Details on 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! Inc.
74 N Pasadena Ave., 3rd Floor
Pasadena, CA 91103, USA
jonesr@yahoo-inc.com

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

ABSTRACT

A typical web search engine returns a mix of introductory and advanced documents (around 50%) in response to a random selection of queries. 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. We conceptualize the notion of *introductory* and *advanced* documents in a way that obviates additional user-interaction and changes to existing search engine architectures. We show that topic familiarity required to understand a document (*familiarity level*) is a notion that people can agree on, as borne out by high inter-rater agreement (70%). We also show that this familiarity level is not predicted by reading level, so new methods of identifying it are needed. 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 *topic-independent* and *user-independent* 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 experiments indicate that we can perform this search result biasing for arbitrary users on arbitrary queries.

Categories and Subject Descriptors

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

1. INTRODUCTION

Different users searching for information on a topic have varying familiarity with it. There is currently no way for a user to inform a search engine of her background knowledge on a topic so that only documents appropriate to her level of expertise are returned. For example, a student searching for help with a linear algebra homework requires a different set of documents than say, a professor of mathematics interested in staying abreast with the latest in the field. The query *linear algebra* will return relevant documents, but will fail to address the backgrounds and requirements of the two users. Documents returned in response to web search queries are a mix of introductory and advanced documents, as we will describe in Section 2.3.

Familiarity was introduced into the TREC High-Accuracy Retrieval From Documents (HARD) track in 2003 [1]. It was defined as meta-data associated with a query, and is a property of the user who issues the search query, representing how much background knowledge the searcher has of the query's topic, on a five-point scale, ranging from no prior knowledge (1) to knowing details of the topic (5). In order to match a user's topic-familiarity with appropriate documents, we need to provide a way of defining the amount of topic-familiarity required to read a given document. The HARD track does not provide any guidelines for associating a degree of familiarity with documents. We provide definitions of *introductory* and *advanced* documents, corresponding to how much background knowledge is required to read the document, in Section 2. In order to verify that our definition of familiarity required to read a document is something that people can recognize, we carried out a user study. The details of the study are provided in Section 2.1.

One approach to retrieving introductory and advanced documents is to automatically modify the query so that only *introductory* documents are returned, by finding keywords indicative of introductory documents for the query. Query modification has been successfully used to retrieve documents requiring significant topic familiarity [7]. Another approach is to assume there is some relationship between document reading level and the assumed knowledge of the document. We found that both of these methods did not work well in our experiments, so a more sophisticated method is required.

Our approach is to use a classifier to label documents as *introductory* or *advanced.* This method could be used to tag documents as *introductory* or *advanced* at index time using a classifier, and simply retrieve documents from either set according to the user's preference. This runs the risk of returning less-relevant documents in favor of documents at an appropriate familiarity level. Instead, we re-rank the search results, using the classifier score. Our classifier method is both query-independent and user-independent. This means we can use the same model to classify all documents into appropriate familiarity levels. Given an arbitrary new query for a new user we can rank results from low-to-high or high-to-low on the topic familiarity scale. This option could be selected through a button or slider on the search page.

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. In practice we would need to integrate familiarity and relevance ranking into a combined framework, but this simplification allows us to study the search for *introductory* and *advanced* documents in isolation.

Our contributions are (1) a definition of the topic familiarity required to read a document (2) demonstration that this definition is valid by showing inter-rater agreement (3) demonstration that standard reading level features do not predict whether a document is introductory or advanced (4) a method of re-ranking web search results based on machine learning which significantly outperforms the default search result ranking, increasing the number of introductory documents shown on the first page.

This paper is organized as follows: in Section 2 we define introductory and advanced documents according to the background knowledge required to read them, and describe our user study and data collection. In Section 3 we describe standard reading level metrics and show that they vary widely across both introductory and advanced documents. In Section 4 we describe the three types of features we experimented with, and the classifier we use for re-ranking documents. In Section 5 we describe our evaluation methods and results. In particular, we compare the feature subtypes for re-ranking, and conclude that stopword features work best in isolation, but best performance is obtained by using all three feature types. We discuss related work in Section 6.

