# Document Representation Methods for Clustering Bilingual Documents

**Shutian Ma**
Department of Information Management, Nanjing University of Science and Technology
No. 200 Xiaolingwei Street, Nanjing, China
mashutian0608@hotmail.com

**Chengzhi Zhang†**
Department of Information Management, Nanjing University of Science and Technology
No. 200 Xiaolingwei Street, Nanjing, China
zhangcz@njust.edu.cn

**Daqing He**
School of Information Science and Intelligent System Program, University of Pittsburgh
135 North Bellefield Avenue, Pittsburgh, PA 15260
dah44@pitt.edu

## ABSTRACT

Globalization places people in a multilingual environment. There is a growing number of users to access and share information in several languages for public or private purpose. In order to deliver relevant information in different languages, efficient multilingual documents management is worthy of study. Generally, classification and clustering are two typical methods for documents management. However, lack of training data and high efforts for corpus annotation will increase the cost for classifying multilingual documents which needs to bridge language gaps as well. Clustering is more suitable to implement in such practical applications. There are two main factors involved in documents clustering, document representation method and clustering algorithm. In this paper, we focus on document representation method and demonstrate that the choice of representation methods has impacts on quality of clustering results. In our experiment, we use parallel corpora (English-Chinese documents on topic of technology information) and comparable corpora (English and Chinese documents on topics of mobile technology and wind energy) as dataset. We compare four different types of document representation methods: Vector Space Model, Latent Semantic Indexing, Latent Dirichlet Allocation and Doc2Vec. Experimental results show that, accuracy of Vector Space Model were not competitive with other methods in all clustering tasks. Latent Semantic Indexing is overly sensitive to corpora itself, for it behaved differently when clustering two different topics of comparable corpora. Latent Dirichlet Allocation behaves best when clustering documents in small size of comparable corpora while Doc2Vec behaves best for large documents set of parallel corpora. Accordingly, characteristics of corpora should be under considerations for rational utilization of document representation methods to have better performance.



*ASIST 2016, October 14-18,* 2016, Copenhagen, Denmark.

[†Corresponding author. +86-84315963]

**Keywords**

Bilingual documents clustering, document representation, text mining, digital library.

## INTRODUCTION

With the globalization of economy and development of Internet, large-scale information in different languages is distributed on the World Wide Web. Currently, multinational corporations, libraries or information service institutions have to manage online information resources in multiple languages (Lesk, 2004). For building multilingual collection in digital library, some libraries involved staff even users (Budzise-Weaver et al., 2012). Wu, He, & Lu's survey (2012) showed that when accessing academic databases or web information, multilingual information is often required. Demand of information access on a global scale and increasing number of cross-lingual users foster the birth of multilingual information portals, such as International Children's Digital Library[1] with literature collection in 59 languages, World Digital Library [2] which interface is available in seven languages, and EMM News Explorer[3], a news summary website in 21 languages. Management of multilingual information resources have been investigated in many scenarios, such as, multilingual news aggregation websites, like Google News[4] and Yahoo! News[5], collecting news from various sources and providing an aggregate view from news around the world, web-based biosecurity intelligence systems, like BioCaster (Collier et al., 2008) and HealthMap (Freifeld et al., 2008), etc. Relevant applications are also proposed for digital library, such as, thesauri automation construction (Zeng, 2012), multilingual

[1] Available at: http://childrenslibrary.org
[2] Available at: http://wdl.org
[3] Available at: http://www.newsexplorer.eu/NewsExplorer/home/en/latest.html
[4] Available at: https://news.google.com/
[5] Available at: https://www.yahoo.com/news/

information access service via CLIR technique (Petrelli & Clough, 2012).

Currently, multilingual documents classification and clustering are two typical methods for multilingual documents management. However, documents classification needs annotated corpus to train classifiers which have expensive labor and effort while document clustering is more suitable to implement for massive documents management when in lack of annotated data. How to improve multilingual documents clustering is worthy of study. As the basis of text mining tasks, document representation method is an important factor to influence performance. Generally, it aims to convert documents into vectors. So far, various documents representation methods have already been investigated separately (Shafiei et al., 2007; Yetisgen-Yildiz & Pratt, 2005) in documents classification and clustering tasks. There are many comparative researches of classification tasks for better document representations, such as, feature selection (Keikha et al., 2009; Y. Yang & Pedersen, 1997), feature extraction (Jiang et al., 2010) or using different latent semantic based methods (Ayadi et al., 2015; Guan et al., 2013), etc. However, systematic studies to compare different document representation methods for clustering are little done. Moreover, when choosing dataset, monolingual corpora or parallel corpora are mainly used (Boyack et al., 2011; Huang & Kuo, 2010).

In order to investigate impacts caused by representation methods in the task of bilingual documents clustering, we compare four different types of methods Vector Space Model (VSM), Latent Semantic Indexing (LSI), Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Doc2Vec (D2V) (Le & Mikolov, 2014). Two typical kinds of bilingual corpora, namely bilingual parallel corpora and comparable corpora are all used as dataset. Bilingual parallel corpora contain documents in two different languages and they're translations to each other. Bilingual comparable corpora contain documents in two different languages but covering the same or similar topic. In this experiment, parallel and comparable corpora are transferred into monolingual one, represented by different methods and clustered finally. Evaluations are made based on two evaluation indexes. Experimental results indicate that representation methods differ from each other and characteristics of corpora should be under considerations for rational utilization of methods.

This paper is organized as follows: Related works give an overview of document representation methods used in this paper. Methodology section shows frameworks of clustering parallel corpora and comparable corpora. Experimental results and analysis are then followed. Conclusion and future works are given in the last section.

## RELATED WORKS

Multilingual information resources are more likely to bring users their wanted information in a comprehensive way. Information portal will have contents in more than one language. Retrieving relevant documents in these situations are almost doing multilingual documents clustering. For example, multilingual news summarization system collect news in multiple languages from different websites and then cluster them after translation (Evans et al., 2004). Multilingual digital libraries also provide multilingual query access to a monolingual collection using machine translation and clustering (Diekema, 2012). Basically, multilingual documents clustering includes two steps. First is to represent documents without language gaps. Then documents are clustered into groups based on the representation results. Related works about multilingual documents representation are mainly divided into two strategies, one is translating multilingual documents into monolingual documents first and then representing them into computable forms, like vectors. There are many options to cross from one language to another (C. Yang & Li, 2004). Another strategy is representing documents via language-independent features, e.g. name entities (Montalvo et al., 2012).

