An Efficient Approach to Web Query Classification using State Space Trees

[1]S.Lovelyn Rose, [2]K.R.Chandran, [3]M.Nithya
[1,2,3]Department of IT, PSG College of Technology, Coimbatore, Tamilnadu, India

Abstract
In web search engines, the retrieval of information is quite challenging due to short, ambiguous and noisy queries. This can be resolved by classifying the queries to appropriate categories. In this paper we propose a web query classification system by using a state space tree based approach which is a hierarchical arrangement of categories as states at different levels. The user given query is passed into yahoo directory search and we extract the resulting categories as features for further processing. The extracted features are mapped to the target categories using direct mapping, wordnet mapping and, glossary mapping. The frequency with which a target category term is matched when the various mapping techniques are involved is recorded at the various nodes in state space tree. Performing a best first search on the state space tree yields a ranked list of categories. This technique when compared with manual classification was found to produce a precision of 0.66.

Keywords
Web Query, Classification, Intermediate categories, State space tree

I. Introduction

Web search engines pave the way for the internet users to find information from the world of internet. Search engines are based on pre-determined algorithms which return the set of user intended web pages. Users may not be aware of the proper usage of keywords or queries to retrieve the required web result and the user intention behind the query is always a mystery to the search engine. The user query contains words of various languages, noisy terms and words of a short average length of 2.6 terms per query. The changing nature of the web language is another factor that degrades the performance of a search engine to produce sub-optimal search results. Once categorization is performed, the information embedded in the resulting categories would enhance the search engine to perform effectively in order to return more relevant web pages.

Using web search results and directory search results for web query classification is an active area of research [1-2]. The KDDCUP participants started a trend of using the search engine directory services and the categories of the open directory project to classify against the KDDCUP 2005s 67 two level target categories[1,3-4]. Shen et.al used an ensemble of classifiers. They tried classifying the queries onto the 67 target categories based on the categories returned by Google directory service and ODP. But due to the low recall achieved by synonym based mapping, SVM with a linear kernel was used to classify the web documents returned for the query [1]. But this model required the classifier to be retrained every time the target category was changed. So along with web knowledge a bridging classifier which needed to be trained only once was used as an intermediate taxonomy [2].

This paper deals with using the web knowledge, semantic knowledge and the state space tree to categorize each user query. Our approach begins with extracting textual features of the category information from web pages through the directory search result and they are considered as the intermediate categories. Direct or exact mapping of the singular and plural forms of the categories was obtained by Shen et. al to map the intermediate categories to obtain the required target category [1-2]. This resulted in low recall due to the mismatch that existed between the directory services keywords and the target category. So in our approach, the technique was expanded by using wordnet to obtain more meaningful terms among relevant categories.

Vogel et.al [3] emerged with a new classification system for the frequently used query terms to identify the subject context by using web directory. They used a taxonomy mapping component, a conjecturing component and a logistic regression model for the classification. Shen et.al [1] used wordnet to convert noisy user queries into meaningful terms by extracting words from them using wordnet and the maximum length matching method. To solve the mismatch problems when wordnet is used, we obtain synonyms from the thesaurus for the target category terms and we construct a manual glossary for the target category. In this paper we have finally combine all the three approaches and use a state space tree based concept to get the ranked target categories. That is, we obtain the mapping frequency using the various mapping techniques and populate the state space tree in order to obtain more appropriate target categories with high accuracy. This was found to enrich the search engine to capture the meaning of the queries and to produce the search results that are most likely intended by the user.

A. Problem definition and overall approach

The objective is to map a query q to an ordered set of target categories tci. Each tci is a multilevel category with the specificity increasing as we traverse down the hierarchy. In this paper, we map the intermediate categories to the target categories suggested in the KDDCUP 2005 competition. The KDDCUP 2005 competition proposed 67 two-level target categories with 7 first level generalized categories and 60 second level categories as shown in Fig.1.

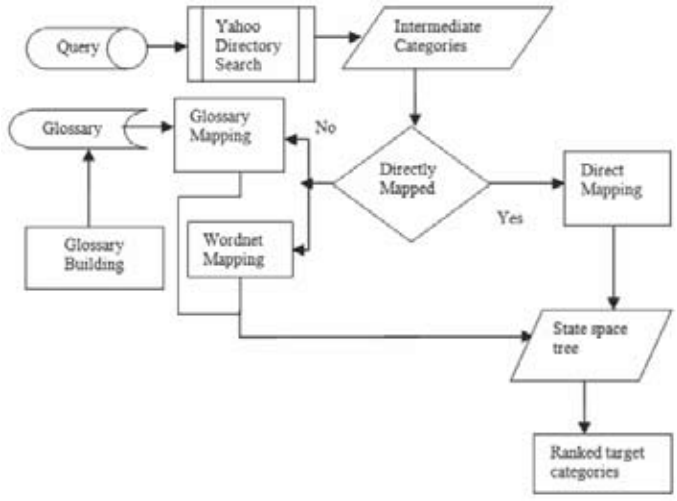

Fig. 1 : Target Category

The overview of our proposed method is as follows: The query to be classified is passed through the Yahoo directory search. The returned categories that are referred to as the intermediate categories are noted for a maximum of 50 search results. The

