Using Metadata for Query Refinement and Recommendation

Jian Qin
Syracuse University, Syracuse, NY 13244, USA
jqin@syr.edu

Xiaozhong Liu
Syracuse University, Syracuse, NY 13244, USA
xliu12@syr.edu

Xia Lin
Drexel University, Philadelphia, PA 19104, USA
xlin@drexel.edu

Miao Chen
Syracuse University, Syracuse, NY 13244, USA
Mchen14@syr.edu

Abstract

Lengthy lists of search results are the fruit of both short queries and conventional Web search result displays. They are problematic for meeting user's information needs. This paper describes the first part, topic extraction and representation from metadata, of a project that will develop an interactive visual query refinement and recommendation (QRR) service to alleviate the problems due to lengthy lists of search results. The topic extraction uses the Latent Dirichlet Allocation (LDA) algorithm to mine the intra- and inter-document relations and represent them in topic and features. The paper presents how the LDA algorithm extracts topics and features from metadata records contained in NSDL search results, which will be used by an interactive QRR service in the next step of the project.

Keywords: query refinement; LDA algorithm; topic detection and tracking; query refinement and recommendation (QRR).

1. Introduction

Metadata provides the foundation necessary for resource discovery and for building multi-facet search interfaces. While digital library search systems such as National Science Digital Library's (NSDL) help users formulate and refine queries, the result display still resembles to a great extent the conventional format in which brief metadata descriptions are organized in a linear listing style. The problem, however, rises when the system returns large numbers of hits. As a result, wading through the lengthy lists of results to select relevant resources can be difficult and time consuming. A recent study discovered that, among the help-seeking situations in digital library searches, search refinement situations took place almost three times more frequently than other types of situations, counting for 41.5% of the total, and the next largest percentage (18%) was the inability to create the search statement (Xie & Cool, 2009). There is "a lack of support and expertise that learners need to select an appropriate resource, incorporate it into a coherent learning experience, and evaluate the impact of the new approach" (NSF, 2009). As Lagoze et al. (2006) point out, "collection building and metadata aggregation are necessary but not sufficient activities for building an information-rich digital library." To solve this problem, the key lies in transforming the linear listing of search results into a more informative visualization of the results with which users can view and interact. The computationally processed, visualized search results would allow search systems to present the results more intuitively and in greater depth, which builds the necessary data for making recommendations for users. In this sense, the query refinement and recommendation we proposed will be computationally based and visually intractable. This short paper reports on preliminary research that uses a topic detection algorithm to reveal the semantic constructs inside metadata search results from the NSDL repository. While the project has two parts—topic extraction and visualization, we will discuss only the topic extraction part of the project in this paper.

2. Related Research

Research has investigated ways to improve users' interaction with search systems from query previews (Doan et al., 1997; Jones, 1998) to various versions of query refinement, e.g., query reformulation (Arens, Knoblock, & Shen, 1996), query feedback (Radlinski & Joachims, 2005),

OPEN ACCESS
This article is licensed under a **Creative Commons Attribution 4.0 International License**, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and cite the source.

https://doi.org/10.23106/dcmi.952109526

and query suggestion (Jones et al., 2006). The sources that these studies used fall into three groups: 1) metadata records, 2) full text corpus, and 3) query logs. These types of sources may also be combined to achieve better results. Query previews provide aggregated information about a search result set based on subject categories, formats, resource types, and publication years that exist in metadata records. Query refinement refers to the process that helps users disambiguate, refine, and recommend queries representing their information needs. NSDL's search engine offers multiple facets of options that can dynamically change the search results as the searcher check or uncheck search options (http://nsdl.org/search/). While query previews and refinement help users narrow down search results and alleviate the difficulties in formulating and refining queries, they solve only half of the problem. The lengthy list of search results displayed on screen still remains as a barrier for users to select the results that best match their information needs.

The query refinement and recommendation (QRR) service is proposed to address the other half of the problem that query previews and multiple-option search interface cannot solve. Rather than simply aggregating metadata information on the display screen, we use automatic topic detection techniques to mine the topic construct inside a search result set and present the semantic pattern to users. What is a topic? In topic detection and tracking research, a topic is defined as "a set of news stories that are strongly related by some seminal real-world event" (Allan, 2002). Another commonly held notion of topic is subject—what a document or resource is about. The subject-based topic may be represented by natural language (keywords or phrases) or controlled vocabulary (terms in thesauri or taxonomies). Although the meaning of topic has been mainly limited to news stories in topic detection and tracking (TDT) research, the notion of topic in both TDT and subject-based representation fits into the context of our study: a topic is a semantic class that consists of terms sharing some common subject relationships or similar meanings.

