Detecting Temporal Pattern and Cluster Changes in Social Networks: A Study Focusing UK Cattle Movement Database

Puteri N.E. Nohuddin[1], Frans Coenen[1], Rob Christley[2], and Christian Setzkorn[2]

[1] Department of Computer Science,
University of Liverpool,
L69 3BX Liverpool
+44 (0)151 795 4275
puteri@liverpool.ac.uk,
frans@liverpool.ac.uk
[2] School of Veterinary Science,
University of Liverpool and National Center for Zoonosis Research,
Leahurst, Neston
+44 (0)151 794 6003
robc@liverpool.ac.uk,
christian@setzkorn.eu

Abstract. Temporal Data Mining is directed at the identification of knowledge that has some temporal dimension. This paper reports on work conducted to identify temporal frequent patterns in social network data. The focus for the work is the cattle movement database in operation in Great Britain, which can be interpreted as a social network with additional spatial and temporal information. The paper firstly proposes a trend mining framework for identifying frequent pattern trends. Experiments using this framework demonstrate that in many cases a large number of patterns may be produced, and consequently the analysis of the end result is inhibited. To assist in the analysis of the identified trends this paper secondly proposes a trend clustering approach, founded on the concept of Self Organizing Maps (SOMs), to group similar trends and to compare such groups. A distance function is used to compare and analyze the changes in clusters with respect to time.

Keywords: Temporal Data Mining, Social Networks, Trends, Temporal Patterns and Clusters.

1 Introduction

Many data mining techniques have been introduced to identify frequent patterns in large databases. More recently, the prevalence of large time stamped databases, facilitated by advances in technology, has increased. As such, time series analysis techniques are of increasing significance. A time series, at its simplest, consists of a sequence of values associated with an attribute. The work described in this paper

Z. Shi et al. (Eds.): IIP 2010, IFIP AICT 340, pp. 163–172, 2010.

considers time series to comprise several sub-series. As such, the sub-series may be compared to identify changes. For example, we can imagine a time series covering N years, where each twelve month period represents a sub-series; as such the sub-series can be compared to identify (say) seasonal changes or anomalies.

The work described in this paper is specifically directed at the comparison of sequences of time series that exist in social network data. In this respect, the time series are defined in terms of the changing frequency of occurrence of combinations of attributes that exist across particular social networks. We refer to such time series as *trends*.

Social networks are collections of interacting entities typically operating in some social setting (but not necessarily so). The nodes in a social network represent the entities and the arcs the interactions. The focus for the work described in this paper is the cattle movement database in operation within Great Britain (GB). The identification of trends in cattle movements, and changes in trends, is of particular interest to decision makers who are concerned with the potential spread of cattle disease, however the trend analysis techniques described have more general applicability.

A particular issue in trend analysis in social networks (and more generally) is the large number of trends that may be discovered. One solution, and that propose in this paper, is to cluster similar trends using Self Organizing Map (SOM) technology. The use of SOMs provides a visualization technique. Moreover, since we are interested in identifying anomalies in trends, we wish to compare individual SOMs (describing sub-series) so as to be able to observe the "dynamics" of the clustered trends.

2 Background

This section provides some brief background regarding the work described. The section is divided into three sub-sections: temporal frequent pattern mining, social networks and trend clustering and comparison.

2.1 Temporal Frequent Pattern Mining

Temporal data mining is directed at data that comprises sequence of events (Antunes *et al.* 2001). The introduction of advanced computer technologies and data storage mechanisms has afforded many organizations the opportunity to store significant amounts of temporal (and spatio-temporal) data. Consequently, there is a corresponding requirement for the application of temporal data mining techniques. The main aim of temporal data mining is to discover the relationship between non-trivial patterns or events in the temporal database (Roddick *et al.* 2002). This then allows the identification of trends or change points within the data. Many approaches have been explored in the context of temporal data mining. Two common methods are time series analysis (Brockwell *et al.* 2001) (Keogh *et al.* 2003) and sequence analysis (Zaki 2001).

