CLGVSM: Adapting Generalized Vector Space Model to Cross-lingual Document Clustering

Guoyu Tang[1,2], Yunqing Xia[1], Min Zhang[2], Haizhou Li[2],Fang Zheng[1]
[1]Dept. of Comp. Sci. & Tech., Tsinghua University, Beijing 100084, China
sweetyuer@gmail.com,{yqxia,fzheng}@tsinghua.edu.cn
[2]Institute for Infocomm Research, A-STAR, Singapore
{mzhang, hli}@i2r.a-star.edu.sg

Abstract

Cross-lingual document clustering (CLDC) is the task to automatically organize a large collection of cross-lingual documents into groups considering content or topic. Different from the traditional hard matching strategy, this paper extends traditional generalized vector space model (GVSM) to handle cross-lingual cases, referred to as CLGVSM, by incorporating cross-lingual word similarity measures. With this model, we further compare different word similarity measures in cross-lingual document clustering. To select cross-lingual features effectively, we also propose a *soft-matching* based feature selection method in CLGVSM. Experimental results on benchmarking data set show that (1) the proposed CLGVSM is very effective for cross-document clustering, outperforming the two strong baselines vector space model (VSM) and latent semantic analysis (LSA) significantly; and (2) the new feature selection method can further improve CLGVSM.

1 Introduction

The globalization of business environment urges organizations to maintain documents in different language. Obviously, organizations and research communities nowadays encounter the challenge of cross-lingual document clustering (CLDC). Document clustering seeks to automatically organize a large collection of documents into groups of similar documents. Various document clustering technologies have been proposed to deal with monolingual documents.

The classical solution to monolingual document clustering is vector space model (VSM), which explores bag of words (BOW) to construct feature space. Each document is converted to a VSM vector. Serious problem occurs when words are matched to the features using the hard matching strategy. For example, when the word *coast* is selected as a feature, word *seashore* will not contribute to hard matching unless it is also selected as a feature.

Different semantic document representation models have been proposed to address the shortcoming of VSM. Some semantic document representing models such as Latent Semantic Analysis (LSA) (Landauer et al., 1998) and Latent Dirichlet Allocation (Blei et al., 2003) implicitly capture statistical semantics by mapping documents to a lower dimension space. Other models such as Generalized Vector Space Model (GVSM) (Wang et al., 1985) extract statistical semantics in an explicit way by directly estimating measures of correlations between words.

The above models are designed for monolingual document sets and cannot be applied to cross-lingual scenario unless a bridge is created to connect cross-lingual features. Carbonell et al. (1997) use semantic models in parallel corpus. As features are selected from a common parallel/comparable corpus, which is usually different from the test cross-lingual document, over-fitting problem inevitably happens.

Other researchers propose to translate features or documents with bilingual dictionary or machine translation tools (Mathieu et al., 2004).However, ambiguity happens constantly and it is difficult to determine translation of a word. Meanwhile, if one translation of a word is selected as feature, the hard matching problem still occurs and becomes more serious.

In this paper, we extend the monolingual GVSM to handle cross-lingual cases, referred to as Cross-lingual GVSM (CLGVSM). Besides term correlation, we make use of word similarity in CLGVSM. For the cross-lingual we use statistical word similarity measure with the parallel corpus. We further improve cross-lingual word

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 580–588,
Chiang Mai, Thailand, November 8 – 13, 2011.

similarity by incorporating dictionary or translation probability. Experimental results show that the best result is achieved when combining the Second Order Co-occurrence Pointwise Mutual Information (SOCPMI) measure on the test dataset and translation probability in development dataset.

Selecting cross-lingual features is a key issue in cross-lingual document clustering. In this work, we propose a *soft-matching* based feature selection method in CLGVSM. In the new feature selection method, most representative terms are selected in semantic space according to Soft Term Frequency and Soft Document Frequency. In this way, a non-feature word can improve weight of the semantically similar features and make contribution to document clustering. Experimental results show that CLGVSM outperforms both LSA and VSM significantly with the help of proper word similarity measure.

The rest of this paper is organized as follows. In section 2, related work is surveyed. In section 3, the CLGVSM model is discussed. Experimental results as well as discussion are presented in Section 4. We conclude this paper in Section 5.

2 Related work

2.1 Monolingual Document Representation Models

The most commonly used model for document representation is the vector space model (VSM). It is assumed in VSM that terms are independent of each other and thus any semantic relations between them are ignored. Proposed by Landauer et al. (1998), LSA seeks to decompose the term-document matrix using singular value decomposition, in which each feature is a linear combination of all words.

