Journal of Computer Science 7 (4): 548-553, 2011
ISSN 1549-3636

Context Disambiguation Based Semantic Web Search for Effective Information Retrieval

[1]M. Barathi and [2]S.Valli
[1]SMK Fomra Institute of Technology, Kelambakkam,
Chennai, 603103, India
[2]Department of Computer Science and Engineering,
Anna University, Chennai, 25, India

Abstract: Problem statement: Search queries are short and ambiguous and are insufficient for specifying precise user needs. To overcome this problem, some search engines suggest terms that are semantically related to the submitted queries, so that users can choose from the suggestions based on their information needs. **Approach:** In this study, we introduce an effective approach that captures the user's specific context by using the WordNet based semantic relatedness measure and the measures of joint keyword occurrences in the web page. **Results:** The context of the user query is identified and formulated. The user query is enriched to get more relevant web pages that the user needs. **Conclusion:** Experimental results show that our approach has better precision and recall than the existing methods.

Key words: Context disambiguation, information retrieval, WordNet based semantic, similarity measures, K-core algorithm, semantic similarity, search engine

INTRODUCTION

As the web keeps expanding, the number of pages indexed in a search engine also increases. With such a large volume of data, finding relevant information is a difficult task. Queries submitted to a search engine tend to be short and ambiguous. The average query length on a popular search engine was only 2.35 terms. These short queries do not express precisely what the user really needs (Jansen *et al*., 1998). As a result, lots of pages retrieved may be irrelevant and the users need to reformulate their queries using more search terms. To improve the user's search experience, major commercial search engines provide query suggestions to help users formulate more effective queries. When a user submits a query, a list of terms that are semantically related to the submitted query is provided to help the user identify terms that the user wants so as to improve the retrieval effectiveness. Yahoo's "Also Try" (www.yahoo.com) and Google's "Search related to provide related queries for narrowing the search, while "Ask Jeeves" (www.ask.com) suggests both more specific and more general queries to the user. Unfortunately, these systems provide the same suggestions for the same query without considering the user's specific interest. Assume that the user is trying to find out information about apple computers. When the user query is given as apple in Google's search engine it gives suggestion like 'apple fruit", "apple iPod", "apple history", "apple pictures". But in a semantically enhanced web search, the system would consult the semantically indexed cluster and choose the correct one. This narrows the user's search and more relevant information is returned to the user.

SWSIR uses a Lucene indexer to index the web page collection. (http://lucene.apache.org). By measuring the Joint keyword occurrence of web pages and the WordNet based semantic related measure; the user specific contexts are identified and disambiguated according to their needs. The motivation of our research is that queries submitted to a search engine may have multiple meanings. The ambiguous queries would be disambiguated and help us to focus the web search according to their needs. For example, the query apple may refer to a fruit, the company apple computer or the name of a person and so forth. Users may be interested in "apple" as a fruit or "apple" as a company. The ambiguous query apple would be disambiguated according to the user's context and help the users to formulate more effective queries according to their needs

SWSIR follows 4 major steps. First, a set of web pages is retrieved from the web and indexed using Lucene. Stop words are removed and keywords are stored for efficient retrieval of the web pages. Secondly,

Corresponding Author: M. Barathi, SMK Fomra Institute of Technology, Kelambakkam, Chennai, India Tel: 91+ 9940323813

this approach uses the Leacock-Chodorow Measure (lch), since noun words are more suitable. So, noun words are extracted and frequency of occurrences of each noun is calculated. Thirdly K-cores are generated using the K-core algorithm (Ramirez and Brena, 2006). Finally, the given user's query is disambiguated and the user's specific context is identified by using WordNet ontology (wordnet.princeton.edu) and WordNet semantic similarity measures (Leacock and Chodorow, 1998). Next, the user's query is enriched and passed to the web searcher to narrow the web search to get more relevant web pages.

