Web knowledge and Wordnet based Automatic Web Query Classification

S. Lovelyn Rose
Department of IT, PSG College of Technology
Peelamedu, Coimbatore, Tamilnadu, India

K. R. Chandran
Department of IT, PSG College of Technology
Peelamedu, Coimbatore, Tamilnadu, India

ABSTRACT
Web search queries are the starting point to access the contents in the WWW for most of the users. Capturing the user intent behind a query statement is crucial for any search engine and is equivalent to figuring out the category to which the query belongs to. In this paper, we analyze a classification system that uses web directory search results as an extended feature of the query. A comparison with glossary based mapping showed that our work outperforms it by a reasonable margin. We also show by experimentation that choosing the right parameter for the search results gives a reasonable improvement in ranking.

Key words
Web Query, Classification, Intermediate categories, Wordnet

1. INTRODUCTION
Web queries through search engines act as the window to the world of internet for most of the internet users. The intention behind the query though trivial to the user is a mystery to the search engine. Ambiguity inherent in the words of various languages, a short average length of 2.6 terms per query and changing nature of the web language escalates the problem. So identifying the category of the web search query is the major challenge in returning the right set of web search results and in inserting the most appropriate advertisement along with the web search results.

Using web search results and directory search results for web query classification is an active area of research [1][2]. Shen et.al enriched the query by taking the categories returned by the Google directory service and ODP (Open Directory Project). The categories were then mapped to target categories through a process of extended and glossary based mapping. But due to the low recall achieved by the above mentioned 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. In a later work, along with web knowledge a bridging classifier which needed to be trained only once was used with the intermediate taxonomy [2].

This paper deals with using the web knowledge and specifically directory services to categorize queries. The main objective of this paper is to classify web search queries with directory search result and to analyze the best practices to be adopted when directory services are involved. The first hurdle to cross is the translation of the categories returned by the directory services to the required target category. Direct or exact mapping of the singular and plural forms of the categories was exploited by Shen et. al to map the directory search result categories to the required target category [1][2]. This resulted in low recall due to the fact that not all categories of the directory services share keywords in the target category. So categories were expanded using wordnet and subsequently mapped. This when experimented by us showed that the numerous possibility of unrelated words creeping in resulted in a reduction of the number of relevant categories. Vogel et.al decided to opt for a semi-automatic category mapping strategy where the categories themselves were passed as queries to the web directory search engine to get a list of recommended mappings which were manually scanned to form appropriate rules for subsequent mapping [3]. But this obviously needs a training phase, every time the intermediate category changes. So we have adopted an approach which eradicates the need for training the intermediate category every time the intermediate category changes.

Another major objective of the paper is to analyze the number of search results to consider since it is crucial in improving the precision and recall. When it comes to finding the target categories of the search result, it is also mandatory that the categories are ranked in the increasing order of relevance to the query. This necessitates the ranking of the categories for which three factors of the search results are considered. The previous researchers have given high priority to the order of the categories in the directory search result, assuming that the position of the search result denotes the degree of relevance of the query to the search result [1][2][3]. In our paper we have taken three factors namely the position, the frequency of occurrence and a combination of position and frequency to figure out that factor which gives the best result.

2. PROBLEM DEFINITION AND OVERALL APPROACH
The problem can be stated as follows:
Map a query q to an ordered set of categories tc_i where the ordering denotes the rank of the category. Each tc_i is a multilevel category with the specificity increasing as we traverse down the hierarchy.

In this paper, for the purpose of uniformity, it was decided to 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.

