TopicVis: A GUI for Topic-based Feedback and Navigation

Debasis Ganguly Manisha Ganguly Johannes Leveling Gareth J. F. Jones
School of Computing, Centre for Next Generation Localisation
Dublin City University, Dublin 9, Ireland
{dganguly, mganguly, jleveling, gjones}@computing.dcu.ie

ABSTRACT

This paper describes a search system which includes topic model visualization to improve the user search experience. The system graphically renders the topics in a retrieved set of documents, enables a user to selectively refine search results and allows easy navigation through information on selective topics within documents.

Categories and Subject Descriptors

H.3.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval; H.5.2 [**USER INTERFACES**]: Graphical User Interfaces

Keywords

Topic Visualization, Latent Dirichlet Allocation

1. INTRODUCTION

Queries often encompass a wide range of different sub-information needs. For instance, the query *relativity theory* may seek for relevant information on the history of relativity, the special or the general versions of the theory, criticisms of the theory and so on. This multi-faceted nature of the information need expressed in a query is manifested in the retrieved documents, as they tend to form clusters of topics [3]. For instance, a search on *Einstein relativity theory* may retrieve documents with information on specific sub-topics of relativity theory. A searcher can benefit from a graphical interface, which provides a visualization of these topics in the retrieved set of documents, due to easier navigation through these topics. Some existing visualization tools such as the *Clusty*[1] and *Carrot*[2], can categorize the retrieved set of documents into groups of topics, and hence provide a more organized information access to the searcher in comparison to standard web search engines. Both these systems rely on clustering the set of retrieved documents. However, limitations of the cluster hypothesis mean that each document can only belong to a single cluster (in this case a topic class), and that clusters are mutually exclusive. However, in practice, it is observed that expository documents, such as Wikipedia articles, contain information on several topics, and hence a document can be classified into multiple categories [2]. For example, the document titled *Theory of relativity*, comprising of a mixture of both the special and the general theories, can be classified to both these classes. Documents thus have to be considered as a mixed bag of topics. The generative model, namely the Latent Dirichlet Allocation (LDA) [1], achieves this by treating each document as a mixture of topic distributions. Previous work on visualizing topic models involved application of LDA to categorize each Wikipedia article as a mixture of topics, and allowing navigate through documents related to a chosen topic [2]. However, the limitations of this system[3] are that it has no provision for: a) query-based information search, and b) visualization of the topics within a document. The motivation of our approach is to help the searcher to locate information on his topic of interest from within a document which may be comprised of multiple topics. Existing cluster-based search systems, such as *Clusty* and *Carrot*, simply categorize the ranked list of documents by topics, but do not provide the topic-based navigational feature within sections of documents. The objective of our work is to demonstrate a user-friendly information access approach within the list of retrieved documents through topic visualization. To this end, we developed a web interface, named *TopicVis*[4].

[1]http://clusty.com/
[2]http://search.carrotsearch.com/carrot2-webapp/search

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.
ACM 978-1-4503-2034-4/13/07.

2. SYSTEM DESCRIPTION

After receiving a query from the user the TopicVis web application retrieves a set of documents from a collection of over $3.69M$ English Wikipedia articles. The system then executes LDA on the top ranked documents to obtain the document-topic and topic-word distributions, i.e. θ and ϕ respectively [1]. The ϕ distribution, along with a list of top 10 most likely words in each topic, is rendered as a pie-chart on the left part of the screen. Along the right pane of the TopicVis screen, the ranked list of retrieved documents are displayed. In addition to showing the title and the snippet of a document, as in a standard search engine, TopicVis shows the distribution of topic in each document, with the help of a stacked bar chart. We now explain the features of TopicVis

