Scientific Research and Essays Vol. 7(48), pp. 4118-4122, 10 December, 2012
Available online at http://www.academicjournals.org/SRE
DOI: 10.5897/SRE11.1204
ISSN 1992-2248 ©2012 Academic Journals

Full Length Research Paper

Directory knowledge, query stream and weighted state space tree based automatic web query classification

S. Lovelyn Rose* and K. R. Chandran

PSG College of Technology, Coimbatore, India.

Accepted 7 June, 2012

Optimal automatic classification of web queries aid in improving the performance of search engines. To automatically classify web queries into predefined topical categories, a novel approach which transforms a classification problem to a search problem in the state space tree is proposed. The topical categories are converted to a state space tree. The directory search results and query stream are used in determining the criterion function for a node. A final traversal through the tree using best first search yields a ranked list of target categories. Experimentation on unique users of an AOL query log with previous queries yielded a comparable result to the conventional automatic web query classification techniques.

Key words: Web query classification, directory knowledge, query stream, weighted state space tree, best first search.

INTRODUCTION

Web search engines mostly pave the way for the WWW users' retrieval of information. With 50.5% users inclined towards viewing only the first result page, designing efficient information retrieval techniques which take the relevant search result to the first result page is inexorable (Spink et al., 2002). The domino effect of the dissatisfaction with the search result is the rate at which a query is modified by the user which is as high as 44.6% (Spink et al., 2002). As concluded by Kowalczyk et al. (2004), an improved topical classification of the web query would aid in the efficient retrieval of information.

The problem is to classify a query q_i to a set of target categories tc_1, tc_2, … tc_n where each tc_i denotes the possible topical category of q_i. The ordering of the tc_i's quantify the amount of relatedness to the query with tc_i being more related to q_i than tc_j, when $i<j$.

In this proposed methodology, the objective is to map a given user web query to a ranked list of 67 target categories as proposed in KDD Cup 2005 (Shen et al., 2006). The underlying principle behind the approach is that the target categories are hierarchical structures which can be modeled as state space trees with each sub-category posing as a state. The classification problem is now transformed into a search problem in the state space tree with bounds for the states set using the query stream and directory knowledge. A subsequent best first search on the state space tree yields a ranked list of target categories.

Automatic web query classification is characterized by hurdles in the form of a mean query length of 2.6 (Spink et al., 2002), polysemy, a large web vocabulary with the English vocabulary as a miniscule subset and time-variant meanings of terms (Shen et al., 2006). The spatial-temporal nature of the web queries exacerbates the problem.

Generating a feature set is the focal challenge in the problem of automatic web query classification. Earlier attempts to populate the feature set mostly involved the post-retrieval features like categories returned in directory search (Shen et al., 2006a, b; Kardkovacs et al., 2005), the web documents returned in a web search (Shen et al., 2006a) and the tags associated with search results (Venkatesh et al., 2010). These augmented features have been found to improve the accuracy of the classification of the web query. The web query is passed through a directory search engine like Yahoo. Along with the search results, directory based search engines

*Corresponding author. E-mail: lovelyndavid@gmail.com. Tel: +91 97863 00365.

