Visualizing Data Trend and Relation for Exploring Knowledge

Ting-Yen Lee[*]
National Taiwan University

Chad Jones[†]
University of California at Davis

Bing-Yu Chen[‡]
National Taiwan University

Kwan-Liu Ma[§]
University of California at Davis

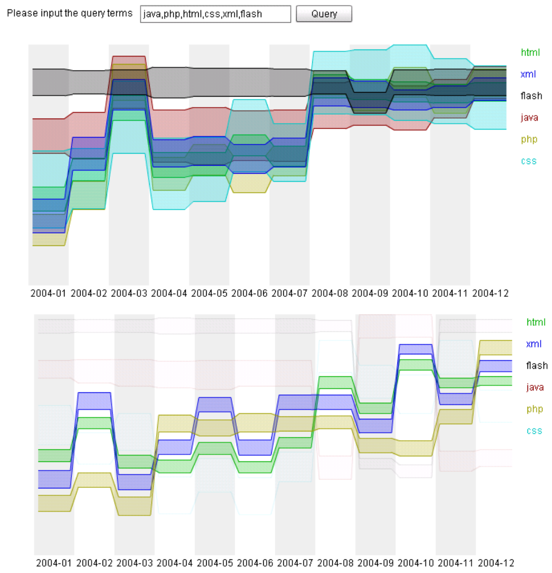
Abstract

When making decisions, it is often critical to understand the trends and relationships in the task-related information. Complicating matters is the time-varying nature of this information. Specifically, topics gain or lose prominence over time and relationships within the data grow, fade, or disappear altogether. This paper presents a novel visualization to help people address this information evolution. Our technique categorizes information into topics, which are visualized in a two-dimensional stream graph. As a topic waxes or wanes in importance, the associated stream in the graph gains or loses thickness. Further, the vertical distances between streams changes to indicate the strength of the relationship between the topics. We provide an interface which leverages multiple views to help users quickly switch between multiples data sets. We present a case study examining the tagging history from the social bookmarking site deliciȯus suggests that our visualization is helpful in concisely describing the aforementioned information evolution.

Index Terms: H.3.5 [Information Storage and Retrieval]: Online Information Services—Web-based services; I.2.4 [Artificial Intelligence]: Knowledge Representation Formalisms and Methods—Representations (procedural and rule-based)

1 Introduction

The enormity of human knowledge available on the Internet today can help people perform research and make decisions. To help users understand the retrieved knowledge, some present a list of discovered information (e.g. Google), while others provide a clustered view of the results based on derived relationships (e.g. Clusty [12]). Different from the traditional keyword extract technique, other sites, such as the social bookmarking site del.ici.ous, rely on user tagging to effectively categorize information on the web. Yet the above systems focus on the current state of the knowledge, and neglect an additional dimension: time. Knowing how topics evolve can more effectively direct our searching, and uncover useful information that might otherwise be overlooked due to current, transient weakened or strengthened relationships. It allows us to see, for example, how topics grow from insignificant to dominant (or vice-versa) in a certain subject. Therefore, we provide a new visual technique in which both aspects of the data are represented though time. A user enters the topics of interest, and our system produces a compact, two-dimensional visual representation in which time flows along the x-axis, and inter-topic relationship strength is represented along the y-axis. Note that our technique is unlike traditional trend views which redundantly use both stream width and the y-axis to represent importance.



Please input the query terms java,php,html,css,xml,flash

Figure 1: A snapshot of our system. For the visualization, different streams represent different topics, and each topic is displayed in different color. The width of the stream represents the importance at each time and the distance indicates the relations of the items.

2 Related Work

To represent the knowledge on the web, keyword is used in many famous search engine, such as Google and Yahoo!. Many studies based on keywords they found can build the topic model and analysis the trend [3, 4, 10]. The tagging system is also another popular way recently. Mika applied these tags in building the semantic knowledge on web [9]. While some people trying to find out the trend of the knowledge, some people try to visualize it. To depict the knowledge relation at a specific time period, generally people used node-link diagram [2, 11]. When it comes to visualize the time-series data, Kumar and Garland used the animation to represent the relation change of the knowledge [8]. Ahmed *et al.* growled the traditional node-link diagram into a 2.5D visualization, while adding Z-axis to represent the time series [1]. Moreover, ThemeRiver used a river to represent the topic flow [6]. However, it assumes the knowledge relation would not be changed over time.

3 System Overview

Figure 1 depicts an overview of our system. The visualization contains two parts. The upper part gives the ”General View”, and each topic can overlap with others. In this way, it is easy for people to visualize the overall clustering trend of the topics. The lower part is used for ”Focus view”, and every topic is listed separately, so that users may more easily focus on the two-two relation. For each visualization, the time moves right with the X-axis, and the Y-axis represents the relation of different streams in different time periods. Namely, the related topic placed near in the Y-axis. Each stream represents a topic and is filled by a unique color while the

[*] e-mail:lydian@cmlab.csie.ntu.edu.tw
[†] e-mail:cejjones@ucdavis.edu
[‡] e-mail:robin@ntu.edu.tw
[§] e-mail:ma@cs.ucdavis.edu

stream width is used to represent the importance in the related time period.

To test the visualization, we used the tagging log gathered from the most popular social bookmarking site, del.icio.us, in 2004. Based on Mika's work [9], we use tags to be a discussion topic on the web, and compute its importance by the tf-idf score. While considering the relation, we used the similar formula of computing the Google similar distance [5]:

$$NGD(x,y) = \frac{max(logf(x), logf(y)) - logf(x,y)}{logN - min(logf(x), logf(y))},$$

where x and y are the topics we interested, and $f(x)$ is the number of pages that we retrieved with the topic term x.

