

Visual Analytics for Interactive Exploration of Large-scale Document Data via Nonnegative Matrix Factorization

Jaegul Choo
Georgia Institute of
Technology
jaegul.choo@cc.gatech.edu

Barry L. Drake
Georgia Tech Research
Institute
barry.drake@gtri.gatech.edu

Haesun Park
Georgia Institute of
Technology
hpark@cc.gatech.edu

1. OVERVIEW

Due to an ever increasing amount of document data and the complexities involved in their analyses that will reveal meaningful insights, it is crucial to guide users in their decision-making processes using advanced methods that are both interactive and information preserving. Numerous computational approaches from machine learning, data mining, information retrieval, and natural language processing, etc., can be useful for this purpose, but they are usually performed in a fully automated manner, leaving little room for human knowledge and interaction to be incorporated.

Visual analytics, defined as the science of analytical reasoning facilitated by interactive visual interfaces, has recently emerged to improve this issue. Generally, visual analytics aims at leveraging both computational approaches and human-computer interaction to effectively provide users with deep understanding about data. In this demo, we present the visual analytics systems for large-scale document data that we recently developed, UTOPIAN [1] and VisIRR [3].

These two systems are built upon a fundamental system called FODAVA testbed¹ [2]. As a core method, they utilize nonnegative matrix factorization (NMF) to compute a high-level topic summary for given document data. Documents are then clustered with respect to these topics and visually represented as a 2D scatter plot along with their topics. The 2D scatter plot is computed via dimension reduction methods such as linear discriminant analysis and t-distributed stochastic neighborhood embedding (t-SNE). In the following, we describe the main capabilities of UTOPIAN and VisIRR.

2. UTOPIAN

Topic modeling has been widely used for analyzing text document collections. Popular techniques such as Latent Dirichlet Allocation (LDA) have been successfully applied in many topic modeling applications, but they are prone to drawbacks in terms of algorithmic consistency and user interactivity in real-world applications.

Motivated by this, UTOPIAN (User-driven **T**opic Modeling Based on **I**nteractive **N**M²)² [1], as shown in Fig. 1, adopts NMF as a better alternative topic modeling method that gives more consistent results and better interactivity.

¹<http://fodava.gatech.edu/fodava-testbed-software>

²<http://tinyurl.com/2013utopian>

Utilizing its semi-supervised formulation, UTOPIAN enables users to interact with the topic modeling method and steer the result in a user-driven manner. We demonstrate the capability of UTOPIAN via several usage scenarios with real-world document corpora such as product review data and InfoVis/VAST paper data.

3. VISIRR

In the context of information retrieval, the notion of high recall, which tries to avoid the cases where any potentially relevant documents are missed by users, have been becoming more important than ever. For instance, when filing a patent, one would not want to miss any relevant prior art. For researchers, it is crucial to track new publications related to their research.

To better serve these high-recall types of tasks, we present our visual analytics system, VisIRR (an Interactive **V**isual **I**nformation **R**etrieval and **R**ecommendation for document data)³ [3], as shown in Fig. 2. VisIRR enables users to assign their preference feedback to the system and perform document recommendations out of half a million document data using a real-time pagerank-style algorithm. The recommendation can be performed based on content, a citation network, and co-authorship relations. We illustrate how the system works by providing detailed usage scenarios.

Acknowledgments

The work of these authors was supported in part by NSF CCF-0808863 and DARPA XDATA grant FA8750-12-2-0309.

4. REFERENCES

- [1] J. Choo, C. Lee, C. K. Reddy, and H. Park. UTOPIAN: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE Transactions on Visualization and Computer Graphics*, 2013. To appear.
- [2] J. Choo, H. Lee, Z. Liu, J. Stasko, and H. Park. An interactive visual testbed system for dimension reduction and clustering of large-scale high-dimensional data. In *Proc. SPIE 8654, Visualization and Data Analysis (VDA)*, pages 1–15, feb 2013.
- [3] J. Choo et al. VisIRR: Interactive visual information retrieval and recommendation for large-scale document data. Technical report, Georgia Institute of Technology, 2013.