Obviously, translation-based approach is the traditional strategy for multilingual documents clustering. Irrespective of the translation quality, how to choose document representation methods is a fundamental problem to solve. Great efforts have been made for a long time. First of all, the most classical method is Vector Space Model (VSM) (Salton et al., 1975). Document is represented as a vector in the vector space. However, VSM doesn't consider semantic relations among different words while each of them represents an independent dimension. To model the sematic relations between words, like synonymy and polysemy, improved methods are proposed. Latent Semantic Indexing (LSI) (Deerwester et al., 1990) approximates the source space with fewer dimensions which uses matrix algebra technique termed SVD. Probabilistic Latent Semantic Indexing (PLSI) (Hofmann, 1999) has a more solid statistical foundation compared with LSI, since it's based on the likelihood principle and defines a proper generative model of the data. Blei, et al. (2003) proposed a more widely used topic model, Latent Dirichlet Allocation (LDA) after PLSI. It can recognize the latent topics of documents and use topic probability distribution for representations. In recent years, machine learning algorithms have also put much effort in preprocessing pipelines and data transformations when referring to data representation. It results in a representation of the data which can support effective machine learning methods (Bengio et al., 2013). One well-known method for distributed representations of sentences and documents, Doc2Vec (D2V) is proposed by Le and Mikolov (2014). It is based on Word2Vec (Mikolov et al., 2013) , which trains distributed representation in a skip-gram likelihood as the input for prediction of words from their neighbor words (Taddy, 2015) while Doc2Vec is to learn distributed vector representations for variable-length pieces of texts, from a phrase or sentence to a large documents.

Although it has been a long time since methods were generated, VSM, LSI and LDA are still popular over these years (Anandkumar et al., 2012; Chen et al., 2013; Sidorov

et al., 2014), even combined with recent proposed D2V, such as Topic2Vec, which can learn topic representations based on Doc2Vec and LDA (Niu & Dai, 2015). Also, in the task of multilingual documents clustering, modified models have also been built. Tang et al. (2011) adapted generalized VSM to cross-lingual document clustering by incorporating cross-lingual word similarity measures. Wei et al. (2008) designed a LSI-based multilingual document clustering which can generate knowledge maps (i.e., document clusters) from multilingual documents. Kim et al. (2013) proposed a frame-based document representation for comparable corpora to capture semantic of documents. Topic models also have many extensions to do multilingual documents clustering, such as MuTo (Boyd-Graber & Blei, 2009), PLTM (Mimno et al., 2009), MuPTM (Vulić et al., 2015) etc.

So far, different document representation methods have been explored separately, there is short of systematic research to compare these methods. In this paper, we use VSM, LSI, LDA and D2V to represent documents for the task of bilingual documents clustering. Two types of bilingual corpora are clustered to find out the suitable method for bilingual documents clustering.

## METHODOLOGY

### Frameworks of Bilingual Documents Clustering

In this paper, framework of bilingual documents clustering is divided into three steps. First step is to transfer bilingual documents into monolingual one. Then, documents are represented with different methods. The last step is using clustering algorithm to divide them into groups with the same sub-topic.

### Corpora Transformation for Bridging Language Gaps

As mentioned in the Related Works, we adapt the first strategy to do multilingual documents representation. In the experiment, corpora in source languages are all translated into monolingual corpora in target language with machine translation. Then, original corpora and translated corpora are combined according to the corpora type. Frameworks of clustering parallel corpora and comparable corpora are shown in Figure 1 and 2 respectively.

As Figure 1 shown, documents set $\{T_1, T_2, \cdots, T_i, .., T_n\}$and $\{S_1, S_2, \cdots, S_i, \dots, S_n\}$ represent documents sets which are translations to each other in target language $T$ and source language $S$, $n$ is the number of documents. The following sentences pair is an example of transformation of parallel document $T_i$ and $S_i$, and target language is English, source language is Chinese:

$T_i$: *Android and iOS users have long been using their Facebook account for single click logging in to apps, and soon Windows 8 and Windows Phone users will be able to do the same.*

$S_i$: *目前 Android 和 iOS 的 app 都已可支援 Facebook 的一键式登入功能，现在就连 Windows 8 和 Windows Phone 8 的用户都可以用了。*

To cluster these documents, transformation are made via text translation and combination. Firstly, document $S_i$ is translated from source language into target language, we obtain document $T_i'$ and set $\{T_1', T_2', \cdots, T_i', \dots, T_n'\}$ represents documents set in language $T$ after translation in Figure 1.

$T_i'$ : *Currently Android and iOS app has one-click login functionality to support Facebook, Windows 8 and now Windows Phone 8 users can use.*

Then we combine document $T_i$ and translated document $T_i'$ into one document, we obtain document $T_iT_i'$ as below.

$T_iT_i'$ : *Android and iOS users have long been using their Facebook account for single click logging in to apps, and soon Windows 8 and Windows Phone users will be able to do the same. Currently Android and iOS app has one-click login functionality to support Facebook, Windows 8 and now Windows Phone 8 users can use.*

