Available online at www.sciencedirect.com

SCIENCE DIRECT®

Knowledge-Based Systems 17 (2004) 219–227

Knowledge-Based SYSTEMS

www.elsevier.com/locate/knosys

A multilingual text mining approach to web cross-lingual text retrieval

Rowena Chau*, Chung-Hsing Yeh

School of Business Systems, Faculty of Information Technology, Monash University, Clayton, Vic. 3800, Australia

Received 26 August 2003; accepted 6 April 2004

Available online 28 May 2004

Abstract

To enable concept-based cross-lingual text retrieval (CLTR) using multilingual text mining, our approach will first discover the multilingual concept–term relationships from linguistically diverse textual data relevant to a domain. Second, the multilingual concept–term relationships, in turn, are used to discover the conceptual content of the multilingual text, which is either a document containing potentially relevant information or a query expressing an information need. When language-independent concepts hidden beneath both document and query are revealed, concept-based matching is made possible. Hence, concept-based CLTR is facilitated. This approach is employed for developing a multi-agent system to facilitate concept-based CLTR on the Web.

Keywords: Multilingual text mining; Cross-lingual text retrieval; Agent; Fuzzy clustering; Fuzzy classification

1. Introduction

The exponential growth of the World Wide Web over the globe is the most influential factor that contributes to the increasing awareness of cross-lingual text retrieval (CLTR) in recent years. Relevant information exists in different languages. A user may want to find documents in languages other than the one the query is formulated in. Among various CLTR techniques developed recently, query translation is the most extensively studied one. Such CLTR approaches are developed mainly to facilitate term-based lexical transfer between a single pair of source and target languages. However, a bilingual lexical transfer is not sufficient for fully supporting the user's need of multilingual information seeking.

Within a multilingual information community, users often rely on CLTR to explore global knowledge relevant to a certain topic/area. Instead of looking for some specific documents that can be characterized by a few translation equivalents of the query terms, users are often interested in a broader view of a particular domain. They are thinking in terms of concepts and expecting to receive all relevant documents existing in any language. In such cases, concept-based CLTR capable of identifying multilingual documents about the concept of a query is necessary.

Documents and queries about the same concept do not necessarily contain matching sets of translation equivalents of each other. Conceptual relevance between documents and queries is not to be determined in an explicit way. To realize concept-based CLTR, the development of a conceptual interlingua to support lexical transfer across multiple languages is required. To encode a conceptual interlingua, terms from multiple languages describing the same concept should be mapped to a language-independent scheme. In this way, it is possible to match a term to its corresponding counterparts in all other languages and to achieve concept-based CLTR.

Multilingual thesaurus (e.g. EuroWordNet) encoding conceptual relationship among multilingual terms is such a conceptual interlingua that has been used to achieve this goal [7]. However, the manual construction of multilingual thesauri is very labor expensive and their coverage is not domain specific. An automatic and possibly unsupervised approach for generating such linguistic knowledge for CLTR by discovering structures of lexical relationships among multilingual terms from analyzing text of relevant domain is highly desirable.

To provide better support to CLTR, a knowledge discovery technology, known as text mining, looks promising in discovering such kind of in-depth multilingual linguistic knowledge. Typically, text mining concerns the discovery and extraction of hidden relationships, such as

* Corresponding author.
E-mail address: rowena.chau@infotech.monash.edu.au (R. Chau).

0950-7051/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2004.04.001

conceptual associations, among textual items, including terms and documents.

To enable concept-based CLTR using multilingual text mining, our approach will first discover the multilingual concept–term relationships from linguistically diverse textual data relevant to a domain. Second, the multilingual concept–term relationships, in turn, are used to discover the conceptual content of the multilingual text, which can be either a document containing potentially relevant information or a query expressing an information need. When language-independent concepts hidden beneath both documents and queries are revealed, concept-based matching is made possible, thus facilitating concept-based CLTR. This approach is employed for developing a multi-agent system to facilitate concept-based CLTR on the Web.

2. Current CLTR techniques

Given a query expressed in one language, the objective of CLTR is to search for relevant documents in other languages. To break the language barrier, either document or query translation is required. As query translation is less resource demanding than document translation, it has proven to be a more feasible approach to CLTR. There are three major approaches to query translation: (a) machine translation, (b) knowledge-based methods using machines-readable dictionary [2,8], and (c) corpus-based methods using parallel corpus [14].

Despite translating query using machine translation being straightforward, it is argued that machine translation and CLTR have divergent concerns [13]. Machine translation aiming at syntactically accurate translation is redundant to CLTR. Since query is short, grammatically invalid and is just formulated with a few terms, it offers little context for the machine translation system to translate accurately. Besides, machine translation always replaces the original query term with only one of its many possible synonymous translations in the target language. This prevents a query expansion by which all synonymous terms are considered to enhance recall.

