Recommending Documents for Complex Question Exploration by Analyzing Collective Browsing Behavior

Alya Asarina
MIT Lincoln Laboratory
Lexington, MA, USA
alya@mit.edu

Olga Simek
MIT Lincoln Laboratory
Lexington, MA, USA
osimek@ll.mit.edu

I. INTRODUCTION

Most information retrieval tools available today require the user to precisely specify the information they’re looking for. We’re interested in developing systems that actively help connect users to the information they need by recommending documents of interest. Recommendation systems for other types of content (movies, products, etc.) are widely used today. Recommending documents is especially valuable when the user is trying to answer broad, challenging questions, and is uncertain as to what relevant information may be available to them.

In our work, we focus specifically on the information needs of intelligence analysts, though our approaches apply to the needs of the general public as well. Intelligence analysts are constantly looking for answers to challenging questions. Moreover, they generally work with private information repositories, and the tools for finding information within those repositories lag far behind what is available for the open internet. On the other hand, analysts are experts in searching for information with the tools that are available to them. Our goal is to take advantage of that expertise by developing algorithms that capture human insight as to what information is relevant.

Based on a relevant information browsing data set we collected, we developed novel algorithms to predict what documents users will find useful, and what information browsing paths they will follow. In addition to features of individual documents (e.g. word count, number of entities, etc.), we used similarities between pairs of documents to make the predictions. We used various dimensionality reduction techniques to represent each document using hundreds, as opposed to tens of thousands, of dimensions, and demonstrated that our methods for predicting browsing behavior and document usefulness outperform baseline approaches.

II. RELATED WORK

The rapidly growing amount of information available on the internet and in other digital repositories poses a serious challenge for information retrieval systems. User modeling, context analysis and personalization are becoming crucial to effective information retrieval. Recommender systems guide users to information and products specifically of interest to them from the large space of options.

Our ultimate goal is to analyze user behavior in order to develop a personalized recommender system for documents. Recommender systems are in wide commercial use today. Their applications range from products [1] and movies [2] to search results [3] and friends [4]. These systems rely on large amounts of data to make inferences about individual users – Netflix, for instance, released a dataset that included 100 million movie ratings from 480 thousand users. Our problem domain differs in that we need to make predictions based on data sets that are orders of magnitude smaller than those generally available for the development of commercial recommender systems.
Earlier work on recommender systems can be classified into three categories: content-based, where the user will be recommended items similar to the ones the user preferred in the past, collaborative, where the user will be recommended items that people with similar tastes liked in the past, and hybrid, which combines the previous two approaches [5]. More modern approaches include context-aware approaches, which utilize information about user’s situation and environment [6], semantic-based approaches, which typically take advantage of taxonomies or ontologies [7], peer-to-peer approaches, where a peer relates to a group with similar interests and obtains recommendations from the users of that group [8], and cross-lingual approaches [9].

We use a novel semantic approach and dimensionality reduction techniques to create a hybrid recommender system. Our approach incorporates the use of organizational knowledge in the form of browsing history and a novel question answering data set we collected to prioritize documents when the user is interested in exploring complex issues. Our specific challenges include the use of limited browsing history data and our particular task of recommending documents for challenging question answering since we are not interested in just presenting documents similar to the ones the user has viewed in the past. There have been various approaches proposed in the literature for recommending novel items of interest to the user rather than just similar items. These include introduction of randomness, designing similarity cutoffs and excluding items that are too similar, use of redundancy measures, and anomaly search [10]. These existing approaches don’t quite apply to our case since our task is different from simply presenting novel items of interest. In our case, relevant browsing history for concepts similar to user interest is used instead.

Our paper is organized as follows: Section I introduces the problem. Section II discusses relevant work. Section III describes the dataset collection. Section IV explains the algorithmic approach. Section V presents the experiments and the results, and Section VI concludes and discusses future directions.

III. USER STUDY

In order to develop and evaluate algorithms for document recommendation based on browsing behavior, we needed a data set that includes (a) individual browsing data, and (b) the text of the documents viewed. We were able to identify only a single publicly available data set that includes both types of data [11], and this data set is unfortunately not well suited to our analysis requirements.

We therefore conducted a study to collect the right data for developing our algorithms. We asked participants to find resources online that would be relevant to the questions shown in Table I.