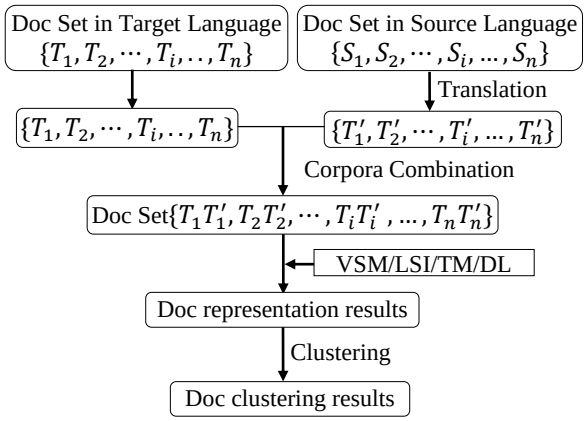



**Figure 1 Clustering framework of parallel corpora**

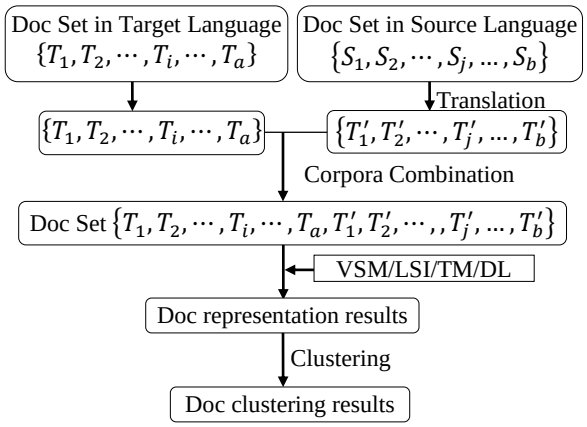



**Figure 2 Clustering framework of comparable corpora**

As Figure 2 shown, in the framework of comparable corpora clustering, $\{T_1, T_2, \cdots, T_i, \dots, T_a\}$ represent documents set of comparable corpora in target language $T$ and $a$ is the documents number of this set. $\{S_1, S_2, \cdots, S_j, \dots, S_b\}$ represent documents sets of comparable corpora in source language $S$ and $b$ is the documents number of this set. Set $\{T_1', T_2', \cdots, T_j', \dots, T_b'\}$ represents documents set in language $T$ after translation. Different from parallel corpora, we directly combine original documents set $\{T_1, T_2, \cdots, T_i, \cdots, T_a\}$ and translated documents set $\{T_1', T_2', \cdots, T_j', \dots, T_b'\}$ into a new set.

Documents number of new corpora is $a + b$ and corpora are organized like$\{T_1, T_2, \cdots, T_i, \cdots, T_a, T_1', T_2', \cdots, , T_j', \ldots, T_b'\}$.

After transformation from original corpora to new combined corpora, the remaining steps of these two frameworks are all the same. After preprocessing of documents, corpora are represented by four document representation methods and clustered. Detailed information will be described in the following subsections.

**Document Representation after Corpora Transformation**

In this section, four different documents representation methods are introduced respectively.

*Vector Space Model*

Vector Space Model (Salton, et al., 1975) is used to represent documents as vectors. Since each dimension in a vector is independent of other dimensions, terms represented in the vector space are assumed to be mutually independent. It can't be viable when facing words which are related semantically, such as synonymy and polysemy. Although VSM might in fact hurt the overall performance (Baeza-Yates & Ribeiro-Neto, 1999), the convenient computation framework makes it the most classical representation method. Moreover, dimensions reduction needs to be done for there will be curse of dimensionality when facing large documents set and some words just have little effects on documents classification and clustering. Dimensions reduction is often based on the statistical measurement of term itself, such as document frequency, information gain, mutual information, etc.

In this paper, we firstly use the VSM model to represent documents. The weight value of each term is computed via *TF*IDF* (term frequency–inverse document frequency) (Salton & Buckley, 1988). The $doc_i$ can be described as $[W_{i1}, W_{i2}, \cdots W_{ij}, \cdots, W_{im}]$, where $W_{ij}$ is *TF*IDF* value of the *j*th term in the *m*-dimensional vector space. Without doing any dimension reduction, all disparate terms in the corpora will represent documents with VSM. Then, in order to compare with the other three models at the same dimensions, we represent documents at smaller dimensions. Words occurring in fewer than 2 documents are moved and top features are chosen from terms which are ordered by term frequency across the corpus. For parallel corpora, VSM model is tested at following dimensions: 10, 20, 30, ⋯, 180, 190, 200 respectively, with 10 as interval. For comparable corpora, since the corpora is of small size and dimensions from 5 to 50 is not suitable for representing documents so we just use all disparate terms.

*Latent Sematic Indexing*

Representation in the vector space model doesn't model the sematic relations of terms, some methods are proposed to solve this problem, such as Latent Semantic Indexing (Deerwester, et al., 1990). LSI is the approach that using singular value decomposition (SVD) with a raw term-by-document matrix to get reduced document matrix under certain singular vectors. From a semantic perspective, it creates the latent concept space to represent documents (La Fleur & Renström, 2015). Statistical patterns of terms are explored so that related documents which may not share terms are still represented by neighboring vectors. As a model using numerical algorithm to reduce dimensions, how to determine the optimal rank to compute low-rank approximations is also a question.

The $doc_i$ can be described as $[W_{i1}, W_{i2}, \cdots W_{ij}, \cdots, W_{ik}]$, where $W_{ij}$ is value of the *j*th feature in the *k*-dimensional semantic space. *TF*IDF* matrix are used as input. To lessen the error of limited dimension numbers, we represent documents in several dimensions. For parallel corpora, LSI model is tested at following dimensions: 10, 20, 30, ⋯, 180, 190, 200 respectively, with 10 as interval. For comparable corpora, LSI model is tested at following dimensions: 5, 10, 15, ⋯, 40, 45, 50, with 5 as interval.

*Latent Dirichlet Allocation*

After generation of LSI, latent topic modeling has become very popular as a technique for topic discovery in document collections, such as Latent Dirichlet Allocation (Blei, et al., 2003). LDA is a generative probabilistic model and its basic idea is documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. Similar with LSI, LDA exploits the co-occurrence patterns of words to represent documents in the latent topic space, but it models a topic to be a distribution over a fixed vocabulary rather than reducing dimensions using numerical algorithm in LSI.

In LDA model, collapsed Gibbs sampling (Griffiths & Steyvers, 2004) is used and topic probability is used to be the weight value of each feature. The $doc_i$ can be described as $[p_{i1}, p_{i2}, \cdots p_{ij}, \cdots, p_{ik}]$, where $p_{ij}$ is probability of the *j*th topic feature in the *k*-dimensional topic space. We set up the same dimension numbers with LSI.

*Doc2Vec*

As one of the significant results achieved in many NLP and ML tasks, Doc2Vec (Para2Vec) is to obtain representations for larger blocks of texts, such as sentences, paragraphs even the entire documents (Le & Mikolov, 2014). In Doc2Vec, first key stage is unsupervised training to get word vectors which is same with Word2Vec, then the inference stage is to get paragraph vector. The third stage is to turn the paragraph vector to make a prediction about some particular labels using a standard classifier.

