Available online at www.sciencedirect.com
ScienceDirect
Decision Support Systems 45 (2008) 606–620
Decision Support Systems
www.elsevier.com/locate/dss

A Latent Semantic Indexing-based approach to multilingual document clustering

Chih-Ping Wei [a,*], Christopher C. Yang [b], Chia-Min Lin [c]

[a] Institute of Technology Management, College of Technology Management, National Tsing Hua University, Hsinchu, Taiwan, ROC
[b] Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
[c] Openfind Information Technology Inc., 4F, No. 222, Sec. 2, Nan-Chang Rd., Taipei, Taiwan, ROC

Available online 1 August 2007

Abstract

The creation and deployment of knowledge repositories for managing, sharing, and reusing tacit knowledge within an organization has emerged as a prevalent approach in current knowledge management practices. A knowledge repository typically contains vast amounts of formal knowledge elements, which generally are available as documents. To facilitate users' navigation of documents within a knowledge repository, knowledge maps, often created by document clustering techniques, represent an appealing and promising approach. Various document clustering techniques have been proposed in the literature, but most deal with monolingual documents (i.e., written in the same language). However, as a result of increased globalization and advances in Internet technology, an organization often maintains documents in different languages in its knowledge repositories, which necessitates multilingual document clustering (MLDC) to create organizational knowledge maps. Motivated by the significance of this demand, this study designs a Latent Semantic Indexing (LSI)-based MLDC technique capable of generating knowledge maps (i.e., document clusters) from multilingual documents. The empirical evaluation results show that the proposed LSI-based MLDC technique achieves satisfactory clustering effectiveness, measured by both cluster recall and cluster precision, and is capable of maintaining a good balance between monolingual and cross-lingual clustering effectiveness when clustering a multilingual document corpus.

Keywords: Document management; Text mining; Document clustering; Multilingual document clustering; Multilingual knowledge management; Latent Semantic Indexing

1. Introduction

Among various knowledge management initiatives, the creation of knowledge repositories has emerged as a prevalent approach in current knowledge management practices [8,14,36,38]. Many knowledge repository-building endeavors have focused on managing tacit knowledge, which includes lessons learned and/or best practices. For example, Ford Motor Company maintains a "TGRW" repository for capturing events that facilitate (i.e., things gone right, TGR) or impede (i.e., things gone wrong, TGW) task accomplishment. Its TGRW repository enables Ford to establish records of events that must be addressed and monitored during future projects [23,31]. Similarly, several leading consulting firms, including Arthur Andersen, Ernst & Young, and Price Waterhouse, have developed best practices knowledge repositories to capture information and knowledge about the best way to do things [24]. Ernst & Young, for example, retains a knowledge repository of more than 5000 leading practices that cover categories such as

* Corresponding author. Tel.: +886 3 574 2219; fax: +886 3 574 5310.
E-mail addresses: cpwei@mx.nthu.edu.tw (C.-P. Wei), yang@se.cuhk.edu.hk (C.C. Yang), alucard.lin@gmail.com (C.-M. Lin).

0167-9236/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2007.07.008

finance, new business development, order management, supply chain management, and so forth.

Conceivably, a knowledge repository may contain vast amounts of formal knowledge elements, which generally are available as documents. Thus, effective management of this ever-increasing volume of documents within a knowledge repository is essential to knowledge sharing, reuse, and assimilation. To facilitate users' navigation of documents within a knowledge repository, knowledge maps, typically created by document clustering techniques, offer an appealing and promising approach [25].

Document clustering automatically organizes a large document collection into distinct groups of similar documents and discerns general themes hidden within the corpus [15,16,26,39]. Understandably, applications of document clustering go beyond organizing document collections into knowledge maps that facilitate subsequent knowledge retrievals and accesses. Document clustering, for example, has been applied to improve the efficiency of text categorization [1] and discover event episodes in temporally ordered documents (e.g., news stories) [35]. In addition, instead of presenting search results as one long list, some prior studies [29,40] and emerging search engines (e.g., Teoma,[1] vivisimo clustering engine,[2] MetaCrawler,[3] WebCrawler[4]) employ a document clustering approach to automatically organize search results into meaningful categories and thereby support cluster-based browsing.

Various document clustering techniques have been proposed [5,7,11,19,21,28,33,39], but most deal with monolingual documents (i.e., all target documents are written in the same language). However, the globalization of business environments and advances in Internet technology often cause an organization to maintain documents in different languages in its knowledge repositories. Evidently, organizations face the challenge of multilingual document clustering (MLDC). Such MLDC requirement is also prominent in other scenarios. For example, with advances in cross-lingual information retrieval (CLIR) technology, many search engines now offer a functionality that retrieves, for a user query expressed in one language, relevant documents in different languages. In this case, to facilitate cluster-based browsing, it would be preferable if the search engine were capable of clustering search results in different languages into distinct categories, each of which contains documents similar in their contents. Thus, effective MLDC support is also essential to search engines' cluster-based browsing capability in multilingual environments.

The major challenge facing MLDC is achieving cross-lingual semantic interoperability, which builds a bridge among representations of target documents written in different languages. Interoperability, to some extent, relates to CLIR, which processes a user query in one language and retrieves relevant documents in other languages. Several prior studies have applied Latent Semantic Indexing (LSI) [9] to address the cross-lingual semantic interoperability facing CLIR [2,20]. Specifically, LSI is a statistical method that analyzes and represents important associative patterns among terms. Previously proposed approaches first employ LSI to "train" a multilingual indexing system on the basis of a parallel corpus that contains a set of parallel documents prepared in two (or more) languages. Accordingly, the documents in any language can be represented in the language-independent LSI space. Similarly, a user query can be treated as a pseudo-document and represented as a vector in the same LSI space. Its similarity with other documents (in the same or a different language as that of the target query) thus can be derived, and the relevant documents in any languages can be retrieved. Empirical evaluation results show satisfactory cross-lingual retrieval effectiveness [2,20].

Despite substantial research studies into CLIR, MLDC has not been duly examined by prior research. Among the proposed MLDC techniques, most rely on a cross-lingual dictionary [6,12] or a multilingual ontology [27] for language translations. On the basis of a cross-lingual dictionary, the dictionary-based MLDC approach performs translations of documents from one language to another and then applies an existing, monolingual document clustering technique. However, the dictionary-based MLDC approach incurs several shortcomings. First, terms may be ambiguous or have multiple meanings that may be difficult to disambiguate with a dictionary. Second, the dictionary used may be limited in its vocabulary and phrases. For example, abbreviations, technical terms, and names of persons, corporations, or events may not appear in a particular cross-lingual dictionary, nor might new or slang terms.

The multilingual ontology-based MLDC approach requires the use of a classification scheme, which contains multilingual documents as a training set for each class. To facilitate MLDC, each document to be clustered is mapped onto the multilingual classification scheme. Specifically, using the training documents written in the same language as the target document,

[1] http://www.teoma.com.
[2] http://vivisimo.com.
[3] http://www.metacrawler.com.
[4] http://www.webcrawler.com.

this approach measures the relevance between the target document and each class, then uses this measure to represent the target document. Accordingly, the documents to be clustered are mapped in a language-independent space, which addresses the cross-lingual semantic interoperability challenge of MLDC. In addition, because the multilingual ontology-based MLDC approach does not involve the use of a cross-lingual dictionary for language translation, it does not share the same problems as the dictionary-based MLDC approach. However, its shortcomings include the limited availability of a multilingual ontology and potential domain incompatibility between the multilingual ontology employed and the set of multilingual documents to be clustered.

Because of these limitations of existing MLDC approaches and the demonstrated feasibility of LSI for CLIR, we propose an LSI-based MLDC technique. Our proposed MLDC technique employs the LSI analysis of a parallel corpus to construct a multilingual indexing system. Subsequently, the target multilingual documents to be clustered are indexed in the language-independent LSI space, and a monolingual document clustering technique is then employed to cluster the target multilingual documents.