Query can easily be translated by replacing every query term with a set of all its possible translations as encoded in a machine-readable dictionary. However, this approach is ineffective mainly due to the translation ambiguity of polysemous terms (i.e. terms with multiple meanings). A polysemous term may have several alternative translations carrying different senses (meanings) in any foreign language. Translating a query by including every possible translation of every query term can greatly increase the set of possible meanings in the translated query, thus contributing to poor precision. Moreover, inadequate coverage of specific terminology and phrases is also a serious shortcoming of such machine-readable dictionary.

An alternative to machine-readable dictionary is using a parallel corpus. A parallel corpus is a set of identical text written in multiple languages. Corpus-based query translation is based on the idea that terms are represented as points in a multi-dimensional semantic space, and terms (in different languages) mapped to the same set of points in that semantic space are used to describe the same concept. Geometric relationships between terms within the semantic space are automatically extracted by analyzing co-occurrence statistics of terms across a parallel corpus. By substituting every query term with its geometrically close translations in the semantic space, query translation is then facilitated [6,12]. The corpus-based approach is most effective for CLTR when the document collection is domain-specific. In this paper, a corpus-based approach to CLTR that applies multilingual text mining using a parallel corpus is proposed.

3. A multilingual text mining approach to cross-lingual text retrieval

Our work for enabling CLTR with multilingual text mining is focused on exploiting the knowledge discovery capability of text mining over multilingual text. This is a logical approach due to the complementary nature of these two areas. Both CLTR and multilingual text mining analyze multilingual textual data employing techniques from information retrieval, natural language processing and machine learning. In terms of the functions they perform, CLTR facilitates multilingual information access while multilingual text mining enables knowledge discovery from multilingual texts. The objective of CLTR is to locate relevant documents from a multilingual document collection in response to a query represented by a set of terms, while the objective of multilingual text mining is to reveal concepts and their relationships embedded within a collection of multilingual texts. To determine the conceptual relevance between documents and a query written in different languages, CLTR requires understanding of their semantics. Multilingual text mining has the potential to complement CLTR by discovering intrinsic meanings of multilingual texts. Our approach to concept-based CLTR with multilingual text mining is depicted in Fig. 1.

Within an integrated framework, multilingual text mining yields knowledge that supports CLTR. First, the multilingual concept–term relationships, which are necessary for a CLTR system to associate documents and query across languages, are mined from a parallel corpus. This is achieved by a fuzzy multilingual term clustering algorithm. By grouping conceptually related multilingual terms into clusters, the multilingual concept–term relationships are revealed. Second, using the conceptual relationship among multilingual terms discovered in the previous step as the linguistic knowledge base, conceptual content exhibiting ideas hidden beneath the multilingual texts is also mined. This is facilitated by a fuzzy multilingual text categorization algorithm. As a result, both documents and query in

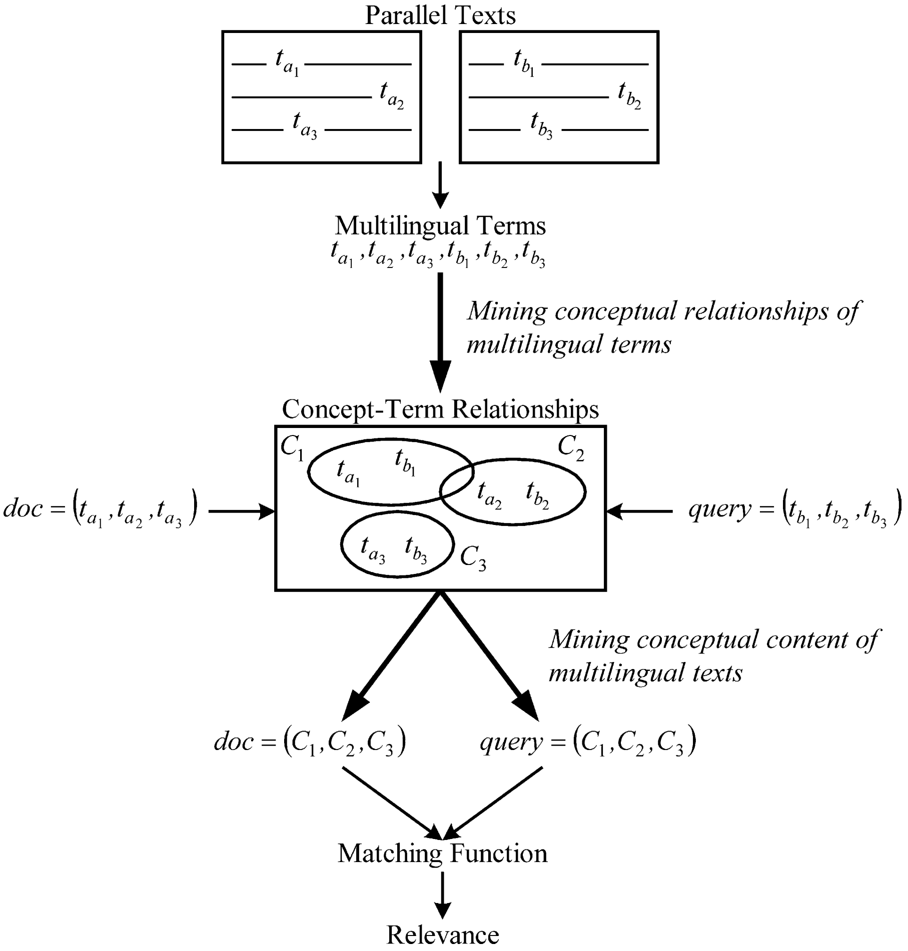
Parallel Texts
t_{a_1}
t_{a_2}
t_{a_3}
t_{b_1}
t_{b_2}
t_{b_3}
Multilingual Terms
$t_{a_1}, t_{a_2}, t_{a_3}, t_{b_1}, t_{b_2}, t_{b_3}$
Mining conceptual relationships of multilingual terms
Concept-Term Relationships
C_1
C_2
C_3
$doc = (t_{a_1}, t_{a_2}, t_{a_3})$
$query = (t_{b_1}, t_{b_2}, t_{b_3})$
Mining conceptual content of multilingual texts
$doc = (C_1, C_2, C_3)$
$query = (C_1, C_2, C_3)$
Matching Function
Relevance