To sum up, this representation vector is learned by predicting the surrounding paragraphs in contexts sampled from the documents set. In Doc2Vec model, $doc_i$ can be described as $[W_{i1}, W_{i2}, \cdots W_{ij}, \cdots, W_{ik}]$ where $W_{ij}$ is value of the *j*th document feature in the *k*-dimensional feature space. We set up the same dimension numbers with LSI. The maximum distance between predicted document and context documents used for prediction within corpora is set to be 5.

**Documents Clustering Algorithm**

In this paper, Affinity propagation (AP) (Frey & Dueck, 2007) is used to cluster documents. By considering all the data points as potential cluster centers (called exemplars), it

works based on similarity between pairs of all the data points. There are two kinds of messages exchanged between data points, and each takes into account a different kind of competition. Messages can be combined at any stage to decide which points are exemplars and, for every other point, which exemplar it belongs to. In the iterative process of searching for clusters, identified exemplars start from the maximum exemplars to fewer exemplars until the number of exemplars doesn't change any more (Wang et al., 2008).

AP avoids many of the poor solutions caused by unlucky initializations and hard decisions and the input can be general nonmetric similarities. Without requiring that the number of clusters be predefined, preference is the only value to be set manually, it indicates the preference that data point is chosen as a cluster center, and influences output clusters and number of clusters. In this paper, similarity between documents are calculated with squared Euclidean distance and preference is set to be a few numerical values which are around the median of similarities. For example, when median value is 2, the preference value is set up at following values: 1, 1.2, 1.4, ⋯, 2, ⋯, 2.6, 2.8, 3, with 0.2 as interval.

## EXPERIMENT AND RESULTS ANALYSIS

### Experimental Dataset

Parallel corpora and comparable corpora are tested in our experiments respectively.

*Parallel corpora*

Parallel corpora are collected from Engadget, a technology blog website with multilingual version. We downloaded 4, 904 pairs of blogs (Between 2006 and 2013) from Chinese[6] and English[7] versions to be dataset and these blogs have few typos or spelling errors to take into account. In this experiment, we make use of the title, full text of the blog and the category of each blog. Totally, there are 20 categories of all the blogs[8]. Documents distribution of this corpora is shown in Table 2.

*Comparable corpora*

Chinese-English comparable corpora are downloaded from TTC project[9] which aims at generating bilingual terminologies from comparable corpora. TTC project released Chinese-English comparable corpora about mobile technology and wind energy respectively[10]. The number of final combined documents in English and Chinese language is shown in Table 3.

In this experiment, Microsoft Translator[11] is used to do translating works via API in September 2015. For parallel and comparable corpora, we translate the original English documents into Chinese and original Chinese documents into English. Documents preprocessing are done on the corpora after transformation. For the corpora in Chinese language, we segment Chinese sentences by a Chinese segmentation tool, namely ICTCLAS[12] and remove the stop words. For the corpora in English language, we remove the stop words and stem words to base forms by Porter Stemmer algorithm[13]. Then, we apply VSM, LSI and D2V model in Genism[14] and python package[15] of LDA model to represent documents. Affinity propagation clustering is done via Scikit-learn[16] python package.

| Category | Sum | Category | Sum | Category | Sum |
|---|---|---|---|---|---|
| Mobile products | 918 | Game products | 319 | Locating products | 21 |
| Tablet PC | 188 | Playing facilities | 248 | Wearing products | 105 |
| Internet network | 462 | Digital camera | 244 | Home appliance | 103 |
| Computer products | 442 | Handheld device | 27 | Intelligent machine | 102 |
| Peripheral equipment | 345 | Transportation related | 175 | Display products | 94 |
| Software application | 82 | Wireless application | 52 | Desktop products | 45 |
| News and report | 827 | High-definition television | 105 | | |

**Table 1 Documents distribution of dataset from Engadget**

| Topic | Wind Energy | Mobile Technology |
|---|---|---|
| Combined Corpora in Chinese | 208 | 128 |
| Combined Corpora in English | 207 | 129 |

**Table 2 Documents distribution in each topic**

### Evaluation Metrics

Supervised and unsupervised evaluation methods are used separately for clustering results of parallel corpora and comparable corpora. For parallel corpora, category of each blog is used as the true label for evaluation. V-measure which based on the true label of documents and predicted label is computed to evaluate clustering performance. For comparable corpora without true labels, we use Silhouette Coefficient to evaluate. These two indexes can all be calculated via Scikit-learn python package.

*V-measure*

Knowing the true labels of samples, we can define some supervised metric to evaluate performance of documents

[6] Available at: http://cn.engadget.com/
[7] Available at: http://www.engadget.com/
[8] Originally there are 34 categories. We checked similar cat egories and made a new summary with 20 categories.
[9] Available at: http://www.ttc-project.eu/
[10] Available at: http://www.lina.univ-nantes.fr/?Reference-Term-Lists-of-TTC.html
[11] Available at: http://www.microsoft.com/translator/api.aspx
[12] Available at: http://ictclas.nlpir.org/
[13] Available at: http://tartarus.org/~martin/PorterStemmer/
[14] Available at: http://radimrehurek.com/gensim/index.html
[15] Available at: https://pypi.python.org/pypi/lda
[16] Available at: http://scikit-learn.org/stable/index.html

clustering. For example, V-measure is an entropy-based measure combined with two aspects of clustering, homogeneity and completeness. In particular, Rosenberg and Hirschberg (2007) define these two indexes for cluster assignment. To satisfy homogeneity criteria, a clustering must assign only those data points that are members of a single class to a single cluster. To satisfy completeness criteria, a clustering must assign all of those data points that are members of a single class to a single cluster. V-measure is the harmonic mean between homogeneity and completeness. A higher V-measure score means better clustering results. Formulation of V-measure is as follows (Rosenberg & Hirschberg, 2007):

$$V = 2 \times \frac{(homogeneity \times Completeness)}{(homogeneity + Completeness)}$$

*Silhouette Coefficient*

If labels are not known, evaluation needs to be performed based on the data. The Silhouette Coefficient (Rousseeuw, 1987) can be used for unsupervised kind of evaluation. The Silhouette Coefficient is also composed of two scores. Suppose $a$ is the mean distance between a sample and all other points in the same class, $b$ is the mean distance between a sample and all other points in the next nearest cluster. A higher Silhouette Coefficient score means better clustering results. Formulation of Silhouette Coefficient is as follows (Rousseeuw, 1987):

$$Sil = \frac{b - a}{max(a, b)}$$

**Experimental Results Analysis**

In this section, we compare performances between different methods based on V-measure (denote as '*V*' in the table) and Silhouette Coefficient (denote as '*Sil*' in the table). Results analyses are shown in following subsections separately.

