Analysis on Multifaceted Information Dynamics of Rich Context Social Media

Conglei shi
shiconglei@gmail.com

Lisha Ye
yelisha5@gmail.com

Nan Cao
nancao@cse.ust.hk

ABSTRACT

Rich context social medias such as online news, documental database, encyclopedias as well as micro-blogs are often huge in amount, dynamic in nature and contain information with various aspects. For example, in the DBLP dataset, three information facets, conferences, authors and topics are aligned together within each paper. The information is dynamically changed year by year. Although several data mining approaches exist, the analysis results are usually obscure and difficult for common users to understand. There is a lack of the efficient analysis tool that helps to analyze and reveal the complicated information dynamics in both contents and relations from different perspectives. In this proposal, we present a visual analysis system. It targets on detecting the multifaceted dynamic information patterns for rich context social medias. The system combines several key analysis techniques including: (1) a dynamic multifaceted entity-relational data model which transfers documental data into the multifaceted dynamic graphs, (2) the optimal graph series segmentation which analyzes the significant of structure change within time-series graphs and (3) a refined dynamic spectrum clustering algorithm which represents the community dynamic over time. In addition, we will demonstrate the power of research through a case study based on real social media dataset.

Keywords

Social Media, Rich Context Analysis, Dynamic Clustering

1. MOTIVATIONS

The increasing amount of information is becoming available through collections of rich context media over the Internet. These collections contain a wealth of multifaceted interconnected resources, ranging from digital libraries, scholarly repositories, online news articles and blog postings to community generated content on social media platforms such as Facebook and Twitter. For example, a set of DBLP articles under a center research area often consists of various topics and relations over different facets such as time, research topics, authors and conferences, which inherently constitutes a multitude of complex, multifaceted networks. How to reveal the dynamics of multifaceted connectivity in such rich-context data is a challenging quest due to the following data properties.

First, data sizes are enormous – recent technological advances allow hundreds of millions of user generated content being aggregated and archived within online social networks or data repositories. Second, there are many dimensions of potential interest, from the textual content to the metadata of users or documents. Finally, the data is dynamic – changes can occur at multiple time scales and be localized to a subset of region in the data. Consequently, we need a framework that facilitates finding useful information from such data collections as well as reducing the complexity and massiveness of information, probing relations across the number and diversity of the facets of the data, and drawing users' attention to some unusual and "interesting" phenomena latent in the information dynamics.

2. RELATED WORK

A formal survey on the analysis of information dynamics is made, however considering the page limitations, here we only list few key research that most related with our proposal.

2.1 Monitoring time-varying networks

Ferlež et al. [6] proposed an MDL based approach for monitoring network evolution. Their method finds cut-points in time by extracting network communities and identifying when the communities change abruptly. Chen [3] introduced the methods for interpret the evolution of research co-citation networks by merging a sequence of networks into a single time-encoded network and emphasizing the visually salient nodes in the merged network. These approaches, while can be useful in finding globally interesting patterns, do not support drilling down to a subset of data according to the users' needs and tasks. Kumar and Garland [7] proposed a hierarchical clustering algorithm for representing time-varying graphs based on the hierarchical structures of graphs. They also provided tools for filtering edges and nodes according to the graph hierarchy. Their method allows for rending time-varying graphs smoothly but does not focus on finding patterns that are potentially interesting in the temporal domain. Several dynamic clustering algorithm [4, 8, 13, 14] are also proposed to tracing the structure dynamics of the social networks. Moreover, most of prior

work focuses on single mode network and lacks of ability for exploring multi-relational network data.

