Available online at www.sciencedirect.com
ScienceDirect
Decision Support Systems 45 (2008) 596–605
Decision Support Systems
www.elsevier.com/locate/dss

Cross-lingual thesaurus for multilingual knowledge management

Christopher C. Yang [a,*], Chih-Ping Wei [b], K.W. Li [c]

[a] Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, China
[b] Institute of Technology Management, National Tsing Hua University, Taiwan, ROC
[c] Department of Information Systems, City University of Hong Kong, China

Available online 27 July 2007

Abstract

The Web is a universal repository of human knowledge and culture which has allowed unprecedented sharing of ideas and information in a scale never seen before. It can also be considered as a universal digital library interconnecting digital libraries in multiple domains and languages. Beside the advance of information technology, the global economy has also accelerated the development of inter-organizational information systems. Managing knowledge obtained in multilingual information systems from multiple geographical regions is an essential component in the contemporary inter-organization information systems. An organization cannot claim itself to be a global organization unless it is capable to overcome the cultural and language barriers in their knowledge management. Cross-lingual semantic interoperability is a challenge in multilingual knowledge management systems. Dictionary is a tool that is widely utilized in commercial systems to cross the language barrier. However, terms available in dictionary are always limited. As language is evolving, there are new words being created from time to time. For examples, there are new technical terms and name entities such as RFID and Baidu. To solve the problem of cross-lingual semantic interoperability, an associative constraint network approach is investigated to construct an automatic cross-lingual thesaurus. In this work, we have investigated the backmarking algorithm and the forward evaluation algorithm to resolve the constraint satisfaction problem represented by the associative constraint network. Experiments have been conducted and show that the forward evaluation algorithm outperforms the backmarking one in terms of precision and recall but the backmarking algorithm is more efficient than the forward evaluation algorithm. We have also benchmarked with our earlier technique, Hopfield network, and showed that the associate constraint network (either backmarking or forward evaluation) outperforms in precision, recall, and efficiency.

Keywords: Cross-lingual concept space; Cross-lingual thesaurus; Associate constraint network

1. Introduction

With the advance of information technology and the global economy, organizations are now managing information in multiple languages that may be generated or gathered from multiple sources at different geographical regions. Inevitably, users have to search across the languages to extract information to fulfill their information needs. Managing knowledge in monolingual documents is not a trivial task. Multilingual knowledge management further extends the problem with language barrier among queries and documents.

The followings are some scenarios of the needs of multilingual knowledge management:

1. The enterprise regional repository has a collection of local monolingual documents but it supports employees from other branches of the enterprise at different locations around the world. These employees may

* Corresponding author.
E-mail address: yang@se.cuhk.edu.hk (C.C. Yang).

0167-9236/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2007.07.005

speak different languages. In this case, users may submit a query in the language that they are familiar with but not in the language of documents in the regional repository. The queries must be translated into the document language before retrieval.

2. The enterprise centralized repository has a collection of multilingual documents and the documents are organized in a hierarchical category tree. A regional branch of the enterprise has a collection of local documents recently aggregated from the local sources. The chief knowledge officer wants to make this collection of regional monolingual documents available to all employees in the enterprise through the centralized repository. In this case, we need to map the regional documents of the local language to the hierarchical category tree of the centralized repository in which the organization of the documents are conducted in different language.

In these scenarios, it is apparent that cross-lingual information retrieval and cross-lingual text categorization are important tasks in multilingual knowledge management.

Cross-lingual information retrieval (CLIR) processes a user query in one language and retrieves relevant documents in other languages [11], [16]. Three major approaches commonly employed for cross-lingual information retrieval include controlled vocabulary, knowledge-based, and corpus-based [11], [17]. The controlled vocabulary approach adopts a predetermined set of vocabularies for user queries and document indexing. However, the retrieval effectiveness highly depends on the selection of vocabularies. The knowledge-based approach adopts an ontology and dictionary to translate queries from one language to another language. However, many technical terms, abbreviation, names of persons, organizations, and events may not be included in the dictionary. Translation can also be ambiguous in some cases. The corpus-based approach makes use of the statistical information of term usage in a parallel or comparable corpus to automatically construct a statistically based cross-lingual thesaurus to overcome the limitations of knowledge-based approach.

Cross-lingual text categorization (CLTC) learns from a set of training documents in one language and classifies new (i.e., unclassified) documents in other languages [1]. As with CLIR, the CLTC technique proposed by Bel et al. [1] uses a cross-lingual thesaurus for translations. On the basis of a particular term selection metric, their technique first selects the best k terms per category from a set of preclassified documents in language L_1. Each set of terms then serves as the class profile for the respective category. Afterward, with the use of the cross-lingual thesaurus, a translation table from L_2 to L_1 is constructed, comprising only those terms in L_1 that can be translated into the k terms identified previously. Accordingly, the translation table translates each unclassified document written in language L_2 into L_1. Finally, on the basis of the similarity between the class profiles and the translated feature vector of each unclassified document, the document is assigned to the closest category.

