Cluster Computing (2019) 22:S14897–S14903
https://doi.org/10.1007/s10586-018-2454-y

# English corpus and literary analysis based on statistical language model

Bo Huang[1] · Xijun Lan[1]

Received: 11 February 2018 / Revised: 3 March 2018 / Accepted: 6 March 2018 / Published online: 13 March 2018


**Abstract**
In this paper, the cross-language retrieval model based on statistical language model, cross-lingual text categorization method and cross-lingual text clustering method are studied systematically and deeply. Without any help of cross-lingual resources such as machine translation and bilingual dictionaries, this paper can solve the many-to-many problem of word translation in CLIR and solve the problem of unregistered words partially. Under a unified framework, a series of topics are extracted from bilingual parallel corpora to form the thematic space for each language. Thematic space of each language exists independently, and the bilingual subject space is established through the bilingual semantic correspondence. The bilingual subject space reflects the semantic correspondence between documents and documents, words and words. It reveals the inherent structure and internal relations among languages and languages.

**Keywords** Statistical language model · English corpus · Machine translation

## 1 Introduction

Searching for information has become a part of our daily lives and people often use native language [1]. However, with the rapid development of the Internet and the progress of globalization, the information resources provided by the Internet are no longer concentrated in a few languages such as English, and people can access information in various languages [2]. According to W3Techs' statistical analysis, English accounted for 54.8% of the content languages used in the global Internet Web site as of April 2013, with less than 7% in other languages, of which Chinese accounted for 4.4% [3]. As for May 31, 2011, the number of Internet users worldwide released by Internet World Stats was 2.01 billion, of which 26.8% were English users (29.4% in 2008) and 24.2% were Chinese users in the second place (2008 18.9%) [4]. The growth of Internet users using Chinese is very significant, and other Internet users who use non-English also have different degrees of growth. Therefore, there is a growing demand for using native language to query information expressed in different languages [5].

## 2 State of the art

A key issue in information retrieval is the multi-semantic representation of an object, which is manifested as polysemy, parasynonyms and synonyms in linguistic terms [6]. Even if the words or phrases contained in the query appear in the document, they may represent another meaning depending on the context. This rich semantic representation of natural language increases the difficulty of retrieving and querying related documents in the IR system [7]. In cross-language information retrieval or multi-language information retrieval, queries and documents are expressed in different languages. In addition to the semantic combination of words or phrases in a single language, cross-language semantic combinations exist, which increases the difficulties of retrievaling related documents. To overcome the language barrier we have to answer four core questions: What should be translated, queries or documents, or translate queries and documents into some kind of internal intermediate representation? Which type of symbol form should be translated in stem, word or phrase? How to use

✉ Xijun Lan
GilesogQw@yahoo.com

[1] School of Foreign Languages, China University of Geosciences, Wuhan, China

translation? That is, one word in the source language L1 may correspond to multiple words in the target language L1. How many words shoule we choose, one, some or all of the words? For a translation of trust, we should give it a greater weight value. How to remove inappropriate translation? Improper translation can severely affect the performance of cross-language searches, we need different censoring methods to reduce their negative impact.

## 3 Methodology

### 3.1 Information retrieval model

The main difference between the various retrieval systems is what kind of information retrieval model the system uses. Information retrieval model is the core part of the retrieval system. The information retrieval model gives a representation of the documents and queries and defines their scoring functions. We define the information retrieval model formally using a quadruplet $[D, Q, F, R(q_i, d_j)]$, where: D is the logical view (or representation) of the document set; Q is the logical view (or representation) of the user's information requirements, which are called queries; F is defined as model document representation, query and the framework of their relationship, $R(q_i, d_j)$ is defined as the ordering function of the relevance between query $q_i \in Q$ and document $d_j \in D$,which gives the document a sequence of queries about qi. Researchers have proposed many information retrieval models. This section describes only commonly used search models: boolean, vector space, probability, and language models. In the information retrieval model, terms are the basic units used to represent. Terms can be words (e.g., computer), stem (e.g., comput), or phrases (e.g., computer system). The form in which the term is taken depends on the indexing method used to represent the basic semantic unit of the information content. Terms using different methods can construct more complex representations. Boolean model is the first proposed classic information retrieval model. In a Boolean model, the document is represented as a collection of lexical items or as a Boolean combination, and the query is represented as a Boolean combination of lexical items. E.g:

