High-Recall Document Retrieval from Large-Scale Noisy Documents via Visual Analytics based on Targeted Topic Modeling

Hannah Kim*
Georgia Tech, Atlanta, USA

Jaegul Choo
Korea University, Seoul, Korea

Alex Endert
Georgia Tech, Atlanta, USA

Haesun Park
Georgia Tech, Atlanta, USA

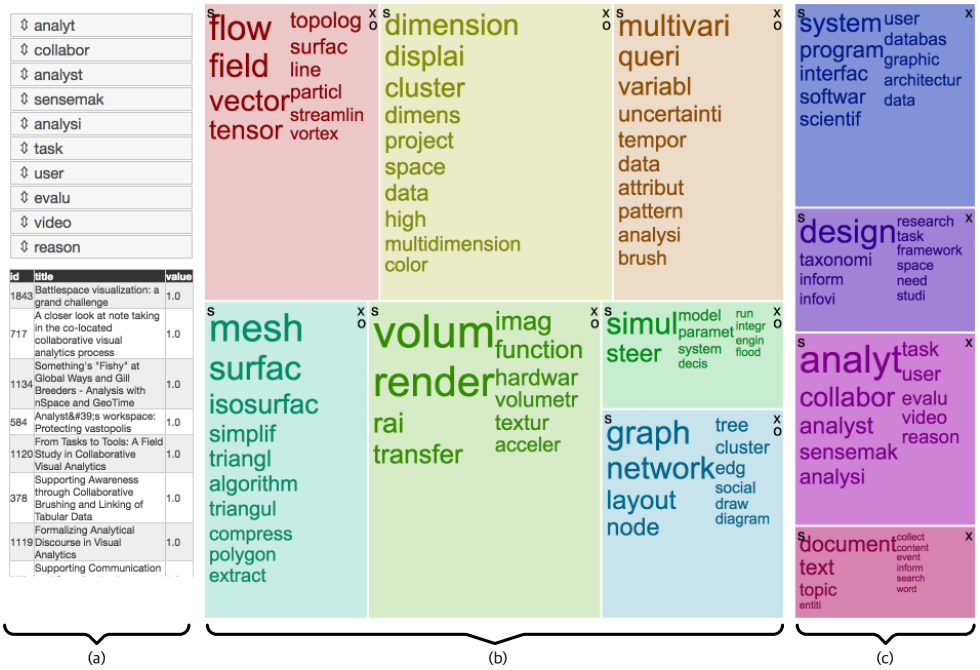


Figure 1: Overview of the proposed system. A topic summary is visualized as a treemap in the main panel (b). Each cell represents a topic with its representative keywords. Users can mark a topic cell as relevant, which will then be moved into the right panel (c), which shows confirmed relevant documents in another treemap visualization with a color darker than the middle one. Details of each topic is shown on the left panel (a), where users can adjust the importance of each keyword in the topic and see those documents belonging to the user-selected topic.

Abstract

We present a visual analytics system for large-scale document retrieval tasks with high recall where any missing relevant documents can be critical. Our system utilizes a novel user-driven topic modeling called *targeted topic modeling*, a variant of nonnegative matrix factorization (NMF). Our system visualizes a topic summary in a treemap form and lets users keep relevant topics and incrementally remove uninteresting topics in our treemap view without losing potentially relevant documents.

Keywords: Text analytics, search space reduction, targeted topic modeling, document retrieval, nonnegative matrix factorization, treemap

1 Introduction

Over the past decades, there has been a deluge of text data from traditional sources such as news and research articles as well as recent sources such as social networking services, online forums, and online encyclopedia. Due to the noisy unstructured nature of these data, there have been increasing needs for *retrieving only a subset of data items of interest*, e.g., documents related to a particular event, subject, brand, or product, from them. This task is not only about collecting a small number of most relevant documents but also about retrieving as many relevant documents as possible. Recently, because of the sheer volume of incoming data, this task has become time-consuming and burdensome.

To support this task, we propose an interactive visual retrieval system for large-scale documents. Our system dynamically groups documents with respect to their topics, which we treat as exploration units. Topics are visualized in treemap for users to explore and interact with. For instance, users can combine multiple topics into one, split a topic into multiple sub-topics, refine a topic, mark a topic relevant, and remove a topic from the treemap. Leveraging these interactions, our system allows users to divide and conquer one group of documents at a time with a flexible granularity. As a result, our system can retrieve relevant topics and documents (targets) in an efficient manner. We refer to this novel topic modeling approach as targeted topic modeling.

2 Interactive Visual Retrieval of Large-Scale Documents via Targeted Topic Modeling

2.1 System Overview

Fig. 1 shows the visual interface of our system. On the left is the detail panel with an interactive list of keywords and a document table displaying detailed information about a topic (Fig. 1(a)). On

*e-mail: hannahkim@gatech.edu

the right is the main panel with interactive treemap visualization for topic exploration (Fig. 1(b)). We use the Nmap technique [1] to put semantically similar topics close to each other in our treemap visualization for easier exploration.

In the treemap, users can investigate topic keywords and merge, split, and modify topics to find topics of interest (targets). When some topics are marked as relevant or interesting by users, our system creates an additional treemap on the right (Fig. 1(c)) to keep track of relevant topics and their corresponding documents. If some topics are considered irrelevant, users can delete them from the treemap. After several interactions, only relevant and interesting topics will finally appear in the treemap on the right.

