Varying Approaches to Topical Web Query Classification

Steven M. Beitzel
Telcordia Technologies, Inc.
One Telcordia Drive
Piscataway, NJ 08854
steve@research.telcordia.com

Eric C. Jensen, Abdur Chowdhury, Ophir Frieder
Information Retrieval Laboratory
Illinois Institute of Technology
Chicago, IL 60616
{ej,abdur,ophir}@ir.iit.edu

ABSTRACT
Topical classification of web queries has drawn recent interest because of the promise it offers in improving retrieval effectiveness and efficiency. However, much of this promise depends on whether classification is performed before or after the query is used to retrieve documents. We examine two previously unaddressed issues in query classification: pre versus post-retrieval classification effectiveness and the effect of training explicitly from classified queries versus bridging a classifier trained using a document taxonomy. Bridging classifiers map the categories of a document taxonomy onto those of a query classification problem to provide sufficient training data. We find that training classifiers explicitly from manually classified queries outperforms the bridged classifier by 48% in F1 score. Also, a pre-retrieval classifier using only the query terms performs merely 11% worse than the bridged classifier which requires snippets from retrieved documents.

Categories and Subject Descriptors
H.3.5 [**Information Storage and Retrieval**]: Online Information Services – *Web-based services*

General Terms
Measurement, Reliability, Experimentation

Keywords
Query Classification, Web Search

1. INTRODUCTION
Topical web query classification can be leveraged by search services to improve efficiency and effectiveness. But, whether this classification is available for use in the retrieval process, or only after, is a key concern. If it is to be used in query routing, for example, pre-retrieval classification is by definition required. Most text classification research focuses on classifying documents, which contain enough terms to adequately train machine learning approaches. The task of classifying web queries is different in that web queries are short, providing very few inherent features. Therefore, most approaches use the documents retrieved by a query as features to classify it. The 2005 KDD Cup focused on the topical classification of web queries. The lack of substantial training data led many participants to turn to external sources to train their systems [2]. This typically consisted of training a document classifier using taxonomies of web pages such as the Open Directory Project (http://www.dmoz.org), and then bridging the categories of that taxonomy onto those desired for the query classification. Many participants achieved satisfactory F1 scores, the harmonic mean of precision and recall, but did not go any further to analyze success and failure.

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007.

We focus on two unaddressed questions: the effect of bridging a document classifier to the query classification problem, and the relative effectiveness of pre versus post-retrieval query classification techniques. We hypothesize that the concepts described by queries of a certain class ("news" queries, for example) do not necessarily correspond with those of documents classified into a category of the same name. We know, for example, that relevance feedback is effective because the language of queries and that of documents often differs.

2. PRIOR WORK
The KDD Cup dataset consisted of 800,000 web queries each to be classified into up to five of 67 possible topical categories. A training set of 111 classified queries was provided, and three human assessors independently judged 800 randomly selected queries for the test set. Several runs made use of external information. Shen and colleagues used an ensemble of several bridged classification techniques to create the winning submission [3]. These included mapping web taxonomies from Google™, Looksmart™, and a crawl of the ODP hierarchy to the 67 categories employed at the KDD Cup based on synonymy via WordNet (http://wordnet.princeton.edu) and submitting category names as queries to Google™. The pages in these taxonomies, their snippets (query-biased summaries), and titles were used as training data. Each test query was then processed to retrieve its snippets which are submitted to each classifier and their results are combined. This approach resulted in an F1 score of 0.44, not far from the mean F1 score of 0.50 when evaluating manual labelers against one-another. However, their baseline of pre-retrieval performance (using only the query terms without the snippets from Google) performed 40-50% worse in F1 than their bridged post-retrieval techniques. Also, they found that using only the snippets of documents in training consistently outperformed using their full text, which they attribute to the introduction of noise.

3. METHODOLOGY
We compare and combine query classifiers that can be applied before gathering the retrieved documents (pre-retrieval classifiers), a bridged document classifier trained from pages in the ODP (as used in the KDD cup), and explicit query classifiers trained on the retrieved documents of classified queries. We use the 20,000 queries manually classified into 18 general topical categories available in previous work by Beitzel, et al [1]. This provides us enough training data to effectively test our explicit classifiers, as compared to only the 111 training queries in the KDD dataset. We partitioned the queries into 1/3 training, 1/6 tuning, and 1/2 testing. For the post-retrieval classifiers (all support vector machines) we used the training queries to build the model and the tuning queries to select the threshold at which we report F1 in testing. To train the explicit classifiers and test