Related work: The effective use of context information in computing applications still remains an open and challenging problem. Several researchers have tried over the years to categorize context-aware applications and features, including contextual searching, adaptation, resource discovery and augmentation (Pascoe, 1997; Schilit *et al*., 1994). Our work is concerned with exploiting context disambiguation information retrieval by using the WordNet. Different subsets of the user's interest are used at runtime to discard the out-of-context preference automatically (Valet *et al*., 2007). The approach (Liu *et al*., 2004) learns the user's profile from the search histories and constructs a general profile based on the Open Directory Project category hierarchy as the context for each query, whereas, the SWSIR uses similarity measures in finding the context.

A lot of research in meta search and distributed retrieval investigates mapping user queries to a set of categories or collections (Dolin *et al*., 1998; Fuhr, 1999; Gauch *et al*., 1996; Gravano and Garcia-Molina, 1995; Howe and Dreilinger, 1997; Powell *et al*., 2003; Xu and Croft, 1999; Yu *et al*., 2001). However, all of the above techniques return the same results for a given query regardless of the submitted query. Our work uses the WordNet and the standard web search algorithm to identify the intention of the user. The Context model for a user is represented as an instance of a reference domain ontology in which concepts are annotated by interest scores, derived and updated implicitly, based on the user's information access behavior (Sieg and Mobasher, 2007).

Another way of improving the search result is by means of personalized search using ontology. An ontology model for personalization is built by considering the user information, interest, preference and other internet profile (Golemati *et al*., 2007). This approach needs to collect and preserve different information and predicting quick user interest change is difficult in this approach.

The approach (Herrera *et al*., 2010) describes the impact of using several features extracted from the document collection and query logs for automatically identifying the user's goal behind their queries. This approach classifies the query into classes such as navigational, informational and transactional. For an ambiguous query, this approach fails to classify the query. SWSIR disambiguates user's queries and identifies the user's goal without using the user profile. A Keyword is used to find the similarity among word pairs (Barathi and Valli, 2010). Some keywords may be adjectives or adverbs. Since there is no organized IS-A hierarchies for adjectives or adverbs, similarity measures cannot be applied .The SWSIR uses noun pairs to find the relatedness between noun words, since the synsets for nouns are more structured.

Another way of improving web search result is by query reformulation. The approach (Bouramoul *et al*., 2010) learns user's context according to user's profile for query reformulation, whereas, the SWSIR uses similarity measures in finding the context. The approach (Thangamani and Thangaraj, 2010) uses K-mean clustering algorithm and Feature selection for grouping text document, whereas, the SWSIR uses joint frequency of any keywords set of any size.The approach uses Association rule Mining Based on vector and Matrix algorithm to mine the relationships between feature words for query expansion in Web information retrieval, whereas, the SWSIR uses semantic similarity measures to determine the relationship between words.

MATERIALS AND METHODS

Measure of semantic similarity: WordNet: Similarity is a freely available software package that makes it possible to measure the semantic similarity or relatedness between a pair of concepts. The Leacock and Chodorow (1998) Measure is used to measure the relatedness between a pair of concepts. This is based on simple edge counts in the IS-A hierarchy of the WordNet. It deals only with nouns. The path length is calculated as given by Eq. 1:

$$\text{Relatedness } (c_1, c_2) = -\log (\text{path\_length}/2D) \quad (1)$$

Where:
c_1 and c_2 = The concepts
D = The depth of the taxonomy

The lexical database: WordNet is particularly well suited for similarity measures, since it organizes nouns and verbs into hierarchies of an IS-A relation. The IS-A relations in the WordNet do not cross parts of speech boundaries. So, WordNet-based similarity measures are limited to making judgments between noun pairs, such as cat and dog and verb pairs, such as run and walk. However, concept can be related in many ways beyond being similar to each other. For example, a wheel is a

part of a car, night is the opposite of day, snow is made up of water, a knife is used to cut bread and so forth. The WordNet provides additional relations, such as has-part, is-made-of, is-an-attribute-of. The synsets for nouns are more structured than those for verbs. Since, the WordNet has limitation on noun and verb pairs, the SWSIR consider only the noun word pair.