2. DOCUMENT TOPIC FAMILIARITY

The TREC HARD track defined topic familiarity as meta-data associated with a user's query. We would like to have a model of topic-familiarity required to read a document, so it can be matched to a given user in response to a query. That is, in response to a request for introductory documents on a given topic, we return documents only on that topic, and rank the most introductory ones highest.

In our preliminary study of labeling documents, we found that it is easy to confound document familiarity rating with query content and expectations of the searcher's information need and familiarity with that topic. For example, viewing a page describing molecular biology, a user would label it differently depending on whether the query was *biology* (labeling the page as *advanced*) or *molecular biology* (labeling the document as *introductory*), assuming that the expectations for these two queries would be different, ie that the web searchers issuing these two queries would have different levels of topic familiarity.

However, we conjecture that introductory documents on molecular biology share properties with introductory documents on biology, in that terms are defined and explained in the text itself. In this spirit we concealed the search queries from our annotators, and defined difficulty levels for web pages based on the *type* of information they contain. An **introductory** web page is thus defined:

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.

An **advanced** web page is defined:

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.

It should be noted that our definition of familiarity level of a document is **query independent** and **user independent**.

2.1 Data Collection

Since our techniques and their analysis require labeled data we had three annotators tag documents as *introductory* or *advanced.* To obtain documents to be tagged, we selected 40 queries uniformly at random from a log of web search queries from 2004. Minor processing of the queries was done. This includes 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. In the last two steps we are assuming a human-quality spell correction module. While such technology may not be available currently, it helps us separate the issues of spelling correction and familiarity re-ranking.

For each query, we issued both the raw query, and the query modified with a trigger word. In initial explorations, we developed a local feedback technique (*TRIGWORDS*) that reissued a modified query to the search engine. The modified query was created by appending the trigger word that co-occured most with the query terms in the top fifty documents initially returned. These *trigger* terms were members of a hand-crafted list of terms we thought were suggestive of introductory content. Examples of such terms were *describe*, *characteristic* and *outline.* By adding these terms to the query, we expected to direct the search towards introductory content. However, analysis of the results (Table 1) showed that this procedure didn't result in a statistically significant change in the mix of introductory and advanced documents returned. We use the documents returned in response to the modified queries for training data, but not for testing.

The queries were sent to a search engine, and the top 20 documents returned were randomly permuted, and presented to three annotators to label as *introductory* or *advanced.* The annotators were instructed to first identify the document's topic, and base their decision on the amount of information on the topic contained in the document. The annotators were permitted to label documents as "inapplicable" if neither introductory nor advanced were appropriate labels. Ex-

	p@1		p@5		p@10	
	Base	TRIGWORD	Base	TRIGWORD	Base	TRIGWORD
Micro Avg.	0.558	0.547	0.540	0.536	0.498	0.494

Table 1: Baseline and query-modification-based TRIGWORDS **performance. There was no statistically significant difference between the results in terms of introductory and advanced documents. We were able to use the data collected using the** TRIGWORDS **method as training data for our classifier.**

Annotator		*Number of common documents*	*Agreement (%)*
1	2	37	28*(75.6%)*
2	3	38	28*(73.68%)*
3	1	38	26 *(68.4%)*

Table 2: Inter-rater agreement when annotators were asked to choose between two levels of familiarity *introductory* **and** *advanced***.**

ample of documents that were labeled *inapplicable* include documents that were not in the English language, contained just an image with a caption "Click here to continue", consisted of just lists of numbers, were in PDF or PS format and so on. About 10% of documents were labeled *inapplicable* and removed from our dataset, leaving a total of 1737 unique labeled documents. In order to study inter-rater agreement, all the annotators were asked to label results for just under 40 documents in common. The remainder of the documents labeled were unique for each annotator, and the three annotators labeled 508, 766, and 463 documents respectively. The labeling process took over 40 person-hours.

