Web Query Expansion by WordNet

Zhiguo Gong, Chan Wa Cheang, and Leong Hou U

Faculty of Science and Technology, University of Macau,
P.O.Box 3001 Macao, PRC
{zggong, ma36600, ma36575}@umac.mo

Abstract. In this paper, we address a novel method of Web query expansion by using WordNet and TSN. WordNet is an online lexical dictionary which describes word relationships in three dimensions of Hypernym, Hyponym and Synonym. And their impacts to expansions are different. We provide quantitative descriptions of the query expansion impact along each dimension. However, WordNet may bring many noises for the expansion due to its collection independent characteristic. Furthermore, it may not catch current state of words and their relationships because of the explosive increase of the Web. To overcome those problems, collection-based TSN (Term Semantic Network) is created with respect to word co-occurrence in the collection. We use TSN both as a filter and a supplement for WordNet. We also provide a quantitatively study as what is the best way for the expansion with TSN. In our system, we combine the query expansions along each semantic dimension as our overall solution. Our experiments reveal that the combined expansion can provide a satisfied result for the Web query performance. The methodologies in this paper have been already employed in our Web image search engine system.

1 Introduction

In recent years, huge amount of information is posted on the Web and it continues to increase with an explosive speed. But we cannot access to the information or use it efficiently and effectively unless it is well organized and indexed. Many search engines have been created for this need in current years. Web users, however, usually submit only one single word as their queries on the Web [5], especially for a Web Image queries. It is even worse that the users' query words may be quite different to the ones used in the documents in describing the same semantics. That means a gap exists between user's query space and document representation space. This problem results in lower precisions and recalls of queries. The user may get an overwhelming but large percent of irrelevant documents in the result set. In fact, this is a tough problem in Web information retrieval. An effective method for solving the above problems is query expansion. In this paper, we provide a novel query expansion method based on the combination of WordNet [2], an online lexical system, and TSN, a term semantic network extracted from the collection. Our method has been employed in our Web image search system [4].

WordNet [2], like a standard dictionary, contains the definitions of words and their relationships. But it also differs from a standard dictionary in that, instead of being

K.V. Andersen, J. Debenham, and R. Wagner (Eds.): DEXA 2005, LNCS 3588, pp. 166 – 175, 2005.

organized alphabetically, WordNet is organized conceptually. The basic unit in WordNet is a synonym set, or synset, which represents a lexicalized concept. For example, the noun "software" in WordNet 2.0 has the synsets {software, software system, software package, package} and also Nouns in WordNet are organized in a hierarchical tree structure based on hypernym/hyponymy. The hyponym of a noun is its subordinate, and the relation between a hyponym and its hypernym is an 'is a kind of' relation. As in Fig. 1, "freeware" is a hyponym of "software", or more intuitively, a "freeware" is a kind of "software". Hypernym (supername) and its inverse, hyponym (subname), are transitive semantic relations between synsets.

We use those various semantic relations between words in our query expansion. However, these three relations have different semantic relevances to the query word. The concept at the upper layers of the hierarchy has more general semantics and less similarity between them, while concepts at lower layers or at the same layer have more concrete semantics and stronger similarity [1]. To determine, in quantity, how to expand the query word along each direction, the average precision is used as the objective function for computing the optimal factor for each direction. On the other hand, some terms added to the query will bring some noises and the search may return large amount of irrelevant results, thus decrease the precision [3]. To solve this problem, we use TSN (Term Semantic Network) extracted from our collection to filter out the words with lower supports and confidences to the query word. By this way, noises can be well controlled.

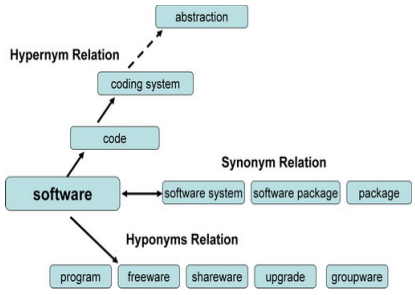

Fig. 1. Hierarchy relation of word "software" in WordNet

Besides noise controlling for WordNet expansion, TSN is also another important supplement for semantic describing between terms. WordNet is systematically created on the base of lexical analysis. It has two intrinsic weaknesses—poor current and collection independent. The first weakness may result in that newly created word semantic relations may not be used in the expansion, and the second one may generate some noise from the expansion. Since TSN is directly extracted from the collection, it can overcome the above shortages. To create TSN, we define association rule between two words in terms of 'Support' and 'Confidence' [12], then the semantic network is extracted with respect to the definition.