Each participant worked on one question at a time, but participants could choose to answer multiple questions over the course of a data collection period of about a month. All participants were asked to spend no more than 7 minutes on each page viewed. They were also asked to assign usefulness ratings to the pages they viewed on a 1–5 scale in a bar inserted at the bottom of the page. By using an HTTP proxy, we collected information about the pages viewed by study participants, including the content of the pages and the time when each page was viewed.

Participants were instructed that their goal was to find resources online that would help with answering the question they selected. Participants were also asked to assign usefulness ratings to the pages they viewed on a 1–5 scale in a bar inserted at the bottom of each page, but they were not required to rate each page.

By using an HTTP proxy, we collected information about all the pages viewed by study participants, including the content of the pages and the time when each page was viewed.

Table II summarizes the data we collected. For our analysis, we focused on the pages that were assigned ratings by the participants. Note that the size of our data set is somewhat larger than the set collected by [11], who gathered 666 rated pages from the Web. A key difference between our study and [11] is that while for our tasks, participants rated an average of 8.3 pages, for [11] the average is just 3.2 rated pages. The absence of a time limit and a more challenging question selection thus enabled us to obtain longer browsing histories.

Figure 1, which is described in Table III provides an overview of the browsing histories of the study participants. This graph highlights the fact that different users chose very different information paths. The great majority of pages were rated by just one participant.

Figure 2 provides another overview of the data set we collected, as described in Table IV. Again, in this graph, each node corresponds to a rated document from our data set. The weights of the edges are based on a measure of content similarity.

---

### Table I. Browsing Behavior User Study Questions

<table>
<thead>
<tr>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Insider Threats: What are the most consistent and identifiable char-</td>
</tr>
<tr>
<td>acteristics displayed by potential insider threats to defense agencies?</td>
</tr>
<tr>
<td>(e.g. CIA, NSA, ...)</td>
</tr>
<tr>
<td>2) Al-Qaeda’s Relevancy: Is Al-Qaeda Central still relevant?</td>
</tr>
<tr>
<td>3) Russian Economic Integration: What is the best way to integrate</td>
</tr>
<tr>
<td>Russian economy into the global economy?</td>
</tr>
<tr>
<td>4) Moderate Leadership</td>
</tr>
<tr>
<td>a) Leadership in Palestine: Which moderate leaders should the U.S.</td>
</tr>
<tr>
<td>support in Palestine?</td>
</tr>
<tr>
<td>b) Leadership in Egypt: Which moderate leaders should the U.S. support</td>
</tr>
<tr>
<td>in Egypt?</td>
</tr>
<tr>
<td>c) Leadership in Syria: Which moderate leaders should the U.S.</td>
</tr>
<tr>
<td>support in Syria?</td>
</tr>
<tr>
<td>5) Chinese Policy</td>
</tr>
<tr>
<td>a) Chinese Policy in Tibet: What are the key policies in Tibet and how</td>
</tr>
<tr>
<td>will they impact stability in the next five years?</td>
</tr>
<tr>
<td>b) Chinese Policy in Xinjiang: What are the key policies in Xinjiang</td>
</tr>
<tr>
<td>and how will they impact stability in the next five years?</td>
</tr>
<tr>
<td>c) Chinese Policy in Hong Kong: What are the key policies in Hong Kong</td>
</tr>
<tr>
<td>and how will they impact stability in the next five years?</td>
</tr>
</tbody>
</table>

---

### Table II. Data Collected in User Study

<table>
<thead>
<tr>
<th>Data</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td># questions</td>
<td>9</td>
</tr>
<tr>
<td># browsing paths</td>
<td>90</td>
</tr>
<tr>
<td># content pages</td>
<td>2973</td>
</tr>
<tr>
<td># search result pages</td>
<td>481</td>
</tr>
<tr>
<td># pages rated (1–5 scale)</td>
<td>744</td>
</tr>
</tbody>
</table>

---

2 Participants in the study were asked to spend no more than 7 minutes to answer highly specific questions (e.g., “What is the average temperature in Springfield, IL for winter?”), whereas we are interested in the browsing patterns of users with more complex information needs.

3 This number excludes redirect pages and second visits by the same participant.
similarity between each pair of documents. Specifically, we use term frequency-inverse document frequency (tf-idf) cosine similarity, which is described in Section IV, with a minimum weight cutoff of 0.1. The layout was generated using the Force Atlas 2 algorithm provided by the open source Gephi ([13]) graph visualization tool. This view of the document collection shows that document similarity reflects intuitive relationships between documents. Documents related to the same question cluster together. Furthermore, clusters for related questions (e.g. the three questions about Chinese policy) are grouped together as well. In the following section, we discuss our use of document similarity measures as a tool for predicting document relevance.

