VisIRR: Visual Analytics for Information Retrieval and Recommendation with Large-Scale Document Data

Jaegul Choo*, Changhyun Lee†, Hannah Kim*, Hanseung Lee‡, Zhicheng Liu∨, Ramakrishnan Kannan*, Charles D. Stolper*, John Stasko*, Barry L. Drake∧, Haesun Park*

*Georgia Institute of Technology, †Google Inc., ‡University of Maryland, ∨Adobe Research, ∧Georgia Tech Research Institute

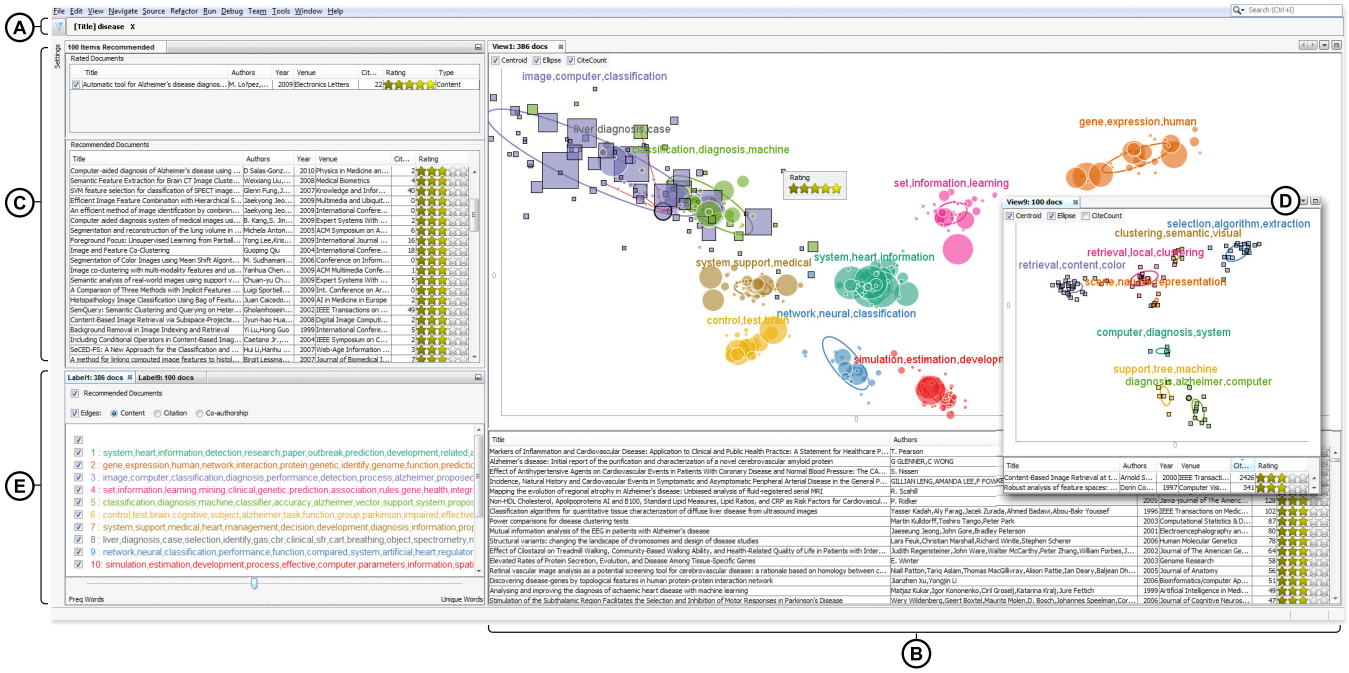


Figure 1: An overview of VisIRR. Starting with a user-initiated query (A) (e.g., a keyword 'disease'), VisIRR visualizes the retrieved documents (circles) (B) along with a topic cluster summary (E). A node size encodes a citation count, and a color represents a cluster index. Now, a user can assign his/her preference in a 5-star rating scale to the documents of interest. VisIRR then recommends potentially relevant documents (C), which are projected back as rectangles to the existing view. Optionally, VisIRR generates a new visualization on recommended items, which provides a much clearer summary of them (D).

