Biasing Web Search Results for Topic Familiarity

Giridhar Kumaran
Center for Intelligent
Information Retrieval
Dept. of Computer Science
University of Massachusetts
Amherst, MA 01003, USA
giridhar@cs.umass.edu

Rosie Jones
Yahoo! Research
74 N Pasadena Ave., 3rd Floor
Pasadena, CA 91103, USA
jonesr@yahoo-inc.com

Omid Madani
Yahoo! Research
74 N Pasadena Ave., 3rd Floor
Pasadena, CA 91103, USA
madani@yahoo-inc.com

ABSTRACT

Depending on a web searcher's familiarity with a query's target topic, it may be more appropriate to show her *introductory* or *advanced* documents. The TREC HARD [1] track defined topic familiarity as meta-data associated with a user's query. We instead define a user-independent and query-independent model of topic-familiarity required to read a document, so it can be matched to a given user in response to a query. An **introductory** web page is defined as
A web page that doesn't presuppose any background knowledge of the topic it is on, and to an extent introduces or defines the key terms in the topic.
while an **advanced** web page is defined as
A web page that assumes sufficient background knowledge of the topic it is on, and familiarity with the key technical/ important terms in the topic, and potentially builds on them.
We develop a method for biasing the initial mix of documents returned by a search engine to increase the number of documents of desired familiarity level up to position 5, and up to position 10. Our method involves building a supervised text classifier, incorporating features based on reading level, the distribution of stop-words in the text, and non-text features such as average line-length. Using this familiarity classifier, we achieve statistically significant improvements at reranking the result set to show introductory documents higher up the ranked list. Our classifier can be seamlessly integrated into current search engine technology without involving any major modifications to existing architectures.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms: Algorithms

Keywords: Familiarity, Web Search, Personalization

1. DATA

We decouple our evaluation of familiarity re-ranking from the evaluation of topic-relevance. We assume that the top 20 documents returned by a search engine are all relevant to the query, and evaluate our familiarity reranking on this set. Data was obtained by querying the Inktomi search engine with 40 queries drawn uniformly at random from 2004 Yahoo! Search query logs. Minor processing of the queries like removal of adult queries and queries not in English, manual correction of spelling mistakes, and manual insertion of spaces between words if they were absent was done. The top 20 documents returned for each query were randomly permuted, and presented to three annotators to label as *introductory* or *advanced*. The number of queries for user 1, 2 and 3, were respectively 16, 24, and 13, and the total number of documents labeled (training instances) were 508, 766, and 463. Two queries were in common to enable measurement of inter-annotator agreement, for which we used Cohen's Kappa statistic [3], and obtained coffecients of 0.32, 0.59, 0.48 between annotators 1 and 2, 1 and 3, and 2 and 3 respectively (who co-annotated 36, 38, and 37 documents respectively). These coffecients indicate fair to moderate levels of agreement.

2. FAMILIARITY CLASSIFIER

We examined three kinds of features which could be predictive of familiarity, and built a classifier FAMCLASS to combine them. The three different feature types were:

1. Stop-word features, which are predictive in several text categorization tasks, numbering around five hundred. We used the rainbow library [6] to obtain the stop-word frequencies. An advantage of using stop-words is that we can be confident that we are not over-fitting to the topics of our training data, and are building a topic-independent model.

2. Eleven features we designed based on various characteristics of web page documents [5].

3. Features used to determine reading level [4] [7]

We sought an algorithm that could handle non-linearity as well as a mix of different feature types with different value ranges. Random forests [2] proved to be the most convenient choice. We used a forest of thousand trees in our experiments.

3. EVALUATION

People expect to find the information they are looking for in the first page of search engine results i.e. within

CIKM'05, October 31–November 5, 2005, Bremen, Germany.
ACM 1-59593-140-6/05/0010.

	p@1				
Measure	Base	ALL	StW	NtF	RL
Micro Avg.	0.558	0.615	0.635	0.404	0.538
Pooled	0.565	0.5	0.610	**0.630**	0.565
	p@5				
Micro Avg.	0.540	**0.616**	**0.608**	0.516	0.544
Pooled	0.536	0.618	0.641	0.627	0.536
	p@10				
Micro Avg.	0.498	**0.590**	**0.576**	0.53	0.508
Pooled	0.5	**0.580**	**0.570**	0.545	0.527

