Finding "Interesting" Trends in Social Networks Using Frequent Pattern Mining and Self Organizing Maps

Puteri N. E. Nohuddin, Frans Coenen[a,*], Rob Christley, Christian Setzkorn[b],
Yogesh Patel, Shane Williams[c]

[a]*Department of Computer Science, University of Liverpool, UK*
[b]*School of Veterinary Science, University of Liverpool and National Centre for Zoonosis Research, Leahurst, Neston, UK*
[c]*Deeside Insurance Ltd., Deeside, UK*

Abstract

This paper introduces a technique that uses Frequent Pattern Mining and Self Organising Maps (SOMs) to identify, group and analyse trends in sequences of time stamped social networks so as to identify "interesting" trends. In this study, trends are defined in terms of a series of occurrence counts associated with frequent patterns that may be identified within social networks. Typically a large number of frequent patterns, and by extension a large number of trends, are discovered. Thus, to assist with the analysis of the discovered trends, the use of SOM techniques is advocated so that similar trends can be grouped together. To identify "interesting" trends a sequences of SOMs are generated which can be interpreted by considering how trends move from one SOM to the next. The further a trend moves from one SOM to the next, the more "interesting" the trend is deemed to be. The study is focused on two types of network, *star* networks and *complex* star networks, exemplified by two real applications: the Cattle Tracing System in operation in Great Britain and a car insurance quotation application.

Keywords: Trends, Social Networks, Frequent Pattern Mining, Self Organizing Maps

[*] Corresponding author

Email addresses: `puteri, frans@liverpool.ac.uk` (Puteri N. E. Nohuddin, Frans Coenen), `robc, c.setzkorn@liverpool.ac.uk` (Rob Christley, Christian Setzkorn), `yogesh, shane@deesideinsurance.co.uk` (Yogesh Patel, Shane Williams)

1. Introduction

The quantity of social network traffic has increased significantly over the past few years, especially with respect to www applications such as Facebook, Bebo and Flicker. In the wider context, we can also identify other forms of social networks such as business communities, file sharing systems and co-authoring frameworks. Social networks are typically conceptualised as graphs comprising nodes and links, where the nodes represent individuals (network users) and the links communications (traffic). These communications often take the form of text (emails) but can be files (photographs, movies, etc.). As such, social network mining has become a popular area of study. The aim is to extract knowledge so as to understand the behaviour of the users of such networks with respect to some specific application context and/or improve the service provided.

With respect to the work described in this paper, the authors have widened the concept of social networks to include any form of identifiable community that interchanges information. More specifically the work is focused on two types of social networks: *Star* networks and *Complex Star* networks. The first is exemplified by a car insurance quote application, the Deeside Insurance Quote network. The second is exemplified by the Cattle Tracing System (CTS) in operation in Great Britain (GB). The Deeside Insurance Quote network was generated from a database, maintained by Deeside Insurance Ltd., that records requests for insurance quotations from customers at different geographical locations. Thus, the network nodes represent geographical locations which all communicate with a central insurance "broker" node (and the links represent the amount of traffic - requests for insurance quotes). The network is therefore described as a Star network. CTS incorporates a database that records cattle movements. A large scale social network can be generated from this database such that the nodes represent cattle holding areas (farms, markets, abattoirs, etc) and the links cattle movements between locations. The network in this case features small sub-communities of nodes the regularly communicate with one another, with little traffic crossing between sub-communities. The network is therefore described as a Complex Star network because, in effect, it comprises many small Star networks.

Social Network mining is usually conducted in a static manner by taking a "snap shot" of the network and then applying mining techniques to this snap shot. The objective is usually to identify communities within the network. In this paper, the authors are interested in the dynamic aspects of social networks. The authors are particularly interested in mechanisms for identifying and analysing trends in social networks. In the context of the Deeside Insurance Quote network, this analysis will highlight variations in insurance quotation trends which may be used for marketing purposes. In the context of the CTS network, the identification of trends, and variations in trends, will provide knowledge of (say) the effect of the introduction of new legislation, or indicate changes in working practices, it will also give an insight into the way that cattle infections may spread.

In the study described in this paper, trends are defined in terms of the changing frequencies with which common patterns occur across social network data. Trends are collected according to sequences or *epochs*, which can then be compared. The nature (duration) of an epoch is application dependent. However, for the cattle movement and Deeside social networks, it made sense to consider trends in terms of years (i.e. the duration of an epoch is 12 months) because this will serve to capture seasonal variations. Whatever the case, the epoch length is a user supplied variable that can be easily adjusted to fit alternative applications.

Using the proposed trend mining mechanism, a significant number of trends may be identified, too many to allow simple inspection by decision makers. Some mechanism was therefore required to allow the simple presentation of trend lines. The first technique advocated in this paper is to group (cluster) trends that display similar contours. For this process, Self Organising Map (SOM) technology has been adopted. Once the trends have been identified and grouped, we wish to determine how these trends change from epoch to epoch. The nature of the changes which we might be interested in will vary from application to application. For some applications, we may be interested in trends that remain constant, for others we may be interested in trends that change radically. To identify changes in trends, the advocated approach is to generate a sequence of SOM maps, one per epoch, and analyse how trends "move" (or do not move) from SOM to SOM (epoch to epoch).

The contribution of this paper may thus be summarised as: (i) an unusual application of social network mining with respect to the CTS and Deeside Insurance database, (ii) a mechanism for generating frequent pattern trends, (iii) a process for assisting the analysis of the identified trends using SOM technology and (iv) an approach to identify "interesting" changes in trends. The rest of this paper is organised as follows. Some previous work is described in Section 2. The proposed social network trend mining approach is described in Section 3. Section 4 presents an evaluation of the proposed technique, firstly using the cattle movement social network, and secondly the Deeside social network. Section 5 presents some conclusions concerning frequent pattern trend analysis.

2. Previous Work

High availability of advanced computer information systems have resulted in the rapid growth of temporal databases together with a corresponding desire to generate trends in these collections and analyse the trends. This comparative analysis is often used to identify patterns of movement and relationship in a certain domain such as financial investment, medical monitoring and commercial products. For example, Google Trends, a public web facility that supports the identification of trends associated with keyword search volume [1]. Trend recognition processes can be applied to both qualitative and quantitative data, such as the forecasting of financial market trends based on numeric financial data, and usage of text corpi in business news [2]. Raza and Liyanage [4] proposed a trend analysis approach to mine and monitor data for abnormalities and faults in industrial production.

