Knowledge representation of large medical data using XML

Vassiliki Somaraki
Computing and Engineering
University of Huddersfield
Huddersfield, United Kingdom
v.somaraki@hud.ac.uk

Zhijie Xu
Computing and Engineering
University of Huddersfield
Huddersfield, United Kingdom
z.xu@hud.ac.uk

Abstract—**SOMA uses longitudinal data collected from the Ophthalmology Clinic of the Royal Liverpool University Hospital. Using trend mining (an extension of association rule mining) SOMA links attributes from the data. However the large volume of information at the output makes them difficult to be explored by experts. This paper presents the extension of the SOMA framework which aims to improve the post-processing of the results from experts using a visualisation tool which parse and visualizes the results, which are stored into XML structured files.**

Keywords—Medical data; Visualisation; XML; Knowledge Representation.

I. INTRODUCTION

Longitudinal data are data that are repeatedly sampled and collected over a period of time with respect to some set of subjects. Typically values for the same set of attributes are collected at each sample points. The sample points are not necessarily evenly spaced. Similarly, the data collection process for each subject need not necessarily be commenced at the same time. A regular longitudinal data set is one where data at each sample point is collected simultaneously for all subjects. Most longitudinal data sets are not regular. The most common example of irregular longitudinal data sets are patient medical records where patients enter and leave the "system" continuously and data are collected during consultations which occur at irregular intervals. One example of an irregular longitudinal database, and the focus of the research described here, is the Diabetic Retinopathy screening dataset maintained by The Royal Liverpool University Hospital (RLUH). Longitudinal data thus provide a record of the "progress" of some set of features associated with the subjects. Medical longitudinal data, such as the Diabetic Retinopathy data, typical plots the progress of some medical condition. Longitudinal data thus implicitly contains information concerning trends. However, the question that arises is how to represent the knowledge of the trend in a meaningful way. The medical records in size are huge and their representation even through trends is difficult. When, for example, the support threshold value (minimum frequency of appearance) is small the result is hundreds of trends in the output file. Thus, the output file cannot interpreted by medical staff since the amount of trends is large and it is difficult to extract the information that they need to evaluate the condition of their patients. To address this problem the SOMA framework [1] extended to include an XML-based visualization tool build in Javascript, in order to improve the representation of the results. This paper describes a data mining mechanism for extracting and focuses on visualizing trends from longitudinal data. To present this approach we used the diabetic retinopathy databases which contain data which have been collected from the Saint Paul Eye Clinic of the Royal Liverpool University Hospital collected from 1991. It includes 150,000 records, comprising some 450 fields (of various types: categorical, quantitative, text, etc.), distributed over two databases each composed of a number of tables. The SOMA framework was developed in MATLAB.

II. BACKGROUND

In this section we provide the necessary background. In particular, we describe the association rule and trend mining processes, the complexity of typical medical data, and the SOMA framework. Additionally, we describe the advantages of XML and how the use of XML can be beneficial in the representation of medical data.

A. Trend Representation

A trend can be described in a number of ways but the most obvious is as a plotting of time against some value.

From a longitudinal data trend mining perspective, we are interested in identifying all "interesting" trends in the data. The definition of "interesting" is of course subjective but we can look to work on interestingness measures conducted in the field of Association Rule Mining (ARM), such as that of [2] and [3]. The most commonly used interestingness framework in ARM is the support-confidence framework, although it has its critics. ARM is concerned with the discovery of relationships between disjoint sets of attributes that feature in the input data. The relationships are defined in terms of Association Rules (ARs) of the form "if X occurs in a record then it is likely that Y also occurs" (where X and Y are disjoint subsets of some global set of attributes A). ARs are generated from identified frequent itemsets. A number of ARs may be generated from a single frequent itemset. A frequent itemset is a subset of A that occurs frequently in the input data. The frequency of a frequent itemset X is measured in terms of a support count. A frequent itemset is deemed interesting if its support count is greater than some user specified threshold s. Consequently the number of generated frequent itemsets increases as the value of s decreases. Association Rule Mining (ARM) is a popular, and well researched, category of data mining for discovering interesting relations between variables in large databases. In (ARM) [4], an observation or transaction (e.g. the record of a clinical consultation) is represented as a set of items where an item is an attribute - value pair.