$$D = t_1 \wedge t_2 \wedge t_3 \tag{1}$$

Indicates that document D contains the terms t1, t2, and t3. This formula is also equivalently expressed by a set expression. The terms that are not in Boolean expressions are assumed not to exist in the document. Similarly, the query can also use Boolean expressions. E.g, $Q = (t_1 \wedge t_2) \vee t_3$. If and only if $D \rightarrow Q$, you can determine whether a document related to the query. The main drawback of the Boolean model is that it does not give different weights to terms. So, some extended boolean models incorporate term weights, which consist of terms of different weights. Therefore, it is also possible to use a weighted Boolean model $D \rightarrow Q$. For example, fuzzy set expansion logic Boolean model, p-norm and so on.

Vector space models use vectors to represent documents or queries. The vector space consists of system-recognizable terms in the document. In the document vector and the query vector, each element value ($d_i$ or $q_i, 1 \leq i \leq n$) indicates the weight of the corresponding term in the document or query. The concrete expression is as follows, the vector space is $< t_1, t_2, \ldots, t_n >$.The document vector is $< d_1, d_2, \ldots, d_n >$, the query vector is $< q_1, q_2, \ldots, q_n >$, the weight of $d_i$ or $q_i$ can be a binary value. For example, if a term appears in a document or query, it is 1, otherwise 0. However, the current use of Tf-idf weight calculation mode. Tf (Term Frequency) is the term frequency in the document or query, Idf (Inverse Document Frequency) is the inverse document frequency. Idf weight is calculated as follows:

$$Idf = \log \frac{N}{n(t_i)} \tag{2}$$

where is the term in the vocabulary, N is the number of documents in the document set, and $n(t_i)$ is the number of documents (also called document frequency) that include the term $t_i$. The basic idea of the Tf-idf weighting model is that the more frequently a term appears in a document or query, the more important it is (Tf factor); the more a term appears in a document with more documents, the more important it is Low (Idf factor). Although the formal definition of Boolean model and vector space model can not solve the problem of uncertainty in retrieval. In the probability model, the relevance of document D to query Q is achieved by estimating the probability of P (rel| D, Q), where rel indicates the correlation. The probabilistic model is based on the principle of probability ordering: if a reference retrieval system responds to each requirement, it is sorted by descending order of relevance of the document and the requirement, where the correlation probability is based on all the data the system can get May be estimated accurately, then the system is based on known data and can obtain the overall effect of the optimal system. The simplest probability model is the Binary Independence Retrieval (BIR) model. The BIR model assumes that the terms are independent of each other. The document is sorted according to the optimal ratio of P (rel | D, Q) to P (irrel | D, Q):

$$score(Q,D) = \log\frac{P(rel|D,Q)}{P(irrel|D,Q)} = \log\frac{P(D|Q,rel)P(rel|Q)}{P(D|Q,irrel)P(irrel|D)} \propto \log\frac{P(D|Q,rel)}{P(D|Q,irrel)} \quad (3)$$

Among them, irrel in P (irrel| D, Q) is irrelevant. Documents D expressed as the collection $\{x_1, x_2, \ldots, x_n\}, x_i =$ 1 of independent binary events indicates that the term $x_i$ appears in the document, $x_i = 0$ indicating that the term $x_i$ does not appear in the document. Therefore, the derived BIR model is transformed into the estimated conditional probabilities P (xi = 1| Q, rel) and P (xi = 1| Q, irrel). Ideally, we can get a sample document set that gives a correlation. Given such a sample set of documents, we can calculate the contingency table for each term. Suppose N is the total number of documents in the sample set, R is the number of related documents, $r_i$ is the number of related documents contained $t_i$, and $n_i$ is the number of documents included $t_i$.