Abstract

We present VisIRR, an interactive visual information retrieval and recommendation system for large-scale document data. Starting with a query, VisIRR visualizes the retrieved documents in a scatter plot along with their topic summary. Next, based on interactive personalized preference feedback on the documents, VisIRR collects and visualizes potentially relevant documents out of the entire corpus so that an integrated analysis of both retrieved and recommended documents can be performed seamlessly.

Keywords: Recommendation, document analysis, dimension reduction, clustering, information retrieval, scatter plot.

1 Introduction

Various visual analytics systems for document data have been proposed, e.g., In-Spire [4]. However, when they are large, e.g., millions of documents, visualizing all of them is not effective, and thus one has to first reduce them by filtering operations, e.g., keyword search, before visualization. However, such operations may exclude some of potentially relevant documents to users. In response, we propose VisIRR, an interactive **Vis**ual **I**nformation **R**etrieval and **R**ecommender system for large-scale document data, which effectively combines traditional query-based information retrieval and personalized recommendation that can interactively expand the document set in users' scope. In the following, we show how the system works using several usage scenarios and briefly describe the analytical approaches used in the system.

*e-mail: jaegul.choo@cc.gatech.edu

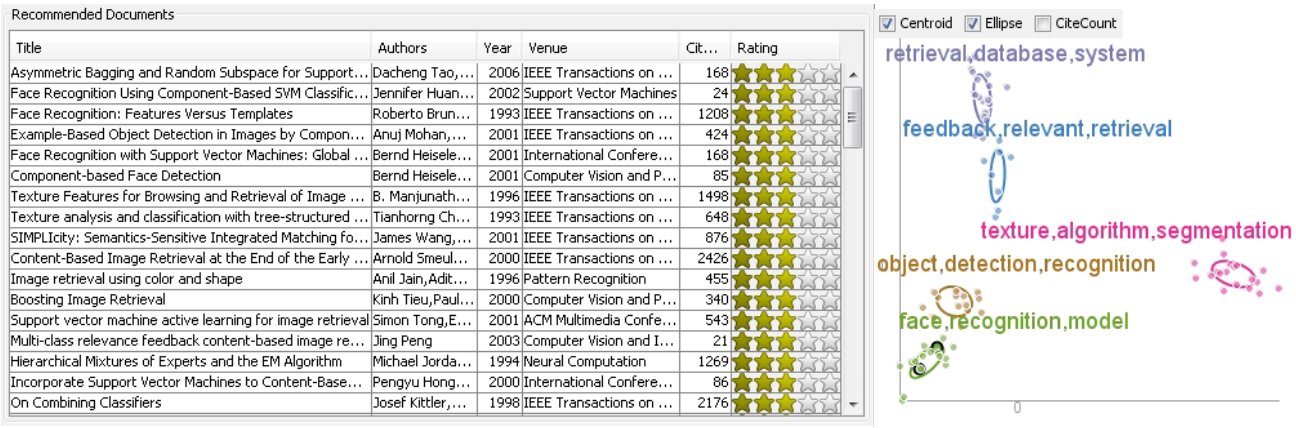
IEEE Symposium on Visual Analytics Science and Technology 2014
November 9-14, Paris, France
978-1-4799-6227-3/14/$31.00 ©2014 IEEE

2 Usage Scenarios

VisIRR[1] currently contains more than 400,000 academic papers and books published in the computer science domain. As shown in Fig. 1, a user can start with a particular query, e.g., a keyword 'disease'. Then the system visualizes the retrieved documents in a scatter plot form, together with their topic summary.

Content-based Recommendation. While exploring, the user finds an interesting paper 'Automatic tool for Alzheimer's disease diagnosis using PCA and Bayesian classification rules' based on his/her interest and assigns it a five-star rating. Now, VisIRR collects potentially relevant documents out of the entire corpus and visualizes them (Fig. 1(B)(C)). By examining the recommended

[1]A demo video: `http://tinyurl.com/visirr`.



(a) The top-ranked recommended documents

(b) The scatter plot of recommended documents

Figure 2: Citation-based recommendation results obtained by assigning a 5-star rating to the paper, 'Automatic classification system for the diagnosis of Alzheimer disease using component-based SVM aggregations'. VisIRR recommends relevant papers mostly with high-citation counts.