A social network represents the structure of some social entity, and normally comprises actors who are connected through one of more links [18]. To analyze this structure, techniques have been proposed which map and measure the relationships and flows between nodes. Social network mining 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 a static context, we typically wish to: (i) find patterns that exist across the network, (ii) cluster (group) subsets of the networks, or (iii) build classifiers to categorize nodes and links. In the dynamic context, we wish to identify relationships between the nodes in the network by evaluating the spatio-temporal co-occurrences of events [6]. The latter is thus the focus of the work described in this paper.

As noted above, in this work, we define trends in terms of the changing frequency of patterns with time. A frequent pattern, as first defined by Agrawal et al. [7], is a subset of attributes that frequently co-occur according to some user specified support threshold. 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 [8], frequent episodes [9], emerging patterns [10] and jumping and emerging patterns [3]. Many alternative frequent pattern mining algorithms, that seek to improve on Agrawal's original Apriori algorithm, have also been proposed. TFP (Total from Partial) [11] is one established algorithm that extends Apriori. For the work described here, TFP has been adapted for the purpose of trend mining.

There has been some work on social networks trend analysis. Gloor *et.al*. introduced a novel trend analysis algorithm to generate trends from Web resources [24]. The algorithm calculates the values of temporal *betweeness* of online social network node and link structures to observe and predict trends concerning the popularity of concepts and topics such as brands, movies and politicians. Likewise, some research directed at recommender systems [29] [30] and online market research [28] focuses on trends describing online social interactions and trusts so as to improve online marketing and sales strategies. Research in social networks trend mining has also provided advantages for online viral marketing [22], stock market activities [23] and many more. There has been some work on the identification of trends in time stamped sequences of binary valued (Association Rule Mining (ARM)) data sets. Furthermore, there are several studies on Jumping and Emerging Patterns. Emerging Patterns describe patterns whose support count becomes significant between time stamps [25]. Jumping Patterns describe patterns whose support counts change or jump drastically from one time stamp to another. There are many more examples, however, in this paper, we are interested in mining trends which are defined in terms of the changing frequency of occurrence of individual patterns presented in the data.

Self Organising Maps (SOMs) were first introduced by Kohonen [13, 12]. Fundamentally, SOMs are a neural network based technique designed to reduce the number of data dimensions in some input space by projecting it onto a $n \times m$ "node map", which plots the similarities of the input data by grouping (clustering) similar data items together at nodes. SOMs have been utilized in many other research areas for data clustering. For example SOMs have been used for clustering gene expression data [14], glaucoma image clustering [15], image retrieval system [27] and stock price forecasting expert systems [26]. The SOM learning process is unsupervised, in other words no predefined number of clusters is specified. Currently, there is no scientific method for determining the best values for $n \times m$, i.e. to identify how many clusters should be represented by a SOM. However, the $n \times m$ value does define a maximum number of clusters; although on completion some nodes may be empty [16]. Since SOM are based on competitive learning, the output nodes on the map compete among each other to be stimulated to represent the input data. With respect to the work described in this paper, we have adopted a SOM approach to group similar trends, and thus provide a mechanism for analysing social network trend mining results.

When dealing with temporal trends, it is of interest to find and analyse changes that occur in these trends. From the sequence of SOM maps, a trend clusters analysis is conducted. Cluster analysis is concerned with discovery of information about the relationship and/or similarity between clusters. Cluster analysis allows practioners to investigate temporal cluster size enlargement and reduction. Cluster analysis may also be used to identify temporal cluster migrations. Several methods have been introduced to detect cluster changes and cluster membership migration. For example, Lingras et. al. [19] proposed the use of Temporal Cluster Migration Matrices (TCMMs), to visualize cluster changes in e-commerce sites usage, that reflected changes in user spending patterns. A straight forward technique for detecting temporal changes is simply to compare two datasets with the same data structure [20]. An example of a more advanced technique is that of Hido et. al. [21] who proposed the use of decision trees to model and identify changes in clusters. A simple Euclidean distance measure is adopted in this paper.

3. The Trend Mining Mechanism

As noted in Section 1, a trend is defined as a sequence of *support* values, associated with a specific pattern, over a sequence of time stamps. The support of a pattern is the number of occurences of that pattern in the data set for some time stamp. The sequence of time stamps is referred to as an *epoch*. Thus, a trend t comprises a set of values $\{v_1, v_2, \ldots, v_n\}$ where n is the number of time stamps in the epoch. A trend associated with a particular pattern i is indicated by t_i. The jth value in a trend t_i is indicated by t_{ij}. We wish to identify changes in the trends associated with individual patterns and thus we wish to compare trends over two or more epochs. A sequence of trends T comprises a set of trends $\{t_1, t_2, \ldots, t_e\}$, where e is the number of epochs described by the sequence. The proposed approach (Figure 1) comprises three stages: (i) frequent pattern trend mining, (ii) trend clustering, and (iii) trend clusters analysis.

3.1. Frequent Pattern Trend Mining

The input to the trend mining system comprises a binary valued, time stamped, data set $D = \{d_1, d_2, \ldots, d_{n\times e}\}$ (where n is the number of time stamps per epoch, and e is the number of epochs under consideration). The records in each dataset in D comprise some subsets of a global set of binary valued attributes $A = \{a_1, a_2, \ldots, a_m\}$. The number of records in each data set need not be constant across the collection. The patterns we are interested in are thus also subsets of A. To limit the overall number of patterns a *support threshold* is used, in the same way as in ARM. A pattern is not deemed to be "interesting" unless its number of occurrences in an individual dataset d is greater than this threshold. Some examples patterns are given in Figure 1. Thus, 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 epoch, similar sequences may be