Proposed by Wang et al. (1985) and further improved by Farahat and Kamel (2010), GVSM is proved an effective document representation model to address limitation of VSM. The model estimates similarity between documents based on how much their terms are related. Wang et al. (1985) pointed out orthonormal basis in VSM and proposed a new model to remove the assumption. Farahat and Kamel (2010) improved GVSM by developing better estimation of term correlation and applying dimension reduction techniques in a semantic space. Results show that the improved GVSM is advantageous over other representation models such as LSA.

Other document representation models are based on lexical ontologies such as WordNet, to represent documents in the concept space (Hotho et al., 2003). Similar representation models are seeks to exploit knowledge within an encyclopedia. Explicit Semantic Analysis (Cimiano et al., 2009) is a famous model that represents words as vectors in a space of concepts represented by articles from Wikipedia.

Most of those semantic models are designed for monolingual document sets, and cannot be used in cross-lingual scenario directly.

2.2 Cross-lingual Document Clustering

The difficulty of CLDC is how to deal with cross-language issue. The straightforward solution is document translation. In TDT3[1], four systems attempted to used Machine Translation systems (Leek et al., 1999). The results show that using a machine translation tool leads to around 50% performance loss, compared with monolingual topic tracking. This ascribed mainly to the poor accuracy of machine translation systems.

Dictionary and corpus are two popular ways to get cross-language information. Some researches (Evans and Klavans, 2003) use dictionary to translate documents. Others (Mathieu et al., 2004) use dictionary to translate features or keywords. But it is hard to select proper translation of ambiguous words. Mathieu et al. (2004) use bilingual dictionaries to translate named entities and keywords and modify the cosine similarity formula to calculate similarity between bilingual documents. Pouliquen et.al (2004) rely on a multilingual thesaurus called Eurovoc to create cross-lingual article vectors.

Wei et al. (2008) use LSA to construct a multilingual semantic space onto which words and document in either language can be mapped and dimensions are reduced again according to documents to be cluster. Yogatama and Tanaka-Ishii (2009) use propagation algorithm to merge multilingual spaces from comparable corpus and spectral method to cluster documents. Li et al. (2007) use Kernel Canonical Correlation Analysis, a method of finding the maximally correlated projections of documents in two languages for cross-language Japanese-English patent retrieval and document classification. Unlike document classification, document clustering lacks training data. So semantic space is constructed from the parallel/comparable corpus, and the dimensions are selected on the basis of their importance in parallel/comparable corpus, which is usually different from the target multilingual documents.

[1]http://www.itl.nist.gov/iad/mig//tests/tdt/1999/index.html

In this work, our proposed CLGVSM use semantic similarity to solve word matching problem caused by different languages. Semantic space is constructed based on word similarity and in our feature selection method, features are select on the basis of their importance in documents to be clustered.

2.3 Cross-lingual Word Similarity

In both monolingual GVSM (Wang et al, 1985) and improved GVSM (Farahat and Kamel, 2010), correlation, correlation between words is computed in documents to be clustered and correlation of the best performance in document clustering is calculate as covariance of words with the assumption that words are random variables with Gaussian distributions. In this work we use word similarity which is calculated as cosine similarity of term vector covariance. This measure can be called as COV measure.

But this similarities method is estimated in test documents which lacks cross-lingual information.

Various measures for cross-lingual word semantic similarity have been proposed to explore statistical techniques and semantic network.

Research works propose to use *WordNet* by Resnik (1999) to measure similarity between English words. Liu and Li (2002) adopt HowNet calculate word similarity in machine translation. Xia et al. (2011) propose to explore cross-lingual word similarity by observing concept definition provided by HowNet.

Corpus-based measures for semantic similarity are found more interesting. The classical method is Pointwise Mutual Information (PMI) (Church and Hanks, 1990). Many researches are based on PMI, such as PMI-IR (Turney, 2001) and Second Order Co-occurrence PMI (SOCPMI) (Islam and Inkpen, 2006). SOCPMI is proved better than PMI-IR and some other similarity measures (Islam and Inkpen, 2006).

In this work, we implement three representative measures: HowNet-based measure (Xia et al., 2011), SOCPMI measure (Islam and Inkpen, 2006) and COV measure (Farahat and Kamel, 2010).