Evidently, the heart of multilingual knowledge management, including CLIR and CLTC, is the cross-lingual semantic interoperability issue. In our previous work, the cross-lingual semantic interoperability problem has been modeled as a neural network (specifically, the Hopfield network) [8], [9] to generate the cross-lingual thesaurus [6], [17]. A cross-lingual concept space is a semantic associate network consisting of concepts and related concepts in multiple languages, and is computed based on co-occurrence relationships from a parallel corpus [5], [17]. As shown in Fig. 1, the documents in the parallel corpus are first aligned to form document pairs [7], [13]. Terms in L_1 and L_2 are then extracted [14], [17]. Each document pair is considered as a unit in the co-occurrence analysis. During the co-occurrence analysis, we compute the relevance weights between all pairs of terms (term i and term j in either L_1 and L_2) denoted as W_{ij}. A meaningful and understandable concept space (a network of terms and weighted associations) can represent the concepts (terms) and their associations for the underlying information space (i.e., documents in the corpus) [4], [17]. Such cross-lingual concept space overcomes the limitations of the traditional dictionary-based approach of cross-lingual information retrieval and text categorization.

The Hopfield network is a promising technique in generating cross-lingual thesaurus as shown in our previous work [6], [17]. The Hopfield network has two phases, the storage phase and the retrieval phase. The synaptic weights between neurons are generated in the storage phase and the activation of neurons is applied in the retrieval phase [2,3,17]. However, the Hopfield network is rather weak in efficiency and consistency. The convergence process in the Hopfield network is time consuming. In some cases, it may not converge especially when the parameters are not tuned appropriately. The random processes in the Hopfield network also cause the inconsistency. Results generated in different convergence processes may be different. In this work, we investigate the associate constraint network approach to generate the cross-lingual thesaurus based on the result of the co-occurrence analysis. Two

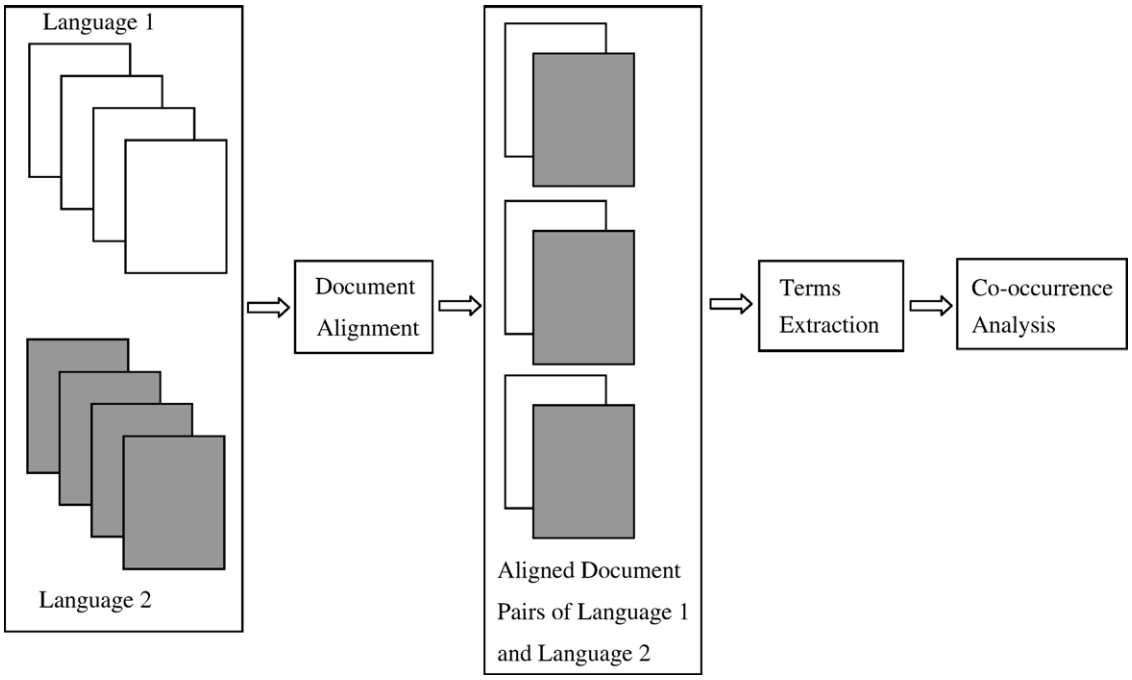

Fig. 1. Construction of a network of cross-lingual concepts from parallel corpus.

constraint propagation techniques, backmarking and forward evaluation, are compared in our experiments.