intermediate categories with words matching with the target category are mapped using direct mapping. The remaining words are mapped using the path length based semantic similarity measure as proposed by Wu and Palmer [5] by using wordnet and finally glossary mapping is performed using a manual glossary. The target categories are finally ranked using a state space tree based on the frequency with which the intermediate categories are mapped to the various target categories.

Algorithm

1. Pass query through yahoo directory search
2. Retrieve a minimum of 50 search results if available
3. Map the intermediate categories to the target category using

a. Direct Mapping
b. Glossary Mapping
c. Wordnet Mapping

4. Populate the state space tree with the frequency with which each node is mapped
5. Perform best first search and retrieve the ranked list of categories

II. Classification methodology

The classification methodology can be fragmented into the following phases.

A. Feature Extraction

In feature extraction, sufficient intermediate category terms are extracted as features using the results of Yahoo directory search. A major difference between text classification and query classification is the unavailability of sufficient features to train the data set [6]. The most important feature in the problem of classification is the query terms. But web queries are generally short with the mean number of terms per query being 2.6 [7]. The major task at hand is to find features to aid in the process of classification.The feature used to aid us in the classification is the categories returned by web directories. There are numerous directory search engines and they classify the web pages into multi-leveled taxonomies. These web directory search engines return web pages along with the category to which the web pages belong. All major search engines are powered by web directories because it is vital for the search engines to return the web pages and project the advertisements of the right user intended category. The KDDCUP participants using the search engine directory services and the categories of the open directory project to classify against the KDDCUP 2005s 67 two level target categories [2-4,8]. The query is therefore passed through the Yahoo directory Search to retrieve the categories of the returned web pages. These categories are referred to as the intermediate categories. Once they are retrieved they need to be transformed into the target categories.

B. Mapping intermediate categories to target categories

The features extracted in the first phase are mapped onto various target categories in this second phase.

1. Direct Mapping

In this, intermediate categories are tokenized into individual terms and string matching is performed with the target categories. The frequencies with which the strings are mapped are recorded in the states in the state space tree. When the direct mapping does not result in achieving a target category, glossary mapping and wordnet mapping are employed.

2. Glossary Mapping

Glossary based mapping is a very effective technique to map the terms which could not be mapped using direct mapping. The efficiency of the glossary based mapping is determined by the glossary building phase. The glossary building phase is carried out by creating a manual "Glossary" using the related words gathered from thesaurus, wordnet and the terms occurring in the representative categories. The intermediate categories need not necessarily be synonyms of the target categories and the singular and plural forms and abbreviation and expansions are also considered as a part of glossary. The glossary is now a rich repository of terms indicative of the target category terms. The intermediate category terms are now mapped to the target category terms used the glossary.

3. Wordnet Mapping

Wordnet is a huge lexical reference system that can be used to map the intermediate category terms which do not map directly to some target category term. The semantic similarity of the words in the intermediate and target categories forms the basis for utilizing wordnet. In this mapping technique, Wu and Palmer's semantic similarity measures are used to obtain the frequency with which the semantically similar terms in the intermediate and target categories are mapped.

Fig. 2: Automatic Web Query Classification Architecture

C. Semantic Similarity Measure

Three information content based semantic similarity measures proposed by Resnik [9], Lin [10] and Jiang and Conrath 11] and two path length based techniques proposed by Leacock and Chodorow [12] and Wu and Palmer [5] were taken and their performance was measured with the Miller –Charles and Finkelstein datasets. The tables below measure the performance of the various similarity measures.

