

2007 Winter Conference on Business Intelligence

Paper Summary for

C-TREND: A New Technique for Identifying Trends in Transactional Data

Gediminas Adomavicius and Jesse Bockstedt
{gadamavicius, jbockstedt}@csom.umn.edu

Department of Information and Decision Sciences
Carlson School of Management, University of Minnesota

1. Research Question and Motivation

The research field of data mining has developed sophisticated methods for identifying patterns in data in order to provide insights to users. Identifying *temporal* relationships (e.g., trends) in data constitutes an important problem that is relevant in many business and academic settings, and the data mining literature has provided analytical techniques for some specialized types of temporal data, e.g., time series analysis (Brockwell and Davis 2001, Keogh and Kasetty 2003, Roddick and Spiliopoulou 2002) and sequence analysis (Pei et al. 2004, Zaki 2001) techniques. Temporal data can take many forms, most commonly being *general* transactional (multi)attribute-value data, for which time series or sequence analysis methods are not particularly well suited.

The ability to identify trends in general temporal data can provide significant benefits, such as competitive advantages to a firm performing forecasts or making decisions on future investments and strategies. In particular, the research presented in this paper is focused on answering the question: *How should transactional attribute-value data be presented so that trends can be clearly identified and analyzed?* Consider the problem of technology forecasting for a firm. Technologies possess many features that change over time and understanding how a technology evolves requires trend analysis of multiple attributes at once. Similar issues arise in the trend analysis of consumer purchasing behavior and many other business intelligence applications; however, current temporal analytical techniques do not provide rich visualization of such trends. In this paper, we present a new approach that uses data mining techniques to provide visualization capabilities for trend analysis in multi-attribute transactional data.

2. Approach

In this paper, we present C-TREND, *Cluster-based Temporal Representation of EveNt Data*, a new method for discovering and visualizing trends and temporal patterns in transactional attribute-value data that builds upon standard data mining clustering techniques. In particular, C-TREND separates data into user-defined partitions based on time periods and then identifies clusters of the dominant transaction types occurring within each partition. Clusters are then compared to the clusters in adjacent time periods to identify cross-period similarities and, over many time periods, trends are identified. Trends are presented in an output graph that uses nodes to represent dominant transaction types and edges to represent cross-time relationships.

The proposed C-TREND technique consists of two major processes: offline preprocessing of the data and online interactive analysis and visualization of the trends. Offline preprocessing includes the calculation of a dendrogram solution using agglomerative hierarchical clustering for each data partition and the calculation of the potential graph constructs (Duda et al. 2000). Graph nodes are determined for each partition by first extracting a *k*-sized clustering solution

from the dendrogram and then filtering the identified clusters using the *within-period trend strength* parameter α . Edges are determined by calculating pair-wise similarity metrics between clusters in adjacent data partitions and filtering them using the *cross-period trend strength* parameter β . Interactive analysis includes the presentation of output graphs in a graphical user interface (GUI) that allows the user to adjust k for each partition and the α and β parameters, which prompts C-TREND to redraw the output graph based on these new values in real-time.

3. Findings and Expected Contributions

To demonstrate the use of C-TREND for identifying trends in transactional attribute-value data and show how modifying input parameters affects the trend graph output, we analyzed the data on over 2,400 certifications for wireless networking (802.11) technologies awarded by the Wi-Fi Alliance (wi-fi.org). By representing this data using C-TREND, we could clearly identify evolutionary paths that Wi-Fi technologies have followed over the past six years. More specifically, the C-TREND output graphs provided the means for identifying the emergence of new products, the death of old products, and both the splitting and integration of similar products over time.

C-TREND provides three advantages over existing techniques. First, C-TREND presents temporal data in a unique and intuitive manner that emphasizes trends between dominant transaction types over time, and its output graphs resemble evolutionary diagrams and naturally portray the changes in data characteristics over time. Second, C-TREND is a meta-analysis tool for data mining results (specifically, hierarchical clustering) and, therefore, is designed to provide the domain expert with substantial control over the data presentation. In particular, C-TREND provides the user with the ability to adjust all key parameters for creating output trend graphs, which allows a domain expert to visualize the data in a manner that provides the most value. Third, C-TREND presents a set of graph statistics (e.g., cluster center, size, radius, and cluster similarity metrics), which, in our future work, will provide a means for developing new trend metrics and a framework for performing hypothesis testing on the existence and characteristics of trends.

4. Current Status of the Manuscript

A conference-style draft of this paper and its initial findings were presented at our department workshop in October, 2006. A journal version of this paper is currently under development and should be ready for review by February 2007.

References

- Brockwell, P., R. Davis, *Time Series: Theory and Methods*, Springer-Verlag, New York, 2001.
- Duda, R., P. Hart, D. Stork. *Pattern Classification*, 2nd ed., Wiley-Interscience, 2000.
- Keogh, E., S. Kasetty, "On the need for time series data mining benchmarks: A Survey and Empirical Demonstration," *Data Mining and Knowledge Discovery*, 7(4):349-371, 2003.
- Pei, J., J. Han, B. Mortazavi-Asl, J. Wang, Helen Pinto, Qiming Chen, Umeshwar Dayal, Mei-Chun Hsu, "Mining sequential patterns by pattern-growth: The PrefixSpan approach," *IEEE Transactions on Knowledge and Data Engineering*, 16(10):1-17, 2004.
- Roddick, J., M. Spiliopoulou, "A survey of temporal knowledge discovery paradigms and methods," *IEEE Transactions on Knowledge and Data Engineering*, 14(4):750-767, 2002.
- Zaki, M., "SPADE: An efficient algorithm for mining frequent sequences," *Machine Learning*, 42(1/2):31-60, 2001.