World Applied Sciences Journal 8 (8): 965-971, 2010
ISSN 1818-4952

Concept-Based Information Retrieval with Ontology Approach for Cross-Language Searching

[1]Morteza Poyan rad, [2]Hamid Hassanpour and [3]Reza Pourshaikh

[1]Qazvin Islamic Azad University, Electrical & Computer Engineering Faculty, Member of Young Research Club
[2]Islamic azad university, Ghaemshahr Branch, Iran
[3]Qazvin Islamic Azad University, Electrical & Computer Engineering Faculty

Abstract : Internet web is a rich source of information. Language of information may restrict their usability to many users. This paper proposes a method to retrieve information written in a language different from the user's language. This kind of retrieval is done by concept induction from the user's queries and it is equipped with ontology; it translates the phrases instead of the words and does semantic disambiguation by using probability calculation of combinational translation, which has good effects on the results.

Key words: Information retrieval · Cross-language · Ontology · Semantic disambiguation

INTRODUCTION

Cross-language information retrieval (CLIR) is a retrieval process in which the user presents queries in one language to retrieve documents in another language. Due to the 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 [1, 2].

Simple approaches in multi -lingual dictionary- based information retrieval are using all the different meanings of words in queries during the translation which leads to ambiguities [3]. Some approaches are trying to solve the problem by using correlation criteria of one translated word to all the other translated words of that query. In this kind of approaches, when a translation appears similar to most of other translations of that word in the query, high degree of correlation would be assigned to that specific query translation. After selecting existing equivalent words, retrieving will be carried out based on their correlation degree. In some other approaches, only the translations with high degree of correlation would be selected for any word of query [4]. In some other approaches, the word will be selected when its correlation degree is higher than a certain limit [5]. The problem with the approaches based on degree correlation is that, different translations may exist for a word.

This paper proposes a method to retrieve information written in a language different from the user's language. This kind of retrieval is done by concept induction from the user's queries and it is equipped with ontology; it translates the phrases instead of the words and does semantic disambiguation by using probability calculation of combinational translation, which has good effects on the results.

Remaining parts of this paper are organized as follows: section 2 focuses on Ontology and WordNet, Section 3 describes our proposed approach for English-Persian bilingual information retrieval. The evaluation and experimental results are discussed in section 4. Finally we conclude our paper in section 5.

Ontology & Wordnet: Ontologies can be regarded as general tools of information representation about a topic. They can have different roles depending on the application domain and the level of specificity at which they are being used. In general, ontologies can be distinguished into domain ontologies, representing knowledge of a particular domain and generic ontologies representing common sense knowledge about the world [6,16].

Corresponding Author: Morteza Poyan rad, Qazvin Islamic Azad University, Electrical & Computer Engineering Faculty, Member of Young Research Club

Table 1: defined connections in WordNet

Semantic Relation	Meaning	Example
Hypemym	X is a kind of F(X)	Apple is a kind of fruit
Hyponym	F(X) is a kind of X	Zebra is a kind of Horse
Holonym	X is a part/member of F(X)	Wheel is a part of a car
Meronym	X has part/member F(X)	Table has part leg
Antonym	F(X) is the opposite of X	Wet is the opposite of dry

