Mining User Interest from Search Tasks and Annotations

Sampath Jayarathna, Atish Patra, and Frank Shipman
Computer Science & Engineering
Texas A&M University
College Station, TX 77843-3112 USA
{sampath,apatra,shipman}@cse.tamu.edu

ABSTRACT
Interactive web search involves selecting which documents to read further and locating the parts of the documents that are relevant to the user's current activity. In this paper, we introduce UIMaP: **U**ser **I**nterest **M**odeling **a**nd **P**ersonalization, a search task based personal user interest model to support users' information gathering tasks. The novelty of our approach lies in the use of topic modeling to generate fine-grained models of user interest and visualizations that direct user's attention to documents or parts of documents that match user's inferred interests. User annotations are used to help generate personalized visualizations for user's search tasks. Based on 1267 user annotations from 17 users, we show the performance comparisons of four different topic models: LDA+H, LDA+KL, LDA+JSD, and LDA+TopN.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Information filtering, Retrieval models*

Keywords
User Interest Modeling, Topic Models, Search Personalization

1. INTRODUCTION
Detailed knowledge about a user's interests is beneficial in web search, advertising, and personalized recommendations as well as in content targeting. The goal of personalized recommendations is to support users by identifying documents or the parts of a document that best match user's interests during an open-ended information gathering task. Such recommendations can result in a more efficient use of the user's time, e.g. that their time is spent on the most relevant documents.

Our past research shows that time is frequently a limiting factor in web search tasks: there are too many documents to assess and too much reading to do. The problem in such a search task is that even with the best web search engines, and the most effective query formulations, these tasks require people to work through long list of documents to examine potentially relevant documents or part of a document. Most users skim early documents, find portion of a document relevant to the current query, and determine additional information needs that result in further queries and more documents to process [1].

This paper describes the usage of user interest models using topic modeling as a basis for visualizations that draw a user's attention to similar documents or to a part of documents that match these interests. The paper describes the overall architecture of UIMaP, and the topic modeling based algorithms we developed for user interest modeling. Section 2 surveys related work and section 3 presents our system components and topic modeling algorithms. In section 4 we discuss the results of initial user evaluations, and section 5 presents conclusions and points out some future work.

CIKM'13, October 27–November 1, 2013, San Francisco, CA, USA

http://dx.doi.org/10.1145/2505515.2507878

2. RELATED WORK
Relevance feedback has a history in information retrieval systems that dates back well over thirty years and has been used for query expansion during short-term modeling of a users' immediate information need [2]. Explicit feedback requires users to assess the relevance of documents or portions of documents or to indicate their interest in certain aspects of the content (e.g. identifying nouns or phrases within search results). Explicit feedback has the advantages that it can be easily understood, is fairly precise and requires no further interpretation [3]. Annotations can be interpreted as one form of explicit feedback.

Reading documents happens for many reasons: we read for fun, for general knowledge, or for some specific activity. When reading as part of an activity, we have a particular task in mind. Not all reading results in annotations. Annotations are most likely when people read materials crucial to a particular task at hand and are infrequent when reading for fun [4]. Explicit interest indicators such as annotations are based on users directly identifying which documents or portions of it are interesting.

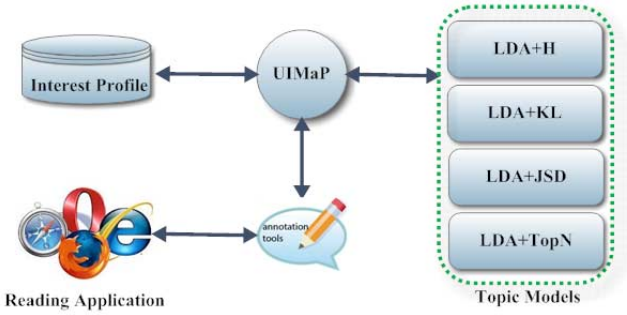

Figure 1. UIMaP: Overview and System Components

As users read through a particular document, they begin to identify the content relevant to the task in hand. If the document text content is large, users will frequently skim or stop reading when they feel what they have is good enough. Consequently, potentially better document contents are left having never been reviewed [5]. A potential solution for this particular problem is to provide visualizations to draw user's attention to similar documents or document parts relevant to their search task [1]. Users' attention to passages of potential interest can be drawn by using colors and icons to highlight them in a document overview application [6]. XLibris [7], and Spatial hypertext systems such as

VIKI[8] and VKB [9], use a similar visualization techniques to provide system-identified "interesting document contents" to provide visualization aided navigation.