[3]http://bit.ly/wiki100
[4]`http://www.cngl.ie/TopicVis/`

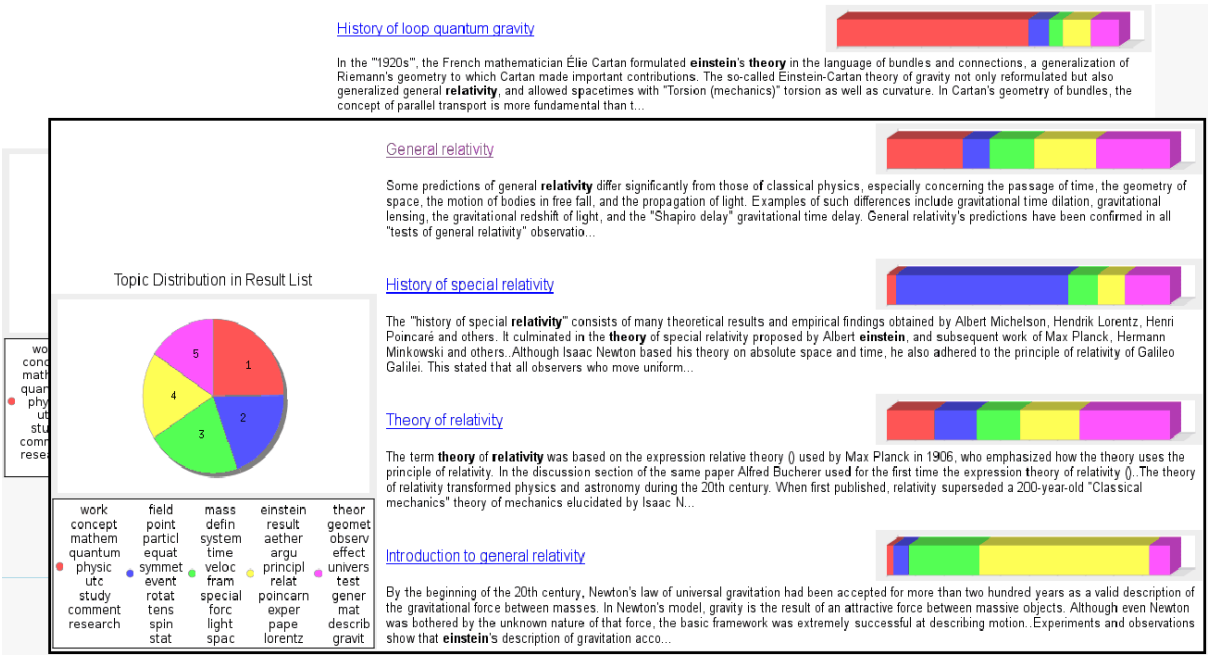

Figure 1: Screenshot of initial retrieval overlayed on that of the topic-based feedback using topic 1.

with reference to a screenshot shown in Figure 1, for the query "Einstein relativity theory", for which the number of topics was preconfigured to 5.

Topic distribution in the retrieved set. Each region in the pie-chart represents a topic in the retrieved set of documents. The topic number increments clock-wise starting from 1 as shown in Figure 1. From the list of words displayed below the pie-chart, it can be seen that topics 1 and 5 broadly represent the concept related to the general theory of relativity, whereas the topics 2 and 3 correspond to special theory of relativity.

Topic distribution in a single document. Alongside the title and snippet of a document, TopicVis shows the distribution of topics in its content with the help of a stacked bar chart. The intention is to provide a visual cue about the content of a document without the necessity of reading the snippet. Each region in a stacked bar chart shows the proportion of text on the corresponding topic. The left most region of a stacked bar chart corresponds to topic 1 and so on. It can be observed from Figure 1 that as per expectation, the first document titled *General Relativity* has more contribution from topics 1 and 5, whereas the second document titled *History of special relativity* has a high proportion of topics 2 and 3.

Topic-based feedback. A click event in a particular region of the pie-chart re-ranks the resultlist on the basis of the topic, the region corresponds to. Since each document is a mixture model of its constituent topics, it can be considered as a vector, where the proportion of each topic is a component. The document vectors are thus sorted by the component value corresponding to the selected topic. Figure 1, shows that when a user clicks on the region in the pie-chart corresponding to topic 1, i.e. the topic on general theory of relativity, the document titled *History of loop quantum gravity* is shown on top. Note that this document has the highest proportion of text on topic 1.

Topic-based navigation. The title of a document is hyperlinked to a standard text-based view. However, in such a case, it is difficult for a user to locate the sought information from long expository articles. Since Wikipedia articles are structured into segments such as sections and paragraphs, a guided walk through the segments of a document, relevant to a given topic, can be beneficial. For example, consider that the user is interested in topic 3, i.e. the topic on frames of references in the special theory of relativity. Figure 1 shows that the first document in the retrieved list of documents, i.e. *General relativity*, has some contribution from topic 3. TopicVis facilitates a user to directly jump into the first segment of a document predominated by his topic of interest, simply by clicking an area of the stacked bar chart corresponding to the topic. The segmentation units, specifically used for the demonstration, are the paragraphs contained within the " " and " " tags.

For each segment of a document, we compute the number of words in each topic with the help of the ϕ matrix. A segment is then classified to the topic with the maximum number of words. The text of each topic classified segment is bordered on the right with the corresponding colour. Each segment within a document is annotated with links to the next and the previous segments on the same topic.

Acknowledgments

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the CNGL project.

3. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[2] A. J.-B. Chaney and D. M. Blei. Visualizing topic models. In *Proceedings of ICWSM*, 2012.

[3] D. Ganguly, J. Leveling, and G. J. F. Jones. Topical relevance model. In *Proceedings of AIRS '12*, 2012.