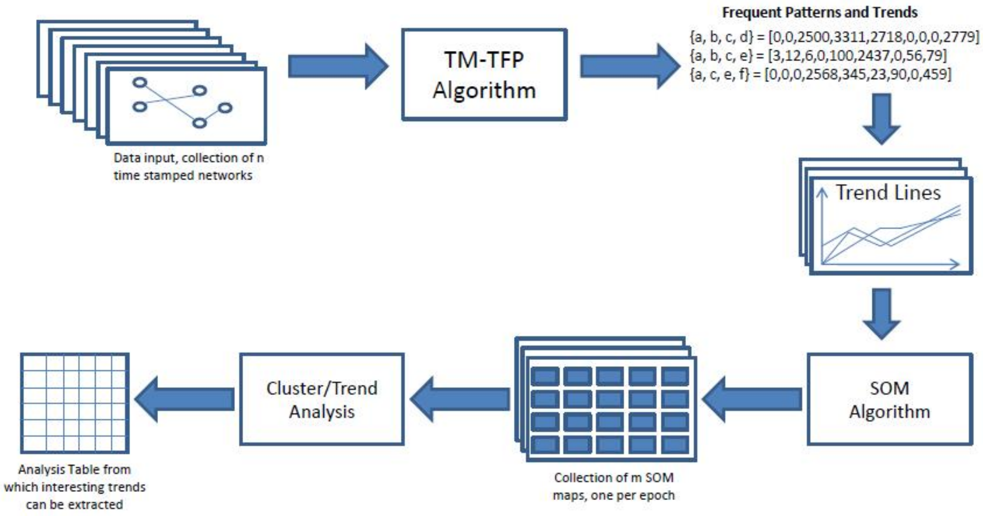
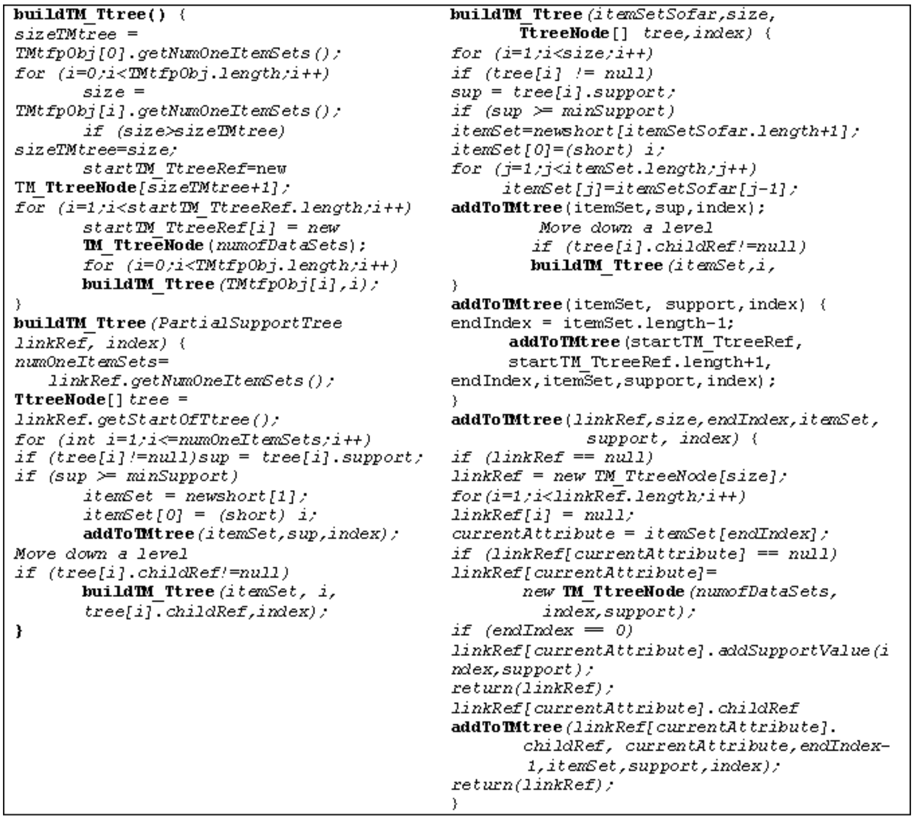

Frequent Patterns and Trends
{a, b, c, d} = [0,0,2500,3311,2718,0,0,0,2779]
{a, b, c, e} = [3,12,6,0,100,2437,0,56,79]
{a, c, e, f} = [0,0,0,2568,345,23,90,0,459]
TM-TFP Algorithm
Data input, collection of n time stamped networks
Trend Lines
SOM Algorithm
Collection of m SOM maps, one per epoch
Cluster/Trend Analysis
Analysis Table from which interesting trends can be extracted

Figure 1: Trend Mining Framework

extracted for all e epochs. Note that a 0 support value indicates a support value below the support threshold.

The trend mining system is an extended version of the TFP algorithm [11]. TFP is an established frequent pattern mining algorithm distinguished by its use of two data structures: (i) a P-tree to encapsulate the input data and hold a partial pattern count, and (ii) a T-tree to store identified frequent patterns. The T-tree is essentially a reverse *set enumeration tree* that allows fast *look up*. The TFP algorithm, in its original form, was not designed to address the temporal aspect of frequent pattern mining. The algorithm was therefore extended so that a sequence of $n \times e$ data sets could be processed and the frequent patterns stored in a way that would allow for differentiation between individual time stamps and epochs. The resulting algorithm was called Trend Mining TFP (TM-TFP) which incoporated a TM T-tree to store the collected temporal frequent patterns from $d_{n \times e}$ data sets. An overview of TM-TFP is given in Figure 2. The $buildTM\_Tree$ method processes the collection of T-trees built from the input data sets. The $addToTMtree$ method adds an item set node to the TM T-tree with its support value. The resulting trends are the input data for the clustering process which is described in the following subsection.



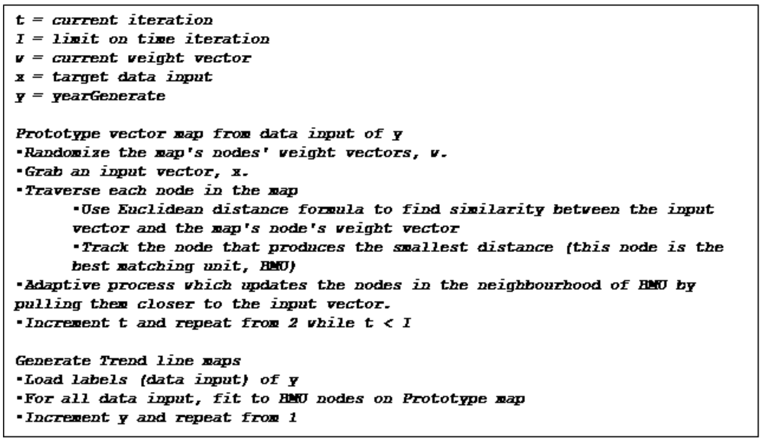
```
buildTM_Ttree() {
sizeTMtree =
TMtfpObj[0].getNumOneItemSets();
for (i=0;i<TMtfpObj.length;i++)
        size =
TMtfpObj[i].getNumOneItemSets();
        if (size>sizeTMtree)
sizeTMtree=size;
        startTM_TtreeRef=new
TM_TtreeNode[sizeTMtree+1];
for (i=1;i<startTM_TtreeRef.length;i++)
        startTM_TtreeRef[i] = new
        TM_TtreeNode(numofDataSets);
        for (i=0;i<TMtfpObj.length;i++)
        buildTM_Ttree(TMtfpObj[i],i);
}
buildTM_Ttree(PartialSupportTree
linkRef, index) {
numOneItemSets=
    linkRef.getNumOneItemSets();
TtreeNode[] tree =
linkRef.getStartOfTtree();
for (int i=1;i<=numOneItemSets;i++)
if (tree[i]!=null)sup = tree[i].support;
if (sup >= minSupport)
        itemSet = newshort[1];
        itemSet[0] = (short) i;
        addToTMtree(itemSet,sup,index);
Move down a level
if (tree[i].childRef!=null)
        buildTM_Ttree(itemSet, i,
        tree[i].childRef,index);
}
```

```
buildTM_Ttree(itemSetSofar,size,
       TtreeNode[] tree,index) {
for (i=1;i<size;i++)
if (tree[i] != null)
sup = tree[i].support;
if (sup >= minSupport)
itemSet=newshort[itemSetSofar.length+1];
itemSet[0]=(short) i;
for (j=1;j<itemSet.length;j++)
     itemSet[j]=itemSetSofar[j-1];
addToTMtree(itemSet,sup,index);
        Move down a level
       if (tree[i].childRef!=null)
       buildTM_Ttree(itemSet,i,
}
addToTMtree(itemSet, support,index) {
endIndex = itemSet.length-1;
      addToTMtree(startTM_TtreeRef,
      startTM_TtreeRef.length+1,
endIndex,itemSet,support,index);
}
addToTMtree(linkRef,size,endIndex,itemSet,
                support, index) {
if (linkRef == null)
linkRef = new TM_TtreeNode[size];
for(i=1;i<linkRef.length;i++)
linkRef[i] = null;
currentAttribute = itemSet[endIndex];
if (linkRef[currentAttribute] == null)
linkRef[currentAttribute]=
       new TM_TtreeNode(numofDataSets,
         index,support);
if (endIndex == 0)
linkRef[currentAttribute].addSupportValue(i
ndex,support);
return(linkRef);
linkRef[currentAttribute].childRef
addToTMtree(linkRef[currentAttribute].
       childRef, currentAttribute,endIndex-
        1,itemSet,support,index);
return(linkRef);
}
```