ARM procedures contain two stages: (i) frequent itemset identification, and (ii) AR generation. Piatetsky-Shapiro [5] defined ARM as a method for the description, analysis and presentation of ARs, discovered in databases using different measures of interestingness. ARM is concerned with the discovery, in tabular databases, of rules that satisfy defined threshold requirements. Of these requirements, the most fundamental one is concerned with the support (frequency) of the item sets used to make up the ARs: a rule is applicable only if the relationship it describes occurs sufficiently often in the data. Whether an item set is frequent or not, is determined by its support count. Whether an AR is relevant or not is determined by its confidence value. An item set is deemed to be frequent if its support count is above a user specified support threshold. Similarly, an AR is deemed relevant if the confidence value is above a user specified confidence threshold [6]. Although minimum support and confidence threshold help remove the uninteresting rules, many of the remaining rules are still not interesting to the users. Strong rules (high support and confidence) are not necessarily interesting rules, since interestingness may depend on a range of factors.

The idea of ARM can be applied to temporal data. In that case it is called Temporal ARM, or trend mining. 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. Trend mining is implemented using mathematical prototypes on the vectors of support in order to show how the support for each rule changes at every time stamp and thus helps the visualization tool to identify how the changes on the support may be linked or not with changes to the values of attributes either at the left or the right hand side of the rule.

B. *XML and data transformation*

An important task that SOMA framework is able to perform is to output the discovered knowledge in a clear and exploitable way. This fosters the discussion with medical experts, and allows a quick evaluation of the discovered knowledge. , and this is achieved in two ways. One way is the generation of a text output where the outcome is recorded as the name of the rule and what the values of certain parameters are for every time stamp. These parameters are support, confidence, lift, and the criteria of interestingness measures. SOMA can provide text output: for every rule, the name and corresponding support, confidence, lift, and the criteria of interestingness measures are listed. The disadvantages of this text output is the high volume of information that contains, subject to the threshold values. The lower the threshold values are, the number of trends increases, making impossible for end-users to find critical information. Therefore, the use of XML can improve the output of the SOMA significantly and turn it into a user friendly format. XML is a suitable technology for transforming unstructured data into a database. XML allows the designer to define his/hers own tags which may be organised in a hierarchical manner to structure data. Moreover, to help in structuring the data, the tags which are used in the XML contain semantic information. This built-in structure can be used both into the visualization of the data and to process the data for more advance functionality. Since XML itself is text based, it follows that it should provide a suitable way to capture textual data. XML uses terms to describe texts that are not linked to a specified formatter and therefore, makes documents platform-independent. [7].

An XML database facilitates complex searches, for example for loops or if conditions. A query language could automate the process of searching for data on more than one parameter within an XML document. The benefits of XML can be summarized below:

- XML is a well-defined, well-understood, widely used industrial standard. This means that there are a range of standard tools which can be employed to manipulate the data: one simple example is a sample application available with many XML parsing suites.
- As a generic data representation language XML files are also easily translatable into other formats through the application of stylesheets. A good example of the possible uses of stylesheets is in the creation of tailored HTML web files. Thus, arguments can be automatically summarised or made navigable for online provision.
- The acceptance of XML as a de facto industry standard facilitates data sharing: with a single common format or interlingua, separate applications can share data in the tasks of input, manipulation and output.

Although XML is very suitable for storing data, it should, however, be noted that the visualization is not actually done by the XML document itself but by another program that operates on the data in the XML file. XSLT (eXtensible Stylesheet Language Transformations) is the recommended style sheet language for XML. XSLT is far more sophisticated than CSS. With XSLT elements and attributes can be added/removed to or from the output file. You can also rearrange and sort elements, perform tests and make decisions about which elements to hide and display, and a lot more. However, in this work Javascript was selected because it is more flexible for the required work.

In SOMA framework the XML tags represent, the input of the user and the output of the framework in a meaningful and structure way. For each trend, that fulfils the criteria which are set by the user, such as support and confidence, the XML generation creates a new entry taking into account the following the selections that the user has made :

- Number of time stamps
- Variables on the left hand side of the rule
- Variables on the right hand side of the rule

From the SOMA framework the following information is passed into the XML generator for every time stamp, for each rule described in the trend and its inverse (i.e. Y=>X):

- Support Value
- Confidence Value
- Lift Value

XML visualization is implemented through a query-based tool where the user selects what kind of information wants to retrieve and visualize.