3 Cross-Lingual Generalized VSM

3.1 Generalized VSM

Let $D = \{d_j; j = 1, ...N\}$ be a set of N documents which contain *M* terms, X be a $M \times N$ matrix whose element x_{ij} represents the weight of term t_i in document d_j. GVSM (Wang et al, 1985) estimates correlation between documents based on how their terms are related. The GVSM model presents document in a non-orthogonal space and similarity between two documents is calculated as follows.

$$Sim^{\mathrm{GVSM}}(d_1, d_2) = \frac{d_1^T G d_2}{\sqrt{d_1^T G d_1}\sqrt{d_2^T G d_2}}, \quad (1)$$

where G is an $M \times M$ association matrix which represents correlations between terms and is usually computed as inner-products of term vectors in some space. An example of 3×3 association matrix is given as follows.

$$\begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \end{array} \begin{array}{c} \begin{array}{ccc} t_1 & t_2 & t_3 \end{array} \\ \begin{pmatrix} 0.1 & 0.2 & 0.3 \\ 0.4 & 0.5 & 0.3 \\ 0.4 & 0.2 & 1 \end{pmatrix}_{3\times 3}, \end{array}$$

where every row and column represents a term, respectively.

In tradition GVSM (Wang et al, 1985), terms are represented as vectors in the dual space of documents and the association measures between terms are calculated as the cosine of the angle between their vectors in the dual space. Accordingly,G can be calculated as:

$$G = L^{-1/2} X X^T L^{-1/2} \quad (2)$$

where L is a diagonal matrix whose elements are the lengths of term vectors in the dual space.

And in improved GVSM (Farahat and Kamel, 2010), the best G which is covariance matrix of terms is calculated as

$$G_{\mathrm{COV}} = \frac{1}{n_c - 1} Q H Q^T \quad (4)$$

where Q is random sample of $X$and

$$H = 1 - \frac{1}{n_c} e e^T \quad (5)$$

G_{COV} maps uncorrelated terms to near-orthogonal directions and negatively correlated terms to opposite directions in the semantic space, while traditional G maps both uncorrelated terms and negatively correlated terms to near-orthogonal directions. Thus we use cosine similarities between term vectors in case of G_{COV} as one of our similarity measures. As the GVSM models proposed, G is estimated from documents to be clustered; they cannot acquire cross-lingual information and cannot deal with cross-lingual issues directly. So we extend GVSM to cross-lingual GVSM by using cross-lingual word similarity measures in section 3.2

3.2 Cross-Lingual GVSM

Note that before (Farahat and Kamel, 2010); term correlation was used in GVSM to construct association matrixes G with the inner-product of term vectors in some semantic space. Length of term vectors quantifies how important of the term in the documents. Thus correlation of terms is not totally the same with similarity of terms. But it is difficult for term correlation to adapt into cross-lingual case. Term vectors generated from test data lack cross-lingual information. Carbonell et.al (1997), generate term vectors from development data. Noise occurs because term importance differs in two dataset.

For those reasons mentioned above, we choose word similarity instead of term correlation to construct word association matrix. And the other advantage of using word similarity is that we can ignore noisy similarity values which contribute little to document similarity calculation. In that case, the associate matrix becomes sparse and computational time can be saved. Therefore in this work, we explore the following several word similarity measures in constructing word association and setup a similarity threshold.

Knowledge-based Similarity Measures

We choose to use cross-lingual word similarity based on HowNet (Xia et al., 2011) which makes use of concept graph in HowNet. HowNet is concept based and the atom unit is sememe, so similarity between words is actually reflected by the sememes they carry. The key idea of cross-lingual word similarity calculation is to locate bilingually definitions for given words so that the language barrier is overcome. For details please refer to Xia et al. (2011).

Statistical Similarity Measures

Statistical similarity actually reflects conceptual relevance between words as it considers merely word co-occurrences within a corpus. We evaluate two statistical similarity measures: SOCPMI and COV in this work.

SOCPMI was proposed by Islam and Inkpen (2006), in which PMI is applied to rank the neighboring words with in a corpus. The measure is proved accurate because it calculates relevance between two words that do not co-occur frequently. Note that the original SOCPMI measure is designed to deal with monolingual word similarity. We extend this measure in this work to calculate similarity between cross-lingual words. The goal is achieved by counting neighboring words with the same language in the corpus and computing the cross-lingual PMI in a parallel corpus.