### 3.2 Thematic dual space model algorithm

As described above, the overall cross-lingual information retrieval framework based on bilingual subject spaces and the cross-language retrieval process need to first establish thematic duality space. The process of establishing this space was achieved by training the TSD model on a bilingual corpus. In the implementation of PLS algorithm, the classical non-linear iterative partial least squares algorithm (NIPALS) is generally adopted. The Do while loop body in the concrete algorithm converges to $u_i$ of the first correct singular vector for calculation $X_{i-1}^T t_i$, and calculates $v_i$ at the same time. After TSD model training algorithm is used to establish thematic dual space, the query and document represented by each language can be mapped to this space to achieve the task of cross-language and single language retrieval.

In terms of both space and time complexity, we analyze the algorithm efficiency of the TDS model proposed in this chapter. As mentioned above, suppose m is the number of documents in a bilingual parallel document set, n is the number of terms in language L1, r is L2.

Number of terms in the language. Without loss of generality hypothesis n > r, k is the number of pairs of bilingual subjects pre-given by the model. In the training phase of TDS model, we first need to store the bilingual document matrix X and Y, the space complexity is m × n and m × r respectively. Second, the output of the algorithm needs to store a set of vectors that represent dual space $< u_i, v_i >$, namely the matrices U and V, whose spatial complexity is n × k and r × k, respectively. Together with the central vector of the stored two-document matrix, the spatial complexity of the TDS model during the training phase is:

$$mn + mr + (r+n)k + n + r \quad (4)$$

In the training phase of TDS model, its time complexity is klmn and reaches O (n4). However, when we actually calculated the experiment,because the vector group $< u_i, v_i >$ needs to be calculated by loop l times. we found that the convergence rate was faster and the number of cycles was less than 15 times, at about 10 times. In the retrieval stage algorithm of TDS model, the algorithm stores the matrices U and V, document vector d and query vector q, the central vectors AvgX and AvgY, Pi and Qi of the original document matrix as a numerical variable, and finally outputs the projection vector. So, when retrieving a document, its spatial complexity is 2n + k for language L1 and 2r + k for language L2. Because the projection functions ProjectToU (x, U, k) and ProjectToV (y, V, k) have only one cycle, the complexity of the algorithm is n × k or r × k (Table 1).

## 4 Result analysis and discussion

### 4.1 Document paired search performance comparison analysis

We run TDS and CL-LSI two models in WSJFT bilingual parallel corpus for document matching search experiment. The paired search results when bilingual subject numbers are 5, 10, 20, and 50 are listed in Table 2. "C → E" in Table 2 means that the Chinese document retrieves the English paired document for the query, and "E → C" means the English document retrieves the Chinese paired document for the query. Figures 2 and 3 show the matching results of the two models in the range of the number of topics [100, 1000] and the step size of 100, and Fig. 1 shows the Chinese documents for English language paired documents (C → E), and Fig. 3 Search for Chinese Paired Documents for English Documents (E → C).

As can be seen from Table 2, when the number of topics is only 10, the accuracy of the Chinese document pairing search for English documents (C → E) and the English document pairing search for Chinese (E → C) is 0.5546 and 0.6056, while the accuracy of the two paired search tasks for the CL-LSI model is only about 0.34. When the number of topics is 20, the accuracy of the TDS model is about 0.85 and the CL-LSI model is about 0.73. When the number of subjects is 50, the accuracy of the two models are more than 0.90, and the TDS model is about 0.95, the advantage is more obvious. Therefore, we can conclude that, compared with the CL-LSI model, the TDS model can

**Table 1** Comparison of document pairing search performance between TDS and CL-LSI at a small number of topics

| Algorithm phase | Space complexity | Time complexity |
|---|---|---|
| Training stage | $mn + mr + (r + n)k + n+r$ | kmnl |
| Retrieval phase (retrieving a document) | $(r + n)k + 2n + k$ | nk |

**Table 2** Comparison of document pairing search performance between TDS and CL-LSI at a small number of topics

| Model | TDS | | CL-LSI | |
|---|---|---|---|---|
| Number of subjects | C → E | E → C | C → E | E → C |
| 5 | 0.1243 | 0.1356 | 0.0463 | 0.0324 |
| 10 | 0.5546 | 0.6056 | 0.3409 | 0.3385 |
| 20 | 0.8444 | 0.8691 | 0.7281 | 0.7409 |
| 50 | 0.9435 | 0.9508 | 0.9073 | 0.9391 |

obtain more abundant bilingual semantic information by extracting a smaller number of bilingual topics, and the document matching search achieves a higher accuracy and can be retrieved Most of the paired documents.