Table 1. Semantically related cluster set for the word apple

Cluster No.	Semantically related cluster
C1	{apple, river, valley}
C2	{apple, fruit, tree}
C3	{apple, billy, artist}
C4	{apple, laptop, ipod}

Fig. 1: Semantically related cluster (K-core) for the word bat

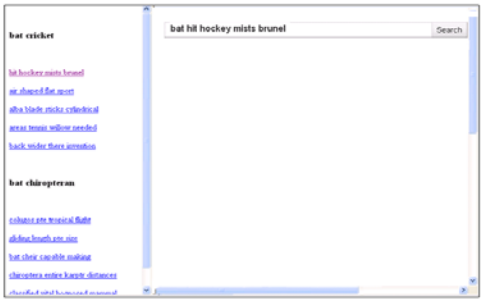

Fig. 2: User context refined query

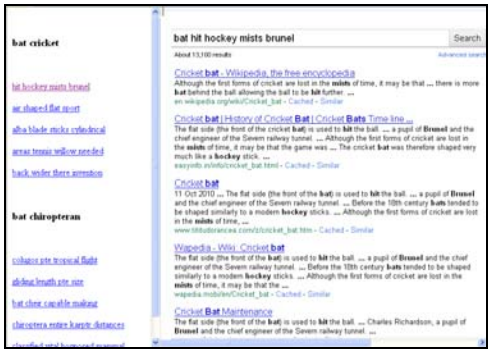

Fig.3: User context relevant web pages

Semantic cluster generation: A semantic cluster (K-core) is a kind of noun word cluster. A semantic cluster is a cluster of interrelated terms in the sense that they appear together in a number of web pages. The joint frequency for each noun word is calculated and the forces for the set of k noun words are calculated for the generation of a set of k-cores. Here, k is the size of the cluster. For example if k = 3, a cluster consists of 3 related words as shown in Table1.

The generation of the k-core depends mainly on the noun word frequencies and maximal force. The "force" of a noun word set $\{_{w1,.....wk}\}$ is given by Eq. 2:

$$f(\{w1,....,wk\}) = \frac{cj(\{w1,....,wk\})}{g(J(wi))} \quad (2)$$

Where:

$_{w1....\ wk}$ = Represents the set of k noun words,
$J(_{w1...wk})$ = The joint count of noun words
$g(J(_{wi}))$ = Joint count of individual noun words
c = Constant and in this implementation, it is 10^{12}

Context based semantic information retrieval: The SWSIR system uses Lucene to create an index file and a table of keyword frequencies. In order to consider only the keywords, stop words and other unwanted words are removed .Then a semantically related cluster (K-core) is generated using the K-core algorithm (Ramirez and Brena, 2006) and it is shown in Fig. 1. The similarity measure is calculated for the user query Q and the available K-cores K. If the similarity value is greater than the threshold value and non zero, a match exists and a new query is constructed by augmenting the user query with the selected K-cores to enrich the user queries with the words from the K-core list. This identifies the context and narrows the search. If the similarity value is less than or equal to the threshold value and non zero, then the user has to choose one of the K-cores from the K-core list, since the query is a highly ambiguous word. This selected core is augmented with the user query and passed to the web searcher to retrieve the results of the enriched queries. For instance, if the user query Q is given as {bat} the similarity measure sim (Q, K) is calculated using Eq. 1.

Algoritlun

1. Calculate the force for the set of K Keywords
2. Maximal force of top n K-cores are selected as K-core list for each sense
3. Calculate the similarity measure between the user query and the set of K-cores
4. If the similarity value is greater than threshold value and non zero, a match exists and the user context is identified.(for ambiguous word) then
5. Construct a new query by augmenting the user query and the selected K-core and passed to the web searcher
6. Go to step 13
7. else
8. If the similarity value is less than or equal to the threshold value and non zero, then
9. The user selects one of the K-cores from the K-core List
10. Construct a new query by augmenting the user query and the selected K-core
11. Pass this newly constructed query to the web searcher to retrieve the relevant web pages and go to step 13
12. If the similarity value is zero, a match does not exist and the actual query is passed to the web searcher (for unambiguous word)
13. End.
14. The user receives the result of the expanded queries.

