Obstetric Medical Record Processing and Information Retrieval

Miroslav Bursa[1], Lenka Lhotska[1], Vaclav Chudacek[1], Michal Huptych[1], Jiri Spilka[1], Petr Janku[2], and Martin Huser[2]

[1] Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic
miroslav.bursa@fel.cvut.cz
http://bio.felk.cvut.cz
[2] Obstetrics and Gynaecology clinic, University Hospital in Brno, Czech Republic

Abstract. This paper describes the process of mining information from loosely structured medical textual records with no apriori knowledge. In the paper we depict the process of mining a large dataset of ∼50,000–120,000 records × 20 attributes in database tables, originating from the hospital information system (thanks go to the University Hospital in Brno, Czech Republic) recording over 10 years. This paper concerns only textual attributes with free text input, that means 613,000 text fields in 16 attributes. Each attribute item contains ∼800–1,500 characters (diagnoses, medications, etc.). The output of this task is a set of ordered/nominal attributes suitable for rule discovery mining and automated processing.

Information mining from textual data becomes a very challenging task when the structure of the text record is very loose without any rules. The task becomes even more difficult when natural language is used and no apriori knowledge is available. The medical environment itself is also very specific: the natural language used in textual description varies with the personality creating the record (there are many personalized approaches), however it is restricted by terminology (i.e. medical terms, medical standards, etc.). Moreover, the typical patient record is filled with typographical errors, duplicates, ambiguities and many (nonstandard) abbreviations.

Note that this project is an ongoing process (and research) and new data are irregularly received from the medical facility, justifying the need for robust and fool-proof algorithms.

Keywords: Swarm Intelligence, Ant Colony, Textual Data Mining, Medical Record Processing, Hospital Information System.

1 Introduction

1.1 Motivation

In many industrial, business, healthcare and scientific areas we witness the boom of computers, computational appliances, personalized electronics, high-

P. Kostkova, M. Szomszor, and D. Fowler (Eds.): eHealth 2011, LNICST 91, pp. 26–33, 2012.

speed networks, increasing storage capacity and data warehouses. Therefore a huge amount of various data is transferred and stored, often mixed from different sources, containing different data types, unusual coding schemes, and seldom come without any errors (or noise) and omissions. Massively parallel distributed storage systems are used nowadays to provide computational nodes with data in reasonable time.

1.2 Nature Inspired Methods

Nature inspired metaheuristics play an important role in the domain of artificial intelligence, offering fast and robust solutions in many fields (graph algorithms, feature selection, optimization, clustering, etc). Stochastic nature inspired metaheuristics have interesting properties that make them suitable to be used in data mining, data clustering and other application areas.

With the boom of high-speed networks and increasing storage capacity of database clusters and data warehouses, a huge amount of various data can be stored. *Knowledge discovery* and *Data mining* is not only an important scientific branch, but also an important tool in industry, business and healthcare. These techniques target the problematic of processing huge datasets in reasonable time – a task that is too complex for a human. Therefore computer-aided methods are investigated, optimized and applied, leading to the simplification of the processing of the data.

Ant Algorithms. Ant colonies inspired many researchers to develop a new branch of stochastic algorithms: *ant colony inspired algorithms*. Based on the ant metaphor, algorithms for both static and dynamic combinatorial optimization, continuous optimization and clustering have been proposed. They show many properties similar to the natural ant colonies, however, their advantage lies in incorporating the mechanisms, that allowed the whole colonies to effectively survive during the evolutionary process.

1.3 Knowledge Extraction

Several techniques to extract knowledge from raw data have been developed in the past. These techniques have various and multiple origins: some result from the statistical analysis of the data, the regressions, decision trees, etc.; some resulting from the artificial intelligence such as the expert systems, intelligent agents, fuzzy logic, etc.

Text Extraction. The accuracy for relation extraction in journal text is typically about 60 % [1]. A perfect accuracy in text mining is nearly impossible due to errors and duplications in the source text. Even when linguists are hired to label text for an automated extractor, the inter-linguist disparity is about 30 %. The best results are obtained via an automated processing supervised by a human [2].

Ontologies have become an important means for structuring knowledge and building knowledge-intensive systems. For this purpose, efforts have been made to facilitate the ontology engineering process, in particular the acquisition of ontologies from texts.

2 Input Dataset Overview

The dataset consists of a set of approx. 50 to 120 thousand records (structured in different relational DB tables; some of them are not input, therefore the range is mentioned) × approx. 20 attributes. Each record in an attribute contains about 800 to 1500 characters of text (diagnoses, patient state, anamneses, medications, notes, references to medical stuff, etc.). For textual mining, 16 attributes are suitable.

The overview of one small (in field length) attribute is visualized in Fig. [2]. Only a subsample (about 5 %) of the dataset could be displayed in this paper, as the whole set would render into a uncomprehensible black stain. The vertices (literals) are represented as a green circle, the size reflects the literal (i.e. word) frequency. Edges represent transition states between literals (i.e. the sequence of 2 subsequent words in a sentence/record); edge stroke shows the transition rate (probability) of the edge. The same holds for all figures showing the transition graph, only a different visualization approach has been used.