Figure 2: TM-TFP Algorithm

3.2. Trend Clustering

TM-TFP has successfully identified frequent pattern trends but produces a great many trend lines when a low support threshold is used (the option to use a higher threshold does reduce the number of trends but entails the risk of missing potentially interesting trends). The large number of trends produced makes it difficult for decision makers to interpret the result. Some mechanism for assisting the desired interpretation was therefore desirable. The idea of clustering similar trends allows decision makers to focus on particular groups of trends. The concept of clustering is well established in the data mining community, however little work has been directed at clustering time series (trend lines). The approach advocated in this paper is to use Self Organising Maps (SOMs). Using the SOM concept one map was created per epoch. The SOM was initialized with $n \times m$ nodes such that each node represented a category of trend; the map was then trained and the remaining examples assigned to nodes using a distance function. The authors experimented with different mechanisms for training the SOM, including: (i) devising specific trends to be represented by individual nodes, (ii) generating a collection of all trends that are arithmetically possible and training the SOM using this set and (iii) using some or all of the trends in the first epoch to be considered. The first required prior knowledge of the trend configurations in which we might be interested. It was discovered that the second resulted in a map for which the majority of nodes were empty. The third option was therefore adopted, the SOM was trained using the trend lines associated with the first epoch. The resulting *prototype map* was then populated with data for all *e* epochs to produce a sequence of *e* maps. Figure 3 outlines the basic SOM algorithm.

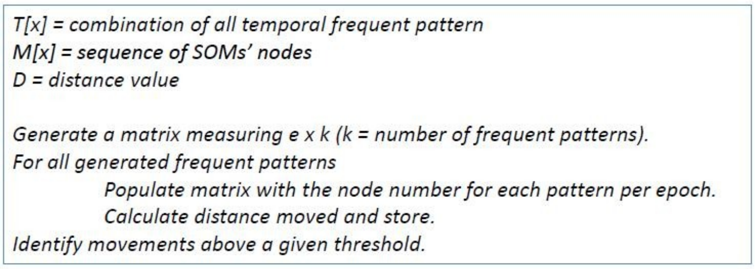


```
t = current iteration
I = limit on time iteration
v = current weight vector
x = target data input
y = yearGenerate

Prototype vector map from data input of y
•Randomize the map's nodes' weight vectors, v.
•Grab an input vector, x.
•Traverse each node in the map
      •Use Euclidean distance formula to find similarity between the input
      vector and the map's node's weight vector
      •Track the node that produces the smallest distance (this node is the
      best matching unit, BMU)
•Adaptive process which updates the nodes in the neighbourhood of BMU by
pulling them closer to the input vector.
•Increment t and repeat from 2 while t < I

Generate Trend line maps
•Load labels (data input) of y
•For all data input, fit to BMU nodes on Prototype map
•Increment y and repeat from 1
```

Figure 3: Basic SOM Algorithm

3.3. Trend Clusters Analysis

Change points in trend analysis can be interpreted in a number of ways. At its simplest, they may be interpreted as an abrupt change in direction of a trend line. A more complex interpretation may be the existence of changes in amplitude and/or frequency of fluctuating (seasonal) trends. Alternatively, an end user may be interested in an absence of change points. The interpretation applied to the cattle movement and Deeside databases is that we are interested in trends, associated with particular patterns, that change from epoch to epoch, i.e. are not consistent across the sampled temporal range. To this end, a simple trend clusters analysis technique was applied to identify trends that change location in the SOM associated with one epoch to the SOM associated with a subsequent epoch. The change can be measured by translating the trend line maps into a rectangular (D-plane) sets of coordinates and applying a Manhattan or Euclidean distance function to observe the similarities and differences of trends across the epochs. The greater the distance moved, the more significant the change. Thus, given a sequence of trend-line maps (SOMs) comparisons can be made to see how trends associated with individual frequent patterns change by analyzing the nodes in which they appear. The trend cluster analysis process is described in Figure 4.



```
T[x] = combination of all temporal frequent pattern
M[x] = sequence of SOMs' nodes
D = distance value

Generate a matrix measuring e x k (k = number of frequent patterns).
For all generated frequent patterns
        Populate matrix with the node number for each pattern per epoch.
        Calculate distance moved and store.
Identify movements above a given threshold.
```

Figure 4: Trend Clusters Analysis Pseudo code

4. Experimental Analysis

This section presents an experimental analysis of the above described approach to trend mining. The evaluation is directed at both the CTS social network and the Deeside social network. This section commences with an overview of the cattle movement database and its transformation into a social network, followed by an analysis of the trend mining process as it applied to this network. This is followed by disccussion of the Deeside insurance, trend mining and clustering process.