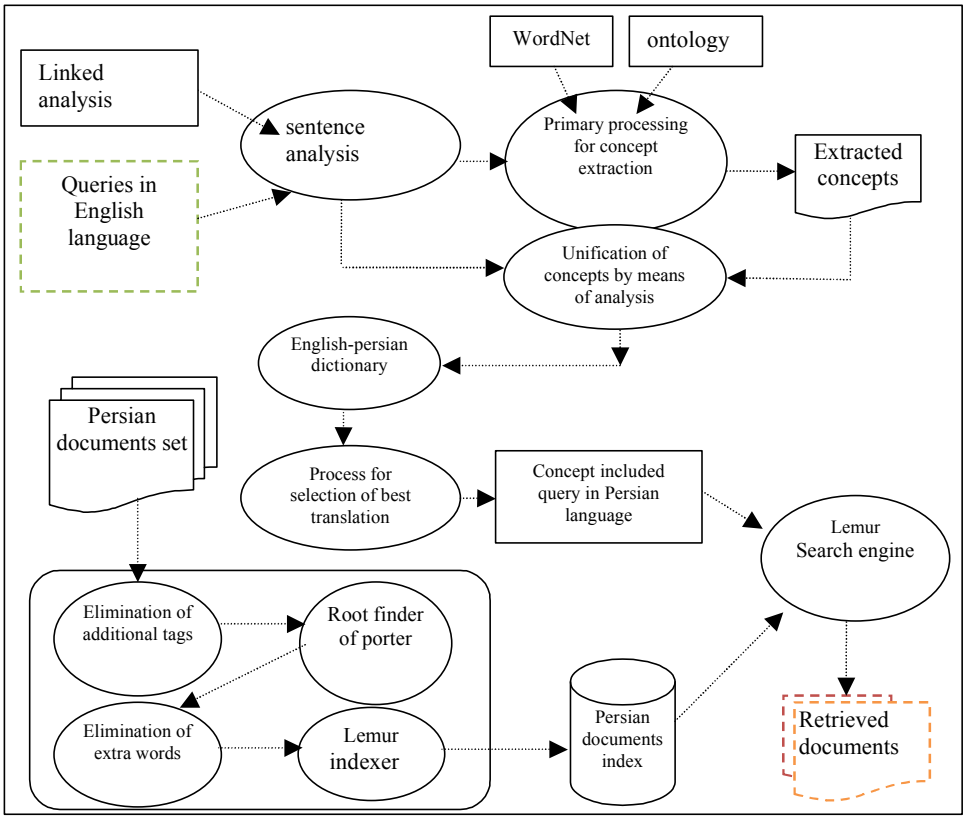

Fig. 1: Proposed approach architecture for improvement of English - persian bilingual retrieval

WordNet [7] is a lexical system manually constructed by a group of people led by George Miller at the Cognitive Science Laboratory at Princeton University. WordNet is organized in sets of synonyms (*synsets*), the words with the same meaning. There are different relations among the words in a synset. The relation of hypernymy/hyponymy (is-a relation) is the principal relation and creates a hierarchic structure. Meronymy/holonymy (part-of relation) is another kind of relation. In addition, WordNet is divided into four taxonomies by the type of words: nouns, verbs, adjectives and adverbs. We only use the taxonomy of nouns because nouns are the most content-bearing words. Expansion terms are selected from the suitable synsets for each noun in the query. In Table 1 some of semantic connections in WordNet with some examples are presented.

Proposed Approach for English-Persian Bilingual 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 that 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 which 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 proposed method is shown in Figure 1.

It is worth mentioning that this approach is not specific for 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 [9] and for this work IC (Information Content) is defined as follow.

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

Benefits of IC are: 1) It shows the degree of a concept specialty in the domain of its topic. 2) A concept with high information content is considerably specific. 3) 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:

$$p(concept) = \frac{1}{number\ of\ hyponym\ for\ concept} \quad (2)$$

For calculating the standard of similarity, we apply a formula which Lin discussed in [10] 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:

$$\begin{aligned} &IC(x) = IC(hypernym(x))* \\ &IC(hypernym(hypernym(x)))* \\ &IC(hypernym(hypernym(hypernym(x))))*...* \\ &IC(Rote) \end{aligned} \quad (4)$$

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 [11, 12 and 13], corpus-based methods [14] and machine-based translation methods [15]. In dictionary-based approaches, bilingual dictionaries which are readable for machine are used for translation. We use this method in our approach with a little 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 Using Calculation of Combinatorial Translation Probability: As it was mentioned one of the problems of using multi- lingual dictionaries is the presence of different meanings for a special word or phrase. In the proposed method, exploiting the contextual information and reciprocal information of words in textual corpus are used both for disambiguation during the translation of the Persian queries and for proper semantic selection from among the extracted semantic sets from the dictionary. To do this, different meanings in the English query will be extracted out by means of a bilingual English- Persian dictionary and by using a textual corpus. And then the number of simultaneous occurrence of all different two-word phrases in query would be found. In a way that for a query like Q, the translations set of Q members would be considered as follow: E (Q) = {T (q_1), T (q_2), T (q_3)… T (q_n)} in which q_i is a one-word or multi-word phrase and T would be the translation function. Now the probability that q_1 and q_{i+1} in the query will be translated into X and Y respectively, would be calculated by using the following formula:

$$p(q_i \to x, q_{i+1} \to y) = \frac{p(x,y)}{c + p(x)p(y)} \quad (5)$$

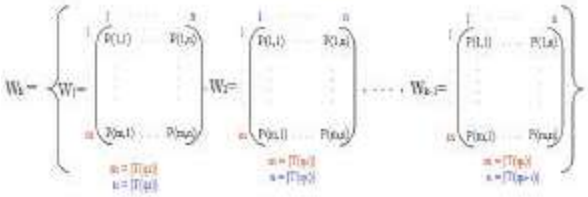

Fig. 2: Calculation of probability matrix for every two phrases q_k and q_{k+1}

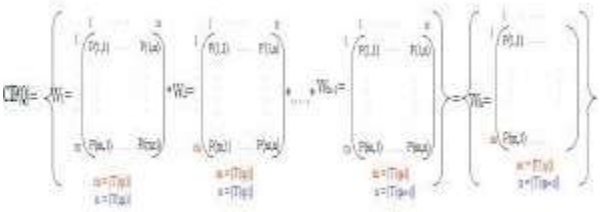

Fig. 3: Calculation of combinatorial translation probability matrix

In which P(x,y) is the number of simultaneous occurrence of two words of X and Y in a sentence, P(x) is the number of repetition of word X, P(y) is the number of repetition of word Y and C is a numeric constant for preventing the denominator to be zero. So translation probability matrix, W_k, for every two consequent phrase of query translation would be constituted as follow:

$$W_k = \{w_{m,n}\} \ \ (m = 1...|T(q_k)|, n = 1...|T(q_{k+1})|) \quad (6)$$

So, compound translation probability matrix would be made by multiplying in W_K like below. Size of this matrix is $|T(q_1)|*|T(q)|$.

$$CTP(Q) = W_1 * W_2 * ... * W_{k-1} \ \ (k = 1...m) \quad (7)$$

In other words, CTP (Q) is a matrix which may have the probability of translating the query elements of Q into possible Persian equivalents and by using this matrix we can consider those translations which have the highest amount of probabilities and at the same time exceed a certain threshold. Since in this method the most probable meanings are always selected, it will be possible that sometimes incorrect meanings will be considered. Although these cases rarely happen.

Evaluation and Experimental Results: For evaluating the proposed method, we used 50 English queries and textual corpus of HAMSHAHRI newspaper [8]. 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. In information retrieval systems, Precision and Recall are two highly important parameters. These Parameters would be defined as follow:

Precision: ratio of related retrieved documents to all the retrieved documents.

Recall: Ratio of related retrieved documents to all the related documents.

By defining the table of distribution for retrieved documents and it's relation to user's requirement in Table 2, Precision and Recall is defined as follow:

$$\Pr ecision = \frac{T_p}{(T_p + f_n)} \tag{8}$$

$$\mathrm{Re}call = \frac{T_p}{(T_p + f_n)} \tag{9}$$

Precision and Recall parameters which are presented in formulas (8) and (9) are not so suitable for information retrieval systems, because these two parameters in such systems are completely dependant on each other, while formulas (8) and (9) are completely independent from each other. To make these formulas proper for information retrieval systems, different amounts of precision would be calculated for different percentages of recall and results are shown in form of a diagram on which the horizontal axis represents recall percentages and vertical axis represents the amount of precision. It should be mentioned that this diagram shows the visual results of all the calculated queries and the final result would be the average of all numbers. In Figure 4 and Table 3 different experiments for retrieval of information in bilingual methods has been shown. As it is obvious from this figure, proposed method in this paper has a good ratio of precision-recall and it is close to the ideal condition.

Table 2: Distribution of retrieved documents and it's relation to the requirement of users

Not related	Related		
f_p	T_p	Retrieved	Not retrieved
T_n	f_n		

Experiments in Figure 4 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 BANEPTCTPM, BANEPTCG, BAEPTAW, BAEPTCTPM, BAEWTCTPM and BAEWTCG experiments, queries are responded by machine and all these efforts are done to be closer to this ideal experiment.

BANEPTCTPM: In this experiment all queries are used without any kind of expansion and for translating the queries into and responding them in destination language, the proposed translation algorithm is used. Then, the best combinations are calculated by means of combinatorial translation probability matrix and are sent to the search engine.

BANEPTCG: In this experiment all queries are used without any kind of expansion and for translating the queries into and responding them in destination language, the proposed translation algorithm is used. Then, the best combinations are calculated by means of conceptual graph [1] and sent to the search engine.

BAEPTAW: In this experiment queries are expanded by using WordNet ontology and for translating the queries into and responding them in destination language, the proposed translation algorithm is used. Then, all different combinations are used and sent to the search engine.

BAEPTCTPM: In this experiment queries are expanded by using WordNet ontology and for translating the queries into and responding them in destination language, the proposed translation algorithm is used. Then, the best combinations are calculated by means of combinatorial translation probability matrix and sent to the search engine.