The remainder of this article is organized as follows: In Section 2, we review the literature relevant to this study, including a brief review of LSI and existing monolingual and multilingual document clustering techniques. In Section 3, we depict the detailed development of our proposed LSI-based MLDC technique, including its overall process and specific designs. Next, we describe our empirical evaluation design and summarize important evaluation results in Section 4. Finally, we conclude with a summary, discussion of our research contributions, and some further research directions in Section 5.

2. Literature review

In this section, we review the literature relevant to this research, including a brief review of LSI and existing monolingual and multilingual document clustering techniques.

2.1. A brief review of LSI

In essence, LSI [9] analyzes term correlations by discerning the usage patterns of terms in documents. The particular statistical technique that LSI employs is the Singular Value Decomposition (SVD). Specifically, LSI applies SVD to a term-document matrix to construct a new semantic space formed by orthogonal vectors (referred to as LSI dimensions) and represents both terms and documents in this space in such a way that those terms and documents that are closely associated are placed close to one another.

Assume a term-document $t \times s$ matrix $X=[x_1, x_2,\ldots, x_s]$ that represents a collection of s documents described with a vocabulary of t terms (i.e., $x_i=[x_{1i}, x_{2i},\ldots, x_{ti}]$ is the vector for document i, and x_{ji} is the weight of term j in document i). Suppose that the rank of the term-document matrix X is r. In this case, LSI decomposes X, using SVD, into the product of three matrices:

$$X = U\Sigma V^{\mathrm{T}}, \tag{1}$$

where $U=[u_1,\ldots, u_r]$ is a $t \times r$ matrix, and u_j is called the left singular vector; $\Sigma=\mathrm{diag}(\sigma_1,\ldots, \sigma_r)$ is a $r \times r$ matrix that forms a r-dimensional LSI space, and $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r$ are the singular values of X; and $V=[v_1, v_2,\ldots, v_r]$ is a $s \times r$ matrix, V^{T} is the transpose of V, and v_j is called the right singular vector.

To capture the major associative patterns in documents while ignoring smaller variations that may be due to idiosyncrasies in the term usage of individual documents, a dimension reduction procedure that reduces the number of LSI dimensions generally is applied. Specifically, LSI reduces the original LSI space into a k-dimensional space by retaining only the first k columns of U and V and the k largest singular values in Σ. That is, LSI approximates X with a rank-k matrix as:

$$X_k = U_k \Sigma_k V_k^{\mathrm{T}}, \tag{2}$$

where $U_k=[u_1, u_2,\ldots, u_k]$, $\Sigma_k=\mathrm{diag}(\sigma_1, \sigma_2,\ldots, \sigma_k)$, and $V_k=[v_1, v_2,\ldots, v_k]$.

Semantically, the rows of the matrices $U_k\Sigma_k$ and $V_k\Sigma_k$ are representations of the original terms and documents, respectively, in the reduced k-dimensional semantic space. In addition, the document vector d_i for a new document i originally represented in the term space (i.e., described with the vocabulary of the t terms) can be mapped easily onto the k-dimensional semantic space by a "folding-in" procedure using Eq. (3) or (4) [32]. The difference between the two equations is whether to scale the vector by the inverse of the singular values.

$$\widehat{d_i} = U_k^{\mathrm{T}} d_i, \text{ or} \tag{3}$$

$$\widehat{d_i} = \Sigma_k^{-1} U_k^{\mathrm{T}} d_i, \tag{4}$$

where $\widehat{d_i}$ is the vector of document i in the k-dimensional semantic space.

Depending on the context of an intended application, an appropriate manipulation of the documents represented in the k-dimensional semantic space can then be

developed. For example, to support information retrieval, we would consider a user query a pseudo-document and represent it in the k-dimensional semantic space using the described folding-in procedure. Accordingly, we can estimate the relevancy of documents to the user query on the basis of the documents and the user query in the same semantic space. Because the constructed semantic space captures the major associative patterns in the documents, the word mismatch problem [9] that commonly impedes the effectiveness of information retrieval can be alleviated.

2.2. Monolingual document clustering techniques

Document clustering groups similar documents into clusters on the basis of their contents. The documents in the resultant clusters exhibit maximal similarity to those in the same cluster and, at the same time, share minimal similarity with documents from other clusters [39]. A review of extant literature suggests that most existing document clustering techniques deal with monolingual documents and are anchored in document content analysis. In addition, existing monolingual document clustering techniques can be classified broadly into non-LSI-based and LSI-based approaches.

Given a collection of monolingual documents, the general process of non-LSI-based document clustering techniques generally comprises three main phases: feature extraction and selection, document representation, and clustering [37,39]. Feature extraction begins with the parsing of each source document to produce a set of features (typically nouns and noun phrases) and exclude a list of prespecified stop words that are non-semantic-bearing words. Subsequently, representative features are selected from the set of extracted features. Feature selection is important for clustering efficiency and effectiveness because it not only condenses the size of the extracted feature set but also reduces any potential biases embedded in the original (i.e., non-trimmed) feature set [10,28]. Previous research commonly has employed feature selection metrics such as TF (term frequency), TF × IDF (term frequency × inverse document frequency), and their hybrids [5,21].

According to the top-k selection method, the k features with the highest selection metric scores are selected to represent each target document in the document representation phase. Thus, each document is represented as a feature vector jointly defined by the previously selected k features. A review of prior research suggests the prevalence of several representation methods, including binary (presence or absence of a feature in a document), TF (i.e., within-document term frequency), and TF × IDF [5,21,28].

In the final phase of non-LSI-based document clustering, the target documents are grouped into distinct clusters on the basis of the selected features and their respective values in each document. Common clustering approaches include partitioning-based [5,7,21], hierarchical [11,28,33], and Kohonen neural network [13,19,22,28]. A partitioning-based approach (e.g., k-means) partitions a set of documents into multiple non-overlapping clusters, whereas a hierarchical approach builds a binary clustering hierarchy whose leaf nodes represent the documents to be clustered. A representative hierarchical clustering algorithm is the hierarchical agglomerative clustering (HAC) method, which starts with as many clusters as there are documents; that is, each cluster contains only one document [33]. On the basis of a specific intercluster similarity measure of choice (e.g., single link, complete link, group-average link, Ward's method), the two most similar clusters are merged to form a new cluster. This merging process continues until either a hierarchy emerges with a single cluster at the top or a termination condition (e.g., intercluster similarity is less than a prespecified threshold) holds. The Kohonen neural network approach, also known as a self-organizing map, uses an unsupervised two-layer neural network [17,18]. In a Kohonen neural network, an input node corresponds to a feature in the selected feature vector space, whereas an output node represents a cluster in a two-dimensional grid. The network is fully connected, such that each output node connects to every input node with a particular connection weight. During the training phase, the source documents are fed into the network multiple times to train or adjust the connection weights so the distribution of the output nodes would accurately represent that of the input documents.

Unlike the non-LSI-based document clustering approach, which typically involves a feature selection phase, the LSI-based approach to clustering monolingual documents employs LSI to reduce the dimensions and thereby improve both clustering effectiveness and efficiency [30]. Its process generally commences with feature extraction, followed by document representation. Accordingly, the target documents to be clustered are represented as a term-document matrix. Subsequently, when applied to the term-document matrix, LSI produces a reduced k-dimensional semantic space in which the target documents are represented. According to a chosen clustering algorithm, the target documents are segmented into distinct clusters on the basis of their positions in the k-dimensional semantic space. The empirical evaluation results reported by Schuetze and Silverstein [30] suggest that compared with the non-LSI-based approach, the efficiency of LSI-based clustering improves dramatically without sacrificing clustering effectiveness.