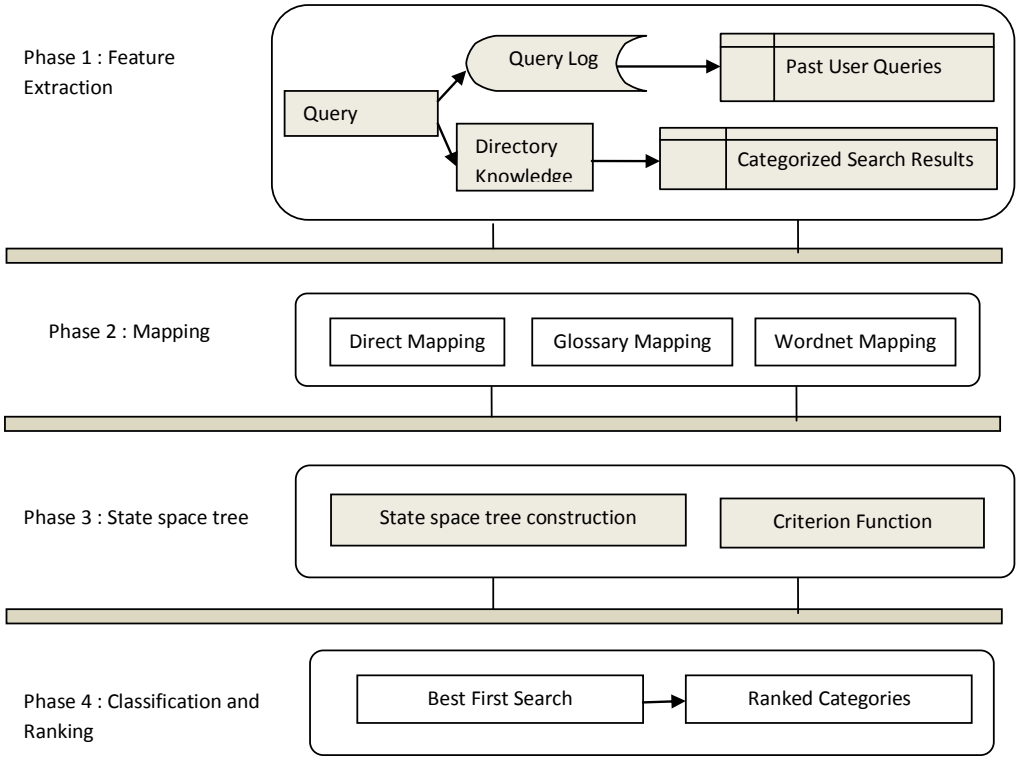

Figure 1. Web Query classification system architecture.

return category to which the search result belongs to. These categories are called intermediate categories and are used in the feature set of our classification model. Pre-retrieval features like the click stream and the web query log have been extensively studied and their influence on attaining the query class has been substantiated (Eugene and Zijian, 2006; Bernard et al., 2008; Xiaofei and Pradhuman, 2008). Beitzel et al. (2007) classified queries solely on the basis of query logs and leveraged it with machine learning and linguistic techniques and achieved recall as high as 0.55. In the research using state space trees, a combination of intermediate categories and query log details are used to allocate the cost of a node in the state space tree. Eventually a search in the state space using best first search gives the final set of ranked target categories. The objective of the research is to conceive a classification model which can be converted to a search problem and gives a comparable result to previous methods.

MATERIALS AND METHODS

The classification methodology can be fragmented into the following 4 phases.

Phase 1: Feature extraction

A major difference between text classification and query classification is the unavailability of sufficient features to train the data (Isak et al., 2007). The most valuable feature in the problem of classification is the query terms which are scant in our problem. So the first objective is to populate the feature set to augment the classification process.

For this purpose, the contents of web directories are tapped. A web directory categorizes and sub-categorizes the various websites. When a query is submitted to a directory based search engine, search results along with the category to which the search results belong to are returned. Query classification is an indispensable labor for any search engine and all major search engines are powered by web directories. Yahoo's own human-edited web directory is used to produce an intermediate category corpus $C = \{ic_1, ic_2, \ldots ic_n\}$ for a query q_i.

Another viable feature is the data rich, fine grained query log which has historical and time variant information about the user. The intermediate categories of the clicked user documents form a user click corpus $U = \{uc_1, uc_2, \ldots uc_m\}$. The user click behavior of 'n' previous queries is exploited for this purpose. The feature set is $F = C \cup U$ where $C \subseteq I$ and $U \subseteq I$ where I is the complete Yahoo taxonomy.

Phase 2: Mapping

The intermediate categories in F need to be mapped to the required target category T to accomplish the final classification.

Direct mapping

A straightforward approach is to compare the category terms by direct string matching. Given an intermediate category corpus C, the distinct category terms d, form a corpus D, where each $d \in D$.

The frequency with which an intermediate category term maps to a target category term by string matching is denoted by freq(dm) = { |d| ϵ D | string_compare(d,t) == True}.

Glossary mapping

If direct mapping fails, the terms attempt to match through a glossary. The glossary building stage is a two-step process involving an automated and a manual step. In the automated glossary building stage the glossary G is built for every term in the target category composing the singular or plural terms, abbreviations or expansions, synonyms from thesaurus and wordnet. The comparison is between the terms in the intermediate and target categories. In the second step, the glossary for a target category is manually stuffed with related intermediate category terms. The frequency with which a d ϵ D maps to a target category term t ϵ T based on the glossary is denoted by freq(gm) = { |d| ϵ D | glossary_compare(d,t) == True}.

Wordnet mapping

If the two above-mentioned mapping techniques fail, a wordnet based mapping technique is used. The remaining intermediate category terms are mapped to the lowest subcategory of the target category to which it has the maximum similarity. The freq(wm) = { |d| ϵ D | wordnet_compare(d,t) == True}. In Wu and Palmer (1994), a path length based semantic similarity measure was used to show the similarity between the intermediate and target category term. The d and t are either nouns or verbs which narrows the semantic similarity measure search to six measures of wordnet based semantic similarity. Three information content based semantic similarity measures proposed by Resnik (1995), Jiang and Conrath (1997) and Lin (1998); two path length based techniques proposed by Leacock and Chodorow (1998) and Wu et al. (1994) and gloss based Adapted Lesk (Satanjeev and Ted, 2002) were contenders. Table 1 shows the performance of the various semantic similarity measures against various benchmark datasets.