4 Stream Drawing Algorithm

The main problem of the visualization is really the layout. Since after we add the time dimension, we lose 1 dimension to visualize the relation. To solve this problem, we decided to display the relation in two different visualizations. The overlap visualization for displaying the grouping information, and the non-overlap visualization for visualizing the topic distance. By the 2-viewpoint visualization, the users could not only determine the cluster of the topics, but also understand how these topics affected others.

4.1 The General View

Make all streams overlap visually helps user detect the clustering effect. We adopted a loose restriction on forming the topic groups, so that the users would be able to consider all the possible related topics when they think of the group. Moreover, when located these topics, we want to preserve the most related relation in the main visualization. Therefore, we borrowed the idea from the "single-link clustering" algorithm [7].

The single-link clustering is a bottom-up algorithm. The clusters got after using this algorithm is quite loose. For a topic to join an existed cluster, it requires only the new topic related to any member of the current cluster. Therefore, for all two-two relation sets, we first execute the single-link algorithm. After the single-link clustering, we could have a tree structure, but if we directly draw the graph, we would find out the graph is not smooth, and leads to a bad readability. Therefore, we need to adjust the nodes, make the node with more descendants to stands in the center of the tree, and we could get a smoother visualization.

To take Figure 1-upper as an example, the user queried on topics: php, css, xml, java, html and flash. With the overlap visualization (the upper part), we can find that the topics may be divided into two clusters at first. The first group includes: html, xml, java, php, css, which are traditional web related topics. The second group is flash only (the gray stream), which is a technique that different from the previous group. But in the end of 2004, these two groups mere together, which means more and more people willing to combine or consider the two groups together. If we continuously detect such trend in the following time, we can predict the trends of the web skill development is to use the flash and traditional html together.

4.2 The Focus View

The main task of the focus view is to show the relation between topics. When we used the transitivity on the relation property, the displayed distance would be shorter than the real distance. In fact, we know the transitivity may not be true for the topic relations, therefore we need to be more careful on depicting the relation. Hence, for the focus view, we used a strict algorithm to prevent the uncertain circumstance. Relative to the overlap visualization, there is also a strict algorithm for building a strict connected cluster in information retrieval area, which is the complete-link clustering algorithm. [7] Similar to previous visualization, the complete-link algorithm also returns a tree to us. We have to make the similar modification on the tree structure to make the graph smooth.

Take the similar query as an example, since in the previous section, we found that PHP is something worth to give a deep look. After focusing on PHP in July, 2004. The highlighted items depict those related topics at that time. As we can see in Figure 1-lower part, "html" and "xml" often related to "PHP", except in October, 2004. Therefore, we find that October, 2004 is some time periods that worth to make a further investigation.

5 Conclusion And Future Work

The data presented in the del.ici.ous case study is flat - there exists no categorical hierarchy. In the future, we want to test on a data set with some categorical depth. Given such a hierarchical structure, we might uncover more complex, obfuscated trends.

Our system utilizes a multi-view display to present the data. Inter-cluster relations were displayed using the overlap visualization, while intra-cluster relations were displayed using the non-overlap variant. By highlighting the topics related to the selected topic for a given time period, the focus view helps to uncover underlying details. The interaction mechanisms described allow the revelation of even complex network relationships using our visualization.

We have presented a novel visualization for simultaneously examining the time-varying aspects of inter-topic relations and topic importance. Based on our case study of data scraped from del.ici.ous, we believe our system can help people to understand trends in the informational development of a subject on the web. Such understanding increases the overall strength of their grasp of the subject, and utilizes the massive repository of human knowledge that the Internet represents.

References

[1] A. Ahmed, T. Dywer, S.-H. Hong, C. Murray, L. Song, and Y. X. Wu. Visualisation and analysis of large and complex scale-free networks. In *Proceedings of the 2005 Eurographics / IEEE VGTC Symposium on Visualization*, 2005.

[2] W. Aigner, S. Miksch, W. Müller, H. Schumann, and C. Tominski. Visual methods for analyzing time-oriented data. *IEEE Transactions on Visualization and Computer Graphics*, 14(1):47–60, 2008.

[3] A. L. Barabai, H. Jeong, Z. Nea, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3-4):590–614, 2002.

[4] T. Y. Berger-Wolf and J. Saia. A framework for analysis of dynamic social networks. In *ACM KDD 2006 Conference Proceedings*, pages 523–528, 2006.

[5] R. L. Cilibrasi and P. M. Vitanyi. The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383, 2007.

[6] S. Havre, B. Hetzler, and L. Nowell. Themeriver: Visualizing theme changes over time. In *IEEE Information Visualization 2000 Conference Proceedings*, pages 115–123, 2000.

[7] S. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, September 1967.

[8] G. Kumar and M. Garland. Visual exploration of complex time-varying graphs. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):805–812, 2006.

[9] P. Mika. Ontologies are us: A unified model of social networks and semantics. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(1):5 – 15, 2007.

[10] S. Morinaga and K. Yamanishi. Tracking dynamics of topic trends using a finite mixture model. In *ACM KDD 2004 Conference Proceedings*, pages 811–816, 2004.

[11] L. T. Nowell, R. K. France, D. Hix, L. S. Heath, and E. A. Fox. Visualizing search results: some alternatives to query-document similarity. In *ACM SIGIR 1996 Conference Proceedings*, pages 67–75, 1996.

[12] I. Vivísimo. Clusty, 2008.