Fig. 1. A multilingual text mining approach to concept-based CLTR.

different languages can then be encoded with language-independent concepts, instead of language-specific terms. As such, concept-based matching is made possible and concept-based CLTR is facilitated.

3.1. Mining the conceptual relationship of multilingual terms

Successful application of text mining in supporting monolingual information retrieval has been well reported [1]. To facilitate CLTR, our first multilingual text mining task is to discover the conceptual relationships among multilingual terms. Towards this end, a fuzzy multilingual term clustering algorithm is developed using a fuzzy clustering technique, known as fuzzy c-means [3]. Its purpose is to generate a partition of a set of multilingual terms for revealing their concept–term relationships with additional concept membership degrees. Application of the multilingual term clustering algorithm thus results in a collection of concepts represented by clusters of conceptually related multilingual terms. This collection of clusters, analogous to a multilingual thesaurus, represents a compression and reflection of the usage of multiple languages. Its importance in concept-based CLTR is in providing a concept-oriented frame of lexical reference. A cluster of conceptually related multilingual terms helps enormously in focusing solely on relevant lexical alternatives by establishing a virtual semantic domain.

Clustering is an unsupervised method for automatic class formation. It offers the advantage that a priori knowledge of classes is not required. Typically, clustering algorithms (e.g. k-means) [9] aim to maximize inter-clustering distance and minimize intra-cluster distances of some similarity measure. In the context of mining conceptual relationships among multilingual terms, clustering looks at building up clusters of semantically related multilingual terms.

As concepts tend to overlap in terms of meaning, crisp clustering algorithms like k-means that generate partitions such that each term is assigned to exactly one cluster is inadequate for representing the real textual data structure. In this aspect, fuzzy clustering methods that allow objects (terms) to be classified to more than one cluster with different membership values are more appropriate. With the application of fuzzy c-means, the resulting fuzzy multilingual term clusters, which are overlapping, will provide a more realistic representation of the multilingual semantic space.

The fuzzy c-means algorithm aims at minimizing the objective function $J(X, U, v) = \sum_{i=1}^{c} \sum_{k=1}^{n} (\mu_{ik})^m d^2(v_i, x_k)$

under the constraints $\sum_{k=1}^{n} \mu_{ik} > 0$ for all $i \in \{1,\dots,c\}$ and $\sum_{i=1}^{c} \mu_{ik} = 1$ for all $i \in \{1,\dots,c\}$ where $X = \{x_1,\dots,x_n\} \subseteq R^p$ is the set of objects; c the number of fuzzy clusters; $\mu_{ik} \in [0,1]$ the membership degree of object x_k to cluster i; v_i the prototype (cluster center) of cluster i, and $d(v_i, x_k)$ the Euclidean distance between prototype v_i and object x_k. The parameter $m > 1$ is the fuzziness index. For $m \to 1$, the clusters tend to be crisp, i.e. either $\mu_{ik} \to 1$ or $\mu_{ik} \to 0$; for $m \to \infty$, $\mu_{ik} \to 1/c$.

On the basis of the objective function optimization, fuzzy c-means is most suitable for finding optimal groupings of objects that best represent the structure of the data set. By minimizing the sum of within-group variance, the strength of associations of objects is maximized within clusters and minimized between clusters. In this aspect, fuzzy c-means is particularly useful in text mining applications, such as term clustering, where intrinsic conceptual structure and semantic relationships among terms must be revealed in order to gain knowledge for better text categorization and retrieval.

Statistical analysis of parallel corpus has been proven to be an effective means of extracting useful multilingual lexical knowledge for CLTR and this has been successfully applied to the development of translation models for CLTR [12]. Text in parallel translation is increasingly available as a result of the global explosion of the World Wide Web. Toward using the World Wide Web as a source of parallel text, effective techniques for automatically identifying parallel translated documents on the Web have also been developed [4,15].