In this work, trends are defined in terms of the changing frequency of frequent patterns with time. A frequent pattern, as first defined by Agrawal *et al.* (1993), is a subset of attributes that frequently co-occur in the input data according to some user specified support threshold. Since then, the frequent pattern idea has been extended in many directions. A number of authors have considered the nature of frequent patterns with respect to the temporal dimension, for example sequential patterns (Agrawal *et al.* 1995), frequent episodes (Mannila *et al.* 1997) and emerging patterns (Dong *et al.* 1999). Many alternative frequent pattern mining algorithms, that seek to improve on Agrawal's original Apriori algorithm, have also been proposed. One example is the TFP (Total From Partial) algorithm (Coenen *et. al.* 2001). The authors have adapted TFP to identify trends as defined above.

2.2 Social Network

A Social Network (SN) describes a social structure of individuals, who are connected directly or indirectly based on a common subject of interest, conflict, financial exchange or activities. A SN depicts the structure of social entities, *actors,* who are connected through ties, links or pairs (Wasserman 2006). Social Network Mining (SNM) has become a significant research area within the domain of data mining. Many SN analysis techniques have been proposed which map and measure the relationships and flows between people, organizations, groups, computers and web sites. SNM can be applied in a static context, which ignores the temporal aspects of the network; or in a dynamic context, which takes temporal aspects into consideration. In the static context, we typically wish to find patterns that exist across the network, or cluster sub-sets of the networks, or build classifiers to categorize nodes and links. In the dynamic context, we wish to identify trends or change points within networks.

2.3 Trend Clustering and Comparison

Trend mining techniques typically identify large numbers of trends. To aide the analysis of the identified trends, the technique proposed in this paper suggests the clustering of trends so that similar trends may be grouped. When this process is repeated for a sequence of sub-trends, the clusters can be compared so as to identify changes or anomalies. Other than simply comparing pairs of data sets (Denny *et al.* 2008), there are several methods that have been proposed to detect cluster changes and cluster membership migration. For example, Lingras *et al.* (2004) proposed the use of Temporal Cluster Migration Matrices (TCMM) for visualizing cluster changes in e-commerce site usage. Hido *et al.* (2008) suggested a technique to identify changes in clusters using a decision tree method. This paper proposes the use of Self Organizing Maps (SOMs).

3 Problem Definition

The work described in this paper assumes a time stamped data set D such that $D=\{D_1, D_2, \ldots D_n\}$, where n is the number of time stamps. Each data set in D comprises a set of records such that each record is a subset of some global attribute set I. A frequent pattern occurring in data set D_k is then some subset of I that occurs in given percentage of the records in D_k, this percentage is termed the support threshold (α). The number of occurrences of a given frequent pattern in D_k is termed its support (s). The sequence of support values for a given pattern can be conceptualized as time series T comprising a number of *episodes* such that $T = \{E_1, E_2 ,\ldots, E_n\}$.The sequence of support values represented in a particular episode the describes a trend comprising m time stamps, thus $E_i = \{t_1, t_2, \ldots, t_m\}$ $(0<i\leq m)$.The problem addressed in this paper is firstly the effective identification of these trends, and secondly the comparison of these trends so as to identify interesting information.

4 Frequent Trend Mining and Analysis

An overview of the proposed trend mining process is given in Figure 1. The process comprises two stages: (i) trend mining and (ii) visualization. Separate software units have been developed for each stage, in the figure; these are identified as the *trend mining unit* and the *visualization unit*. The process commences (top left) with a time stamped data set covering a sequence of N episodes. Sequences of N trends can be then identified from within the data using appropriate trend mining software (see below). The trends are stored in a compressed from in a "reverse" set enumeration tree structures (top right of Figure 1). The tree structures allow fast "look up" to extract the actual trends. Some examples are given in the figure. Thus (from the figure), the pattern {a,b,c,d} has a sequence of support values of {0,0,2500,3311,2718,0,0, 0,2779} describing a nine time-stamp trend associated with a single episode, similar sequences may be extracted for all N episodes associated with the pattern {a,b,c,d}. Note that a 0 support value indicates a support value below the support threshold. The trend values are the input for visualization unit which produces several maps that cluster yearly trends which are orderly processed by the unit (bottom left of Figure 1).

