Automatic Web Query Classification Using Labeled and Unlabeled Training Data

Steven M. Beitzel, Eric C. Jensen,
Ophir Frieder, David Grossman
Information Retrieval Laboratory
Illinois Institute of Technology
{steve,ej,ophir,dagr}@ir.iit.edu

David D. Lewis, Abdur Chowdhury,
Aleksandr Kolcz
America Online, Inc.
davelewis@daviddlewis.com
{cabdur,arkolcz}@aol.com

ABSTRACT
Accurate topical categorization of user queries allows for increased effectiveness, efficiency, and revenue potential in general-purpose web search systems. Such categorization becomes critical if the system is to return results not just from a general web collection but from topic-specific databases as well. Maintaining sufficient categorization recall is very difficult as web queries are typically short, yielding few features per query. We examine three approaches to topical categorization of general web queries: matching against a list of manually labeled queries, supervised learning of classifiers, and mining of selectional preference rules from large unlabeled query logs. Each approach has its advantages in tackling the web query classification recall problem, and combining the three techniques allows us to classify a substantially larger proportion of queries than any of the individual techniques. We examine the performance of each approach on a real web query stream and show that our combined method accurately classifies 46% of queries, outperforming the recall of the best single approach by nearly 20%, with a 7% improvement in overall effectiveness.

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

General Terms: Algorithms, Experimentation

Keywords: Query Classification, Web Search

1. INTRODUCTION
Understanding the meaning of user queries is a problem at the heart of web search. Successfully mapping incoming general user queries to topical categories, particularly those for which the search engine has domain-specific knowledge, can bring improvements in both the efficiency and the effectiveness of general web search. These improvements can be realized in applications such as query routing, topic-specific query reformulation, and targeted advertising for commercial search services. To fully realize these gains, a classification system is required that can automatically classify a large portion of the general query stream with a reasonable degree of accuracy. In the case of large-scale web search, this task is particularly challenging due to the sheer size and dynamic nature of the web's content, users, and query traffic, as well as the difficulty of accurately determining a user's desired task and information need from short web queries.

SIGIR'05, August 15–19, 2005, Salvador, Brazil.
ACM 1-59593-034-5/05/0008.

We examine three methods for classifying general web queries: exact match against a large set of manually classified queries, a weighted automatic classifier trained using supervised learning, and a rule-based automatic classifier produced by mining selectional preferences from hundreds of millions of unlabeled queries. The key contribution of this work is the development of a query classification framework that leverages the strengths of all three individual approaches to achieve the most effective classification possible. Our combined approach outperforms each individual method, particularly improves recall, and allows us to effectively classify a much larger proportion of an operational web query stream.

2. PRIOR WORK
Initial efforts in automatic web query classification have focused mostly on task-based classification and unsupervised clustering. A number of researchers have automatically classified queries by task into Broder's informational / navigational / transactional taxonomy (or very similar taxonomies) and applied different query rewriting and weighting strategies for each category. Kang & Kim used query term distribution, mutual information (co-occurrence), usage rate in anchor text, part-of-speech info, and combinations of the above to automatically classify queries into Broder's taxonomy [2]. They used this classification to decide which query processing algorithms to use for retrieval, achieving at best modest improvements in retrieval effectiveness. This study exhibits a problem common to all query classification studies: the difficulty of achieving accurate classification, and in particular high recall, for the inevitably short queries.

Gravano, et al. used machine learning techniques in an automatic classifier to categorize queries by geographical locality [1]. Their analysis revealed that data sparseness was a problem for their classifier, and concluded by positing that for such an automatic classification system to be successful, it would have to draw on auxiliary sources of information to compensate for the small number of available features in a single query.

3. METHODOLOGY
Our framework for classifying web queries combines a precision-oriented exact match approach with higher recall machine learning approaches. The exact match component uses 18 lists of popular web queries manually classified by a team of editors at AOL™. These lists, while covering a tiny fraction of unique queries seen, actually represent approximately 12% of the general query stream. The exact match component provides high precision and known behavior for high frequency queries.

However, it is expensive to maintain, has low recall, and both its recall and precision degrade over time.

Therefore, we leverage the labeled queries from the exact match system in two ways. First, we treat the manually classified queries as training data for supervised learning. We train a perceptron classifier for each of the 18 categories, treating each as a binary classification task. The input features are simply the query words (2.6 per query on average) so inputs are very sparse. Each perceptron was first trained to distinguish queries in one of the 18 categories from the other 17 categories, using the several hundred thousand queries on these lists. The thresholds of all perceptrons were then tuned to a common value, so as to optimize the micro-averaged F_1 measure on a validation set of 5,283 queries randomly sampled from the query stream and manually classified. We found that while the perceptron classifier could in theory achieve very high recall, it could do so only at the cost of very low precision. Striking a balance by optimizing F_1 yielded substantially higher recall than exact match, and substantially lower (but not disastrous) precision.

