Fuzzy-neuro Web-Based Multilingual Knowledge Management

Rowena Chau[1], Chung-Hsing Yeh[1], and Kate Smith-Miles[2]

[1] Clayton School of Information Technology,
Faculty of Information Technology,
Monash University, Clayton, Victoria 3800, Australia
{rowena.chau, chunghsing.yeh}@infotech.monash.edu.au
[2] School of Engineering and Information Technology
221 Burwood Highway, Burwood, Victoria 3125, Australia
katesm@deakin.edu.au

Abstract. This paper presents new methodology towards the automatic development of multilingual Web portal for multilingual knowledge discovery and management. It aims to provide an efficient and effective framework for *selecting* and *organizing* knowledge from voluminous linguistically diverse Web contents. To achieve this, a concept-based approach that incorporates text mining and Web content mining using neural network and fuzzy techniques is proposed. First, a concept-based taxonomy of themes, which will act as the hierarchical backbone of the Web portal, is automatically generated. Second, a concept-based multilingual Web crawler is developed to intelligently harvest relevant multilingual documents from the Web. Finally, a concept-based multilingual text categorization technique is proposed to organize multilingual documents by concepts. As such, correlated multilingual Web documents can be gathered/filtered/organised/ based on their semantic content to facilitate high-performance multilingual information access.

1 Introduction

A portal is a Web site that serves as an entry point into the Web information space. A typical portal has a hierarchical subject catalogue associated with categories of relevant documents, and/or a search engine combined with other services and interactive content. The rapid expansion of the WWW has made electronically accessible information available in almost all natural languages. With majority of this Web data being unstructured text [1], effective multilingual Web portal development technology capable of automatically discovering and organizing relevant multilingual Web documents thus holds the key to exploit the vast human knowledge hidden beneath this largely untapped multilingual text. In this paper, a concept-based approach towards the automatic development of multilingual Web portal, is presented. This approach represents an integration of the text mining and Web content mining paradigms using artificial intelligence techniques.

In the rest of this paper, Section 2 will introduce the concept-based approach and the rationale behind it. This is followed by the generation of a hierarchical

L. Wang et al. (Eds.): FSKD 2006, LNAI 4223, pp. 1229 – 1238, 2006.
© Springer-Verlag Berlin Heidelberg 2006

concept-based theme taxonomy, which will act as the backbone of the multilingual Web portal in Section 3. Section 4 will give an account the process of concept-based multilingual Web crawling, and Section 5 will explain the development a concept-based multilingual text categorization system for organizing multilingual Web documents by their concept-based thematic content. Finally, Section 6 will conclude the work.

2 The Concept-Based Approach

The focus of this research is to automate the development of a concept-based multilingual Web portal capable of discovering, filtering and organizing multilingual documents relevant to a set of themes underlying a target area of information interest. By associating correlated multilingual documents to their corresponding themes, content-focused information exploration is facilitated. As such, the objective of enabling global knowledge acquisition is achieved.

On discovering and organizing multilingual Web documents relevant to the themes of a target area of interest, multilingual Web portal development is a typical application of text mining and Web content mining that requires an understanding of the natural language text. Text mining concerns the discovery of knowledge from textual data by applying various knowledge discovering techniques. Text clustering and categorization are the two most extensively studied areas in text mining [2]. Using appropriate clustering techniques, terms (i.e. keywords) or documents related to the same concepts can be effectively grouped by their semantic similarity and thus revealing their conceptual content. Emerged as an area of text mining specific to Web documents focusing on analyzing and deriving meaning from textual collection on the Internet [3], Web content mining has attracted much research attention in recent years [4]. Popular Web content mining activities include the exploitation of retrieval results returned by search engines and automatically harvesting relevant Web documents using intelligent Web crawlers [5].

Currently, both text mining and Web content mining technologies are still primarily focused on processing monolingual text. The challenge of discovering knowledge from textual data, which are significantly linguistically diverse, has early been recognised by text mining research [6]. In a monolingual environment, the conceptual content of documents can be discovered by directly detecting patterns of frequent features (i.e. terms) without precedential knowledge of the concept-term relationship. Documents containing an identical known term pattern thus share the same concept. However, in a multilingual environment, *vocabulary mismatch* among diverse languages implies that documents exhibiting similar concept will not contain identical term patterns. This *feature incompatibility* problem thus makes the inference of conceptual contents using term pattern matching inapplicable.