III. SOMA FRAMEWORK

SOMA [8] is a trend-mining framework for knowledge discovery from large databases, the development of a validation framework for trend mining, and the application of trend mining in medical data. The framework exploits three steps: pre-processing, association rule mining, and trend mining. During pre-processing, data from different sources are brought together after applying logic rules to deal with problems arising from the nature of data and to create a time-stamped subset for analysis.

During the association rule mining process, rules are extracted by considering a set of variables. Finally, trend mining takes information from the previous step, evaluates the attitude of the rules over time and estimates their interestingness.

In SOMA time-stamp datasets are passed through the main processing, in which the ARM technique of matrix algorithm is applied to identify frequent rules with acceptable confidence. Mathematical conditions are applied to classify the sequences of support values into trends.

In this work we will focus on the evaluation of interestingness of identified rules. Strong rules, according to support and confidence, are not always interesting. In order to determine which trend is interesting, SOMA exploits five measures introduced by Han et al. [9]. If Lift value equals 1 the itemsets of a rule are independent, if a Lift value is less than 1 then they are negatively related, and if Lift is greater than 1the they are positive correlated. Other measures considered are: all confidence, max confidence, Kulczynski and cosine. The reader is referred to [10] for a description of those measures.

Each measure is only influenced by the supports of X, and Y, but not by the total number of transactions. The values of the measures range from 0 to 1. The higher the value the closer the relationship between X and Y is.

In SOMA, rules are deemed to be interesting if, for each considered time stamp, the sum of the values of the 5 considered measures is over a threshold given by the final user.

A set of measures used in this study to examine how some parameters affect the self-consistence of trend mining framework and how finally they affect to the results. In this paper we are only focus to time related measures and also some parameters whose values control the effectiveness of the framework. Generally, to this framework a number of parameters used:

A. *Number of time stamps*

The first parameter for the framework is the number of time stamps. The size of the dataset is determined by the number of time stamps. Also, another factor that is very important is the time window which determines how data recorded in different times can be collocated into a time stamp.

B. *Completeness of dataset*

This measure refers to the degree of complexity of the datasets and how much information they contain. Even using logic rules at the pre-processing stage, it is not possible to fill all the empty values. The numerator of the ratio, of the support count of an item set over the total number of transactions of a dataset, that is used to calculate the support value, is the number of occurrence of an item set. The more complete a dataset is, the more information can be extracted from it.

C. *Rule conflict*

Sometimes, if a dataset is very dense (very large), there is the probability that an item set of “variable attributes” (the antecedent part of the rule) belongs to two or more different “key-variable” values (consequent).

If $X = \{x_1, x_2, ..., x_n\}$ is an item set, the following rules are a set of conflict rules:

- If X then Y1
- If X then Y2

where Y1 and Y2 are item sets of “key attributes”. In such a case, the methodology to deal with this problem will affect the final results of trend mining. One way to deal with the conflict rules is to discard both of them from the results. Another way is to perform a comparison between the conflict rules in terms of support and confidence across all time stamps.

D. *Banding of continuous attributes*

Medical data values are usually from continuous domains. The presence of continuous domains makes it difficult to apply the frequent item set techniques. For this reason, in SOMA pre-processing discretisation is applied. Discretisation allocates continuous values into a limited number of intervals, called bands [11], [12].Bands can be either defined by domain experts -in our analysis, physicians- or automatically identified

E. Parameterization

In the trend mining framework, there are four parameters whose values control the effectiveness of the framework:

- support threshold : the minimum support required for an item set;
- confidence threshold: the minimum confidence required for an association rule;
- growth rate : the rate that shows how the support increases across all time stamps;
- tolerance : parameter to determine a constant trend.

All the above parameters are user-specified. The support threshold and confidence threshold control which rules from a dataset are kept and which are discarded. Also, the support threshold controls the type of trend: if support for a rule is above the threshold at all time stamps, the trend could be increasing, decreasing, or fluctuating. Otherwise it would fall into the category of jumping or disappearing.