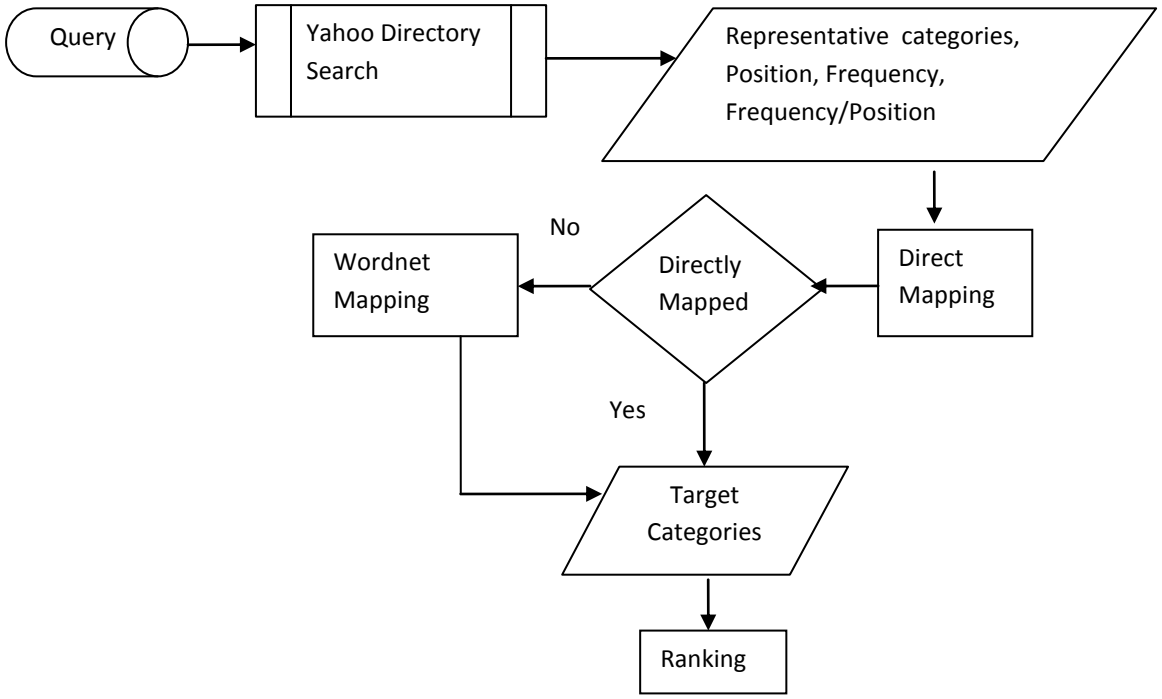

Fig 1. Automatic Web Query Classification Architecture

The following section presents a brief overview of the methodology employed for the classification of the queries which return search results with yahoo directory search. The query to be classified is passed through the Yahoo directory search. The returned categories are referred to as the intermediate categories. The position and the frequency of occurrence of 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 [4]. The target categories were finally ranked based on position, frequency and a combination of position and frequency.

Algorithm

1. Pass query through yahoo directory search
2. Retrieve a minimum of 50 search results if available
3. Record the position of the intermediate categories returned by yahoo directory search
4. Map the intermediate categories to the target category
    a. Direct and Wordnet based Mapping (given in section 3.2)
5. Rank the target categories
    a. Rank (given in section 3.3)

This procedure is summarized in figure 1.

3. CLASSIFICATION METHODOLOGY

The classification methodology adopted is discussed in detail in this section.

3.1 Feature Extraction

For any form of categorization, sufficient features are required for categorization. While the text themselves are viable features in text categorization, queries are typically very short text. So additional features extracted from external resources are inevitable. Of them query log, web search results and directory search results are favorable entities of which the best practices in directory search results are researched by us. In our paper, we have taken the categories returned by the yahoo directory search as the main feature and call them the intermediate categories. When it was decided to consider directory search results, the first choice was to consider ODP. But when the queries were passed through ODP, it was found that it returned results for far less number of queries than the popular search engines. It was then decided to consider Google which uses ODP to extract the features due to the wide popularity it enjoys [5]. But the fact that the Googleapis did not support the mining of the categories resulted in us considering the next popular search engine, namely Yahoo.

3.2 Direct and Wordnet based Mapping

The next step is to map the intermediate categories to the required target categories. Glossary based mapping [1][2] and semi-automatic mapping [3] though viable options have their own flaws as mentioned earlier. So we decided to tap Wordnet which has a vast potential in natural language processing. Earlier works have used Wordnet to expand the intermediate categories using the synonyms in Wordnet while we have used similarity measures [1][6]. The similarity of the words in the intermediate and target categories forms the basis for utilizing wordnet. Six methods were considered for the mapping and the method which correlated most to the manual ranking of the similarity between the words was chosen.

The problem at hand is to map a query q to a set of target categories $f(ic_i)$ using the intermediate categories ic_i. Here f: $ic_i \longrightarrow tc_i$ where $tc_i \in f(ic_i)$ and can be neither injective or

surjective. Let ics_{ij} be the intermediate category sub-category which refers to the different levels of categories in the intermediate category ic_i and let icw_{ijk} be the terms (terms are actually words since wordnet can process only words) in each ics_{ij}. Each ics_{ij} was found to be composed of an average of 2 words. Similarly we consider tcs_{ij} and tcw_{ijk}. For wordnet based mapping, the path length based semantic similarity measure as proposed by Wu and Palmer is employed.