Due to the feature incompatibility problem contributed by the vocabulary mismatch phenomenon across multiple languages, documents describing a common theme in different languages are represented by different sets of features (i.e. terms) in separate feature spaces. This language-specific representation has made multilingual text incomparable. Therefore, monolingual text representation techniques that rely on shared syntactic terms will not work for multilingual text/Web content mining

application, such as multilingual Web portal development. To overcome this problem, a concept-based approach is proposed. The idea is: by generating a taxonomy of language-independent concept-based thematic profiles encapsulating all feature variations among multiple languages, semantically-relevant multilingual Web documents can then be scanned, filtered, retrieved and categorized regardless of the syntactic terms they contain. The automatic development of multilingual Web portal typically involves three major tasks, including (1) concept-based theme taxonomy generation, (2) concept-based multilingual Web crawling and (3) concept-based multilingual text categorization.

3 Concept-Based Theme Taxonomy Generation

Web portal development concerns the organization of documents into a pre-defined taxonomy of related themes. In a multilingual environment, documents about a common theme used synonymous terms in different languages. To generate taxonomy of themes capable of accommodating these linguistic variations, concept-based theme profiles, which are linguistically comprehensive as well as semantically rich, are necessary. Each theme profile should be capable of capturing its corresponding semantic content while encapsulating the syntactic variations among term usage. Towards this end, a novel concept-based taxonomy generation approach using a training parallel corpus (i.e. a collection of translated documents) is proposed.

This concept-based taxonomy generation approach involves three constituent tasks: (a) Multilingual term clustering (b) Concept-based document representation (c) Concept-based document clustering. *Multilingual term clustering* aims at forming a set of concepts by grouping related multilingual terms extracted from the parallel documents based on their semantic similarity using self-organizing maps [10]. *Concept-based document representation*, in turn, takes these concepts as a kind of language-independent indexing features to produce a concept-based representation (i.e. a concept-based document vector) for characterizing each training documents. Finally, *concept-based multilingual document clustering*, given the concept-based document vectors as inputs, generates a concept-based theme taxonomy by clustering thematically-related training documents using a hierarchical clustering algorithm. As such, this taxonomy, providing a hierarchical schema, will act as the backbone of the multilingual web portal to guide the collection/filtering/organization of relevant multilingual Web documents for the multilingual Web portal development.

3.1 Multilingual Term Clustering

To cluster multilingual terms with an aim of acquiring a set of concepts by grouping related multilingual terms, parallel corpora containing sets of documents and their translations in multiple languages are ideal sources of multilingual lexical information. Parallel documents basically contain identical concepts expressed by different sets of terms. Therefore, multilingual terms used to describe the same concept tend to occur with very similar inter- and intra-document frequencies across a parallel corpus. An analysis of paired documents has been used to infer the most likely translation of terms between languages in the corpus [7,8,9]. As such,

co-occurrence statistics of multilingual terms across a parallel corpus can be used to determine clusters of conceptually related multilingual terms.

To acquire domain-specific concepts, which can effectively characterize the knowledge context of a multilingual Web portal to be developed, a training parallel corpus from the corresponding subject domain is used. Given a parallel corpus *D* consisting *N* pairs of parallel documents, meaningful terms from every languages covered by the corpus are extracted. They form the set of multilingual terms for the multilingual term clustering process. Each term is represented by an *n*-dimensional term vector. Each feature value of the term vector corresponds to the weight of the *nth* document indicating the significance of that document in characterizing the meaning of the term. Parallel documents which are translated versions of one another within the corpus, are considered as the same feature. To determine the significance of each document in characterising the contextual content of a term based on the term's occurrences, the following weighting scheme is used. It calculates the feature value w_{kp} of a document d_p for $p = 1,...,N$ in the vector of term t_k.