2.3. Multilingual document clustering techniques

Chen and Lin [6] propose a multilingual news summarization system that is capable of summarizing news documents from multiple sources in different languages (in their study, English and Chinese). In their multilingual news summarization system, they propose a MLDC technique and use it to group multilingual news documents into clusters (events), each of which becomes the input to a news summarizer. Specifically, they adopt a two-phase approach to cluster the multilingual news documents, in which the first phase (i.e., monolingual clustering) segments news documents in the same language into several clusters. In the second phase, the resultant clusters in different languages are matched to form multilingual clusters. To address the cross-lingual semantic interoperability, these authors perform a unidirectional translation that translates Chinese news documents (verbs, nouns, and named entities only) to English on the basis of a cross-lingual dictionary. If a Chinese phrase has more than one translation, the one with the highest frequency is selected. In addition, they develop a name transliteration mechanism to translate Chinese named entities into English. The clusters initially in different languages therefore are in the same language (i.e., English), and similar clusters can be merged to form larger clusters, each of which pertains to a single event [6].

An alternative multilingual news summarizer, Columbia Newsblaster [12], relies on cross-lingual dictionaries and a machine-translation system for document (or feature) translations. Similar to the aforementioned MLDC technique [6], non-English documents are first translated into English, and then both the translated documents and the documents originally written in English are clustered using an existing monolingual document clustering algorithm. Several cross-lingual dictionaries (e.g., Japanese, Russian) have been developed to translate non-English documents to English. If a particular cross-lingual dictionary is not available, the machine-translation system (specifically, Altavista babelfish, http://babelfish.altavista.com) is employed to perform the translation.

Pouliquen et al. [27] propose a multilingual and cross-lingual news topic tracking system that employs *Eurovoc*, a multilingual ontology, to address the multilingual or cross-lingual challenges. *Eurovoc* offers a wide coverage classification scheme with approximately 6000 hierarchically organized classes, each of which contains multilingual documents as a training set. To facilitate MLDC, each document to be clustered is mapped onto the multilingual classification scheme. Using the training documents written in the same language as the target document, this method measures the relevance between the target document and each class, then uses the measure to represent the target document. Thus, the target documents are mapped in a language-independent space, in which the *Eurovoc* classes serve as the dimensions. Subsequently, a specific clustering algorithm (i.e., HAC) is adopted to cluster the target documents.

As we mentioned, the dictionary-based MLDC approach [6,12] has difficulty translating terms that are ambiguous or have multiple meanings, as well as those that do not appear in the cross-lingual dictionary (e.g., abbreviations, technical terms, named entities, slang). For example, existing dictionary-based MLDC techniques fail to address [12] or just adopt a simple heuristic rule [6] to deal with word sense disambiguation. More sophisticated disambiguation methods need to be developed to enhance translation quality and, in turn, improve the multilingual document clustering effectiveness attained by the dictionary-based MLDC approach. Although the multilingual ontology-based MLDC approach [27] seems promising, the limited availability of a multilingual ontology and potential domain incompatibility between the multilingual ontology employed and the set of multilingual documents to be clustered constrain its applicability.

3. Design of LSI-based multilingual document clustering (MLDC) technique

In this section, we detail our proposed LSI-based MLDC technique. As Fig. 1 illustrates, the overall process of our proposed technique comprises three main phases: multilingual semantic space analysis, document folding-in, and clustering. In the multilingual semantic space analysis phase, we rely on a parallel corpus and use LSI to construct a multilingual semantic space onto which terms and documents in either language can be mapped. For the document folding-in phase, we exploit the multilingual semantic space to fold in the target multilingual documents (in this study, Chinese and English) to be clustered in this semantic space. Finally, we choose a specific clustering algorithm to determine the appropriate clusters for the folded-in multilingual documents. In the following subsections, we depict each phase involved in our proposed LSI-based MLDC technique.

3.1. Multilingual semantic space analysis

Given a parallel corpus, the multilingual semantic space analysis employs LSI to construct automatically a multilingual semantic space that captures the major

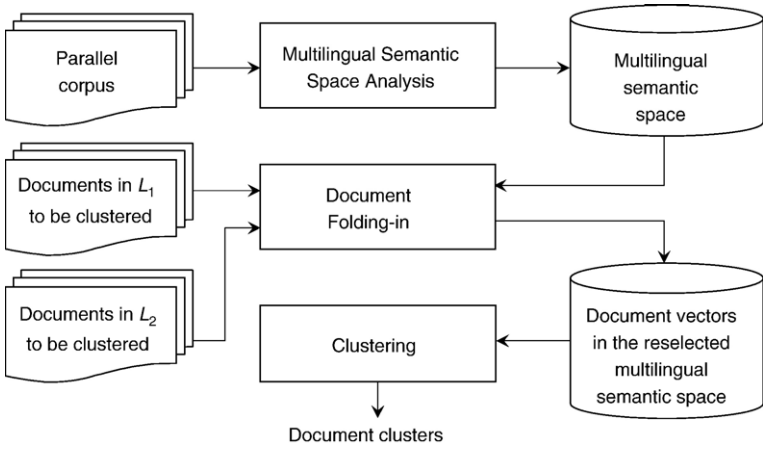

Fig. 1. Overall process of LSI-based multilingual document clustering (MLDC) technique.

associative patterns of monolingual and cross-lingual terms exhibited in the parallel documents. As Fig. 2 shows, the multilingual semantic space analysis starts by extracting terms (particularly, nouns and noun phrases) from each parallel document in the corpus. Because this study deals with English and Chinese documents, we adopt the rule-based part-of-speech tagger developed by Brill [3,4] to syntactically tag each word in the English documents in the parallel corpus. Subsequently, we employ the approach proposed by Voutilainen [34] to implement a noun phrase parser to extract noun phrases from each syntactically tagged English document. For the Chinese documents in the parallel corpus, we employ a hybrid approach that combines dictionary-based and statistical approaches (i.e., the mutual information measure) to extract the Chinese terms [41].

Following term extraction, we conduct the document representation process, in which the parallel documents in the corpus are represented as a term-document matrix. Assume that the parallel corpus contains s parallel documents and that the number of English and Chinese terms appearing in the parallel corpus is t. The document representation step generates a $t \times s$ term-document matrix, in which each parallel document is described by the t terms using a representation scheme of choice. Specifically, we adopt the TF × IDF representation scheme, mainly due to its popularity in document clustering and cross-lingual information retrieval research. Subsequently, we apply LSI to the term-document matrix to construct a multilingual semantic space and reduce the number of LSI dimensions. The resultant multilingual semantic space therefore comprises only the most important k LSI dimensions.

3.2. Document folding-in

The document folding-in phase (whose process is illustrated in Fig. 3) folds the target multilingual

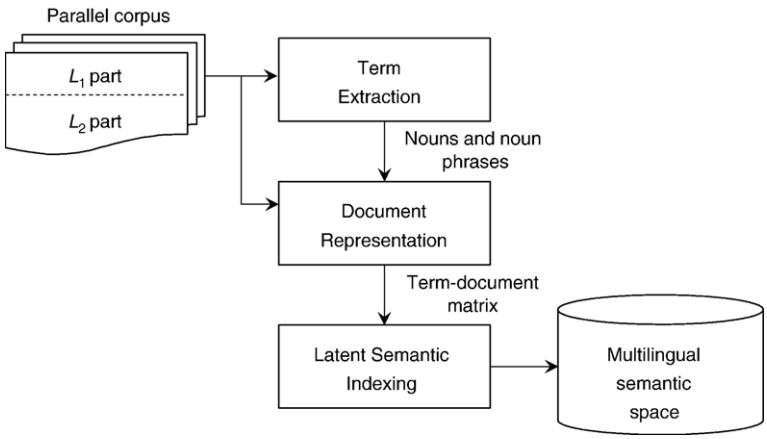

Fig. 2. Process of multilingual semantic space analysis.

documents (in L_1 or L_2) to be clustered into the multilingual semantic space on the basis of their constituent terms. In the term extraction step, we first extract nouns and noun phrases from each target document using the same techniques employed for the parallel corpus (described in Section 3.1), then employ binary, TF, and TF×IDF as alternative representation schemes to represent each target document as a term-document vector, and empirically evaluate their effectiveness. In the vector projection step, we fold in each target vector using Eq. (3) to project the corresponding document onto the multilingual semantic space. We choose Eq. (3) primarily because it is more efficient computationally than Eq. (4) and can achieve comparable effectiveness to that of Eq. (4) in an information retrieval context [32]. As a result, each target document i is represented as a vector $\widehat{d_i}=<w_{1i}, w_{2i},\ldots, w_{ki}>$, where w_{ji} is the weight of document i on the jth LSI dimension.