Detecting topics from a text corpus (including metadata) can be decomposed into a series of tasks, including extraction of words and/or entities, detection or modeling of topics by using algorithms, clustering or grouping features of topics, and evaluation of the results. A classic approach is using the term frequency–inverse document frequency (tf-idf) scheme (Salton and McGill, 1983). This frequency-based weighting approach is limited to the intra-document topic relations and provides little information on inter-document relations. The Latent Semantic Indexing (LSI) (Deerwester et al, 1990) and probabilistic Latent Semantic Indexing (pLSI) (Hofmann, 1999) have been proposed to address the limitation. A more recent development is the Latent Dirichlet Allocation (LDA) algorithm (Blei et al., 2003). It is a dimensionality reduction technique while providing "proper underlying generative probabilistic semantics that make sense for the type of data that it models" (Blei et al., 2003, p. 1014).

3. Topic Detection and Representation from Metadata

Short queries in information search systems have become a well-known user search behavior. They make it difficult for rank and match algorithms to detect the necessary information for locating the best result sets for users. They are also prone to generating overwhelmingly large result sets that can be hard to wade through and select the relevant ones by average users. Our strategy to solve the problems is using the LDA topic extraction algorithm (Blei et al., 2003) to extract semantic topics and features from the search result set to reduce the complexities and dimensionalities of overwhelmingly large search result sets while providing a picture of the overall content construct in the sets. This solution is based on the assumption that a search result set contains information related to user's information needs and the post-search topic extraction from the resulting metadata record set will help users discover the topics and features that they might otherwise be unable to see. This approach differs from the aggregative display of metadata records mentioned earlier in this paper in that the topics are extracted based on probability distribution of topical words and entities. The term "topic" in this research refers to a semantic class that consists of terms sharing some common subject relationships or similar meanings. The features of a topic are words, keywords or phrases that semantically belong to the topic.

OPEN ACCESS
This article is licensed under a **Creative Commons Attribution 4.0 International License**, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and cite the source.
https://doi.org/10.23106/dcmi.952109526

Traditional document indexing systems represent each document as a vector of features (the doc-by-feature matrix on the left in Figure. 1).

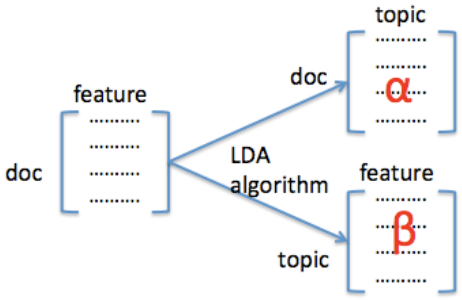

FIG. 1. A model of topic and feature probability distribution

The LDA algorithm decomposes the traditional doc-by-feature matrix into two new matrices: a document-by-topic matrix α and a topics-by-feature matrix β (Figure. 1). A benefit of this method is that it transfers the classical statistical vector space into a semantic topic-based vector space and the probability distributions create the necessary data for visualizing the intra- and inter-document semantic constructs in search result sets. Topic representation in the context of this project is a process of transforming the word occurrence probabilities into human understandable subject terms or keywords. During this process, the extracted topics are mapped with the subject keywords and other features (words in title, description, etc.) existing in NSDL metadata records. Such mapping is done through calculating the probability at which a given keyword in a metadata record belongs to a topic. The computation can be expressed by the formula below:

$$P(keyword_X \mid topic_Y) = \frac{\sum_{m=1}^{N} Weight(keyword_X, m) * P(topic_Y \mid doc_m)}{n}$$

The $P(keyword_x|topic_y)$ is the probability that a keyword x in a metadata record maps to a topic, m stands for the training set, and $P(topic_y|doc_m)$ for the probability that a given topic occurred in the search result set obtained from the matrix α. The workflow of topic extraction and representation using the LDA method is shown in Figure. 2 (following page), which includes following steps:

1. Input a query to obtain a list of results containing snippets, collection names, and keywords as shown in Figure. 3 on the following page.
2. Randomly sample search results (snippets) from the retrieved collection to extract topic distribution probabilities.
3. Compute the document-topic probability distribution for each candidate result.
4. Calculate the topic-keyword and topic-collection probability distributions.
5. Visualize the topics by showing top words, keywords, and collections covered by each topic through an interactive user interface, which will allow users to view the topic constructs and easily reconstruct their queries.