Fig. 4: Algorithm for user context refined query

Since the query bat is a highly ambiguous word and less than the threshold value, the user has to select core according to their needs. The selected core is augmented with the user query and passed to the web searcher to narrow the search as shown in Fig. 2. The refined query is sent to the web searcher to retrieve more relevant pages as shown in Fig. 3. The algorithm for user context refined query is given in Fig. 4. Current web search just shows some words as suggestion but does not refine the user query with semantically related words of that theme. So, the SWSIR is more efficient than the existing method for the retrieval of user context relevant pages.

RESULTS

To evaluate the SWSIR, the Lucene indexer is implemented to index the web page collections. Around 1500 web pages are collected for each topic. For the purpose of indexing, each web page is read and then the pages are parsed to get the required information to be indexed. The HTML Parser tool is used to parse the html document into fields. It parses the html page and gives the various fields like title, content, summary, URL and so on. Each document stores the following fields as Document (title, content, summary, modified, URL).

The HTML parser is used for extracting the text portion of each web page. The stop words and other unwanted words are removed; the words (noun) are stemmed and indexed into the Lucene index. Then the frequency count for each noun term is calculated and the term frequency inverted document frequency (tfidf) for the terms are calculated (Baeza-Yates and Ribeiro-Neto, 2001). K-cores are generated from the set of web page collections and the best set of top n cores is considered for each theme. The similarity measure is calculated for the user query Q and the available K-cores K to retrieve more user context relevant results.

The effectiveness of an Information retrieval system is evaluated using precision and recall as given by Eq. 3 and 4. The precision measures the exactness of the search (i.e.), the percentage accuracy of the retrieved documents. The recall measures the completeness of the search (i.e.), the percentage of the relevant documents retrieved:

$$\text{Precision} = \frac{\text{Retrieved relevant documents}}{\text{Retrieved documents}} \quad (3)$$

$$\text{Recall} = \frac{\text{Retrieved relevant documents}}{\text{All relevant documents}} \quad (4)$$

The SWSIR was tested for each query against the web page collection and also some famous search engine. The search query results using SWSIR and without using SWSIR are shown in Table 2. The maximum retrieved pages for each query is shown in Table 3. The comparison of the precision graph for the unrefined and refined query results is shown in Fig. 5. This shows a significant improvement over the context of the result than the existing method. The recall measure of the existing and proposed method doesn't vary much between the unrefined and refined query results.

Table 2.Comparison of the Refined and Unrefined Query Results

Q.No	Ambiguous query	Unrefined results (without using SWSIR)	Context	Refined query (using SWSIR)	Refined query results (using SWSIR)
Q1	Apple	Apple inc-wikipedia, apple-wikipedia, apple-itunes	Computer	Apple+company +product	Apple inc-wikipedia, apple computers inc, apple macintosh
Q2	Bat	British american tobacco, bat-wikipedia, cricket bat	Flying mammals	Bat+mammals +chiroptera	Bat-wikipedia, flying mammals
Q3	Java	Java software, Java programming language-wikipedia, java wikipedia	Island	Java+island +history	Java-wikipedia, java island
Q4	Port	Port number wikipedia, port wikipedia	Network	Port+network +serial	Port number-Wikipedia, what is a Network port
Q5	Cancer	Cancer wikipedia, cancer horoscope, cancer medline plus	Disease	Cancer+disease +tumor	Cancer-wkipedia oncogenes and cancer cell
Q6	Jaguar	Jaguar international market, jaguar wikipedia, jaguar cars-wikipedia	Cars	Jaguar+car+ company	Jaguar international-market selector page, jaguar cars wikipedia
Q7	Pop	Pop office protocol pop wikipedia pop music wikipedia	Music	Pop+music +albuam	Pop music-wikipedia, videos for pop music
Q8	Bank	Bank wikipedia, riverbank wikipedia	River	Bank+river +flood wikipedia	Stream bed-wikipedia, riverbank-
Q9	Man	Man wikipedia, metropolitan area network	Network	Man+network +internet	Metropolitan area network-wikipedia, types of network
Q10	Duck	Duck wikipedia, duck cricket wikipedia	Cricket	Duck+cricket+ player	Duck cricket-Wikipedia duck out in cricket