3. USER INTEREST MODELING

Figure 1 shows the overall architecture of UIMaP. The reading application (web browser) communicates with the UIMaP via the browser plug-in annotation tool. The interest profile stores inferred user interests; records of user activity in reading application. UIMaP then drives the visualizations (system generated underlines of text content) of documents based on the inferred interests the topic models generated.

3.1 Explicit User Annotations

During information gathering search task, useful documents may be long, and cover multiple subtopics; users may read some segments and ignore others. In order to record which portions of the document pique the user's interests most, an explicit interest expressions capturing tool can be used. The WebAnnoate [1] provides basic annotation capabilities, collect data on user's interactions with web documents, and uses interest data returned from UIMaP to create visualizations(see Figure 2) that enhance document skimming and reading. With annotation tool, users can provide explicit feedback via annotations and convey it to UIMaP with terms associated with the annotation.

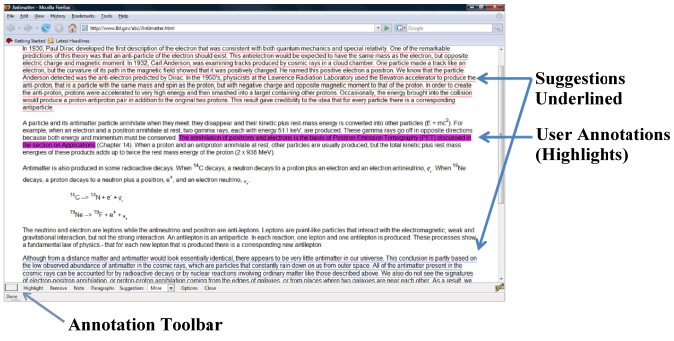

Figure 2. User Generated Annotations (highlights) and System-generated visualizations (underlines)

3.2 Interest Profile

The interest profile plays the central role in the UIMaP. It collects and stores information about interest related activity from document reading application and this information are processed to create a user interest profile based on UIMaP topic modeling algorithms. The UIMaP then estimates the user interest based on the inferred user interest profile and broadcast it to the document reading application to generate visualizations. Any application that can be modified to include the interest profile client software API can communicate with the UIMaP enabling multi-application user interest modeling capability.

3.3 Topic Models

Before introducing our topic model algorithms for inferring user interests, we first give a brief review of the statistical model Latent Dirichlet Allocation (LDA) and its parameters used in this research paper. LDA is a hierarchical probabilistic generative model which can be used to model a collection of documents by topics [10]. Given LDA parameters, a number of topics K, a document corpus of W distinct words, two smoothing parameters α and β, and prior distribution over document corpus, LDA can create random documents whose contents are a mixture of topics. As words are the only observable variables in an LDA model, conditional independence holds true for the outputs of LDA model which are document and topic distributions θ and $\emptyset$.

For a corpus containing D documents, the parameters θ, the $D \times K$ matrix of topic probability distribution per each document and $\emptyset$, the $K \times W$ matrix of topics must be learned from the data. The remaining parameters α and β, and K are specified by UIMaP. For the LDA models used in this paper, parameter fitting is performed using collapsed Gibbs sampling [11] to estimate θ , and $\emptyset$. We use $\alpha = 0.01$ and $\beta = 0.01$ [12]. Two additional parameters for the Gibbs sampling are the number of sampling and burn-in iterations, which we set to 1 and 5 respectively.

In our experiments with LDA models, we will create similarity matrices to compare the user-generated annotations (Source S) to document components (Target T); hence we define proposed measures as similarities. The granularity in this scenario is a paragraph/passage of the document.

3.3.1 LDA+Hellinger

The Hellinger distance is computed over two positive vectors. Since we are dealing with probability distributions in document-topic distribution, we chose hellinger distance [13] to measure their divergence. The main idea of our approach is to use the hellinger distance between document topic distributions to find the similarity of target T to the user generated source S.

$$D_{LDA+H}(S||T) = \sqrt{\frac{1}{2}\sum_{i=1}^{K}\left(\sqrt{s_i} - \sqrt{t_i}\right)^2} \quad \textbf{(1)}$$

where S is a K-dimensional multinomial topic distribution and s_i is the probability of the i^{th} topic.