In LSI-based monolingual document clustering, the k LSI dimensions selected by the dimension reduction procedure essentially are the most representative dimensions for the target monolingual document corpus, because the semantic space has been constructed from the same set of documents. Thus, as we discuss in Section 2.2, the use of LSI for clustering monolingual documents does not require a dimension selection step. For clustering multilingual documents, the k LSI dimensions are selected on the basis of their importance in the parallel corpus. However, their importance with regard to the target multilingual documents to be clustered may not be the same as that to the parallel corpus, especially when the domain covered by the target multilingual corpus is only a subset of the parallel corpus. Hence, it is desirable to re-rank, with respect to the target multilingual documents, the k LSI dimensions and retain only the top h LSI dimensions (where $h \le k$) to represent the target multilingual documents. In this study, we adopt the sum of absolute dimension loadings across all target multilingual documents as our measure to re-rank the importance of the LSI dimensions:

$$\mathrm{DL}(D_j) = \sum_i |w_{ji}|, \tag{5}$$

where D_j denotes the LSI dimension j, and w_{ji} is the weight of document i in D_j.

Accordingly, we select the top h LSI dimensions with the highest DL scores to represent the folded-in documents.

3.3. Clustering

In this phase, we employ a clustering algorithm to cluster the target multilingual documents represented in the multilingual semantic space. As we mention in Section 2.2, common document clustering approaches include partitioning-based, hierarchical, and Kohonen neural network. The hierarchical clustering approach

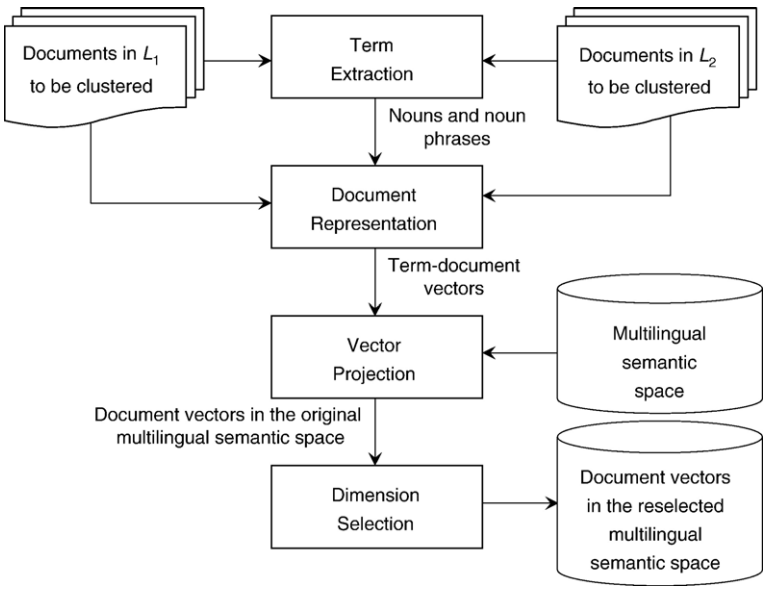

Fig. 3. Process of document folding-in.

has an advantage over the partitioning-based approach because the number of clusters need not be prespecified and can be decreased (increased) by simply moving up (down) the resultant clustering hierarchy. Furthermore, the hierarchical approach can achieve comparable clustering effectiveness to that of the Kohonen neural network [28]. Accordingly, we adopt the hierarchical approach, specifically, the HAC algorithm.

We estimate the similarity of two documents according to their cosine similarity measure. In each step of the HAC, the two most similar clusters are merged to form the next larger cluster. In this study, we employ the group-average link method to measure the similarity between two clusters. That is, the two clusters whose average similarity among all intercluster pairs of documents is the highest are joined first. The merging process continues until a hierarchy of clusters emerges (with a single cluster that contains all target documents at the top of the hierarchy) or a termination condition (e.g., a desired number of clusters is obtained or the intercluster similarity is less than a prespecified threshold) holds.

4. Empirical evaluation

In this section, we report on our empirical evaluation of our proposed LSI-based MLDC technique. In the following subsections, we detail the design of our empirical experiments, including data collection, evaluation procedure, and evaluation criteria. Subsequently, we discuss several important empirical evaluation results.

4.1. Data collection

To construct a multilingual semantic space, which is essential to our proposed LSI-based MLDC technique, we collected a parallel corpus in both English and Chinese. Our experimental parallel document corpus, collected from a dissertation and thesis digital library in Taiwan (accessible at http://datas.ncl.edu.tw/), contains 2949 Chinese- and English-language abstracts of theses and dissertations from the management information systems (MIS) departments of major universities in Taiwan between 1992 and 2003.

To evaluate the effectiveness of our proposed LSI-based MLDC technique, we also collected two additional monolingual document corpora with the same categorization scheme. The English document set (referred to as the English corpus) includes abstracts collected from a science Web site (http://sdos.ejournal.ascc.net/) and consists of 490 research articles from 1994–2004. Two doctoral students majoring in MIS reviewed the 490 English articles and collaboratively developed six categories (i.e., data mining, electronic commerce, e-learning, information retrieval, knowledge management, and wireless network) for the English corpus. The Chinese document set (referred to as the Chinese corpus) consists of two subsets: one collected from a science and technology information center Web site in Taiwan (http://sticnet.stic.gov.tw/) that includes 350 research reports from 1996–2003, and another from the proceedings of an international conference on information management that contains 109 research articles from 2000–2003. The two doctoral students also reviewed the 459 articles in the Chinese corpus and categorized them into the same six categories as those of the English corpus. The six categories therefore are considered the "true categories" for both the English and the Chinese corpus. For each research article in these corpora, we use only the title, abstract, and keywords for our evaluation purpose. Table 1 provides a summary of these document corpora.

4.2. Evaluation procedure and criteria

We use all documents in the parallel corpus to construct a multilingual semantic space. To expand the number of trials when evaluating the effectiveness of the proposed LSI-based MLDC technique, we randomly select 80% of the documents in the English and Chinese corpora to be clustered. Subsequently, we use this document subset to estimate clustering effectiveness. To minimize the potential biases that may result from our sampling process and obtain a reliable performance estimate, we perform the described sampling-and-clustering process 30 times and estimate the overall effectiveness of the proposed MLDC technique by averaging the performance estimates obtained from the 30 individual sampling-and-clustering processes.

We employ cluster recall and cluster precision [28,39], defined according to the concept of associations, to measure the effectiveness of a document clustering technique. An association refers to a pair of documents that belong to the same cluster. Accordingly,

Table 1
Summary of document corpora for evaluation

Data set	Number of documents	Average number of words per document	
Parallel corpus	2949	English	213
		Chinese	347
English corpus	490	162	
Chinese corpus	459	384	

cluster recall (CR) and cluster precision (CP) are defined as:

$$\mathrm{CR} = \frac{|\mathrm{CA}|}{|\mathrm{TA}|}, \text{ and} \tag{6}$$

$$\mathrm{CP} = \frac{|\mathrm{CA}|}{|\mathrm{GA}|}, \tag{7}$$

where TA is the set of associations in the true categories, GA is the set of associations in the clusters generated by the document clustering technique, and CA is the set of correct associations that exists in both the clusters and the true categories.

