

Ontology Driven Query Expansion for Better Image Retrieval

Payal Gulati

Department of Computer Engineering
YMCA University of Science &
Technology, INDIA

Dr. A.K. Sharma

Department of Computer Engineering
YMCA University of Science &
Technology, INDIA

ABSTRACT

As more information becomes available on the Web it is more difficult to provide effective search services for Internet users. Since, it is assumed that users do not always formulate search queries using the best terms. So, search engines invoke query expansion to increase the quality of user search results. Query expansion is useful in reducing query/document mismatch by expanding the query using words or phrases with a similar meaning or some other statistical relation to the set of relevant documents. This paper proposes an approach for ontology driven query expansion for better image retrieval. In this work the association rules; generated by mining web logs and WorldNet have been used for query expansion process. Further, an algorithm has been proposed and the performance has been measured.

Keywords

Query Expansion, Association rule mining, WorldNet, web log.

1. INTRODUCTION

The size of publically indexed World Wide Web has probably surpassed 23.99 billion pages in July 2010 [1] and as yet growth shows no sign of leveling off. Even the number of digital images over the web is exponentially rising. As a result prevailing search engines like Yahoo, AltaVista, Google return thousands of matches in response to a user query. Since, it is assumed that users do not formulate search queries using the best terms. Therefore, search engines invoke query expansion (QE) to increase the quality of user search results. Query expansion is the process of reformulating a seed query to improve retrieval performance in information retrieval operations. In the context of web search engines, query expansion involves evaluating a user's input (what words were typed into the search query area, and sometimes other types of data) and expanding the search query to match additional documents. Thus, it is useful in reducing query/document mismatch, resulting in a higher ratio of relevant documents in the retrieval results (precision) and a higher ratio of the relevant documents from the collection that are retrieved (recall) [5]. The ontology's can be used for query expansion in which firstly the query terms must be disambiguated so that they map to a unique ontology concept and then terms related in the ontology to the disambiguated concepts are added to the query.

This paper is organized in the following way. Section 2 discusses related work done in this domain. Architecture and Algorithm of the proposed system is presented in section 3 and 4 respectively. Finally Section 5 and 6 comprise of the results and conclusion.

2. RELATED WORK

Query expansion has improved the effectiveness of ranked retrieval by automatically adding additional terms to a query. In [6] an original query is run using conventional information retrieval techniques [7]. Then, related terms are extracted from the top documents that are returned in response to the original query using statistical heuristics. This approach has been shown to be effective on some collections, but results on large collections of web data have been mixed. The work in [8] chooses expansion terms from past user queries directly, rather than using them to construct sets of full text documents from which terms are then selected. The method consists of three phases: ranking the original query against the collection of documents; extracting additional query terms from the highly ranked items; then ranking the new query against the collection. The results show relative improvements over unexpanded full text retrieval of 26%–29%. Liu et al. [2] used noun phrases for query expansion. Specifically four types of noun phrases were identified: proper names, dictionary phrases, simple phrases, and complex phrases. A document has a phrase if all the content words are in the phrase within the defined window, and these documents that have matched phrases are considered to be relevant. They also apply a similarity measure to select the content words in the phrases to be positively correlated in the collection. The work in [3] utilizes the semantics as well as linguistic properties of unstructured text corpus.

In this work the association rules; generated by mining web logs and WorldNet have been used for query expansion process.

3. PROPOSED SYSTEM

The architecture of the proposed system is shown in Figure 1. Following subsection describes its various components.

3.1 Search Interface

It is a user interface, where user types the search query. The user search query is forwarded to recommendation engine, which then give user, suggestions, by returning various contexts of the query been fired by the user.

3.2 Query log

The query log provides detailed and summarized information about queries. The query log lists the time that each search occurred, the IP address of the web user performing the search, the number of hits for the search, and the user's query.

