Research Journal of Applied Sciences, Engineering and Technology 4(12): 1714-1720, 2012
ISSN: 2040-7467

Submitted: December 28, 2011 Accepted: January 25, 2012 Published: June 15, 2012

# Conceptual Information Retrieval in Cross-Language Searches

[1]Morteza Poyan Rad, [2]Reza Pourshaikh and [3]Hamid Alinejad-Rokny
[1]Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran
[2]Qazvin Islamic Azad University, Electrical and Computer Engineering Faculty, Qazvin, Iran
[3]Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran

**Abstract:** Extending the internet, need of information management and conceptual information system has grown. On the other hand high amount of information are presented in various languages and learning all the languages is not possible for users. Therefore, multilingual information retrieval systems are the great necessity. The purpose of this paper is conceptual implementation of cross-language systems. Using ontology in concept extraction of user query, phrase translation instead of word by word translation and conceptual disambiguation by graph of concepts are the proposed methods here.

**Key words:** Conceptual search, cross-language information retrieval, graph of concepts, ontology

## INTRODUCTION

Cross-Language Information Retrieval (CLIR) is a kind of information retrieval in which query language is different from user's required documents. Statistics show that most of internet users are speaking Non-English languages. This issue leads to explosive increase of existing information of the web in different languages and caused more need for cross-language information retrieval; so we can say: Cross-language information retrieval is a retrieval process in which the user presents queries in one language to retrieve documents in another language. Due to increasing availability of electronic documents written in various languages from all over the world, Cross Language Information Retrieval has gained popularity among Information Retrieval (IR) researchers in recent years. Since the existing Web search engines only support the retrieval of documents written in the same language as the query, there is no efficient way for monolingual users to retrieve documents written in nonnative languages (Aleahmad *et al*., 2007; Teymoorian *et al*., 2009).

A lot of common information retrieval systems are as simple as a pattern matching system. On the other hand, by receiving the informational needs of user, without a correct realization of concepts in user query and existing documents, they just try to find patterns similar to that of user query among a lot of existing documents. Thus, common search engines are not able to realize the intended concepts of user and cannot realize the concepts of documents. These problems show that conceptual information retrieval is very important (Jadidinejad and Keyvanpour, 2008).

In cross-language information retrieval, documents and queries are presented in different languages, so in these kinds of systems, translation techniques should be combined with mono-lingual information retrieval systems. Due to this reason, researchers are using natural language processing techniques in cross-language information retrieval. For example dictionaries, corpus, encyclopedias and machine translation systems use these techniques a lot for translation of queries and documents. But these techniques are not efficient enough, therefore in practice, cross-language information retrieval has reached 50 to 70% of efficiency in monolingual retrieval systems (Ata *et al*., 1995; Jagarlamudi and Kumaran, 2007; Sadat *et al*., 2002).

This study proposes a method to retrieve information written in a language different from the language of user. This kind of retrieval is done by concept induction from the user query and it is equipped with ontology; it translates the phrases instead of the words and does semantic disambiguation by using graph of concepts, which has good effects on the results.

## ONTOLOGY

Ontology is an important and extended tool in knowledge presentation. It makes a logical relation among a lot of data. Using of domain knowledge and designing

**Corresponding Author:** Morteza Poyan Rad, Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran

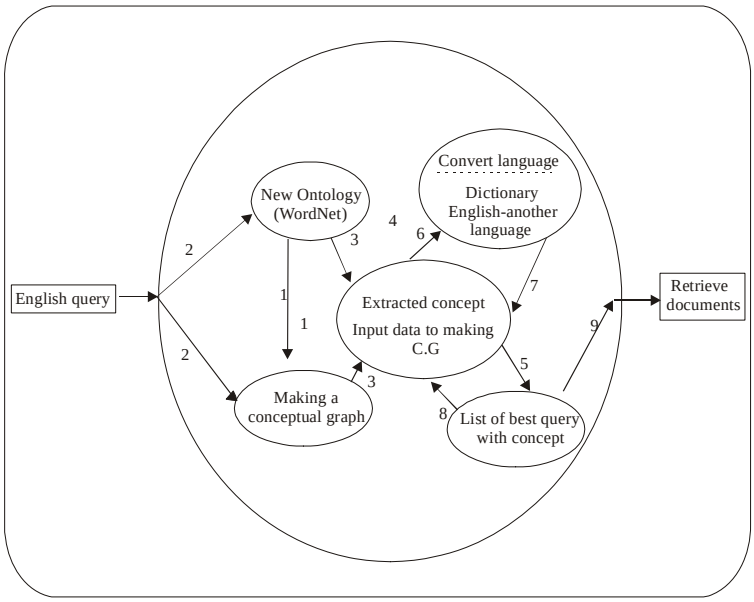


Fig. 1: Proposed approach architecture for improvement of English-Persian bilingual retrieval