4.1. Cattle Movement Database

The CTS database records 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-of” 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. The 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 comprises a number of tables, the most significant of which are the animal, location and movement tables. For the analysis reported here the data from 2003 to 2006 was extracted to form 4 epochs 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 (epoch) and month (time stamp). The number of CTS records represented in each epoch 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 breed, (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. After discretisation and normalization process, we have about 445 attributes.

4.2. Cattle Movement Trend Mining

The TM-TFP algorithm was applied to the cattle movement social network and frequent pattern trends generated. For experimental purposes three support threshold values, 0.5%, 0.8% and 1.0% were used. Table 1 presents the number of frequent pattern trends discovered for each of the 4 epochs using the three support thresholds. As expected, the lower the support threshold the greater the number of generated trends. Note also that the number of trends increases exponentially. An example of the nature of a frequent pattern, in the context of the cattle movement social network, is:

$$\{numberAnimalsMoved \leq 5, SenderPTI = 4, ReceiverArea = 54,\\ SenderLocationType = Agricultural\ Holding, SenderArea = 53,\\ AnimalAge \leq 1 year\ old\}$$

(the values for the *ReceiverArea* and *SenderArea* are grid square numbers). The associated sequence of support values (for 2003) representing the trend line for that year were:

$$[2391, 2609, 3218, 3009, 3890, 2759, 2298, 3124, 2911, 3331, 3791, 2417]$$

The generated trends were clustered using the SOM technique. The SOM was initializing with 7×7 nodes, and trained using the frequent pattern trends

Table 1: Number of frequent pattern trends identified using TM-TFP for sequence of four cattle movement social network epochs and a range of support thresholds

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

produced for the (earliest) 2003 year. The resulting prototype map is shown in Figure 5. Inspection of this map shows, for example, that node 1 (top-left) represents trend lines associated with patterns with higher support in spring (March to May) and autumn (September to November). Alternatively, 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 them; the greatest dissimilarity is thus between nodes at opposite ends of the diagonals. Once the initial prototype map has been generated, a sequence of trend line maps can be produced, one for each epoch. Figure 6 gives the 2003 trend lines map. Note that in Figure 6, each node has been annotated with the number of trends in the "cluster" and that the "darker" trend lines indicate a greater number of trend lines within that cluster.

The trend clusters analysis mechanism highlighted interesting information beneficial to decision makers. Table 2 shows some example trends (representing frequent patterns) migrate from one cluster to another. Thus, the trend line $\{339\ 301\ 196\ 110\ 4\ 2\}$ representing the pattern $\{ReceiverArea = 24,\ SenderLocationType = AgriculturalHolding,\ BreedType = beef,\ CattleBreed = LimousinCross, AnimalAge \leq 1yearold, Gender = male\}$ (refer to pattern schema in Table 3) was in node 44 (bottom left in Figure 6) in 2003 and moved to node 36 in 2004, but then migrated to node 28 in 2005 and disappeared in 2006. The trend experienced a drastic change, shown by the $distance\ value$ 6.3 between 2004 and 2005.

Using the above trend clusters analysis technique, decision makers can "focus in" on particular types (clusters) of trends. In terms of further reducing the overall number of trend lines this can be achieved by considering only a subset of the detected frequent patterns according to particular attributes of interest. The term *meta-pattern* is introduced to represents a way of considering groups of patterns. In the context of the cattle movement social network, we are interested in patterns that include spatial information (i.e. sender and receiver locations).

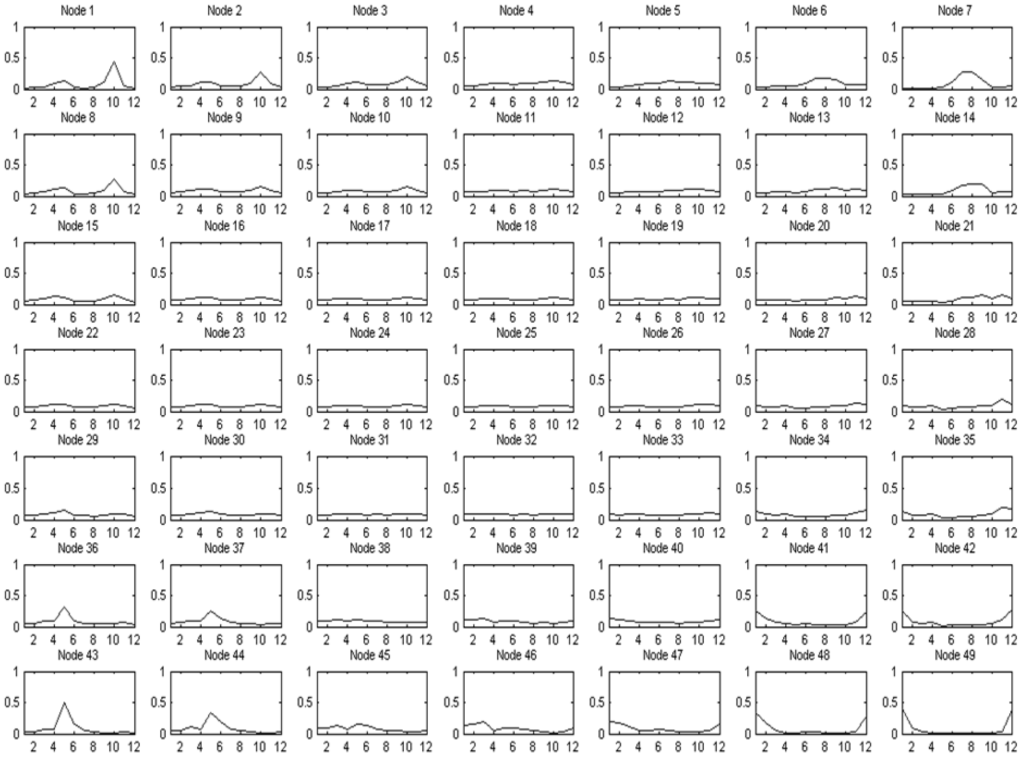


Figure 5: CTS prototype map

Four categories of meta-pattern were therefore identified:

1. **Movement from start points**: patterns that include movement and sender attributes/columns.
2. **Movement to end points**: patterns that include movement and receiver attributes/columns.
3. **Movement from start to end points**: patterns that include movement and both sender and receiver attributes/columns.
4. **Movement for other non spatial attributes**: patterns which do not feature the above.

Meta-patterns form smaller groups of patterns for cluster and trend analysis thus simplifying the cluster analysis task. Table 4 provides numbers of frequent patterns and trends according to meta-patterns category.