Table 3: Precision measure for the refined and unrefined queries

	Precision without using SWSIR			Precision using SWSIR		
Query No.	Relevant Documents	Retrieved Documents	Precision (%)	Relevant documents	Retrieved documents	Precision (%)
Q1	42	87	48.28	66	75	88.00
Q2	18	73	24.66	51	65	78.46
Q3	31	83	37.35	68	72	94.44
Q4	12	65	18.46	44	52	84.62
Q5	27	76	35.53	48	61	78.69
Q6	15	68	22.06	52	57	91.23
Q7	10	56	17.86	38	45	84.44
Q8	19	62	30.65	41	50	82.00
Q9	23	71	32.39	46	52	88.46
Q10	19	68	27.94	44	53	83.02

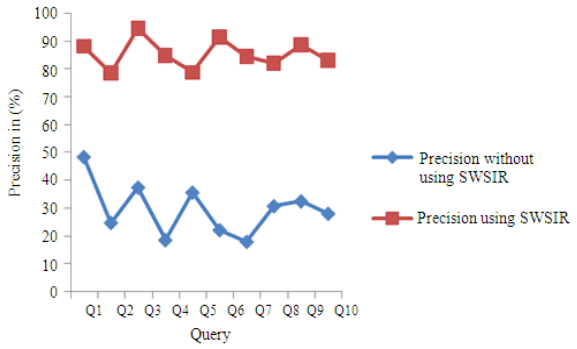

Fig. 5: Precision of unrefined queries Vs. refined queries

DISCUSSION

The comparison of the precision using SWSIR and without using SWSIR are shown in Fig. 4, Table 2 and 3. This shows a significant improvement over the context of the result than the existing method.

CONCLUSION

The SWSIR introduces an effective approach that captures the user's specific context by using the WordNet based semantic relatedness measure and the measures of joint keyword occurrences in the web page. The context of the user query is identified and formulated. The user query is enriched to get more relevant web pages that the user needs. When compared to the existing search results, this system improves results by avoiding other unrelated pages returned by the search engine. Future work would focus on improving the cluster set.

REFERENCES

Baeza-Yates, R. and B. Ribeiro-Neto, 2001. Modern Information Retrieval. 1st Edn., Machinery Industry Press, China, ISBN-10: 7111137043, pp: 513.

Barathi, M. and S. Valli, 2010. Ontology based query expansion using word sense disambiguation. Int. J. Comput. Sci. Inform. Security, 7: 022-027.

Bouramoul, A., M.K. Kholladi and B.L. Doan, 2010. PRESY: A context based query reformulation tool for information retrieval on the web. J. Comput. Sci., 6: 470-477. DOI: 10.3844/jcssp.2010.470.477

Dolin, R., D. Agrawal, A. E1 Abbadi and J. Pearlman, 1998. Using automated classification for summarizing and selecting heterogeneous information sources. D-Lib Mag. http://www.dlib.org/dlib/january98/dolin/01dolin.html

Fuhr, N., 1999. A decision-theoretic approach to database selection in networked IR. ACM Trans. Inform. Syst., 17: 229-249. DOI: 10.1145/314516.314517

Gauch, S., G. Wang and M. Gornex, 1996. Profusion: Intelligent fusion from multiple, distributed search engines. J. Univ. Comput. Sci., 2: 637-649.