As Fig. 4 illustrates, we assume that English documents e_1 and e_2 and Chinese documents c_1 and c_2 belong to the true category T_1, and the English documents e_3 and e_4 and Chinese documents c_3, c_4, and c_5 belong to the true category T_2. In this case, the set of associations in the true categories is TA = {(e_1-e_2), (c_1-c_2), (e_1-c_1), (e_1-c_2), (e_2-c_1), (e_2-c_2), (e_3-e_4), (c_3-c_4), (c_3-c_5), (c_4-c_5), (e_3-c_3), (e_3-c_4), (e_3-c_5), (e_4-c_3), (e_4-c_4), (e_4-c_5)}, such that six associations pertain to T_1 and ten to T_2. Let the clusters produced by the proposed MLDC technique for the same set of documents be G_1, G_2, and G_3, where G_1 = {e_1, e_2, c_1, c_3}, G_2 = {e_3, e_4, c_2}, and G_3 = {c_4, c_5}. Thus, the set of associations in the clusters generated by the MLDC technique is GA = {(e_1-e_2), (c_1-c_3), (e_1-c_1), (e_1-c_3), (e_2-c_1), (e_2-c_3), (e_3-e_4), (e_3-c_2), (e_4-c_2), (c_4-c_5)}, such that the first six associations {(e_1-e_2), (c_1-c_3), (e_1-c_1), (e_1-c_3), (e_2-c_1), (e_2-c_3)} refer to cluster G_1, the following three associations {(e_3-e_4), (e_3-c_2), (e_4-c_2)} are for cluster G_2, and the last association {(c_4-c_5)} is for cluster G_3. Consequently, the set of correct associations is CA = {(e_1-e_2), (e_1-c_1), (e_2-c_1), (e_3-e_4), (c_4-c_5)}. Hence,

$$\mathrm{CR} = \frac{|\mathrm{CA}|}{|\mathrm{TA}|} = \frac{5}{16} = 0.3125, \text{ and } \mathrm{CP} = \frac{|\mathrm{CA}|}{|\mathrm{GA}|} = \frac{5}{10} = 0.5.$$

4.3. Tuning experiments and results

As we depicted previously, the proposed LSI-based MLDC technique involves two parameters and a design choice: k (i.e., number of LSI dimensions determined in the multilingual semantic space analysis phase) and h (i.e., number of re-selected LSI dimensions determined in the document folding-in phase), as well as the document representation scheme (i.e., binary, TF, or TF × IDF) to represent the target multilingual documents to be clustered (in the document folding-in phase). We set k to 200 and investigate the number of re-selected dimensions h, ranging from 5 to 25 in increments of 5 and from 50 to 200 in increments of 25. We examine the number of clusters produced by the proposed LSI-based MLDC technique, ranging from 1 to 50 in increments of 1. To address the inevitable trade-offs between cluster recall and cluster precision, we employ precision/recall trade-off (PRT) curves, which represent the effectiveness of a document clustering technique across different numbers of clusters [39]. Evidently, a larger number of clusters tends to result in an increase of cluster precision at the cost of cluster recall. Overall, a document clustering technique whose PRT curve is closer to the upper-right corner is more desirable.

As we show in Fig. 5 (to reduce the complexity of the figure, we show only a subset of values for h), the PRT curve of the proposed LSI-based MLDC technique using the binary representation scheme moves toward to the upper-right corner as h increases from 5 to 50, but remains comparable or becomes inferior when h is beyond this range. Accordingly, the most upper-right PRT curve is attained when h is set to 50. In addition, the breakeven point measure, defined as the value at which cluster recall equals cluster precision, also suggests that h as 50 achieves the best clustering effectiveness for the binary representation scheme. Specifically, the breakeven points attained by h = 5, 10, 50, 100, and 200 are 0.4699, 0.5426, 0.7095, 0.6751, and 0.6038, respectively. Likewise, when using an alternative representation scheme (i.e., TF or TF × IDF), the proposed technique exhibits similar behavior across the range of h investigated; the best clustering effectiveness of the proposed technique occurs when h is set to 20 for both the TF and TF × IDF representation schemes.

Using the parameter value determined previously, we evaluate the effectiveness of the proposed LSI-based

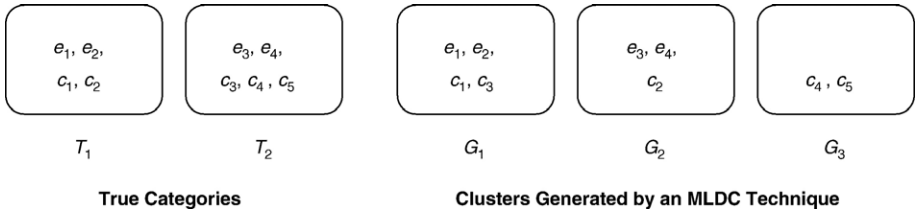

Fig. 4. Examples of true categories and clusters generated by an MLDC technique.

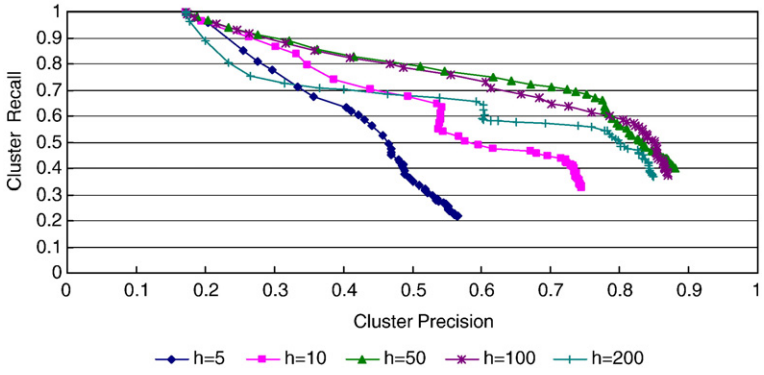

Fig. 5. PRT curves of the LSI-based MLDC technique (with binary representation scheme).

MLDC technique across different representation schemes. As we show in Fig. 6, across the range of number of clusters examined, the proposed technique with the binary representation scheme generally results in better clustering effectiveness than that achieved with TF or TF × IDF, and the use of the TF representation scheme results in the worst clustering effectiveness. Thus, in subsequent experiments, we use the binary representation scheme for our proposed LSI-based MLDC technique.

The inferior performance resulting from the use of TF and TF × IDF representation schemes may be because our multilingual documents are short in length. As Table 1 depicts, the average number of words per document in our English corpus is 162, and the average number of Chinese characters per document in our Chinese corpus is 384. As a result, the within-document term frequency may not be a reliable metric for estimating the representativeness of a term within a document. In other words, given a document that is short in length, a term with a higher occurrence frequency may not be more representative than another term with a lower frequency. Accordingly, the use of the binary representation scheme that considers only the presence or absence of a term within a target document achieves a better clustering effectiveness than its counterparts (i.e., TF and TF × IDF).

4.4. Effects of dimension selection

Recall that in our previous tuning experiments, we set the number of LSI dimensions k to 200 in the multilingual semantic space analysis phase and then re-selected h dimensions among those 200 in the dimension selection step of the document folding-in phase. To investigate the effects of our proposed dimension selection mechanism, we empirically compare the clustering effectiveness of the proposed LSI-based MLDC technique with and without the use of our dimension selection mechanism. Specifically, for the proposed LSI-based MLDC technique without dimension selection, we do not perform the dimension selection step in the document folding-in phase; i.e., the number of LSI dimensions (k) determined in the Latent Semantic Indexing step of the multilingual

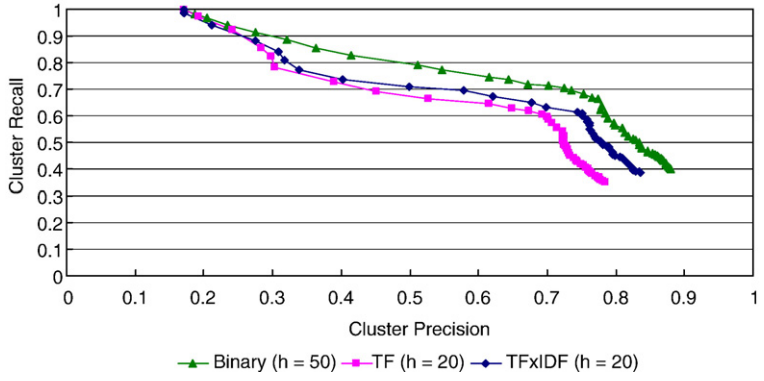

Fig. 6. Comparisons of different representation schemes.