Algorithm : Direct and Wordnet based Mapping

Functions :

DM(icw_{ijk},tcw_{ijk}) – performs a string matching of icw_{ijk} with all tcw_{ijk}

FLSM(icw_{ijk},tcw_{ijk}) – matches the first level icw_{ijk} with the first level tcw_{ijk} which gives the maximum similarity measure with Wu and Palmer

SM(icw_{ijk},tcw_{ijk}) - matches the icw_{ijk} with the tcw_{ijk} which gives the maximum similarity measure with Wu and Palmer for all i, j and k in tcw_{ijk}

Input :

Intermediate categories ic_i retrieved by passing query through yahoo directory search.

Direct_count initially 0, stores the number of times a tcw_{ijk} is mapped directly

First_level_target_category initially 0 records the first level target category which is mapped in FLSM(icw_{ijk},tcw_{ijk})

Similarity_measure_target_category_count initially 0, stores the number of times a tcw_{ijk} is mapped using wordnet, i.e. SM(icw_{ijk},tcw_{ijk})

Output :

Required target category

Step 1: Tokenize ic_i's into ics_{ij}'s

Step 2 : For every j in ics_{ij}, perform steps 3 to 10

Step 3: Tokenize ics_{ij}'s into icw_{ijk}

Step 4: For every k in icw_{ijk} which are not stopwords, perform steps 5 to 10

Step 5: Increment Direct_count of the tcw_{ijk} which matches by Direct Matching DM(icw_{ijk},tcw_{ijk}) for every i,j,k in tcw_{ijk}

Step 6: For every k which has no result in DM(icw_{ijk},tcw_{ijk}), perform steps 7 to 10

Step 7: Calculate FLSM(icw_{ijk},tcw_{ijk}). This would give an unambiguous viewpoint of the possible first category.

Step 8: Increment First_level_target_category for tcw_{ijk} which returns Max(FLSM (icw_{ijk},tcw_{ijk}))

Step 9: Calculate SM(icw_{ijk},tcw_{ijk}) between the intermediate category term and every tcw_{ijk}

Step10: Increment Similarity_measure_target_category_count of the tcw_{ijk} which returns Max(SM(icw_{ijk},tcw_{ijk}))

Step 11: Arrange tcs_{ij} in descending order of Direct_count, First_level_target_category, Similarity_measure_target_category_count

Step 12: If tcs_{ij} in the top position is of first level, then

a. Search for a high ranking subcategory of tcs_{ij} and thus decide the tc_i
b. If no such tcs_i is found, put the subcategory as "Others"

Step 13: Else, take the corresponding first level category of tcs_{ij}

3.2.1 Semantic Similarity Measure

The mapping of the intermediate categories to their target categories is a crucial phase which can lead to a reduced precision and recall, if not done pertinently. As done in the previous works[1][2] direct matching of the intermediate and target category terms help in correlating the intermediate categories and the target categories. But the fact remains that there is a high possibility of not finding target categories which map directly to the intermediate categories. So the glossary based mapping was performed for the remaining intermediate categories whereby a glossary of terms associated with the target categories was used. When experimented, it was found that due to ambiguity the thesaurus and other external lexicon had a high level of straying. This is attributed to the reduced precision and recall as has been shown under experimentation and results.

The problem under consideration is to find the similarity between the terms in the intermediate and the target categories to facilitate the category mapping. Six wordnet based techniques were taken and based on their performance in the benchmark Miller-Charles and Finkelstein dataset, one technique was ultimately selected. Based on Wordnet, two classes of measures namely, similarity and relatedness can be used to compare words [7]. A measure is said to be similarity based if the relatedness between concepts (represents a single distinct sense) is measured based on hypernyms and hyponyms. These relationships are specified only for noun and verb pairs and so these measures are limited to only these parts of speeches. Since they are based on the is-a hierarchy, path length based approaches which calculate the shortest distance between the edges in a graph are used for this purpose. The shorter the distance, the more similar are the words. Another approach used in the measures of semantic similarity is based on information content. This approach has its foundation on the reasoning that semantic similarity between words increases as the amount of information shared between them increases. Since wordnet is ontology, the common information is contained in the lowest common concept that subsumes both. The intermediate and target category terms are purely nouns or verbs and so five measures of semantic similarity are considered in this paper. Three information content based semantic similarity measures proposed by Resnik[8], Lin[9] and Jiang and Conrath[10] and two path length based techniques proposed by Leacock and Chodorow [11] and Wu and Palmer [4] were contenders.