2.2 Novelty detection and event tracking

Novelty detection, or anomaly detection, on temporal data has been an important research topic in different areas ranging from fraud detection [5] to medical informatics [11, 12]. Many data analysis techniques have been employed in order to identify unusual and "interesting" phenomena. An extensive review of novelty detection techniques using statistical approaches and neural networks has been presented by Markou and Singh [9, 10]. Cox et al. [5] used the statistical profiling based visualization for fraud detection in mobile phone networks. Pizzi et al. [11] proposed a method for detecting and visualizing novel events in a set of time-varying fMRI images based on fuzzy cluster analysis. Suzuki et al. [12] proposed a color coding scheme that visually summarizes multi-dimensional time-series medical data as a sequence of probabilistic prototypes for detecting exceptional patterns, based on a probabilistic mixture model. Most of these methods deal with numerical or homogeneous network data.

Tracking events in online social media like blogs and Flickr has drawn considerable research interests [1, 2, 15]. With rich contents available in social media, there has been some works utilizing a variety of context features, such as authors, tags, hyperlinks, times and locations, to improve event mining from social media sources [2]. Zhao and Mitra [15] combined text-based clustering, temporal segmentation, and graph cuts to detect events in social networks where the event is defined as a group of actors communicating with each other on a specific topic over a certain time period. Many of the proposed methods focus on global patterns and lack of ability of searching for local patterns with respect to users' interests (e.g. on a particular facet or relation).

3. METHODOLOGY

In this proposal, we propose an new analysis system, which aims to analysis information dynamics of rich context social medias from multiple perspectives. In particular, three key information patterns - temporal patterns, multifaceted content patterns and relational patterns - are analyzed in an interactive way driven by users' interests. Generally speaking, our system will provide following three key features:

Visual analysis on information dynamics.. We analyze the information dynamics based on a time-varying multifaceted entity-relational data model. We reduce the complexity and massiveness of information by displaying the salient data entities on a contour map. It generates statistical estimation of the huge among of underlying data and provides a intuitive view based on few data samples. It also seamlessly integrates a smooth dynamic clustering as well as a temporal structure segmentation. Our algorithm target on generating stable clustering over time that help keep the users' mental map and reveal the salient dynamics of the data.

Multifaceted content exploration.. We extract facet information from rich text social medias based on our multifaceted entity-relational data model. It helps to separate information details into multiple layers that facilities the understanding of the complex information content from different perspective. A multifaceted Tag Cloud that combines with the a dynamic contour is designed in our system to reveal the topic patterns across different information facets and time slots.

Multi-Relational analysis.. Analysis and revealing relational patterns is another key feature of our system. We focus on two kinds of relations: the internal relations which connect the entities within the same facet and the external relations which are across different facets. They are used to represent rich contexts of the underlying data in an efficient and intuitive way.

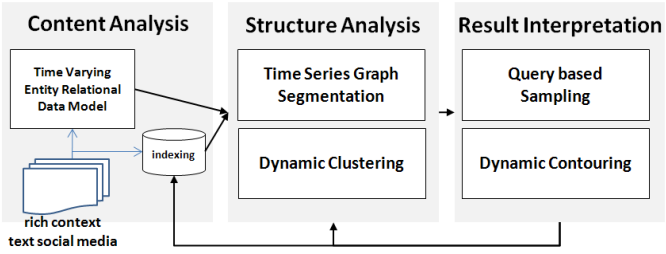

Figure 1: The overview of the proposed analysis system.

Generally speaking, the initial design of our system contains three major parts as illustrated in Fig. 1. With a large set of raw social media dataset, we first transform it into a dynamic multifaceted entity-relational data model. In this model, text documents are first segmented by their timestamps. After that, the documents within each time slot are segmented into different information facets using topic modeling techniques such as LDA. Entities will be further extracted from each information facets. Two kinds of relations will be build to connect the entities together: internal relation that connects entities within the same information facet to represent their co-occurrence within information facets and the external relations that connects entities over different facets to represent a sematic relationship between entities. All the information contents will be indexed facet by facet that facilities information query and data sampling.

The second part of our system is the dynamic analysis of the time series graph based on the above data model. In this part we are going to analysis the dynamic features of the rich context media from two aspects: the significant structure changes as well as the smooth transitions of social communities. The first problem will be achieved by the optimal time series graph segmentation and the second part will be based on a refined dynamic spectrum graph clustering technique.