To improve recall further, we leveraged the labeled queries in a second fashion, by treating them as a lexicon of semantically classified lexemes (words and fixed phrases). We then applied the computational linguistics notion of *selectional preferences* [3] to these data. The idea is that a word *x* may strongly prefer that a word *y* following it (or preceding it) belong to class *u*. If so, then *u* is a good prediction for the class of an ambiguous, or previously unknown, *y* in that context. We mine selectional preferences from a large unlabeled query log as follows:

1. Convert queries in the log to a set of head-tail (x,y) pairs.
2. Convert the (x,y) pairs to weighted (x,u) pairs, discarding y's for which we have no semantic information
3. Mine the (x,u) pairs to find lexemes that prefer to be followed or preceded by lexemes in certain categories (preferences)
4. Score each preference using Resnik's Selectional Preference Strength [4] and keep the strongest ones

The preferences are then applied to a new query to estimate the probabilities that particular components in the query belong to particular categories. We tune a threshold on these probability estimates for components, so as to optimize effectiveness of classifying entire queries, as measured by micro-averaged F_1.

Simply tuning the threshold of a classifier cannot make it do well on queries whose important features were simply not present in the training data. Table 1 shows that our three classifiers vary substantially in the category assignments they get correct. This suggested combining them for greater effectiveness. We tested a simple combined classifier that assigned a query to each category that any of the above three classifiers predicted. The component classifiers were individually tuned to optimize micro-averaged F_1, but no tuning of the combined classifier as a whole was done. More sophisticated approaches are certainly possible.

	Exact Match	Perceptron
Perceptron	.1243	N/A
Selectional Preferences	.0894	.6513

Table 1: Positive Example Overlap

4. RESULTS & ANALYSIS

Table 2 shows the effectiveness, as measured by the micro-averaged F_1, of the three individual classification approaches and our disjunctive combination scheme.

	Micro F1	Micro Precision	Micro Recall
Exact Match	.0749	.3079	.099
Perceptron	.1135	.1913	.279
SP	.1093	.1524	.3856
Combined	.1215	.1651	.4602
Over Best	7.05%	-46.38%	19.35%
Over Worst	62.22%	8.33%	364.85%
Over Mean	22.44%	-23.99%	80.80%

Table 2: Classification Effectiveness for Each Technique

To provide additional insight we also show the micro-averaged recall and micro-averaged precision, though it is important to remember that averaged F_1 values are not simple functions of averaged recall and precision values. These results show that the combined approach outperforms all three individual techniques, and in particular it achieves approximately 20% higher recall than any single approach. This large increase in recall suggests that the combined approach may in fact be a tenable solution to the recall problem that has hindered past efforts at query classification, and allow this method to address a much larger proportion of the web query stream.

5. CONCLUSIONS & FUTURE WORK

We proposed a framework for automatic web query classification that combines a small seed manual classification with techniques from machine learning and computational linguistics. This framework is able to outperform each of its component approaches, and achieves high recall in particular, while still maintaining reasonable precision. This high degree of recall allows the combined approach to classify a much larger portion of the query stream than would be possible using any of the individual approaches along. Moreover, our hope is that by leveraging unlabeled data we can minimize the need for periodically labeling new training data to keep up with changing trends in the query stream over time. There are several potential areas for future work, including expanding the initial seed-manual classification using queries classified by the framework, improving the linguistic properties of our SP algorithm, incorporating ideas from traditional rule-learning, examining the effectiveness of each approach on popular versus rare queries, and examining the performance of each approach on specific topical categories.

6. REFERENCES

[1] Gravano, L., Hatzivassiloglou, V. and Lichtenstein, R., Categorizing Web Queries According to Geographical Locality. in *ACM CIKM*, (2003), ACM.

[2] Kang, I.-H. and Kim, G., Query Type Classification for Web Document Retrieval. in *ACM SIGIR*, (2003), ACM.

[3] Manning, C.D. and Schutze, H. *Foundations of Statitical Natural Language Processing*. MIT Press, 1999.

[4] Resnik, P. Selection and Information: A Class-Based Approach to Lexical Relationships, Ph.D Thesis, University of Pennsylvania, 1993.