3.3.2 LDA+KL

Kullback-Leibler divergence (KL divergence)[14] is a non-symmetric measure of the difference between two probability distributions. In our LDA+KL model, the association of source and target in the document topic distribution can be measured using the KL-divergence. The smaller the score is, the stronger the associated similarity is. For two probability distributions, from target to the user generated source, KL divergence is calculated as follows:

$$D_{LDA+KL}(S||T) = \sum_{i=1}^{K} s_i \log_2 \frac{s_i}{t_i} \quad \textbf{(2)}$$

3.3.3 LDA+JSD

We use Jensen-Shannon divergence (JSD) measure as a smoothed and symmetric alternative to the KL divergence. The measure is 0 only for identical distributions and approaches infinity as the two differ more and more. Formally it is defined as the average of the KL divergence of each distribution to the average of the two distributions [15].

$$D_{LDA+JSD}(S||T) = \frac{1}{2}D_{KL}(S||R) + \frac{1}{2}D_{KL}(T||R) \quad \textbf{(3)}$$

$$R = \tfrac{1}{2}(S+T)$$

3.3.4 LDA+TopN

The simplest way to support Top-N topic probabilities is to sort the resultant document topic probability distribution in the desired order and then discard all but the first N topic tuples. Then compare the Top-N topics between document topic distributions

to find the similarity of target T to the user generated source S. The main motivation behind this method is to find document-based results, such as finding main topics of a document or finding the top topics that are most related to a specific document content or user annotation.

Table 1. Confusion matrix for system evaluation

		User Generated	
		Annotated	**Not-Annotated**
System Generated	**Underlined**	tp	fp
	Not-Underlined	fn	tn

4. EXPERIMENTS

In this section we discuss user experiments we have done to evaluate our proposed methods. We first describe our evaluation metrics, and then experimental setup. Next we present the results from our user survey that measures the perceived quality of our user interest model.

4.1 Evaluation Metrics

How can we evaluate the effectiveness of our proposed methods? Given that our primary goal is to learn the user's preference from her explicit feedback and use these user generated annotation results to visualize relevant document content, we may consider the standard information retrieval domain evaluation metrics such as precision, recall, accuracy, F1 measure, false positive and true positive. Precision is the ratio of correctly underlined as a class to the total document content as the class. For example, the precision (P) of the underlined class in Table 1 is . Recall (R) is the ratio of correctly underlined document content as a class to the actual user generated annotations in the class. The recall of the underlined class in the table is . Accuracy is the proportion of the total number of underlines that were correct. The accuracy in the table is . F1 is a measure that trades off precision versus recall. F1 measure of the underlined class is .

4.2 Data and Setup

Since our approaches are based on annotated document contents, we need to collect user's annotations for a set of search tasks. In the meantime, users are required to supply a set of annotations using the annotation tool that reflects relevance to the main idea of the given search tasks. The data is composed of five search tasks and twenty web documents. Documents are preprocessed and removed graphics and annotations before experiments. We recruited 17 students to annotate the documents relevant to the given search tasks. Users are told to make annotations freely which reflects the main idea of the given task and relevance to the given documents. We collected total of 1267 annotations.

4.3 Experimental Results

It is important to identify the optimum number of topics in each LDA model as they determine the quality of the user interest modeling. We calculate the values at first 5, 10, 15, 20, 25 topics respectively. The results are shown in Figure 3. From these results, we first observe that the effect on the final performance is not quite large in all four models. K=8 gives the best average F1 measure for all four models.

We evaluate the results sensitivity to the similarity threshold in the LDA+H, LDA+KL and LDA+JSD. Figure 4 shows how the model threshold influences the performance. As the threshold increases from 0.1 to 0.5, the performance keeps on improving and reaches the average optimal value at 4.5 for all three models. Model LDA+TopN shows a similar trend and reaches optimum F1 measure at N=2.

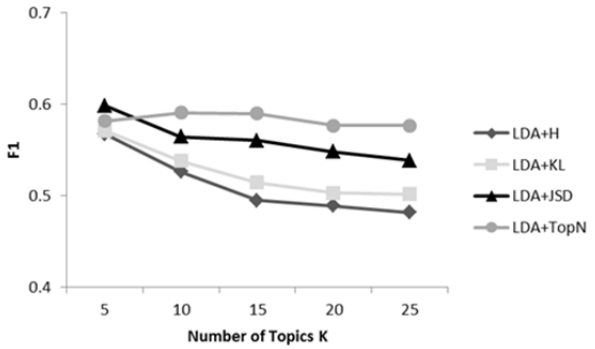

Figure 3. Impact of varying the number of latent topics

The Figure 5.(b) shows the overall performance of all four algorithms. The improvement on recall and F1 of LDA+JSD and LDA+TopN is very encouraging since recall is a more important factor in generating user interest models to provide relevant content as suggestions/recommendations. The results demonstrate that the LDA+JSD and LDA+TopN consistently outperform the other two methods in terms of hit recall and F1 measure. From this comparison, it can be concluded that the proposed approach is capable of making accurate and effective search suggestions.