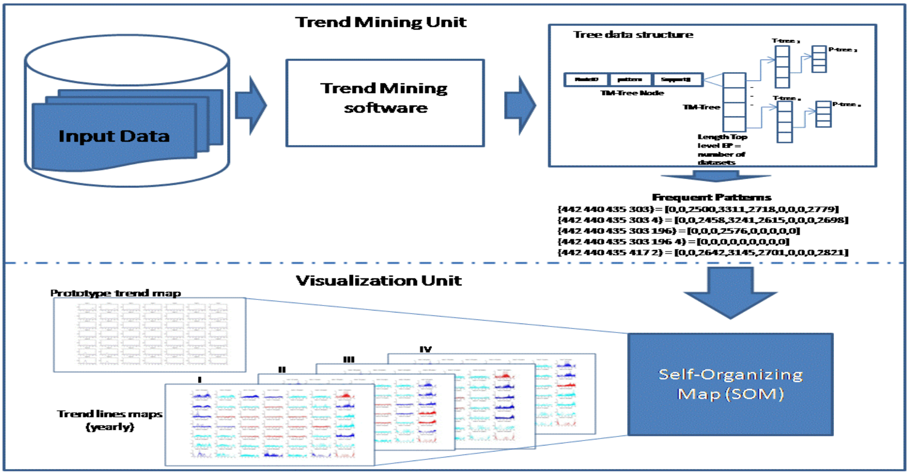

Fig. 1. Trend Mining Analysis

The trend mining unit software identifies and extracts the desired trends. The software was founded on the Total From Partial (TFP) association rule mining algorithm extended to give Trend Mining TFP (TM-TFP) so that sequences of support values could be identified.

The visualization unit is responsible for analyzing the output from TM-TFP and presenting the results. The objective is, other than clustering of the trends, to identify cluster changes. The process commenced with the clustering of the trends in each episode (so that N sets of clusters are produced). The clustering was undertaken using a Self Organizing Map (SOM) (Kohonen 1998).The process commences with the generation of a prototype map using some proportion (or all) of the trends associated with one of the episodes. The SOM map was initialized with p x q = N nodes such that each node represented a “type” (category) of trend line. Once the prototype map had been generated, the trends associated with each of the episodes were fitted to this map so as to give a sequence of “trend line” maps. In the figure, four episodes are assumed (labeled I, II, III and IV). Once the trend line maps have been derived the change in trends was determined by considering how individual trends, associated with particular frequent patterns, moved (or did not move) across the sequence of maps. A simple Euclidean distance measure was used for this purpose. The maximum change is the diagonal distance across the map.

5 Experimental Evaluation

This section presents and discusses sample results obtained using the proposed trend mining and analysis process. For the experiments, the Cattle Tracing System (CTS) database in operation in Great Britain (GB) was used. The CTS database records, for monitoring purposes, all the movements of cattle registered within or imported into GB. The database is maintained by the Department for Environment, Food and Rural Affairs (DEFRA). Cattle movements can be one off movements to final destinations, or movements between intermediate locations. Movement types include: (i) cattle imports, (ii) movements between locations, (iii) movements in terms of births and (iv) movements in terms of deaths. CTS was introduced in September 1998, and updated in 2001 to support disease control activities. Currently the CTS database holds some 155 Gb of data. The CTS database can be interpreted as a social network where the nodes represent cattle holding areas and the arcs between nodes cattle movements.

The CTS database comprises a number of tables, the most significant of which are the animal, location and movement tables. For the experiments reported here the data from 2003 to 2006 was extracted to form 4 episodes each comprising 12 (one month time stamps). The data was stored in a single data warehouse such that each record represented a single cattle movement instance associated with a particular year (episode) and month (time stamp). The number of CTS records represented in each data episode was about 400,000. Each record in the warehouse comprised: (i) a time stamp (month and year), (ii) the number of cattle moved, (iii) the bread, (iv) the senders location in terms of easting and northing grid values, (v) the "type" of the sender's location, (vi) the receivers location in terms of easting and northing grid values, and (vii) the "type" of the receiver's location. If two different breeds of cattle were moved at the same time from the same sender location to the same receiver location this would generate two records in the warehouse. The maximum number of cattle moved between any pair of locations for a single time stamp was approximately 40 animals.