Table 1. Comparison of Miller-Charles dataset and Six Similarity Measures

Similarity Measure	Spearman Rank Correlation Coefficient	Kendall Rank Correlation Coefficient
Adapted Lesk	0.96	0.85
Jiang and Conrath	0.87	0.75
Leacock and Chodorow	0.94	0.83
Lin	0.84	0.71
Resnik	0.90	0.76
Wu and Palmer	0.95	0.83

Table 2. Comparison of Finkelstein dataset and Six Similarity Measures

Similarity Measure	Spearman Rank Correlation Coefficient	Kendall Rank Correlation Coefficient
Adapted Lesk	0.61	0.45
Jiang and Conrath	0.40	0.29
Leacock and Chodorow	0.51	0.37
Lin	0.39	0.29
Resnik	0.5	0.37
Wu and Palmer	0.53	0.38

From the comparison table, though Adapted Lesk[13] was found to be the best technique in terms of accuracy, it was found to produce the result much slower than Wu and Palmers technique. Since Wu and Palmers technique was almost on par with Adapted Lesk in terms of accuracy, it was decided to use the Wu and Palmer technique to match intermediate category terms with the target category terms based on semantic similarity. Adapted Lesk took a longer time due to usage of definitional glosses while Wu and Palmer is a path based similarity measure.

If the relatedness between distinct concepts is measured on the basis of hypernyms and hyponyms, then it is a similarity based measure. Similarity measures are limited to nouns and verbs and since the intermediate category terms are nouns it was apt for our mapping. Path length based approaches calculate the shortest distance between the edges in the ontology and base the similarity on the is-a hierarchy. The similarity between the words increases with shorter distances. This is the basis for the Wu and Palmer technique.

D. Constructing the state space tree

A state space tree is a hierarchical arrangement of categories as states at different levels. In our technique we allow the various states in the tree structure to represent the 67 target categories as proposed in KDD Cup 2005. The state space tree is designed with the major seven categories fixed in the first level and the remaining categories are fixed in the next level of the tree structure. This forms a hierarchical arrangement of target categories represented in the state space tree as shown in Fig. 1. Each state contains the individual category along with the frequency with which the three mapping technique map the target categories, i.e. υd, υw and υg. Based on these values, we classify each user query by traversing from the root of the state space tree by using the best first branch and bound technique to obtain the target category. By using this state space tree based approach we can retrieve the subsequent five target categories.

E. Populating the state space tree

Each intermediate category terms are matched to one of the target category terms either directly or using the glossary or wordnet. The frequency with which a term maps directly to a target category is taken as υd. If they are matched using the glossary the frequency υg is updated and υw is updated for Wordnet based mapping. To disallow the creeping in of unrelated categories, the first level intermediate category is mapped to a target category using wordnet. To obtain the target category we compare the value retrieved from direct mapping to traverse in that portion of the state space tree by using best first branch and bound technique. If two states contain the same values, then we proceed with the glossary mapping frequency and the wordnet mapping frequency.

Algorithm

1. Tokenize the intermediate category into individual terms
2. If the intermediate category terms map with the target category
a. Calculate the direct mapping frequency count υd
3. Else if direct mapping does not result in achieving target category
a. Perform wordnet mapping and glossary mapping
b. Calculate their respective mapping frequency count υw and υg
4. Construct a state space tree with each node containing target categories with the three frequency counts υd, υw and υg.
5. Find max (υd) by using the best first branch and bound searching technique and obtain the first level target category.
a. If two categories have same max (υd) then find max (υw) and max (υg)
b. Obtain the first level target category comparing max (υw) and max (υg)
6. Fix the first level target category and traverse down to obtain second level category

Using the target categories achieved by our proposed technique, the precision and recall measures are calculated.

F. Best first branch and bound search

Our classification mechanism is carried out by using best first branch and bound search method by traversing each state in the state space tree using a bounding function. For each node of a state space tree, a bound is computed using the frequencies obtained from the mapping techniques. The node with the maximum value is expanded next to obtain the ranked result.

III. Experimentation and results

This paper uses the web as a source of knowledge for categorization. The experiments were carried out by using a data set which consists of queries from an AOL query log with a 500k user session collection. It consists of fields like anon id, the given query, user clicks, date and time. Of the 1012 queries, almost 16% of

the queries produced only web result and the number of noisy queries which had neither web search nor directory search result was less than 1%.

Table 3. Test Dataset

Description	Number
Original Set	1012
Noisy Queries	5
Directory Search Result	844
Only Web Search Result	168

To validate our method, the user queries are to be compared with a known database. For this purpose, the 1012 queries were given to 2 human classifiers who were asked to classify the queries into the given 67 target categories. The manual classification produced astonishing results since there was considerable difference in the classification of the 2 human classifiers. This also shows the inherent difficulty in understanding the user query. The retrieved categories were checked against the manual classifications.
The methodology was measured using precision, recall and F1 measure. The metrics can be defined as follows:
RetC = number of categories returned for a query Q
RelC = number of categories relevant for the query Q
ExpC = number of categories that should have been returned

$$\text{Precision} = \frac{\text{Re}lC}{\text{Re}tC} \quad (1)$$

$$\text{Recall} = \frac{\text{Re}lC}{ExpC} \quad (2)$$

F1 is the harmonic mean between precision and recall. Based on the above mentioned formula, the results were achieved and are as tabulated in table 4.