In reminder of this paper, we discuss related works in section 2. Section 3 provides our detail methodologies for query expansion with WordNet and TSN. The experiment results are illustrated in section 4. Finally, we conclude our work in section 5.

2 Related Work

As mentioned, query expansion is one of the promising approaches to deal with the word mismatch problem in information retrieval. It can be roughly classified into two groups: global analysis and local analysis.

The basic idea in global analysis is that the global context of a concept can be used to determine similarities between concepts. Context can be defined in a number of ways, as can concepts. The simplest definitions are that all words are concepts and that the context for a word is all the word that co-occurs in document. Similarity thesauri [6] and Phrase Finder [8] are used with those global techniques. Another related approach uses clustering to determine the context for document analysis [9]. Global analysis techniques are relatively robust. However, since the co-occurrence information for every pair of terms in the whole corpus are normally needed, the processing is generally rather computational resource consuming. Moreover, global analysis cannot handle ambiguous terms effectively. It did not show consistent positive retrieval results unless further strategies for term selection could be suggested.

Local technologies analyze only the information in some initial documents, which are retrieved for the original query. Terms are extracted from these documents for query expansion [7]. Local analysis can be further divided into two groups. The first one is called relevance feedback [10], which relies on user's relevance judgments of the retrieved documents. Relevance feedback can achieve great performance if users cooperate well. However, this method is seldom deployed in practice because users are not always willing or able to give sufficient and correct relevance judgment about documents. Local feedback methods [11] are developed to solve this problem. They work without user interaction by assuming the top ranked documents to be relevant. The drawback of these methods is: if the top-ranked documents happen to be irrelevant (this situation is very common in the Web query world), the suggested terms from these documents are also likely to be unrelated to the topic and the query expansion will fail.

Our approach in this paper belongs to the first category. But it is based on combination of the online dictionary WordNet and the collection related TSN. We use WordNet as the similarity thesaurus to expand the Web query. To solve the disadvantage of it, we use collection related TSN to filter out some noise words to improve the precisions of the queries. Furthermore, we provide a qualitative description on word expansion along different semantic dimensions.

3 System Design

In [4], we introduced our Web search system which include a crawler, a preprocessor, an indexer, a knowledge learner and query engine. Web document is gathered by crawler and loaded into the document database by document preprocessor. The indexer creates the inverted index for retrieval. In order to solve the problem of low precision of the query results, we design and implement a query expansion subsystem which performs functions such as keyword expansion, keyword filtering and keyword weighting. We will introduce each process in detail in the following.

3.1 Keyword Expansion

The query keyword used by users is the most significant but not always sufficient in the query phase. For example, if a user query with "computer", he only can get the object indexed by "computer". We use WordNet and TSN to expand the query. With

WordNet, we expand the query along three dimensions including hypernym, hyponymy and synonym relation [2]. The original query "computer", for instance, may be expanded to include "client, server, website, etc." In other words, with those expanded words together, the system could raise both the query precision and recall.

To extract TSN from the collection, we use a popular association mining algorithm – Apriori [12] — to mine out the association rules between words. Here, we only consider one-to-one term relationship. Two functions—*confidence* and *support*— are used in describing word relations. We define *confidence* (*conf*) and *support* (*sup*) of term association $t_i \rightarrow t_j$ as follows, let

$$D(t_i, t_j) = D(t_i) \cap D(t_j) \tag{1}$$

where $D(t_i)$ and $D(t_j)$ stand for the documents including term t_i and t_i respectively. Therefore, $D(t_i) \cap D(t_j)$ is the set of documents that include both t_i and t_j. We define

$$Conf_{ti->tj} = \frac{\| D(t_i, t_j) \|}{\| D(t_i) \|} \tag{2}$$

where $\| D(t_i, t_j) \|$ stands for the total number of documents that include both term t_i, and t_j; and $\| D(t_i) \|$ stands for the total number of documents that include t_i ,

$$Sup_{ti->tj} = \frac{\| D(t_i, t_j) \|}{\|D\|} \tag{3}$$

where $\|D\|$ stands for the number of document in the database.

Those relationships are extracted and represented with two matrixes, we could use them to expand the query keywords. For example, the keyword "computer" has the highest confidence and support with the words "desktop, series, price, driver…etc" which are not described in WordNet but can be used to expand the original query.