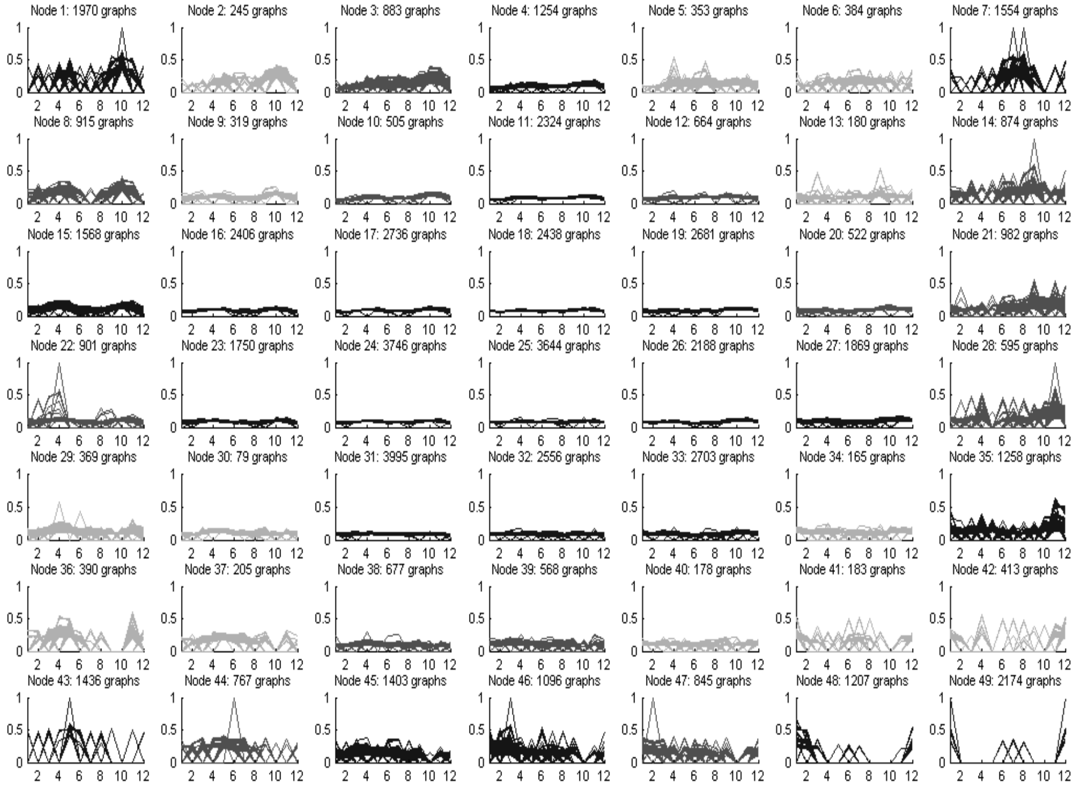


Figure 6: CTS trend line map for 2003 frequent pattern trends

4.3. Deeside Insurance Database

The Deeside insurance quotation data set [1] was extracted from a sample of records taken from the customer database operated by Deeside Insurance Ltd. Twenty-four months of data were obtained comprising, on average, 400 records per month. In total, the data set comprised 250 records with (after discretisation and normalisation) 314 attributes. The data was processed to produce a sequence of 24 networks, one per month; divided into two epochs comprising 12 months each, 2008 and 2009. The data can be viewed as representing a “star” network with Deeside at the center as a *super node* and all other nodes radiating out from it. The outlying nodes represent geographical locations defined by the first characters of customer postcodes. The links are labelled with the number of

[1] The data set was provided by Deeside Insurance Ltd, Deeside, UK.

Table 2: Examples of CTS Frequent Patterns migrating from one SOM node to another

Frequent Pattern Code	Node 2003	Dist	Node 2004	Dist	Node 2005	Dist	Node 2006
{339 301 196 110}	31	0	31	1.4	23	1.4	31
{339 301 196 110 2}	44	2.2	29	6.1	28	-	-
{339 301 196 110 4}	31	0	31	1.4	23	1.4	31
{339 301 196 110 4 2}	44	1.4	36	6.3	28	-	-
{339 301 197}	26	0	26	1.4	32	1	31

Table 3: CTS Pattern schema

Frequent Pattern Code	Pattern Schema
{339}	$ReceiverArea = 24$
{301}	$SenderLocationType = Agricultural\ Holding$
{196}	$BreedType = beef$
{197}	$BreedType = dairy$
{110}	$CattleBreed = Limousin\ Cross$
{4}	$AnimalAge \leq 1 year\ old$
{2}	$Gender = male$

interconnections between individual geographic locations and the center. Each month comprises some 1000 records. Each record consists of 13 attributes: (i) Aggregator [2], (ii) year of insurance contract, (iii) customer gender, (iv) make of car, (v) car engine size, (vi) year of manufacture, (vii) customer postcode, (viii) driver age (ix) conviction code, (x) conviction code number (xi) length of

Table 4: Statistic of CTS Meta-patterns

Meta-patterns	Number of frequent pattern trends
Movement from start points	27611
Movement to end points	22242
Movement from start to end points	12563
Movement for other non spatial attributes	23478

[2] An aggregator is a web application or search facility that allows users to obtain and compare a number of insurance quotes/prices.

Table 5: Number of frequent pattern trends identified using TM-TFP for sequence of two Deeside Insurance epochs and a range of support thresholds

Year	Support Threshold		
	2%	3%	5%
2008	314,471	142,175	55,241
2009	284,871	122,371	49,983

disqualification, (xii) fault and (xiii) penalty (note that the value for some of the attributes is *null*).

4.4. Deeside Insurance Trend Mining

Using the Deeside network, the frequent patterns and trends were again generated using TFP-TM algorithm. Since the number of Deeside insurance records was much less that of the CTS network, a slightly higher support threshold values were used, 2%, 3% and 5% respectively.

Table 5 presents the number of trends generated by applying TM-TFP to the Deeside data. Note that lower support thresholds were used than in the case of the CTS dataset because the Deeside data was smaller. The results presented in Table 5 corroborate those presented previously in Table 1. An example of a frequent pattern found in the Deeside data is:

$$\{Fault = NoBlame, LengthOfDisqualify \leq 5, Age \leq 50, \\ PostCodeArea = CH, CustomerGender = female\}$$

The associated sequence of trend occurrence values (for 2008) was:

$$[23, 0, 31, 18, 0, 4, 0, 7, 25, 9, 16, 19]$$

A 7 × 7 SOM was again used and trained using the 2008 data. The prototype map is presented in Figure 8 and the trend lines maps for insurance quotation in 2008 in Figure 8. From the figure, it can be seen, for example, that node 1 indicates a trend line with high support mainly in February, whilst node 7 shows a trend line with high support mainly in March. It is interesting to note that there are more identified patterns in the first and last quarters of the year. The prototype map was then populated with the 2008 and 2009 data to produce a sequence of two maps that could be compared. Comparison of clusters allowed for the identification of changes in customer "quote request" habits.