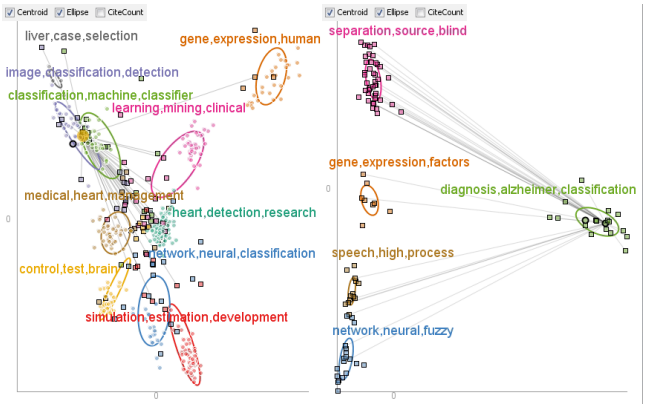
documents (Fig. 1(D)), the user can now see that research on Alzheimer's disease mainly involves automatic classification and image analysis.

Citation-based Recommendation. Now the user wants to know representative papers relevant to particular papers, and thus s/he changes the recommendation type as 'citation' and gives a five-star rating to the paper, 'Automatic classification cystem for the diagnosis of Alzheimer disease using component-based SVM aggregations'. As a result, the recommended items are shown to be highly-cited (Fig. 2(a)). Furthermore, from their own visualization with a topic summary (Fig. 2(b)), these highly-cited papers are related to image retrieval, object recognition, face recognition, and texture analysis. Note that this type of recommendation is not easily obtainable by a simple keyword search since the recommended documents do not share a common keyword and they are only indirectly related through citation networks.

Co-authorship-based Recommendation. Alternatively, the user can get another type of recommended documents by changing the recommendation option to 'co-authorship'. With this option, VisIRR can reveal what other topics or areas the authors of this paper conduct their research in. The recommended documents (rectangles in Fig. 3(a)) are distributed among many different existing topics. However, a new topic summary for recommended documents, as shown in Fig. 3(b), indicates that the authors of the rated paper have written papers in the fields of blind source separation, gene expression, speech processing, and neural networks, in addition to Alzheimer's disease diagnosis in which the initially rated paper was about. If the user worked in a similar domain to Alzheimer's disease diagnosis, such knowledge could lead the user to expanding his/her own research to these domains.

3 Computational Methods

VisIRR adopts various computational methods. To visualize document data along with a topic summary, VisIRR performs a clustering and a dimension reduction steps by using nonnegative matrix factorization [2] and linear discriminant analysis [1], respectively. For recommendation, VisIRR performs a heat-kernel-based graph propagation algorithm [3] on a k-nearest neighbor cosine similarity, a citation, and a co-authorship graphs. The bag-of-words vectors of individual documents and these three graphs have been precomputed and efficiently stored in a spare matrix format.



(a) The scatter plot of retrieved and recommended documents

(b) The scatter plot of recommended documents

Figure 3: Co-authorship-based recommendation results obtained by assigning a 5-star rating to the paper, 'Automatic classification system for the diagnosis of Alzheimer disease using component-based SVM aggregations'. Edges show direct co-authorship relations from the rated document.

4 Conclusions and Future Work

We presented VisIRR, a large-scale document visual analytics that combines information retrieval and recommendation based on personalized preference feedback. We plan to conduct a user study to evaluate the utility of our recommendation capabilities.

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

References

[1] J. Choo, S. Bohn, and H. Park. Two-stage framework for visualization of clustered high dimensional data. In *IEEE VAST*, pages 67 –74, 2009.

[2] J. Choo, C. Lee, C. K. Reddy, and H. Park. UTOPIAN: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE TVCG*, 19(12):1992–2001, 2013.

[3] F. Chung. The heat kernel as the pagerank of a graph. *PNAS*, 104(50):19735–19740, 2007.

[4] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In *IEEE InfoVis*, pages 51–58, 1995.