SOMA: A proposed Framework for Trend Mining in Large UK Diabetic Retinopathy Temporal Databases

Vassiliki Somaraki[1], **Simon Harding**[2], **Deborah Broadbent**[3], **Frans Coenen**[4]

Abstract. In this paper, we present SOMA, a new trend mining framework; and Aretaeus, the associated trend mining algorithm. The proposed framework is able to detect different kinds of trends within longitudinal datasets. The prototype trends are defined mathematically so that they can be mapped onto the temporal patterns. Trends are defined and generated in terms of the frequency of occurrence of pattern changes over time. To evaluate the proposed framework the process was applied to a large collection of medical records, forming part of the diabetic retinopathy screening programme at the Royal Liverpool University Hospital.

1 Introduction

Trend mining is the process of discovering interesting trends in large time stamped datasets. The approach to trend mining advocated in this paper is to measure changes in frequently patterns that occur across time stamped (longitudinal) datasets. The focus of this paper is the longitudinal diabetic retinopathy screening data collected by the Royal Liverpool University Hospital (RLUH), a major centre for retinopathy research. The challenges of this particular data set are: (i) that it is large and complex, 150,000 episodes, comprising some

1 Department of Computer Science, University of Liverpool, UK, L69 3BX and Ophthalmology Research Unit, School of Clinical Science, University of Liverpool, Liverpool L69 3GA, UK.
V.Somaraki@liverpool.ac.uk

2 Ophthalmology Research Unit, School of Clinical Science, University of Liverpool, Liverpool L69 3GA, UK, and St. Paul's Eye Unit, Royal Liverpool University Hospital, Liverpool L7 8XP UK.
sharding@liverpool.ac.uk

3 Ophthalmology Research Unit, School of Clinical Science, University of Liverpool, Liverpool L69 3GA, UK, and St. Paul's Eye Unit, Royal Liverpool University Hospital, Liverpool L7 8XP UK.
D.M.Broadbent@liverpool.ac.uk

4 Department of Computer Science, University of Liverpool, UK, L69 3BX.
Coenen@liverpool.ac.uk

450 fields (of various types); (ii) it does not fit into any standard categorisation of longitudinal data in that the "time stamp" used is the sequential patient consultation event number where the duration between consultations is variable; and (iii) the data, in common with other patient datasets, contains many empty fields and anomalies. This last issue was addressed by developing a set of logic rules. In the context of empty fields the logic rules were used to define where values were not relevant and where data was incomplete. In the case of inter-related data, the logic rules were used to derive additional fields providing relevant definitions. To identify trends in the form of longitudinal data a trend mining framework was developed, SOMA, together with an associated trend mining algorithm (Aretaeus). Both are described in this paper.

2 Diabetic Retinopathy Databases

Diabetic Retinopathy (DR) is the most common cause of blindness in working age people in the UK. DR is a chronic multifactorial disease affecting patients with Diabetes Mellitus and causes damage to the retina [4]. Over 3,000,000 people suffer from diabetes and at least 750,000 of these people are registered blind or partially sighted in the UK. The remainder are under the risk of blindness. The RLUH has been a major centre for retinopathy research since 1991. Data collected from the diabetic retinopathy screening process is stored in a number of databases. The structure of these databases, and the tables that comprise them, reflect the mechanism whereby patients are processed, and also includes historical changes in the process [1]. The Liverpool Diabetic Eye Screening Service currently deals with some 17,000 people with diabetes registered, with family doctors, within the Liverpool Primary Care Trust per year. Consequently, a substantial amount of data is available for analysis.