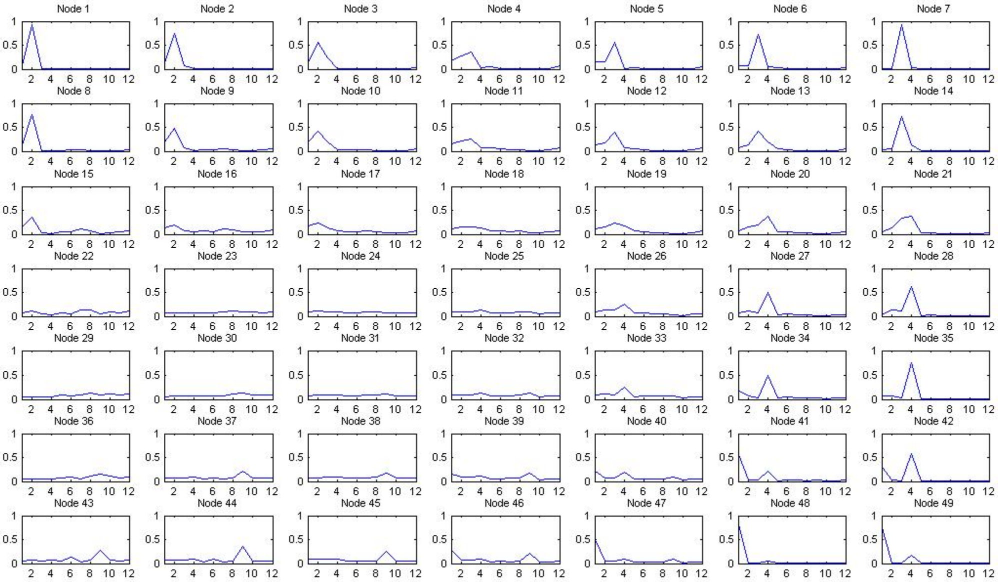


Figure 7: Deeside Insurance prototype map

Table 6 presents some examples of trend migrations identified from within the Deeside insurance data set. For example, the trend line representing the pattern $\{301\ 282\ 270\ 98\}$ which translates to $\{LengthOfDisqualify \leq 5,\ 31 \leq DriverAge \leq 40,\ PostCodeDistrict \leq 10, CarEngineSize \geq 2001\}$ (refer to pattern schema in Table 7) which was in node 23 (center left in Figure 8) in 2008 migrated to node 39 in 2009. This signifies that the pattern has changed from a trend with high support in September to the trend with high support in February and March. Moreover, the pattern trend $\{301\ 275\ 164\ 122\}$ which translates to $\{LengthOfDisqualify \leq 5,\ PostCodeSector \leq 10, PostCodeArea = CH, CarMake = MG\}$ has a high *distance value* between 2008 and 2009 to indicate the trend changed significantly from node 1 in 2008 to node 44 in 2009.

The number of Deeside frequent pattern trends is larger than that found in the CTS data. Thus, we are also interested in grouping the patterns into meta-patterns. Since the nature of the links in the Deeside data is different to that found in the CTS data, the following two meta pattern types were identified:

1. **Link with spatial attributes**: patterns that include link between Deeside node and customer nodes attributes/columns.
2. **Link with other non spatial attributes**: patterns which do not feature the above.

Table 6: Example of Deeside Insurance Frequent Patterns that migrated to other clusters

Frequent Patterns Code	Node 2008	Dist	Node 2009
$\{301\ 282\ 270\ 98\}$	23	2.8	39
$\{301\ 282\ 270\ 98\ 1\}$	24	3.2	46
$\{301\ 282\ 270\ 98\ 93\}$	24	2.2	9
$\{301\ 282\ 270\ 98\ 93\ 1\}$	11	-	-
$\{301\ 275\ 164\ 122\}$	1	6.1	44

Table 7: Deeside Pattern schema

Frequent Pattern Code	Pattern Schema
$\{301\}$	$LenDisQualify \leq 5months$
$\{282\}$	$31 \leq DriverAge \leq 40$
$\{275\}$	$PostCodeSector \leq 10$
$\{270\}$	$PostCodeDistrict \leq 10$
$\{164\}$	$PostCodeArea = CH$
$\{122\}$	$CarMake = MG$
$\{98\}$	$CarEngineSize \geq 2001cc$
$\{93\}$	$DriverGender = female$
$\{1\}$	$Aggregator = NULL$

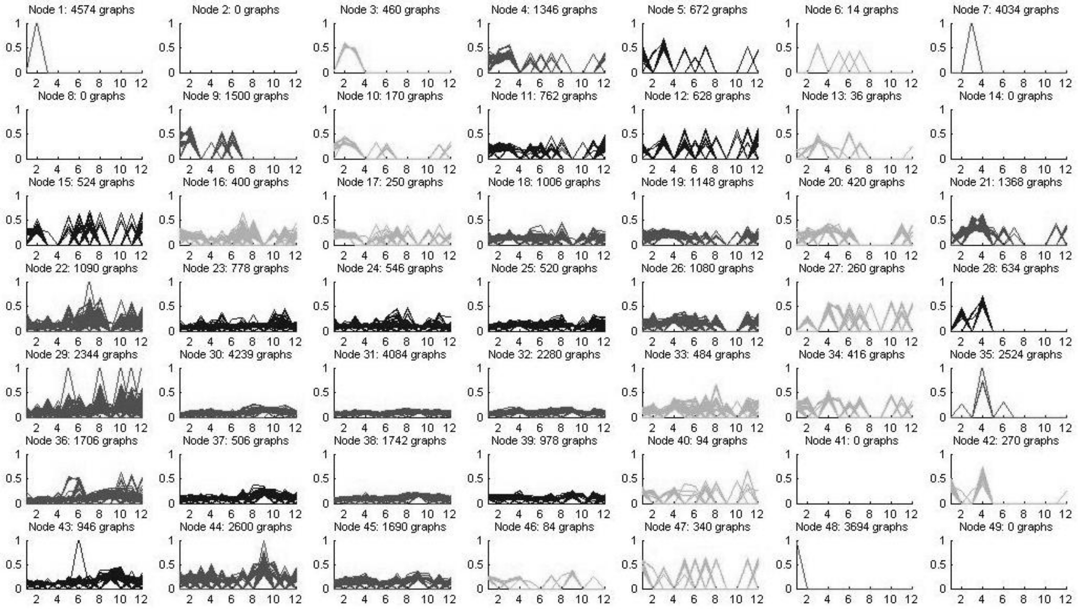


Figure 8: Deeside Insurance trend line map for 2008 frequent pattern trends

Table 8: Statistic of Deeside insurance Meta-patterns

Meta-patterns	Number of frequent pattern trends
Link with spatial attributes	11122
Link with other non spatial attributes	83280

Note that links with spatial attributes can be divided into specific geographical areas using postcodes. This can be achieved by applying constraints to the data sets during pre-processing. Table 8 provides statistics concerning the number of frequent patterns and trends according to meta-patterns category.