The growth-rate threshold determines if the increase in the support of a rule is sufficient to be characterized as an increasing trend. However, tolerance is the threshold at which, when the growth rate is less than the tolerance, the trend is characterized as constant.

F. Validation

Researchers have found that diabetic patients who are able to maintain appropriate blood sugar levels have fewer eye problems than those with poor control. Diet and exercise play important roles in the overall health of those with diabetes. These are some examples of common-sense clauses that can be justified from the trends. The more trends SOMA produces (smaller support), the more clauses can be "exported". The confidence of the trend also plays an important role, as it measures the validity of the trend. The higher the confidence of the role, the more valid it is.

The validation of the entire SOMA framework is based on known associations between the selected attributes. Among the attributes that have been selected, there should be at least one that has the role of the "key attribute". In this research, a "key attribute" could be an attribute that characterizes the status of diabetic retinopathy for each patient. The other attributes play the role of variables, "variable attributes" whose values affect the "key attribute". At the end of the SOMA framework, the rules that are produced are compared with known associations, given by the experts. For more information on the validation of the framework the reader is referred to [8].

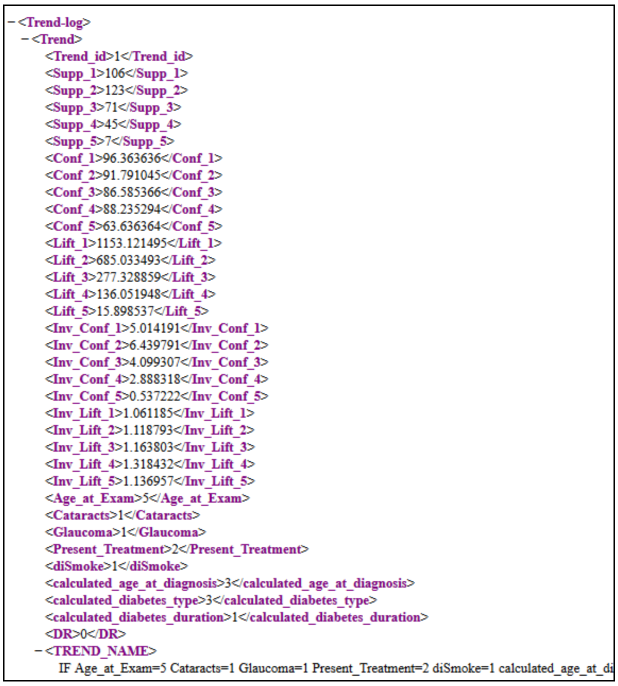
G. Complexity of medical data

Modern medicine generates a great deal of information stored in medical databases, and it has become increasingly necessary to extract useful knowledge and provide scientific decision-making for the diagnosis and treatment of disease from the database. Because the medical information is characteristic of redundancy, multi-attribution, incompletion and closely related with time, the medical data mining differs from others.

In particular, four peculiar characteristics of medical data have been identified [13]: (i) Redundancy: a typical medical database is a huge data resource which collects data from different sources. It may contain repeated, irrelevant, and even contradictory records; (ii) Complexity: as the medical data obtained from medical imaging, laboratory data and the exchange between doctors and patients, they are in various forms. These include images (SPECT), signals (ECG), pure data (the signs of parameters, test results), and text; (iii) Privacy: medical information is related to patients, and must be handled confidentially. Moreover, most of the data cannot be shared; (iv) Missing values: Medical data collection is always out of line with the stage of processing. The main purpose of medical data collection is to cure sickness and save patients' lives. However, the purpose of medical data processing is to determine regular patterns in certain diseases. In this case, the collected data may not meet the need to cover all the information. In addition, human factors may lead to errors and incomplete information in patients' records and the expression of many medical data is uncertain.