Based on the hypothesis that semantically related multilingual terms representing similar concepts tend to co-occur with similar inter- and intra-document frequencies across a parallel corpus, fuzzy c-means can be applied to sort a set of multilingual terms into clusters (concepts) such that terms belonging to any one of the clusters (concepts) should be as similar as possible while terms of different clusters (concepts) are as dissimilar as possible in terms of the concepts they represent.

To realize the idea of mining the multilingual concept–term relationship using fuzzy c-means, a fuzzy multilingual term clustering algorithm is developed. To begin with, a set of multilingual terms, which are the objects to be clustered, is first extracted from a parallel corpus of N parallel documents. Each term is then represented as an input vector of N features where each of the N parallel documents is regarded as an input feature with each feature value representing the frequency of that term in the nth parallel document. Details of the fuzzy multilingual term clustering algorithm is presented as follows:

The fuzzy multilingual term clustering algorithm:

1. Initialize the membership values μ_{ik} of the k multilingual terms x_k to each of the i concepts (clusters) for $i = 1,\dots,c$ and $k = 1,\dots,K$ randomly such that

$$\sum_{i=1}^{c} \mu_{ik} = 1 \qquad \forall k = 1,\dots,K \tag{1}$$

and

$$\mu_{ik} \in [0,1] \qquad \forall i = 1,\dots c; \;\; \forall k = 1,\dots k \tag{2}$$

2. Calculate the concept prototype (cluster centers) v_i, using these membership values μ_{ik} :

$$v_i = \frac{\sum_{k=1}^{K} (\mu_{ik})^m x_k}{\sum_{k=1}^{K} (\mu_{ik})^m}, \qquad \forall i = 1,\dots,c \tag{3}$$

3. Calculate the new membership values μ_{ik}^{new} using these cluster centers v_i :

$$\mu_{ik}^{\text{new}} = \frac{1}{\sum_{j=1}^{c} \left(\frac{\|v_i - x_k\|}{\|v_j - x_k\|} \right)^{2/(m-1)}}, \tag{4}$$

$$\forall i = 1,\dots,c; \;\; \forall k = 1,\dots,K$$

4. If $\|\mu^{\text{new}} - \mu\| > \varepsilon$, let $\mu = \mu^{\text{new}}$ and go to step 2. Otherwise, stop.
5. Concept labeling. As a result of clustering, every multilingual term is assigned to various concepts (clusters) with various membership values. To apply these found clusters as a multilingual concept directory, concepts can be labeled by giving meaningful tags. This can be done manually using expert knowledge or by selecting the term being assigned the highest membership in each cluster for every language involved. As a result, a fuzzy partition of the multilingual term space acting as a multilingual linguistic knowledge base is now available for mining the conceptual content of all multilingual text.

3.2. Mining the conceptual content of multilingual text

Aiming at discovering the conceptual content of both multilingual document and query, our second multilingual text mining task concerns the mapping of multilingual text to concepts This process is considered a text categorization task.

Text categorization is conducted based on the cluster hypothesis [16], which states that documents with similar contents are relevant to the same concept. To accomplish the task, the crisp k-nearest neighbor algorithm [5] is among the most widely used method [11,17]. It determines the membership of an unclassified text d to a concept c by examining whether the k pre-classified texts, which are closest to d have also been classified to c.

Two problems exist in applying the crisp k-nearest neighbor algorithm in text categorization. First, when the concepts are overlapping, the contribution of a pre-classified text, which actually belongs to more than one concept with different degrees of membership, is not weighted accordingly to differentiate its uneven impact in determining the concept memberships of an unclassified text among various concepts. Second, text categorization decision based on the crisp k-nearest neighbor algorithm is arbitrary and binary. Although text is commonly associated with different concepts to various extents, with the crisp k-nearest neighbor text categorization approach, such extent of membership of a text within each concept is neither considered nor indicated. A text is categorized as either belonging or not belonging to a concept.

To overcome these problems, the fuzzy k-nearest neighbor algorithm [10] that gives a class membership degree to a new object in each class, instead of assigning it to a specific class, is more appropriate for multilingual text categorization. In the fuzzy k-nearest neighbor algorithm, the assignment of the membership degree to an unclassified object depends on the proximity of the unclassified object to its nearest neighbors and the strength of the membership of these neighbors in the corresponding classes. This provides the advantage of avoiding an arbitrary assignment with the additional benefit of a degree of relevance from the resulting categorization.

Fuzzy multilingual text categorization concerns the assignment of a membership value in the range of [0,1] to each entry of the categorization matrix as illustrated in Fig. 2, where $C = \{c_1, \ldots, c_m\}$ is a set of concepts, $D = \{d_1, \ldots, d_n\}$ is a set of multilingual texts to be categorized, and $\mu_i(d_j) \in [0, 1]$ is the degree of membership of text d_j in concept c_i.