2. Cross-lingual thesaurus

In this research work, an associate constraint network approach is utilized to simulate the associate memory in a human brain in order to construct a cross-lingual concept space. The associate constraint network is the associate network of the extracted terms from a parallel corpus [13] with constraints imposed on the nodes of the associate network. Searching techniques are developed to search for a feasible solution that satisfies the constraints in order to generate a cross-lingual thesaurus for an input term in any languages.

A constraint satisfaction problem (CSP) is a problem composed of a finite set of variables, each of which is associated with a finite domain, and a set of constraints that restricts the values that the variables can simultaneously take [12]. The task is to assign a value to each variable satisfying all the constraints.

The construction of the cross-lingual thesaurus is modeled as a constraint satisfaction problem [10,12,15] and the constraints are depicted by the associate constraint network. The nodes of an associate constraint network (x_1, x_2,..., x_n) represent the extracted terms of the parallel corpus, where x_i can be a term in L_1 and L_2. The values of the nodes are binary, $x_j = \{0, 1\}$:

- $x_j = 1$ if x_j is a term in the cross-lingual concept space and
- $x_j = 0$ if x_j is not a term in the cross-lingual concept space.

The arcs of the associate network represent the association between the extracted terms.

The constraint c_j is applied on x_j. A term, x_j, is considered to be relevant to the other terms in the thesaurus if the sum of the associate weights between the other terms and itself is sufficiently large; otherwise, it should not be included in the thesaurus. c_j is given as below.

$$x_j = \begin{cases} 1 & \sum_{i=1, i\neq j}^{n} W_{ij} x_i \geq \text{threshold} \\ 0 & \sum_{i=1, i\neq j}^{n} W_{ij} x_i < \text{threshold} \end{cases}$$

where W_{ij} denotes the relevance weight from node i to node j, Node i and node j can be terms in L_1 and L_2, and n is the number of nodes in the network.

The threshold is determined empirically based on the distribution of relevance weights W_{ij}. Moreover, W_{ij} is derived as follows:

$$W_{ij} = \frac{\sum_{k=1}^{N} d_{kij}}{\sum_{k=1}^{N} d_{ki}}$$

where N is the number of document pairs in the parallel corpus.

$$d_{ki} = tf_{ki} \times \log(N/df_i \times l_i) \quad \text{where } l_i \text{ is the length of term } i$$

$$d_{kij} = tf_{kij} \times \log N/df_{ij}$$

where tf_{kij} is the minimum of term frequency of term i and term frequency of term j in document pair k and df_{ij} is the document frequency of term i and term j.

The constraint satisfaction problem (CSP) for constructing a cross-lingual thesaurus is then defined in terms of the node consistency and the satisfaction of the associate constraint network as follows:

Definition 1. Node consistency

x_j is consistent if and only if c_j is satisfied in the associate constraint network.

Definition 2. Associate constraint network satisfaction

The associate constraint network is satisfied if and only if all nodes in the associate constraint network are consistent and

$$\sum\nolimits_j x_j < C$$

where C is a threshold and determined statistically based on the distribution of an input concept and their associated concepts.

A solution to a CSP is an assignment of a value from its domain to every variable, in such a way that every constraint is satisfied. A solution tuple of a CSP is a compound label for all those variables which satisfy all the constraints [12]. Fig. 2 illustrates an associate constraint network. In this network, there are four nodes representing four terms in either L_1 and L_2. The directed arcs correspond to the relevance weights from one term to another. The relevance weights for the associate constraint network in Fig. 2 are given in Table 1.

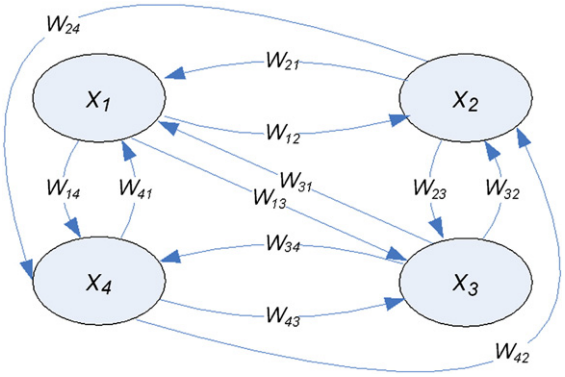

Fig. 2. An illustration of associate constraint network with four nodes, x_1 to x_4, representing terms in L_1 and L_2.