From Figs. 2 and 3, it can be observed that starting from the number of topics of 100 and the number of topics, the accuracy of document paired search increases greatly. When reaching about 500, the search performance tends to be stable. When the number of topics is around 800, the performance changes hardly change. We paired the English documents (C → E) and Chinese documents (E → C) of the two models with a paired *t* test with a 95% confidence level, and Chinese documents paired with English documents (C → E) The bilateral p value is 0.070, and the bilateral p value of the English document pairing search Chinese (E → C) is 0.027. This shows that the TDS model performs significantly better than the CL-LSI model when the English documents are paired search for Chinese (E → C), but the advantages of pairing search for English documents (C → E) are not obvious. The possible reason for this phenomenon is that training Chinese documents contains English special nouns and other content, while English documents do not have Chinese content. When pairing searches using Chinese documents, both models can retrieve relevant documents by using English content in Chinese documents, and these English words are

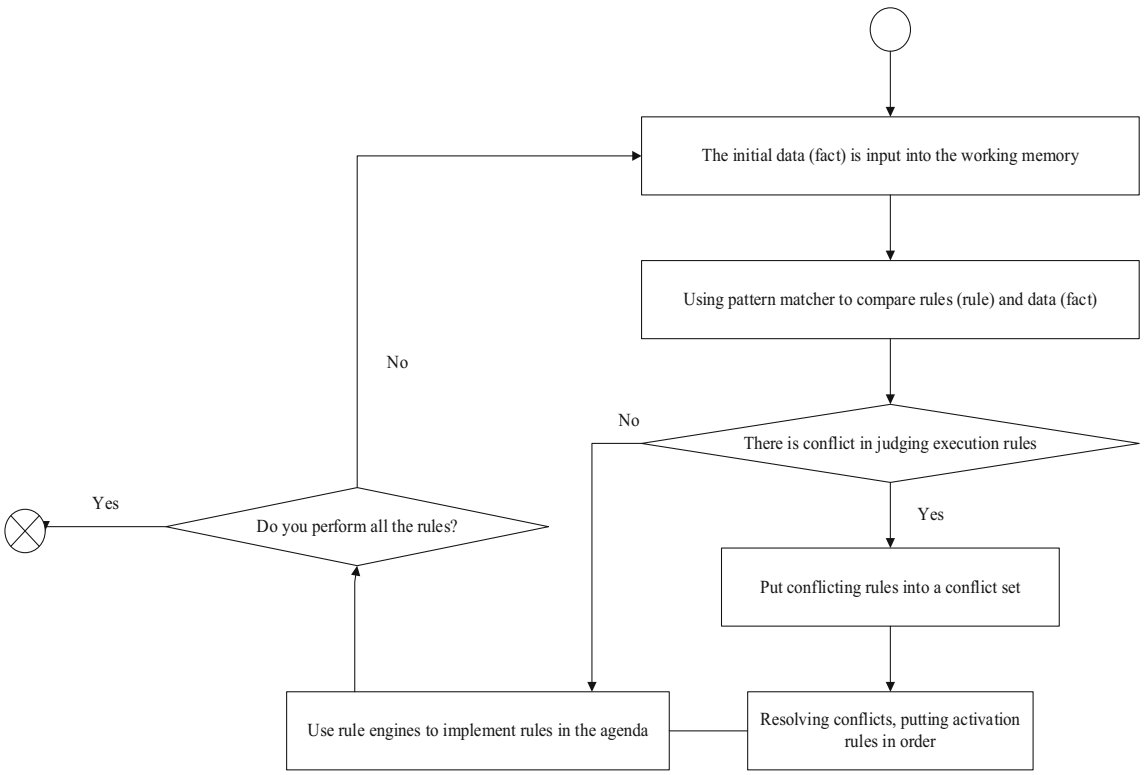