As we mentioned above， associate matrix G constructed by covariance of term vectors achieve the best performance in monolingual document clustering. In this paper we use the cosine similarity instead of inner-product in COV similarity measure.

The two above word similarity measures both need to be calculated in a cross-lingual parallel/comparable corpus.

Combining Similarity with Dictionary or Translation probability

The statistical word similarity measures are developed with general development corpus. However, we believe word co-occurrence in the test documents is also very useful. Thus we improve cross-lingual word similarity in the following way: statistical monolingual word similarity is computed from test documents first and a dictionary or translation probability as a bridge is used to get cross-lingual word similarity.

When using dictionary as bridge, assuming word w_i which has a translation list T_i^{T}={ $t_{i_k}^{\mathrm{T}}$; $k = 1...N_i^{\mathrm{T}}$} and word $w_j$which has a translation list T_j^{T}={$t_{j_k}^{\mathrm{T}}$;$k = 1...N_j^{\mathrm{T}}$}, we choose the highest value between similarities of $w_i$and each words in $T_j^{\mathrm{T}}$and similarities of w_j and each words in T_i^{T}as similarity between w_i and w_j.

When using translation probability as bridge, assuming word w_i which has a translation probability list P_i^{T}={$t_{i_k}^{\mathrm{T}}$: p_{i_k};$k = 1...N_i^{\mathrm{W}}$} and word w_j which has a translation list P_j^{T} ={ $t_{i_k}^{\mathrm{T}}$: p_{j_k} ; $k = 1...N_j^{\mathrm{W}}$}, the similarity of $w_i$and w_j can be calculated as follows:

$$Sim^{\mathrm{PR}}(w_i \to w_j) = \sum_k p_{i_k} \times Sim^{\mathrm{Mono}}(w_j, t_{i_k}^{\mathrm{W}}) \quad (6)$$

$$Sim^{\mathrm{PR}}(w_j \to w_i) = \sum_k p_{j_k} \times Sim^{\mathrm{Mono}}(w_j, t_{i_k}^{\mathrm{W}}) \quad (7)$$

$$Sim^{\mathrm{PR}}(w_i, w_j) = \max\{Sim^{\mathrm{PR}}(w_i \to w_j), Sim^{\mathrm{PR}}(w_i \leftarrow w_j)\} \quad (8)$$

where Sim^{Mono}returns monolingual word similarity.

3.3 Feature selection for GVSM

Term Importance based on GVSM

In the VSM model, importance of a term is proportional to Term Frequency (*TF*) in a single document, and inversed proportional to Document Frequency (*DF*) in a document set. We argue that the theory can be improved.

Consider a document that contains term *criminal* for 3 times and term *imprisonment* for 10 time. We find that term *criminal* is still very important though its *TF* is low. This is because term *imprisonment* is semantically similar to *criminal* and it appears many times. If the classical term matching method is named *hard matching*, we can call our method *soft matching*. We incorporate the *soft-matching* idea to the GVSM model.

In the GVSM model, importance of a term can be reflected by the following statistics.

(1) Soft Term Frequency

The soft term frequency (TF^s) considers the term and the semantically similar terms. Given term t and document d, we first retrieve the semantically similar terms $T=\{t_i\}_{i=1...M^T}$ from document d. We define the soft term frequency (TF^s) as follows.

$$TF^s(t,d) = \sum_i TF_i \times Sim(t,t_i) \quad (9)$$

where TF_i denotes term frequency of term t_i within document d.

Note that the soft term frequency is calculated within a single document.

(2) Soft Document Frequency

The soft document frequency (DF^s) considers not only number of documents that contain the term, but also the number of documents that contains the semantically similar terms. Given term t and document set $D = \{d_j\}_{j=1...N}$. Let $d_j = \{t_{i,j}\}_{i=1...M^D}$ denote terms within document d_j. We define the soft document frequency (DF^s) as follows.

$$DF^s(t) = \sum_{d_j \in D} \max_i \{Sim(t,t_{i,j}))\} \quad (10)$$

Note that the soft document frequency is calculated within a document set. We use maximum instead of summation in order to reduce the effects of word pairs with low similarities.

Feature Selection in GVSM

With GVSM-based term importance, features can be selected appropriately. Following the idea of TF-IDF, we refine inverse document frequency as follows.

$$IDF^s(t) = log\Big(\frac{N}{DF^s(t)}\Big), \quad (11)$$

where N denotes number of documents.