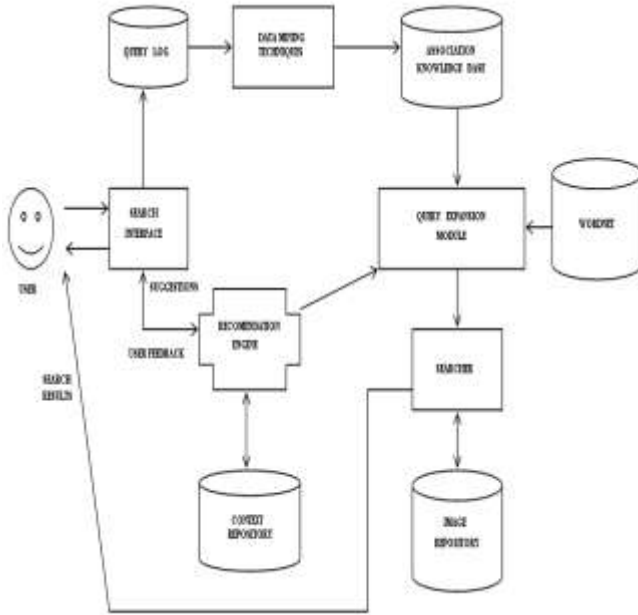


Figure 1. Proposed System Architecture

3.3 Data Mining Techniques

Data mining is the process of extracting hidden patterns from data. In this paper, association rules are applied. Association rules mining are a major pattern discovery technique [6]. The original goal of association rule mining is to solve market basket problem but the application of association rules are far beyond that [7]. The major statistics computed for the association rules, support and confidence are given in equation (i) and (ii).

$$\text{Support } (X \Rightarrow Y) = P(X \cup Y) \quad \dots (i)$$

$$\text{Confidence } (X \Rightarrow Y) = P(Y | X) \quad \dots (ii)$$

3.4 Association Knowledge Base

The knowledge base is a repository of extracted rules which have been derived using the association rule mining. A knowledge base containing rules is shown in Figure 2.

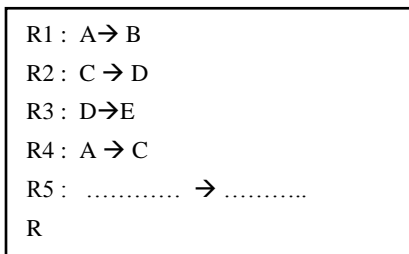


Figure 2 Knowledge Base

3.5 Query Expansion Module

Query expansion (QE) is the process of reformulating a seed query to improve retrieval performance in information retrieval operations. Query expansion involves techniques such as searching for the synonyms as well finding all the various

morphological forms of words by stemming each word in the search query. It also deals in fixing spelling errors and automatically searching for the corrected form or suggesting it in the results. In this work, query is expanded by all the distinct contextual pairs of the words using Wordnet.

3.6 Wordnet

WordNet is a lexical database for the English language. It groups English words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets. Since Google maintains only the word lexicon, it returns synonyms words pair in alphabetical order in response to a given query keyword as discussed by Brin and Page [4]. In comparison to this, Wordnet produce more meaningful pair of words. Hence Wordnet has significance in this work to make search proceed in specific direction and retrieving relevant images.



Figure 3. Contextual meaning for keyword “mouse”

Different senses for keyword “mouse” and its corresponding meaning as per WordNet Search 3.0 is given in Table 1 below.

Table 1 Keywords and their Interpretations

Keyword	Interpretation	Meaning
Mouse	Interpretation 1	Rodent
Mouse	Interpretation 2	Shiner, black eye
Mouse	Interpretation 3	Quiet, timid
Mouse	Interpretation 4	Computer Mouse

3.7 Recommendation Engine

The user search query is forwarded to recommendation engine, which then give user, suggestions, by returning various contexts of the query been fired by the user. User then feedbacks the recommendation engine by selecting the context, for instance for the user search term shown in Figure 4 and the suggestions by the recommendation engine and user feedback are shown in Figure 5 and Figure. 6 respectively.