*Parallel corpora*

Four representation methods are compared according to V-measure. '*Chinese Combined Corpora*' is the corpora made up by combining original Chinese documents with translated English documents. '*English Combined Corpora*' is the corpora made up by combining original English documents with translated Chinese documents. '*Original Combined Corpora*' is the corpora made up by combining original Chinese documents with original English documents. Firstly, Kruskal-Wallis nonparametric test is conducted via SPSS 19.0 [17]. V-measure values between methods are all statistically significantly different when clustering three different Corpora ($p$. Table 4 shows the best clustering results (the highest V-measure value) of each method.

As we can see, LDA and D2V behave best whose V-measure value reach maximum value in all the tasks. For VSM, the two best clustering results are obtained when using all disparate terms in the corpora to represent documents. Results of VSM using all disparate terms are shown in Table 5. LSI didn't have outstanding performances compared with other methods, sometimes even worse than classical VSM.

| Corpora / Methods | Chinese Combined Corpora | English Combined Corpora | Original Combined Corpora |
|---|---|---|---|
| VSM | 0.39 | 0.44 | 0.45 |
| LSI | 0.41 | 0.39 | 0.43 |
| LDA | **0.46** | **0.46** | **0.46** |
| D2V | **0.46** | **0.46** | **0.46** |

**Table 3 Best clustering results (the highest V-measure value) of each representation method**

| Corpora / Methods | Chinese Combined Corpora | English Combined Corpora | Original Combined Corpora |
|---|---|---|---|
| Dimension | 66652 | 40034 | 72709 |
| V | 0.39 | 0.44 | 0.45 |

**Table 4 Clustering result of VSM using all terms**

Figure 3 and 4 shows results of combined Chinese corpora and English corpora. Figure 5 shows results of combined original corpora. Corpora are all represented by VSM, LSI, LDA and D2V at dimensions from 10 to 200 respectively. According to V-measure, values of D2V are higher than those of VSM, LSI and LDA when the dimension is over 50. For VSM, results are all lower than those of D2V. But when we use over 100 terms to represent documents, it shows good performance which are close to LSI and even better than LDA. Results of LSI have little fluctuations at first but the line gets smooth when dimension is bigger than 80. Although LDA behaves best when the dimension is 10 at the first point and shows unexceptionable performance at some dimensions (Dimension Size=190 in Figure 3, Dimensions Size=150 in Figure 4, Dimension Size=150 and 180 in Figure 5), trend shows stabile totally. Additionally, VSM using all terms behaves better over the VSM with dimensions reducing.

Moreover, Figures 3-5 reveal that clustering performance will not get improved when dimension number is higher than some certain values. When choosing representation dimensions, if diversity of data set is not big enough, we don't need to set dimension size too large. Like for this parallel corpora, labels of data set have 20 classes and we set dimensions from 10 to 200, while most methods will not get better performance when dimension size is bigger than 80. Except topic diversity, size of corpora might also make a difference on the clustering performance of methods. In order to investigate effects made by corpora size, blogs are selected randomly in quantities of 1226, 2452 and 3678 which are a quarter, a half and three quarters of corpora, respectively. We made the same clustering experiments on these corpora and compared with results which used the whole corpora. '*CH_1*' represents Chinese Combined Corpora in quantity of 1226, '*EN_1*' and '*CE_1*' represents

[17] Available at: http://www.spss.co.in/

English Combined Corpora and Original Combined Corpora in the same quantity, '*_2*', '*_3*' and '*_4*' represents the corpora is in quantity of 2452, 3678 and 4904 respectively. Figure 6, 7 and 8 shows the best clustering results (the highest V-measure value) of each method when clustering Chinese Combined Corpora, English Combined Corpora and Original Combined Corpora in different size of corpora.