We conducted a test of the LDA algorithm for three queries: "global warming," "pollution," and "ontology." Using the global warming query as an example, the original query "global warming" to the NSDL search engine returned 4,735 hits. After removing duplicate results, 2,137 remained in the search result set. The processing of this result set by the LDA algorithm generated k topics based on the semantic proximity of words and/or entities (in this experiment, k = 10. Table 1 shows the fist three topics only due to page limit). It should be mentioned that the words came from all text fields in metadata records, rather than just the subject keywords. The resulting topics recommend that searching information for global warming can be looked at under

OPEN ACCESS
This article is licensed under a **Creative Commons Attribution 4.0 International License**, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and cite the source.

https://doi.org/10.23106/dcmi.952109526

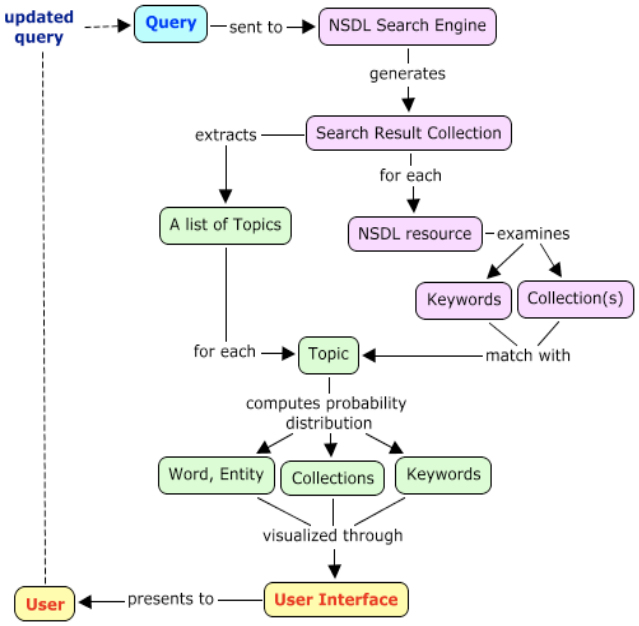

FIG. 2. Workflow of topic extraction and representation

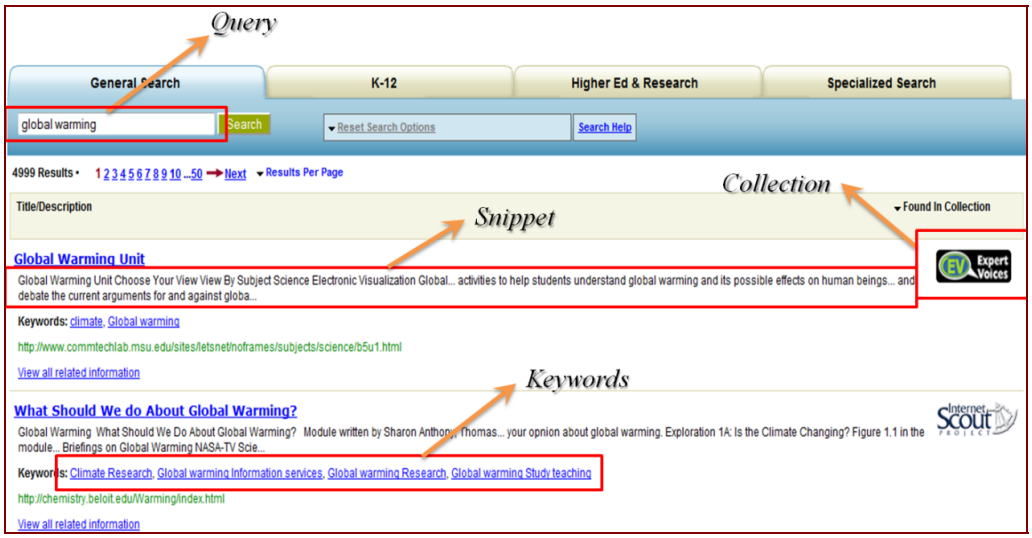

FIG. 3. Search result display interface of NSDL

earth science, government policy, and climate change topics (Table 1). The numerical value next to each word or entity in Table 1 is the probability that a word or entity belongs to a topic.

TABLE 1: Top word (stem), entity distribution (β Matrix) for query "global warming" calculated by the LDA algorithm

Earth Science Topic		Government Policy Topic		Climate Change Topic	
scienc	0.03678588	energi	0.0377121	research	0.032611
earth	0.03404345	carbon	0.0282291	inform	0.025435
student	0.01522045	develop	0.0174741	nation	0.018136
includ	0.01497114	technolog	0.0136578	issu	0.014974

OPEN ACCESS
This article is licensed under a **Creative Commons Attribution 4.0 International License**, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and cite the source.
https://doi.org/10.23106/dcmi.952109526

project	0.01409855	emiss	0.0135421	state	0.01473
resourc	0.01409855	fuel	0.0131952	center	0.014487
educ	0.01247803	power	0.0130795	news	0.013027
activ	0.01210406	dioxid	0.0125013	public	0.012906
learn	0.00873836	product	0.0109979	polici	0.011933

Similarly, the numerical values in Table 2 are the probabilities at which keywords match a topic and those in Table 3 are the probabilities that a collection has resources for the topic. The data in Table 1 presents the top-word probability distribution from the β matrix. As keywords in metadata records are originally assigned (or extracted) to describe resources in the NSDL repository, we actually compute the probability that each keyword is assigned for a topic. The top keywords can be viewed as the "pseudo keywords" for topics. Similar probability computations can be performed for other metadata facilitated features, e.g., between topics and NSDL collections to show which collections are more likely to contain resources about which topics, that is, the probability that a keyword appears in different collections and belongs to different topics.