H. XML Visualisation tool

As stated in the previous sections, depending on the parameters selection from the user, the extracted trends from SOMA can be large in amounts (>1000). Moreover, the output in text format was not the appropriate for the user to explore the results, since those consist of a high amount of pages. Here it is presented an extension of the framework an XML visualization tool. The principal is that SOMA exports its results in an XML file using a predefined format (predefined tag names for the XML structure) which is programed and it is embedded into the main code of the SOMA framework. Based on than predefined format and using JAVASCRIPT a platform has been created for the visualization of the results. Figure 1 depicts the structure of an XML file which contains the results from a run using SOMA. The XML file contains information regarding certain parameters such as support, confidence and lift values, which fields have been selected from the user and their values, too. Finally, in the XML file it is recorder the trend name in the format if X then Y. The tool can provide statistical information about the trends like maximum values and metrics on the types of trends, can return back specific type of trends and also allows the user to search for specific trends by their field.



```
- <Trend-log>
  - <Trend>
      <Trend_id>1</Trend_id>
      <Supp_1>106</Supp_1>
      <Supp_2>123</Supp_2>
      <Supp_3>71</Supp_3>
      <Supp_4>45</Supp_4>
      <Supp_5>7</Supp_5>
      <Conf_1>96.363636</Conf_1>
      <Conf_2>91.791045</Conf_2>
      <Conf_3>86.585366</Conf_3>
      <Conf_4>88.235294</Conf_4>
      <Conf_5>63.636364</Conf_5>
      <Lift_1>1153.121495</Lift_1>
      <Lift_2>685.033493</Lift_2>
      <Lift_3>277.328859</Lift_3>
      <Lift_4>136.051948</Lift_4>
      <Lift_5>15.898537</Lift_5>
      <Inv_Conf_1>5.014191</Inv_Conf_1>
      <Inv_Conf_2>6.439791</Inv_Conf_2>
      <Inv_Conf_3>4.099307</Inv_Conf_3>
      <Inv_Conf_4>2.888318</Inv_Conf_4>
      <Inv_Conf_5>0.537222</Inv_Conf_5>
      <Inv_Lift_1>1.061185</Inv_Lift_1>
      <Inv_Lift_2>1.118793</Inv_Lift_2>
      <Inv_Lift_3>1.163803</Inv_Lift_3>
      <Inv_Lift_4>1.318432</Inv_Lift_4>
      <Inv_Lift_5>1.136957</Inv_Lift_5>
      <Age_at_Exam>5</Age_at_Exam>
      <Cataracts>1</Cataracts>
      <Glaucoma>1</Glaucoma>
      <Present_Treatment>2</Present_Treatment>
      <diSmoke>1</diSmoke>
      <calculated_age_at_diagnosis>3</calculated_age_at_diagnosis>
      <calculated_diabetes_type>3</calculated_diabetes_type>
      <calculated_diabetes_duration>1</calculated_diabetes_duration>
      <DR>0</DR>
    - <TREND_NAME>
        IF Age_at_Exam=5 Cataracts=1 Glaucoma=1 Present_Treatment=2 diSmoke=1 calculated_age_at_di
```

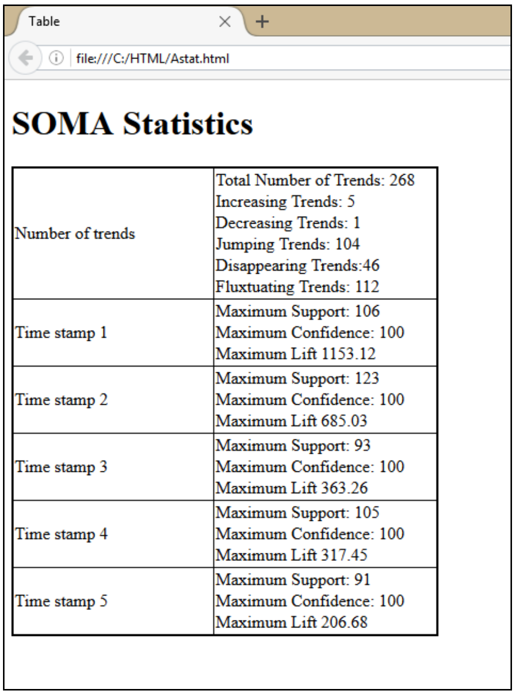
Fig. 1. Example of and XML SOMA output file

IV. Experimental Analysis

In this Section we firstly introduce the data that have been exploited in our experimental analysis, and then discuss the results obtained by using the proposed SOMA-XML framework for presenting the results.