the semantic models such as semantic network, is a sufficient method in which intelligent terminal of information retrieval can understand the content of query and documents and make a conceptual relation between query and documents (Alpcan *et al*., 2007; Studer, 2000). There are different relations in ontology; the relation of Synonym (means the same as), Hypernymy (is the general term), Hyponymy (is a kind of), Meronymy (is part substance/member of), Holonymy (has part), Antonym (is the contrary); through these relations different parts of knowledge can be extracted.

In this study WordNet which is a public ontology has been used. For more detail about WordNet ontology refer to (Fellbaum, 1998).

**Proposed approach for English-Persian cross-language information retrieval:** In the proposed method a combined way for concept retrieving of documents in both Persian and English, for queries which are in English, has been considered. In this combinational method, ontology has been used to extract the concepts out of documents and queries. Since in multi-lingual information retrieval systems query translation is a common issue for unification of source and destination languages, query translation by using a bilingual dictionary is used. But the problem is a word may have different meanings in a dictionary; on the other hand, word by word translation of query doesn't have a good precision and causes a lot of ambiguity in translation. In this proposed method, different ways like phrase recognition, phrase translation instead of word translation and extending the queries by using ontology and semantic disambiguation are proposed as solutions for this problem. Results show the increasing efficiency in the case of using this bilingual retrieving system. These methods will be explained more below. The architecture of the proposed method is shown in Fig. 1.

It is worth to mention that this approach is not specific to English-Persian bilingual information retrieval and with a little change it could be applied for other cross-language systems.

**Extension of English query phrases:** Operation of query extension could be done either before or after translation, or in both. Extension of query before translation leads to a good query and includes more phrases in query language. Extension of query after the translation with adding some more conceptual phrases would decrease the effect of unrelated query terms. However, in this proposed approach we do the query extension before the translation. For achieving this goal we use WordNet as a general ontology and at first, we extract all the synsets of any entry for that word. (Note: we don't consider just nouns). If, for any word more than one Synset exists, calculation of semantic similarity will be used among Synsets of this word and the words before and after that word. After the selection of the most related Synsets to this word, these stages for query extension are done:

**Stage 1:** All the synonyms are added.
**Stage 2:** All the hyponyms of that word are added. These hyponyms are children which share all the features of their parents and they increase precision rate.
**Stage 3:** If just one hypernym exists for that word, it is added, too.

Since existence of more than one hypernym may lead us to a broader domain of concepts and hence this increases the ambiguity; for this reason, only when a specific hypernym for that intended word exists, it will be added. This issue will increase the retrieval rate.

For computation of semantic similarity of two words WordNet has been used (Sebti and Barfroush, 2008) and for this work IC (Information Content) is defined as follow.

$$IC(Concept) = -\log(P(Concept)) \quad (1)$$

Benefits of IC are:

- It shows the degree of a concept specialty in the domain of its topic.
- A concept with high information content is considerably specific.
- Concepts with low information content have general meanings and they have a low degree of specialty.

In formula (1), p (concept) will be calculated as follow:

$$P(concept) = 1/\text{ number of hyponym for concept} \quad (2)$$

For calculating the standard of similarity, we apply a formula which Lin discussed in (Lin, 1998) with a little change:

$$related\ (c_1, c_2) = \frac{2 * IC(Lcs(c_1, c_2))}{IC(c_1) + IC(c_2)} \quad (3)$$

Lcs (C1, C2) is the common parent or the hypernym which both of the two words (C1 and C2) are common in and it is a word itself. In formula (3) amount of IC for any word like X with padding of MySql data base, WordNet will be calculated as follow:

$$IC(x) = IC(hypernym(x))* IC(hypernym(hypernym(x)))* IC(hypernym(hypernym(hypernym(x))))*\ldots* IC(Rote) \quad (4)$$