One simple approach to find a solution tuple is use of the exhaustive search. Table 2 illustrates 4 possible tuples for the associate constraint network illustrated in Fig. 2, $\{x_1=1, x_2=0, x_3=0, x_4=0\}$, $\{x_1=1, x_2=1, x_3=0, x_4=0\}$, $\{x_1=1, x_2=0, x_3=1, x_4=0\}$, $\{x_1=1, x_2=1, x_3=1, x_4=0\}$. Assuming threshold=0.7 and $C=3$, only $\{x_1=1, x_2=1, x_3=0, x_4=0\}$ is a solution tuple in which x_1, x_2, x_3 and x_4 are all consistent and $x_1+x_2+x_3+x_4<3$. For the other three tuples, either one or more nodes are inconsistent and/or $x_1+x_2+x_3+x_4\geq 3$. When the size of the associate constraint network is small, the exhaustive search is computationally feasible because the number of possible solution tuples is manageable. However, in a practical case, the number of nodes extracted from a parallel corpus is typically several thousands. As the associate constraint network expands, the computational cost of the exhaustive search increases exponentially. As a result, the exhaustive search in an associate constraint network becomes impractical to find a solution tuple when generating a cross-lingual thesaurus.

Solutions to CSPs can be found efficiently and effectively by searching systematically through possible assignments of values to variables. The searching algorithms can be categorized into two major schemes: backtracking scheme and look-ahead scheme. The backmarking algorithm, a typical backtracking strategy, continues to search for a solution until an infeasible solution is determined. When an infeasible solution is found, it backtracks to the previous assignment that has caused the infeasibility. Such assignment will be marked and then it continues to search for an alternative solution. The searching continues until a feasible solution is

Table 1
Relevance weights, W_{ij}, for the association constraint network in Fig. 2

W_{12}	0.8
W_{13}	0.1
W_{14}	0.2
W_{21}	0.7
W_{23}	0.4
W_{24}	0.1
W_{31}	0.3
W_{32}	0.4
W_{34}	0.1
W_{41}	0.1
W_{42}	0.2
W_{43}	0.1

determined. A backmarking algorithm is investigated in this work.

An alternative approach is the look-ahead scheme. Look-ahead scheme detects the inconsistency earlier than backtracking scheme and thus it allows branches of the search tree that will lead to failure to be pruned earlier than with look-back schemes. It reduces the search tree. It is reasonable to assume that the forward evaluation is more efficient than the backmarking because the search tree is pruned and therefore the search space is smaller. However, the efficiency also depends on the computational cost of the evaluation process and the amount of the search tree is pruned. A forward evaluation algorithm is also investigated in this work. We are interested in identifying if the forward evaluation is more efficient than the backmarking. In addition, we are also interested to determine whether the forward evaluation produces better cross-lingual thesaurus because less relevant terms are pruned without any consideration as part of the cross-lingual thesaurus.

3. Search strategies in associate constraint network

Given an associate constraint network and an input term x_i, the searching algorithm exploits the associate constraint network and determines the cross-lingual thesaurus of x_i (specifically, a set of terms in another language associated to x_i).

Backtracking and lookahead are two major search strategies in solving constraint network problem. The backtracking strategy (e.g., the backmarking propagation algorithm) checks the compatibility of the new assignment. If the new assignment is incompatible, it is rejected. Such assignment will be marked and an alternative assignment will be searched. On the other hand, the lookahead strategy (e.g., the forward evaluation algorithm) checks forward to see if the current assignment may cause future incompatibility instead of continuing the iterations until an incompatible assignment is found.

Table 2
Illustrations of solution tuples for the associate constraint network in Fig. 2

$x_1=1, x_2=0, x_3=0, x_4=0$		$x_1=1, x_2=1, x_3=0, x_4=0$	
$\sum_{i=1,i\neq1}^{4} W_{i1}x_i = 0$	x_1 is inconsistent	$\sum_{i=1,i\neq1}^{4} W_{i1}x_i = 0.7$	
$\sum_{i=1,i\neq2}^{4} W_{i2}x_i = 0.8$	x_2 is inconsistent	$\sum_{i=1,i\neq2}^{4} W_{i2}x_i = 0.8$	
$\sum_{i=1,i\neq3}^{4} W_{i3}x_i = 0.1$		$\sum_{i=1,i\neq3}^{4} W_{i3}x_i = 0.5$	
$\sum_{i=1,i\neq4}^{4} W_{i4}x_i = 0.2$		$\sum_{i=1,i\neq4}^{4} W_{i4}x_i = 0.3$	
$x_1=1, x_2=0, x_3=1, x_4=0$		$x_1=1, x_2=1, x_3=1, x_4=0$	
$\sum_{i=1,i\neq1}^{4} W_{i1}x_i = 0.3$	x_1 is inconsistent	$\sum_{i=1,i\neq1}^{4} W_{i1}x_i = 1.0$	
$\sum_{i=1,i\neq2}^{4} W_{i2}x_i = 1.2$	x_2 is inconsistent	$\sum_{i=1,i\neq2}^{4} W_{i2}x_i = 1.2$	
$\sum_{i=1,i\neq3}^{4} W_{i3}x_i = 0.1$	x_3 is inconsistent	$\sum_{i=1,i\neq3}^{4} W_{i3}x_i = 0.5$	x_3 is inconsistent
$\sum_{i=1,i\neq4}^{4} W_{i4}x_i = 0.3$		$\sum_{i=1,i\neq4}^{4} W_{i4}x_i = 0.4$	