The soft weighting equation of term t in document d is as follows.

$$w^s(t,d) = TF^s(t,d) IDF^s(t) \quad (12)$$

If we select features for a document based solely on term weight, some semantically similar terms might be selected because they hold close weights. This results in fewer representatives feature set. Before feature selection we update term TF^s as follows:

1) Setup an initial term list.
2) Sort terms in the list according to their TF^s in descending order.
3) Move the first term t_0 into the tentative feature list.
4) For each remaining term in the list, update its TF^x using the following equation.

$$TF^s_{g+1}(t,d) = TF^s_g(t,d) - \sum_{t_k \in d} (Sim(t,t_k) Sim(t_0,t_k) TF^s(t_k,d), \quad (13)$$

where t_k denotes a term in document d, and g the iteration round number.

5) Delete terms a weight less than 0 in the list.
6) Repeat step 2) ~ 5) until the term list becomes empty.

Once the tentative feature list is obtained, we then calculate weight for each candidate features using Eq (12). The features of each document are then ranked according to the weight and joined together to represent document set.

Document representation in GVSM

With the feature set available, we describe how a document is represented with the features. Let $F = \{f_i\}_{i=1...M^F}$ denote the feature set, and $T = \{t_j\}_{j=1...M^D}$ denote terms within document d. We now try to map d to the feature space.

For each feature, it should be mapped no matter whether it appears in the document set. But in order to avoid redundant information, we map only one term with each feature.

So for feature list F sort by TF^s in document d, the actually weight of feature f_i in document d is as follows:

First, retrieve the most similar term t^* which is not included in T^M, which stores terms that have been matched.

$$t^* = argmax_{t_i \in d, t_i \notin T_M} Sim(t, f_i) \quad (14)$$

Then we re-calculate weight w^a of term f_i as follows.

$$w^a(f_i,d) = TF(t^*) IDF^s(f_i) \quad (15)$$

Finally, put t^* into T^M and repeat until all features are matched once.

3.4 Document Clustering based on GVSM

With document similarity, we employ certain clustering algorithm to manage the cross-lingual documents with a few clusters. As clustering algorithm is not core of this work, we simply choose the classic document clustering algorithm, i.e., HAC (Hierarchical Agglomerative Clustering) algorithm (Voorhees, 1986). To measure cluster-cluster similarity, we adopt the group-average link algorithm (Voorhees, 1986). The merging procedure repeats until a desired number of clusters are obtained.

4 Evaluation

4.1 Setup

Development dataset: We randomly extract 1M parallel sentence pairs from LDC corpora (i.e., LDC2004E12, LDC2004T08, LDC2005T10, LDC2003E14, LDC2002E18m LDC2005T06, LDC2003E07 and LDC2004T07) as our development data to train the bilingual corpus-based term similarity and get translation probability.
Dictionary: Translation pairs are extracted from HowNet.
Translation Probability: We compute translation probability by Giza++ (Och and Ney, 2000) in development data.
Test dataset: Four datasets are tested in this paper.

Corpus	TDT41 (2002) (Topic#/Story#)	TDT42 (2003) (Topic#/Story#)
English	38/1270	33/617
Chinese	37/657	32/560
Common	40/1927	37/1177

Table1. Statistics on the twoTDT4 datasets.

Corpus	CLTC1 (Topic#/Story#)	CLTC2 (Topic#/Story#)
English	20/200	20/600
Chinese	20/200	20/600
Common	20/400	20/1200

Table2. Statistics on the two CLTC datasets.

TDT4 datasets

We first extract two datasets from the TDT4 evaluation dataset (see statistics in Table 1).

CLTC datasets

The second dataset is extracted from our own cross-lingual topic corpus (CLTC). The news articles are retrieved from Gigaword (English and Chinese), and the topics are labeled by human (see statistics in Table 2).

Evaluation criteria

We adopt the evaluation criteria proposed by Steinbach et al. (2000). The calculation starts from maximum F-measure of each cluster. Let A_i represent the set of articles managed in a system-generated cluster c_i, A_j the set of articles managed in a human-generated cluster c_j. F-measure of the system-generated cluster c_i is calculated as follows.

$$\begin{aligned} p_{i,j} &= \frac{|A_i \cap A_j|}{|A_j|} \quad p_i = \max_j\{p_{i,j}\} \\ r_{i,j} &= \frac{|A_i \cap A_j|}{|A_i|} \quad r_i = \max_j\{r_{i,j}\} \\ f_{i,j} &= \frac{2 \cdot p_{i,j} \cdot r_{i,j}}{p_{i,j} + r_{i,j}} \quad f_i = \max_j\{f_{i,j}\} \end{aligned} \tag{16}$$

where $p_{i,j}$, $r_{i,j}$ and $f_{i,j}$represent precision, recall and measure of cluster when compared with clusterc_j , respectively.