A. Settings and results

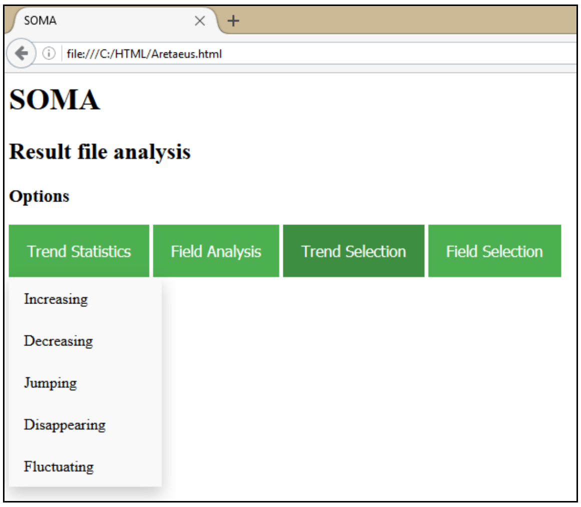
In this work we considered the data of the Saint Paul's Eye Clinic of the Royal Liverpool University Hospital, UK. The data (anonymised in order to guarantee patients' privacy) was collected from a warehouse with 22,000 patients, 150,000 visits to the hospital, with attributes including demographic details, visual acuity data, photographic grading results, data from biomicroscopy of the retina and results from biochemistry investigations. Stored information had been collected between 1991 and 2009. The data collection is large and complex; comprising about 450 attributes distributed over two databases each composed of a number of datasets. Data are noisy and in that contain missing and anomalous information. The data are longitudinal; they are repeatedly sampled and collected over a period of time with respect to some set of subjects. Typically, values for the same set of attributes are collected at each sample points. The sample points are not necessarily evenly spaced. Similarly, the data collection process for each subject need not necessarily be commenced at the same time. In our experimental analysis we considered all the 2820 patients who had readings over 5 time stamps. The percentage of missing values is 6.57% for this test. The stored attributes that are analysed in this work include: (i) background information on patients (ii) general demographic patient details, like age, sex, etc.; (iii) visual acuity data; (iv) photo details, results from the photographic grading; (v) biomicroscopy, results from the slit lamp biomicroscopy in cases where this has been conducted; and (vi) risk factors, blood pressure and biochemistry investigations known to be associated with an increased risk of progression of retinopathy. Here the selected fields are age at exam, presence of glaucoma or cataracts, smoking, type of diabetes, treatment of diabetes, duration of diabetes age, age when diabetes firstly diagnosed and if the patient suffers from diabetic retinopathy or not. Figure 2 shows the statistical analysis for this example. The depicted table gives abstract information about the trends. Figure 3 shows the trend selection function. The user can select what kind of trend wants to see (e.g. increasing) and the tool will return all the trends that belong into this category. Figure 4 shows the "search engine" for targeting trends, where the user selects one or more attributes by typing the desired value (or values when looking for more than one attribute), and Figure 5 the outcome of the search engine. The advantage of the advocated search engine is that the user can search quickly trends with specific attributes. When SOMA firstly introduced [8], the results were stored into a document file. As a result, the required trend could have been within 100 pages or more. With the introduced visualization tool, this drawback has been successfully dealt since the search engine allows the user to run "targeted" investigations into the trends he/she considers to be more important or more interest.



SOMA Statistics

Number of trends	Total Number of Trends: 268 Increasing Trends: 5 Decreasing Trends: 1 Jumping Trends: 104 Disappearing Trends:46 Fluxtuating Trends: 112
Time stamp 1	Maximum Support: 106 Maximum Confidence: 100 Maximum Lift 1153.12
Time stamp 2	Maximum Support: 123 Maximum Confidence: 100 Maximum Lift 685.03
Time stamp 3	Maximum Support: 93 Maximum Confidence: 100 Maximum Lift 363.26
Time stamp 4	Maximum Support: 105 Maximum Confidence: 100 Maximum Lift 317.45
Time stamp 5	Maximum Support: 91 Maximum Confidence: 100 Maximum Lift 206.68

Fig. 2. Summary table.