3 The SOMA Trend Mining Framework

Figure 1 depicts the operation of the SOMA framework from the input of data, via the Aretaeus algorithm, to the final output. The raw data first goes to the warehouse; and then to the Data Pre-processing Software where data cleansing, creation of data timestamps, selection of subsets for analysis and the application of logic rules takes place. The data, after pre-processing, then goes to the data normalization stage, after which the frequent patterns are generated by applying the Total From Partial (TFP) frequent pattern mining algorithm [2,3] to every episode (defined by a unique time stamp) in the given data set. Then the frequent patterns and their frequency of occurrence are passed to Aretaeus algorithm to apply trend mining in order to produce different kind of prototype trends across the datasets based on the changes of the support.

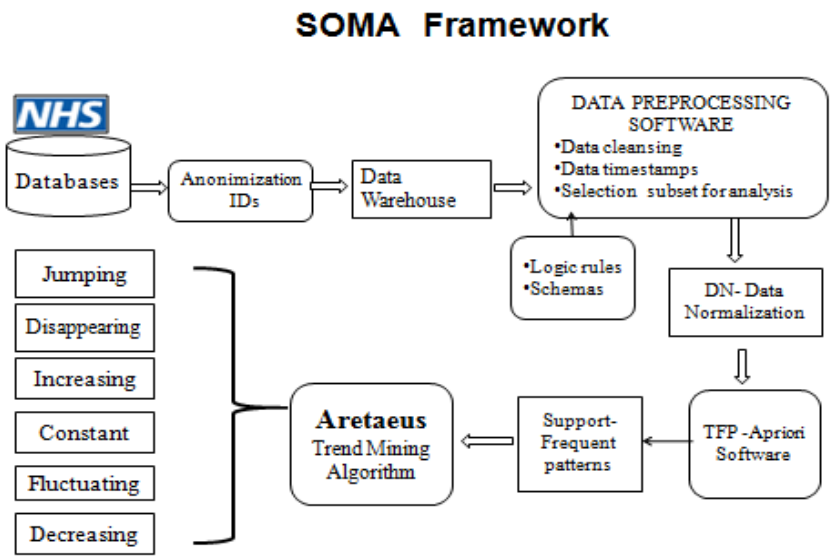


Figure 1: Representation of SOMA Framework

The Aretaeus algorithm uses mathematical identities (prototypes) to categorize trends. Let I be a frequent item set, identified within a sequence of time stamped data sets D_1, D_2… D_n, with support values of S_1, S_2, ..., S_n (where n is the number of timestamps). The *growth rate* (GR) associated with a trend is then defined as:

$$GR = \sum_{i=1}^{n-1} \frac{S_{i+1} - S_i}{S_i} + 1$$

The mathematical identities used by Aretaeus are presented in Table 1. The Aretaeus algorithm comprises the following basic steps:

1. Read, as input, the frequent patterns and their support values generated by the TFP algorithm.
2. Define the trends as vectors where the length of each vector is equal to the number of time stamps, so that each element of the vector represents a time stamp.
3. Where the support for an itemset, at any time stamp is less than the support threshold, the support value is recorded as 0.
4. Categorize the trends according to a predefined set of trend prototypes (see Table 1) to create clusters (groups) of trends.

With reference to Table 1 the Jumping and Disappearing trends can be categorized further by considering trend sub-sequences. For example a Jumping trend can be Jumping-Increasing, Jumping-Constant or Jumping-Decreasing.

Similarly the increasing, constant and decreasing categories can be combined by pairing trend sub-sequences as shown in Table 2.