**Translation of English queries to their equivalents in Persian:** Different approaches for translation would be introduced for multi-lingual information retrieval among which three common approaches are: dictionary based methods (Adriani and Van Rijsbergen, 1999; Ballesteros and Croft, 1997; Gao *et al*., 2001), corpus-based methods (Littman *et al*., 1998) and machine-based translation methods (Yamabana *et al*., 1998). In dictionary-based approaches, bilingual dictionaries which are readable for machine are used for translation. We use this method in our approach with a partial change. For promotion of translation and achieving higher precision, the phrasal translation is used. It means that instead of word by word translation, we use phrasal translation and phrases would be given to the bilingual English- Persian dictionary and its Persian equivalent will be substituted. Compared with words, phrases have less numbers of equivalents. It's an advantage which decreases the ambiguity in translation. The procedure is to give the phrase to the dictionary and if the phrase doesn't exist, it will be broken into smaller words or phrases in a way that words on the beginning and end of the phrase are separated and the remaining parts will be tested in four stages:

**Stage 1:** If the phrase doesn't exist, the last word of that phrase would be separated from it and the rest of the phrase would be tested. If it exists in dictionary, it will go to stage 4 and if it doesn't exist, it will go to stage 2.

**Stage 2:** In this stage, the last word which has been separated in the previous stage would be returned to its place and the first word of it would be separated. The rest of the phrase would be tested. If it is in dictionary, we will go to stage four and if it is not, we will go to stage 3.

**Stage 3:** In this stage, both the first and the last words would be separated from the phrase and the rest of the phrase would be tested. If it exists, it will go to stage four and if it does not exist, all of the above stages will be carried out again on this phrase.

**Stage 4:** In this stage, the considered phrase exists in the dictionary, so, it will be substituted with an equivalent phrase and would be deleted and the process will go on for other phrases.

The point is that in stage 1 and 2, the phrase will be divided into two parts and in stage 3; the phrase would be divided into three parts. When we come to a phrase which exists in the dictionary, after substitution with its equivalent in the dictionary and the deletion of that phrase, the process will be repeated for all of the words before and after that phrase, which together constitute another phrase. The worst case is the one in which no compound phrase is found and the system will be forced to translate word by word.

**Semantic disambiguation by creating the graph of concepts:** One of the problems of using dictionaries is offering different meanings for a specific word or phrase. In the proposed method, creating the graph of concepts technique is used. This graph is a tool for disambiguating of query translation into Persian. Also, it is used for selecting the proper meaning of extracted set from dictionary. We start with a word from a set of Hamshhri's index words (Darrudi *et al*., 2004). This word is considered as a concept. For creating the final graph of

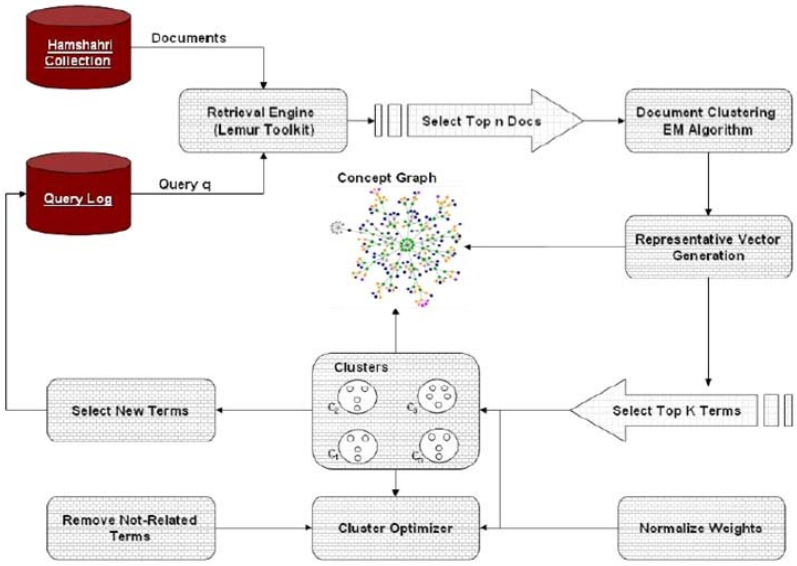


Fig. 2: Process of assessing concepts and their relations (Amiri *et al*., 2008)

concepts, related concepts to primary concept are recognized and distinguished. This operation is done for Wikipedia corpus (Fig. 2) and our work is similar to the process proposed in this reference