**Fig. 1** Drools schematic diagram

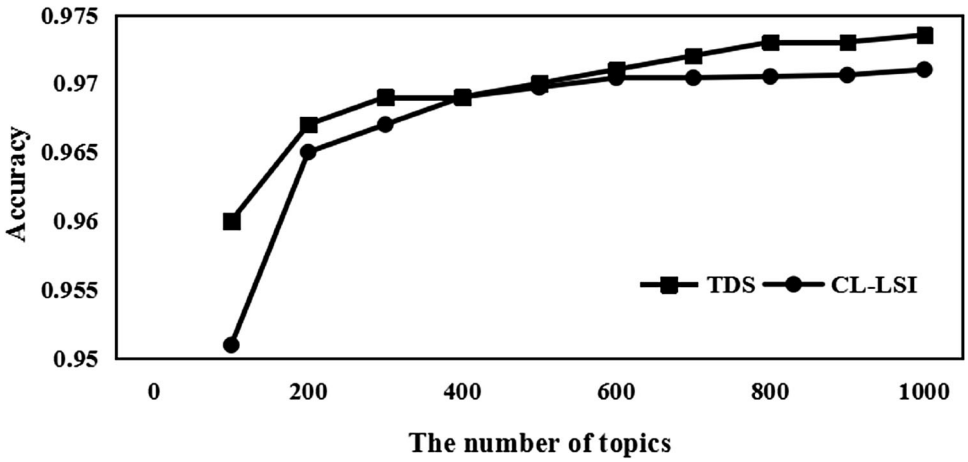


**Fig. 2** Chinese search English (C → E)

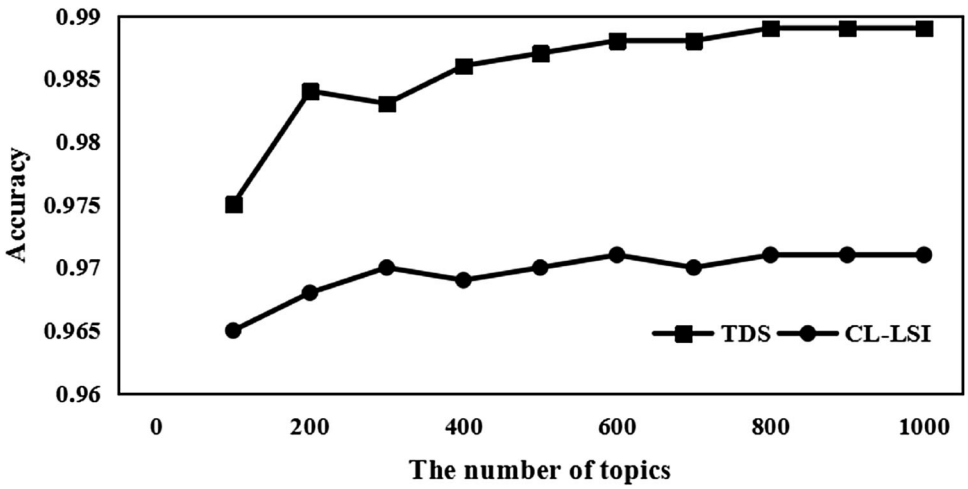


**Fig. 3** English search Chinese (E → C)

translated into a single document with more translations, so there is not much difference between the two models. However, when searching for Chinese pairing using English documents as a query, one needs to search for the corresponding translation or related term in Chinese, and on the other hand, English terms in Chinese documents also affect the similarity of the documents. The TDS model because of the training of the two languages separately modeled, the ability to distinguish more powerful language, so paired search performance advantages more models. Overall, the paired search performance of the TDS model is generally superior to that of the CL-LSI model.

## 4.2 Model parameter sensitivity analysis

In the TDS model, two important parameters that affect performance are involved: the number of bilingual topics k and the TF threshold. Figures 4 and 5 show the effect on the MAP performance of the TDS model when the number of bilingual topics k varies for the three document sets WSJFT, TREC-5 & 6 and TREC -9. The value of k in the figure is in the range of [100, 2000] and the step size is 100.

