A Visual Approach for Interactive Expertise Finding and Exploration

Ehsan Sherkat
Dalhousie University, Canada
ehsansherkat@dal.ca

Seyednaser Nourashrafeddin
Dalhousie University, Canada
nourashr@cs.dal.ca

Rosane Minghim
University of São Paulo, Brasil
rminghim@icmc.usp.br

Evangelos Milios
Dalhousie University, Canada
eem@cs.dal.ca

ABSTRACT

Members of knowledge-intensive organizations (e.g. universities) is publishing considerable amounts of unstructured and semi-structured textual data such as research papers, tutorials, technical reviews. This data is a comprehensive description of both the members' expertise and the organization's goals. Exploring this document collection helps the organization to find the appropriate skillsets and expertise to improve their business. As the size of this document collection increases, analyzing this data to gain an insight becomes quite challenging. One solution is applying a clustering and topic modeling technique to these documents in order to help the organization to quickly find its major direction in the past months. Even the best clustering approach is not completely satisfactory and may produce some unintuitive results. To tackle this problem we have designed an interactive document clustering and keyword mining system using visualization which can be utilized to find the skillsets of an organization. Keyterm-based interactive approaches are arguably very intuitive for the users to guide the text clustering process and adapt the clustering results to the various applications in the text analysis. To demonstrate the usefulness of the proposed system, we have conducted a case study on a collection of data belonging to the School of Law and we supported the exploration process of the School in identifying its research directions over the past thirty years.

1. INTRODUCTION

The first step for expertise finding is uncovering the organization's set of expertise to recommend appropriate experts demanded by the organization. This set of expertise is mostly mined from documents such as project reports, human resource data, technical articles, or the employee's resume and descriptions [1]. One approach for extracting the expertise set from an organization's documents is clustering.

Recently using user supervision for interactive document clustering gained significant attention [3, 2]. In [3] an interactive clustering system called iVisClustering is introduced. In that system, user can impact the clustering result by changing the term weights. One of the problems associated with changing the weight of terms is the difficulty of guessing the appropriate weight by the user which may guide the clustering result. In our system, the user can interact with the clustering algorithm by only selecting topic keyterms of each cluster. A single term can appear in a number of documents and have considerable impacts on the result. The user of iVisClustering system is not able to change the number of clusters and is only allowed to split and merge the existing clusters. In [2] the user can define the must-link and cannot-link constraints over the words in order to change the result of clustering. In that system, the main focus is on the algorithmic side of clustering to reflect the user's feedback in a real time; while less attention has been given to the visualization techniques that support the user in finding appropriate words to guide the clustering process.

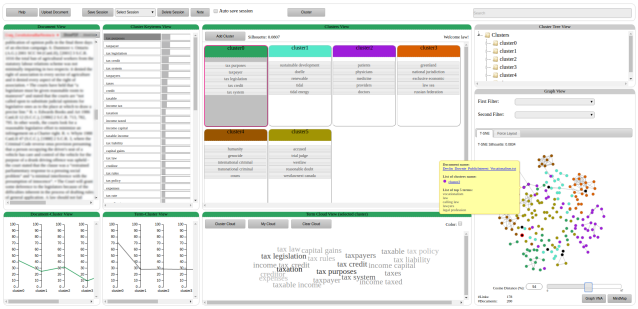

Figure 1: General view of visualization modules of proposed system.

In this research work, we have demonstrated the usefulness of combination of document clustering, visualization, and user interaction for finding the major talents and needs of organizations. We use a real dataset as a case study to support the effectiveness of the proposed system.

2. SYSTEM COMPONENTS

The proposed system consists of two major parts. The first part is designed to process the raw textual documents and automatically clustering data into a reasonable number of clusters automatically. The second part is visualizing the results to provide a comprehensive view of the data. The user can benefit from different views and interact with the system.

In the processing phase, important terms of the document collection are extracted. The dimension reduction techniques are used to filter out noisy and useless terms. A novel approach is employed to determine the initial number of clusters. At the end, the data is clustered and the top key terms and phrases of each cluster are displayed as the

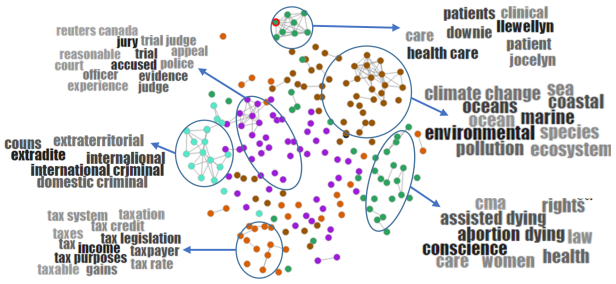

Figure 2: Initial Clustering results. Each node is a document. Color of the node represent its cluster.

major topics of each cluster based on LDC algorithm [4].