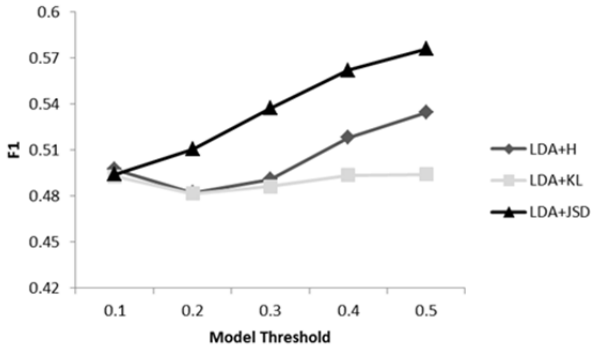

Figure 4. Impact of varying the threshold in topic models

5. CONCLUSION AND FUTURE WORK

This paper introduces UIMaP, a novel search task based user interest model based on user's annotations. Annotations are used to help generate personalized visualizations for user's search tasks. Four different topic models are produced: LDA+H, LDA+KL, LDA+JSD, and LDA+TopN. Performance comparisons between these four topic models are made. This paper also describes the usage of user interest models using topic modeling as a basis for visualizations that draw a user's attention to similar documents and to portions of documents that match these interests.

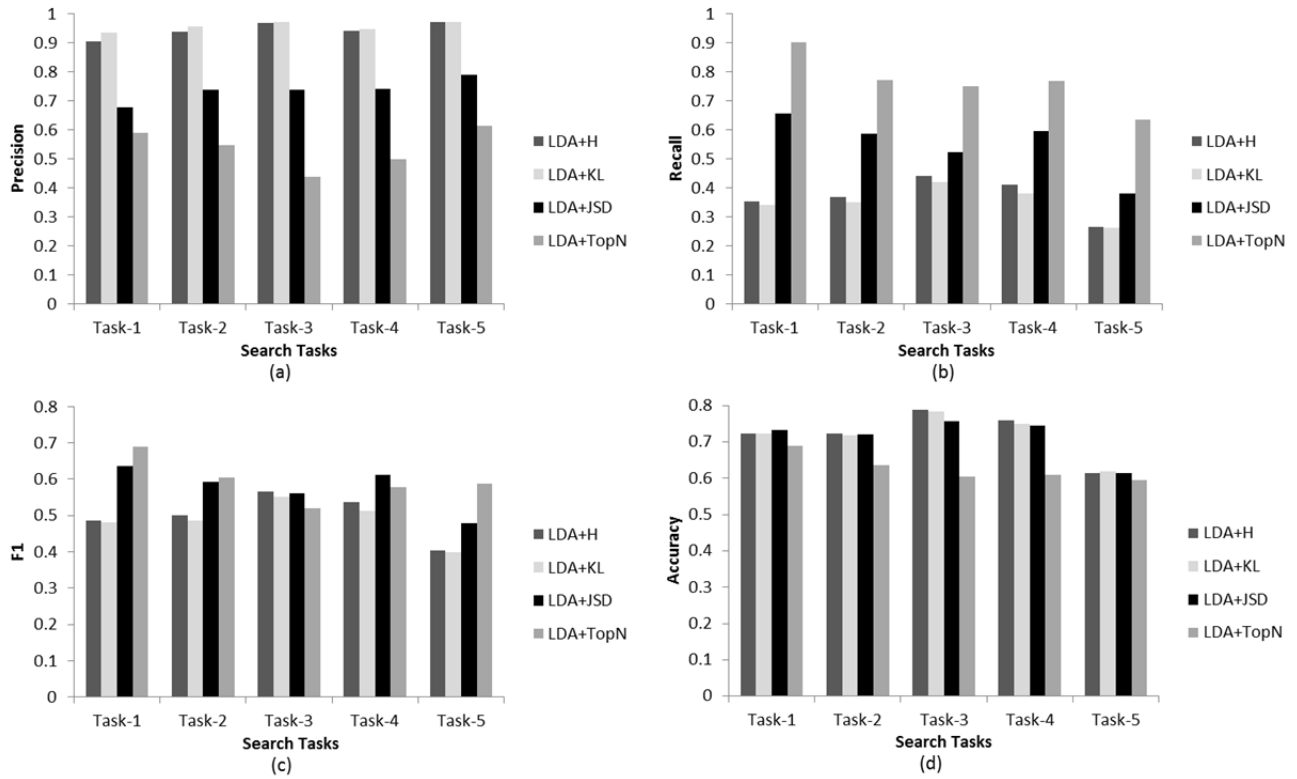

Figure 5. Performance comparisons of different models. (a) Precision, (b) Recall, (c) F1 measure, and (d) Accuracy

We have evaluated the effectiveness of the visualizations in recommending interesting new documents and passages within documents based on what the user has explicitly indicated their interests using annotations. In the future, we expect to employ Automatic Query Expansion (AQE) to identify pseudo-implicit-feedback to generate similar visualization to support users search task activity. The classification of documents and parts of a document into different user interests in the current UIMaP is based on explicit user annotations in a single application. The current work can be easily extended to support multi-application environment with weighting schemas to detect the relative importance of different applications to the users' search tasks. For example, if the user is writing a paper while she is performing search task relevant to the writing task, the topics that emerge in the paper may be a very effective source of interest profile data.

6. ACKNOWLEDGMENTS

This work was supported in part by National Science Foundation grant DUE-0938074.

7. REFERENCES

[1] S. Bae, D. Kim, K. Meintanis, J. M. Moore, A. Zacchi, F. Shipman, *et al.*, "Supporting document triage via annotation-based multi-application visualizations," in *Proceedings of the 10th annual joint conference on Digital libraries*, 2010, pp. 177-186.

[2] D. Kelly and J. Teevan, "Implicit feedback for inferring user preference: a bibliography," in *ACM SIGIR Forum*, 2003, pp. 18-28.