In contrast to the document set WSJFT, the TDS model has a slightly different impact on performance over TREC-5 & 6 and TREC-9 document sets. The results also list the MAP curves for these two documents, with TREC-5 & 6 on the left and TREC-9 on the right. Each graph shows the changes of MAP values of five different length queries of T, D, N, TD and TDN. When the number of subjects is less than 1000, the MAP value of the model grows faster; after 1000, the MAP of TREC5 & 6 increases relatively slowly, but the MAP value of TREC-9 keeps increasing when the number of the subjects is small. The possible reason for this phenomenon is that the data distribution of the TREC document set and the training document set is quite different, which makes the model unable to extract the semantic information more relevant to the query. Some queries (such as CH79 in TREC-9) do not have sufficient semantic information in the bilingual subject space. Because the TDS model does not have enough capacity for bilingual semantic interpretation of these queries, the

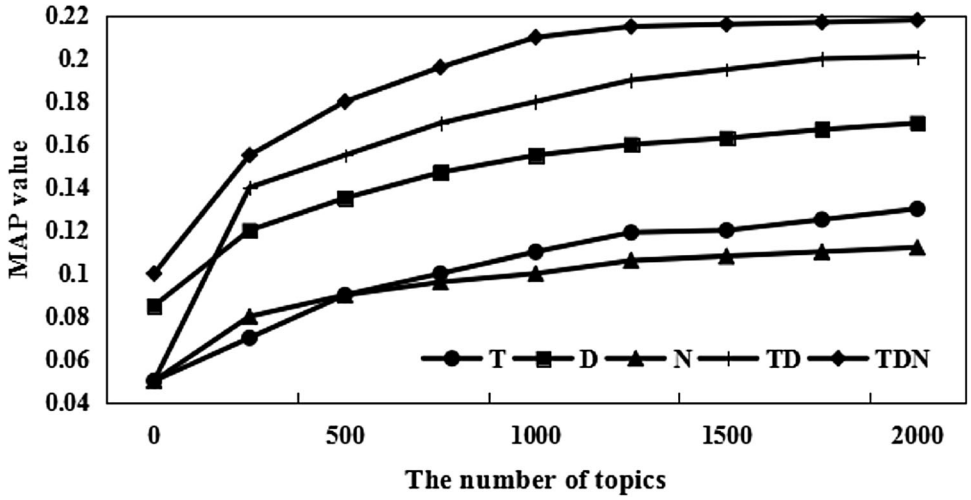


**Fig. 4** TREC-5&6

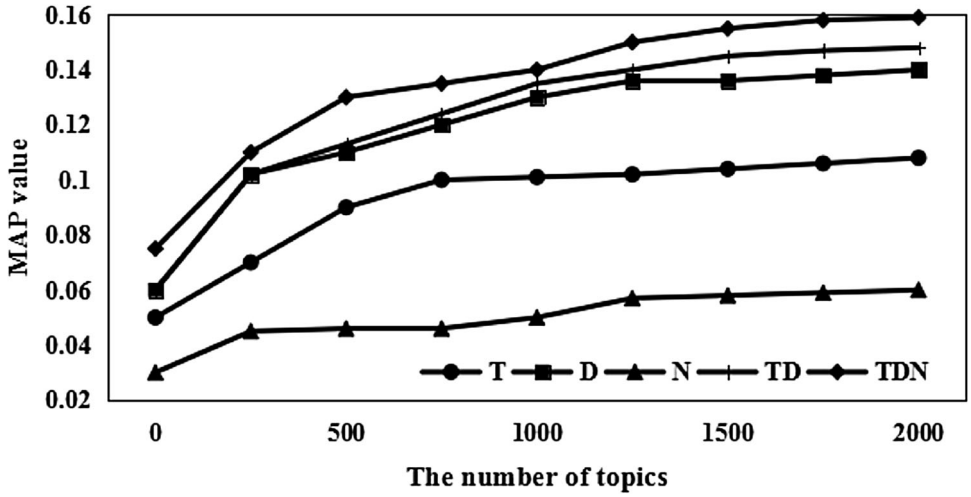


**Fig. 5** TREC-9

model can extract semantic information related to the topic as much as possible even if the number of topics is large, but the relevance is not high enough to establish an effective bilingual relationship.