We also use relative F-measure to compare systems over all dataset which is used by Farahat and Kamel (2010).In this approach, the F measure for a particular data set are normalized relative to the best value obtained using different representation models when applying the same clustering algorithm to the same data set:

$$F_r = \frac{F}{\max_i\{F_i\}}, \tag{17}$$

where F_i denotes F-measure values obtained using different representation models.

The relative F measures are then averaged for different data sets.

4.2 Evaluation

Experiment 1: Different word similarity calculation measures

This experiment seeks to compare different cross-lingual word similarity (CLWS) measures. Seven CLWS measures are implemented:
HN: HowNet-based cross-lingual word similarity measure.
SOCPMI^DEV: SOCPMI similarity measure learned from development data.
SOCPMI&DIC: SOCPMI similarity measure calculated in test documents and dictionary as cross-lingual bridge.
SOCPMI&TranPro: SOCPMI similarity measure directly computed in test documents and translation probability on development set as cross-lingual bridge.
COV^DEV: COV similarity measure learned from development data.
COV&DIC: COV similarity measure computed in test documents and dictionary as cross-lingual bridge.

System / Dataset	HN	SOCPMI ^DEV	SOCPMI &DIC	SOCPMI &TranPro	COV ^DEV	COV &DIC	COV &TranPro
TDT41	0.783	0.880	0.854	0.892	0.824	0.868	0.907
TDT42	0.797	0.880	0.835	0.880	0.860	0.840	0.851
CLTC1	0.764	0.818	0.834	0.877	0.782	0.854	0.874
CLTC2	0.667	0.856	0.804	0.839	0.805	0.833	0.840

Table 3. Highest F-measure of CLDC systems with different CLWS measures.

System	HN	SOCPMI ^DEV	SOCPMI &DIC	SOCPMI &TranPro	COV ^DEV	COV &DIC	COV &TranPro
ARF	0.855	0.976	0.945	**0.991**	0.929	0.965	0.986

Table 4. Average of relative F-measure (ARF) of CLDC systems with different CLWS measures.

COV&TranPro: COV similarity measure in test documents and translation probability as cross-lingual bridge.

All the CLDC systems use HAC algorithm to do clustering documents. The thresholds of similarity measures in this paper is all set 0.4 based on our empirical study. Experiment results on four datasets are as Table 3. Table 4 computed from Table 3 shows the average of relative F-measure (ARF) over all data sets for different CLWS measures.

We can observe from Table 3 and Table 4 that the performance of HowNet is much worse than other systems in all dataset. We look into the intermediate results to check the reasons. We find semantic similarities between words computed based on HowNet are too high. For example, word similarity between *Federal Reserve* and *bank* is assigned 1 by HowNet. Error analysis shows that HowNet-based CLWS measure puts much emphasis upon the semantic property of given word rather than semantic itself. So it tends to assign bigger CLWS values to semantically similar word pairs, no matter how semantically relevant they are. This would obviously jeopardize document clustering. With such an observation, we conclude that HowNet-based CLWS measure is not suitable for document clustering.

We can also observe that systems with translation probability outperform those with dictionary when the same monolingual word similarity measures are used. For instance, SOCPMI&TranPro outperforms SOCPMI&DIC by 4.5% on average on relative F-measure and COV&TranPro outperforms COV&DIC by 1.9%. Two reasons are worth noting. First, dictionary extracted from HowNet have more OOV than translation probability computed from development corpus. Translation probability is more discriminative than dictionary when word is ambiguous. It tries to get word similarity from the most frequency translation.

Seen from Table 4 that SOCPMI&TranPro outperforms SOCPMI^DEV by 1.5% on average relative F-measure and COV&TranPro outperforms COV^DEV by 5.5% on average relative F-measure. As both systems use the same development data to get cross-lingual information and the different is that systems computed word similarity in test dataset take use of word occurrence information in test dataset so we can conclude that with combining word similarity in test dataset and translation probability can is useful in cross-lingual document clustering.

And over all seven systems, SOCPMI&TRAN achieves the best result on average, so we select SOCPMI&TRAN as our word similarity measures and next experiments both use this word similarity measure.