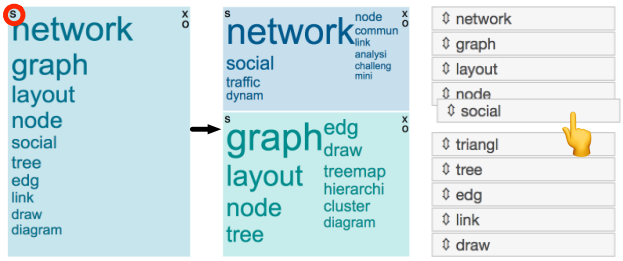
2.2 Targeted Topic Modeling

While traditional topic modeling has been effective in providing a keyword-based topic summary over an entire dataset, it fails to achieve the same level of effectiveness in providing detailed analysis on a subset of documents of interest, namely targets. Our targeted topic modeling approach based on nNonnegative matrix factorization (NMF) [2], however, aims to discover relevant topics and documents (targets) by leveraging user interactions. With our system, users may indicate whether a certain topic is relevant and to be kept; or irrelevant and to be removed. To avoid removing *potentially* relevant documents when the associated topic (marked as irrelevant) is removed, we remove only those documents strongly related to the topic and conservatively re-assign moderately related ones to the remaining topics.

2.3 Supported Interactions

This section describes the interactions supported by our system. For easy and intuitive interaction, simple click or drag-and-drop operations are adopted for all functions.

Getting the details of a topic. In some cases, it may not be easy to understand a topic solely based on its top keywords. For better interpretation, users can examine which documents belong to the topic and how strongly they are related to the topic in the table on the detail panel (Fig. 1(a)).



(a) Split (b) Refine

Figure 2: Examples of supported interactions. (a) Splitting a topic 'graph' reveals more coherent topics 'graph layout' and 'social network vis'. (b) One can refine a topic by re-ordering topic keywords.

Merging topics. When two topics are similar, users may want to combine them into a single topic. To do so, users can simply drag one topic cell onto another.

Splitting a topic. When a topic is too general or its size is too big, users can split it to obtain more detailed subtopics by clicking 's' button in the top left corner of the topic cell (Fig. 2(a)).

Refining a topic. Users can change the importance weight of each keyword by re-ordering the keyword list to emphasize or de-emphasize a particular aspect of a topic. For example, in Fig. 2(b), we want the 'graph, network' topic to be about 'social network', so we drag 'social' keyword upwards.

Keeping/marking a topic. If a user considers some topics interesting and relevant, she may want to keep them separate from unexplored topics. By clicking 'o' button in the top right corner of a topic cell, the selected topic is moved into the second treemap along with other marked topics (Fig. 1(c)).

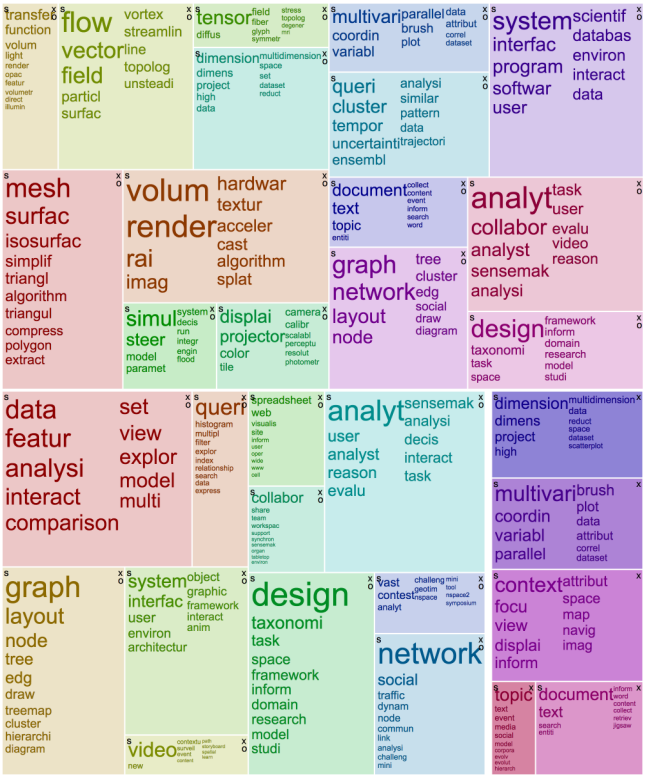


Figure 3: (Top) Initial topic overview. (Bottom) After interactions.

Removing a topic. If a user consider a topic irrelevant or uninteresting, they can remove the topic by clicking 'x' button in the top right corner of a topic cell. This allows more screen space to be used for the exploration of the remaining topics.

3 Usage Scenario

We use our system to analyze IEEE VIS publication dataset[1]. In the initial topic overview, SciVis-related topics are shown on the left side while InfoVis/VAST-related topics are shown on the right side of the treemap (Fig. 3(Top)).

Suppose a user is mainly interested in interactive visualization for high-dimensional data such as text data. She first removes topics related to scientific visualization such as 'volume rendering', 'mesh rendering', etc. Unsure about which topics to delve into, she splits some of the remaining topics in treemap and inspects associated documents on the left panel. She finds topics such as 'dimension projection/reduction' and 'multivariate visualization' interesting and marks them by clicking 'o' button. As a result, the marked topics are shown in the second treemap on the right panel (Fig. 3(Bottom)). She continues to keep interesting topics and remove uninteresting topics until only relevant topics remain. She is satisfied that the document subset of the marked topics are indeed what she is looking for.

References

[1] F. S. L. G. Duarte, F. Sikansi, F. M. Fatore, S. G. Fadel, and F. V. Paulovich. Nmap: A novel neighborhood preservation space-filling algorithm. *TVCG*, 20(12):2063–2071, 2014.

[2] J. Kim and H. Park. Toward faster nonnegative matrix factorization: A new algorithm and comparisons. In *ICDM*, pp. 353 –362, 2008.

[1] http://www.vispubdata.org/