The final part of our system is the dynamic contouring which will summarize and represent the analysis results in an intuitively manner. It converts the underlying raw data and its analysis results into knowledge which is acceptable by

common users.

4. PROJECT PLANS AND GOALS

This project is conduced by a group of three students (two first year Ph.D students and one second year Ph. D student). We will generally separate the tasks based on the different modules of our system. Some modules such as the dynamic analysis model are more complex than others. We are going to investigate this part all together.

The project starts at a survey as described in section 2. In the next step we will start to design the analysis algorithms and finish the whole system. Finally we will deploy our system onto a real social dataset and do some experiments to verify our designs and the correctness of our algorithms.

The final goal of this project is to submit a paper to a data mining conference by extending this course project.

5. CONCLUSIONS

In this proposal we proposed an analysis system for rich context social medias. The system focuses on analysis the multifaceted information dynamics of both information contents and structures. Three unique components are proposed: the content analysis, structure analysis and result interpretation. All these components will work together seamlessly to reveal underlying data patterns of information dynamics.

6. REFERENCES

[1] R. Balasubramanyan, F. Lin, W. Cohen, M. Hurst, and N. Smith. From episodes to sagas: Understanding the news by identifying temporally related story sequences. 2009.

[2] H. Becker, M. Naaman, and L. Gravano. Event identification in social media. In *Proc. of the ACM SIGMOD Workshop on the Web and Databases (WebDB?9)*. Citeseer, 2009.

[3] C. Chen. Searching for intellectual turning points: Progressive knowledge domain visualization. *Proceedings of the National Academy of Sciences*, 101(Suppl 1):5303, 2004.

[4] Y. Chi, X. Song, D. Zhou, K. Hino, and B. Tseng. Evolutionary spectral clustering by incorporating temporal smoothness. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 153–162. ACM, 2007.

[5] K. Cox, S. Eick, G. Wills, and R. Brachman. Visual data mining: Recognizing telephone calling fraud. 1997.

[6] J. Ferlež, C. Faloutsos, J. Leskovec, D. Mladenić, and M. Grobelnik. Monitoring network evolution using mdl. In *Proceedings of the IEEE 24th International Conference on Data Engineering ICDE*, pages 1328–1330. Citeseer, 2008.

[7] G. Kumar and M. Garland. Visual exploration of complex time-varying graphs. *IEEE Transactions on Visualization and Computer Graphics*, pages 805–812, 2006.

[8] Y. Lin, H. Sundaram, M. De Choudhury, and A. Kelliher. Temporal patterns in social media streams: theme discovery and evolution using joint analysis of content and context. In *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, pages 1456–1459. IEEE, 2009.

[9] M. Markou and S. Singh. Novelty detection: a review–part 1: statistical approaches. *Signal Processing*, 83(12):2481–2497, 2003.

[10] M. Markou and S. Singh. Novelty detection: a review–part 2::: neural network based approaches. *Signal Processing*, 83(12):2499–2521, 2003.

[11] N. Pizzi, R. Vivanco, and R. Somorjai. Evident: a functional magnetic resonance image analysis system. *Artificial Intelligence in Medicine*, 21(1-3):263–269, 2001.

[12] E. Suzuki, T. Watanabe, H. Yokoi, and K. Takabayashi. Detecting interesting exceptions from medical test data with visual summarization. In *Third IEEE International Conference on Data Mining, 2003. ICDM 2003*, pages 315–322, 2003.

[13] C. Tantipathananandh and T. Berger-Wolf. Constant-factor approximation algorithms for identifying dynamic communities. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 827–836. ACM, 2009.

[14] C. Tantipathananandh, T. Berger-Wolf, and D. Kempe. A framework for community identification in dynamic social networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 717–726. ACM, 2007.

[15] Q. Zhao and P. Mitra. Event detection and visualization for social text streams. *proceedings of ICWSM?007*, pages 26–28, 2007.