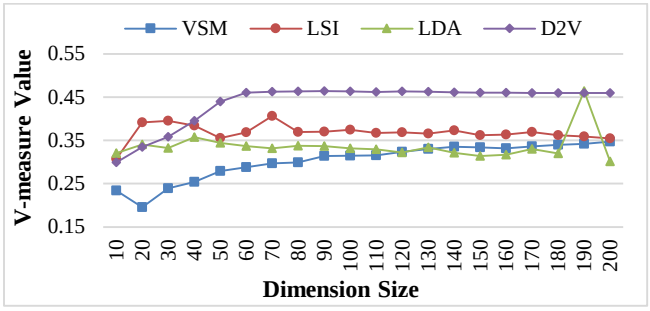



**Figure 3 Results of Chinese Combined Corpora**

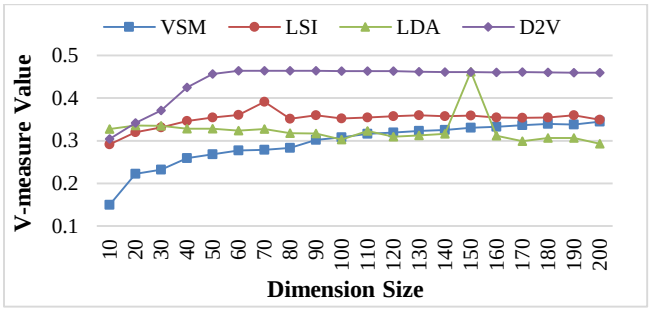



**Figure 4 Results of English Combined Corpora**

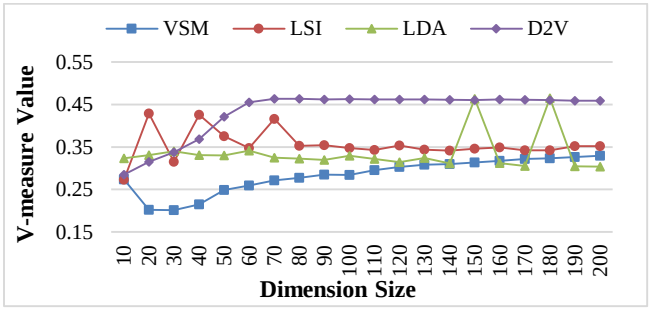



**Figure 5 Results of Original Combined Corpora**

As we can see, when clustering corpora in different size, methods perform differently. For VSM, V-measure value shows a descending trend with the increasement of corpora size, D2V shows the similar situation but V-measure values are all higher than the other three methods. V-measure values of LSI and LDA will rise up from some points, corpora size will definitely have affections on clustering performance of these two methods, especially for LDA, which reaches the highest V-measure values when clustering the whole corpora. Although, these findings can be observed from figures, due to the limitation of dataset, effects of corpora size didn't show clearly, datasets in more orders of magnitude are required for further study.

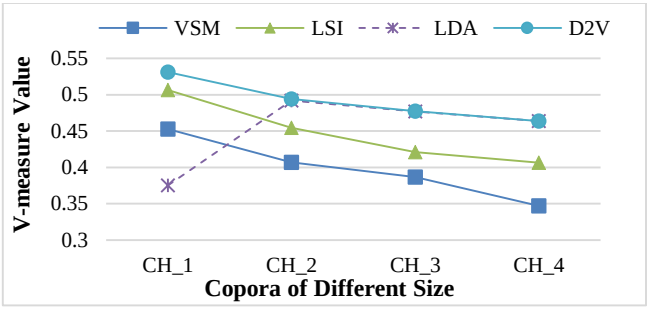



**Figure 6 Results of Chinese Combined Corpora in Different Size**

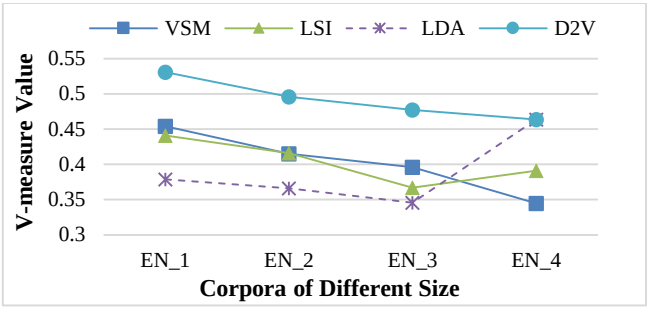



**Figure 7 Results of English Combined Corpora in Different Size**

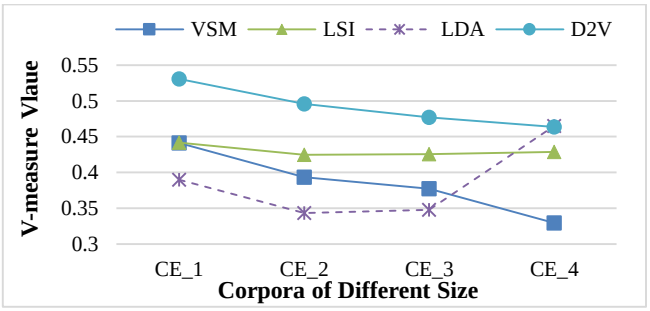



**Figure 8 Results of Original Combined Corpora in Different Size**

*Comparable corpora*

Four different methods are compared according to the Silhouette Coefficient values. Column '*wind_ch*' denotes corpora made by combining Chinese documents with translated English documents on wind energy topic. Column '*wind_en*' represents corpora made by combining English documents with translated Chinese documents on wind energy topic. Column '*mobile_ch*' and column '*mobile_en*' denote the mobile technology topic corpora which processed by the same way as '*wind_ch*' and '*wind_en*' do. Firstly, Kruskal-Wallis nonparametric test is also conducted and Silhouette Coefficient values between three methods except VSM are all statistically significantly different when clustering corpora on wind energy topic ($p$ and mobile technology topic($p$ for both languages. Table 6 shows the best clustering results (the highest Silhouette Coefficient value) of each method. Results of VSM using all terms are shown in Table 7. Figure 9-12 show results of combined

corpora which are represented by LSI, LDA and D2V respectively

| Corpora / Evaluation | wind_ch | wind_en | mobile_ch | mobile_en |
|---|---|---|---|---|
| VSM | 0.16 | 0.14 | 0.10 | 0.11 |
| LSI | 0.64 | 0.63 | 0.65 | 0.69 |
| LDA | **0.66** | **0.69** | **0.88** | **0.84** |
| D2V | 0.40 | 0.42 | 0.67 | 0.58 |

**Table 5 Best clustering result (the highest Silhouette Coefficient value) of each representation method**

| Corpora / Evaluation | wind_ch | wind_en | mobile_ch | mobile_en |
|---|---|---|---|---|
| Dimension | 30017 | 19388 | 24766 | 18194 |
| Sil | 0.16 | 0.14 | 0.10 | 0.11 |

**Table 6 Clustering results of VSM using all terms**

According to Table 6, values of VSM model are obviously lower than LSI, LDA and D2V. When looking at Figure 9-12, values of VSM model are still lower than most results of LSI, LDA but higher than some of D2V. Moreover, results of LDA are higher among the other two models and D2V behaves the worst. For this comparable corpora, the amount of data set is not large enough for D2V to play its leading role when representing large-scale documents. Oppositely, LDA will have better performance when doing this task. For LSI, it shows different performance on two topics. When it comes to wind energy topic, Silhouette Coefficient values of LSI are close to LDA. But when it comes to mobile technology topic, LSI behaves moderately. Moreover, with dimensions increasing, Silhouette Coefficient get reduced.

From all experimental results, we found D2V performs best in parallel corpora clustering but performs worst in comparable corpora clustering. The size of data set might lead to this contrast for the comparable corpora are too small for good training. When increasing size of parallel corpora, performance of VSM and D2V all get worse while LSI and LDA trend to get improved. However, the size of parallel corpora ranges from 1000 to 4000, and a larger amount of data changes will better illustrate effects on clustering performance caused by size according to different methods.