The measures of semantic relatedness are based on supplementary information like the gloss, meronyms and holonyms. The adapted Lesk technique which uses the definitional glosses and the relations in Wordnet was tested

for possibly employing it in query classification [12] based on the correlation of the similarity measures against human ranking, Adapted Lesk was singled out ignoring the similarity measures proposed by Jiang and Conrath, Leacock and Chodorow, Lin, Resnik and Wu and Palmer. But it was found that Adapted Lesk took a much longer time to calculate the similarity measure than Wu and Palmer which came second. So considering the trade-off between the time to get the result and accuracy Wu and Palmer algorithm was chosen. But not much of a compromise was made since Wu and Palmers method was almost on par with Adapted Lesk as can be seen in section 4.0.

3.3 Ranking

Each target category is assigned a weight $w_p(tc_i)$ based on its position in the search results. With the highest priority assigned to the $\max(w_p(tc_i))$, the target categories are ranked. When two or more tc_i occur in different positions, the first occurring tc_i is considered. This parameter is usually used by researchers [1][2][6] and we considered two more parameters for ranking and analyzed the best parameter. The second parameter is the frequency of occurrence of the various target categories, namely $w_f(tc_i)$. The target categories are now ranked with the tc_i with $\max(w_f(tc_i))$ getting the top rank.

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 $p_1, p_2, \ldots p_n$ 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.

$$w_{fp}(tc_i) = \sum_{i=1}^{n} \alpha_i$$

4. EXPERIMENTATION AND RESULTS

4.1 Comparison between the wordnet similarity measures

To filter the best wordnet based semantic similarity measure to consider, two commonly used datasets were taken. The Miller-Charles dataset with 38 human subjects evaluating 30 pairs of words and the Finkelstein dataset with 13 human experts evaluating 352 pairs of words were considered for the purpose of comparison. Spearman's rank correlation coefficient and Kendall's rank correlation coefficient were used to find the correlation between the base datasets and the values returned by the six similarity measures under consideration. The results are as tabulated in table 1 and table 2.

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

Similarity Measure	Spearman Rank Correlation Coefficient	Kendall Rank Correlation Coefficient
Adapted Lesk	0.956943936630477	0.848377208148486
Jiang and Conrath	0.874563066677181	0.751053338006296
Leacock and Chodorow	0.942078711192773	0.83230185214418
Lin	0.843883087327051	0.715029264684776
Resnik	0.902880437921394	0.763285245951998
Wu and Palmer	0.947487446358392	0.832781233761188

Table 2. Comparison of Finkelstein dataset and Six Similarity Measures

Similarity Measure	Spearman Rank Correlation Coefficient	Kendall Rank Correlation Coefficient
Adapted Lesk	0.6079913270771	0.448450734819099
Jiang and Conrath	0.402441357997534	0.294184536668673
Leacock and Chodorow	0.509401950119176	0.374869415658709
Lin	0.392537001680256	0.293349453104224
Resnik	0.499302290585459	0.366479022470691
Wu and Palmer	0.527698512691584	0.38353456972069

The results as tabulated in table 1 and table 2, suggest that the Adapted Lesk technique has the highest correlation with the rating of the human experts. So based on the correlation coefficients, it was decided to use the Adapted Lesk technique to find the similarity between words in the intermediate and target categories. But it was found that Adapted Lesk took a considerably long time when executed on a machine. So it was decided to take the next best algorithm and ultimately Wu and Palmer was used as the Wordnet similarity measure.

4.2 Training Dataset and Test Dataset

There is no available benchmark dataset to check the category into which a query falls. The KDDCUP competition held in 2005, gave 67 target categories into which the queries had to be ranked. Also a sample of 111 queries and their ranked categories were made available. These 111 queries with their categories were taken as the training data. The test dataset is from an AOL query log with a 500K 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 3. 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.

Table 3. Test Dataset

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

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-averaged precision, recall and 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

$$\text{Precision} = \frac{\text{Re}lC}{\text{Re}tC} \quad (5)$$

$$\text{Recall} = \frac{\text{Re}lC}{ExpC} \quad (6)$$

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 4. 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 4. A Comparison of the Manual Classifiers

Set1	Set2	Precision	Recall	F1
Manual1	Manual2	0.45	0.42	0.44

Table 5 is used to compare the performance of the two automated methods namely, glossary based and wordnet based mapping. The wordnet based mapping has proved itself to outperform the glossary based method adopted by previous researchers.