2.2 Inter-rater Agreement

In preliminary experiments, the authors labeled a set of documents on a scale from one to five, one being *very introductory* and five being *very advanced.* We found it difficult to use the five-point scale, with some of us preferring the upper end of the range, and others preferring the lower end of the range. Measuring annotations differing by at most 1, we found pair-wise agreement at 60-70%, and annotations differing by at most 2, we found pair-wise agreement at 75-90%. We do not include this data in the experimental section of the paper.

This experience led us to use the binary classification labeling for our main data collection. Table 2 shows inter-rater agreement for our non-author labelers. They agree around 70% of the time on labels for documents. Comparing this data to our own inter-rater agreement, we can observe that even in the situation in which we allowed people greater freedom to choose among categories, the agreement levels didn't deteriorate. This shows that the familiarity definitions were sound, and captured the general perception of what people consider *introductory* and *advanced.*

2.3 Default Search Engine Ordering

With the documents returned by a search engine labeled as *introductory* or *advanced*, we can begin to understand how search engines currently behave. While we use only a single search engine for these experiments, it provides a data-point in the current space of available search engines.

Annotator	Num. Queries	ratio	p@1	p@5	p@10
1	16,15,15	0.425	0.375	0.467	0.407
2	23,22,22	0.637	0.739	0.645	0.655
3	13,13,13	0.387	0.462	0.446	0.338
Micro Avg.		0.509	0.558	0.540	0.498

Table 3: Baseline performance. For each annotator we calculate the proportion of introductory documents at 1 (p@1), p@5 and p@10 for the default ordering of documents returned by a search engine. We discarded ***inapplicable*** **documents, so for p@10 the averages may be computed over fewer queries. We see that on average the search engine returns slightly more introductory documents at position one than further down in the list. We also show the mix of introductory and advanced documents over all documents labeled by each user (ratio). Overall the mix of introductory and advanced documents returned was 50.1%**

Table 3 shows the distribution of documents returned by the search engine, according to our annotators. Assuming we were looking only for introductory documents, we evaluated the proportion of introductory documents out of all documents returned for a query (ratio), as well as precision[1] at position 1 (p@1), in positions 1 through 5 (p@5) and in positions 1 through 10 (p@10). The average proportion (ratio) of introductory documents over queries for each annotator are also shown in the table. Note that a random ordering of documents would yield familiarity-precisions equal to this ratio, on average. About 10% of documents were labeled *inapplicable.* We calculated precision at cut-offs for the remaining documents by respecting the original order and collapsing any empty slots. Thus there may be fewer queries considered for p@10, if fewer than 10 documents were available for some queries. The number of queries which had 1, 5 and 10 applicable documents for each annotator is given in the "Num. Queries" column in Table 3.

We see in Table 3 that annotator 2 labeled a larger proportion of documents as introductory. We saw in Section 2.1 that annotator 2 had high agreement with the other two annotators on the 37 and 38 documents retrieved in response to two common queries. The remainder of the queries labeled were unique for each annotator. In order to check whether this disparity was due to the queries used to retrieve documents for annotation, we measured the average query length for each user, but found no major differences. We also examined average query length for documents labeled introductory and advanced, and found no major differences (average length of 2.7 words and 17.1 characters for queries retrieving introductory documents, and 2.8 words and 18.4 characters for queries retrieving advanced documents, calculated with micro-averaging).

Overall we see that the search engine returns a mix of intro-

[1] As explained in the Introduction, we assume the first 20 documents returned for a query are relevant. Precision is measured with respect to *familiarity*, and not *relevance.*

ductory and advanced documents. The document in the first position is slightly more likely to be introductory: labeling documents in randomized order, 56% of the time our annotators labeled the document which was originaly in position 1 as introductory. For documents originally in the first 5 positions, our annotators labeled them as introductory just over 50% of the time. The mix over the first 10 documents contains fewer introductory documents (47%). So the search engine ranking function may already include a slight bias to return introductory documents in the top-ranked position. Our goal in this work is to re-rank the top 20 documents, to increase precision at 1, precision at 5, and precision at 10.

3. READING LEVEL AND DOCUMENT FAMILIARITY

A natural question to ask about models of documents which are introductory or advanced is whether measures of reading level [9] [12] provide sufficient resolution. In this section we describe these reading level measures and their constituent features, and show that their distribution across the classes *introductory* and *advanced* does not distinguish the classes.