3.2 Keyword Filtering

In the next step, we use TSN to eliminate some noise words in the keyword expansion of WordNet. Actually, the comprehensive WordNet often expands a query with too many words. And some of them are low-frequency and unusual words. They may bring in some noises and detract from retrieval performance, thus lead to precision decrease. So it is very important to avoid noises when expanding queries. We use the association rules to remove the expansion words that have lower support and confidence to the original word. In our system, we use the expanded words which have minimum confidence over 0.3 and support over 0.01 with the original query keyword into our query expansion. As the keyword "robot" in Fig 2, we filter out the words "golem, humanoid, mechanical man".

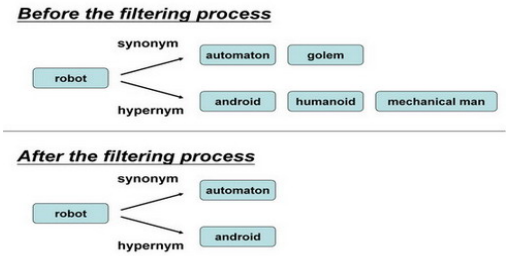

Fig. 2. Keyword filtering process of word "robot"

3.3 Keyword Weighting

In TFIDF model, term *t*'s semantic relevance to web page *p* is measured by *tf(t)*idf(t)*, where *tf(t)* is the frequency of *t* occurring in *p* and *idf(t)* is the inverted document frequency of term *t*. In this paper, we use terms (or concepts) in *p* to derive semantics of the Web image *i*. However, above TFIDF approach can not be directly applied to index the embedded image.

3.3.1 Semantic Relevance of Terms to the Embedded Images

In this paper, we modify TFIDF model regarding following two arguments:

(1) $idf(t_i)$ is used to enhance significances of the terms which appear in less documents, thus can discriminate the corresponding documents effectively. In our system, terms are used to derive the semantics of the embedded images other than discriminate images. Furthermore, image users only use one term or concept for image retrievals in most of the cases. As for those reasons, *idf* is not used in our system.

(2) As we mention in previous sections, we use WordNet and TSN to expand our original query keyword. If a Web page not only contains the original query keyword but also contains the words which expand by WordNet and TSN, it should be more relevant than the ones which only contain the original query keyword. With this observation, we define the total term weight over the whole *p* as

$$ttf(t) = tf(t) + \alpha \cdot tf(t_{Hypernyms}) + \beta \cdot tf(t_{Hyponyms}) + \gamma \cdot tf(t_{Synonyms}) + \delta \cdot tf(t_{LocalContext}) \quad (4)$$

where $t_{Hypernyms}$ is the term set which use WordNet hypernyms relation to expand words, $t_{Hyponyms}$ is the term set which use WordNet hyponyms relation to expand words, $t_{Synonums}$ is the term set which use WordNet synonyms relation to expand words and $t_{LocalContext}$ is the term set which has top-rank TSN expansion words with our original keyword to expand our query. The factor α, β, γ and δ are used to indicate different effects from each expansion direction. And it is natural to suppose $0 < \alpha, \beta, \gamma, \delta < 1$.

In our approach, $ttf(t)|_p$ indicates the semantic relevant value of term *t* to the embedded image *i* embedded in *p*. Thus, the important and challenging task for using this measure is how to determine the values for the factors α, β, γ and δ.

3.3.2 Objective Function of Retrieval Performance

In the area of information retrieval, precision/recall is well accepted evaluation method for the performance of the systems [7, 13]. An ideal information retrieval system is trying to raise the values for both of the two objectives. Since the result of a retrieval is usually long list in size, especially in the World Wide Web environment, a figure of precision versus recall changing is commonly used as a performance measurement for a retrieval algorithm. However, this metric can not be used as an objective function in determining those factor values. Instead, in this study, we employ the single value summaries as our objective function in order to determine the values for the factors [7, 13]. The average precision is defined as

$$AP = \frac{1}{R}\sum_{k=1}^{R_k}\frac{k}{N_k} \tag{5}$$

where R is the total number of all relevant results with respect to $ttf(q_i)|_p$ and N_k is the number of results up to the *k-th* relevant result in the result list. As a matter of the fact, AP is the single value metric which indicates the performance of querying q_k.. In this paper, we assume that all expansion dimensions are independent, thus, we can determine them one by one. For example, the optimal values for α is determined when AP of queries with respect to rank function $ttf(t) = tf(t) + \alpha \cdot tf(t_{Hypernyms})$ reaches its maximum. And we also use the same way for determining other factor values.