Since any word in different fields may have different concepts, applications and values, we need to recognize the domain of each word (Huang *et al*., 2007; Jain *et al*., 1999; Sowa, 1976). For this, EM clustering algorithm (Witten and Frank, 2005) is used. In each cluster there is a set of documents. To select a vector of words for each cluster, existing documents of that cluster is used. For recognizing the best words of the vector we need to weight them. The words with higher weight show the field of that cluster better. For weighting the words the TF/IDF criteria (formula (5)) can be used:

$$w_{d,t_i} = \frac{tf(t_i,d)*(C_{doc} - df(t_i))}{\sum_i tf(t_i,d)*C_{doc}} \quad (5)$$

where $tf(t_i,d)$ is repetition of word $t_i$ in document d, $\sum_i tf(t_i,d)$ shows all the words of document d, $df(t_i)$ is the number of documents which has the word of $t_i$, $C_{doc}$ is the total number of documents in corpus and $w_{d,ti}$ shows the weight of $t_i$ in document d.

Formula (5) is calculated for existing words in each document separately and it is not appropriate to compare weights of two documents in one cluster. For this formula (6) is used:

$$w'_{d,t_i} = \frac{w_{d,t_i} - Min(w_{d,t})}{Max(w_{d,t}) - Min(w_{d,t})} + c \quad (6)$$

where $Min(w_{d,t})$ and $Max(w_{d,t})$ specify the lowest and highest weight of words of document d respectively and C constant is for avoiding weights to become zero. Formula (2) makes the weights normalize and provide the possibility of fuzzy retrieval (Amiri *et al*., 2007).

Since it may possible that a word repeats in several documents, the final weights for words of each cluster are calculated with formula (7):

$$w_{c,t_i} = \frac{\sum_{j=1}^{NoDocs} w'_{d_j,t_i}}{NoDocs_c} \quad (7)$$

where $w_{C,t_i}$ the weight of ti is in C cluster, $w'_{d_j,t_i}$ is the weight of word of $t_i$ in $d_j$ document and $No\ Docs_c$ is the number of existing documents in C cluster. In formula (7) if a document has not the word $t_i$, weight of the $t_i$ on that document is considered zero. This equation helps to find the more valuable words by devoting more weight to them. Thus, the algorithm would be the same as below: First one word is selected from the repository stochastically. Then, all the documents which include the selected word are found clustered and the word vector of each cluster is determined. At this stage the graph of concept for selected word is created. In order to create the graph of concept of the total repository, for each word of first word vector, the graph of concept should be created. Through this operation, several new word vectors will be appeared. Again for all new words of these word vectors, the graph of concept should be created. This procedure is repeated until the graphs of concept of all the word of repository are created.

This graph is used for estimating the query translation probability in cross-language information retrieval. Relation of two concepts in this graph is evaluated by calculating the weights of relations among them. Relation degree of two concepts is calculated by finding the shortest path between the concepts. To calculate the weights of concepts and existing path between them neighborhood matrix multiplication of graph is used. This method has a high time complexity and is time consuming. But in this research, graph of concepts which is used is not too big to use. Of course there are many of efficient methods which are proposed by researchers and they can be used for big graphs (Hanghang Tong, 2006).

## EVALUATION AND EXPERIMENTAL RESULTS

For evaluating the proposed method, we used 50 English queries and textual corpus of Hamshahri newspaper (Darrudi *et al*., 2004). These queries are translated into Persian by means of a dictionary, disambiguating method and some other methods proposed here. And the outcome is given to the LEMUR context-free search engine which has been applied for search on the Hamshahri textual corpus.

Hamshahri corpus is one of the biggest test corpora in Persian language which is based on the features of TREC conference. In TREC conference, a technique has been used in which a thesaurus of documents was made for every topic and then those documents were evaluated in terms of being related or not being related to the intended topic. In TREC conference, there is a human-judged file which all things in it are related to specific topics. So, they are judged by human beings and are considered to be ideal. Different proposed methods in bilingual retrieval systems are evaluated and graded in terms of their correspondence to this human- judged file. In this paper different experiments have been graded in the same way.