We used a publicly available package Fathom [14] to obtain reading level scores. The three reading level scores are shown in Equations 1, 2 and 3.

$$fog = (words\_per\_sentence + \%complex\_words) * 0.4 \tag{1}$$

$$flesch = 206.835 - (1.015 * words\_per\_sentence) - (84.6 * syllables\_per\_word) \tag{2}$$

$$kincaid = (11.8 * syllables\_per\_word) + (0.39 * words\_per\_sentence) - 15.59 \tag{3}$$

In Table 4 we see means and standard deviations for the three reading level metrics over introductory and advanced documents. The Fog index, in particular, is designed to indicate the number of years of formal education required to read the document once and understand it. We see that both introductory and advanced documents score an average of 20 (unreadable) on the Fog index. This is partly due to outlier documents (the standard deviations are extremely high), and may be due to a mismatch between these indices, the form of web documents, and the automated way we calculated reading level features such as syllables per word. Though the introductory documents average slightly lower, the difference is much less than the standard deviation. The Flesch index rates documents on a 100 point scale, with higher scores indicating greater readability and 60-70 considered optimal. On the Flesch scale, our documents averaged around 20, with introductory documents slightly higher (slightly more readable), but the standard deviation again dwarfs the difference. Finally, the Kincaid measure scores reading level in terms of US grade school level. On average a score indicative of slightly better readability was obtained by introductory documents, but again this difference was much less than the standard deviation. We will see in our experimental section 5 that this slight difference in mean reading level is insufficient to show an improvement in reranking documents.

	Fog	Flesch	Kincaid
introductory	19.7 (68.7)	22.6 (175.9)	17.2 (67.0)
advanced	20.6 (23.3)	18.15 (63.0)	18.1 (22.4)

Table 4: Means and standard deviation for reading level metrics on introductory and advanced documents. On average documents were rated as unreadable. While introductory documents were marginally more readable, the standard deviation was much greater than the difference between the means for introductory and advanced.

Reading Level Features	
1. Fog measure	2. Flesch measure
3. Kincaid measure	4. Num. of characters
5. Number of words.	6. Percentage of complex words
7. Num. of sentences	8. Num. of text lines
9. Num. of blank lines	10. Num. of paragraphs
11. Num. syllables per word	12. Num. words per sentence.

Table 6: Before extracting the above features from the Lynx rendering of each webpage, we performed some preprocessing. We removed content enclosed by square brackets (indicates links), and excluded content after the term *References:*.

4. FAMILIARITY CLASSIFIER

In order to build a query-independent, user-independent model of the introductory nature of a document, we examine three kinds of features which could be predictive, and built a classifier FAMCLASS to combine them. The ranked sets for each query were re-ranked using the familiarity classifier with goal of moving introductory documents towards the top of the list.

We used three different feature types:

1. Stop-word features, which are predictive in several text categorization tasks [3][13] [2], numbering around five hundred. We used the rainbow library [11] to obtain the stop-word frequencies. An advantage of using stopwords is that we can be confident that we are not overfitting 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 (Table 5).

3. Features used to determine reading level (Table 6).

We postulated that some subset of the above features, for example the usage patterns of stop words together with reading level and web page document characteristics (*e.g.*, average word and sentence length), might be predictive of whether a document is introductory or advanced.

We experimented with a number of different classifiers including SVMs (with polynomial kernels), decision trees, and random forests [3]. As the learning problem was potentially nonlinear, we sought an algorithm that could handle nonlinearity. Furthermore, we sought an algorithm that could handle a mix of different feature types with different value ranges[2]. As the purpose of our current study was mainly