Golemati, M., A. Katifori, C. Vassilakis, G. Lepouras and C. Halatsis, 2007. Creating an ontology for the user profile: Method and applications. Proceedings of the 1st IEEE International Conference on Research Challenges in Information Science, Morocco. http://sdbs.cst.uop.gr/?q=node/198

Gravano, L. and H. Garcia-Molina, 1995. Generalizing GIOSS to vector-space databases and broker hierarchies. ProceedingS of the 21st International Conference Very Large Databases, (VLDB'95), Morgan Kaufmann Publishers Inc., USA., pp: 78-89.

Herrera, M.R., E.S. de Moura, M. Cristo, T.P. Silva and A.S.D. Silva, 2010. Exploring features for the automatic identification of user goals in web search. Inform. Process. Manage., 46: 131-142. DOI: 10.1016/j.ipm.2009.09.003

Howe, A.E. and D. Dreilinger, 1997. Savvy search: A meta-search Engine that learns which search engines to query. AI Maga., 18: 19-25.

Jansen, B.J., A. Spink, J. Bateman and T. Saracevice, 1998. Real life information retrieval: A study of user queries on the web. Proc. ACM SIGIR Forum, 32: 5-17.

Leacock, C. and M. Chodorow.1998. Combining Local Context and WordNet Similarity for Word Sense Identification. In: WordNet: An Electronic Lexical Database, Fellbaum, C. (Ed.). Mit Press, pp: 265-283.

Liu, F., C. Yu and W. Meng, 2004. Personalized web search for improving retrieval effectiveness. IEEE Trans. Knowl. Data Eng., 16: 28-40. DOI: 10.1109/TKDE.2004.1264820

Pascoe, J., 1997. The stick-e note architecture: Extending the interface beyond the user. Proceedings of the 2nd International ACM Conference Intelligence User Interfaces, (ICIUI'97), ACM, New York, pp: 261-264.

Powell, A.L., J.C. Frech, J.P. Callan, M.E. Cornell and C.L. Viles, 2003. The impact of database selection on distributed searching. Proceeding of the 23rd Annual International ACM SIGIR Conference Research and Development in Information Retrieval, (SIGIR'03), ACM New York, NY, USA., pp: 232-239. DOI: 10.1145/345508.345584

Ramirez, E.H. and R.F. Brena, 2006. Semantic contexts in the internet. Proceedings of the 4th Latin American Web Congress, Oct. 25-27, Cholula, Mexico, pp: 74-81.

Schilit, S., N. Adams and R. Want, 1994. Context-aware computing applications. Proceedings of the IEEE 1st Workshop Mobile Computing System Application, Dec. 8-9, Santa Cruz, CA., pp: 85-90.

Sieg, A. and B. Mobasher and R. Burke, 2007. Ontological user profiles for representing context in web search. Proceedings of the IEEE International Conferences on Web Intelligence and Intelligent Agent Technology, (WI-IATW'07), IEEE Computer Society Washington, DC, USA., pp: 91-94.

Thangamani, M. and P. Thangaraj, 2010, Integrated clustering and feature selection scheme for text documents. J. Comput. Sci., 6: 536-541 DOI: 10.3844/jcssp.2010.536.541

Valet, D., P. Casteless, M. Fernandez and P. Mylonas, 2007. IEEE Trans. Circuits Syst. Video Technol., 17: 336-346. DOI: 10.1109/TCSVT.2007.890633

Xu, J. and W.B. Croft, 1999. Cluster-based language models for distributed retrieval. ProceedingS of the 22nd Annual International Acm SIGIR Conference Research and Development in Information Retrieval, (SIGIR'99), ACM New York, NY, USA., pp: 254-261. DOI: 10.1145/312624.312687

Yu, C., W. Meng, W. Wu and K.L. Liu, 2001. Efficient and effective metasearch for text databases incorporating linkages among documents. Proceedings of the ACM SIGMOD International Conference on Management of Data, (SIGMOD'01), ACM New York, NY, USA., pp: 187-198. DOI: 10.1145/375663.375684