$$w_{kp} = \begin{cases} \dfrac{tf_{kp} \cdot log\left(\dfrac{|T|}{|d_p|}\right)}{\sqrt{\sum_{q=1}^{N}\left(tf_{kq} \cdot log\left(\dfrac{|T|}{|d_q|}\right)\right)}} & for\ tf_{kp} > 0 \\ 0 & for\ tf_{kp} = 0 \end{cases} \quad (1)$$

where

tf_{kp} is the occurrence of term t_k in document d_p;

$log\left(\dfrac{|T|}{|d_p|}\right)$ is the inverse term frequency of document d_p; $|T|$ is the number of terms in the whole collection, and $|d_p|$ is the number of terms in document d_p. The longer the document d_p, the smaller the inverse term frequency;

$1\Big/\sqrt{\sum_{q=1}^{N}\left(tf_{iq} \cdot log\left(\dfrac{|T|}{|d_q|}\right)\right)}$ is the normalisation factor. With this normalisation factor, the feature value relating a document to a term t_k is reduced according to the total number of documents in which the term occurs.

When contextual contents of every multilingual term are well represented, they are used as the input into the self-organizing map algorithm [10] for multilingual term clustering. Let $\mathbf{x}_i \in R^N$ $(1 \le i \le M)$ be the term vector of the i^{th} multilingual term, where *N* is the number of documents in the parallel corpus for a single language (i.e. the total number of documents in the parallel corpus divided by the number of languages supported by the corpus) and *M* is the total number of multilingual terms. The self-organizing map algorithm is applied to discover the multilingual term

clusters, using these term vectors as the training input to the map. The map consists of a regular grid of nodes. Each node is associated with an *N*-dimensional model vector. Let $\mathbf{m}_j = [m_{jn} | 1 \le n \le N]$ ($1 \le j \le G$) be the model vector of the j^{th} node on the map. The algorithm for multilingual term clustering is given below.

Step 1: Select a training multilingual term vector $\mathbf{x}_i$ at random.
Step 2: Find the winning node *s* on the map with the vector $\mathbf{m}_s$ which is closest to $\mathbf{x}_i$ such that

$$\|\mathbf{x}_i - \mathbf{m}_s\| = \min_j \|\mathbf{x}_i - \mathbf{m}_j\| \tag{2}$$

Step 3: Update the weight of every node in the neighbourhood of node *s* by

$$\mathbf{m}_t^{new} = \mathbf{m}_t^{old} + \alpha(t)(\mathbf{x}_i - \mathbf{m}_t^{old}) \tag{3}$$

where $\alpha(t)$ is the gain term at time *t* ($0 \le \alpha(t) \le 1$) that decreases in time and converges to 0.
Step 4: Increase the time stamp *t* and repeat the training process until it converges.

After the training process is completed, each multilingual term is mapped to a grid node closest to it on the self-organizing map. A partition of multilingual term space, represented by a multilingual term cluster map, is thus formed. This process corresponds to a projection of the multi-dimensional term vectors onto an orderly two-dimensional concept space where the proximity of the multilingual terms is preserved as faithfully as possible. Consequently, conceptual similarities among multilingual terms are explicitly revealed by their locations and neighbourhood relationships on the map. Multilingual terms that are synonymous are associated to the same node. In this way, conceptual related multilingual terms are organised into term clusters, representing all existing concepts, within a common semantic space. The problem of feature incompatibility among multiple languages is thus overcome.

3.2 Concept-Based Document Representation

A taxonomy of themes relevant to a particular domain of interest can be generated by finding hierarchy of thematically-related document clusters using a domain-specific training corpus. However, document clustering depends heavily on the document representation (i.e. indexing) scheme. To form the taxonomy of themes that effectively reflects the conceptual content among documents, a suitable method for document indexing must be devised. Contextual contents of documents need to be expressed explicitly in a computationally meaningful way.

In information retrieval, several approaches for document indexing and representation have been suggested. Among them, the vector space model [11] represents documents conveniently as vectors in a multi-dimensional space defined by a set of language-specific index terms. Each element of a document vector corresponds to the weight (or occurrence) of one index term. However, in a multilingual environment, the direct application of the vector space model is infeasible due to the

feature incompatibility problem. Multilingual index terms characterising documents of different languages exist in separate vector spaces.