Experiments in Fig. 3 can be explained as follow:

- **Monolingual:** it's an experiment in which queries are responded by human agents rather than machines. This experiment is considered as an ideal one. In BAEWTCG, BAEPTAW, BAEPTCG and BANEPTCG experiments, queries are responded by machine and all these efforts are done to be closer to this ideal experiment.
- **BAEWTCG:** In this experiment queries are expanded by using WordNet ontology and for translating the queries; word by word translation method is used. Then, the best combinations are calculated by graph of concepts (Aleahmad *et al*., 2007) and sent to the search engine.

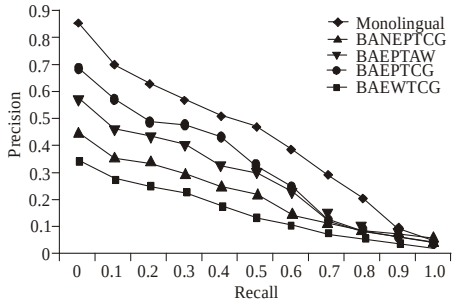


Fig. 3: Comparison of different methods in cross-language information retrieval

Table 1: Average of precision and recall for different methods of cross-language information retrieval

| Test Id | Test name | Tot-ret | Rel-ret | Tool | Map |
|---|---|---|---|---|---|
| * | Monolingual | 5161 | 1970 | 26.41 | lemur |
| 1 | BAEWTCG | 5161 | 83 | 1.14 | lemur |
| 2 | BAEPTAW | 5161 | 586 | 7.85 | lemur |
| 3 | BAEPTCG | 5161 | 1112 | 15.37 | lemur |
| 4 | BANEPTCG | 5161 | 304 | 4.21 | lemur |

- **BAEPTAW:** In this experiment queries are expanded by using WordNet ontology and for translating the queries; the proposed translation algorithm is used. Then, all different combinations are used and sent to the search engine.
- **BAEPTCG:** In this experiment queries are expanded by using WordNet ontology and for translating the queries, the proposed translation algorithm is used. Then, the best combinations are calculated by graph of concepts and sent to the search engine.
- **BANEPTCG:** In this experiment all queries are used with no expansion and for translating the queries; the proposed translation algorithm is used. Then, the best combinations are calculated by graph of concepts and sent to the search engine.

In Table 1, "Test Id" is the number of experiments, "Test Name" is the name of experiment, "Tot-Ret" is total number of related documents in assessment corpus, "Rel-Ret" is the number of retrieved related documents, "MAP" is the average of precision and "Tool" is the name of search engine. Table 1 show that BAEPTCG with the average precision of 15.37 is the closest one to the ideal condition with the average precision of 26.41.

## CONCLUSION

Present information retrieval systems without understanding existing concepts in user query and source documents find patterns similar to that of user query in existing source documents. Also, in cross-language retrieval systems which are using query translation to

extract the information, ambiguity of words caused some problems for selecting the exact translation of a word. All these issues lead to decrease of precision and recall factors in this type of systems. In this paper for understanding the user informational needs and to improve the rate of precision and recall factors, a new conceptual information retrieval method is proposed.

The specifications of this new method are:

- Using the ontology and extending the phrases based on extracting the semantic similarity in ontology
- Phrase translation instead of word by word translation
- Semantic disambiguation of phrases translation by creating the graph of concepts

## REFERENCES


Adriani, M. and C.J. Van Rijsbergen, 1999. Term Similarity-Based Query Expansion for Cross-Language Information Retrieval. In Proceeding of the Third European Conference on Research and Advanced Technology for Digital Libraries (ECDL'99). Paris, France, pp: 311-322.

Aleahmad, A., P. Hakimian, F. Mahdikhani and F. Oroumchian, 2007. NGram and Local Context Analysis for Persian Text Retrieval. International Symposium on Signal Processing and its Applications ISSPA 2007, Sharjah, UAE.

Amiri, H., A. AleAhmad, F. Oroumchian, C. Lucas and M. Rahgozar, 2007. Using OWA Fuzzy Operator to Merge Retrieval System Results. Computational Approaches to Arabic Script-based Languages (CAASL 2007), Stanford University, USA.