Experiment 2: Different feature selection vs. dimension reduction methods

This experiment aims to compare the proposed feature selection method and the existing ones**.** Three CLDC systems are implemented.

SFS: feature selection we proposed by TF^s-$IDF^s$and soft matching is used in GVSM.

HFS: feature selection by TF-IDF and hard matching is used in GVSM.

NFS: feature selection is not used, which equals to system SOCPMI&TranPro in Experiment 1.

System / Dataset	SFS	HFS	NFS
TDT41	0.900	0.903	0.892
TDT42	0.899	0.881	0.880
CLTC1	0.876	0.869	0.877
CLTC2	0.891	0.847	0.839

Table5. Highest F-measure of CLDC systems with/without feature selection.

System	SFS	HFS	NFS
F-measure	**0.998**	0.980	0.976

Table6. Average of relative F-measure of CLGVSM systems with/without feature selection.

Experiment results on four test datasets are given in Table 5. Table 6 computed from Table 5 shows the ARF over all data sets for different feature selection methods.

Seen from Table 6, system with feature selection we proposed using soft matching outperforms system with feature selection using hard matching by 1.8% on average relative F-measure It also outperforms system without feature selection by 2.2% on average relative F-measure.

This reveals that feature selection does improve GVSM. The reason why it works is that with TF^s and DF^s, it can select the most representative terms as feature set and with proper document representation method, documents are properly matched into feature space.

Experiment 3: Different document representation models

This experiment aims to compare CLGVSM with VSM and LSA. Three CLDC systems are implemented:

CLGVSM: Our system, which equals FS system in Experiment 2.

VSM: A baseline system that uses VSM to represent documents and cosine similarity to compute document systems. HowNet dictionary is used to match terms in different languages.

LSA: LSA uses dictionary to match terms in different languages and make use of LSA in test dataset. The number of LSA dimensions is set to 200,

Experiment results on two data sets are given in Table 7. Table 8 computed from Table 7 shows the ARF over all data sets for different document representation models.

System	CLGVSM	VSM	LSA
TDT41	0.900	0.877	0.885
TDT42	0.899	0.835	0.881
CLTC1	0.876	0.792	0.867
CLTC2	0.891	0.776	0.841

Table7. Highest F-measure of CLDCsystems with different document representation models.

System	CLGVSM	VSM	LSA
F-measure	**1**	0.920	0.974

Table8. Average of relative F-measure of CLDCsystems with different document representation models.

We can observe from Table 8 that CLGVSM outperforms VSM by 8.0% on average relative F-measure. It means CLGVSM improve cross-lingual document clustering by using SOCPMI^TRAN similarity measure. Observation shows that the word similarity measure makes significant contribution to document clustering. Using second order co-occurrence information of words, SOCPMI assigns word pair with higher PMI a higher similarity. This coincides perfectly with the real demand in word similarity measuring. For example, word similarity between 犯罪分子 (*criminal*) and *imprisonment* is assigned 0.49 by SOCPMI. When 犯罪分子 is chosen as a feature, document containing *imprisonment* holds a reasonable similarity with document containing 犯罪分子 even though they do not contain common word.

Results also show that CLGVSM outperforms LSA by 2.6% on average relative F-measure. It means CLGVSM is better than LSA in cross-lingual document clustering by using SOCPMI^TRAN similarity measure. The follow reason is worth noting. When dictionary is used to match words in LSA, the semantic relation between different translations of one term in one document is added and this brings much noise. While in CLGVSM, cross-lingual terms are soft matched by SOCPMI^TRAN term similarity.

5 Conclusion

In this paper, we extend monolingual generalized VSM (GVSM) to handle cross-lingual cases, referred to as CLGVSM, by incorporating cross-lingual word similarity measures.Under GVSM, we compare different word similarity measures in cross-lingual document clustering. We propose new feature selection method for CLGVSM and experiments show it improves document clustering. We also compare CLGVSM and other well-known document representation models such as VSM and LSA and experiments show it outperform both VSM and LSA significantly.

Three conclusions can be drawn in this paper. Firstly, HowNet-based word similarity method is less suitable for document clustering. Secondly, translation probability computed from a development dataset as a cross-lingual bridge performs better than HowNet dictionary. At last, combining word similarity in test dataset and translation probability in development dataset can help cross-lingual document clustering.