To apply the fuzzy k-nearest neighbor classification algorithm to the task of multilingual text categorization, decisions regarding the set of pre-classified multilingual texts and the value of parameter k must be made. For many operation-oriented text categorization tasks such as document routing and information filtering, a set of pre-classified texts determined by the user or the operation is always necessary. These pre-classified texts are used as training samples for the text classifier to learn the specific text categorization task. However, multilingual text categorization may not require a set of pre-classified texts. This is because the categorization of multilingual texts by concepts is a concept-oriented decision. It is made on the basis of a text's conceptual relevance to a concept and not on how a similar text is previously categorized during a sample operation. As long as the conceptual context of both concepts and texts are well represented, a decision on the conceptual categorization of multilingual text can then be made.

	d_1	…	…	d_j	…	…	d_n
c_1	$\mu_1(d_1)$	…	…	$\mu_1(d_j)$	…	…	$\mu_1(d_n)$
…	…	…	…	…	…	…	…
c_i	$\mu_i(d_1)$	…	…	$\mu_i(d_j)$	…	…	$\mu_i(d_n)$
…	…	…	…	…	…	…	…
c_m	$\mu_m(d_1)$	…	…	$\mu_m(d_j)$	…	…	$\mu_m(d_n)$

Fig. 2. Categorization matrix of fuzzy multilingual text classification.

In fact, given the result of the fuzzy multilingual term clustering in the previous stage, concept memberships of all multilingual terms are already known. Interpreting each term as a document containing a single term, a virtual set of pre-classified multilingual texts is readily available. Given the concept membership of every multilingual term, the class membership values of every single-term document in the corresponding concepts are also known. For fuzzy multilingual text categorization, conceptual specifications provided by fuzzy multilingual term clustering are considered reasonably sufficient and relevant for supporting the decision.

To categorize multilingual text using the fuzzy k-nearest neighbor algorithm, a threshold k specifying the number of neighboring multilingual texts to be considered for calculating the membership degree $\mu_i(d_j)$ for an unclassified text d_j in concept c_i should also be determined. In our multilingual text categorization problem, the nearest neighbor to an unclassified text with k index terms will be the k single-term virtual documents where each of them contains one of the unclassified text's k index terms, respectively. This is based on the assumption that a single-term document should contain at least one index term of another document to be considered related or conceptually close. As a result, the categorization decision of an unclassified text with k index terms will be a function of its distance from its k single-term neighboring documents (each containing one of the k index term) and the membership degree of these k neighboring documents in the corresponding concepts. Details of the fuzzy multilingual text categorization algorithm are presented as follows:

The fuzzy multilingual text categorization algorithm:

1. Determine the k neighboring texts for text d_j.
2. Compute $\mu_i(d_j)$ using:

$$\mu_i(d_j) = \frac{\sum_{s=1}^{k} \mu_i(d_s)\left(\dfrac{1}{\|d_j - d_s\|^{2/(m-1)}}\right)}{\sum_{s=1}^{k}\left(\dfrac{1}{\|d_j - d_s\|^{2/(m-1)}}\right)}, \tag{5}$$

$$\forall i = 1, \ldots, m$$

where $\mu_i(d_s)$ are the membership degrees of the kth nearest neighboring sample text d_s in concept c_i and m is the weight determining each neighbor's contribution to $\mu_i(d_j)$. When m is 2, the contribution of each neighboring text is weighted by the reciprocal of its distance from the text being categorized. As m increases, the neighbors are more evenly weighted, and their relative distances from the text being categorized have less effect. As m approaches 1, the closer neighbors are weighted far more heavily than those farther away, which has the effect of reducing the number of texts that contribute to the membership value of the text being categorized. Usually, $m = 2$ is chosen.

The result of this computation assigns a degree of conceptual relevance to each text being categorized to a particular concept. Hence, when any multilingual text is being categorized to every existing concept, its degrees of conceptual relevance to all concepts are known and its conceptual content is totally revealed.

3.3. Concept-based cross-lingual text retrieval

Given a set of query terms, a user engaged in CLTR expects to receive the most relevant set of documents that discusses the concept encapsulated in the query terms, rather than documents that just contain the bag of translations of the original query terms which may not be truly relevant to the information needed due to translation ambiguities. To support such concept-based CLTR, a concept-based matching between documents and query is required.

The basic idea of concept-based matching is: when both the language-specific documents and query are represented as a function of language-independent concepts, they can be compared. With the fuzzy multilingual text categorization algorithm, any document in any language can be mapped to the relevant concepts with a membership degree indicating its strength of belonging. Regarding the query, it is interpreted as the representation of an ideal document specified by the user. In this way, it can also be categorized using the fuzzy multilingual text categorization algorithm as other documents. In other words, regardless of the language used for a query, it will be assigned to its relevant concepts with the corresponding membership degrees, based on its conceptual similarity with the concepts.