5. Conclusion

The paper has described a frequent trend mining technique for social network analysis. The mechanism is founded on frequent pattern mining, SOM clustering and trend clusters analysis. The aim is to discover dynamic knowledge, namely, the nature of the traffic that occurs across a social network and how trends in this traffic change with time. The proposed mechanism has been evaluated using two applications: a “complex star” social network derived from the CTS database, and a “star” network derived from the Deeside insurance data. The proposed TM-TFP algorithm is able to generate frequent time stamped patterns which can be sub-divided into epochs, which may then be compared. By employing the SOM clustering technique, the large number of trend lines that are typically identified may be grouped to facilitate a better understanding of the nature of the trends. Using the proposed cluster comparison/analysis technique, trend migrations can be discovered. For future work, the research team is currently developing methods for the visualization of the clustering result and techniques to support trend prediction that can be more effective with respect to the requirements of decision makers and stakeholders.

References

[1] Google Trends, `http://www.google.com/intl/en/trends/about.html` (2010).

[2] O.Streibel, Trend Mining with Semantic-Based Learning, Proceedings of CAiSE-DC (2008).

[3] M.S. Khan, F. Coenen, D. Reid, H. Tawfik, R. Patel, A. Lawson, A Sliding Windows based Dual Support Framework for Discovering Emerging Trends from Temporal Data, Research and Development in Intelligent Systems XXVIl, Springer London (2010), pp. 35-48.

[4] J. Raza, J.P. Liyanage, An integrated qualitative trend analysis approach to identify process abnormalities: a case of oil export pumps in an offshore oil and gas production facility, Proceedings of the Institution of Mechanical Engineers, Part E: Journal of Process Mechanical Engineering, Professional Engineering Publishing, vol 223 (4) (2008), pp. 251-258.

[5] S. Wasserman, K. Faust, Social Network Analysis: Methods and Applications, Cambridge University Press (2006).

[6] H. Lauw, E. Lim, H. Pang, T. Tan, Social Network Discovery by Mining Spatio-Temporal Events, Computational and Mathematical Organization Theory, vol 11(2), Springer Netherlands (2005), pp. 97-118.

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

[8] R. Agrawal, R. Srikant, Mining sequential patterns, 11th International Conference on Data Engineering (1995).

[9] H. Mannila, H. Toivonen, A. Verkamo, Discovery of Frequent Episodes in Event Sequences, Data Mining and Knowledge Discovery 1 (1997), pp. 259-289.

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

[11] F. Coenen, G. Goulbourne, P. Leng, Computing Association Rules Using Partial Totals, Principles of Data Mining and Knowledge Discovery. LNCS, vol. 2168, Springer Berlin / Heidelberg (2001), pp. 54-66.

[12] T. Kohonen, The Self Organizing Maps, Neurocomputing Elsevier Science, vol. 21 (1998), pp. 1-6.

[13] T. Kohonen, The Self Organizing Maps, Series in Information Sciences, vol. 30. Springer, Heidelberg (1995).

[14] J. Wang, J. Delabie, H.C. Aasheim, E. Smel, O. Myklebost, Clustering of the SOM easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study, BMC Bioinformatics, vol 3(36) (2002).

[15] S. Yan, S.S.R. Abidi, P.H. Artes, Analyzing Sub-Classifications of Glaucoma via SOM Based Clustering of Optic Nerve Images, Studies in Health Technology and Informatics, vol 116 (2005), pp. 483-488.

[16] M. Cottrell, P. Rousset, A powerful Tool for Analyzing and Representing Multidimensional Quantitative and Qualitative Data, In Proceedings of IWANN 97. LNCS, vol. 1240, Springer Berlin / Heidelberg (1997), pp. 861-871.

[17] T. Kohonen, E. Oja, O. Simula, A. Visa, J. Kangas, Engineering applications of the Self-Organizing Map, Proceedings of the IEEE, vol. 84(10) (1996), pp. 1358-1384.

[18] S. Wasserman, K. Faust, Social Network Analysis: Methods and Applications, Cambridge University Press (2006).

[19] P. Lingras, M. Hogo, M. Snorek, Temporal Cluster Migration Matrices for Web Usage Mining, In Proceedings of IEEE/WIC/ACM InternationalConference on Web Intelligence (2004).

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

[21] S. Hido, T. Id, H. Kashima, H. Kubo, H. Matsuzawa, Unsupervised changes analysis using supervised learning, Advances in Knowledge Discovery and Data Mining, 12th Pacific-Asia Conference. PAKDD. LNCS, vol. 5012 (2008), pp. 148-159.

[22] M. Richardson, P. Domingos, Mining Knowledge Sharing Sites for Viral Marketing, Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (2002), pp. 61 - 70.

[23] M.D. Choudhury, H. Sundaram, A. John, D.D. Seligmann, Can blog communication dynamics be correlated with stock market activity? Proceedings of the nineteenth ACM conference on Hypertext and hypermedia (2008), pp. 55-60.

[24] P. Gloor, J. Krauss, S. Nann, K. Fischbach, D. Schoder, Web Science 2.0: Identifying Trends Through Semantic Social Network Analysis, Social Science Research Network (2008).

[25] G. Dong, J. Li, Efficient Mining of Emerging Patterns: Discovering Trends and Differences, In Proceeding of fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (1999), pp. 43-52.

[26] E. Hadavandi, H. Shavandi, A. Ghanbari, Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting, Knowledge-Based Systems 23 (8), 2010, pp. 800-808.

[27] V.P. Subramanyam Rallabandi, S.K. Sett, Knowledge-based image retrieval system, Knowledge-Based Systems 21 (2), 2008, pp. 89-100.

[28] C. Kaiser, S. Schlick, F. Bodendorf, Warning system for online market research - Identifying critical situations in online opinion formation, Knowledge-Based Systems available online Article in Press 2011.

[29] J. Bobadilla, F. Serradilla, J. Bernal, A new collaborative filtering metric that improves the behavior of recommender systems, Knowledge-Based Systems 23 (6), 2010, pp. 520-528.

[30] W. Yuan, D. Guan, Y.-K. Lee, S. Lee, S.J. Hur, Improved trust-aware recommender system using small-worldness of trust network, Knowledge-Based Systems 23 (3), pp. 232-238.