Table 1. Trend Categorisation Identities

Type	**Mathematical conditions**
Increasing (Inc)	$\frac{S_{i+1}}{S_i} > 1, \forall\ i \in [1, n-1]$, GR>ρ
Decreasing (Dec)	$\frac{S_{i+1}}{S_i} < 1, \forall\ i \in [1, n-1]$
Constant (Const)	$\frac{S_{i+1}}{S_i} = 1 \pm k, \forall\ i \in [1, n-1]$, k : tolerance threshold
Fluctuating (Fluct)	$\frac{S_{i+1}}{S_i} = 1 \pm k, \forall\ i \in [1, n-1]$ *and* $\frac{S_{j+1}}{S_j} > 1, \forall\ i \in [1, n-1], j \neq i$ $\frac{S_{i+1}}{S_i} = 1 \pm k, \forall\ i \in [1, n-1]$ *and* $\frac{S_{j+1}}{S_j} < 1, \forall\ i \in [1, n-1], j \neq i$ $\frac{S_{i+1}}{S_i} > 1, \forall\ i \in [1, n-1]$ *and* $\frac{S_{j+1}}{S_j} < 1, \forall\ i \in [1, n-1], j \neq i$ $\frac{S_{i+1}}{S_i} > 1, \forall\ i \in [1, n-1]$ *and* $\frac{S_{j+1}}{S_j} < 1, \forall\ i \in [1, n-1], j \neq i$ *and* $\frac{S_{l+1}}{S_l} = 1 \pm k, \forall\ l \in [1, n-1]\ l \neq j, l \neq i$
Jumping (Jump)	*for* $m < n$: $S_i = 0, \forall\ i \in [1, m]$ *and* $S_i > 0\ \forall\ i \in [m+1, n]$
Disappearing (Disp)	*for* $m < n$: $S_i > 0, \forall\ i \in [1, m]$ *and* $S_i = 0\ \forall\ i \in [m+1, n]$

Table 2. Combinations of Increasing, Decreasing and Constant of trends subsequences

	Increasing	**Decreasing**	**Constant**
Increasing	Inc	Inc-Dec	Inc-Const
Decreasing	Dec-Inc	Dec	Dec-Const
Constant	Const-Inc	Const- Dec	Const

4 Experimental Evaluation

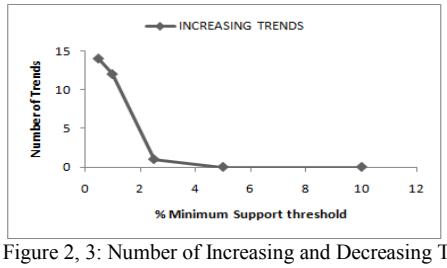
This section presents an evaluation of SOMA. The evaluation was directed at an analysis of: (i) the number of trends that might be discovered and (ii) the nature of the trend categorisation. The RLUH Diabetic Retinopathy database was used for the evaluation. The RLUH database has recorded details of some 20,000 patients spanning an eighteen year period. Patients with diabetes are screened annually. Patients enter and leave the screening programme at different times. The average

time that a patient spends within the screening process is currently six years. Thus, for the evaluation, only those patients that had taken part in the programme for at least six years were selected. Where patients had been in the programme for more than six years, data from the first six consultations was selected. This gave a dataset comprising six time stamps with 1430 records per time stamp. 7 data fields were used for the evaluation, which, after normalisation and discretisation, resulted in 215 attributes. It is worth noting that the data required significant "cleansing" to remove noise and to address the issue of empty fields.

Table 3 presents an analysis, using a sequence of support thresholds (S), of: (i) the total number of trends generated, (ii) the number of trends in each category and (iii) the run time in seconds required by the trend mining software to generate and categorise the trends. The k tolerance threshold was set to 0.05, and the ρ growth/shrink rate threshold to 1.1. It is interesting to note that no constant trends were identified (because the nature of the K threshold value used). Figures 2 to 7 plot the data presented in Table 3 so as to demonstrate the increase in the number of trends, assigned to the six categories (prototypes), as the value for S is reduced. Inspection of the figures indicates, as expected, that the number of trends decreases as the support threshold increase. Note that in Figures 2 to 7 the X-axis represents a sequence of support thresholds and the Y-axis the number of Increasing, Decreasing, Total, and Fluctuating, Jumping and Disappearing trends respectively.