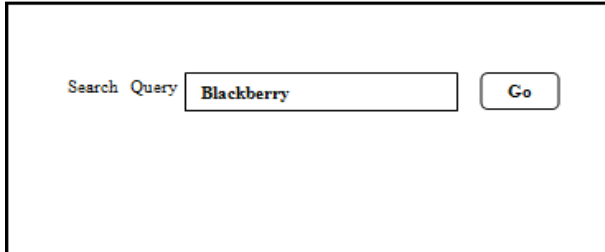


Figure 4. Search Interface

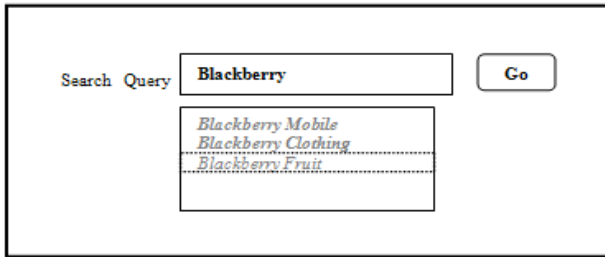


Figure 5. Suggestions by Recommendation Engine

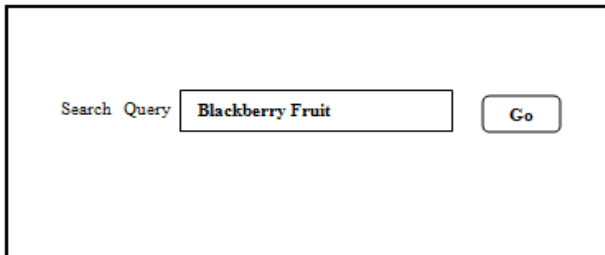


Figure 6 User Feedback and Updated query with context

3.8 Context Repository

It stores all the context information corresponding to the keyword. When user enters the search query, recommendation engine returns various contexts for the query been fired.

Sample context repository is shown in Table 2 below.

Table 2 Keyword & Context

CID	Keyword	Context
C01	Mouse	Rodent., Skinner, black eye, Quiet, timid, Computer Mouse
C02	Blackberry	Clothing, Smartphone, Fruit
C03	Apple	Fruit, Computer, Mobile, tree
.....
.....

3.9 Searcher

Searcher searches Image repository for images corresponding to the user query expanded by query expansion module.

3.10 Image Repository

Search engine component, web crawler collects, parse, and stores images and URL in an image repository to facilitate fast and accurate image retrieval. This repository is then indexed to optimize speed and performance in finding relevant documents for a search query. Without an index, the search engine would scan every document in the corpus, which would require considerable time and computing power. Sample indexed document is shown in Table.3.

Table 3. Indexed Documents

ID	Keyword	Context	File	URL
E01	Apple	Fruit	apple-fruit-3d-	www.mrcad.cc
E02	Apple	iPhone	iphone_apple.	www.letsgomc
E03	Blackberry	Phone	blackberry-bol	www.techgadg
E04	Mouse	animal	Mouse-1.jpg	http://www.ar
E05	Mouse	computer	conclude.jpg	www.tweakne

4. ALGORITHM

This section discusses about the work done by the proposed system.

Algorithm 1 Query_Input(q_{i+1})

Input : user query q_{i+1}

begin

$qlog_{i+1} \leftarrow qlog_i \cup q_{i+1}$

for each new query q_{i+1}

Association_Rule_Mining ($qlog_{i+1}$) // will return rules

$KB \leftarrow rules$

RecommendationEngine(q_{i+1});

end for

end

Query_Suggestions($q_{i+1} \wedge C_1/C_2/ C_3 | \dots | C_k$)

End

Algorithm 3 UserInteractiveQuery ($q_{i+1} \cup C_m$) // where $C_m = C_1, C_2, C_3, \dots, C_k$

begin

for each query $q_{i+1} \cup C_m$

$qeByAsso \leftarrow q_{i+1} \cup C_m + \text{"Extracted_Rules"}$

$qeByWordNet \leftarrow q_{i+1} \cup C_m + t_p$ // where $p = \{1, 2, \dots, n\}$
terms from wordnet

Search ($qeByAsso$)

Search ($qeByWordNet$)

end for

end

5. SIMULATED RESULTS

Results of prevailing search engines like Google, Yahoo!, Altavista, Flickr and Picsearch for user query “blackberry” are shown in Figure 7. For a user who is searching for “blackberry” in context of fruit the result seem to be irrelevant and act as a noise. Since user generally does not formulate query along with context, therefore efforts are made in this direction in this work.



Figure 7. Prevailing Systems Results

The proposed system is effective as it will return most relevant results as per the user query. As it is an interactive process and user feedback the context of the query been fired by them. For instance for the same example where user wants to search for blackberry in context of fruit; user feedbacks the context prior before the commencement of the actual search process. Moreover query is expanded based on rule set and wordNet dictionary thus increasing relevancy. Expected results of the proposed system are shown in Figure 8 below.