TABLE 2: Keyword distribution based on metadata in the result set for query "global warming"

Earth Science Topic		Government Policy Topic		Climate Change Topic	
earth_space_science	0.16215084	policy_economy	0.2684802	climatic_changes	0.161443
energy	0.13454888	energy_planning	0.2669252	air_pollution	0.142076
physics	0.13390858	carbon_dioxide	0.2039896	global_warming	0.130123
climate	0.12670773	greenhouse_effect	0.1782857	science_earth_science	0.129516
technology	0.12003977	environmental_impacts	0.1625156	climate_change	0.114684
space_science	0.11821817	greenhouse_gases	0.1591277	climate	0.114438
education_(general)	0.11688358	environmental_sciences	0.143155	health	0.103269
natural_hazards	0.10845419	climatic_change	0.1347704	greenhouse_gases	0.099786
atmospheric_science	0.10832897	progress_report	0.1288264	hydrology	0.097104

TABLE 3: Probability of top 3 collections that contain resources on the global warming topic

Earth Science Topic	Government Policy Topic	Climate Change Topic
NSTA: 0.19214482	OSTI: 0.1873345	Infomine: 0.178953
On the Cutting Edge: 0.17101018	DSpace at MIT: 0.1339377	NSDL Expert Voices: 0.128294
Compadre: 0.14753963	Directory of Open Access Journals: 0.1247221	Internet Scout Project: 0.126464

4. Next Steps

The research reported in this paper is still in its early stages. The LDA topic extraction and representation algorithms have been developed and tested with sample queries. The next step will be obtaining query log data from NSDL to conduct a larger scale testing and tuning of the algorithms and, once the algorithms stabilize, a visualization interface will be developed and evaluated with users. A unique contribution of this research is that, by using the LDA topic extraction techniques, metadata's role will be greatly enhanced from mechanically matched and presented to semantically classified and linked. Users will be able to view the semantic constructs that are covert and difficult to see in large search result sets due to the linear display.

OPEN ACCESS
This article is licensed under a **Creative Commons Attribution 4.0 International License**, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and cite the source.

https://doi.org/10.23106/dcmi.952109526

References

Allan, J. (2002). Introduction to topic detection and tracking. In: J. Allan (Ed.), Topic Detection and Tracking: Event-Based Information Organization, pp. 1-16. Boston, MA: Kluwer Academic Publishing.

Arens, Y., Knoblock, C.A., & Shen, W. (1996). Query reformulation for dynamic information integration. Journal of Intelligent Information Systems, 6(2-3): 99-130.

Blei, D. M., Ng, A. Y., & Jordan, M. J. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3: 993-1022.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6): 391-407.

Doan, K., Plaisant, C., Shneiderman, B., & Bruns, T. (1997). Query previews for networked information systems: A case study with NASA environmental data. SIGMOD Record, 26(1): 75-81.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In: Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 1999, Berkeley, CA, pp. 50-57. New York: ACM Press.

Jones, R., Rey, B., Madani, O., & Greiner, W. (2006). Generating query substitutions. In: Proceedings of the 15th International Conference on World Wide Web, pp. 387-396. New York: ACM Press.

Jones, S. (1998). Dynamic query result previews for a digital library. In: Proceedings of the Third ACM conference on Digital Libraries, Pittsburg, PA, June 23 - 26, 1998, pp. 291-292. New York: ACM Press.

Lagoze, C., Kraft, D., Cornwell, T., Eckstrom D., Jesuroga, S., and Wilper, C. (2006). Representing contextualized information in the NSDL. In: Research and Advances for Digital Libraries, pp. 329-340. Berlin: Springer.

NSF. (2009). National STEM Education Distributed Learning (NSDL): Program solicitation NSF 09-531. http://www.nsf.gov/pubs/2009/nsf09531/nsf09531.html

Radlinski, F. & Joachims, T. (2005). Query chains: Learning to rank from implicit feedback. In: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, pp. 239-248. New York: ACM Press.

Salton, G. & McGill, M. (1983). Introduction to Modern Information Retrieval. New York: McGraw-Hill.

Xie, I. & Cool, C. (2009). Understanding help seeking within the context of searching digital libraries. Journal of the American Society for Information Science and Technology, 60(3): 477-494.

OPEN ACCESS
This article is licensed under a **Creative Commons Attribution 4.0 International License**, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and cite the source.

https://doi.org/10.23106/dcmi.952109526