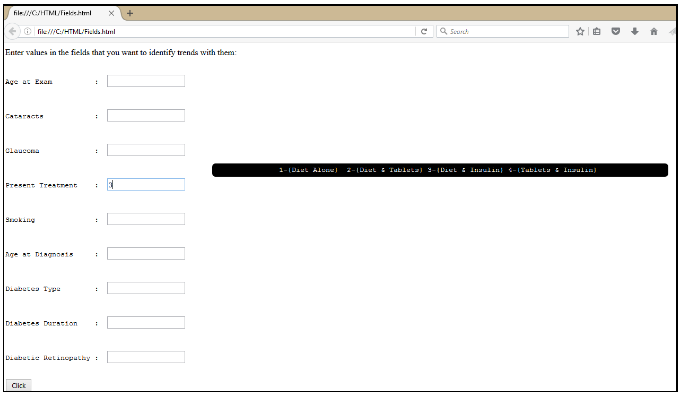


SOMA

Result file analysis

Options

Fig. 3. Trend selection function



Enter values in the fields that you want to identify trends with them:

Age at Exam :

Cataracts :

Glaucoma :

Present Treatment : 3

1-{Diet Alone} 2-{Diet & Tablets} 3-{Diet & Insulin} 4-{Tablets & Insulin}

Smoking :

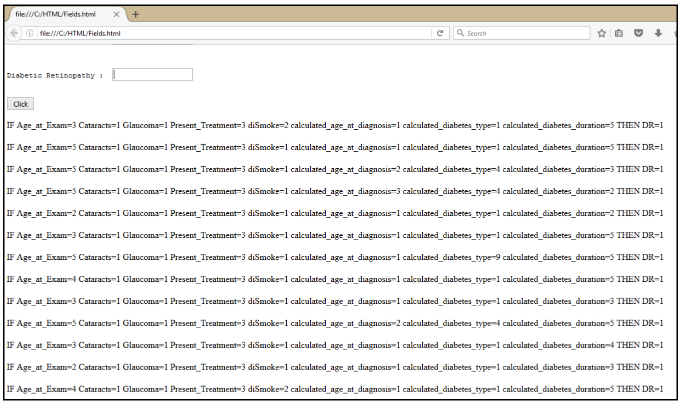
Age at Diagnosis :

Diabetes Type :

Diabetes Duration :

Diabetic Retinopathy :

Fig. 4. Trends search engine.



Diabetic Retinopathy :

```
IF Age_at_Exam=3 Cataracts=1 Glaucoma=1 Present_Treatment=3 diSmoke=2 calculated_age_at_diagnosis=1 calculated_diabetes_type=1 calculated_diabetes_duration=5 THEN DR=1
IF Age_at_Exam=5 Cataracts=1 Glaucoma=1 Present_Treatment=3 diSmoke=1 calculated_age_at_diagnosis=1 calculated_diabetes_type=1 calculated_diabetes_duration=5 THEN DR=1
IF Age_at_Exam=5 Cataracts=1 Glaucoma=1 Present_Treatment=3 diSmoke=1 calculated_age_at_diagnosis=2 calculated_diabetes_type=4 calculated_diabetes_duration=3 THEN DR=1
IF Age_at_Exam=5 Cataracts=1 Glaucoma=1 Present_Treatment=3 diSmoke=1 calculated_age_at_diagnosis=3 calculated_diabetes_type=4 calculated_diabetes_duration=2 THEN DR=1
IF Age_at_Exam=2 Cataracts=1 Glaucoma=1 Present_Treatment=3 diSmoke=1 calculated_age_at_diagnosis=1 calculated_diabetes_type=1 calculated_diabetes_duration=2 THEN DR=1
IF Age_at_Exam=3 Cataracts=1 Glaucoma=1 Present_Treatment=3 diSmoke=1 calculated_age_at_diagnosis=1 calculated_diabetes_type=1 calculated_diabetes_duration=5 THEN DR=1
IF Age_at_Exam=5 Cataracts=1 Glaucoma=1 Present_Treatment=3 diSmoke=1 calculated_age_at_diagnosis=1 calculated_diabetes_type=9 calculated_diabetes_duration=5 THEN DR=1
IF Age_at_Exam=4 Cataracts=1 Glaucoma=1 Present_Treatment=3 diSmoke=1 calculated_age_at_diagnosis=1 calculated_diabetes_type=1 calculated_diabetes_duration=5 THEN DR=1
IF Age_at_Exam=3 Cataracts=1 Glaucoma=1 Present_Treatment=3 diSmoke=1 calculated_age_at_diagnosis=1 calculated_diabetes_type=1 calculated_diabetes_duration=3 THEN DR=1
IF Age_at_Exam=5 Cataracts=1 Glaucoma=1 Present_Treatment=3 diSmoke=1 calculated_age_at_diagnosis=2 calculated_diabetes_type=4 calculated_diabetes_duration=5 THEN DR=1
IF Age_at_Exam=3 Cataracts=1 Glaucoma=1 Present_Treatment=3 diSmoke=1 calculated_age_at_diagnosis=1 calculated_diabetes_type=1 calculated_diabetes_duration=4 THEN DR=1
IF Age_at_Exam=2 Cataracts=1 Glaucoma=1 Present_Treatment=3 diSmoke=1 calculated_age_at_diagnosis=1 calculated_diabetes_type=1 calculated_diabetes_duration=3 THEN DR=1
IF Age_at_Exam=4 Cataracts=1 Glaucoma=1 Present_Treatment=3 diSmoke=2 calculated_age_at_diagnosis=1 calculated_diabetes_type=1 calculated_diabetes_duration=5 THEN DR=1
```