5.1 Frequent Patterns and Trends

Table 1 presents some statistics indicating the number of trends discovered in the CTS data warehouse uisng TM-TFP. Recall that TM-TFP was used to identify frequent patterns and their associated support values over a sequences of time stamps. The results presented in Table 1 were generated using support thresholds of 0.5%, 0.8% and 1% respectively. Each row in Table 1 represents the number of trends identified for each of the 4 episodes (12 time stamps per episode). The lower the support threshold the greater the number of discovered frequent patterns and hence the greater the number of trends. However, use of a low support threshold ensures that no potentially interesting trends are omitted. The results presented in Table 1 indicate that a large number of trends can be identified, in a realistically sized dataset, when low support thesholds are used. It is also interesting to note that the variation between years is relatively small.

Table 1. Number of trend lines identified using TM-TFP

Year	Support Threshold		
	0.5%	0.8%	1%
2003	63,117	34,858	25,738
2004	66,870	36,489	27,055
2005	65,154	35,626	25,954
2006	62,713	33,795	24,740

5.2 Temporal Clustering

Figure 2 depicts prototype trend map trained using the 2003 data. With reference to the figure, node 1 (top-left) represents trend lines that have for patterns with high support in spring (March to May) and autumn (September to November) while node 43 (bottom-left) indicates trend lines with high support in spring only (March to April). Note that the distance between nodes indicates the dissimilarity between nodes; the greatest dissimilarity is thus between nodes at opposite ends of the diagonals. Once the initial proto-type map has been generated a sequence of trend line maps can be produced, one for each episode. An example is given in Figure 3 for the 2003 cattle movement data of the trend line maps which have been produced. These trend line maps are referring to 2003 prototype map for the same cluster structure. Each node has been annotated with the number of trends in the “clusters”. Thus, from Figure 3, there are 1970 trend lines in node 1. The shading used in Figure 3 indicates the number of trend lines in each node, the darker the shading, the greater the number of trends.

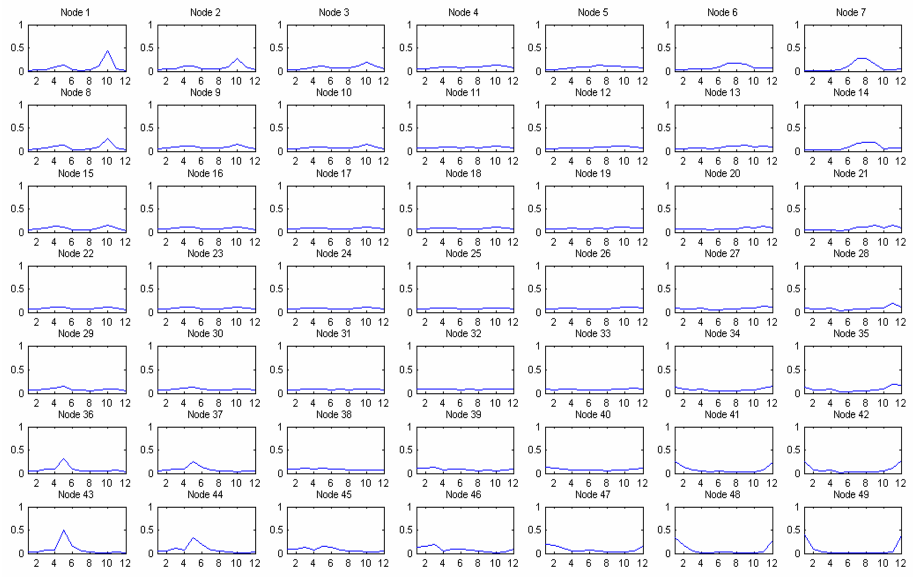

Fig. 2. Prototype map

5.3 Temporal Cluster Changes

Figure 4 indicates the number of trends in each node (cluster) for the 4 years (episodes) included in the described study. From Figure 4, the greatest differences are observed for nodes 23 and 31. Whatever the case, from the figure, it is clear that the number of trends per node is not static. Given a sequence of trend-line maps comparisons can be made to see how trends associated with individual frequent patterns change by analyzing the nodes in which they appear. Some trends may remain within the same node for the entire sequence of episodes. Some other trends may oscillate between nodes, while some further trends may slowly migrate across the map. By translating the trend line maps into a rectangular (D-plane) set of coordinates a Euclidean distance function was applied to observe the similarities and differences of trends within each node across the episodes. By comparing the values produced by the distance function, the degree of movement could be determined. This could be interpreted in a number of ways, for example the greater the distance moved the more interesting the change may be deemed to be.