It is clear, that human interpretation and analysis of the textual data is very fatiguing, therefore any computer aid is highly welcome.

3 Graph Explanation

In this paper we describe *transition graphs*. These are created for each attribute. An attribute consists of many records in form of a sentence. By *sentence* we hereby mean a sequence of literals, not a sentence in a linguistic form. The records are compressed – unnecessary words (such as verbs *is*, *are*) are omitted. In this paper, only the atribute delivery_anesthetics is visualized, as it is the simplest one.

Vertices of the transition graph represent the words (separated by spaces) in the records. For each word (single or multiple occurence) a vertex is created and its potence (number of occurences is noted). For example, the words *mesocaine*, *anesthetics*, *not*, *mL* form a vertex. Note that also words as *mesocain*, *mezokain* and other versions of the word *mesocaine* are present. For a number (i.e. sequence of digits) a special literal *_NUMBER_* is used.

Edges are created from single records (sentences entered). For example the sentence *mesocaine 10 mL* would add edges from vertex *mesocaine* to vertex *_NUMBER_* and from vertex *_NUMBER_* to the vertex *mL* (or the edge count is increased in case it exists). For all records, the count of the edges is also useful. It provides an overview on the inherent structure of the data – the most often word transitions.

4 Motivation

The task of this work is to provide the researchers with a quick automated or semi-automated view on the textual records. Textual data are not easy to visualize. The word frequency is inapropriate, although it is very simple. Therefore we decided to extract information in the form of a transition graph.

Using these graphs a set of rules for information retrieval is bein created (defined). These rules serve for extraction of (boolean) attributes from the textual rules. These attributes are used in automated rule discovery and can be further used for recommendation.

5 Nature Inspired Techniques

Social insects, i. e. ant colonies, show many interesting behavioral aspects, such as self-organization, chain formation, brood sorting, dynamic and combinatorial optimization, etc. The coordination of an ant colony is of local nature, composed mainly of indirect communication through pheromone (also known as *stigmergy.* The high number of individuals and the decentralized approach to task coordination in the studied species means that ant colonies show a high degree of parallelism, self-organization and fault tolerance. In studying these paradigms, we have high chance to discover inspiration concepts for many successful metaheuristics.

5.1 Ant Colony Optimization

Ant Colony Optimization (ACO) [3] is an optimization technique that is inspired by the foraging behavior of real ant colonies. Originally, the method was introduced for the application to discrete and combinatorial problems.

Ant Colony Methods for Clustering. Several species of ant workers have been reported to form piles of corpses (cemeteries) to clean up their nests. This aggregation phenomenon is caused by attraction between dead items mediated by the ant workers. The workers deposit (with higher probability) the items in the region with higher similarity (when more similar items are present within the range of perception). For example, the *Messor sancta* ants organize dead corpses into clusters; brood sorting has been studied in ant colony of *Leptothorax unifasciatus.*

This approach has been modeled in the work of Deneubourg et al. [4] and in the work of Lumer and Faieta [5] to perform a clustering of data.

Methods using pheromone also exist, namely A^2CA [6]. Another approach can be found the work of J. Handl in [7] (an ATTA algorithm), which introduce modified neighborhood function (penalizing high dissimilarities), short-term memory with lookahead (jumping ants), increasing radius of perception, time-dependent modulation of the neighborhood function.

ACO_DTree Method. The ACO_DTree method is a hybrid evolutionary approach for binary decision tree construction [8,9]. The tree is induced using the known data and it can be used for unsupervised clustering later: each leaf of the classification tree can be interpreted as a cluster. The algorithm uses a population of classification trees that is evolved using an evolutionary approach. Creation of the trees is driven by a pheromone matrix, which uses the ACO paradigm. The high number of individuals and the decentralized approach to task coordination in the studied species means that ant colonies show a high degree of parallelism, self-organization and fault tolerance.

This approach has been utilized (with improvements) to simplify the structure of the vast dataset by finding the most important state transitions between literals, producing a probabilistic transitional model. The output structure is presented to the analyst for further processing/iteration.

For clustering, the ACO_DTree method [8,10] and ACO inspired clustering [5] variations have been successfully used.

6 Automated Processing

Automated layout of transition graph is very comfortable for an expert, however the contents of the attribute is so complicated, that a human intervention is inevitable. Examples of automated layout can be seen in Fig. [1] and Fig. [2].

The figure Fig. [1] shows a transitional graph where only positioning based on the word distance from the sentence start is used. Althoug it migh look correct, note that the same words are mispositioned in the horizontal axis.