Fig. 5. Results from search engine

V. Conclusion and Future Work

In this paper, the extension of SOMA framework is presented. The extension is a robust visualization tool that uses the advantages of the XML along with a predefined format for storing the results. This allows the user to explore and navigate into the results of large medical databases with the click of buttons in order to identify critical information for medical conditions, here for diabetic retinopathy. Both the XML file and the visualization tool would be extended in the future in order to provide more information combining the patients and their conditions in order to identify critical change points and optimize interval for subsequent visits to their consultants.

References

[1] V. Somaraki, D. Broadbent, F. Coenen, and S. Harding. Finding temporal patterns in noisy longitudinal data: a study in diabetic retinopathy. In Advances in Data Mining. Applications and Theoretical Aspects, p. 418-431, 2010.

[2] V. B. M. P. Lenca, P. and S. Lallich. Association rule interestingness measures: Experimental and theoretical studies. In Quality Measures in Data Mining. Studies in Computational Intelligence, p. 51-76,2007.

[3] Z. L. N. G. Zhang, Y. and Y. Shi. A survey of interestingness measures for association rules. In International Conference on Business Intelligence and Financial Engineering , p. 460-463, 2009.

[4] Piatetsky-Shapiro, Gregory. *Discovery, analysis, and presentation of strong rules*, in Piatetsky-Shapiro, Gregory; and Frawley, William J.; eds., *Knowledge Discovery in Databases*, AAAI/MIT Press, Cambridge, MA, 1991.

[5] G. Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. Knowledge discovery in databases , p. 229-238, 1991.

[6] J. D. Singer and J. B. Willett. Applied longitudinal data analysis: Modeling change and event occurrence Oxford university press, 2003.

[7] S. DeRose. Xml and the TEI.Computers and the Humanities , p. 11-30, 1999.

[8] V. Somaraki. A framework for trend mining with application to medical data, PhD Thesis, 2013.

[9] J. Han, M. Kamber, and J. Pei. Data mining: Concepts and techniques, (The morgan kaufmann series in data management systems). 2006.

[10] V. Somaraki,, M. Vallati and L. McCluskey. Discovering Interesting Trends in Real Medical Data: A study in Diabetic Retinopathy. In: Progress in Artificial Intelligence : 17th Portuguese Conference on Artificial Intelligence, EPIA 2015 Proceedings. Lecture Notes in Computer Science, 9273 . Springer, p. 134-140, 2015.

[11] S. Kotsiantis and D. Kanellopoulos. Discretization techniques: A recent survey. GESTS International Transactions on Computer Science and Engineering, p. 47-58, 2006.

[12] H. Liu, F. Hussain, C. L. Tan, and M. Dash. Discretization: An enabling technique. Data mining and knowledge discovery , p. 393-423, 2002.

[13] J. Zhao and T. Wang. A general framework for medical data mining. In Proceedings of the International Conference on Future Information Technology and Management Engineering (FITME),p. 163-165, 2010