ChronAtlas: A Visualization for Dynamic Topic Exploration

Nan Cao* Yu-Ru Lin† David Gotz‡ Jimeng Sun‡ Huamin Qu*

ABSTRACT

Documents in rich text corpora such as digital libraries and social media often contain complex information. These data resources are huge in amount, dynamic in nature and contain multifaceted information. This poster presents ChronAtlas, a visual analytic technique for visually exploring topics in multifaceted dynamic data. ChronAtlas simultaneously visualizes the topic distribution of the underlying entities from one facet together with keyword distributions that convey the semantic definition of each cluster along a secondary facet. ChronAtlas combines several visual techniques including 1) topic contour clusters and interactive multifaceted keyword topic rings, 2) a global layout optimization algorithm that aligns each topic cluster with its corresponding keywords, and 3) an optimal temporal network summarization and clustering algorithm that renders evolution trends of clusters.

1 INTRODUCTION

An increasing amount of information is becoming available through rich context media over the Internet such as the online digital libraries and social medias. New challenges occur during the analysis of these data resources which are huge in amount, dynamic in nature, contain multifaceted information and complex relationships. Consequently, a framework that facilitates finding information evolution trend from such data collections are needed.

Information visualization can be of great value in addressing these issues. Prior work such as [1, 3] mainly focuses on the information dynamics of a single facet. Several multifaceted visual systems are designed recently. For example, ContexTour [4] maps the dynamics of multifaceted content through the smooth clustering technique with little consideration on relational patterns. FacetAtlas [2] focuses on a static multifaceted network, while the dynamic trend of the information is obscured.

In this poster, we present *ChronAtlas*, a visual analysis system, which reveals the information evolution trend of the DBLP dataset over multiple facets. In this system, we propose a new data model as well as a new data transformation pipeline to automatically convert unstructured documents into a multifaceted structured network. We analysis the evolution trends of the network based on a new dynamic graph summarization and clustering algorithm. Finally, we visualize the transformation and analysis results in a hybrid visualization which combines clustered tag could with a radial tag view.

2 SYSTEM OVERVIEW

ChronAtlas contains three major components. The data transformation module converts the document corpora into the dynamic multifaceted entity relational data model. The dynamic analysis module detects the evolution trends and significant changes of the data. The visualization module allows users to interpret the analysis results and to explore the evolution trend in an interactive way.

2.1 Dynamic Entity Relational Data Model and Analysis

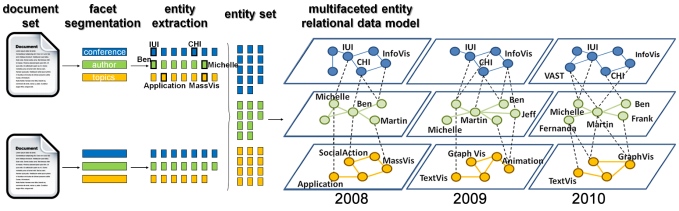

Figure 1: Dynamic multifaceted entity relational data model.

Figure 1 illustrates the document transformation pipeline and the data model. Generally speaking, given a set of documents, we first segment them into multiple information facets. For example, a paper in the DBLP dataset contains information from three facets: authors, topics and conferences. After that, the entities are extracted from each facet snippet. For example, the conference facet contains the entities such as "InfoVis" and "Vast", and the topic facet contains the keywords such as "TextVis" and "Graph Drawing". In the third stage, connections between these entities are established using two types of relations: internal relations and external relations. An *internal relation* connects entities within the same facet. For example, the entities "InfoVis" and "IUI" can be connected by an internal relation according to their cross conference citations. An *external relation* is a connection between entities from different facets such as the "topic - conference" relation. Finally, we organize the data in a temporal order over the facets and segment the data with a specified interval.

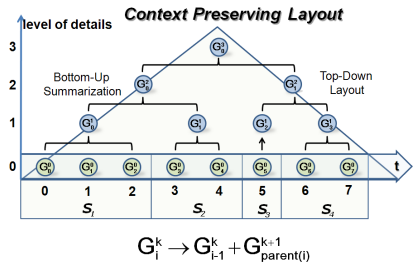

Figure 2: The graph summarization and context preserving layout.

Based on the above data model, we introduce a new visual analysis method that reveals the evolution trends of a dynamic network, detects the dramatic changes as well as provides a smooth clustering over time. As illustrated in Figure 2, it contains two major parts: a bottom-up hierarchical graph summarization and a top-down context preserving clustering and layout.

The summary procedure tries to find segmentations in the given graph sequence target on the following object:

$$sse(S_i) = min \sum_{s_i \le j \le e_i} \|G_j - G_i^k\|^2 \quad (1)$$

*Nan Cao and Huamin Qu are with the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology. E-mail:{nancao, huamin}@cse.ust.hk

†Yu-Ru Lin is with the Institute for Quantitative Social Science, Harvard University and College of Computer and Information Science, Northeastern University. E-mail: yuruliny@gmail.com

‡David Gotz and Jimeng Sun are with IBM T.J. Watson Research Center. E-mail:{dgotz, jimeng}@us.ibm.com

