Poster:Visual Analysis of Stream Texts with Keywords Significance

Jamal Alsakran
Kent State University

Ye Zhao*
Kent State University

Dongning Luo
University of North Carolina - Charlotte

Jing Yang†
University of North Carolina - Charlotte

ABSTRACT

Stream text collections demand a thorough sophisticated method to explore the massive data and effectively extract valuable information. We propose a force-directed method to dynamically visualize stream text documents. The method captures events as they occur, injects them into the system, and clusters them on-the-fly. Typically, keywords are used to calculate similarity and accordingly steer the clustering results, we propose keywords significance to magnify the importance of certain keywords on-the-go. The user can interactively manipulate the significance of keywords and see the impact on the system. A real dataset example is provided to show the effectiveness of our method.

1 INTRODUCTION

In this paper, we present a method to dynamically visualize constantly evolving texts. Our approach employs force-directed placement to resemble documents visualization to vertices and forces between them. Physically, the method models the potential field between vertices and presents forces that tend to group similar documents, and move dissimilar ones apart. The method receives streaming text documents such as news or emails and visualizes them in a 2D plot. The visual result shows documents grouped in related topics and placed in relative view among all other documents.

Documents are represented as vectors of keywords. Keywords are special words that qualify to be representatives of the whole documents. Typically, similarity between documents is calculated using keywords. When a document emerges and keywords are extracted, its similarity stays still over time. In our approach, we propose keywords significance to provide dynamical system furnished with the capability to adjust the similarity as the system is running. A keyword significance present the importance of that keywords at a certain time. Intuitively, keywords with higher significance are highlighted by increasingly attracting more of the documents that have those keywords. To sustain interactivity, keywords significance can be adjusted based on the users preference or can automatically extracted from the hot topics.

Stream text visualization has recently received a great attention. ThemeRiver [4] depicts the strength changes of individual keywords as currents within a river flowing a long time axis. Narratives [3] uses a temporal axis to visualize ways that concepts (keywords) have changed over time in weblog archives. Force directed placement has been employed in InfoSky [1] to enable users to explore large, hierarchically structured document collections. Chalmers et al. [2] propose a prototype to explore word-based information using the physical behavior of the containing particles.

2 OUR APPROACH

Our system receives constantly evolving text documents, groups them into clusters, and places them in a 2D plot based on their similarities, i.e. similar documents are drawn relatively close to each other and far from dissimilar ones. A text document is represented as a vector of keywords. Initially, the system captures recently occurred events, computes the similarity between every pair of documents, and place them randomly in a 2D plot. The pairwise similarity is mapped into a distance in our domain to determine the optimal distance between the documents. Iteratively, the documents start moving closer or farther until they approximately reach optimal distances. When a new document reaches, the system injects it randomly and continues the process. Specifically, the (documents) vertices will exert attractive and repulsive forces on each other, the forces cause the vertices to move closer or farther. At each step, the attractive and repulsive forces are calculated Eq. 1, then they are applied on the vertices to group them or fan them out. Subsequently, the total energy of the system is computed. The process continues until the total energy reaches a preset minimum threshold.