³<http://tinyurl.com/VisIRR>

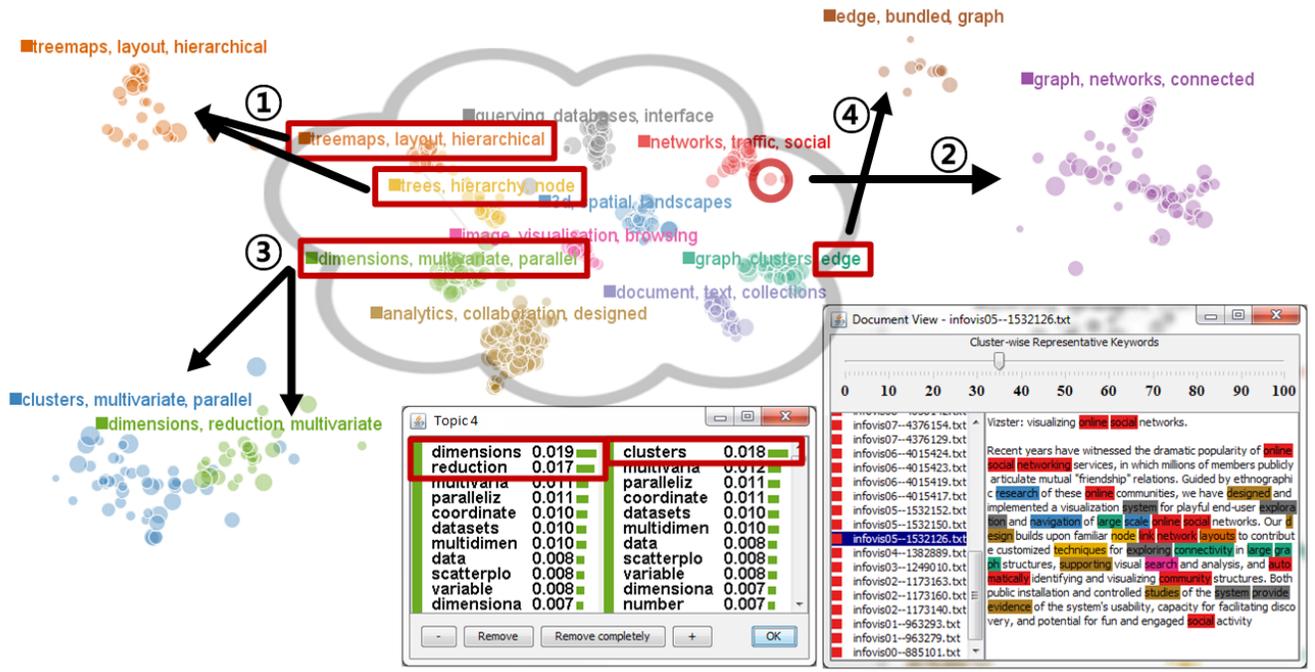


Figure 1: An overview of UTOPIAN. Given a scatter plot visualization generated by the modified t-SNE, UTOPIAN provides various interaction capabilities: (1) topic merging, (2) document-induced topic creation, (3) topic splitting, and (4) keyword-induced topic creation. Additionally, the user can refine topic keyword weights (not shown here). The document viewer highlights the representative keywords from each topic.

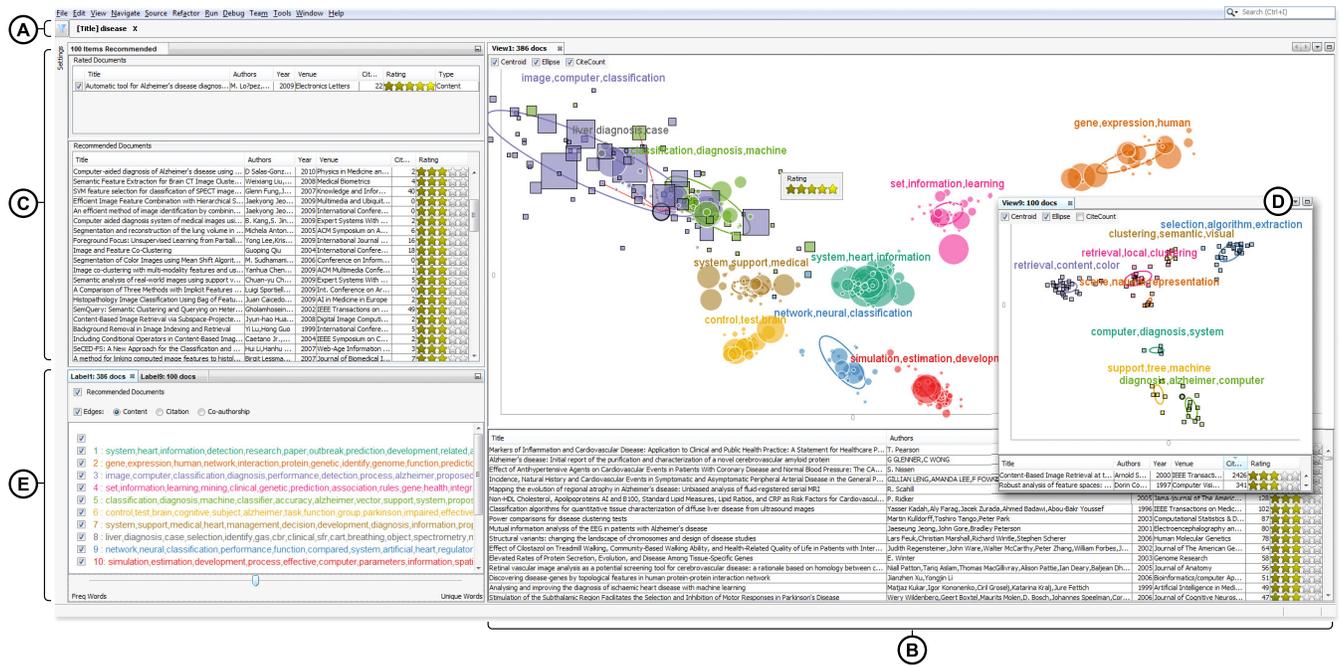


Figure 2: An overview of VisIRR. The user can start by issuing a query (A) (e.g., a keyword 'disease'). VisIRR visualizes the retrieved documents (circles) in a scatter plot and a table view (B) along with a topic cluster summary (E). Node size encodes citation count. The user can assign his/her preference in a 5-star rating scale to documents of interest. Based on the user rating, the system provides a list of recommended items (C), which are also projected back (rectangles) to the existing scatter plot view so that the consistent topical perspective can be maintained. For better understanding, the user can apply computational zoom-in on recommended items providing a much clearer summary (D).