When TDS models train document sets, different TF thresholds mean vocabulary of different sizes. Vocabularies of different lengths result in the missing of some words, thus affecting the semantic expression of bilingual topics. The effect of different length bilingual vocabularies on bilingual presentation ability of TDS model was analyzed after different TF thresholds were set for analysis. TF values were taken 3, 5, 10, the number of terms used in training more obvious impact. Each time the threshold value of TF increases, the total length of Chinese vocabulary is reduced by about 6000 words each time, and the English words are reduced by about 4000 words. The benchmark model is a single-language VSM model, which is represented by "ML" in the table. The MAP values of different query lengths are listed in the third row. The experimental results in the table are divided into three groups, respectively, take different TF values. " % Of ML" is expressed as the ratio of MAP values in single language VSM. On the TREC-5 & 6 document set, the best performance of the model is obtained when the TF value is 10, the TF value is 3, and the TF value is 5, the performance is the worst. On the TREC-9 document set, the MAP values of the TDS model did not change much when the TF thresholds were 3 and 5, but when the TF value was 10, the performance was optimal and there was a significant improvement. The MAP values of the 4 length queries Approaching or exceeding the single-language VSM model. The main reason is that the TDS model can effectively construct the bilingual subject space. The change of the quantity of a certain term does not have a significant impact on the model performance. Too many bilingual words will bring noise and increase the computational complexity of the model. For the change of TF threshold, it has little effect on the performance of TDS model.

## 5 Conclusion

The Internet in the context of globalization has been characterized by multiple languages and the user's information search needs are no longer confined to the use of native languages. How to help users retrieve the knowledge they need from multi-lingual information resources quickly and effectively is a frontier research field in information retrieval. Cross-Language Information Retrieval (CLIR) is one of the effective ways to solve this problem. Users are familiar with the differences between native and non-native languages, which inevitably bring about language barriers for users to utilize the Internet. This paper treats bilingual parallel documents as two views of language description objects and assumes that bilingual parallel documents share the same semantic information. These views are equivalent in essence. In order to modelling these views and extract their semantic abstract content, this paper focuses on the use of Partial Least Squares (PLS) statistical analysis theory to extract a series of bilingual subject pairs with the same semantic information from bilingual parallel corpora, and it constructs a thematic space that represents the correspondence between bilingual semantics. At last, this paper establishes a cross-language information retrieval framework based on bilingual subject space.

## References


1. Otegi, A., Arregi, X., Ansa, O., et al.: Using knowledge-based relatedness for information retrieval. Knowl. Inf. Syst. **44**(3), 689–718 (2015)
2. Rahimi, R., Shakery, A., King, I.: Multilingual information retrieval in the language modeling framework. Inf. Retr. J. **18**(3), 246–281 (2015)
3. Kim, S.: Youngjoong Ko, Oard D W. Combining lexical and statistical translation evidence for cross-language information retrieval. J. Assoc Inf. Sci. Technol. **66**(1), 23–39 (2015)
4. Rahimi, R., Shakery, A., King, I.: Extracting translations from comparable corpora for Cross-Language Information Retrieval using the language modeling framework. Inf. Process. Manag. **52**(2), 299–318 (2016)
5. Mamchich, A.A.: Models and algorithms of information retrieval in a multilingual environment on the basis of thematic and dynamic text corpora. Cybern. Inf. Technol. **16**(1), 99–115 (2016)
6. Narula, G.S., Jain, V.: Improving statistical multimedia information retrieval model by using ontology. Int. J. Comput. Appl. **94**(2), 27–30 (2017)
7. Lupu, M.: Information retrieval, machine learning, and natural language processing for intellectual property information. World Pat. Inf. **49**, A1–A3 (2017)


**Bo Huang** Hubei Qianjiang, a Lecturer in China University of Geosciences (wuhan), Master of Literature, mainly engaged in the study of applied linguistics, translation, and computational linguistics.

**Xijun Lan** Ganzhou in Jiangxi Province, China University of Geosciences (wuhan), Associate Professor, School of Foreign Languages, Master of Law, is mainly engaged in legal linguistics, corpus, translation studies.