Fig 3~6 show the curves of AP values via factor values for α, β, γ and δ respectively. Those figures indicate that query expansion along each dimension can always get some better performance than that without expansion (factor = 0). And the maximums for AP can be obtained with 0<factors<1. Table 1 shows optimal factor values with their corresponding AP values.

Table 1. Factor Values and Average Precision

Factor	*Values*	*Average Precision*
α (Hypernyms)	0.47	0.2406
β (Hyponyms)	0.84	0.3888
γ (Synonyms)	0.70	0.3404
δ (Local Context)	0.94	0.3559

From Table 1, it is clear that different semantic relations have different influences to the search results. Hypemyms relation has less significant impact than Hyponyms and Synonyms relations. With a close study, we find the reason may be due to the fact that Hypermyms relation (abstract concept expansion) may bring more noises than Hyponyms and Synonyms do. So its factor value is less than others. And Local Context (TSN) is closely semantic relevant to the original keyword. Thus, its factor value is very high.

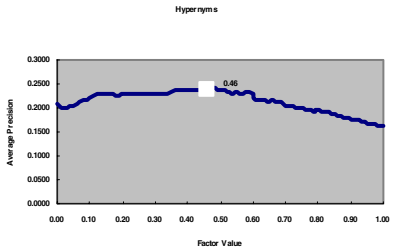

Fig. 3. Average precision versus Factor Values for Hyponyms relation

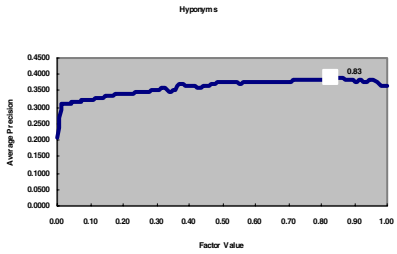

Fig. 4. Average precision versus Factor Values for Hypernyms relation

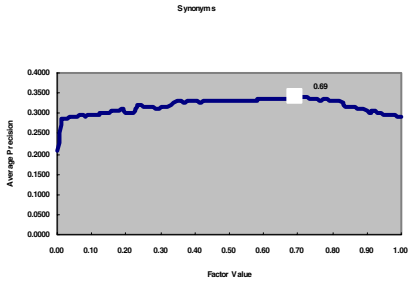

Fig. 5. Average precision versus Factor Values for Synonyms relation

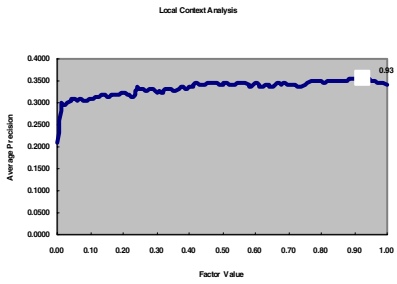

Fig. 6. Average precision versus Factor Values for Local Context Analysis

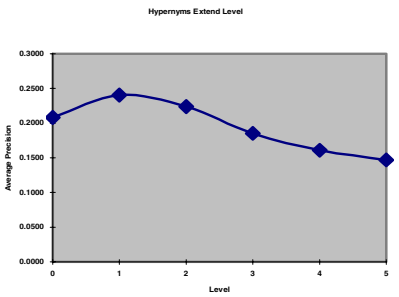

Fig. 7. Average precision versus Extend Level for Hypernyms relation

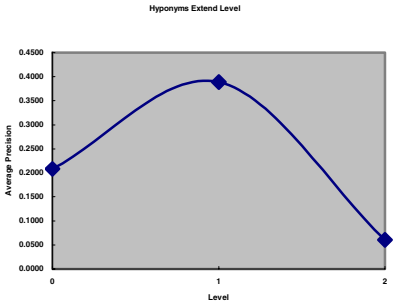

Fig. 8. Average precision versus Extend Level for Hyponyms relation

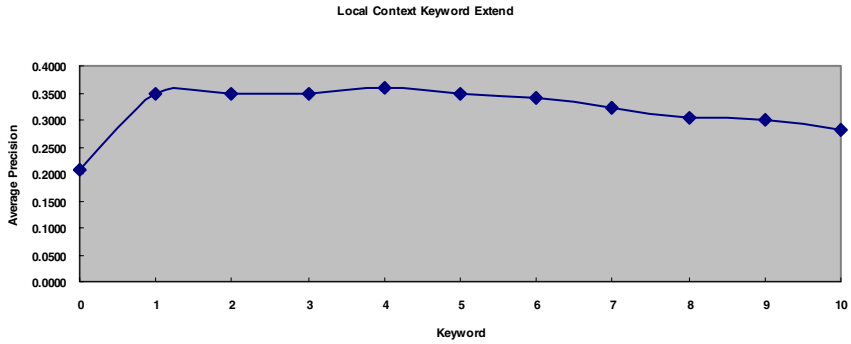

Fig. 9. Average precision versus Number of Keyword expand for Local Context

As a matter of the fact, we can get a word's hypernyms relations recursively up to the top root and hyponyms relations down to the bottom along WordNet. Therefore, it is interesting to know how many levels along hypernyms and hyponyms to expand the queries. For this objective, we use the same method as to calculate the factors by average precision. As in Fig 7~8, our experiments show that one level for expansion will get better results in both the situations. In fact, multiple hyponyms relations expand too many words which may diverge the