A. Data Pre-Processing

In our data collection setup, each HTTP GET request is processed separately, with the result that what appeared as a single page to a study participant is saved as multiple pages in our database. To address this issue, we merged pages which were associated with the same time stamp. We filtered out any downloaded content that is blocked by the EasyList4 of ad domains, as ad content is unlikely to be relevant for the research tasks in our study.

IV. METHODOLOGY

Our overall objective was to develop approaches for recommending relevant documents based on past information browsing behavior. In order to do so, we can use machine learning techniques to combine various types of document features that may be predictive of document relevance. The challenge is to design features that will be effective in distinguishing relevant documents from ones that are not relevant. The features we developed fall into two categories: properties of individual documents, and measures of similarity between pairs of documents.

The basic properties of individual documents we used are shown in Table V. We included features that intuitively indicate documents that are informative (e.g. word count) and specific (e.g. entity counts, list counts).

The other type of features we used was measures of similarity between pairs of documents. The basic idea is to represent each document as a feature vector, and to calculate similarity between documents based on those feature vectors. The set of pairs of documents being compared depended on the specific prediction task. For example, when predicting page usefulness, we compared the page under consideration to other pages viewed by study participants working on the same question.

We use cosine similarity as our metric, where cosine similarity between vectors $A$ and $B$ is defined by

$$\frac{A \cdot B}{||A|| \cdot ||B||}$$

(1)

The key question that arises is how to represent documents as feature vectors. A standard representation found in the information retrieval and text mining literature is term frequency-inverse document frequency (tf-idf), where each document is represented as a sparse vector of term frequencies scaled by

4https://easylist.adblockplus.org/en/
the inverse document frequencies of the terms. We make use of tf-idf representations, but their high dimensionality (equal to the size of the vocabulary across all documents) is a challenge for generalizability. Because the numbers of observations in the labeled data sets we are working with measure in the hundreds or low thousands, we cannot discover the right generalizations in data with tens of thousands of dimensions.\(^5\) In order to address this difficulty, we use several dimensionality reduction approaches, which we discuss in the following section.

### A. Approaches to Dimensionality Reduction

Given a sparse word vector representation for each document, our goal is to find lower-dimensional document representations that will preserve the relevant information encoded in the high-dimensional representations. Since we are using document representations to capture content similarity between documents, we need representations that continue to reflect the content of the documents. Because different representations will preserve somewhat different information, we use several document representations in our model, as listed in Table VI.

Topic modeling jointly models the documents in the corpus of interest and the words they contain. The other approaches in Table VI use separate corpora to derive word representations in an unsupervised way. These representations generally aim to map words judged to be similar by people to vectors that are close (e.g., using a cosine similarity metric) in the mapping space. The representation for each word thus captures semantic information about that word. Each document can then be represented as the sum of the representations for the words in that document.

For details on topic representations, as well as representations 3–6, we refer the reader to the papers cited above. In the following section, we discuss the word representations we developed using Canonical Correlation Analysis (CCA). Section IV-C discusses how we also applied CCA to reduced-dimensionality document representations to derive an additional similarity metric.

### B. Unsupervised Word Representation Learning with Canonical Correlation Analysis (CCA)

Canonical correlation analysis (CCA) \([21, 22]\) is a statistical method for finding linear projections of two random vectors such that their correlation is maximized. CCA has in recent years been used for a variety of applications. For example, Hardoon et al [15] used CCA to learn a semantic representation of web images and their associated text. Faridani [23] applied CCA to derive a model that combines text and numerical ratings on websites like Tripadvisor and Zappos and used it to predict numerical ratings from the text. And Rousus et al [24] employ canonical correlation analysis to find multivariate correlations between the ’omics measurements and the complex clinical phenotypes.