Table 2 shows examples of how some trends (representing frequent patterns) migrate from one cluster to another. For example, the trend line representing the pattern {441 436 329 301 213 4 3} which translates as:{numberAnimalsMoved <=5, Receiver PTI = NULL, Receiver Location Type = Calf Collection Centre, Sender Location Type = Agricultural Holding, Sender area = 14, Animal age <= 1 year old, Gender = female} was in node 49 (bottom right in Figure 2) in 2003 and 2004, but then migrated to node 48 in 2005 and disappeared in 2006. Table 3 gives some further

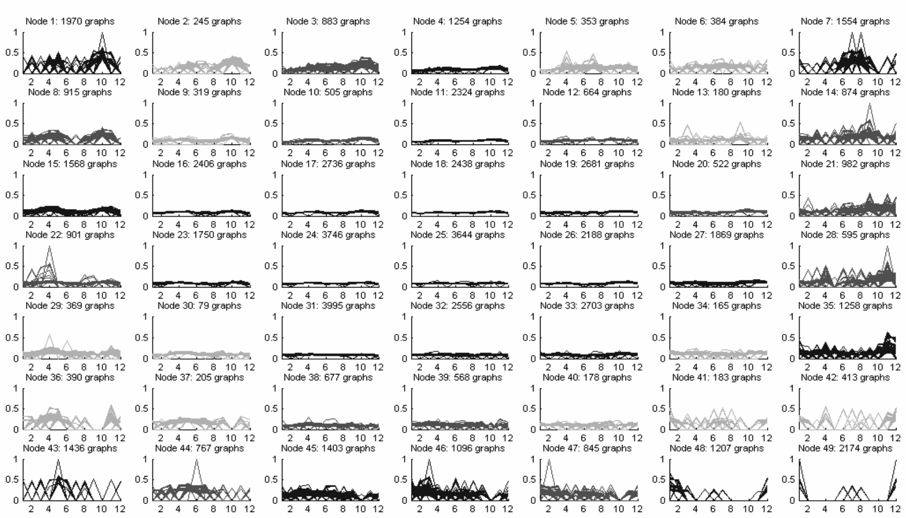

Fig. 3. Trend line map for 2003

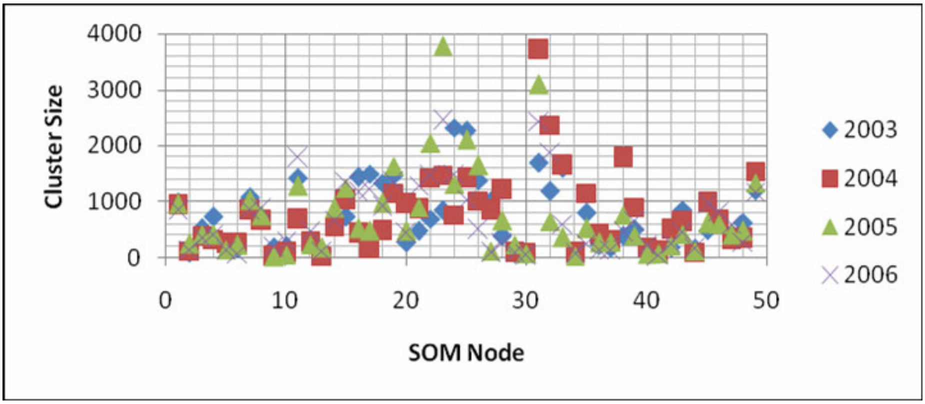

Fig. 4. Comparison cluster size between 2003 and 2006

statistics regarding the movement of trends in the context of the CTS database. There are 79894 distinct frequent patterns generated between 2003 and 2006 data episodes. But only 4193 patterns that remain in the same nodes across the years whereas the rest of the patterns moved to different cluster nodes. Table 3 also shows statistics of spatio-temporal patterns between 2003 and 2006.