original keyword meanings. That is, they will generate many noises in the results, thus reduce *AP* values. So in our system, we only use one level for both hypernyms and hyponyms expansions. In Fig 9, we also use the average precision method to calculate how many keywords are proper for expansions along TSN dimension. As the result, we find that using the first 4 top-rank words in TSN expansions can generate better performances.

4 Evaluation

Our experiments are carried out with our Web image search engine [4]. The crawler of our system gathered about 150,000 Web pages with a given set of seeds which are randomly selected from dot-com, dot-edu and dot-gov domains. After the noise images (icons, banners, logos, and any image with size less than 5k) are removed by the image extractor, about 12,000 web images embedded in the Web pages are left. Then, 5 human experts are assigned to define the subjects of the Web images manually. And sometimes, more than one subject is defined for the same images. For example, concepts 'Laptop', 'Notebook' and 'Computer' may be annotated to the same Web image. We select 30 concepts from the defined set as our query training set, and another 15 terms as our test samples. We measure the query performances with the precision-recall curves.

As shown in Fig 10, we compare the performances of original query, expanded query, keyword filtered query, *TSN* expanded query and the combined query in the experiments. Comparing with the original query, even though pure WordNet expansions can improve query recall dramatically, however, its precision improvement

is limited. WordNet expansion with filtering by *TSN* is much better than both the original query and the pure WordNet expansion. It shows that WordNet with TSN filtering yields a significant increase in the number of correct documents retrieved and in the number of relevant results for the queries. The TSN expansion's recall and precision are both better than that of pure WordNet expansion, however, are lower than WordNet-TSN-Filtering expansion. The last curve in the figure shows the performance of combining WordNet-TSN-Filtering with TSN expansions. As the result, it is much-improved than any other alone. It combines advantages using the words which could not be expand in WordNet but are highly supported by TSN. Therefore, they can provide good retrieval performances for the web images search.