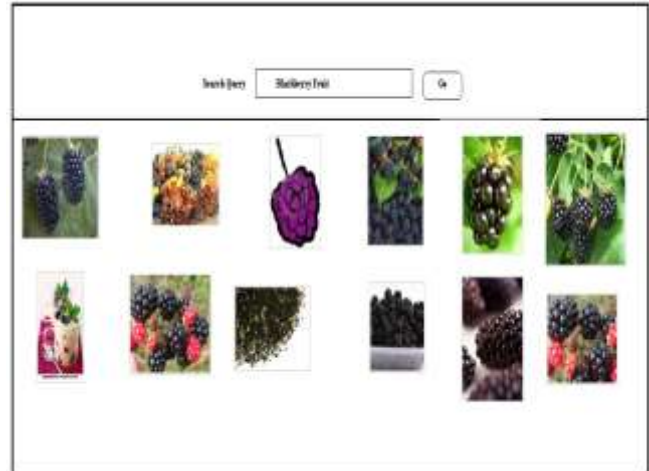


Figure 8. Proposed System Results

Comparison of prevailing search engines and expected results of the proposed system is shown by Graph in Figure 9.

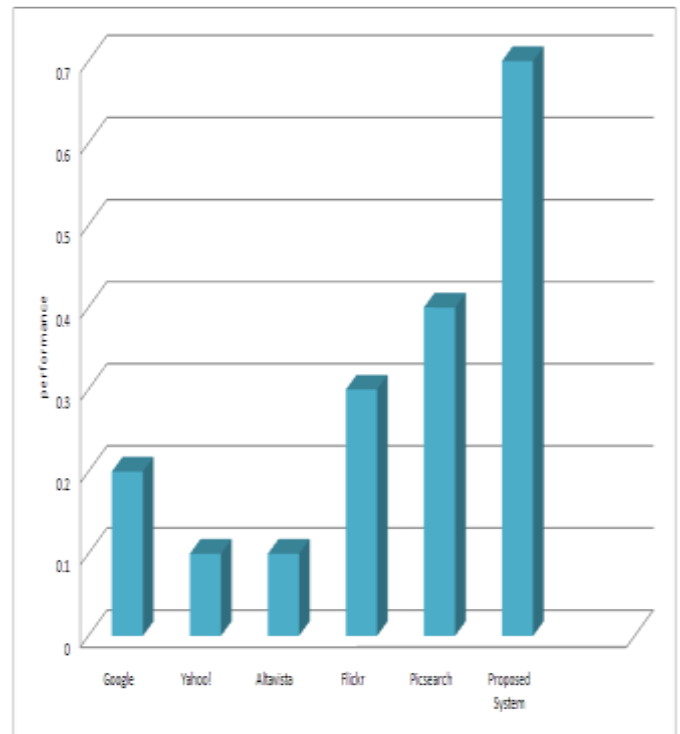


Figure 9. Comparison of Prevailing and Proposed System

6. CONCLUSION

This work proposes a novel approach for better Image Retrieval by query expansion process, overcoming mismatch in query and document vocabularies by addition of terms with similar statistical properties to those in the original query. The query is expanded by the terms from WordNet as well as from previous queries. Moreover this work will increase the relevancy of the returned results, as user feedbacks the context of the query been fired prior to the actual search done by the searcher.

7. REFERENCES

- [1] <http://www.worldwidewebsite.com/index.php?lang=NL>
- [2] Liu, S., Liu, F., Yu, C., and Meng, W. (2004) An Effective Approach to Document Retrieval via Utilizing WordNet and Recognizing Phrases, *Proceedings of the 27th annual international Conference on Research and development in Information Retrieval*: 266-272.
- [3] Min Song , Il-Yeol Song , Xiaohua Hu, Robert B. Allen, “Integration of Association Rules and Ontology for Semantic-based Query Expansion”, MIT Press, 2006
- [4] S. Brin and L. Page, “The Anatomy of a Large-Scale Hypertextual Web Search Engine” *Computer Networks and ISDN Systems* Vol.30, issue 1-7, Pages: 107-117, (April-1998).
- [5] Pahal N, Gulati P, Gupta P, “Ontology Driven Conjunctive Query Expansion based on Mining User Logs”, International Conference on Methods and Models in Computer Science, pp. 224-228 Dec 2009.
- [6] C. Buckley, G. Salton, J. Allan, and A. Singhal. Automatic query expansion using SMART: Trec-3 report. In *Proceedings of the Third Text REtrieval Conference*, pages 69–80, 1995.
- [7] R. Baeza- Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.
- [8] B. Billerbeck, F. Scholer, H. E. Williams, and J. Zobel. Query expansion using associated queries. pages 2–9, New Orleans, USA, 2003.
- [9] Davies, J., Fensel, D., and van Harmelen, F., *Towards the Semantic Web: Ontology-Driven Knowledge Management*, John Wiley and Sons, Ltd., 2002.