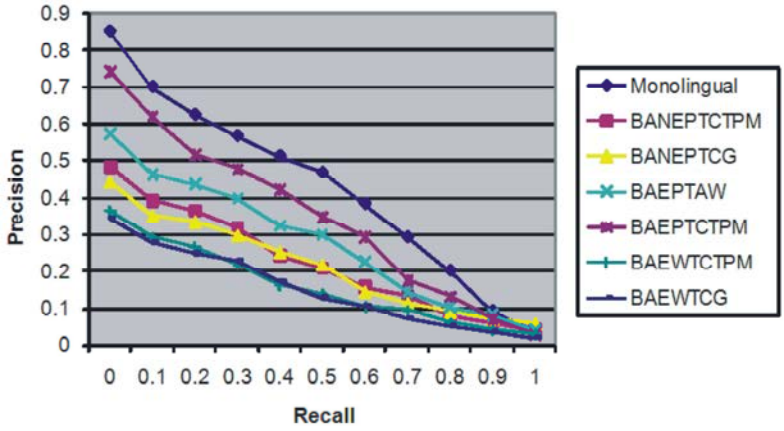

Fig. 4: comparison of different methods in cross-language information retrieval

Table 3: Average of precision and recall for different methods of cross-language information retrieval

Test Id	Test Name	Tot-Ret	Rel-Ret	MAP	Tool
*	Monolingual	5161	1970	26.41	lemur
1	BANEPTCTPM	5161	384	5.31	Lemur
2	BANEPTCG	5161	304	4.21	Lemur
3	BAEPTAW	5161	586	7.85	Lemur
4	BAEPTCTPM	5161	1243	17.19	Lemur
5	BAEWTCTPM	5161	119	1.64	Lemur
6	BAEWTCG	5161	83	1.14	Lemur

BAEWTCTPM: In this experiment queries are expanded by using WordNet ontology and for translating the queries into and responding them in destination language, word by word translation method is used. Then, the best combinations are calculated by means of combinatorial translation probability matrix and sent to the search engine.

BAEWTCG: In t his experiment queries are expanded by using WordNet ontology and for translating the queries into and responding them in destination language, word by word translation method is used. Then, the best combinations are calculated by means of conceptual graph [1] and sent to the search engine.

In Table 3, “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 3 shows that BAEPTCTPM with the average precision of 17.19 is the closest one to the ideal condition with the average precision of 26.41.

CONCLUSION

A lot of current information retrieval systems work in the same way as pattern matching systems. In other words, by obtaining the user information requirements, these systems try to find the patterns similar to the query among the documents in source file. This happens without any realization of concepts in the user's query and existing documents. This matter shows that such search engines can not support the realization of concepts for meeting the informational needs of user and source documents. A substantial approach for this problem is to understand the informational requirements of users and existing source documents. In this paper an approach with the following features is introduced: 1) using ontology and expanding the phrases based on extracting the semantic similarity 2) translating phrases instead of word by word translation and 3) semantic disambiguation of phrase translation by calculating the translation probability.

Using the mentioned features, the proposed approach tries to get the informational needs of users and do the conceptual retrieval. Different experiments done in this paper shows that our proposed method has a good promotion on results.

REFERENCES

1. 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.
2. 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.
3. Davis, M.W., 1996. New Experiments in Cross-Language Text Retrieval, The Fifth Text Retrieval Conference (TREC-5). NIST.
4. Adriani, M., 2000. Using Statistical Term Similarity for Sense Disambiguation in Cross-Language Information Retrieval, Inf. Retr., 2(1): 71-82.
5. Myung-Gil Jang, S.H.M. and S.Y. Park, 1999. Using Mutual Information to Resolve Query Translation Ambiguities and Query Term Weighting, In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 99). Maryland, pp: 145-148.
6. Studer, R., 2000. Knowledge Engineering and Agent Technology, In J. Cuena *et al.* editors, Situation and Perspective of Knowledge Engineering. IOS Press, Amsterdam.
7. Fellbaum, C., editor, 1998. WordNet: An Electronic Lexical Database, MIT Press, Cambridge, USA.
8. 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.
9. Sebti, A. and A.A. Barfroush, 2008. A new word sense similarity measure in wordnet, International Multiconference on Computer Science and Information Technology, IEEE, Poland, pp: 369-373.
10. Lin, D., 1998. An information-theoretic definition of similarity. In Proceedings of the 15th International Conference on Machine Learning, Madison, WI.
11. 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).ACMPress, Philadelphia, PA, USA, pp: 84-91.
12. 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.
13. 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.
14. Littman, M.L., S. Dumais and T.K. Landauer, 1998. Automatic Cross-Language Information Retrieval using Latent Semantic Indexing. In: G. Grefenstette, (eds), Cross-Language Information Retrieval, chapter 5, Kluwer Academic Publishers, Boston.
15. Yamabana, K., K. Muraki, S. Doi and S. Kamei, 1998. A Language Conversion Front-End for Cross-Language Information Retrieval. In: G. Grefenstette, (eds), Cross-Language Information Retrieval, chapter 8, Kluwer Academic Publishers, Boston.
16. Alpcan, T., C. Bauckhage and S. Agarwal, 2007. An Efficient Ontology-Based Expert Peering System, In Proc. IAPR Workshop on Graph-based Representations, pp: 273-282.