Though Adapted Lesk's performance was the best, the huge computation cost incurred, resulted in Wu and Palmer being considered to measure the semantic similarity between d ϵ D and tϵ T.

Phase 3: State space tree

Construction

A state space tree is constructed with the nodes being represented by the categories at the various levels. So the state space tree inherits the hierarchical structure of the target categories for subsequent processing.

Criterion function

Every node is associated with a criterion function based on freq(dm), freq(gm) and freq(wm) for the current and 'n' previous queries. The freq(dm), freq(gm) and freq(wm) are assigned weights by mapping them to positive real numbers.

$$f : X \rightarrow R^{+} \quad (1)$$

where X={ freq(dm), freq(gm), freq(wm)} and R is a number in the geometric sequence {a, ar, ar^2, … }. The scale factor 'a' and common ratio 'r' are assigned 0.5 because as n → ∞, the series converges to a unit value in an infinite series and we approximate our series to an infinite series for the purpose of simplification. The weight of a node η is

$$W(\eta) = f(freq(dm_c)) + f(freq(gm_c)) + f(freq(gm_c)) + \sum_{i=1}^{n}(f(freq(dm)) + f(freq(gm)) + f(freq(gm))) \quad (2)$$

where 'c' refers to the current query and 'n' is the number of previous queries. In each f(x), the position of 'x' in the geometric series is the corresponding position in formula (2).

Phase 4: Classification and ranking

Best first search

The ranked classification of the categories is now formulated as a search for a best first answer node in the state space tree. The best first search algorithm begins with the maximum cost problem state and traverses down each level searching for maximum (W(η)) (Anany, 2007; Ellis, 1998). The traversal is stopped when the required number of target categories is achieved. The end result of the traversal is a ranked list of target categories T.

RESULTS

A subset of a 500K user session from the AOL query log was used as the dataset. From the query log, 50 users with a minimum of 21 queries were gleaned. Due to the unavailability of a bench mark dataset, manual classification of user queries into their relevant and most appropriate target category was performed by two users. The metrics used for our comparison purpose are the micro-averaged precision, recall and F1 measures. The metrics can be defined as follows: If RetC is the number of categories returned for a query Q, RelC is the number of categories relevant for the query Q and ExpC is the number of categories that should have been returned, then

$$\text{Precision} = \frac{RelC}{RetC} \quad (3)$$

$$\text{Recall} = \frac{RelC}{ExpC} \quad (4)$$

F1 is the harmonic mean between precision and recall. The analysis between the manual classifiers is given in Table 2. Since the training set involves manually labeled queries, a k-fold cross validation method is used. The performance of the proposed methodology with varying number of previous queries is tabulated in Table 3.

DISCUSSION

While Table 1 indicates the inherent difficulty in finding the user intent, Table 2 shows the improved performance

Table 1. Spearman Rank Correlation co-efficient.

Similarity measure	Miller-Charles dataset	Finkelstein dataset
Adapted Lesk	0.9569	0.6079
Jiang and Conrath	0.8745	0.4024
Leacock and Chodorow	0.9420	0.5094
Lin	0.8438	0.3925
Resnik	0.9028	0.4993
Wu and Palmer	0.9474	0.5276

Table 2. Performance of the manual classifiers.

Set1	Set2	Precision	Recall	F1
Manual1	Manual2	0.3056	0.4876	0.3757
Manual2	Manual1	0.3099	0.4381	0.3630

Table 3. Performance of the proposed methodology.