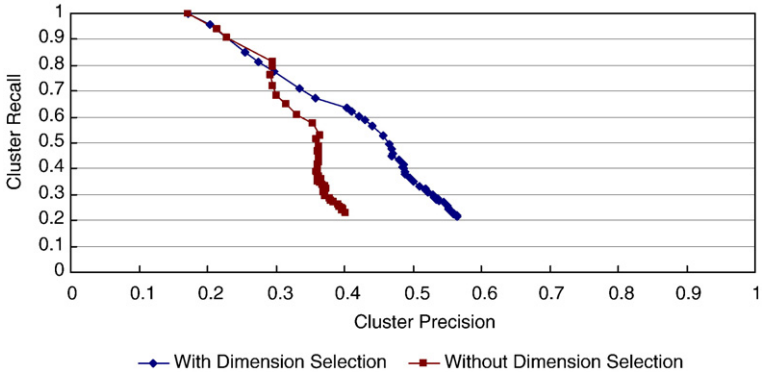

Fig. 7. Effect of dimension selection ($h=5$ for MLDC with dimension selection; $k=5$ for MLDC without dimension selection).

semantic space analysis phase is used to represent the folded-in documents to be clustered. We investigate different values for k, ranging from 5 to 25 in increments of 5 and from 50 to 200 in increments of 25. For the LSI-based MLDC technique with the dimension selection mechanism, we set k to 200 and examine various values for the number of re-selected dimensions h, ranging between 5 and 200.

As Fig. 7 shows, when the number of LSI dimensions is 5, the clustering effectiveness achieved with dimension selection is notably superior to that attained without dimension selection. However, the superiority of dimension selection over its counterpart decreases as the number of LSI dimensions increases. As Fig. 8 illustrates, when the number of LSI dimensions equals 20, the proposed LSI-based MLDC technique with dimension selection only slightly outperforms that without dimension selection. As the number of LSI dimensions increases further, both methods reach comparable clustering effectiveness. The analysis suggests that the proposed dimension selection mechanism is effective (i.e., resulting in better or comparable clustering effectiveness).

Next, we conduct a "best scenario versus best scenario" analysis of the two methods (i.e., with and without dimension selection). As we mentioned previously, with dimension selection, we attain the best clustering effectiveness of the proposed LSI-based MLDC technique when h is set to 50. On the contrary, when we do not employ the dimension selection mechanism, the PRT curve of the LSI-based MLDC technique moves toward to the upper-right corner as k expands from 5 to 75, remains comparable when k ranges between 75 and 125, and then degrades when k is greater than 125. Therefore, setting k to 125 results in the best clustering effectiveness for the LSI-based MLDC technique without dimension selection. As we illustrate in Fig. 9, the best scenarios attained by both methods are largely comparable, but the number of LSI dimensions involved is notably different (50 versus 125). That is, the LSI-based MLDC technique with our proposed dimension selection mechanism achieves its best scenario with fewer LSI dimensions and thus can be considered more efficient than the technique without dimension selection.

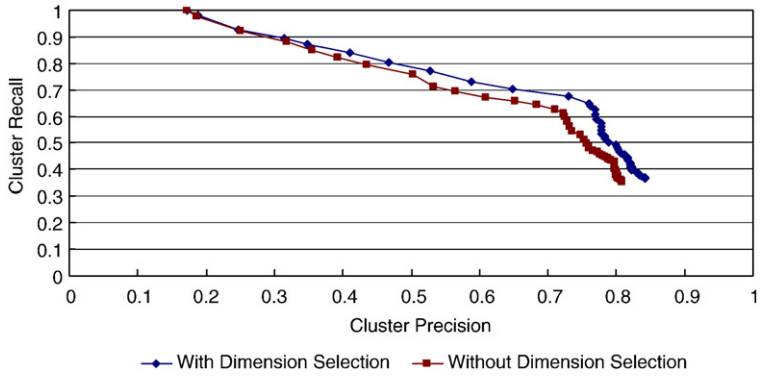

Fig. 8. Effect of dimension selection ($h=20$ for MLDC with dimension selection; $k=20$ for MLDC without dimension selection).

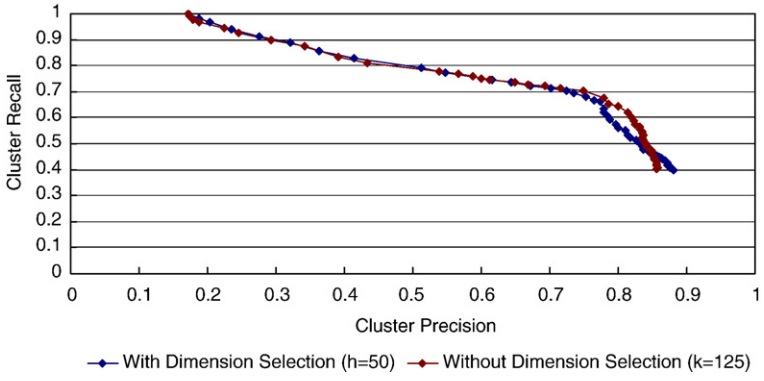

Fig. 9. Best scenario versus best scenario comparison.

4.5. Analysis of monolingual and cross-lingual capability of LSI-based MLDC technique

Given a collection of multilingual documents, an MLDC technique that satisfactorily clusters monolingual documents but inadequately clusters cross-lingual ones in the collection is not desirable, because crossing the language barrier between documents in different languages is a main goal of MLDC. Thus, when performing document clustering on a multilingual corpus, in addition to measuring overall clustering effectiveness by cluster recall and cluster precision (defined in Eqs. (6) and (7)), we must evaluate the capability of the proposed LSI-based MLDC technique to cluster documents in the same language and documents across different languages separately. Therefore, we decompose the cluster recall and cluster precision measures into monolingual and cross-lingual components. Accordingly, we define the monolingual cluster recall (CR_{MN}) and the monolingual cluster precision (CP_{MN}) as:

$$CR_{MN} = \frac{|CA_{MN}|}{|TA_{MN}|}, \text{and} \tag{8}$$

$$CP_{MN} = \frac{|CA_{MN}|}{|GA_{MN}|}, \tag{9}$$

where TA_{MN} is the set of monolingual associations in the true categories, GA_{MN} is the set of monolingual associations in the clusters generated by an MLDC technique, and CA_{MN} is the set of correct monolingual associations that exists in both the clusters and the true categories.

Similarly, the cross-lingual cluster recall (CR_{CL}) and cluster precision (CP_{CL}) are defined as follows:

$$CR_{CL} = \frac{|CA_{CL}|}{|TA_{CL}|}, \text{and} \tag{10}$$

$$CP_{CL} = \frac{|CA_{CL}|}{|GA_{CL}|} \tag{11}$$

where TA_{CL} is the set of cross-lingual associations in the true categories, GA_{CL} is the set of cross-lingual associations in the clusters generated by an MLDC technique, and CA_{CL} is the set of correct cross-lingual associations that exists in both the clusters and the true categories.