[2] Some algorithms such as support vector machines are sensitive

Non-textual Features	*Hypothesis*
1. Avg. num. of words per line with anchor text removed.	Web pages with a lot of non-anchor text are introductory
2. Avg. num. of anchor text words per line with other text removed.	An advanced web page has more anchor text per line
3. Document length excluding anchor text.	Longer documents are introductory
4. Anchor text count.	An advanced web page has more anchor text
5. Fraction of non-anchor text in document.	Lower the fraction, more introductory the document
6. Average word length (excluding anchor text).	Advanced documents have higher average word length due to more complex vocabulary.
7. Fraction of term "the" in text excluding anchor text.	Low fraction implies an introductory document.
8. Fraction of term "a" in text excluding anchor text.	High fraction implies an introductory document.
9. Fraction of term "an" in text excluding anchor text.	High fraction implies an introductory document.
10. Average of the top five highest TFs.	Salient terms are repeated in introductory documents
11. Similarity of WordNet expansion of top 10% of document with remaining 90%	The last 90% of of an introductory document describes the first 10%

Table 5: We used the Lynx browser to render web pages. Lynx automatically "scraped" the web pages by displaying only textual content. All counts used in the features extracted for a page are based on the standard Lynx rendering.

to determine whether a machine learning approach could significantly outperform the baseline in the difficult task of ranking based on familiarity, we did not perform exhaustive experiments in order to identify the best feature representation or machine learning algorithm. Random forests proved to be the most convenient choice in addressing nonlinearity as well as handling a mix of different (numeric) feature types. Preliminary experiments showed that they performed best overall (though linear SVMs and comittees of perceptrons came close), and we report on experiments with random forests only in our evaluation experiments.

Briefly, for our experiments, a random forest is the sum of the scores of k decision trees, where each decision tree is trained on a bootstrap sample of the training fold. At each tree level a random feature is chosen and the best single partitioning value for that feature (minimizing the entropy) is chosen to partition the data at that node. Partitioning is done until all instances at a node have the same label. No pruning is performed. We used a forest of thousand trees in our experiments. Each experiment (training and testing) on a given annotator's data took no more than a few minutes.

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. Targeted retrieval is hence much simplified. Since the features used to determine familiarity level aren't necessarily limited to any specific terms, even documents that don't contain clues in the form of specific terms can be classified.

When used for re-ranking the search results, it can only increase precision at the top of the list. Improvements in recall are not possible. However, since documents can be assigned familiarity levels at indexing /crawl time, in addition to the scalability benefits, the potential recall problem can be addressed too. FAMCLASS can be seamlessly integrated into current search engine technology without involving any major modifications to existing architectures.

	p@1		p@5		p@10	
Annotator	Base	RL	Base	RL	Base	RL
Micro Avg.	0.558	0.538	0.540	0.544	0.498	0.508
Combined	0.565	0.565	0.536	0.536	0.5	0.527

Table 7: Baseline and classifier based on reading level features. Combined refers to the case when we pooled data from all annotators for training. Micro averaging averages results for individual annotator specific classifiers, and was done to provide a better comparison point for the 'combined' results. There was no statistically significant difference between the results. This means that reading level alone does not determine whether a document is introductory or advanced.

	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 9: 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.

5. EVALUATION

People expect to find the information they are looking for in the first page of search engine results i.e. within the first ten documents. This holds true for a familiarity-flavored search too. Thus FAMCLASS should be able to re-rank the results obtained from an ordinary search so that documents of desired familiarity are at the top.

To compare our approach against the baseline, we measured the proportion of introductory documents at ranks one (p@1), five (p@5), and ten (p@10)

to the choice of normalization for each feature, as this affects the dot product operation, while others, such as decision tree inducers, are not as sensitive.

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

Table 8: Baseline performance versus classifier performance using all features (ALL), only stop word features (StW), and only non-textual features(NtF). 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 classifier results we report are all based on *leave-one-query-out* (loqo) validation: we partition the labeled documents based on query, and for every fold we hold out all the documents associated with that query, train on the remainder and test on the hold-out. We do leave-one-query-out as opposed to standard cross-validation as otherwise, similar documents (returned for the same query) could end in both training and test folds in regular cross-validation, leading to misleadingly good results. Indeed, a few regular cross-validation experiments (not reported) yielded much better performance. We report performance when we train for each annotator independently (*micro-averaged*), and also when we pool labeled documents from all annotators together (*combined*). 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. We used labeled documents obtained from both trigger word and non-trigger word queries to train the classifier in each fold, but measured performance only on the documents from non-trigger queries, in order to directly compare against the baseline (default search engine ordering). This yields 254, 381, and 233 test documents respectively for annotators 1 to 3. Note that a random ranking would on average have the same average precision at 1, 5, and 10 as the average proportion of introductory documents, given in table 3.