Table 2. Example of frequent patterns that migrated to other clusters

Frequent Patterns	Node 2003	Dist	Node 2004	Dist	Node 2005	Dist	Node 2006
{441 436 329 301 213 4 3}	49	0	49	1	48	0	0
{441 436 329 301 213 196}	48	1	49	4.1	38	3.2	48
{378 301 263}	39	0	39	3.2	49	3.2	39
{378 301 263 4}	46	0	46	0	0	0	46
{378 301 263 196}	47	1	46	0	0	0	49
{441 318 301 212 4}	14	0	14	5	16	5.4	7
{441 329 214}	47	2	49	0	0	0	49

Table 3. Clusters memberships

Number of Patterns	Quantity
Distinct frequent patterns described by trend lines between 2003 and 2006	79894
Trends stayed in the same node (unchanged) between 2003 and 2006	4193
Trends migrated to other clusters between 2003 and 2006	75701
Spatio-temporal trends stayed in the same cluster between 2003 and 2006	637
Spatio-temporal trends migrated to other clusters with greater distance values (distance>4) between 2003 and 2006	2061

6 Conclusions

This paper has described a trend mining framework, TM-TFP that successfully identifies trends in large social networks. The framework is supported by a SOM technique that provides a powerful mechanism for grouping similar trends, and a trend migration identification mechanism to show changes in the nature of individual trends associated with frequent patterns. The mechanism has been tested and evaluated using data from GB's cattle tracking database. The research team is currently looking at other ways in which change detection can be made more effective in the context of decision makers and stakeholders.

References

Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rules between Sets of Items in Large Databases. In: Proceedings of ACM SIGMOD Conference (1993)

Agrawal, R., Srikant, R.: Mining sequential patterns. In: 11th International Conference on Data Engineering (1995)

Antunes, C.M., Oliveira, A.L.: Temporal Data Mining: An Overview. In: Proc. ACM SIGKDD Workshop Data Mining, August 2001, pp. 1–13 (August 2001)

Brockwell, P., Davis, R.: Time Series:Theory and Methods. Springer, Heidelberg (2001)

Coenen, F.P., Goulbourne, G., Leng, P.: ComputingAssociation Rules Using Partial Totals. In: Siebes, A., De Raedt, L. (eds.) PKDD 2001. LNCS (LNAI), vol. 2168, pp. 54–66. Springer, Heidelberg (2001)

Denny, Williams, G.J., Christen, P.: ReDSOM: relative density visualization of temporal changes in cluster structures using self-organizing maps. In: IEEE International Conference on Data Mining (ICDM), pp. 173–182. IEEE Computer Society, Los Alamitos (2008)

Dong, G., Li, J.: Efficient Mining of Emerging Patterns: Discovering Trends and Differences. In: Proceeding of Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (1999)

Hido, S., Idé, T., Kashima, H., Kubo, H., Matsuzawa, H.: Unsupervised changes analysis using supervised learning. In: Washio, T., Suzuki, E., Ting, K.M., Inokuchi, A. (eds.) PAKDD 2008. LNCS (LNAI), vol. 5012, pp. 148–159. Springer, Heidelberg (2008)

Keogh, E., Kasetty, S.: On the need for Time Series Data Mining Benchmarks: A Survey and Empirical Demostration. Data Mining and Knowledge Discovery 7(4), 349–371 (2003)

Kohonen, T.: The Self Organizing Maps. Neurocomputing 21, 1–6 (1998)

Lingras, P., Hogo, M., Snorek, M.: Temporal Cluster Migration Matrices for Web Usage Mining. In: Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence (2004)

Mannila, H., Toivonen, H., Verkamo, A.: Discovery of Frequent Episodes in Event Sequences. Data Mining and Knowledge Discovery 1, 259–289 (1997)

Roddick, J., Spiliopoulou, M.: A Survey of Temporal Knowledge Discovery Paradigms and Methods. IEEE Trans. Knowledge and Data Eng. 14(4), 750–767 (2002)

Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, New York (2006)

Zaki, M.: SPADE: An Efficient Algorithm for Mining Frequent Sequences. Machine Learning 42(1-2), 31–60 (2001)