In the future, we will apply CLGVSM in more languages pairs and extend it in more than two languages. As GVSM represents document with semantic space, we can utilize GVSM to handle sparse data problem in short text clustering.

Acknowledgment
This work is partially supported by NSFC (60703051) and MOST (2009DFA12970). We thank the reviewers for the valuable comments.

References

D. M. Blei, A. Y. Ng, and M.I. Jordan. 2003. Latent dirichlet allocation. J. Machine Learning Research (3):993-1022.

K. W. Church and P. Hanks. 1990. Word association norms, mutual information, and lexicography. Computational Linguistics, 16(1):22-29.

P. Cimiano, A. Schultz, S. Sizov, P. Sorg, and S. Staab. Explicit vs. latent concept models for cross-language information retrieval. Proc. of IJCAI'09, 2009.

C. Corley and R. Mihalcea. 2005. Measuring the semantic similarity of texts, Proc. of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, p.13-18, June 30-30, 2005, Ann Arbor, Michigan

Z. Dong and Q. Dong. 2006. HowNet and the Computation of Meaning. World Scientific Publishing Co. Inc., River Edge, NJ, USA.

D. K. Evans and J. L. Klavans. 2003. A Platform for Multilingual News Summarization, Technical Report. Department of Computer Science, Columbia University.

A. K. Farahat, M. S.Kamel.2010. Statistical semantic for enhancing document clustering. Knowledge and Information Systems.

A. Hotho, S. Staab,G. Stumme. 2003. WordNet improves text document clustering. Proc. of SIGIR2003 semantic web workshop.ACM, New York, pp. 541-544.

A. Islam and D. Inkpen. 2006. Second order co-occurrence PMI for determining the semantic similarity of words. Proc. LREC'2006: 1033-1038

A. Islam and D. Inkpen. 2008 Semantic text similarity using corpus-based word similarity and string similarity, ACM Transactions on Knowledge Discovery from Data (TKDD) v.2 (2), pp.1-25, July 2008

T. K. Landauer and S. T. Domais. 1997. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. Psychological Review. 104(2):211-240.

T. Landauer, P. W. Foltz and D. Laham. 1998. Introduction to Latent Semantic Analysis. Discourse Processes 25: 259-284.

T. Leek, H. Jin, S. Sista, and R. Schwartz. 1999. The BBN cross-lingual topic detection and tracking system. Proc. of TDT' 1999.

Y. Li, J. Shawe-Tylor, 2007, Advanced learning algorithms for cross-language patent retrieval and classification. Information Processing and Management v.43(5), pp 1183-1199, September, 2007.

Q Liu, S Li. 2002.Word similarity computing base don Hownet. Computational Linguistics and Chinese Language Processing. (in Chinese)

B. Mathieu, R. Besancon and C. Fluhr. 2004. Multilingual Document Clusters Discovery. Proc. of RIAO'2004: 1-10.

F. J. Och, H. Ney. 2000. Improved Statistical Alignment Models. Proc. of the 38th Annual Meeting of the Association for Computational Linguistics: 440-447.

B. Pouliquen, R. Steinberger, C. Ignat, E. Käsper, I. Temnikova. 2004. Multilingual and cross-lingual news topic tracking. Proc. of COLING'2004:959-965.

P. Resnik. 1999. Semantic similarity in a taxonomy: An information based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research, V.11:95–130.

R. Rosenfeld. 1996. A maximum entropy approach to adaptive statistical language modeling. Computer speech and language. 10:187-228.

P. D. Turney. 2001. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. Proc. of ECML'2001: 491-502.

E. M. Voorhees. 1986. Implementing Agglomerative Hierarchic Clustering Algorithms for Use in Document Retrieval. Information Processing and Management, 22(6): 465-76.

C-P. Wei, C. C. Yang and C-M. Lin. 2008. A Latent Semantic Indexing Based Approach to Multilingual Document Clustering. Decision Support System. 45(3):606-620.

S. K. M. Wong, W. Ziarko, P. C. N. Wong. 1985 Generalized vector model in information retrieval. Proc. of the 8th ACM SIGIR:18-25

Y. Xia, T. Zhao, and P. Jin. 2011. Measuring Chinese-English Cross-lingual Word Similarity with HowNet and Parallel Corpus. Proc. of CICling'2011(II):221-233.

D. Yogatamaan, K.Tanaka. 2009. Multilingual Spectral Clustering Using Document Similarity Propagation. Proc. of EMNLP'2009: 871-879.