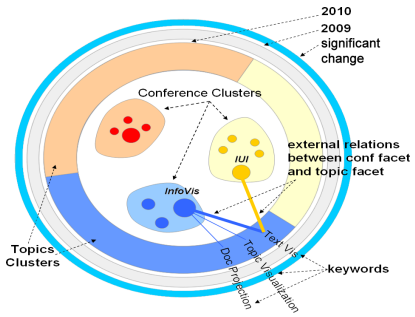

Figure 3: Visual encoding. In this example, we focus on the year 2010 and select *conference* as the primary facet. All the conferences such as "InfoVis" within this time slot are clustered and visualized in the center. *Topic* is selected as the context facets. Its related keywords occurred in 2010 such as "Text Vis" and "Topic Visualization" are visualized along the focused time ring. These entities appear in the blue wedge of the context ring because they are the common keywords of the conferences in the blue cluster. Other time slots are collapsed and visualized as thinner rings such as the rings for year 2009. When click, these collapsed rings will be expand as the focus and accompanied with the dynamic change of the center view. Some significant changes occur at the earlier time, the related time ring is thus highlighted by cyan which provides additional visual cues.

where S_i is a segmentation, G_j is a graph sample in S_i and G_i^k is the summary graph of S_i which can be understood as the union graph that aggregates all the nodes and edges in G_j. Once the graphs are summarized hierarchically, they are clustered and laid out in a top down procedure using the following stress:

$$layout(G_i^k) = \min\{\alpha \cdot \sum_{p<q} \frac{1}{d^2}(||G_i^k(p) - G_i^k(q)|| - d_{pq})^2 + \beta \cdot \sum_p ||G_i^k(p) - G_{i-1}^k(p)||^2 + \gamma \cdot \sum_p ||G_i^k(p) - G_j^{k+1}(p)||^2\} \quad (2)$$

where $G_i^k(p)$ indicates the coordinate of node p based on the layout of graph G_i^k; $\alpha, \beta, \gamma \in (0,1)$ are the weights that balance the three main terms and $\alpha + \beta + \gamma = 1$. The first part of this model is a standard KK layout and the second and third part try to make the layout smooth by adding controls according to its summary graph and its success graph.

2.2 Visualization Design

We visualize the dynamic multifaceted entity relational data model by following the visual encodings illustrated in Figure 3[1].

First, we use rings to encode the time slots. At any time, only one time slot is selected as the focus. Others are represented as surrounding time rings in the temporal order. The focused time slot has a wider ring than other time slots which serves as the history contexts. The time slots contains significant changes are highlighted in cyan. Second, one facet, at any given time, is selected to serve as the *primary facet*. Entities in the primary facet (which we call primary entities) inside the current time slot are rendered as nodes and arranged within the central region of the view. A contour is further rendered to highlight the cluster information. Third, another *context facet* is also selected. The entities (called as context entities) in this facet are displayed as tags placed within a radial tag cloud on the current time ring. It provides secondary contextual information about each cluster. The tags are grouped based on the clusters identified along the primary topic facet. This forms wedge-shaped sections along the focused time ring with one wedge for each cluster. The size of each wedge indicates the size of the corresponding topic cluster, and the correspondence between cluster and wedge is captured using both color and position. The context entities are connected with the primary entities by their external relations. These linkages are colored by its primary entity's cluster and line thickness represents the number of topic entities related to the same keyword entity.

[1]See http://www.youtube.com/watch?v=rPNrwHWCekg for the screen capture video.

3 Case Study on DBLP

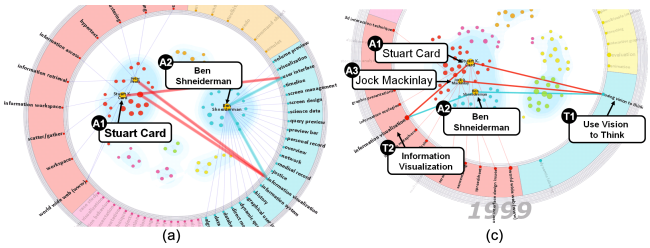

Figure 4: Case study on DBLP data.

Time plays an important role in the DBLP dataset as it captures the evolution of research topics and teams over several years. This case study examined changes in the InfoVis community from 1992 through 2002. As illustrated in Figure 4, we use author names as the topic facet, and paper keywords grouped by year as the keyword facets. The years are ordered allowing easy navigation through time using the keyword rings.

Exploring the data year by year, we found some interesting evolution patterns. In the first years, such as 1994, several isolated author clusters emerged. The largest were led by Ben Shneiderman and Stuart K. Card. Shneiderman's cluster focused most on interaction designs such as "dynamic query" and information exploration such as "information seeking and retrieval". In contrast, Card's group focused more on "graphical representation" and "explorative data analysis".

In 1996, researchers in both clusters began working on a few similar topics such as "Information Visualization" and "User Interface", as indicated by the common links to those keywords on Figure 4(a). However, as shown by the author clusters, the research communities were still not directly collaborating. However, by 1999, the clusters begin to merge. This merger, as shown in Figure 4(b), occurs around the time that Card and Shneiderman join as two of the co-authors on the book "Using Vision to Think".

Acknowledgements

The authors wish to thank the anonymous reviewers for their valuable comments. This work was supported in part by grant HK RGC GRF 619309 and an IBM Faculty Award.

References

[1] S. Bender-deMoll and D. McFarland. The art and science of dynamic network visualization. *Journal of Social Structure*, 7(2), 2006.

[2] N. Cao, J. Sun, Y. Lin, D. Gotz, S. Liu, and H. Qu. FacetAtlas: Multifaceted Visualization for Rich Text Corpora. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1172–1181, 2010.

[3] W. Cui, Y. Wu, S. Liu, F. Wei, M. Zhou, and H. Qu. Context-Preserving, Dynamic Word Cloud Visualization. *IEEE Computer Graphics and Applications*, 30(6):42–53, 2010.

[4] Y.-R. Lin, J. Sun, N. Cao, and S. Liu. Contextour: Contextual contour visual analysis on dynamic multi- relational clustering. In *SIAM Data Mining conference*, 2010.