Entries are in boldface whenever the corresponding number of wins (queries for which the classifier gave a higher precision number than baseline) was significantly higher than the losses according to the paired sign test at $p = 0.05$ level. Table 8 consolidates the results for classifiers trained on the different subsets of features. For the "combined" classifier (last row of table), we remove the 2 queries in common to the annotators (since they may disagree on some of those documents). This gives 46 queries on which we can compare the combined classifier against the baseline.

We observe that the micro-average performance of the classifier, when the classifier is trained on all features, is significantly higher than the baseline at p@5 and p@10. The per-user trained classifier performance results are at least as good as those for the combined classifier, even though there are significantly more training documents available for the latter, as can be seen in Table 9.

Recall that the pair-wise estimated inter-rater disagreement rate is roughly 30% on average, which may explain why the combined classifier is not performing significantly better than the individual classifiers. The accuracy of the combined classifier, using all features, was roughly 66% (averaged over 52 total queries), already close to the inter-rater agreement rates. However, there is potentially more room to improve per-user trained classifiers, pointing to personalization opportunities. The accuracies of FAMCLASS trained per annotator were, respectively, 60%, 68%, and 65%.

Table 7 reports results comparing baseline performance and a classifier learned using the 3 reading level measures as the only features. We can observe that a classifier based on reading level alone did not provide statistically significant improvements against the baseline. Recall from Table 8 that the performance of stop-word features appear to provide most of the mileage. Reading level features give the poorest performance, validating our claim that familiarity required to read a document is complementary to reading level.

5.1 Discussion of Important Features

Stop-words appeared to be the most important features in our classifier, when we broke down the feature sets into the three sub-types. In an attempt to characterize the individual features that contributed the most to the classifier's decision, we trained a linear classifier[3] and examined the coefficients of the features. The performance of the linear classifier was competitive with the performance of random forests. Therefore, high coefficient weight (in magnitude) in the classifier is suggestive of the importance of the feature in the classification task, with the sign of the coefficient being indicative of the class (*i.e.*, introductory vs. advanced). In Table 5.1 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.

6. RELATED WORK

[3] We trained a committee of perceptrons each initialized with a different set of weights, "bag-of-perceptrons", and then added the perceptrons' feature weights and thresholds to obtain one linear classifier.

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

Table 10: The stopwords with highest coefficients across multiple training runs of a linear classifier for introductory and advanced documents.

Personalized information systems [6] based on information filtering aim to provide users with relevant information based on their profiles. These models are aimed at customizing based on a topic-model for users. The task we are considering is based on a property of documents, independent of the user and the query. This means it is applicable across all documents, and across all queries, even for users we have no prior model of.

Wolfe et al [15] showed that learners benefit from documents with content just a little more advanced than their current level of knowledge. Similarly children learning to read benefit from texts with a small percentage of words outside their current vocabulary [4]. These approaches require a model of the user so that appropriate content can be delivered to them. Kelly and Cool [8] reported on the relationship between topic familiarity and information search behavior. They concluded that information searching behaviors like reading time and search efficacy tended to improve with topic familiarity.

Intuitively, the topic familiarity required to read a document with ease differs from reading level. Reading level reflects vocabulary acquisition and mastery of grammatical constructs, and is tailored to measuring children's ability to read. The Flesch-Kincaid reading level score [9] and the SMOG reading level index [12] measure reading level in terms of the average number of syllables per word, and the average number of words per sentence. For topic familiarity level, however, our target population is adults. Even with fixed vocabulary size and grammatical understanding, we would expect varied level of familiarity across topics.