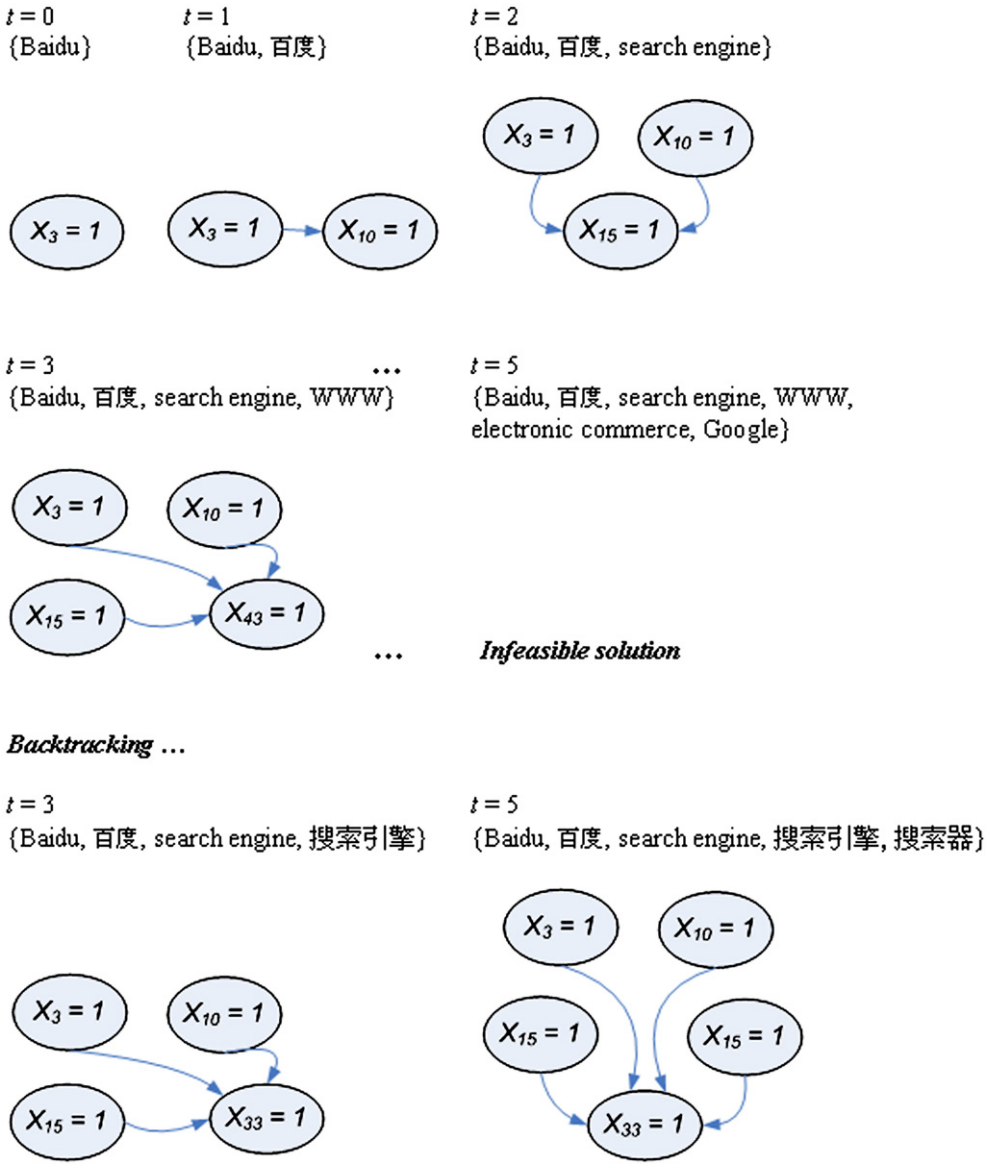

Fig. 3. Illustration of backmarking in associate constraint network.

In this section, we first introduce the backmarking algorithm and provide an illustration.