Table 1: Baseline (default search engine ranking) performance versus classifier performance using all features (ALL), only stop word features (StW), only non-textual features (NtF), and only reading level (RL) features. Pooled refers to the case when we pooled data from all annotators for training (46 queries). Micro averaging averages results for individual annotator specific classifiers, and was done to provide a comparison point to the 'pooled' results. Values in bold are statistically significant improvements compared to the baseline. Statistical significance was measured using the sign test at a 95% confidence level.

	p@5		p@10	
Annotator	Base	ALL	Base	ALL
1	0.467	0.52	0.407	**0.50**
2	0.645	**0.736**	0.655	0.709
3	0.446	0.523	0.338	**0.492**

Table 2: Baseline performance versus classifier performance using all features (ALL) for individual users. Values in bold are statistically significant improvements compared to the baseline. Statistical significance was measured using the sign test at a 95% confidence level.

the first ten documents. This holds true for a familiarity-flavored search too. To compare our approach against the baseline (default search engine ordering), we measured the proportion of introductory documents at ranks one (p@1), five (p@5), and ten (p@10) in top 20. The classifier results we report are all based on *leave-one-query-out* validation: we partition the labeled documents based on query, and for every fold we hold out all the documents associated with a certain query, train on the remainder of the documents and rank the held out documents. Table 1 consolidates the results for classifiers trained on the different subsets of features and documents and reports the performance of the baseline ranking as well. For the classifiers trained on the pooled data, we remove the two queries in common to the annotators, since they may disagree on some of those documents, which yields 46 queries. The performance of per-user (per-annotator) trained classifiers (Micro Avg. in Table 1) when trained on all features or stop words is close to the performance of the pooled classifiers, even though there are significantly more training documents available for the latter (Section 1). We observe that in many cases the performance of the classifiers is significantly higher than the baseline for p@5 and p@10, in particular for the per-user trained classifiers. Table 2 reports performances for per-user classifiers and the baseline for each annotator.

introductory	so, enough, just, in, needs, help, each,away
advanced	if, cause, while, way, through, which, us

Table 3: The stopwords with highest coefficients across multiple training runs of a linear classifier for introductory and advanced documents.

4. DISCUSSION AND CONCLUSIONS

Stop-words appeared to be the most important features in our classifier, when we broke down the feature sets into the three sub-types. To study the contribution of each feature to the classifier's decision, we trained a linear classifier and examined the coefficients of the features [1]. In Table 4 we show the top-ranking stop-words (in the top 20, when features are sorted by decreasing magnitude of coefficient). Note that "help" is indicative of introductory content. The other stop-words are suggestive of differences in genre or writing style, with the advanced words perhaps suggestive of more formal or scientific writing, while the introductory words are suggestive of informal or colloquial writing. Note that appearance of a single highly weighted feature such as "help" in a document does not imply that the classifier will necessarily output "introductory" for the document. The presence of other features in the document, as well as the frequency of the feature in the document (in case of stop words), also affects the classifier's output. Features other than stop words that were often ranked high by the linear classifier included several reading level features (syllables-per-word, percent-complex-words) which had positive coefficients, indicative of introductory documents, and one non-textual feature: average-word-length, which had a negatively coefficient, indicative of advanced documents.

FAMCLASS can re-rank in the order desired i.e. based on advanced or introductory preferences. It can be extended to handle even greater granularity (classes) of familiarity, subject to availability of suitable training data. Our experiments indicate that we can perform search result biasing for arbitrary users on arbitrary queries.

5. REFERENCES

[1] J. Allan. HARD track overview in TREC 2003 high accuracy retrieval from documents. In *Notebook Proceedings of TREC 2003*, 2003.

[2] L. Breiman. Random forests. *Machine Learning*, 45, 2001.

[3] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurements*, (20):37–46, 1960.

[4] J. Kincaid, R. Fishburn, R. R. Jr., and B. Chissom. Derivation of new readability formulas for navy enlisted personnel. Technical Report 8-75, Research Branch Report, Naval Air Station Memphis, Millington, Tennessee, 1975.

[5] G. Kumaran, R. Jones, and O. Madani. Details on biasing web search results for topic familiarity. In *UMass Amherst CIIR Tech. Report – IR-393*, 2005.

[6] A. K. McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow, 1996.

[7] G. McLaughlin. SMOG grading: A new readability formula, 1969.

[1] The performance of the linear classifier was competitive with the performance of random forests