We now briefly describe CCA. Let us consider two multivariate random variables \(x\) and \(y\) with zero mean and assume that they are correlated (zero mean assumption can be relaxed as any feature vector can be transformed by subtracting the mean). Denote each sample observation of \(x\) as \(x_i\). For \(n\) samples we form two vectors \(S_x = [x_1, \ldots, x_n]\) and \(S_y = (y_1, \ldots, y_n)\). We now consider two linear directions \(w_x\) and \(w_y\) but we add the assumption that they have the same column rank such that \(<w_x, x_i>\) and \(<w_y, y_i>\) are projections in the same-dimensional vector space. Here \(<w_x, x_i>\) is the inner product of vectors \(w_x\) and \(x_i\) and equals \(w_x^T x_i\) (\(w_x^T x_i\) is the transpose of \(w_x\)). We can now look at elements of \(S\) in the new coordinate system; thus we will have

\[
S_{x, w_x} = (<w_x, x_1>, \ldots, <w_x, x_n>)
\]

\[
S_{y, w_y} = (<w_y, y_1>, \ldots, <w_y, y_n>)
\]

CCA seeks to maximize the correlation between \(S_{x, w_x}\) and \(S_{y, w_y}\); thus the goal is to find \(w_x\) and \(w_y\) such that the following objective function (\(\rho\)) is maximized.

\[
\rho = \max_{w_x, w_y} \text{corr}(S_{x, w_x}, S_{y, w_y}) = \max_{w_x, w_y} \frac{S_{x, w_x}^T S_{y, w_y}}{|S_{x, w_x}| |S_{y, w_y}|}
\]

Hardoon et al show that the optimization problem of finding \(w_x\) and \(w_y\) linear transformations can be reduced to a symmetric eigenproblem and can be solved by a Cholesky decomposition [15]. \(S_x\) and \(S_y\) in this case are now the canonical representations of \(x\) and \(y\). Meaning that \(S_x\) can be considered as the expected value of some latent variable \(z\) given \(x\), \(E(z|x)\). Similarly \(S_y\) is the expected value of the latent variable \(z\) given \(y\), \(E(z|y)\).

We applied CCA in a novel way to create word representations that are expected to preserve the document meaning. Consider a bag-of-words representation for a document. We would expect that a randomly selected half of the words in the document would capture a good deal of the relevant information about it. Moreover, approximately the same information should be captured by the other half of the words in the document. Motivated by this idea, we use CCA to learn projections that preserve semantic content by mapping the two halves of each document in a large corpus to points that are close together in the lower-dimensional space.

We use Gigaword ([25]), a corpus of 4.1 million English-language newswire documents, to train our CCA model. The set of words in each document is randomly split into two equal-size subsets, \(x_i\) and \(y_i\). The \(x_i\)’s and \(y_i\)’s are represented as sparse word vectors, and we use CCA to find projections of the \(X = [x_i]\) and \(Y = [y_i]\) matrices that maximize the correlations between the \((x_i, y_i)\) pairs.

We find the top 500 projections to obtain 500-dimensional representations. We can then apply one of the CCA projection matrices (given that the splits were random, it does not matter whether we use the “left” or “right” matrix) to word vector representations of the documents in our data set in order to obtain reduced-dimensionality representations.

---

\(^5\) There were 35,472 unique word in our data set.
TABLE VII. SPEARMAN CORRELATION BETWEEN WORD REPRESENTATIONS AND HUMAN-ANNOTATED WORD SIMILARITY DATASETS

<table>
<thead>
<tr>
<th>Representation</th>
<th>WordSim-353</th>
<th>SCWS*</th>
<th>Rare Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>HSIN ([18])</td>
<td>0.626</td>
<td>0.321</td>
<td>0.020</td>
</tr>
<tr>
<td>Modified HSIN ([19])</td>
<td>0.646</td>
<td>0.437</td>
<td>0.223</td>
</tr>
<tr>
<td>C&amp;W ([20])</td>
<td>0.498</td>
<td>0.486</td>
<td>0.268</td>
</tr>
<tr>
<td>Modified C&amp;W ([19])</td>
<td>0.570</td>
<td>0.485</td>
<td>0.344</td>
</tr>
<tr>
<td>Our CCA Representations</td>
<td>0.442</td>
<td>0.346</td>
<td>0.277</td>
</tr>
</tbody>
</table>

TABLE VIII. WORD SIMILARITY DATASETS

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Word Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. WordSim-353 ([26])</td>
<td>353</td>
</tr>
<tr>
<td>2. SCWS* ([19], based on [18])</td>
<td>1762</td>
</tr>
<tr>
<td>3. Rare Words ([19])</td>
<td>2034</td>
</tr>
</tbody>
</table>

1) Evaluation of CCA Word Representations on Similarity Task: In order to verify that the CCA word representations we computed captured some intuitive notion of word similarity, we evaluated these representations on the word similarity task described by [19]. We compared the similarity between our word representations (using the cosine similarity metric) to human-annotated word pair similarity datasets. Our results, as well as the results for other representations reported by [19], are shown in Table VII.