Amiri, H., A. A leAhmad, M. Rahgozar and F. Oroumchian, 2008. Keyword Suggestion Using Concept Graph Construction from Wikipedia rich Documents. ACM SIGIR Forum, ISSN: 0163-5840, pp: 55-58.

Alpcan, T., C. Bauckhage and S. Agarwal, 2007. An Efficient Ontology-Based Expert Peering System. In Proceeding IAPR Workshop on Graph-based Representations, pp: 273-282.

Ata, B.M.A., T. Mohd, T. Sembok and M. Yusoff, 1995. SISDOM: A multilingual document retrieval system. Asian Libraries, 4(3): 37-46.

Ballesteros, L. and W.B. Croft, 1997. Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval. In Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '97). ACM Press, Philadelphia, PA, USA, pp: 84-91.

Darrudi, E., M.R. Hejazi and F. Oroumchian, 2004. Assessment of a Modern Farsi Corpus. The Second Workshop on Information Technology and its Disciplines, WITID.

Fellbaum, C., 1998. Word Net: An Electronic Lexical Database. MIT Press, Cambridge, USA.

Gao, J., J.Y. Nie, E. Xun, J. Zhang, M. Zhou and C. Huang, 2001. Improving Query Translation for Cross- Language Information Retrieval using Statistical Models. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '01). ACM Press, New Orleans, Louisiana, USA, pp: 96-104.

Hanghang Tong, C.F., 2006. Center-piece subgraphs: Problem definition and fast Solutions. 12th ACM SIGKDD international conference on Knowledge discovery and data mining, Philadelphia, PA, USA, pp: 404-413.

Huang, W.C., A. Trotman and S. Geva, 2007. Collaborative knowledge management: Evaluation of automated link discovery in the Wikipedia. The SIGIR 2007 Workshop on Focused Retrieval, pp: 9-16.

Jagarlamudi, J. and A. Kumaran, 2007. Cross- Lingual Information Retrieval System for Indian Languages. 8th Workshop of the Cross- Language Evaluation Forum, pp: 80-87.

Jadidinejad, A.H. and M.R. Keyvanpour, 2008. Extracting and Organizing Purpose-built Scientific Features of Wikipedia the Free. Multilingual and Open Content Encyclopedia as a Valuable Knowledge-Base in Data Mining area. Amir Kabir University, Iran, IDMC 2008.

Jain, A.K., M.N. Murty and P.J. Flynn, 1999. Data Clustering: A review. ACM J. Comput. Survives, 31(3): 264-323.

Littman, M.L., S. Dumais and T.K. Landauer, 1998. Automatic Cross-Language Information Retrieval using Latent Semantic Indexing. In: Grefenstette, G., (Ed.), Cross-Language Information Retrieval, Chapter 5, Kluwer Academic Publishers, Boston.

Lin, D., 1998. An information- theoretic definition of similarity. In Proceedings of the 15th International Conference on Machine Learning, Madison, WI, pp: 296-304.

Sadat, F., M. Yoshikawa and S.H. Uemura, 2002. Cross-language information retrieval using multiple resources and combinations for query expansion. Adv. Inf. Syst. Lect. Notes Comput. Sci., 2457: 114-122.

Sebti, A. and A.A. Barfroush, 2008. A new word sense similarity measure in WordNet. International Multi conference on Computer Science and Information Technology, IEEE, Poland, pp: 369-373.

Studer, R., 2000. Situation and Perspective of Knowledge Engineering. In: Cuena, J., (Eds.), Knowledge Engineering and Agent Technology. IOS Press.

Sowa, J.F., 1976. Concept Graphs for a Data Base Interface. IBM J. Res. Dev., 20(4): 336-357.

Teymoorian, F., M. Mohsenzadeh and A. Seyyedi, 2009. Using Concept Graph to Increase Bilingual Text Retrieval Precision. IEEE International Conference on Digital Ecosystems and Technologies, Istanbul, Turkey.

Witten, I.H. and E. Frank, 2005. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco.

Yamabana, K., K. Muraki, S. Doi and S. Kamei, 1998. A Language Conversion Front-End for Cross-Language Information Retrieval. In: Grefenstette, G., (Ed.), Cross-Language Information Retrieval. Chapter 8, Kluwer Academic Publishers, Boston.