Table 3. Trend Mining Framework Evaluation (p = 1.1, k = 0.05)

Support T'hold	Number of Trends						Total Num. Trend	Run Time (sec)
	Inc	Dec	Const	Disp	Fluct	Jump		
0.5	14	25	0	1827	930	1376	7602	2928.62
1.0	12	12	0	714	638	559	2532	1154.25
2.5	1	2	0	235	134	193	874	410.93
5.0	0	3	0	74	11	59	266	188.99
10.0	0	6	0	25	3	25	108	69.08



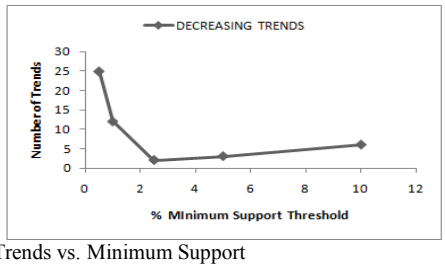
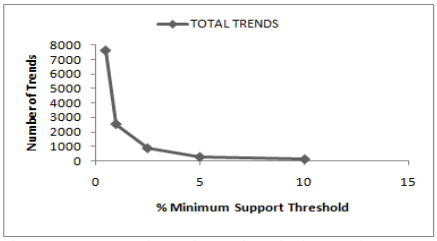


Figure 2, 3: Number of Increasing and Decreasing Trends vs. Minimum Support



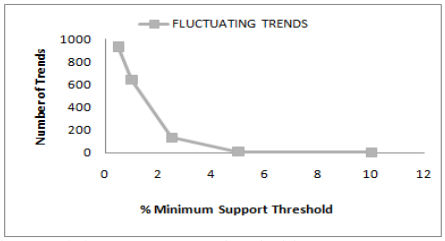
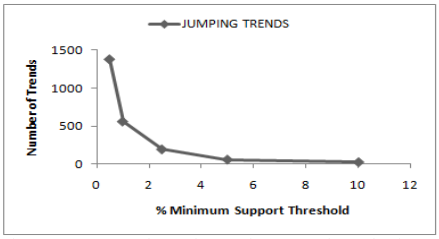


Figure 4, 5: Number of Total and Fluctuating Trends vs. Minimum Support Threshold



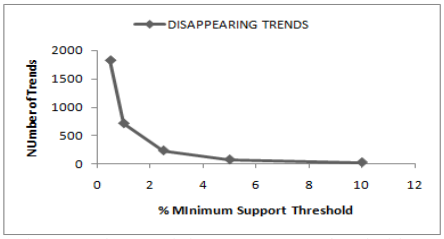


Figure 6, 7: Number of Jumping Trends and Disappearing Trends vs. Minimum Support Threshold

7 Conclusion

In this paper, we have described a novel approach to mine trends from a large amount of data. The Aretaeus algorithm allows us to generate more than 20 different kinds of trends across the datasets and is able to discover hidden, useful information across them. The fundamental idea underlying this paper is to use the support values of item sets across datasets in order to indentify useful trends. The advantage of this method is the classification of trends into categories, which is ideal for large databases. Finally, the development of a mechanism for the appropriate representation of the results using Bayesian networks is a topic of ongoing and future work, which will be particularly suitable for this purpose.

References

1. Somaraki, V., Broadbent, D., Coenen, F. and Harding, S.: Finding Temporal Patterns in Noisy Longitudinal Data: A Study in Diabetic Retinopathy. Proc. 10th Ind. Conf. on Data Mining, Springer LNAI 6171, pp418-431 (2010).
2. Coenen, F.P., Leng, P. and Ahmed, S.: Data Structures for association Rule Mining: T-trees and P-trees. IEEE Transactions on Data and Knowledge Engineering, Vol 16, No 6, pp774-778 (2004).
3. Coenen, F.P. Leng, P., and Goulbourne, G.: Tree Structures for Mining Association Rules. Journal of Data Mining and Knowledge Discovery, Vol 8, No 1, pp25-51 (2004).
4. Kanski, J.: Clinical Ophthalmology: A systematic Approach. Butterworth-Heinemann/Elsevier (2007).