$$F_{ij} = (|p_i - p_j| - l_{ij})^2 \quad (1)$$

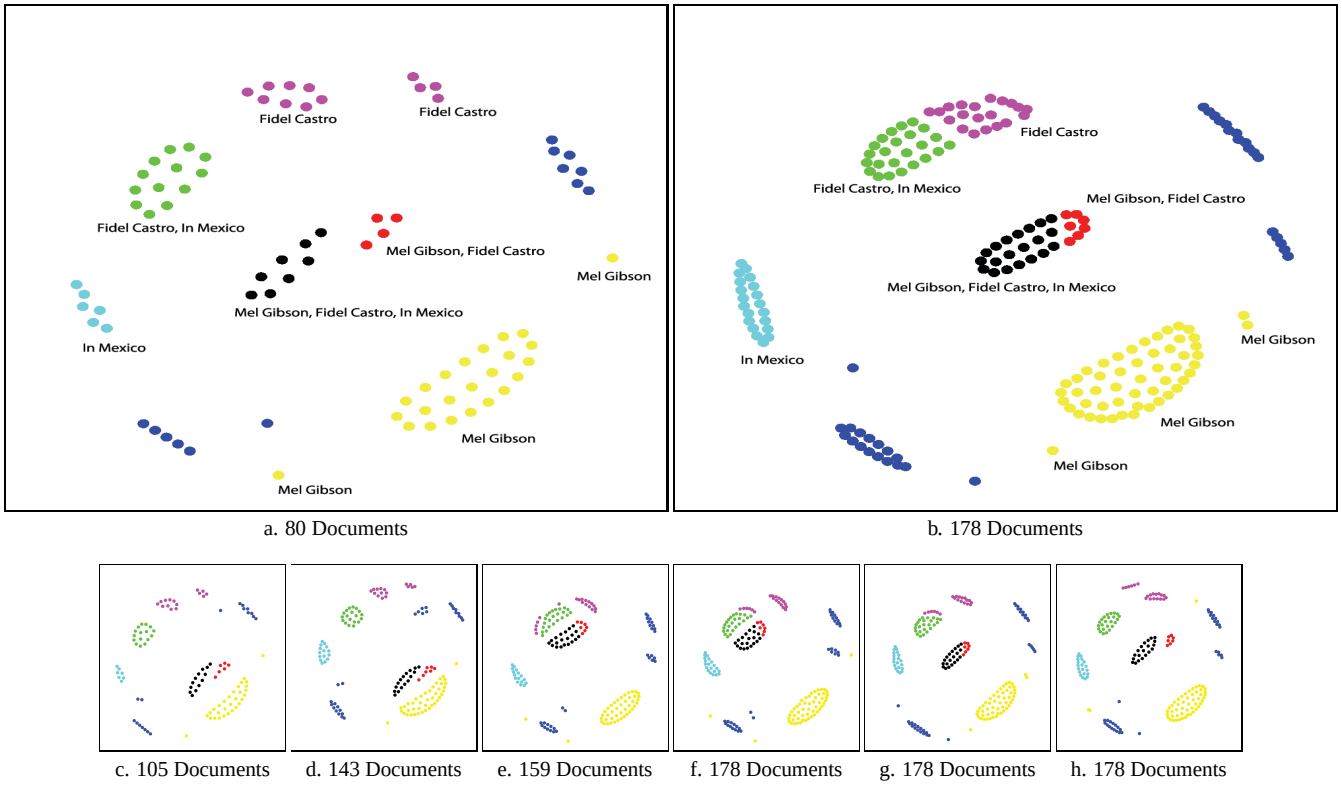
where $p_i - p_j$ is the spatial distance between d_i and d_j, respectively. l_{ij} is the optimal distance that is mapped from the similarity between d_i and d_j.

Usually, keywords are vital words that highly occur in the document, they represent a brief summary that conveys the topic of the document. Similarity is typically calculated from the keywords, e.g. cosine similarity. When a document enters the system, its similarity with other documents is calculated and stays fixed throughout the system. However, keywords that make up the similarity can have different importance and therefore participate variously. For example, cosine similarity uses weights to represent keywords importance. We propose a feature called Keyword Significance to provide the capability of altering keywords importance as the system runs. A keyword Significance retains the importance of a keyword at a certain time. Interactively, keywords significance can be modified to achieve better clustering result and explore different trends of the data. To assign keywords significance, the system investigates the data to extract vital keywords that occur the most in the data and raise their similarity. Alternative, a user can manipulate the significance of specific keywords of his interest to widely explore the documents contain those keywords, and to adjust the clustering results.

