Journal of Computer Science 8 (5): 804-808, 2012
ISSN 1549-3636
© 2012 Science Publications

Normalized Web Distance Based Web Query Classification

Lovelyn Rose, S. and K.R. Chandran
Department of C and IS,
PSG College of Technology, Coimbatore, India

Abstract: Problem statement: The problem is to classify a given web query to a set of 67 target categories. The target categories are ranked based on the degree of similarity to a given query. **Approach:** The feature set is the set of intermediate categories retrieved from a directory search engine for a given query. Using direct mapping and Normalized Web Distance (NWD) the intermediate categories are mapped to the required target categories. The categories are then ranked based on three parameters of the intermediate categories namely, position, frequency and a combination of frequency and position. **Results:** The results proved that the third parameter gave a better result and a maximum of 40 search result pages ensure better results. **Conclusion:** With NWD as the similarity measure, the precision and recall is found to increase by 10% over the previous methods.

Key words: Automatic web query classification, directory search, query log, NWD

INTRODUCTION

In the context of the World Wide Web, the user information need is usually translated into queries which are submitted to the search engine. The search engine processes the queries and returns a ranked list of documents which it finds as being appropriate to the query. Spink *et al*. (2001) showed that the rate of query modification was as high as 44.6%, thereby indicating the dissatisfaction of the user with the results returned. The purpose of this study is to augment the search result by predicting the right category to which the query falls under. As shown by Yamin and Ramayah (2011), this improves user satisfaction with the search result.

The problem can be formally stated as follows:

Classify a query q_i to a set of target categories tc_1, tc_2,…tc_n. The tc_i are ranked in accordance with the relevance of the query to the category. This implies that the similarity of the query to the target category tc_i is more than the similarity to tc_j, where $i<j$.

Automatic classification of web queries is restrained by the inherent nature of the web queries. Web queries are generally short with the mean query length being 2.6 (Spink *et al*., 2001). The web vocabulary increases at a rapid rate and the meanings of the query terms evolve over time. The problem worsens due to the polysemous nature of the queries (Shen *et al*., 2006a). Researchers have extensively used post-retrieval techniques with the feature set including the directory and web search results (Shen *et al*., 2006b). While query logs help to mine the trends at a point of time and the characteristics of the search engine user and the query, they are also extensively used in the prediction of the future need of the user. The user history and click behaviour are extensively used to learn the information need (Liu *et al*., 2006; Lu *et al*., 2006; Li *et al*., 2008; He and Jhala, 2008). Beitzel *et al*. (2007) worked on query logs and identified user intent with the help of selectional preferences. He also concluded that exact match and n-gram matching yielded better precision for popular queries.

MATERIALS AND METHODS

The process of query classification encompasses the following steps:

- Construct the feature set using directory knowledge
- Map the intermediate categories to the target categories using direct matching and NWD
- Assign weights to the target categories based on the following parameters: position, frequency and a combination of position and frequency
- Rank the target categories based on the weights

Feature set construction using directory knowledge: The query by itself has very few index terms and is highly insufficient for the purpose of classifying the

Corresponding Author: Lovelyn Rose, S., Department of C and IS, PSG College of Technology, Coimbatore, India
Tel: +91 97863 00365

query. To augment the feature set, the query is passed through directory search engines and the returned directory search results are used as features. The returned categories are termed intermediate categories to differentiate it from the actual target categories. The intermediate categories may be of varying depth and may or may not have an exact match to the category terms used in the target categories. The top 50 search result pages are considered for the research. The position at which the intermediate category occurs in the search result is saved along with the frequency with which an intermediate category occurs. While the position and frequency are good indicators of the relevance of the category, it was also decided to consider a third attribute combining the position and frequency.

Mapping intermediate categories to target categories.

The following steps are performed to map the intermediate categories to the target categories:

Step 1: Remove the lower most categories in the intermediate category
Step 2: Convert the intermediate and target categories into a bag of words
Step 3: Perform direct mapping
Step 4: If the intermediate category does not match in step 3 and has not been previously mapped using NWD, perform NWD based mapping
Step 5: Perform the above steps for ‘n’ previous queries

Direct mapping: The number of words in a target category is less than the number of words in an intermediate category. A unigram map is performed between the terms in the target category and the terms in the intermediate category. If there is a perfect match between the terms, then the intermediate category is mapped to that target category.

NWD based mapping: NWD is an acronym for Normalised Web distance. It was originally named NGD (Normalised Google Distance) to denote the usage of the Google search engine. NGD is a technique to find the semantic relatedness between two words with the help of a search engine and a database (Cilibrasi *et al*., 2007; 2009). The other measures of semantic relatedness that were considered were LSI and Wordnet based similarity measures (Deerwester *et al*., 1990; Pedersen *et al*., 2004). LSI is a good option for multi-word similarity checks, but due to the inherent complexity in the technique, LSI was not considered. Wordnet based similarity measures were considered next and they were found to exclude numerous words from the web vocabulary. So NGD, a similarity measure which encompassed the web vocabulary was used.