Given all conceptual membership values, both document d and query q can be represented as a function of concepts $c_k \in \{c_1, \ldots, c_m\}$ by a vector, as follows

$$d = (\mu_{c_1}(d), \mu_{c_2}(d), \ldots, \mu_{c_m}(d)) \tag{6}$$

$$q = (\mu_{c_1}(q), \mu_{c_2}(q), \ldots, \mu_{c_m}(q)) \tag{7}$$

where $\mu_{c_k}(d)$ and $\mu_{c_k}(q)$ are membership degrees of document d and query q in concept c_k, respectively. To determine the similarity between concept q and document d, a similarity function based on the cosine similarity measure is defined as:

$$\mathrm{sim}(q, d) = \frac{\sum_{k=1}^{m} \mu_{c_k}(q) \cdot \mu_{c_k}(d)}{\sqrt{\sum_{k=1}^{m} \mu_{c_k}(q)^2 \sum_{k=1}^{m} \mu_{c_k}(d)^2}} \tag{8}$$

This similarity value lies between 0 and 1 and depends on the membership degrees of matching concepts in the vectors. Finally, the results of this similarity computation are used to produce a ranked list of multilingual documents relevant to a particular query with the most relevant one appearing at the top.

One major benefit of concept-based CLTR is that documents and query in different languages are mapped without the need of either document or query translation. The user's original query is compared directly against the documents based on the concepts they exhibit. This avoids the inclusion of irrelevant senses of query terms resulting from translation ambiguity to be included as legitimate in the retrieval process. As a result, less irrelevant documents are returned.

4. An illustration

We use a sample parallel corpus to illustrate the multilingual text mining approach to concept-based CLTR. The corpus includes nine pairs of parallel documents, which are selected articles related to Hong Kong economy written in both English and Chinese.

According to our approach, the first task is to discover the concept–term relationships for concept-based CLTR. To begin with, meaningful terms from both languages are extracted by referring to an English wordlist and Chinese wordlist, respectively. After excluding the most frequently occurring ones, 51 terms, including 26 in English and 25 in Chinese, are retained. They are used as the set of multilingual terms for characterizing the major concepts described in the parallel corpus. These 51 terms and nine parallel documents then form a 51 × 9 matrix where each row is a term vector and each column corresponds to a parallel document. The feature value of a term vector in the nth column corresponds to its frequency in the nth document. The term vectors are used as the input to the fuzzy multilingual term clustering algorithm, as presented in Section 3.1.

When the fuzzy c-means learning process is completed, three clusters are found. Each cluster represents a major concept discovered from the parallel corpus. In terms of the concept–term relationship, each term is also mapped to every cluster found with different membership values indicating its state of belonging to every concept. Given these membership values, each concept is then interpreted as a weighted vector of its conceptually related multilingual terms, as illustrated in Fig. 3. Here, each concept is labeled manually after analyzing the weight vector of each cluster. The three concept classes are 'property market 物業市場', 'foreign exchange market 外匯市場' and 'Hong Kong economy 香港經濟', respectively.

As the multilingual linguistic knowledge base, the fuzzy multilingual term clustering result is applied to discover

Property market 物業市場	μ	Foreign exchange market 外匯市場	μ	Hong Kong economy 香港經濟	μ
銀行信貸	0.8743	exchange rate	0.8995	Asian Currency Turmoil	0.8716
property market	0.8725	匯率	0.8995	亞洲貨幣風潮	0.8716
物業市場	0.8725	美元	0.8955	unemployment rate	0.8616
Bank credit	0.8487	US dollar	0.8817	失業率	0.8616
rental yield	0.8139	Hong Kong dollar	0.6998	depreciation	0.8499
租金回報	0.8139	港元	0.6841	devaluation	0.8436
end-user	0.7061	inflation	0.6603	recession	0.8358
用家	0.7061	通脹	0.6603	經濟衰退	0.8358
speculators	0.6607	arbitrage	0.5453	Asian currencies	0.8250
炒家	0.6607	利率	0.5233	亞洲貨幣	0.8250
mortgage	0.5773	interest rate	0.5199	貶值	0.7749
按揭	0.5773	套戥	0.5021	appreciation	0.7087
物業價格	0.5544	Export	0.3333	升值	0.6988
property prices	0.5342	出口	0.3329	GDP	0.6549
income	0.4263	Import	0.2042	本地生產總值	0.6549
入息	0.4263	入口	0.2042	import	0.5298
investment	0.4091	經濟復甦	0.1965	入口	0.5298
投資	0.4091	investment	0.1930	recovery	0.5239
經濟復甦	0.2868	投資	0.1930	income	0.5183
recovery	0.2852	recovery	0.1909	入息	0.5183
出口	0.2758	property prices	0.1673	經濟復甦	0.5167
export	0.2741	物業價格	0.1555	investment	0.3979
import	0.2660	mortgage	0.1479	投資	0.3979
入口	0.2660	按揭	0.1479	export	0.3926
interest rate	0.2398	GDP	0.1185	出口	0.3913
利率	0.2382	本地生產總值	0.1185	property prices	0.2985
GDP	0.2266	speculators	0.1051	套戥	0.2917
本地生產總值	0.2266	炒家	0.1051	物業價格	0.2901
套戥	0.2063	升值	0.0990	mortgage	0.2748
升值	0.2022	appreciation	0.0981	按揭	0.2748
arbitrage	0.1936	貶值	0.0719	arbitrage	0.2610
appreciation	0.1931	income	0.0555	end-user	0.2529
貶值	0.1532	入息	0.0555	用家	0.2529
inflation	0.1497	Asian currencies	0.0532	interest rate	0.2403
通脹	0.1497	亞洲貨幣	0.0532	利率	0.2384
港元	0.1457	rental yield	0.0525	speculators	0.2342
Hong Kong dollar	0.1366	租金回報	0.0525	炒家	0.2342
recession	0.1219	bank credit	0.0439	inflation	0.1900
經濟衰退	0.1219	depreciation	0.0425	通脹	0.1900
Asian currencies	0.1219	recession	0.0423	港元	0.1702
亞洲貨幣	0.1219	經濟衰退	0.0423	Hong Kong dollar	0.1636
devaluation	0.1163	end-user	0.0410	rental yield	0.1336
depreciation	0.1075	用家	0.0410	租金回報	0.1336
unemployment rate	0.1022	Devaluation	0.0401	bank credit	0.1073
失業率	0.1022	Asian Currency Turmoil	0.0389	property market	0.0927
Asian Currency Turmoil	0.0895	亞洲貨幣風潮	0.0389	物業市場	0.0927
亞洲貨幣風潮	0.0895	unemployment rate	0.0362	銀行信貸	0.0913
US dollar	0.0520	失業率	0.0362	US dollar	0.0663
美元	0.0467	property market	0.0348	美元	0.0579
exchange rate	0.0440	物業市場	0.0348	exchange rate	0.0565
匯率	0.0440	銀行信貸	0.0344	匯率	0.0565