From the example in Fig. 4, the set of monolingual associations in the true categories is $TA_{MN}=\{(e_1-e_2), (c_1-c_2), (e_3-e_4), (c_3-c_4), (c_3-c_5), (c_4-c_5)\}$, whereas the set of cross-lingual associations in the true categories is $TA_{CL}=\{(e_1-c_1), (e_1-c_2), (e_2-c_1), (e_2-c_2), (e_3-c_3), (e_3-c_4), (e_3-c_5), (e_4-c_3), (e_4-c_4), (e_4-c_5)\}$. In contrast, the set of monolingual associations in the three clusters generated by the proposed MLDC technique is $GA_{MN}=\{(e_1-e_2), (c_1-c_3), (e_3-e_4), (c_4-c_5)\}$, and the set of cross-lingual associations is $GA_{CL}=\{(e_1-c_1), (e_1-c_3), (e_2-c_1), (e_2-c_3), (e_3-c_2), (e_4-c_2)\}$. Consequently, the set of correct monolingual associations in this case is $CA_{MN}=\{(e_1-e_2), (e_3-e_4), (c_4-c_5)\}$, and the set of correct cross-lingual associations is $CA_{CL}=\{(e_1-c_1), (e_2-c_1)\}$. Hence, the resultant monolingual cluster recall (CR_{MN}) and cluster precision (CP_{MN}) are as follows:

$$CR_{MN} = \frac{|CA_{MN}|}{|TA_{MN}|} = \frac{3}{6} = 0.5, \text{ and } CP_{MN} = \frac{|CA_{MN}|}{|GA_{MN}|} = \frac{3}{4} = 0.75.$$

The corresponding cross-lingual cluster recall (CR_{CL}) and cluster precision (CP_{CL}) are

$$CR_{CL} = \frac{|CA_{CL}|}{|TA_{CL}|} = \frac{2}{10} = 0.2, \text{ and } CP_{CL} = \frac{|CA_{CL}|}{|GA_{CL}|} = \frac{2}{6} = 0.33.$$

Evidently, for clustering a multilingual document corpus, an MLDC technique with monolingual and cross-

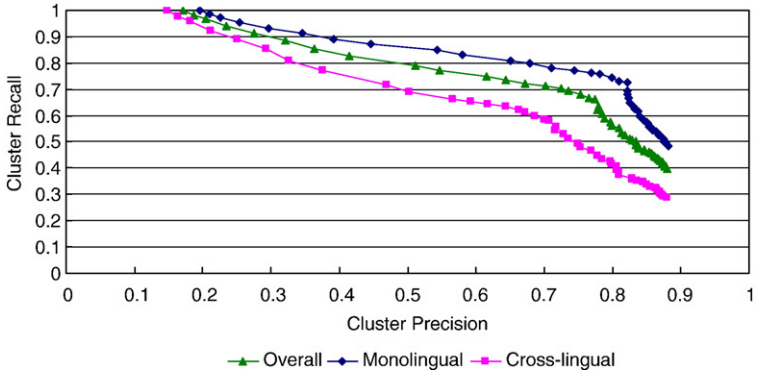

Fig. 10. PRT curves of overall, monolingual, and cross-lingual performance (with binary representation scheme and $h=50$).

lingual PRT curves close to the PRT curve for the overall performance is more desirable. As we show in Fig. 10, the overall clustering effectiveness attained by our proposed MLDC technique does not deviate substantially from its monolingual clustering effectiveness. As can be derived from Fig. 10, the breakeven point of the overall PRT curve is 0.7095, which is 5.57% lower than that of the monolingual PRT curve (i.e., 0.7652). According to its cross-lingual clustering capability, the proposed technique is also acceptable. Specifically, the breakeven point of the overall PRT curve is 7.38% higher than that of the cross-lingual PRT curve (i.e., 0.6357). Overall, the proposed LSI-based MLDC technique achieves a satisfactory balance between monolingual and cross-lingual clustering effectiveness.

5. Conclusion and future research directions

With increasing globalization and Internet technology advances, many organizations maintain documents in different languages in their knowledge repositories, which necessitates MLDC to support their knowledge management. Multilingual document clustering organizes documents in different languages into distinct categories of similar documents on the basis of their contents. Various document clustering techniques have been proposed in the literature; however, most deal with only monolingual documents. In response, we propose an LSI-based MLDC technique that employs LSI analysis to a parallel corpus and constructs a multilingual semantic space. Accordingly, the target multilingual documents to be clustered are mapped in this language-independent space, and a monolingual document clustering technique (particularly, HAC) clusters the target multilingual documents. Our empirical evaluation results show that the proposed LSI-based MLDC technique achieves satisfactory clustering effectiveness and is capable of maintaining a good balance between monolingual and cross-lingual clustering effectiveness when clustering a multilingual document corpus.

Some research extensions to this study might include the following: First, this study intends to demonstrate the feasibility of applying LSI to MLDC, and thus, we focus only on the clustering effectiveness of our proposed LSI-based MLDC technique in our empirical evaluations. Developing and evaluating a thesaurus-based MLDC technique that relies on a statistical thesaurus, constructed from a parallel corpus, for feature translation or a machine-translation-based MLDC technique represents an interesting research direction. Second, the proposed LSI-based MLDC technique does not explore the characteristics of the multilingual documents to be clustered to improve clustering effectiveness. For example, for our target corpus, it seems reasonable to assume that terms appearing in titles or as keywords of research abstracts are more important than those occurring only in abstracts. Therefore, a desirable future research direction would take such document characteristics into account during the document clustering process. Third, our current evaluation study employs only one set of document corpora (parallel, English, and Chinese corpora) in a specific domain (i.e., research articles). Evaluations of the proposed LSI-based MLDC technique using other document corpora that pertain to more diversified application domains (e.g., multilingual news or patent clustering) would improve the generalizability of the evaluation results we report. Fourth, we design a dimension selection mechanism (included in the document folding-in phase of our proposed technique) whose utility has been empirically demonstrated. We employ the sum of the absolute dimension loadings across the target multilingual documents to be clustered as the measure for re-ranking LSI dimensions. However, other measures of dimension selection may improve the

clustering effectiveness of the proposed LSI-based MLDC technique further. Fifth and finally, in addition to MLDC, other multilingual or cross-lingual document management issues, such as multilingual event detection and summarization, cross-lingual text categorization, and cross-lingual question and answer, demand additional research attention.

Acknowledgment

This work was supported by the National Science Council of the Republic of China under the grants NSC 93-2416-H-110-021 and NSC 94-2416-H-110-002.