NGD is a similarity technique that uses Google, the search engine and the WWW, the database. The similarity measure is built on the normalized information distance and normalized compression distance which are based on the Kolmogorov complexity. The extent of relation between two strings q_1 and q_2 is quantified using the page count for the strings when passed through a search engine individually and as a concatenated single.

Let $SRC(q_i)$ be the number of search results (Search Result Count) returned when the web query q_i is passed through the Google search engine. $SRC(q_j)$ and $SRC(q_i,q_j)$ can be defined in a similar manner where $SRC(q_i,q_j)$ passes the concatenated string q_i. q_j as the web query. Let n be the total number of web pages indexed by the search engine, then the NGD is defined by Eq. 1:

$$NGD(q_i,q_j)=\frac{\max(\log SRC(q_j))-\log SRC(q_i,q_j)}{\log n-\min(\log SRC(q_i),SRC(q_j))} \quad (1)$$

NGD was used as a semantic similarity measure in the automatic extraction of taxonomy (Makrehchi and Kamel, 2007) and based on that Eq. 2 is used to find the similarity between the terms q_i and q_j:

$$Sim(q_i,q_j)=\frac{1}{1+NGD(q_i,q_j)} \quad (2)$$

The proposed method uses the Yahoo search engine and the WWW database. With ‘a’ as the web query, 32,500,000,000 results are returned and so ‘n’ was approximated to 3.2×10^9 for the experimentation purpose.

The intermediate categories are preprocessed before checking them with the target categories. The intermediate category is transformed to a string of words and the delineations between the hierarchies are made obsolete. The lowest two categories in the multi-hierarchy of Yahoo are too specific and they are culled. To find the semantic similarity, the intermediate category is considered as q_i and the target category is considered as q_j. The number of target categories is 67 and is a tree of depth 2. Every intermediate category is checked against all the target categories and the results are tabulated. The target category with the highest similarity to the intermediate category is chosen and the result is permanently stored. This learnt target category is used to prevent future NWD calculations for the same intermediate category. This considerably reduces the time complexity involved in the computation of the target category for a given intermediate category.

Weighing factors: Each target category is assigned a weight $w_p(tc_i)$ based on its position in the search results. The highest priority is assigned to min ($w_p(tc_i)$). When two or more tc_i occur in different positions, the first occurring tc_i is considered. This parameter is usually used by researchers (Shen *et al*., 2006a; Kardkovacs *et al*., 2005) and two more parameters were considered for weighing. The second parameter is the frequency of occurrence of the various target categories, namely $w_f(tc_i)$.

While the position and frequency are good indicators of the relevance of the category, it was also decided to consider a third attribute combining the position and frequency. The involvement of the third attribute is due to the following reason. The position of the attributes in the search result is based on the search engine's page ranking algorithm. So a bias in the page ranking algorithm would affect the ranking of the categories to a large extent. But positions are indicators to a reasonable extent. The next major attribute under consideration is the frequency of occurrence of the categories. Consider a category tc_x returned only once but in the first position. Consider another category tc_y which occurs more number of times but in lower positions. For which category should the weightage be more is an aspect to consider? So without making a trade-off between position and frequency a new measure involving both the parameters are considered. Let p_1, p_2, … p_n be the various positions occupied by target category tc_i . That is, tc_i occurs with a frequency n. Assign a high weightage α_1 to the category at the top position and reduce the weightage for the subsequent positions. That is, α_i is inversely proportional to p_i. Combining the values α_i linearly for the same tc_i's in different positions is the third attribute. This is given in Eq. 3:

$$w_{fp}\left(tc_i\right) = +\sum_{i=1}^{n} \alpha_i \tag{3}$$

Ranking: Weight is assigned to the target categories based on $w_p(tc_i)$, $w_f(tc_i)$ and $w_{fp}(tc_i)$ for the current and 'n' previous queries. The $w_p(tc_i)$, $w_f(tc_i)$ and $w_{fp}(tc_i)$ are assigned weights by mapping them to positive real numbers as given in Eq. 4:

$$f: X \rightarrow R^{+} \tag{4}$$

where, X={ $w_p(tc_i)$, $w_f(tc_i)$, $w_{fp}(tc_i)$ } 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 $n \rightarrow \infty$, 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 target category is given in Eq. 5:

$$\begin{aligned} W(\eta) &= f\left(w_p\left(tc_i\right)\right) + f\left(w_f\left(tc_i\right)\right) + f\left(w_{fp}\left(tc_i\right)\right) \\ &+ \sum_{i=1}^{n} f\left(w_p\left(tc_i\right)\right) + f\left(w_f\left(tc_i\right)\right) + f\left(w_{fp}\left(tc_i\right)\right)) \end{aligned} \tag{5}$$

where, 'c' refers to the current query and 'n' is the number of previous queries. The importance of considering the query profile helps the ranking process. In each f(x), the position of 'x' in the geometric series is the corresponding position in formula (2). Based on the weight W(η), the target categories are ranked.

