the expected specialized terminologies. For English corpus, the general concepts in the news reflect the focused topics of the documents so that multi-word has a better performance than TF*IDF. In this sense, the multi-word extraction method is somehow decisive for multi-word performance in IR and TC. Nevertheless, what is also worth our noticing is the computation complexity of these three methods. The computation complexity on TD*IDF score is O(nm), where n is the total number of individual words and m is the total of number of documents in the corpus. For LSI, it has a huge computation as $O(p^2k^3)$, where p is the smaller of the number of documents(m) and the number of terms(n), k is the number of single values. For multi-word, most of the computation is spent on extraction as $O(ms^2)$, where s is the average number of sentences in a document. Take the Chinese corpus for example, n is 5,468,301 and m is1,200, so p is 1200, k as approximately 1200, s as 219,115/1,200, i.e., 182.60, nm as 6.56*109, p^2k^3 as 3.58*1021; ms^2 = 4.00*107. In this sense, multi-word has the least computation complexity.

Although some conclusions have drawn from our theoretical analysis and experiments and the multi-word is proved effective for document indexing in this research, there are still some questions ahead of us as follows:

- In this paper, we gave some experimental evaluation of multi-words on text classification. However, it is lack of a strong theoretical foundation. How to estimate the performance of indexing method in theory instead of practical experiments is also the problem for us.
- The multi-word extraction is conducted by the traditional simple and intuitive linguistic method. Whether or not there will be an improvement for text representation if statistical methods are integrated to elaborate the extraction also needs further experiments.
- The basic criterion of text representation is the semantic quality and statistical quality. Unfortunately, we do not have a standard measure to guage the two kinds of qualities mathematically. Until now, these two qualities are just considered by our intuition instead of theory.
- Actually, we discussed two different points about text representation without discrimination in this paper as term weighting and index term selection. Although, they are all included in representation, their influence on the effectiveness of representation should be different. There are also some innovative researches in this filed such as [21, 22] which combine the two points organically and give us a clear clue to proceed the work in future.

ACKNOWLEDGMENT

This work is partially supported by the National Natural Science Foundation of China under Grant No.70571078 and 70221001 and by Ministry of Education, Culture, Sports, Science and Technology of Japan under the "Kanazawa Region, Ishikawa High-Tech Sensing Cluster of Knowledge-Based Cluster Creation Project".

REFERENCES

[1] D. L. David, "Text representation for intelligent text retrieval: A classification-Oriented view," Paul, S. J. (eds.): Text-based intelligent systems: current research and practice in information extraction and retrieval. Lawrence Erlbaum Associates, Inc., Publishers, Hillsdale, New Jersey, pp. 179-197, 1992.

[2] G. Salton and C. S. Yang, "On the Specification of Term Values in Automatic Indexing," Journal of Documentation, 29(4), 1973, pp. 351-372.

[3] M. G. H. Jose, "Text representation for automatic text categorization," Online: http:// www.esi.uem.es/~jmgomez/tutorials/eacl03/slides.pdf.

[4] K.Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," Journal of Documentation, 28, 1972, pp. 11-21.

[5] K. Sparck Jones, "IDF term weighting and IR research lessons," Journal of Documentation, 60(6), 2004, pp. 521-523.

[6] S. Roberston, "Understanding inverse document frequency: on theoretical argument for IDF," Journal of Documentation, 60(5), 2004, pp. 503-520.

[7] D. M. Christopher and S.Hinrich, Foundations of Statistical natural language processing. MIT Press. Cambridge, Massachusetts, 2001, pp. 529-574.

[8] S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by Latent Semantic Analysis," Journal of American Society of Information Science, 41(6), 1990, pp. 391-407.

[9] J. R. Firth, A Synopsis of Linguistic Theory 1930-1955. Studies in Linguistic Analysis. Philological Society. Oxford: Blackwell. 1957.

[10] S. M. Weiss, N. Indurkhya, T. Zhang, F. Damerau, Text Mining: Predictive Methods for Analyzing Unstructured Information. Springer Science and Business Media, Inc. 2005, pp.1-45.

[11] J. Chen, C. H. Yeh, R. Chau, "Identifying multi-word terms by text-segments," In Proceedings of the seventh international conference on Web-Age information Management Workshops. Hongkong, 2006, pp. 10-19.

V. DISCUSSION AND CONCLUDING REMARKS

In this paper, a series of experiments are carried out to examine the performance of three kinds of document indexing methods as TF*IDF, LSI and multi-word after their basic ideas and motivations are specified. Basically, two kinds of properties should be considered for the indexing term as semantic quality and statistical quality. With this motivation, multi-word is proposed because it is not only representative but also discriminative by our analysis. Most importantly, the experimental results validate our anticipation for multi-word with the conclusion that it is comparable with the benchmark method as TF*IDF for document representation in text classification.