Backmarking

1. Initialization

$x_i = 1$
$x_j = 0 \ \forall j, j \neq i$
$y_j = 0 \ \forall j$
$t = 0$

2. Generate a potential solution

$t = t + 1$

$$v_j = \sum_{i=0, i \neq j}^{n} W_{ij} x_i$$

$x_k = 1$ where $k = \underset{j, x_k = 0}{\text{Arg Max}}\, v_j$ and x_k is not marked in the memory

$y_t = k$

3. Determine if a solution is found
IF the associate constraint network is satisfied
The solution is equal to the current tuple
ELSE IF an infeasible solution is found
$x_k=0$
$y_t=0$
$t=t-1$
Mark x_k in the memory
IF $\forall x_k$ are marked
Unmarked $\forall x_k$ in the memory
$x_{y_{t-1}}=0$
$y_{t-1}=0$
$t=t-1$
Mark $x_{y_{t-1}}$ in the memory
ELSE
Go back to Step 2

For example, as illustrated in Fig. 3, partial associate constraint networks in a few iterations are presented (due the limitation of space). Given x_3=Baidu, x_{10}=百度 (i.e., Baidu in Hong Kong translation), x_{15}=search engine, x_{17}=Google, x_{33}=搜索器 (i.e., search engine in Hong Kong translation), x_{35}=搜索引擎 (i.e., search engine in Mainland China translation), x_{43}=WWW, and x_{49}=electronic commerce. Let x_3 be the input term and is initialized to 1 while all others are initialized to 0. In Step 2, x_{10} is set to 1 and y_1 is set to 10 because $v_{10}=W_{3,10}$ is the largest. {Baidu, 百度} is a partial solution. The associate constraint network is not satisfied yet but the current solution is a feasible partial solution. The algorithm goes back to Step 2. At this iteration, x_{15} is set to 1 and y_2 is set to 15 because $v_{15}=W_{3,15}+W_{10,15}$ is the largest. The partial solution now becomes {Baidu, 百度, search engine} which is feasible but does not satisfy the associate constraint network. The partial solutions in the next few iterations are {Baidu, 百度, search engine, WWW}, {Baidu, 百度, search engine, WWW, electronic commerce} and finally an infeasible solution is found. Based on the marking, we backtrack to the partial solution {Baidu, 百度, search engine, WWW}. Since all the other possible partial solutions generated from {Baidu, 百度, search engine, WWW} are infeasible, we backtrack to {Baidu, 百度, search engine} and identify {Baidu, 百度, search engine, 搜索引擎}. In one further iteration, the solution is found as {Baidu, 百度, search engine, 搜索引擎, 搜索器}. As illustrated, the backmarking algorithm continues to search for a solution until an infeasible solution is found. It marks the node that causes the infeasibility and backtracks to the previous partial solution. A partial solution is a tuple that does not satisfy the associate constraint network but does not cause any infeasibility.

The forward evaluation algorithm is proposed to detect the possible infeasibility in advance so that inappropriate assignment of irrelevant terms to the cross-lingual concept space can be avoided before efforts have been wasted on searching. The forward evaluation algorithm is presented as follow:

Forward evaluation

1. Initialization
$x_i=1$
$x_j=0\ \forall j, j\neq i$
$y_j=0\ \forall j$
$t=0$

2. Generate a potential solution
$t=t+1$

$$v_j=\sum_{i=0,i\neq j}^{n} W_{ij}x_i$$

$x_k=1$ where $k=\underset{j,x_k=0}{\text{Arg Max}}\, v_j$ and x_k is not marked in the memory

$y_t=k$

3. Evaluating forward
Z=the number of node j with relevance weights W_{kj} that is higher than $\text{threshold}_{\text{high}}$
Z'=the number of node j with relevance weights Wji that is lower than $\text{threshold}_{\text{low}}$
IF $Z-Z'$ is small and Z is large
$x_k=0$
$y_t=0$
$t=t-1$
Mark x_k, in the memory
Go back to Step 2

4. Determine if a solution is found
IF the associate constraint network is satisfied
The solution is equal to the current tuple
ELSE IF an infeasible solution is found
$x_k=0$
$y_t=0$
$t=t-1$
Mark x_k in the memory
IF $\forall x_k$ are marked
Unmarked $\forall x_k$ in the memory
$x_{y_t-1}=0$
$y_{t-1}=0$
$t=t-1$
Mark $x_{y_{t-1}}$ in the memory
ELSE
Go back to Step 2

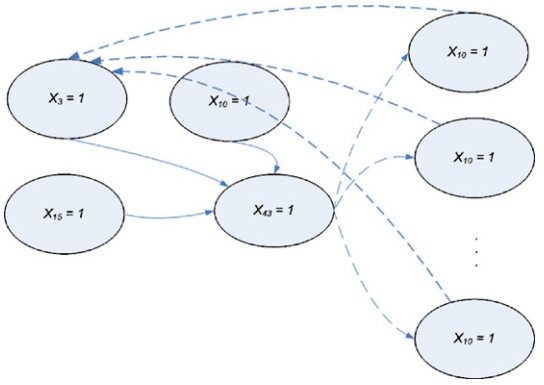

Fig. 4. Illustration of forward evaluation.