RESULTS

Training dataset and test dataset: There is no available benchmark dataset to check the category into which a query falls. Due to the non-availability of a benchmark dataset, a k-fold cross-validation was performed. The KDDCUP competition held in 2005, gave 67 target categories into which the queries had to be ranked. The training and test dataset is from an AOL query log with a 500 k user session collection. It consists of 5 fields namely, anon id, the given query, date and time at which the query was submitted, the rank of the item clicked and the clicked URL. The nature of the test dataset is given in Table 1. Of the 1012 queries, 16.60079% of the queries gave only web Result and 0.49407116% was noisy queries which had neither web search nor directory search result.

To test the data, 1012 queries were given to 2 human evaluators and they were asked to classify the queries into the 67 target categories. To evaluate the manual and automated classifiers, micro-averaging of precision, recall and the F1 measure are used. 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 Eq. 6 and 7:

$$\text{Precision} = \frac{\text{RelC}}{\text{RetC}} \tag{6}$$

$$\text{Recall} = \frac{\text{RelC}}{\text{ExpC}} \tag{7}$$

Table 1: Test dataset

Description	Number
Original set	1012
Noisy queries	5
Directory search result	844
Only web search result	168

Table 2: A comparison of the manual classifiers

Set1	Set2	Precision	Recall	F1
Manual1	Manual2	0.4567	0.4279	0.4417

Table 3: Performance of the automated classifiers

Set1	Set2	Precision	Recall	F1
Manual1	NWD	0.4826	0.6215	0.5433
Manual2	NWD	0.4423	0.5883	0.5049
Manual1	Shen *et al.*	0.4222	0.5534	0.4789
Manual2	Shen *et al.*	0.4157	0.5437	0.4711

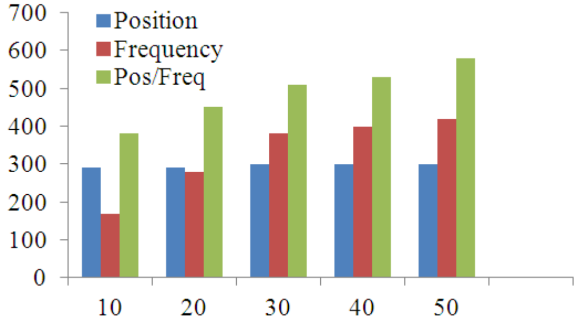

Fig. 1: Performance of the three parameters in ranking categories

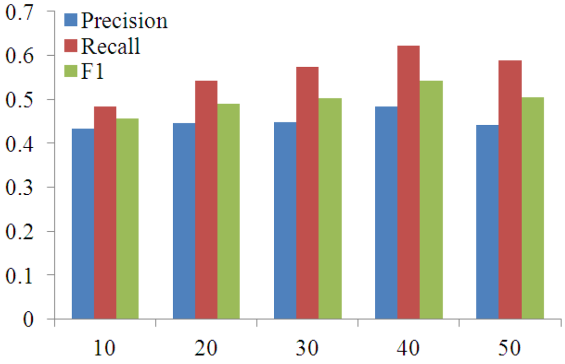

Fig. 2: Average performance of the NWD classifier for varying number of search results

F1 is the harmonic mean between precision and recall.

Based on the manual categorization, the precision and recall of each manual classifier was calculated with respect to every other manual classifier. The results obtained are as tabulated in Table 2. The low precision and recall achieved shows the inherent difficulty in analyzing the web query category. The category differs according to the human perception and so our intention is to create an automated technique which is nearer to 0.5.

Table 3 is used to compare the performance of the automated NWD methods with the manual classifiers.

Figure 1 makes an analysis of the performance of the three parameters position, frequency and position/frequency in ranking the categories. The x-axis has the number of search result pages returned and it is plotted against the number of categories that are ranked the same in the manual and automated classifiers. The third parameter which combined was found to be far above the other two parameters, thereby giving an ideal choice of parameter to consider while considering search results.

Figure 2 makes an analysis of the NWD based classifier on the basis of the above mentioned metrics. The analysis shows that the break off point is 40 search result pages and that the precision improves after 30 result pages. The recall is comparatively higher due to the fact that more the search result pages considered, more the chances of correct categories getting assigned. The relatively lower precision can be due to the difference in the interpretation of a query.