We used three datasets for the evaluation, as described in Table VIII. Despite the fact that our CCA representations were not designed specifically for this task, their performance was right in the ballpark of other word representations found in the literature. This confirms that similar representations correspond to intuitively similar words.

C. CCA as a Second Dimensionality Reduction Step

There is another way in which CCA can be used as a method to capture patterns in our data. Based on the techniques discussed above, we can represent each document in the dataset as an 824 dimensional vector. 14 of the features encode basic properties of the document (see Table V), while the remaining 810 come from the six reduced-dimensionality document representations listed in Table VI. We can use CCA to further project the 824-dimensional representations into a smaller number of dimensions in a way that captures the relevant generalizations about our data.

Our aim is to capture similarities between documents in our data set. Therefore, we can use pairs of documents \((d_i, d_j)\) as inputs to applying CCA, with the selection of pairs depending on the specific prediction task. We use pairs \((d_i, d_j)\) that are positive examples of what we're looking for, e.g. documents that are both useful and relevant for some question; the precise set of document pairs depends on the specific prediction task. The learned projections will then maximize the correlations between representations for documents whose similarity we want to capture.

We always restrict \(d_i, d_j\) to being documents that are in the training set for the task we're evaluating algorithms for, not in the testing set. We experimented with various numbers of dimensions in the reduced representations, and found empirically that representations up to 100 dimensions or so generalize well to the testing set, while higher dimensional representations are overfitted to the training data. In addition to computing standard cosine similarity on the document representations derived with this CCA approach, we also computed a cosine similarity measure with each dimension scaled by the corresponding canonical correlation value \(r\) derived from CCA.

V. EXPERIMENTS AND RESULTS

Our broad objective is to predict what documents will be useful to someone in the context of particular question. With this in mind, we designed three prediction tasks with the goal of identifying which documents study participants chose to look at, and which documents they found to be useful. Broadly speaking, the tasks we considered were:

- **Usefulness prediction**: Based on some initial usefulness rating information, which additional pages will be rated as being useful?
- **Browsing path prediction**: Based on the pages a study participant has looked at so far, which pages will he choose to look at in the future?
- **Usefulness based on browsing path**: Based on the sequence of pages a study participant has looked at so far, will he rate the next page as being useful or not?

In this section, we lay out the specific details of these tasks, and present our results. As discussed in Section IV our approaches combined individual document features with measures of document similarity based on document representations with reduced dimensionality. We compare our results on these tasks against two baseline approaches:

1) **Individual document features**: Prediction accuracy using only the features of individual documents described in Table V.
2) **Term frequency-inverse document frequency (tf-idf) similarity**: Prediction accuracy using only tf-idf similarity based on high-dimensional word vector representations of documents, with no dimensionality reduction.

We show that our methods significantly outperform the baselines on all three tasks. We calculated nine measures of similarity: tf-idf cosine similarity on sparse word vector representations, values corresponding to the six reduced-dimensionality representations listed in Table V, and two similarity based on representations using CCA as a second dimensionality reduction step (see Section IV-C).

A. Usefulness Prediction

For this task, our goal was to predict whether each rated page in our collection was assigned a positive rating (3, 4 or 5) or a negative rating (1 or 2) (see Figure 3). We conducted our evaluation using 10 random 90%-10% train-test splits of the 744 rated documents we collected. For the document similarity attributes (see Section IV), for each document, we calculated its average similarity to three sets of documents using each of the representations listed in Table V, as well as high-dimensional tf-idf representations:
TABLE IX. EVALUATION RESULTS

<table>
<thead>
<tr>
<th>Task</th>
<th>Usefulness</th>
<th>Browsing Path</th>
<th>Usefulness from Browsing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc Features</td>
<td>0.59</td>
<td>N/A</td>
<td>0.61</td>
</tr>
<tr>
<td>Tf-idf cossim</td>
<td>0.61</td>
<td>0.54</td>
<td>0.55</td>
</tr>
<tr>
<td>Our algorithm</td>
<td>0.64</td>
<td>0.78</td>
<td>0.66</td>
</tr>
</tbody>
</table>

- All the training set documents viewed by study participants working on the same question.
- Positively rated training set documents viewed by study participants working on the same question.
- Negatively rated training set documents viewed by study participants working on the same question.

The intuition is that documents that are similar to other documents people chose to view when working on the same question, and especially the documents they rated positively, are more likely to be relevant and informative for that question. Similarity to negatively rated documents may indicate that a document is less likely to be useful.

