Navigation Leads Selection Considering Navigational Value of Keywords

Robert Moro
Faculty of Informatics and Information Technologies,
Slovak University of Technology in Bratislava
Ilkovičova 2, 842 16 Bratislava, Slovakia
robert.moro@stuba.sk

Maria Bielikova
Faculty of Informatics and Information Technologies,
Slovak University of Technology in Bratislava
Ilkovičova 2, 842 16 Bratislava, Slovakia
maria.bielikova@stuba.sk

ABSTRACT

Searching a vast information space such as the Web presents a challenging task and even more so, if the domain is unknown and the character of the task is thus exploratory in its nature. We have proposed a method of exploratory navigation based on navigation leads, i.e. terms that help users to filter the information space of a digital library. In this paper, we focus on the selection of the leads considering their navigational value. We employ clustering based on topic modeling using LDA (Latent Dirichlet Allocation). We present results of a preliminary evaluation on the Annota dataset containing more than 50,000 research papers.

Categories and Subject Descriptors

H.5.4 [**Information Interfaces and Presentation**]: Hypertext/ Hypermedia – *navigation*, *user issues*. H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *information filtering*.

Keywords

Navigation; navigational value; topic modeling; digital libraries.

1. INTRODUCTION

Whether researchers want to get acquainted with a new domain or want to find state-of-the art approaches pertinent to their area of research, they usually conduct searches in a digital library that can be viewed as cognitive traveling in this space [6]. Their goal is not to find specific facts, but to learn about the given domain. The term *exploratory search* was coined for this type of searches [5].

In order to support exploration and sense-making, we have proposed a method of exploratory navigation using the navigation leads. We define *navigation leads* as important terms (automatically) extracted from the documents present in the information space. When users choose to follow a navigation lead, the documents are filtered so that only those related to the selected lead are retrieved. The leads are visualized directly in a document's summary (or abstract) or underneath it.

This way, the proposed navigational approach emulates the browsing behavior which allows the users to investigate the results and follow the links in the text. It also supports the idea of navigation-aided retrieval as defined in [7] by understanding the search results as mere starting points for further exploration.

It is as an alternative to a tag cloud navigation approach whose navigability was explored in [3]. In contrary to a tag cloud, the navigation leads by being placed directly in the text, do not force the users to split their attention, which could otherwise lead to higher cognitive load of the users as shown in [2].

2. SELECTION OF NAVIGATION LEADS

For the terms to be selected as navigation leads, they should be relevant for the document in which they are identified; we denote this as document relevancy R_D. At the same time, they should reflect the information subspace that is covered by the lead, i.e., the size of the subspace, its relevancy for the user (his current query) as well as how the term represents the subspace; we denote this as *navigational value NV* of the term.

Thus, the overall relevancy of the term *t* (a lead candidate) for a document *d* is computed as a product of the term's relevancy for the document and its navigational value:

$$R(t,d) = R_D(t,d) \times NV(t,d) \qquad (1)$$

In order to compute the document relevancy R_D, we can use any method of keyword extraction and weighing (e.g. TF-IDF). However, we do not consider only the texts of the documents, but also their associated metadata, such as user-added tags and keywords added by the authors (which are usually available in the domain of digital libraries of research papers).

When computing the navigational value of a term, we firstly identify the subspace behind the term. For this purpose, we propose to employ clustering; thus, the subspace is represented by a cluster that the document, for which we identify the leads, belongs to. We employ LDA (Latent Dirichlet Allocation) [1] for the clustering: each identified topic represents a cluster of documents, each document is represented as a probability distribution of topics (i.e. the document belongs to multiple clusters with some probability) and each topic is in turn represented as a probability distribution of terms. Overall, the navigational value of a term is computed as follows:

$$NV(t,d) = \sum_{c \in C} R_C(t,d,c) \times S(c) \times R_U(c) \qquad (2)$$

where C is a set of all cluster assignments for the current document *d*, R_C is a relevancy of the term *t* for the cluster *c*, *S* is a function of size of the cluster *c* that penalizes too large or too small clusters and R_U is a relevancy of the documents in the cluster *c* for the current user *u* (his query). The relevancy R_C of the term *t* for the cluster *c* is computed as a product of the probability of the cluster (topic) assignment for the document *d* and the probability (relevancy) of the term for that cluster:

WWW'15 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742764

$$R_C(t,d,c) = P(d \mid c) \times P(t \mid c) \qquad (3)$$

Thus, the idea is that by using the proposed method, the users explore the topics depth-first as opposed to width-first; however, by considering also clusters (topics) associated with the document with lower relevancy, the method leaves space also for exploring the related (less relevant) topics.

3. EVALUATION AND DISCUSSION

In the first phase, we focused on the suitability of the chosen clustering approach using LDA for the identification of the information subspace that is represented by the navigational value of a term. We experimented with the data from the web-based bookmarking system Annota (annota.fiit.stuba.sk). The dataset is publicly available for research purposes [4]; it contains more than 50,000 research papers from the domain of informatics, 140,000 author-added keywords and 3,500 user-added tags.