To overcome the problem, a better representation of document contents incurporating information about semantic/conceptual relationships among multilingual index terms is desirable. Towards this end, the multilingual term cluster map obtained in Section 3.1 is used. On the multilingual term cluster map, semantically related multilingual terms have been organised into clusters (i.e. concepts). These concepts, associating semantically related multilingual terms, are thus used to index multilingual documents in place of the documents' original language-specific index terms. As such, a concept-based document vector that explicitly expresses the conceptual context of a document regardless of its language can be obtained. The term-based document vector of the vector space model, which suffers from the feature incompatibility problem, can now be replaced with the language-independent concept-based document vector.

To realize this, every multilingual document is indexed by mapping its text, term by term, onto the multilingual term cluster map. This is done by counting the occurrence of each term on the multilingual term cluster map at the node to which that term has been associated. This statistics of term cluster (i.e. concept) occurrences is then a kind of transformed 'index' of the multilingual document to produce a language-independent concept-based document vector. Concept-based document vectors thus obtained are essential for enabling concept-based multilingual document clustering. Using these concept-based document vectors as input to some appropriate clustering algorithm, multilingual documents, which are originally syntactically incomparable can then be grouped based on the conceptual similarity they convey.

3.3 Concept-Based Multilingual Document Clustering

The hierarchical concept-based theme taxonomy is generated via concept-based multilingual document clustering. With the application of the complete linkage hierarchical clustering algorithm [12] using the concept-based document vectors of the training parallel corpus as inputs, the concept-based multilingual document clustering algorithm is given below.

Step 1: Construct a $K \times Z$ concept-document matrix, *CD*, to represent *K* concepts, c_k ,obtained from the multilingual term clustering process, and the parallel corpus, *D*, of *Z* documents, d_z , as:

$$CD = \begin{array}{c} \\ c_1 \\ c_2 \\ c_3 \\ \cdot \\ \cdot \\ \cdot \\ c_K \end{array} \begin{array}{c} \begin{array}{cccccccc} d_1 & d_2 & d_3 & . & . & . & d_Z \end{array} \\ \begin{bmatrix} w_{11} & w_{12} & w_{13} & . & . & . & w_{1z} \\ w_{21} & w_{22} & w_{23} & . & . & . & w_{2z} \\ w_{31} & w_{32} & w_{33} & . & . & . & w_{3z} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ w_{k1} & w_{k2} & w_{k3} & . & . & . & w_{kz} \end{bmatrix} \end{array} \tag{4}$$

where w_{kz} is the occurrence of the concept c_k in document d_z

Step 2: Obtain a document association matrix $D \times D$ such that

$$D \times D = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ . \\ . \\ . \\ d_z \end{array} \begin{array}{c} \begin{array}{ccccccc} d_1 & d_2 & d_3 & . & . & . & d_Z \end{array} \\ \begin{bmatrix} A_{11} & A_{12} & A_{13} & . & . & . & A_{1z} \\ A_{21} & A_{22} & A_{23} & . & . & . & A_{2z} \\ A_{31} & A_{32} & A_{33} & . & . & . & A_{3z} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ A_{k1} & A_{k2} & A_{k3} & . & . & . & A_{kz} \end{bmatrix} \end{array} \tag{5}$$

where A_{xy} is the coefficient of association between each pair of documents $d_x, d_y \in D$ calculated using cosine similarity measure as defined by the following equation:

$$A_{xy} = \frac{\sum_{k=1}^{K} w_{kx} \cdot w_{ky}}{\sqrt{\sum_{k=1}^{K} w_{kx}^2} \cdot \sqrt{\sum_{k=1}^{K} w_{ky}^2}} \tag{6}$$

Step 3: Apply the complete linkage hierarchical clustering algorithm to the document association matrix $D \times D$ to determine the document clusters. The hierarchy of document clusters resulted from this process is thus a taxonomy of themes representing the contextual content relevant to the domain of the training parallel corpus. This taxonomy, providing a hierarchical schema, will thus act as the backbone of the multilingual Web portal to facilitate organization of theme-related multilingual Web documents collected by the Web crawler.

4 Concept-Based Multilingual Web Crawling

Web crawler plays the significant role in Web portal development by automatically collecting Web documents. Web crawling focusing on a collection of themes requires a set of good seed URLs that point to other potentially relevant documents to guide the process. In this research, a concept-based multilingual Web crawler, incorporating a novel approach to deduce a set of seed URLs using existing search engine is developed.