We used a Bayesian network classifier (implemented by the free Weka ([27]) tool) to classify useful documents versus ones that are not useful. The results of the evaluation are shown in Table IX. Our approach significantly outperforms the baselines (\(p < .05\)).

B. Browsing Path Prediction

For the browsing path prediction task, our goal was to understand how study participants navigated information. In particular, given that a participant has viewed document \(d_i\), we wanted to predict whether he would also view document \(d_j\), as illustrated in Figure 4. The positive examples were thus all ordered pairs of documents \((d_i, d_j)\) such that some study participant viewed first \(d_i\) and then \(d_j\) when working on some question. For the negative examples, we constructed pairs of documents \((d_i, d_j)\) such that both \(d_i\) and \(d_j\) were viewed by study participants working on the same question, but no participant first viewed \(d_i\) and then viewed \(d_j\).

Because this task involves making predictions about pairs of documents, rather than individual documents, we did not use document features in our model. For each input pair, we computed similarity features using tf-idf on sparse vector representations, as well as all the reduced-dimensionality representations discussed in Section IV.

There were 3,586 positive examples in our data set, and we generated a random set of 3,586 negative example for the evaluation. We trained a random forests classifier, which we empirically found to be effective for this task. The evaluation was conducted using ten random 90%-10% train-test splits, and our approach significantly outperformed the baseline (\(p < .05\)). Note that, because this task makes predictions about pairs of documents, the document features baseline is not applicable.

C. Prediction of Usefulness from Browsing

With our final task, we wanted to model the recommendation system context as closely as possible for a prediction task based on the data set we collected. Our goal was to use an individual study participant’s browsing behavior to predict whether that participant would find a particular page useful for the question he’s working on, as illustrated in Figure 5. The features we used to predict document usefulness were thus the features of that document (entity counts, etc.) together with measures of similarity between that document and the documents previously viewed by the study participant. We expect that a document is more likely to be found useful if it is similar to other documents the same user chose to view in the course of his research. We used each of the representations listed in Table V, as well as high-dimensional tf-idf representations, to calculate similarity.

In order to make the setting realistic, we reserved the last document viewed by each participant for each question for the test set, and used the other documents as training data. This produced a single split of the data, with 5938 training data points and 1208 testing data points. We subsampled the negative examples to create a balanced data set. We empirically found a Bayesian networks classifier to be effective for distinguishing useful vs. non-useful documents, outperforming...
both baseline approaches. Our results are shown in Table IX, and are statistically significant (p < .05).

VI. CONCLUSION

We have presented a novel algorithm for recommending documents to users interested in exploring complex questions. Our approach takes advantage of human insight and information search expertise in order to identify useful documents. We conducted a user study to collect a document browsing dataset that captures human information navigation behavior. We then applied supervised machine learning algorithms together with a variety of semantic dimensionality reduction methods to predict which documents are useful.

Not only are dimensionality reduction techniques necessary to tackle data sparsity issues, they also allow us to create document similarity measures in diverse semantic spaces, which in turn let us recommend relevant documents based on much richer set of features than just bag-of-words similarity. We have demonstrated that our algorithm significantly outperforms baseline approaches on three different prediction tasks.

By using confidence scores produced by the document usefulness classifiers, we can create a recommendation system for documents in order to aid the exploration of challenging questions. Given an initial document selected by a user (e.g. through keyword search), the algorithm would recommend additional documents of interest. The recommendations would get more and more targeted as the user selects documents to read and the algorithm incorporated that information.

In future work, we plan to expand our browsing behavior dataset, which will allow us to utilize a broader range of document usefulness prediction techniques. We will be able to aggregate document ratings to produce ranked documents lists, and then take advantage of algorithms developed specifically for ranked list prediction, e.g. AdaRank [28] and SoftRank [29].

A larger data set will also allow us to further test the robustness of our algorithms, and to use standard metrics like Normalized Discounted Cumulative Gain (NDCG) to evaluate ranking approaches. We also plan to perform user-in-the-loop testing to fully assess the utility of our recommender system.

ACKNOWLEDGMENT

The authors would like to thank the New Technologies Initiative at MIT Lincoln Laboratory for funding this research, and our MIT LL group members Davis King, Garret Bernstein, Alex Broad and Mike Snyder for great suggestions, fruitful discussions and help with the data collection and processing.

REFERENCES