[3] M. Claypool, P. Le, M. Wased, and D. Brown, "Implicit interest indicators," in *Proceedings of the 6th international conference on Intelligent user interfaces*, 2001, pp. 33-40.

[4] F. Shipman, M. Price, C. C. Marshall, and G. Golovchinsky, "Identifying useful passages in documents based on annotation patterns," in *Research and Advanced Technology for Digital Libraries*, ed: Springer, 2003, pp. 101-112.

[5] R. Badi, S. Bae, J. M. Moore, K. Meintanis, A. Zacchi, H. Hsieh, *et al.*, "Recognizing user interest and document value from reading and organizing activities in document triage," in *Proceedings of the 11th international conference on Intelligent user interfaces*, 2006, pp. 218-225.

[6] M. N. Price, B. N. Schilit, and G. Golovchinsky, "XLibris: The active reading machine," in *CHI 98 Cconference Summary on Human Factors in Computing Systems*, 1998, pp. 22-23.

[7] M. N. Price, B. N. Schilit, and G. Golovchinsky, "XLibris: the active reading machine," presented at the CHI 98 Cconference Summary on Human Factors in Computing Systems, Los Angeles, California, USA, 1998.

[8] C. C. Marshall and F. M. Shipman III, "Spatial hypertext: designing for change," *Communications of the ACM,* vol. 38, pp. 88-97, 1995.

[9] F. M. Shipman III, H. Hsieh, P. Maloor, and J. M. Moore, "The visual knowledge builder: a second generation spatial hypertext," in *Proceedings of the 12th ACM conference on Hypertext and Hypermedia*, 2001, pp. 113-122.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research,* vol. 3, pp. 993-1022, 2003.

[11] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling, "Fast collapsed gibbs sampling for latent dirichlet allocation," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 569-577.

[12] A. K. McCallum, "Mallet: A machine learning for language toolkit," 2002.

[13] C. R. Rao, "The use of Hellinger Distance in graphical displays of contingency table data," *New trends in probability and statistics,* vol. 3, pp. 143-161, 1995.

[14] C. M. Bishop, *Pattern recognition and machine learning* vol. 1: springer New York, 2006.

[15] D. Hall, D. Jurafsky, and C. D. Manning, "Studying the history of ideas using topic models," in *Proceedings of the conference on empirical methods in natural language processing*, 2008, pp. 363-371.