Fig. 3. Concepts represented by fuzzy clusters of multilingual terms.

the conceptual content of both English and Chinese texts. Applying our fuzzy multilingual text categorization algorithm as presented in Section 3.2, the following multilingual queries and documents are categorized and their degrees of conceptual relevance with respect to each concept are summarized in Fig. 4.

Queries:

Q_e = (property prices, mortgage)
Q_c = ((物業價格, 按揭))

Documents:

D_1 = ((投資, 租金回報))
D_2 = (exchange rate, interest rate)
D_3 = (investment, GDP)

As the conceptual content represented by the concept membership degrees is revealed, concept-based matching between the documents and queries can be carried out using our method as described in Section 3.3. As presented in

	Property market 物業市場	Foreign exchange market 外匯市場	HK economy 香港經濟
Q_e	0.556	0.15	0.29
Q_c	0.556	0.15	0.29
D_1	0.61	0.12	0.265
D_2	0.14	0.71	0.15
D_3	0.32	0.155	0.525

Fig. 4. Result of fuzzy multilingual text categorization.

Q_e	Q_c
D_1(0.996)	D_1(0.996)
D_3(0.864)	D_3(0.864)
D_2(0.477)	D_2(0.477)

Fig. 5. Ranked list of documents returned by concept-based CLTR.

Fig. 5, both the English and Chinese queries, which are translation versions of each other, receive the same ranked list of documents. This indicated that the concept-based CLTR is successful in retrieving documents based on the concepts the documents exhibit rather than the terms they contain.

5. Developing a multi-agent system for Web cross-lingual text retrieval

We have used this multilingual text mining approach for developing a multi-agent system to facilitate concept-based CLTR on the Web. This multi-agent system, as shown in Fig. 6, consists of an interface agent, a multilingual ontology agent and an information-gathering agent.

The interface agent is the agent interacting with the user. It accepts query from the user and returns the user with the relevant information. It decomposes the CLTR main task into several subtasks and delegates them to other agents. The information-gathering agent is acting as a Web crawler that fetches multilingual Web documents from the Web. It automatically traverses the Web for collecting multilingual Web documents and generating the concept-based search indexes. The multilingual ontology agent provides the knowledge of the multilingual concept–term relationship and the method of multilingual text categorization. With the multilingual linguistic knowledge, it helps the interface agent to transform the user's query from language-specific terms into language-independent concepts. Similarly, it also helps the information-gathering agent to generate concept-based search indexes for the multilingual Web documents it collects.

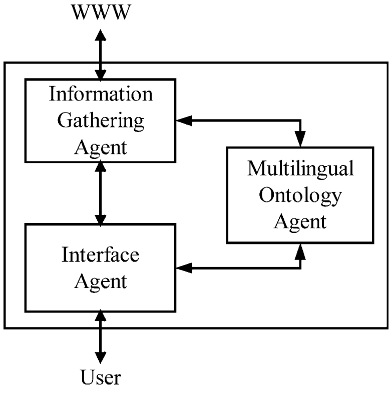

Fig. 6. Architecture of a multi-agent system for Web concept-based CLTR.