Related Web documents are often connected by their in- and out- links [13]. Web crawlers traverse the Web to collect related documents by following links. To start crawling the Web for documents relevant to a theme, a set of seed URLs are required

by the Web crawler. In a monolingual dimension, it is sufficient to use any documents, written in a particular language, previously considered as relevant to the theme as the seeds for gathering other related documents exist in the same language. However, related documents written in different languages are rarely linked. To make a Web crawler capable of collecting and retrieving relevant documents in various languages, a set of seed URLs in every language must be made available.

To address the above issue, a novel concept-based approach is proposed. The idea is: for every language, we use the term that is most prototypical to a concept to run a Web search. Top-ranked Web documents returned by the search engine are then used as the seed URLs to initialise the concept-based Web crawling. To realise this, identification of the multilingual terms, which are prototypical to each concept becomes the major challenge. Recalling that the multilingual term cluster map obtained in Section 3.1 has already encoded the multilingual concept-term relationships by associating all multilingual terms with the concepts to which they belong, it thus provides the essential clues for effectively inferring such concept-specific prototypical multilingual terms. To identify a term that is prototypical with respect to a concept for every language, we find the term, one for each language, which is closest to the output node of a concept on the self-organizing map. This term is then submitted as a query to the search engine to retrieve a set of the most relevant (e.g. top 10) Web document in each language. These documents then become the seed URLs for the Web crawler to collect other related documents describing similar concepts in various languages on the Web. This approach is to ensure the Web crawler will explore the Web context diversely and yet still remained conceptually focused. Multilingual Web documents collected by the Web crawler are then passed on to a concept-based multilingual text categorization algorithm for organizing into the taxonomy to facilitate global knowledge discovery.

5 Concept-Based Multilingual Text Categorization

Multilingual text categorization algorithm performs the crucial task of Web portal development by organizing documents based on the contextual relevance. As document may belong to multiple themes with different membership degrees, multilingual text categorization is essentially a fuzzy classification process. Towards this end, a fuzzy concept-based multilingual text categorization algorithm is proposed. This algorithm regards the determination of a document's conceptual relevance to a particular theme as a process of fuzzy instance-based classification using the fuzzy nearest-prototype classification algorithm [14]. Multilingual Web document passed on from the Web crawler will then categorized to the corresponding themes where its membership degrees exceeds a certain threshold.

Based on the fuzzy nearest prototype classification algorithm, the fuzzy concept-based multilingual text classifier considered multilingual text categorisation task as a task of determining for every multilingual document a membership value in the range of [0,1] to each of the decision matrix with reference to a set of themes as illustrated in Figure 1.

	d_1	…	…	d_j	…	…	d_q
h_1	$\mu_1(d_1)$	…	…	$\mu_1(d_j)$	…	…	$\mu_1(d_q)$
…	…	…	…	…	…	…	…
h_i	$\mu_i(d_1)$	…	…	$\mu_i(d_j)$	…	…	$\mu_i(d_q)$
…	…	…	…	…	…	…	…
h_p	$\mu_p(d_1)$	…	…	$\mu_p(d_j)$	…	…	$\mu_p(d_q)$

Fig. 1. Decision matrix of fuzzy concept-based text categorization

where $H = \{h_1,\ldots, h_p\}$ is a set of themes, $D = \{d_1,\ldots,d_q\}$ is a set of multilingual documents to be classified and $\mu_i(d_j) \in [0,1]$ is the degree of membership of document d_j in theme h_i. The fuzzy concept-based multilingual text categorization algorithm is as follows:

Let $V = \{v_1, v_2, ..., v_p\}$ be the set of i theme prototypes representing i thematic categories. Such theme prototypes are obtained from the concept-based multilingual document clustering result in Section 3.3 by finding the centroid document in each document cluster. The concept-based document vector of this centroid document thus forms the theme prototype. The membership value $\mu_i(d_x)$ of an unclassified text, d_x, is determined by

$$\mu_i(d_x) = \frac{\left(\frac{1}{\|d_x - v_i\|^{2/(m-1)}}\right)}{\sum_{j=1}^{p}\left(\frac{1}{\|d_x - v_j\|^{2/(m-1)}}\right)} \tag{7}$$

where $m = 2$ is chosen.