Using the earlier example, when the partial solution is {Baidu, 百度, search engine, WWW}, the forward evaluation identifies that x_{43} (WWW) may cause infeasibility and immediately back track to the partial solution {Baidu, 百度, search engine} as shown in Fig. 4. It is because there are large number of nodes with $W_{43,j}$ higher than $\text{theshold}_{\text{high}}$ (as illustrated by the arcs pointing from x_{43} to x_{49}, x_{31}, and x_7) and there are large number of nodes with $W_{j,3}$ lower than $\text{threshold}_{\text{low}}$ (as illustrated by the arcs pointing from x_{49}, x_{31}, and x_7 to x_3). "WWW" may make many less relevant terms as part of the partial solution in the next few iterations. The consistencies of these irrelevant terms cannot be maintained; however, the summation of x_i will eventually greater than C.

4. Experiment

We have conducted an experiment to measure the performance of the associate constraint network in generating an automatic cross-lingual thesaurus. Its performance is also compared with the previous technique, Hopfield network.

The data used in the experiment is collected from the Hong Kong government press releases available in the government Web site. Statistics of the corpus is presented in Table 3.

10 subjects were invited from an engineering department to participate in the experiment. All subjects are fluent in both English and Chinese languages. Two evaluation sessions are assigned to each subject. The objective of the first session is to measure the precision of the cross-lingual thesaurus while the objective of the second session is for measuring the recall of the cross-lingual thesaurus.

In the first session, fifty terms (again, twenty-five terms are English and twenty-five terms are Chinese) were randomly selected as input to the Hopfield network and associate constraint network using backmarking and forward evaluation. The input terms and the results generated by the three methods are presented to the subjects. Ten percent of noise terms were added to reduce the bias of the automatically generated result. The order of the terms extracted by the three methods was randomized so that the subjects were not able to identify the source of any terms. The subjects were asked whether the terms were relevant to the input term and marked the translated term if it could be identified. The precision rate was then computed as the number of retrieved relevant terms divided by the number of retrieved term.

In the second session, fifty terms (twenty-five terms are English and twenty-five terms are Chinese) were randomly selected but not necessary the same as those used in the first session. The subjects were asked to suggest relevant terms according to their experience and knowledge. The recall rate was computed as the number of retrieved relevant terms divided by the number of relevant terms.

The result is presented in Tables 4 and 5. It is found that the forward evaluation outperforms the backmarking in both precision and recall and the backmarking outperforms the Hopfield network in both precision and recall. In terms of precision and recall, forward evaluation of associate constraint network is the best. It shows that the forward evaluation is able to prune the less relevant terms in the process of generating the solution and therefore the final precision and recall of the solution is the best among all methods. The backmarking only backtracks to the previous partial solution whenever it finds an infeasible solution; however, the previous partial solution may still

Table 3
Statistics of the corpus

Number of English and Chinese pairs of document	2548
Number of terms extracted	9222
Number of English terms	3635
Number of Chinese terms	5597

Table 4
Precision and recall of Hopfield network, backmarking and forward evaluation

	Hopfield Network	Associate constraint network	
		Backmarking	Forward evaluation
Precision	0.84	0.89	0.96
Recall	0.80	0.82	0.89

Table 5
Efficiency of Hopfield network, backmarking and forward evaluation

	Hopfield network	Associate constraint network	
		Backmarking	Forward evaluation
Efficiency	49 s	20 s	36 s

include terms that are not so relevant. In terms of efficiency, the backmarking of associate constraint network is the best while the Hopfield network is the worst. The forward evaluation is not as efficient as the backmarking because the forward evaluation conducts the evaluation process in every iteration and such evaluation process is rather time consuming. As the number of iterations increase, the amount of time to search the solution tuple increases.

5. Conclusion

Multilingual knowledge management is an important issue as a result of the globalization of economy and the advance of the Internet. In the global enterprise, information is collected or generated from different sources with multiple languages. The knowledge management systems are supporting users who are only familiar with one or a few languages but the knowledge or information that support their information needs are available in many other languages that the users are not familiar with. To manage the knowledge that is obtained from documents in different languages is not an easy task. The major challenge in multilingual knowledge management is the cross-lingual semantic interoperability problem. In this work, we have investigated the associate constraint network approach to construct the cross-lingual thesaurus. Two searching techniques, backmarking and forward evaluation, are considered. The experimental result shows that the associate constraint network approach outperforms the Hopfield network, a technique we have previously investigated, in terms of precision, recall and efficiency. Comparing between backmarking and forward evaluation of the associate constraint network approach, forward evaluation achieves higher precision and recall but it is less efficient. Given the automatic technique to construct cross-lingual thesaurus, it is promising to resolve the cross-lingual interoperability problem in multilingual knowledge management. The cross-lingual thesaurus can be used to expand queries from one language to another language so that users may be able to search across the language boundary. Besides, the cross-lingual thesaurus can be used to relate the terms in text categories of different languages to support the cross-lingual text categorization.