Table 5. Performance of the Automated Classifiers

Set1	Set2	Precision	Recall	F1
Manual1	Wordnet	0.41	0.60	0.49
Manual2	Wordnet	0.43	0.66	0.52
Manual1	Glossary	0.22	0.38	0.28
Manual2	Glossary	0.21	0.40	0.28

Figure 2 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 position and frequency was found to be far above the other two parameters, thereby giving an ideal choice of parameter to choose while considering search results.

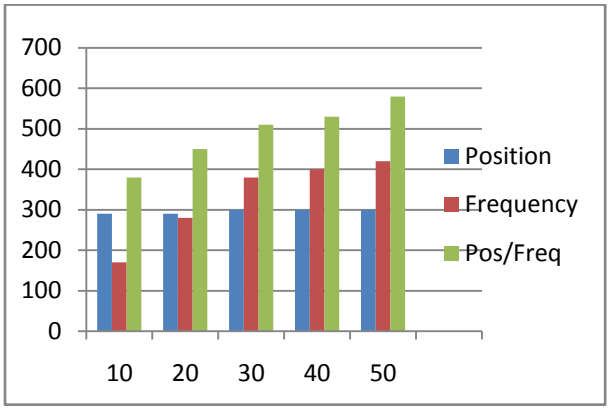

Fig. 2. Performance of the three parameters in ranking categories

Figure 3 makes an analysis of the wordnet 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 low precision obtained can be due to many factors like the difference in the interpretation of a query and the inability of wordnet to process phrases.

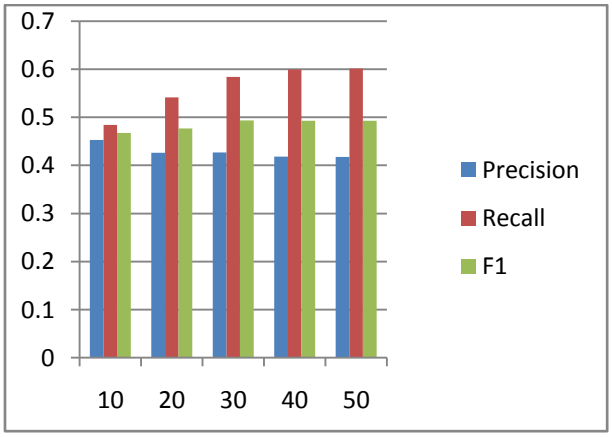

Fig. 3. Average performance of Wordnet based classifier for varying number of search results

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 & Humanities, Information\References & Libraries, "alfred Hitchcock" was mapped to Entertainment\Movies, Entertainment\TV, Entertainment\Celebrities, Living\Book & Magazine and Entertainment\Games & 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". This provides sufficient evidence to the low recall and precision obtained by all researchers in general.

5. CONCLUSION AND FUTURE WORK

Web directory search result categories are mapped to the target category using Wordnet and the results when analyzed with glossary based mapping was found to yield a better precision and recall. Three parameters namely the position of the category, the frequency of occurrence of the category and a combination of the position and frequency was used to rank the categories and the third parameter was found to give the best results. A detailed experimentation of the usage of different number of search result pages showed that the results touched their peak performance at 40 search result pages. These results are useful in the categorization of queries since they would help to return the best possible results in the first few pages of the search engine. In future, web search results will also be taken, since they are factor in achieving a improved precision and recall. Also various other features like the query log can be analyzed for the effective classification of the queries.

6. REFERENCES

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

[2] Shen, D., Sun, J., Yang, Q. and Chen, Z. 2006. 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. 131- 138.

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

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

[5] Sullivan, D., "Searches Per Day, Search Engine Watch", [WWW document] http://searchenginewatch.com/ 2156461 (accessed 4th January 2010).

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

[7] Ted Pedersen, Siddharth Patwardhan, Jason Michelizzi. 2004. WordNet::Similarity: Measuring the relatedness of concepts. In Proceedings of Human Language Technology Conference. Boston. Massachusetts. 38-41.

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

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

[10] J. J. Jiang and D. W. Conrath. 1998. Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings of the International Conference on Research in Computational Linguistics. Taiwan.

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

[12] Satanjeev Banerjee and Ted Pedersen. 2002. An Adapted Lesk Algorithm for word sense disambiguation using wordnet. In Lecture Notes in Computer Science. Volume 2276. 136 - 145.