each of the post-retrieval classifiers, we processed each query with Google to obtain the top ten retrieved documents and their snippets. Each SVM classifier uses the default configuration of LIBSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm). The pre-retrieval classifier we evaluate is an ensemble of exact match, perceptron, and selectional preference classifiers described in Beitzel, et al [1]. These methods leverage both labeled and unlabeled query logs for training, expanding on the training queries based on category phrase statistics. Since they are independently trained, they only require a tuning set to select the optimal threshold for this task. Therefore, we use the training and tuning sets combined to set this threshold for these methods. Classification uses only the query string itself. The bridged post-retrieval document classifier is an SVM trained using web pages in the ODP with their categories manually bridged to one of the 18 in the testing set by the authors. Although these documents were spread across thousands of very specific ODP categories, in most cases, one of their general parent categories corresponded reasonably to one of the 18 in the testing set. To isolate the effect of bridging document categories to query ones, we use the same retrieved documents as for the explicit classifier to train our document classifier, simply replacing their known query class with the bridged class from the ODP. Classification uses the snippets from the retrieved documents. Finally, our explicit post-retrieval query classifiers were trained on the 6,666 queries in our training set based on their manually assigned categories. We evaluate two variants of explicit classifier, one trained and tested using only the snippets of retrieved documents, and one trained and tested using both the snippets and full text of the top ten documents retrieved.

4. RESULTS & ANALYSIS

To determine how these three categories of query classifiers compare to each other, we first examine the overall optimal performance for each classifier. Then, we combine the classifiers to try and exploit their differences for overall improved performance. The performance of each classifier over our 10,000 query testing set using the threshold of optimal F1 from the tuning set is detailed in Table 1. Surprisingly, one can achieve nearly as effective performance from pre-retrieval classifiers that use only the query string itself for classification as that of the generic text classifier which requires the retrieved documents. The post-retrieval classifiers learned explicitly from classified query logs improve upon this substantially, with a 48% relative improvement in F1. Clearly, performance is lost when treating query classification as a generic topical text classification problem by mapping document taxonomies to query ones. Like Shen, et al., we find that using only snippets outperforms including the full text of documents for classification [3]. Further analysis is warranted, but like them we hypothesize the full text introduces too much noise. Based on the results from the individual classifiers, we hypothesized that differences in classifiers would provide for improved performance if they were combined. We fused them together, using the classifications from higher-precision classifiers first, and backing off to higher-recall classifiers when necessary. Despite their very different focus, however, this combination of pre-retrieval classifiers with the best post-retrieval one (explicit using snippets) does not provide substantial improvement. With the additional information available post-retrieval, the imprecision of the pre-retrieval techniques prevents them from adding substantial value. By however slight margin, this fusion does represent the best post-retrieval performance we achieve. To examine the effectiveness of pre- versus post-retrieval classification in more detail, we show the overall precision/recall tradeoffs of the pre- and best post-retrieval (fused) classifiers in Figure 1. The ability of retrieved document classifiers to achieve greater recall than the query-log-based, pre-retrieval, classifiers is expected due to the larger number of features available.

Classifier	F1	Precision	Recall
Pre-retrieval	0.240	0.191	0.322
Bridged	0.266	0.275	0.258
Explicit (snippets)	0.394	0.336	0.476
Explicit (snippets + docs)	0.382	0.395	0.370
Pre-retrieval + Explicit (snippets)	0.396	0.342	0.472

Table 1: Classifier Performance

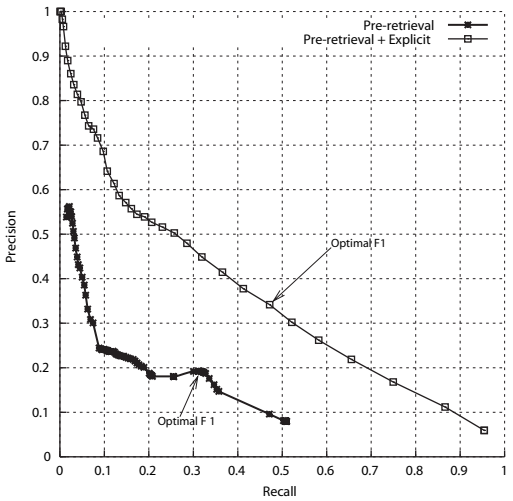

Figure 1: Best Pre- and Post-Retrieval Precision and Recall

5. CONCLUSIONS & FUTURE WORK

We have evaluated three differing approaches to topical web query classification. We find that training explicitly from classified queries outperforms bridging a document taxonomy for training by as much as 48% in F1. We have also shown that pre-retrieval classification using only the query string can provide surprisingly effective results, enabling adjustments to the retrieval process to improve effectiveness and efficiency. However, our fusion of multiple approaches did not yield improved performance. In future work we will analyze the differences between these methods and develop improved combination strategies.

6. REFERENCES

1. Beitzel, S.M., Jensen, E.C., Lewis, D.D., Chowdhury, A., Kolcz, A. and Frieder, O., Improving Automatic Query Classification via Semi-supervised Learning. in *IEEE ICDM,* 2005, 42-49.
2. Li, Y., Zheng, Z. and Dai, H.K. KDD Cup-2005 Report: Facing a Great Challenge. *SIGKDD Explorations*, *7* (2), 2005, 91-99.
3. Shen, D., Sun, J., Yang, Q. and Chen, Z., Building bridges for web query classification. in *ACM SIGIR,* 2006, 131-138.