The Web CLTR process is performed with the cooperation among these three agents. When the user submits a query to the interface agent, it delegates the multilingual text categorization task to the multilingual ontology agent. The multilingual ontology agent will apply its knowledge of multilingual concept–term relationship and the multilingual text categorization method to extract relevant concepts from the user's query. The interface agent then sends the concept-based query returned by the multilingual ontology agent to the information-gathering agent. The information-gathering agent will match the concepts of the user's query against conceptual content of the multilingual Web document with reference to the concept-based search indexes. Finally, the interface agent presents all relevant multilingual Web documents as a ranked list to the user for evaluation.

6. Conclusion

We believe our study has contributed to Web intelligence by generating insights for research towards development of multilingual search engines and Web directories. The multilingual text mining approach has suggested an exciting new direction for discovering interesting knowledge, which is useful for developing multilingual text management systems. In particular, our multilingual text mining approach for automatically discovering the multilingual linguistic knowledge contributes to CLTR by providing a more affordable alternative to the costly manually constructed linguistic resources. By exploiting a parallel corpus covering multiple languages, the automatic construction of language-independent concept space capturing all conceptual relationships among multilingual terms is accomplished. By making multilingual document and query comparable within a common semantic space, concept-based CLTR is realized. Without restricting to bilingual lexical transfer, this concept-based approach to CLTR is significant in enhancing support to global knowledge exploration by allowing multilingual documents relevant to the concept of a query but not necessarily containing the translation equivalents of the query terms to be identified. Finally, the multi-agent architecture for Web concept-based CLTR has introduced a practical framework to realize knowledge discovery from the multilingual World Wide Web.

References

[1] H. Ahonen, O. Heinonen, M. Klemettinen, A.I. Verkamo, Applying Data Mining Techniques in Text Analysis, Report C-1997-23, Department of Computer Science, University of Helsinki, 1997.

[2] L. Ballesteros, W.B. Croft, Statistical methods for cross-language information retrieval, in: G. Grefenstette (Ed.), Cross-Language Information Retrieval, Kluwer Academic Publishers, Boston, 1998, pp. 23–40.

[3] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.

[4] J. Chen, J.-Y. Nie, Web parallel text mining for Chinese–English cross-language information retrieval, in: Proceedings of RIAO2000 Content-Based Multimedia Information Access, CID, Paris, 2000, http://133.23.229.11/~ysuzuki/Proceedingsall/RIAO2000/.

[5] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1973.

[6] T.E. Dunning, M.W. Davis, Multilingual information retrieval, Memoranda in Cognitive and Computer Science, MCCS-93-252, New Mexico State University, Computer Research Laboratory, 1993.

[7] J. Gilarrans, J. Gonzalo, F. Verdejo, An approach to conceptual text retrieval using the EuroWordNet multilingual semantic database, in: Proceedings of the AAAI Symposium on Cross-Language Text and Speech Retrieval, American Association for Artificial Intelligence, March 1997, http://www.clis.umd.edu/dlrg/filter/sss/papers/.

[8] D. Hull, G. Grefenstette, Querying across languages: a dictionary-based approach to multilingual information retrieval, in: Proceedings of the 19th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, 1996, pp. 49–57.

[9] A.K. Jain, R.C. Dubes, Algorithms for Clustering Data, Prentice-Hall, Upper Saddle River, NJ, 1988.

[10] J.M. Keller, M.R. Gray, J.A. Givens, A fuzzy *k*-nearest neighbor algorithm, IEEE Transactions of Systems, Man and Cybernetics 15 (4) (1985) 580–585.

[11] W. Lam, C.Y. Ho, Using a generalized instance set for automatic text categorization, in: Proceedings of the 21st Annual International ACM SIGIR Conference in Research and Development in Information Retrieval, ACM Press, New York, 1998, pp. 81–89.

[12] M.L. Littman, S.T. Dumais, T.K. Landaur, Automatic cross-language information retrieval using latent semantic indexing, in: G. Grefenstette (Ed.), Cross-Language Information Retrieval, Kluwer Academic Publishers, Boston, 1998, pp. 51–62.

[13] J.-Y. Nie, M. Simard, P. Isabelle, R. Durand, Cross-language information retrieval based on parallel texts and automatic mining of parallel text from the web, in: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, New York, 1999, pp. 74–81.

[14] D. Oard, Alternative approaches for cross-language text retrieval, in: Proceedings of the AAAI Symposium on Cross-Language and Speech Retrieval, American Association for Artificial Intelligence, March 1997, http://www.clis.umd.edu/dlrg/filter/sss/papers/.

[15] P. Resnik, Parallel strands: A preliminary investigation into mining the web for bilingual text, in Proc. The Third Conference of the Association for Machine Translation in the Americas (AMTA-98), Lecture Notes in Artificial Intelligence, vol. 1529, Springer, Berlin, 1998, pp. 28–31.

[16] C.J. Van Rijsbergen, Information Retrieval, Butterworth, London, 1972.

[17] Y. Yang, Expert network: effective and efficient learning from human decisions in text categorization and retrieval, in: Proceedings of the Seventh Annual International ACM SIGIR Conference in Research and Development in Information Retrieval, ACM Press, New York, 1994, pp. 13–22.