Table 4. Precision and recall values

Set1	Precision	Recall	F1
Manual1	0.66	0.31	0.42
Manual2	0.69	0.49	0.51

The good precision achieved stands testimony to the fact that the categories returned by the proposed technique were relevant to the query. But the low recall shows that not all relevant categories were retrieved and it might be due to the fact that the manual classifiers mapped the query to a maximum of five target categories.

V. Conclusion

Extracting the semantic information from the user query enhances the performance of search engines. With our approach, the resulting target category was found to yield a better precision and recall. These results are useful in query classification since they would help to yield the best possible user intended results in the first few pages of the search engine. In future, this approach can be extended with a user's previous queries along with the position of the extracted features in the search results to yield an optimized result.

References

[1] Shen, D., Pan, R., Sun, J., Pan, J., Wu, K., Yin, J., Yang, Q. "Query enrichment for web-query classification". ACM Transactions on Information Systems, Vol. 24, issue 3, 2006, pp. 320-352, 2006.

[2] Shen, D., Sun, J., Yang, Q. , Chen, Z., "Building bridges for web query classification". In Proceedings of the 29th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, 2006, pp. 131- 138, 2006.

[3] Vogel, D., Bickel, S., Haider, P., Schimpfky, R., Siemen, P., Bridges, S., Scheffer, T., "Classifying search engine queries using the web as background knowledge". In ACM SIGKDD Explorations, Vol. 7, issue 2, 2005, pp. 117-122, 2005.

[4] Zsolt T Kardkovacs, Tikk, D., Bansaghi, A.,"The ferrety algorithm for the KDD Cup 2005 problem", In ACM SIGKDD Explorations , Vol. 7, Issue 2, 2005, pp 111-116, 2005.

[5] Wu, Z., Palmer, M., "Verb semantics and lexical selection". In Proceeding of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, 1994, pp. 133–138, 1994.

[6] Isak Taksa, Sarah Zelikovitz, Amanda Spink. "Using Web Search Logs to Identify Query Classification Terms". Proceedings of the International Conference on Information Technology, 2007, pp. 469 – 474, 2007.

[7] Spink, A., Jansen, B. J., Wolfram, D. Saracevic, T. "From E-Sex to E-Commerce: Web Search Changes". IEEEComputer, Vol.35, No. 3, 2002, pp. 107–109, 2007.

[8] Vogel, D., Bickel, S., Haider, P., Schimpfky, R., Siemen, P., Bridges, S., Scheffer, T. "Classifying search engine queries using the web as background knowledge". In ACM SIGKDD Explorations, Vol. 7, Issue 2, 2005, pp. 117-122, 2005.

[9] P. Resnik. "Using information content to evaluate semantic similarity in a taxonomy". In Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Quebec, Canada, 1995, pp. 448-453, 1995.

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

[11] J. J. Jiang, D. W. Conrath. "Semantic similarity based on corpus statistics and lexical taxonomy". In Proceedings of the International Conference on Research in Computational Linguistics, Taiwan, 1997.

[12] C. Leacock, M. Chodorow. "Combining local context and wordnet similarity for word sense identification". In WordNet: An electronic lexical database, MIT Press, 1998, pp. 265–283.

[13] Satanjeev Banerjee, Ted Pedersen. "An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet". In Lecture Notes In Computer Science, Vol. 2276, 2002, pp. 136 – 145, 2008.

Ms. Lovelyn Rose S has a MCA and M.E degree in computer science and engineering. She has an experience of 9 years in the academic field and is currently an Assistant Professor(Senior Grade) in PSG College of Technology, India. She is a life member of the "Indian Society for Technical Education". Her main research interest is web information retrieval and has published 3 research papers in international/national conferences and 2 research papers in international journals.

Dr Chandran K R is an M.E in applied electronics and a Ph.D in electrical sciences. He has a rich industrial experience of 24 years and an academic experience of 4 years. He is currently a professor and head of the Computer and Information Sciences department of PSG College of Technology, Coimbatore. He is a fellow of the "Institute of Engineers(India)" and was nominated as a resource person for the "Asian Regional Research Program" on Energy, Environment and Climate-Phase II in 2000. He has coordinated 20 research projects funded by the Central and State Governments, Quasi Government bodies/boards and Private sectors in India

Ms.Nithya M completed her M.Tech degree in Information Technology from P.S.G College of Technology, Tamilnadu, India, in the year of 2011. Her research focuses on web query classification and key word search methods in the field of information retrieval.