LDA tends to have better results when clustering comparable corpora and reaches to its best performance at some dimensions when clustering parallel corpora. It indicates that number of topics is important to determine for better representation performance and among the four methods, LDA has better results when clustering a small scale corpora. Generally, LSI model shows moderate performance and when clustering comparable corpora of two different topics, it performs differently. Moreover, with dimensions increasing, lines of LSI fluctuate at first and trends to stable after some dimensions which shows that dimension determination is also important when using LSI. Although, performance of VSM can be improved when increasing the number of terms used for representation. VSM using all separate terms are still not better than other methods.

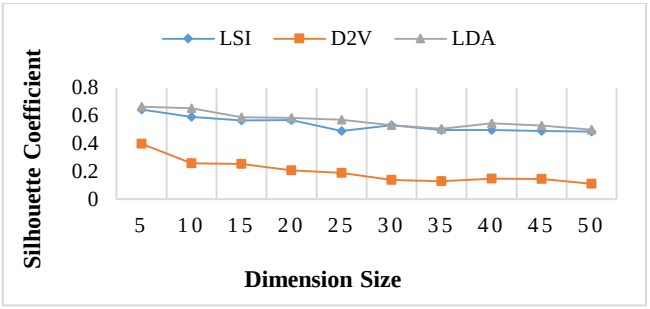



**Figure 9 Silhouette Coefficient of Chinese corpora on wind energy topic represented by models**

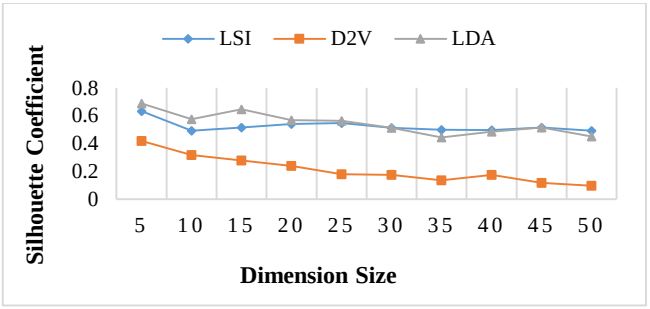



**Figure 10 Silhouette Coefficient of English corpora on wind energy topic represented by models**

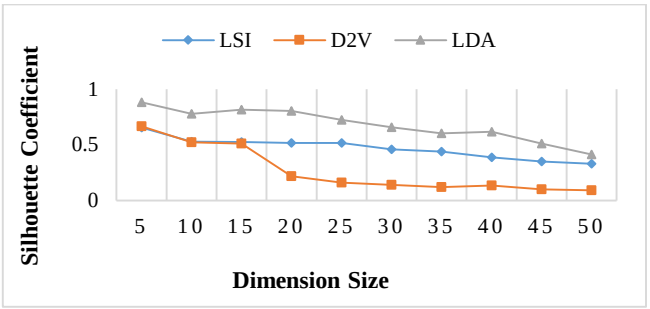
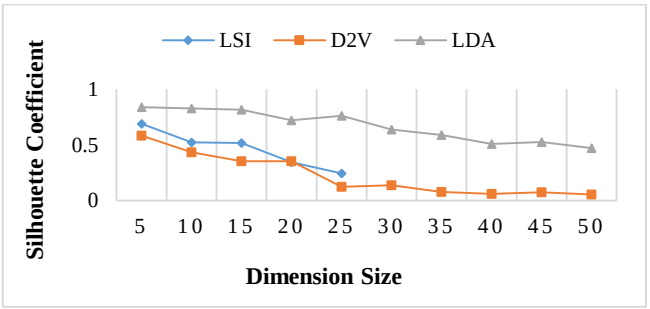



**Figure 11 Silhouette Coefficient of Chinese corpora on mobile technology topic represented by models**




Note: There are no representation results based on LSI when dimension size is bigger than 25.

**Figure 12 Silhouette Coefficient of English corpora on mobile technology topic represented by models**

Four methods behave differently on parallel and comparable corpora. When choosing method for clustering, corpora characteristics should be considered. In this paper, no matter using what method, dimension of representation will affect

clustering quality and each method needs to determine dimensions for representation. Dimensions increasement will not certainly bring the improvement of performance, it also depends on corpora size, topic diversity, etc.

## CONCLUSION

In this paper, four different document representation methods are compared in task of bilingual corpora clustering. We found that representation methods should be chosen according to corpora characteristics to have better clustering performance. For each method, VSM using all disparate terms shows best performance compared with VSM using dimension reduction. Performance of LSI will be affected by corpora itself and the concept space built through statistical dimension reduction technique doesn't work when optimal rank is bigger than some values. LDA behaves better when clustering documents in small size of comparable corpora (hundred magnitude) while D2V behaves better for large documents set of parallel corpora (thousand magnitude). But performance have different trends when changing the corpora size. To sum up, corpora size and topic diversity of corpora should be taken into considerations while languages of corpora don't distinguish methods from each other. How to choose representation method and determine its dimension number based on corpora characteristics is more critical.

Our experiment is a preliminary exploration to detect performance of four document representation methods in bilingual documents clustering. More works are needed to be done. We can use data set in different size and domain to discover the performance of D2V and LDA. Also how to choose the evaluation metrics is another important problem for in this kind of experiment. More clustering algorithm can be used in the future works to test our conclusion on behaviors between different methods. Moreover, further study about the clustering strategies on translating steps also needs to be done.

## ACKNOWLEDGMENTS

This work is supported by Major Projects of National Social Science Fund (13&ZD174), National Social Science Fund Project (No.14BTQ033).

## REFERENCES

Anandkumar, A., Liu, Y.-k., Hsu, D. J., Foster, D. P., & Kakade, S. M. (2012). A spectral algorithm for latent dirichlet allocation. *Proceedings of the Advances in Neural Information Processing Systems* (pp. 917-925).

Ayadi, R., Maraoui, M., & Zrigui, M. (2015). LDA and LSI as a Dimensionality Reduction Method in Arabic Document Classification *Information and Software Technologies* (pp. 491-502): Springer.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval* (Vol. 463): ACM press New York.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 35*(8), 1798-1828.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research, 3*, 993-1022.