References

[1] C.C. Aggarwal, S.C. Gates, P.S. Yu, On the merits of building categorization systems by supervised clustering, Proceedings of Conference on Knowledge Discovery in Databases, San Diego, CA, 1999, pp. 352–356.
[2] M.W. Berry, P.G. Young, Using latent semantic indexing for multilingual information retrieval, Computers and the Humanities 29 (6) (1995) 413–429.
[3] E. Brill, A simple rule-based part of speech tagger, Proceedings of the Third Conference on Applied Natural Language Processing, Association for Computational Linguistics, Trento, Italy, 1992.
[4] E. Brill, Some advances in rule-based part of speech tagging, Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94), Seattle, WA, 1994, pp. 722–727.
[5] D. Boley, M. Gini, R. Gross, E. Han, K. Hastings, G. Karypis, V. Kumar, B. Mobasher, J. Moore, Partitioning-based clustering for web document categorization, Decision Support Systems 27 (3) (1999) 329–341.
[6] H.H. Chen, C.J. Lin, A multilingual news summarizer, Proceedings of 18th International Conference on Computational Linguistics, July–August 2000, pp. 159–165.
[7] D. Cutting, D. Karger, J. Pedersen, J. Tukey, Scatter/gather: a cluster-based approach to browsing large document collections, Proceedings of 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1992, pp. 318–329.
[8] T.H. Davenport, L. Prusak, Working Knowledge: How Organizations Manage What They Know, Harvard Business School Press, Boston, 1998.
[9] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R.A. Harshman, Indexing by latent semantic analysis, Journal of the American Society for Information Science 41 (6) (1990) 391–407.
[10] S. Dumais, J. Platt, D. Heckerman, M. Sahami, Inductive learning algorithms and representation for text categorization, Proceedings of the 1998 ACM 7th International Conference on Information and Knowledge Management (CIKM '98), 1998, pp. 148–155.
[11] A. El-Hamdouchi, P. Willett, Hierarchical document clustering using Ward's method, Proceedings of ACM Conference on Research and Development in Information Retrieval, 1986, pp. 149–156.
[12] D.K. Evans, J.L. Klavans, A Platform for Multilingual News Summarization, Technical Report, Department of Computer Science, Columbia University, May 2003.
[13] V.P. Guerrero Bote, F. Moya Anegón, V. Herrero Solana, Document organization using Kohonen's algorithm, Information Processing & Management 38 (1) (2002) 79–89.
[14] M.T. Hansen, N. Nohria, T. Tierney, What's your strategy for managing knowledge? Harvard Business Review 77 (2) (1999) 106–116.
[15] H.J. Kim, S.G. Lee, A semi-supervised document clustering techniques for information organization, Proceedings of the 2000 ACM 9th International Conference on Information and Knowledge Management (CIKM '00), McLean, VA, 2000, pp. 30–37.
[16] H.J. Kim, S.G. Lee, An effective document clustering method using user-adaptable distance metrics, Proceedings of the 2002 ACM Symposium on Applied Computing, Madrid, Spain, 2002, pp. 16–20.
[17] T. Kohonen, Self-organization and Associative Memory, Springer, 1989.
[18] T. Kohonen, Self-organizing Maps, Springer, 1995.
[19] K. Lagus, T. Honkela, S. Kaski, T. Kohonen, Self-organizing maps of document collections: a new approach to interactive exploration, Proceedings of the 2nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1996.
[20] T.K. Landauer, M.L. Littman, Full automatic cross-language document retrieval using latent semantic indexing, Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research, Waterloo, Ontario, October 1990.
[21] B. Larsen, C. Aone, Fast and effective text mining using linear-time document clustering, Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999, pp. 16–22.
[22] C. Lin, H. Chen, J.F. Nunamaker, Verifying the proximity and size hypothesis for self-organizing maps, Journal of Management Information Systems 16 (3) (1999) 57–70.
[23] D.E. O'Leary, Enterprise knowledge management, IEEE Computer 31 (3) (1998) 54–61.
[24] D.E. O'Leary, Using AI in knowledge management: knowledge bases and ontologies, IEEE Intelligent Systems 13 (3) (1998) 34–39.
[25] T.H. Ong, H. Chen, W.K. Sung, B. Zhu, Newsmap: a knowledge map for online news, Decision Support Systems 39 (4) (2005) 583–597.
[26] P. Pantel, D. Lin, Document clustering with committees, Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, 2002, pp. 199–206.
[27] B. Pouliquen, R. Steinberger, C. Ignat, E. Käsper, I. Temnikova, Multilingual and cross-lingual news topic tracking, Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland, August 2004.
[28] D. Roussinov, H. Chen, Document clustering for electronic meetings: an experimental comparison of two techniques, Decision Support Systems 27 (1–2) (1999) 67–79.
[29] D. Roussinov, H. Chen, Information navigation on the web by clustering and summarizing query results, Information Processing & Management 37 (6) (2001) 789–816.
[30] H. Schutze, C. Silverstein, Projections for efficient document clustering, Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, PA, July 1997, pp. 74–81.
[31] A. Stewart, Under the hood at ford, Webmaster Magazine (June 1997) 26–34.
[32] C. Tang, S. Dwarkadas, Z. Xu, On scaling latent semantic indexing for large peer-to-peer systems, Proceedings of the 27th Annual International ACM SIGIR Conference on Research and

Development in Information Retrieval, Sheffield, South Yorkshire, UK, July 2004, pp. 112–121.
[33] E.M. Voorhees, Implementing agglomerative hierarchical clustering algorithms for use in document retrieval, Information Processing & Management 22 (1986) 465–476.
[34] A. Voutilainen, Nptool: a detector of english noun phrases, Proceedings of Workshop on Very Large Corpora, Ohio, June 1993, pp. 48–57.
[35] C. Wei, Y.H. Chang, Discovering event evolution patterns from document sequences, IEEE Transactions on Systems, Man and Cybernetics. Part A. Systems and Humans 37 (2) (2007) 273–283.
[36] C. Wei, P. Hu, H.H. Chen, Design and evaluation of a knowledge management system, IEEE Software 19 (3) (2002) 56–59.
[37] C. Wei, P. Hu, Y.X. Dong, Managing document categories in e-commerce environments: an evolution-based approach, European Journal of Information Systems 11 (3) (2002) 208–222.
[38] C. Wei, T.H. Cheng, Y.C. Pai, Semantic enrichment in knowledge repositories: annotating reply semantic relationships between discussion documents, Journal of Database Management 17 (1) (2006) 49–66.
[39] C. Wei, C.S. Yang, H.W. Hsiao, T.H. Cheng, Combining preference- and content-based approaches for improving document clustering effectiveness, Information Processing & Management 42 (2) (2006) 350–372.
[40] M. Wu, M. Fuller, R. Wilkinson, Using clustering and classification approaches in interactive retrieval, Information Processing & Management 37 (3) (2001) 456–484.
[41] C.C. Yang, J. Luk, Automatic generation of English/Chinese thesaurus based on a parallel corpus in laws, Journal of the American Society for Information Science and Technology 54 (7) (2003) 671–682.

Chih-Ping Wei received a BS in Management Science from the National Chiao-Tung University in Taiwan, ROC in 1987 and an MS and a Ph.D. in Management Information Systems from the University of Arizona in 1991 and 1996. He is currently a professor of Institute of Technology Management at National Tsing Hua University in Taiwan, ROC. Prior to joining the National Tsing Hua University in 2005, he was a faculty member at Department of Information Management at National Sun Yat-sen University in Taiwan since 1996 and a visiting scholar at the University of Illinois at Urbana-Champaign in Fall 2001 and the Chinese University of Hong Kong in Summer 2006 and 2007. His papers have appeared in *Journal of Management Information Systems* (*JMIS*), *Decision Support Systems* (*DSS*), *IEEE Transactions on Engineering Management*, *IEEE Software*, *IEEE Intelligent Systems*, *IEEE Transactions on Systems*, *Man*, *Cybernetics*, *IEEE Transactions on Information Technology in Biomedicine*, *European Journal of Information Systems*, *Journal of Database Management*, *Information Processing and Management*, and *Journal of Organizational Computing and Electronic Commerce*, etc. His current research interests include knowledge discovery and data mining, information retrieval and text mining, knowledge management, multidatabase management and integration, and data warehouse design. He has edited special issues of *Decision Support Systems*, *International Journal of Electronic Commerce*, and *Electronic Commerce Research and Applications*. He can be reached at Institute of Technology Management, National Tsing Hua University, Hsinchu, Taiwan, ROC; cpwei@mx.nthu.edu.tw.

Christopher C. Yang is an associate professor in the Department of Systems Engineering and Engineering Management at the Chinese University of Hong Kong. He received his B.S., M.S., and Ph.D. in Electrical and Computer Engineering from the University of Arizona. He has also been an assistant professor in the Department of Computer Science and Information Systems at the University of Hong Kong and a research scientist in the Department of Management Information Systems at the University of Arizona. His recent research interests include cross-lingual information retrieval and knowledge management, Web search and mining, security informatics, text summarization, multimedia retrieval, information visualization, digital library, and electronic commerce. He has published over 130 referred journal and conference papers in *Journal of the American Society for Information Science and Technology* (*JASIST*), *Decision Support Systems* (*DSS*), *IEEE Transactions on Image Processing*, *IEEE Transactions on Robotics and Automation*, *IEEE Computer*, *Information Processing and Management*, *Journal of Information Science*, *Graphical Models and Image Processing*, *Optical Engineering*, *Pattern Recognition*, *International Journal of Electronic Commerce*, *Applied Artificial Intelligence*, *IWWWC*, *SIGIR*, *ICIS*, *CIKM*, and more. He has edited several special issues on multilingual information systems, knowledge management, and Web mining in *JASIST* and *DSS*. He chaired and served in many international conferences and workshops. He has also frequently served as an invited panelist in the NSF Review Panel in US. He was the chairman of the Association for Computing Machinery Hong Kong Chapter. Starting from 2008, he is an associate professor in the College of Information Science & Technology at Drexel University.

Chia-Min Lin received a BS and an MBA in Management Information Systems from the National Sun Yat-sen University in Taiwan, ROC in 2003 and 2006. He is currently a software engineer in the Openfind Information Technology, Inc. His current research interests include information retrieval, text mining, and software architecture and design. He can be reached at the Openfind Information Technology Inc., 4F, No. 222, Sec. 2, Nan-Chang Rd., Taipei, Taiwan, ROC; alucard.lin@gmail.com.