Liu et al [10] conducted experiments on predicting the reading level of queries. The features they used for their classifier include sentence length, average number of characters per word, percentage of part-of-speech tags, readability indices, and frequency of unigrams, bigrams and trigrams. They were able to classify queries with greater than 80% accuracy on two-class problems using an SVM, outperforming standard reading level metrics which had accuracies from 10 - 20% on the same data sets. They also showed that search results vary in their grade appropriateness, which suggests that classifying both queries and search results into grade level could be useful for customizing search results for children. In our work we perform classification only on documents, leaving a study of classification of the familiarity level connoted by queries for future work.

Harper et al's HARD 2004 work on familiarity [7] was based on the hypothesis that users unfamiliar with a topic prefer documents with *representative* terms while users familiar with a topic prefer documents with *highly discriminative* terms. Such terms are identified using the clarity measure [5]. For every topic, the top-ranking K documents were considered relevant, and the terms in them were sorted by clarity scores. By interpreting the clarity score in a particular way, representative and discriminative terms were selected, and used to modify the query. They found that using this measure when high topic familiarity documents were requested gave significant improvements against the baseline. However, they were not able to improve results for introductory documents (for queries tagged with meta-data indicating the searcher has little background knowledge on the topic). Our approach to identifying introductory documents may prove complementary.

7. CONCLUSIONS

We have shown that by decoupling familiarity from relevance, defining it as a property of the document independent of query and user, we have made a task that users can agree on. This is a challenging task for a classifier. It is not predicted by reading level, as we showed in Table 7 in Section 5. Additionally, in our initial experiments, query expansion methods adding introductory terms found in the initial retrieval did not improve the proportion of introductory results returned. Furthermore, it is also challenging to develop a method that is topic-independent.

We collected a considerable amount of training data, and with a rich set of features and an advanced classifier, we were able to re-rank the documents, producing a statistically significantly higher proportion of introductory documents at 5 documents retrieved and at 10 documents retrieved, over a baseline search engine retrieval. This kind of topic-independent, user-independent classifier is empowering for personalized search, as with a single change to the retrieval reranking, any user can specify whether they want introductory or advanced documents for any query.

It may be useful to expand the set of features under consideration. For example, adding in features such as clarity [5] or idf which are indicative of the *rarity* of the terms in the document may help identify advanced documents. It is also important to explore the relationship between relevance and document familiarity. Evaluating how useful the documents are for task-completion may help address this issue.

Acknowledgments

We thank Prof. James Allan and Jean Joyce at the University of Massachusetts Amherst for their support with the user study. This work was supported in part by the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the sponsor.

8. REFERENCES

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

[2] S. Argamon, M. Koppel, J. Fine, and A. Shimoni. Gender, genre, and writing style in formal written texts. *Text*, 23(3), 2003.

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

[4] R. P. Carver. Percentage of unknown vocabulary words in text as a function of the relative difficulty of the text: Implications for instruction, 1994.

[5] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 299–306. ACM Press, 2002.

[6] P. W. Foltz and S. T. Dumais. Personalized information delivery: an analysis of information filtering methods. *Commun. ACM*, 35(12):51–60, 1992.

[7] D. J. Harper, G. Muresan, B. Liu, I. Koychev, D. Wettschereck, and N. Wiratunga. The Robert Gordon University's HARD Track Experiments at TREC 2004. In *The Thirteenth Text REtrieval Conference (TREC 2004) Notebook*, Gaithersburg, MD, USA, November 2004.

[8] D. Kelly and C. Cool. The effects of topic familiarity on information search behavior. In *JCDL '02: Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, pages 74–75. ACM Press, 2002.

[9] 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.

[10] X. Liu, W. B. Croft, P. Oh, and D. Hart. Automatic recognition of reading levels from user queries. In *Proceedings of the Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*, pages 548–549, Sheffield, UK, July 2004.

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

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

[13] F. Mosteller and D. Wallace. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. 1964.

[14] K. Ryan. Fathom, 2004. http://search.cpan.org/ kimryan/Lingua-EN-Fathom-1.08/Fathom.pm.

[15] M. Wolfe, M. Schreiner, R. Rehder, D. Laham, P. Foltz, T. Landauer, and W. Kintsch. Learning from text: Matching reader and text by latent semantic analysis, 1998.