Set1	Set2	Precision	Recall	F1
Proposed Methodology with no previous query	Manual 1	0.4400	0.3780	0.4066
Proposed Methodology with no previous query	Manual 2	0.4233	0.3040	0.3538
Proposed Methodology with 5 previous queries	Manual 1	0.4468	0.3772	0.4090
Proposed Methodology with 5 previous queries	Manual 2	0.4431	0.3550	0.3941
Proposed Methodology with 10 previous queries	Manual 1	0.6032	0.5609	0.5811
Proposed Methodology with 10 previous queries	Manual 2	0.5086	0.4665	0.4864
Proposed Methodology with 15 previous queries	Manual 1	0.4983	0.3934	0.4396
Proposed Methodology with 15 previous queries	Manual 2	0.4459	0.3543	0.3948
Proposed Methodology with 20 previous queries	Manual 1	0.3890	0.3457	0.3660
Proposed Methodology with 20 previous queries	Manual 2	0.3676	0.3465	0.3567

of the proposed methodology over the previous methodologies when at least 10 previous queries are considered. The manual classifiers have returned only the top 5 target categories for a query and this reduces the recall when compared to precision. When the number of previous queries is reduced, there is no sufficient distinguishing feature and this does not give an improved performance over a methodology which does not use the query stream. There is a straying of the user intent from the current query when the number of previous queries becomes large and so the performance dips. The advantage of storing the target categories as a tree is easy maintenance of the previous state and only a slight modification needs to be done in the criterion function for the next user query. A best first search on the state space tree results in a rapid sequencing of the target categories.

Conclusion

Despite astounding improvements in search engine technology, impediments continue. The proposed methodology involves the directory search result and previous query user click behavior as criterion function on a state space tree traversed using best first search. While the state space helps to retain the context of a user for subsequent processing, the results reiterate the importance of judiciously using the query stream. A competitive system that incorporates the changing behavior of the user is built and the results reaching an F1 measure of approximately 0.6 is a substantial IMPROVEMENT over the previous research performance of around 0.5. The potential areas for future work include creating an all pervading benchmark dataset for better comparison between the different methodologies and including web search results in the feature set.

REFERENCES

Anany L (2007). Introduction to the design and analysis of algorithms. Addison-Welsey Publishers.

Beitzel SM, Jensen EC, Lewis DD, Chowdhury A, Frieder O (2007). Automatic classification of web queries using very large unlabeled

query logs. ACM Trans. Inf. Syst. 25(2), Article 9.
Bernard J, Jansen L, Booth S (2008). Determining the informational, navigational, and transactional intent of web queries. Inf. Process. Manage. 44(3):1251-1266.
Horowitz E, Sahni S, Rajasekaran S (1998). Computer Algorithms. Galgotia Publications.
Eugene A, Zijian Z (2006). Identifying “best bet" web search results by mining past user behaviour. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 902-908.
Isak T, Sarah Z, Amanda S (2007). Using web search logs to identify query classification terms. Proceedings of the International Conference on Information Technology pp. 469-474.
Jiang J, Conrath D (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In Processing of International Conference Res. Computational Linguistics (ROCLING X), Taiwan pp. 19-33.
Kowalczyk P, Zukerman I, Niemann M (2004). Analyzing the effect of query class on document retrieval performance. In 17th Australian similarity for word sense identification. In Fellbaum, C., ed., WordNet: An electronic lexical database. MIT Press pp. 265-283.
Leacock C, Chodorow M (1998). Combining local context and WordNet similarity for word sense identification. In Fellbaum, C., ed., WordNet: An electronic lexical database. MIT Press pp 265–283.
Lin D (1998). An Information - Theoretic Definition of Similarity. In Proceedings of International Conference on Machine Learning. Madison, Wisconsin pp. 296–304.
Resnik P (1995). Using information content to evaluate semantic similarity in a taxonomy. In Proceedings of the 14th International Joint Conference on Artificial Intelligence pp. 448-453.
Satanjeev B, Ted P (2002). An adapted Lesk algorithm for word sense disambiguation using WordNet. In Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing pp. 136-145.
Shen D, Pan R, Sun J, Pan J, Wu K, Yin J, Yang Q (2006a). Query enrichment for web-query classification. ACM Trans. Inf. Syst. 24:320-352.
Shen D, Sun J, Yang Q, Chen Z (2006b). Building bridges for web query classification. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research Development in Information Retrieval pp. 131-138.
Spink A, Jansen BJ, Wolfram D, Saracevic T (2002). From E-Sex to E-Commerce: Web Search Changes. IEEE Comput. 35(3):107-109.
Venkatesh G, Arnd K, Xiao Li (2010). Precomputing Search Features for Fast and Accurate Query Classification. In Proceedings of WSDM pp. 61-70.
Xiaofei H, Pradhuman J (2008). Regularized query classification using search click information. J. Patt. Recogn. Soc. pp. 2283-2288.
Wu Z and Palmer M (1994). Verb semantics and lexical selection. In 32nd Annual Meeting of the Association for Computational Linguistics pp. 133–138.
Zsolt T, Kardkovacs D, Bansaghi A (2005). The ferrety algorithm for the KDD Cup 2005 problem. SIGKDD Explor. Newslett. 7(2):111-116.