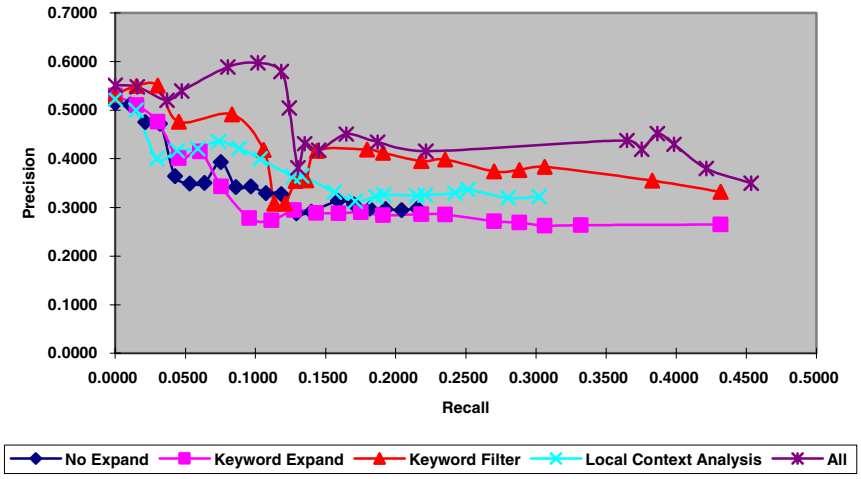

Fig. 10. Performance of Different Query Expansion Method

5 Conclusion and Future Work

In this paper, we propose a method for query expansions. We use WordNet nouns hypernym/hyponymy and synonym relation between words to expand the query words. And we use association rules to find collection dependent term relationships (TSN), and further to use TSN both as a filter and a supplement for WordNet expansion. In our approach, we use average precision (*AP*) as the objective function in calculating the optimal factor values and the expansion levels of hypernyms and hyponyms, and also use it to find the optima number of keywords to expand with TSN. The experiments show that the result of our combined query expansion is much better than only using WordNet or TSN along. The methodologies addressed in this paper are already exploited in our Web image search engine [4]. But some limitations also exist in the current work.

In our current work, we only assume the queries are single-word queries. Even though Web users often use only single words as their queries, multiple word queries

can always dramatically reduce the search scope in the explosive Web document space. However, the words used in the same query may be semantically relevant with each other. That is, they are not independent. If we expand all the words in the same query independently as the overall expansion for the query, the performance may not be the ideal one. We plan to address this problem in our future work. That is, in query expansion, we will consider the semantic similarity between words and determine which word could be used for extension [14] to get a better result.

References

1. Yuhua Li, Zuhair A. Bandar, and David McLean, *An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources*, Knowledge and Data Engineering, IEEE Transactions on , Volume: 15 , Issue: 4 , July-Aug. 2003 Pages:871 - 882
2. Miller, G. A., Beckwith, R., Felbaum, C., Gross, D., and Miller, K., *Introduction to WordNet: An On-line Lexical Database,* Revised Version 1993.
3. Qianli Jin, Jun Zhao, and Bo Xu, *Query Expansion Based on Term Similarity Tree Model*, Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003 International Conference on , 26-29 Oct. 2003 Pages:400 – 406
4. Zhiguo Gong, Leong Hou U and Chan Wa Cheang, *An Implementation of Web Image Search Engine*, Digital Libraries: International Collaboration and Cross-Fertilization: 7th International Conference on Asian Digital Libraries, ICADL 2004, Shanghai, China, December 13-17, 2004. Proceedings Pages:355 – 367
5. S. Lin, M.C. Chen, J. Ho and Y. Huang, *ACIRD: Intelligent Internet Document Organization and Retrieva*l, IEEE Transactions on Knowledge and Data Engineering, 14(3), 2002, pp. 599-614.
6. Qiu, Y. and Frei, H. P., *Concept based query expansion*, In Proceedings of ACM SIGIR International Conference on Research and Development in Information Retrieval, ACM Press, 160-170, 1993
7. R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley, 1999
8. Jing, Y. F. and Croft, W. B., *An Association Thesaurus for Information Retrieval*, In RIAO 94 Conference Proceedings, p. 146-160, New York, Oct. 1994
9. Crouch, C. J., and Yang, B., *Experiments in automatic statistical thesaurus construction*, In Proceeding of ACM SIGIR International Conference on Research and Development in Information Retrieval, 1993, pp. 77-88
10. Rocchio, J.Y., *Relevance Feedback in Information Retrieval*, The SMART Retrieval System. Engelwood Cliff, N.J.: Prentice Hall, PP. 313-323, 1971
11. Buckley, C., Singhal, A., Mitra, M., and Salton, G., New Retrieval Approaches Using SMART: TREC 4, In Harman, D., editor, Proceedings of the TREC 4 Conference. National Institute of Standards and Technology Special Publication. 1996
12. R. Agrawaland R. Srikant, "Fast Algorithms for Mining Association Rules," *Proc. 20th Int'l Conf. Very Large Data Bases, (VLDB)*, Sept. 1994
13. R. K. Sriari, Z. Zhang and A. Rao, Intelligent indexing and semantic retrieval of multimodal documents, *Information Retrieval 2(2)*, Kluwer Academic Publishers, 2000, pp. 1-37.
14. Jun Yang; Liu Wenyin; Hongjiang Zhang; Yueting Zhuang; *Thesaurus-aided approach for image browsing and retrieval,* Multimedia and Expo, 2001. ICME 2001. IEEE International Conference on , 22-25 Aug. 2001 Pages:1135 – 1138.