As a result of this process, every multilingual document is now assigned a set of theme membership values. Concept-based multilingual text categorisation is then achieved by associating each theme with a set of multilingual documents ranked in decreasing order of relevance as indicated by their theme membership values.

6 Conclusion

This paper has presented novel research towards automatic development of multilingual Web portal. The salient feature of this methodology is the concept-based approach that overcomes the feature incompatibility problem by unifying multilingual terms in a common semantic space. As such, Web documents relevant to the same domain, regardless of language, can then be gathered, filtered and organised based on their comparable semantic content rather than the incomparable syntactic terms they

contain. Based on the concept-based approach, a concept-based theme taxonomy representing the hierarchical thematic backbone of a multilingual Web portal is first generated using a training parallel corpus relevant to the domain of a portal to be developed. To collect multilingual documents relevant to the portal's knowledge domain, a concept-based multilingual Web crawler is developed. Multilingual Web documents fetched by the Web crawler, in turn, are organized onto the concept-based theme taxonomy with the fuzzy concept-based multilingual text categorization algorithm. A multilingual Web portal thus developed is particularly significant to multilingual knowledge management where global knowledge relevant to a certain domain of interest need to be scanned, filtered, retrieved and organized to reveal actionable insights. To further evaluate the effectiveness of the proposed approach, a prototype multilingual Web portal is currently being developed, which uses a company's multilingual product catalog as training corpus.

References

1. Chakrabarti, S. (2000) Data mining for hypertext: a tutorial survey. *ACM SIGKDD Exploration*, 1(2), pp. 1 –11.
2. Berry, M. (2003) *Survey of Text Mining: Clustering, Classification, and Retrieval.* Springer-Verlag.
3. Chang, C., Healey, M. J., McHugh, J. A. M. and Wang, J. T. L. (2001) *Mining the World Wide Web: an information search approach.* Kluwer Academic Publishers.
4. Kosala, R. and Blockeel, H. (2000) Web mining research: a survey. *ACM SIGKDD Exploration*, 2(1), pp. 1 –15.
5. Chakrabarti, S. (2002) Mining the Web: Discovering Knowledge from Hypertext Data.Morgan Kaufmann.
6. Tan, A-H. (1999) Text Mining: The state of the art and the challenges. In *Proceedings of PAKDD'99 workshop on Knowledge Disocovery from Advanced Databases*, Beijing, pp. 65-70.
7. Carbonell, J. G., Yang, Y., Frederking, R. E., Brown, R. D., Geng, Y. and Lee, D (1997) Translingual information retrieval: a comparative evaluation. In Pollack, M. E. (ed.) *IJCAI-97 Proceedings of the 15th International Joint Conference on Artificial Intelligence,* pp. 708-714.
8. Davis, M., (1996) New experiments in cross-language text retrieval at nmsu's computing research lab. In Proceedings of the Fifth Retrieval Conference (TREC-5) Gaithersburg, MD: National Institute of Standards and Technology.
9. Landauer, T. K. and Littman, M. L. (1990) Fully automatic cross-language document retrieval. In *Proceedings of the Sixth Conference on Electronic Text Research,* pp. 31-38.
10. Kohonen, T. (1995) *Self-Organizing Maps.* Springer-Verlag, Berlin.
11. Salton, G. (1989) Automatic Text Processing: The Transformation, analysis, and Retrieva of Information by Computer. Addison-Wesley, Reading. MA.
12. Anderberg, M. R. (1973) *Cluster analysis for applications.* Academic Press, Inc., New York.
13. Kumar, R., Raghavan, P., Sridhar Rajagopalan, S., Sivakumar, D., Tompkins, A., and Upfal, E. (2000) The Web as a graph. In Proceedings of the nineteenth ACM SIGMOD SIGACT SIGART symposium on Principles of database systems, pp. 1—10.
14. Keller, J. M., Gray, M R. and Givens, J. A. (1985) A fuzzy *k*-nearest neighbor algorithm. IEEE Transactions of Systems, Man and Cybernetics. Vol. SMC-15, no.4, pp.580-585.