The indefinite nature of the classification can be justified by looking at the following example. In the training data supplied by KDDCUP2005, while “actress hildegarde” was mapped to Entertainment\Celebrities, Online Community\People Search, Entertainment\Movies, Information\Arts and Humanities, Information\References and Libraries, “alfred Hitchcock” was mapped to Entertainment\Movies, Entertainment\TV, Entertainment\Celebrities, Living\Book and Magazine and Entertainment\Games and Toys. Though both are names of persons, Online Community\People Search has been included only in the query “actress hildegarde” and not in the query “alfred hitchcock”.

DISCUSSION

Exact matching gives the result much faster than NWD matching, but is limited to 9% of the result. But NWD matching has the ability to work with terms which do not match directly and through NLP based techniques. The computation time for NWD is comparatively high than the word net based similarity measures due to the limitation in the internet speed. But the results are highly commendable. Also storing the results, help in reducing the number of times NWD is used. This further increases the speed of computation. But the results are better than the results achieved by Shen *et al*. (2006b) as seen in Table 3.

CONCLUSION

Search engines are updating themselves at a rapid pace and still the information need of most of the users is not met. Topical classification of web queries using directory search results is a tried and tested method. The success of the proposed technique over the previous techniques is probably due to the following two factors: 1) Using the web for similarity checking and 2) Treating the multi-termed sub-categories as a single unit. The proposed methodology can be applied to map any two given taxonomies and is robust to the changing nature of the taxonomies. In future various other similarity measures can be considered along with web search results.

REFERENCES

Beitzel, S.M., E.C. Jensen, D.D. Lewis, A. Chowdhury and O. Frieder, 2007. Automatic classification of Web queries using very large unlabeled query logs. ACM Trans. Inform. Syst. DOI: 10.1145/1229179.1229183

Cilibrasi, R.L. and P.M.B. Vitanyi, 2007. The Google similarity distance. IEEE Trans. Knowl. Data Eng., 19: 370-383. DOI: 10.1109/TKDE.2007.48

Cilibrasi, R.L. and P.M.B. Vitányi, 2009. Normalized web distance and word similarity. Computation and Language.

Deerwester, S., S.T. Dumais, G.W. Furnas, T.K. Landauer and R. Harshman, 1990. Indexing by latent semantic analysis. J. Am. Soc. Inform. Sci., 41: 391-407. DOI: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

He, X. and P. Jhala, 2008. Regularized query classification using search click information. J. Patt. Recog. Soc., 41: 2283-2288.

Kardkovacs, Z.T., D. Tikk and A. Bansaghi, 2005. The ferrety algorithm for the KDD Cup 2005 problem. ACM SIGKDD Explorat. Newslett., 7: 111-116. DOI: 10.1145/1117454.1117468

Li, X., Y.Y. Wang and A. Acero, 2008, Learning query intent from regularized click graphs. Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Jul. 20-24, ACM, Singapore, pp: 339-346. DOI: 10.1145/1390334.1390393

Liu, Y., M. Zhang and L. Ru, 2006. Automatic query type identification based on click through information. Commun. TeX Users Group, 4182: 593-600.

Lu, Y., F. Peng, X. Li and N. Ahmed, 2006. Coupling feature selection and machine learning methods for navigational query identification. Proceedings of the 15th ACM International Conference on Information and Knowledge Management, Nov. 5-11, Arlington, Virginia, USA., pp: 682-689.

Makrehchi, M.K. and S. Kamel, 2007. Automatic taxonomy extraction using Google and term dependency. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, Nov. 2-5, IEEE Xplore Press, Fremont, CA., pp: 321-325. DOI: 10.1109/WI.2007.37

Pedersen, T., S. Patwardhan and J. Michelizzi, 2004. WordNet::Similarity - measuring the relatedness of concepts. University of Minnesota.

Shen, D., J. Sun, Q. Yang and Z. Chen, 2006a. Building bridges for web query classification. Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (SIGIR '06), Seattle, Washington, USA, pp: 131-138. DOI: 10.1145/1148170.1148196

Shen, D., R. Pan, J. Sun, J. Pan and K. Wu *et al*., 2006b. Query enrichment for web-query classification. ACM Trans. Inform. Syst., 24: 320-352. DOI: 10.1145/1165774.1165776

Spink, A., B.J. Jansen, D. Wolfram and T. Saracevic, 2001. Searching the Web: The public and their queries. J. Am. Soc. Inform. Sci. Technol., 52: 226-234. DOI: 10.1002/1097-4571(2000)9999:9999<::AID-ASI1591>3.3.CO;2-I

Yamin, F.M. and T. Ramayah, 2011. The impact of user knowledge on web search satisfaction. Am. J. Econ. Bus. Admin., 3: 139-145. DOI: 10.3844/ajebasp.2011.139.145