3 EXAMPLE

Figure 1 shows an example of applying our method to dynamically visualize text documents. The dataset represents CNN news over one day Aug. 1, 2006, it contains 178 documents, and 19 keywords. Figure 1(a) shows the system with 80 documents, it displays how similar documents come together and settle adjacent to each other while they keep relative distance form dissimilar ones. The dataset contains 19 keywords, we run some statistics on the keywords and decided to investigate 3 keywords which occur the most in the dataset, namely Mel Gibson, Fidel Castro, and In Mexico. In figure 1(a) we use different colors to highlight the documents that contain those keywords. Blue vertices represent documents that have none of the 3 keywords. As of figure 1(a) we don't employ keywords significance, and that is achieved by setting keywords significance equally to 1. Figure 1(b) shows a late stage of the system after all documents have been entered. To show the impact of keywords significance we assign higher significance to Mel Gibson and

*e-mail: {jalsakra, zhao}@cs.kent.edu
†e-mail:{dluo2,Jing.Yang}@uncc.edu



a. 80 Documents b. 178 Documents

c. 105 Documents d. 143 Documents e. 159 Documents f. 178 Documents g. 178 Documents h. 178 Documents

Figure 1: A dataset contains 178 documents, and 19 keywords. The figures show different snap shots of the system

Fidel Castro, and set it to 10. As figure 1(b) shows, including the significance effects the clustering results by increasingly attracting more of the documents that contain the keywords of interest. Moreover, users can explore a wider perspective of that data by manipulating significance. For example, increasing the significance of Mel Gibson causes all the documents that contain it (yellow vertices) to be attracted, similarly, increasing the significance of Fidel Castro groups further all the documents that have it (green and purple vertices), those documents were previously spread out. The documents that share both of the keywords (red and black vertices) are placed somewhere in the middle between the two groups.

Figures 1 (c), (d), (e), (f), (g), and (h) present different snap shots of the system (See a video in the supplementary material). Figure 1(c) shows the system with 105 documents, we a little bit increase Mel Gibson significance to 5, clearly, all the documents that have the keyword Mel Gibson (yellow, red, and black vertices) become closer to each other. In figure 1(d), we increase Mel Gibson significance further to 10, as we can see, yellow, red, and black vertices come even closer and form a group. On the other hand, other documents that don't contain Mel Gibson are not effected. Similarly, figure 1(e) shows the clustering results of 159 documents and Fidel Castro significance set to 10. In this figure, all the documents that have Fidel Castro (green, purple, red, and black vertices) are gathered to a group. To further explore the dataset, we set Fidel Castro and In Mexico significance to 10 in figure 1(f), as we expect, the documents that have either Fidel Castro or In Mexico (purple vertices and cyan vertices, respectively) are forming separate groups, while the documents that share both (green, red, and black vertices) are collectively grouped in between to reflect that they belong to both. Figure 1(g) displays the results of 178 documents with Mel Gibson, Fidel Castro, and In Mexico significance set to 10, obviously, the documents that have merely one of those keywords are forming distinguished clusters, also the documents that share more than a keyword are placed relatively in between, for instance, black vertices are placed in the middle of the three clusters since they have all of them, while green vertices are positioned between Fidel Castro and In Mexico clusetrs. Finally, figure 1(h) shows with no significance employed how the documents are grouped based on their explicit similarity.

4 Conclusion

We present a method to dynamically visualize stream texts. The system places the documents in a 2D plot based on their similarity. In addition, we propose keywords significance to dynamically assign importance to keywords. Effectively, keywords significance allows the users to track certain keywords upon their preferences by changing the configuration of the system The computational demand of force-directed placement encourages us to seek out faster method to integrate the system at each step. We also plan on developing an automatic way to extract the significance of the keywords based on the status of the system.

References

[1] K. Andrews, W. Kienreich, V. Sabol, J. Becker, G. Droschl, F. Kappe, M. Granitzer, P. Auer, and K. Tochtermann. The infosky visual explorer: Exploiting hierarchical structure and document similarities. *Information Visualization*, 8(3):166–181, Dec. 2002.

[2] M. Chalmers and P. Chiston. Bead: Explorations in information visualization. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 330–337, 1992.

[3] D. Fisher, A. Hoff, G. Robertson, and M. Hurst. Narratives: A visualization to track narrative events as they develop. In *IEEE Symposium on In Visual Analytics Science and Technology*, pages 115–122, 2008.

[4] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. Themeriver: Visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):9–20, Jan. 2002.