Different views are designed to give the user an elegant insight into the the data collection and enable the user to easily migrate from the document level to more general cluster view and conversely. Screenshot of the system's visual components[1] is shown in Figure 1. In the middle, the clusters which contain the major subjects of the collection are represented with the top five key phrases of each cluster. On the left, the content of the selected document, the list of top key phrases of the selected cluster, and their relatedness to other clusters are represented. On the right, user benefits from the graph view of the document collection, where each node is a document and the links are determined based on the Cosine distance between documents. The position of nodes is calculated based on t-SNE [5] algorithm and the color of nodes corresponds to the computed clusters. On the bottom, there is a Term Cloud view which can represent the cluster terms, document terms or the terms of a set of user selected documents from the Graph view.

The user can add, remove or merge clusters based on his/her preferences. The important interaction between the user and the system is determining the key-phrases of the clusters. The user is able to add terms from document view, Cluster keyterm view, Term Cloud and or search a term in document collection and then add it to the desired cluster.

3. CASE STUDY FINDINGS

This section highlights the effectiveness of the proposed system in a case study conducted on a real dataset related to the Dalhousie University School of Law. The main task is understanding major research directions of this school over its recent history for faculty development. This dataset contains research papers, law reports and trial cases.

The initial result is based on the document clustering with automatically determining the number of clusters. As depicted in Figure 2, there are five different clusters showing the major direction of this school are "International Criminal", "Tax Legislation", "Climate Change", "a trial case" and "health care and abortion law". The user can change the Cosine distance threshold and consequently the number of links between documents. The user can filter the graph view based on a topic, name or year. Figure 3b is the result of filtering based on the name of a faculty member. This faculty member research interest is on legislation about "Climate Change" issues.

The green cluster has two separate groups of documents

[1]Online demo of the system: demeter.research.cs.dal.ca/IC/

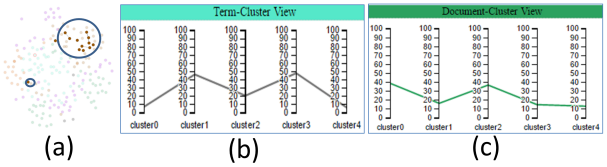

Figure 3: a. Document cluster view of a document. b. Filtering graph view. c. Term Cluster View.

as indicated in Figure 2. Comparing Term Clouds of these two groups proves that they are representing two different issues, beneficial to split the cluster. The user can perform this split by creating a new cluster and adding terms such as "abortion law" to this new cluster.

The user can also take benefits from the Term Cluster view which illustrates the relatedness of each term to the clusters. The Term Cluster view of a term which is related to two clusters is shown in Figure 3c. These clusters may join or split by adding or removing this term. The Document Term view of a selected document which reflects its relatedness to each of clusters is shown in Figure 3a. This view also can help the user to find the appropriate terms to join or split clusters and creating a new one.

4. CONCLUSION AND FUTURE WORK

In this demo paper, we introduced a general system for the interactive clustering of a document collection which can be employed to identify the key direction of knowledge-intensive organizations such as universities in order to find their skillsets demands. It is a need to track the impact of changes and user interactions on the result of clustering. In the current system, the user can visually compare the result of the clustering after each interaction and also compare the Silhouette score of the clustering. In the future, we are looking to use a more intuitive approach to inform the user of his/her impact (negative or positive) on the clustering results. We are also looking forward to add some document level interaction to the graph view. This will help the user to select several documents and change their cluster label or perform further clustering in order to have a hierarchy of clusters. Linking the extracted key topics from the proposed system to the actual skillsets demands is the next step. An application of the system is, a person looking for works on "International Criminal" needs to have certain skills about "international laws".

5. REFERENCES

[1] I. Guy, U. Avraham, D. Carmel, S. Ur, M. Jacovi, and I. Ronen. Mining expertise and interests from social media. WWW '13, pages 515–526, 2013.

[2] Y. Hu, J. Boyd-Graber, B. Satinoff, and A. Smith. Interactive topic modeling. *Machine Learning*, 2014.

[3] H. Lee, J. Kihm, J. Choo, J. Stasko, and H. Park. iVisClustering: An Interactive Visual Document Clustering via Topic Modeling. *Computer Graphics Forum*, 2012.

[4] S. Nourashrafeddin, E. Milios, and D. Arnold. Interactive text document clustering using feature labeling. DocEng '13, 2013.

[5] L. Van Der Maaten. Accelerating t-sne using tree-based algorithms. *J. Mach. Learn. Res.*, 2014.