Acknowledgments

This project was supported by the Direct Research Grant of the Chinese University of Hong Kong, 2050268, and the Earmarked Grant for Research from the Hong Kong Research Grant Council, 4335/02E.

References

[1] N. Bel, C.H.A. Koster, M. Villegas, Cross-lingual text categorization, Proceedings of 7th European Conference on Research and Advanced Technology for Digital Libraries (ECDL '03), August 2003, pp. 126–139, Trondheim, Norway.
[2] H. Chen, K.J. Lynch, Automatic construction of networks of concepts characterizing document database, IEEE Transactions on Systems, Man and Cybernetics 22 (5) (1992) 885–902.
[3] H. Chen, B. Schatz, T. Ng, J. Martinez, A. Kirchhoff, C. Lin, A parallel computing approach to creating engineering concept spaces for semantic retrieval: The Illinois Digital Library Initiative Project, IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (8) (August, 1996) 771–782.
[4] H. Chen, T. Ng, J. Martinez, B. Schatz, A concept space approach to addressing the vocabulary problem in scientific information retrieval: an experiment on the Worm Community System, Journal of the American Society for Information Science 48 (1) (1997) 17–31.
[5] K.W. Li, C.C. Yang, Automatic construction of cross-lingual networks of concepts from the Hong Kong SAR Police Department, Proceedings of the First NSF/NIJ Symposium on Intelligence and Security Informatics (ISI 2003), June 2–3 2003, Tucson, Arizona.
[6] K.W. Li, C.C. Yang, Automatic cross-lingual thesaurus generated from the Hong Kong SAR Police Department Web corpus for crime analysis, Journal of the American Society for Information Science and Technology 56 (3) (2005) 272–282.
[7] K.W. Li, C.C. Yang, Conceptual analysis of parallel corpus collected from the Web, Journal of the American Society for Information Science and Technology 57 (5) (2006) 684–696.
[8] J.J. Hopfield, Neurons with graded response have collective computational properties like those of two-state neurons, Proceedings of the National Academy of Sciences of the United States of America 81 (30) (1984) 88–92.
[9] J.J. Hopfield, Neural network and physical systems with collective computational abilities, Proceedings of the National Academy of Sciences of the United States of America 79 (4) (1982) 2554–2558.
[10] V. Kumar, Algorithms for constraint satisfaction problems: a survey, AI Magazine 13 (1) (1992) 32–44.
[11] D.W. Oard, Alternative approaches for cross-language text retrieval, Proceedings of 1997 AAAI Symposium in Cross-Language Text and Speech Retrieval, AAAI, 1997.
[12] E. Tsang, Foundations of Constraint Satisfaction, Academic Press, London, 1995.
[13] C.C. Yang, K.W. Li, Automatic construction of English/Chinese parallel corpora, Journal of the American Society for Information Science and Technology 54 (8) (June, 2003) 730–742.
[14] C.C. Yang, K.W. Li, A heuristic method based on statistical approach for Chinese text segmentation, Journal of the American Society for Information Science and Technology 56 (13) (2005) 1428–1447.
[15] C.C. Yang, K.W. Li, Cross-lingual semantics for crime analysis using associate constraint network, Proceedings for the Second NSF/NIJ Symposium on Intelligence and Security Informatics (ISI2004), June 10–11 2004, Tucson, Arizona.
[16] C.C. Yang, K.W. Li, Cross-lingual information retrieval: The challenge in multilingual digital libraries, in: Y.L. Theng, S. Foo (Eds.), Design and Usability of Digital Libraries: Case Studies in the Asia Pacific, Idea Group, Inc., 2004.
[17] C.C. Yang, J. Luk, Automatic generation of English/Chinese thesaurus based on a parallel corpus in law, Journal of the American Society for Information Science and Technology, Special Topic Issue on Web Retrieval and Mining: A Machine Learning Perspective 54 (7) (May 2003) 671–682.

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 and Technology at Drexel University.

Chih-Ping Wei received a BS in Management Science from the National Chiao-Tung University in Taiwan, R.O.C. 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, R.O.C. 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, R.O.C; cpwei@mx.nthu.edu.tw.

Kar-Wing Li received his M.Phil. and Ph.D. in the Department of Systems Engineering and Engineering Management from the Chinese University of Hong Kong.