Time aware visualization of document collections

Aretha Barbosa Alencar
Institute of Mathematics and Computer Science
University of São Paulo, São Carlos, SP, Brazil

Abstract

Scientific articles are the major mechanism for researchers to report their scientific results, and several visualization techniques have been proposed specifically to support the exploratory analysis of paper collections (Chen, 2006; Sci2 Team, 2009). An approach being pursued at our research group in Brazil is to generate exploratory visualizations of documents based on their textual content (Paulovich et al., 2008). In these visual maps, each document is identified as a visual marker (e.g., a point or circle) in the two-dimensional space, and (ideally) documents close in the map share similar content, whereas distant documents are uncorrelated in terms of their content. The maps allow visually identifying groups of highly related documents (addressing similar topics) and frontiers between groups. Content-based similarity maps are constructed with multidimensional projection techniques, capable of placing individual data instances in a (typically two or three-dimensional) visualization space, while retaining some data property of interest (e.g., content similarity).

Analysts exploring a set of papers on a research field would find useful to detect and visualize topical events such as the emergence of new topics; the gradual disappearance of topics when researchers stop publishing articles related to that topic; and the merging of two or more topics into new ones. However, the visual maps obtained so far by our research do not explicitly address the temporal evolution of a collection, i.e., they do not convey changes in the similarity relationships occurring along time. We believe such changes are good indicators of topical events, since one or multiple topics are (should be) typically associated to sub-sets of similar documents.

In this project, we intend to prove that general content-based approach may be adapted to generate visualizations that convey such information. Moreover, we hypothesize that the changes in the similarity relationships along time are good indicators of interesting topical events. For this we investigate: (i) how to adapt a content-based projection technique to construct gradual maps of a time-stamped collection, while preserving user context at successive time-stamps, i.e., the maps should not change drastically from one time stamp to the next; (ii) how to identify topics on groups of similar documents based on the sequence of document maps generated and analyze the similarities/dissimilarities of the extracted topics along time; (iii) how to meaningfully integrate the exhibition of topics in the maps, while visually emphasizing important topical changes along time.

We present our tool that currently implements a dynamic multidimensional projection to generate a dynamic document map. The approach adopted to generate a dynamic document map is based on a backward scheme. Given a collection of documents divided into a list of batches \( D = \{D_1, \ldots, D_t\} \), according to the publication year, the idea is to generate the dynamic projection in a backward manner, which projects first the full collection of documents adding it to a list of projections and progressively removes from this projection the documents belonging to the most recent batches (batches at the end of the list). After doing the documents removal, the remaining documents that had in their neighborhood removed documents must be projected again to update their position. The other documents must remain at their current position. With these actions, a new projection is generated and added to the list of projections. By using information from the previous projection, we seek to maintain a spatial coherence between subsequent projections. After the list of projections is obtained, the order is reversed in order to make an animation that depicts the changes in the similarity relationships in subsequent projections. It is expected that documents that address similar subjects and are positioned at a certain region at the projection stay approximately in the same region in the following projection. This type of behavior favors the perception of the topics evolving.

Video available at: http://www.icmc.usp.br/~aretha/TopicEvolutionVis.avi

The author acknowledges the orientation of Prof. Katy Börner, who advises the PhD student Aretha B. Alencar during her current internship at the Cyberinfrastructure for Network Science Center at Indiana University - Bloomington campus.
References


Available at [http://sci.slis.indiana.edu/sci2](http://sci.slis.indiana.edu/sci2)