However, there are also some points in our experiments should be clarified. The fact that LSI has a better recall than precision in IR demonstrates that LSI has captured the semantic relationships but also ignored the discrimination in some extent. This point also can be seen by its poor performance on TC. We can observe from the experimental results that TF*IDF has better performance than multi-word in Chinese document collection but when it goes to the English document collection, TF*IDF has poorer performance than multi-word method. We can regard them as comparable if without further hypothesis testing and conceive their differences in performances as the influences from types of language and genres of documents. In addition, multi-word has the greatest number of indexing terms (dimensionality) for Chinese collection, but it cannot produce the best performance in both IR and TC, because most multi-words extracted from Chinese documents are general concepts instead of the expected specialized terminologies. For English corpus, the general concepts in the news reflect the focused topics of the documents so that multi-word has a better performance than TF*IDF. In this sense, the multi-word extraction method is somehow decisive for multi-word performance in IR and TC. Nevertheless, what is also worth our noticing is the computation complexity of these three methods. The computation complexity on TD*IDF score is O(nm), where n is the total number of individual words and m is the total of number of documents in the corpus. For LSI, it has a huge computation as $O(p^2k^3)$, where p is the smaller of the number of documents(m) and the number of terms(n), k is the number of single values. For multi-word, most of the computation is spent on extraction as $O(ms^2)$, where s is the average number of sentences in a document. Take the Chinese corpus for example, n is 5,468,301 and m is1,200, so p is 1200, k as approximately 1200, s as 219,115/1,200, i.e., 182.60, nm as 6.56*109, p^2k^3 as 3.58*1021; $ms^2 = 4.00*107$. In this sense, multi-word has the least computation complexity.

Although some conclusions have drawn from our theoretical analysis and experiments and the multi-word is proved effective for document indexing in this research, there are still some questions ahead of us as follows:

- In this paper, we gave some experimental evaluation of multi-words on text classification. However, it is lack of a strong theoretical foundation. How to estimate the performance of indexing method in theory instead of practical experiments is also the problem for us.
- The multi-word extraction is conducted by the traditional simple and intuitive linguistic method. Whether or not there will be an improvement for text representation if statistical methods are integrated to elaborate the extraction also needs further experiments.
- The basic criterion of text representation is the semantic quality and statistical quality. Unfortunately, we do not have a standard measure to guage the two kinds of qualities mathematically. Until now, these two qualities are just considered by our intuition instead of theory.
- Actually, we discussed two different points about text representation without discrimination in this paper as term weighting and index term selection. Although, they are all included in representation, their influence on the effectiveness of representation should be different. There are also some innovative researches in this filed such as [21, 22] which combine the two points organically and give us a clear clue to proceed the work in future.

ACKNOWLEDGMENT

This work is partially supported by the National Natural Science Foundation of China under Grant No.70571078 and 70221001 and by Ministry of Education, Culture, Sports, Science and Technology of Japan under the "Kanazawa Region, Ishikawa High-Tech Sensing Cluster of Knowledge-Based Cluster Creation Project".

REFERENCES

[1] D. L. David, "Text representation for intelligent text retrieval: A classification-Oriented view," Paul, S. J. (eds.): Text-based intelligent systems: current research and practice in information extraction and retrieval. Lawrence Erlbaum Associates, Inc., Publishers, Hillsdale, New Jersey, pp. 179-197, 1992.

[2] G. Salton and C. S. Yang, "On the Specification of Term Values in Automatic Indexing," Journal of Documentation, 29(4), 1973, pp. 351-372.

[3] M. G. H. Jose, "Text representation for automatic text categorization," Online: http:// www.esi.uem.es/~jmgomez/tutorials/eacl03/slides.pdf.

[4] K.Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," Journal of Documentation, 28, 1972, pp. 11-21.

[5] K. Sparck Jones, "IDF term weighting and IR research lessons," Journal of Documentation, 60(6), 2004, pp. 521-523.

[6] S. Roberston, "Understanding inverse document frequency: on theoretical argument for IDF," Journal of Documentation, 60(5), 2004, pp. 503-520.

[7] D. M. Christopher and S.Hinrich, Foundations of Statistical natural language processing. MIT Press. Cambridge, Massachusetts, 2001, pp. 529-574.

[8] S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by Latent Semantic Analysis," Journal of American Society of Information Science, 41(6), 1990, pp. 391-407.

[9] J. R. Firth, A Synopsis of Linguistic Theory 1930-1955. Studies in Linguistic Analysis. Philological Society. Oxford: Blackwell. 1957.

[10] S. M. Weiss, N. Indurkhya, T. Zhang, F. Damerau, Text Mining: Predictive Methods for Analyzing Unstructured Information. Springer Science and Business Media, Inc. 2005, pp.1-45.

[11] J. Chen, C. H. Yeh, R. Chau, "Identifying multi-word terms by text-segments," In Proceedings of the seventh international conference on Web-Age information Management Workshops. Hongkong, 2006, pp. 10-19.