The goal of the experiment was to find out which keywords are the best for the clustering of the documents, as well as to assess the optimal number of the topics (clusters). We compared the results of LDA for the keywords from four different sources: the keywords (terms) extracted from the abstracts of the documents, the keywords extracted from the whole content of the documents, the keywords added by the authors of the documents and those added by the users as tags.

We used a measure of perplexity on a held-out test set to compare the results; lower perplexity values indicate better generalization performance of the trained model [1]. The best performance was achieved by the keywords extracted from the whole content of a document followed by the user-added tags (see Fig. 1). The worst performance was achieved by the author-added keywords (not in the figure because the values would be off the scale).

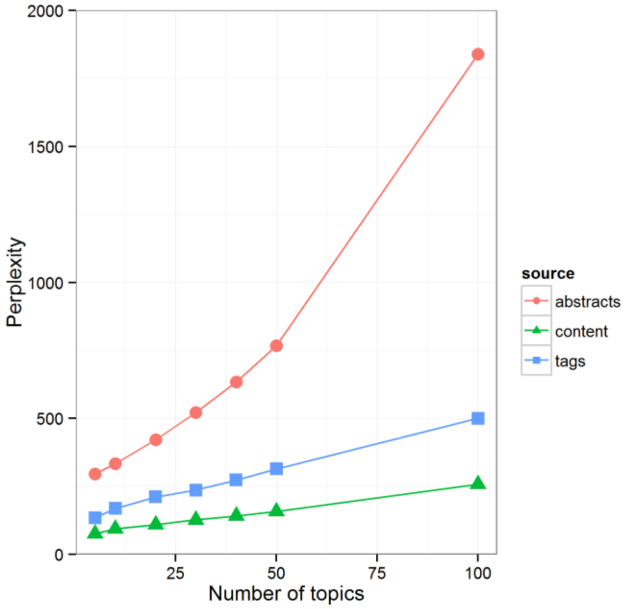

Figure 1. Comparison of perplexity values for different number of topics and different sources of keywords.

However, the perplexity score does not necessarily mean that the model is better from the perspective of human judgment; therefore, we also inspected the identified topics manually. Table 1 shows two topics (represented by the five most relevant words) identified when using papers' abstracts and a combination of keywords extracted from the whole content and user-added tags. We can see that the former contains more general terms, some of which are actually domain-specific stop words (e.g. system, user). On the other hand, the latter gives more sensible output and manages to identify certain relations (e.g. recommendation and collaborative filtering, etc.).

The results of the preliminary evaluation suggest that the clustering approach can be used to assess the navigational value of the terms. Its consideration when computing the relevancy of a term presents the main contribution of the proposed approach as well as the fact that it combines the depth-first navigation with wider exploration by considering also less relevant topics (clusters) associated with the documents. We plan to evaluate the method of the leads selection by a quantitative user study assessing the usefulness of the proposed exploratory navigation approach for sense-making in the researcher novice scenario.

Table 1. Example of identified topics.

Abstracts		Content + Tags	
Topic #1	Topic #2	Topic #1	Topic #2
recommendation, user, system, social, item	model, system, software, use, development	recommender system, collaborative filtering, adult, person, recommend	social network, semantic web, ontology, social media, linked data

4. ACKNOWLEDGMENTS

This work was partially supported by grants No. VG 1/0646/15, No. KEGA 009STU-4/2014, and it was created with the support of the Research and Development Operational Programme for the project "University Science Park of STU Bratislava", ITMS 26240220084, co-funded by the European Regional Development Fund.

5. REFERENCES

[1] Blei, D.M., Ng, A.Y., Jordan, M.I. 2003. Latent Dirichlet Allocation. J. of Machine Learn. Research. 3, 4-5, 993–1022.

[2] Ginns, P. 2006. Integrating information: A meta-analysis of the spatial contiguity and temporal contiguity effects. *Learning and Instruction*. 16, 6, 511-525.

[3] Helic, D., Trattner, C., Strohmaier, M., Andrews, K. 2011. Are tag clouds useful for navigation? A network-theoretic analysis. *Int. J. of Social Computing and Cyber-Physical Systems*. 1, 33-55.

[4] Holub, M., Moro, R., Sevcech, J., Liptak, M., Bielikova, M. 2014. Annota: Towards enriching scientific publications with semantics and user annotations. *D-Lib Magazine*. 20, 11/12.

[5] Marchionini, G. 2006. Exploratory search: From finding to understanding. Communications of the ACM. 49, 41-46.

[6] Návrat, P. 2012. Cognitive traveling in digital space: from keyword search through exploratory information seeking. *Central European J. of Computer Science*, 2, 3, 170-182.

[7] Pandit, S. and Olston, C. 2007. Navigation-aided retrieval. In *WWW '07: Proc. of the 16th Int. Conf. on World Wide Web*. ACM Press, 391–400.