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., et al. (2011). Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PloS one, 6*(3), e18029.

Boyd-Graber, J., & Blei, D. M. (2009). Multilingual topic models for unaligned text. *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence* (pp. 75-82).

Budzise-Weaver, T., Chen, J., & Mitchell, M. (2012). Collaboration and crowdsourcing: the cases of multilingual digital libraries. *The Electronic Library, 30*(2), 220-232.

Chen, H., Martin, B., Daimon, C. M., & Maudsley, S. (2013). Effective use of latent semantic indexing and computational linguistics in biological and biomedical applications. *Front Physiol, 4*(8), 1-6.

Collier, N., Doan, S., Kawazoe, A., Goodwin, R. M., Conway, M., Tateno, Y., et al. (2008). BioCaster: detecting public health rumors with a Web-based text mining system. *Bioinformatics, 24*(24), 2940-2941.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of The American Society for Information Science, 41*(6), 391-407.

Diekema, A. R. (2012). Multilinguality in the digital library: a review. *The Electronic Library, 30*(2), 165-181.

Evans, D. K., Klavans, J. L., & McKeown, K. R. (2004). Columbia newsblaster: multilingual news summarization on the Web. *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL) 2004* (pp. 1-4).

Freifeld, C. C., Mandl, K. D., Reis, B. Y., & Brownstein, J. S. (2008). HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. *Journal of the American Medical Informatics Association, 15*(2), 150-157.

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *science, 315*(5814), 972-976.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences* (pp. 5228-5235).

Guan, H., Zhou, J., Xiao, B., Guo, M., & Yang, T. (2013). Fast dimension reduction for document classification based on imprecise spectrum analysis. *Information Sciences, 222*, 147-162.

Hofmann, T. (1999). Probabilistic latent semantic indexing. *Proceedings of the 22nd annual international ACM SIGIR*

*conference on Research and development in information retrieval* (pp. 50-57).

Huang, H.-H., & Kuo, Y.-H. (2010). Cross-lingual document representation and semantic similarity measure: a fuzzy set and rough set based approach. *Fuzzy Systems, IEEE Transactions on, 18*(6), 1098-1111.

Jiang, C., Coenen, F., Sanderson, R., & Zito, M. (2010). Text classification using graph mining-based feature extraction. *Knowledge-Based Systems, 23*(4), 302-308.

Keikha, M., Khonsari, A., & Oroumchian, F. (2009). Rich document representation and classification: An analysis. *Knowledge-Based Systems, 22*(1), 67-71.

Kim, H., Ren, X., Sun, Y., Wang, C., & Han, J. (2013). Semantic frame-based document representation for comparable corpora. *Proceedings of the Data Mining (ICDM), 2013 IEEE 13th International Conference on* (pp. 350-359).

La Fleur, M., & Renström, F. (2015). Conceptual Indexing using Latent Semantic Indexing: A Case Study.

Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*.

Lesk, M. (2004). *Understanding Digital Libraries, Second Edition (The Morgan Kaufmann Series in Multimedia and Information Systems)*: Morgan Kaufmann Publishers Inc.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Proceedings of the Advances in neural information processing systems* (pp. 3111-3119).

Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., & McCallum, A. (2009). Polylingual topic models. *Proceedings of the 2009 Empirical Methods in Natural Language* (pp. 880-889).

Montalvo, S., Fresno, V., & Martínez, R. (2012). NESM: A named entity based proximity measure for multilingual news clustering. *Procesamiento del lenguaje natural, 48*, 81-88.

Niu, L.-Q., & Dai, X.-Y. (2015). Topic2Vec: Learning Distributed Representations of Topics. *arXiv preprint arXiv:1506.08422*.

Petrelli, D., & Clough, P. (2012). Analysing user's queries for cross-language image retrieval from digital library collections. *The Electronic Library, 30*(2), 197-219.

Rosenberg, A., & Hirschberg, J. (2007). V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. *Proceedings of the EMNLP-CoNLL* (pp. 410-420).

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics, 20*, 53-65.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management, 24*(5), 513-523.

Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613-620.

Shafiei, M., Wang, S., Zhang, R., Milios, E., Tang, B., Tougas, J., et al. (2007). Document representation and dimension reduction for text clustering. *Proceedings of the Data Engineering Workshop, 2007 IEEE 23rd International Conference on* (pp. 770-779).

Sidorov, G., Gelbukh, A., Gómez-Adorno, H., & Pinto, D. (2014). Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas, 18*(3), 491-504.

Taddy, M. (2015). Document Classification by Inversion of Distributed Language Representations. *arXiv preprint arXiv:1504.07295*.

Tang, G., Xia, Y., Zhang, M., Li, H., & Zheng, F. (2011). CLGVSM: Adapting Generalized Vector Space Model to Cross-lingual Document Clustering. *Proceedings of the IJCNLP* (pp. 580-588).

Vulić, I., De Smet, W., Tang, J., & Moens, M.-F. (2015). Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Information Processing & Management, 51*(1), 111-147.

Wang, K., Zhang, J., Li, D., Zhang, X., & Guo, T. (2008). Adaptive affinity propagation clustering. *arXiv preprint arXiv:0805.1096*.

Wei, C.-P., Yang, C. C., & Lin, C.-M. (2008). A Latent Semantic Indexing-based approach to multilingual document clustering. *Decision Support Systems, 45*(3), 606-620.

Wu, D., He, D., & Luo, B. (2012). Multilingual needs and expectations in digital libraries: a survey of academic users with different languages. *The Electronic Library, 30*(2), 182-197.

Yang, C., & Li, K. (2004). Cross-lingual information retrieval: The challenge in multilingual digital libraries. *Design and Usability of Digital Libraries: Case Studies in the Asia Pacific, Idea Group, Inc*.

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. *Proceedings of the ICML* (pp. 412-420).

Yetisgen-Yildiz, M., & Pratt, W. (2005). The effect of feature representation on MEDLINE document classification. *Proceedings of the AMIA*.

Zeng, W. (2012). Exploration and study of multilingual thesauri automation construction for digital libraries in China. *The Electronic Library, 30*(2), 233-247.