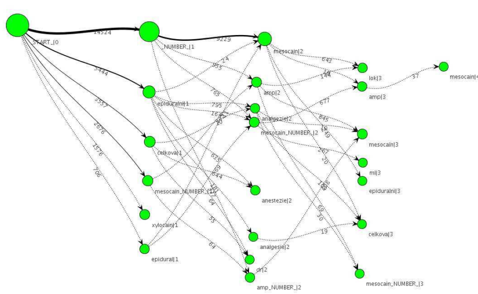

Fig. 1. A fully automated transition graph showing the most important relations in one textual attribute. No clustering has been used. The layout is based on the word distance from the start of the sentence. Note the mis-alignment of the similar/same words. Refer to section [2].

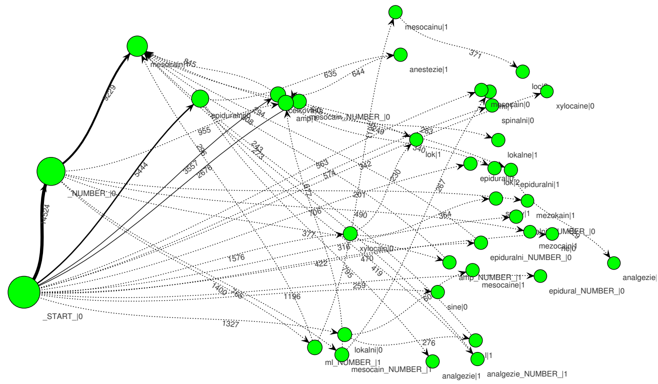

Fig. 2. A fully automated transition graph (sub-graph) showing the most important relations in one textual attribute. The ACO approach has been used to cluster the corresponding vertices. Refer to section [2].

7 Expert Intervention

A human intervention and supervision over the whole project is indiscutable. Therefore also human (expert) visualization of the transition graph has been studied.

The vertices in a human-only organization are (usually) organized depending on the position in the text (distance from the starting point) as the have the highest potence. Number literal (a wildcard) had the highest potence, as many quantitative measures are contained in the data (age, medication amount, etc.). Therefore it has been fixed to the following literal, spreading into the graph via multiple nodes (i.e. a sequence *mesocain 10 mL* become two vertices – *mesocain_NUMBER_* and *mL*). This allowed to organize the chart visualization in more logical manner. Time needed to organize such graph was about 5–10 minutes. The problem is that the transition graph contains loops, therefore the manual organization is not straigthforward.

An aid of a human expert has been used in semi-automated approach (see Fig. [3] where the automated layout has been corrected by the expert. The correction time has been about 20–30 seconds only.

8 Results and Conclusion

The main advantage of the nature inspired concepts lies in automatic finding relevant literals and group of literals that can be adopted by the human analysts and furthermore improved and stated more precisely. The use of induced probabilistic models in such methods increased the speed of loosely structured textual attributes analysis and allowed the human analysts to develop lexical analysis grammar more efficiently in comparison to classical methods. The speedup (from

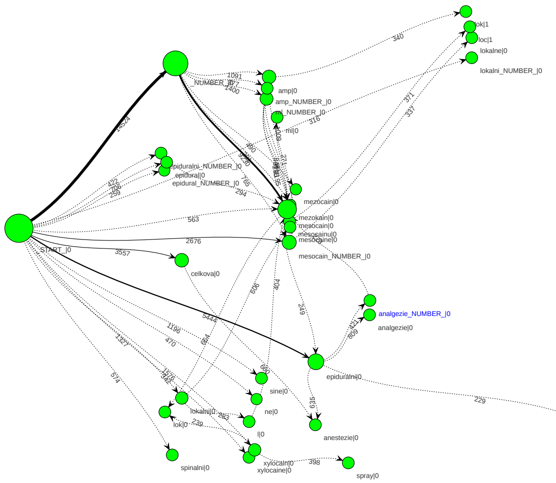

Fig. 3. A semi-automated (corrected by a human expert) organized transition graph showing the most important relations in one textual attribute. Refer to section [2].

about 5–10 minutes to approx 20–30 seconds) allowed to perform more iterations, increasing the yield of information from data that would be further processed in rule discovery process. However, the expert intervention in minor correction is still inevitable. The results of the work are adopted for rule discovery and are designed to be used in expert recommendation system.

9 Discussion and Future Work

The future work is to evaluate the DB analyst's utilization and aid of such graphs in more accurate way. The graphs serve as a bases for extraction rule proposal. However the only relevant measure is the time to reorganize the transitional graphs. The subjective opinion is very expressive and is not coherent. Next, the semantic meaning of the attributes will be extracted and verified followed by rule discovery mining.

Acknowledgements. This research project has been supported by the research programs number MSM 6840770012 "Transdisciplinary Research in the Area of Biomedical Engineering II" of the CTU in Prague, sponsored by the Ministry of Education, Youth and Sports of the Czech Republic and by the project number NT11124-6/2010 "Cardiotocography evaluation by means of artificial intelligence" of the Ministry of Health Care. This work has been developed in the BioDat research group `http://bio.felk.cvut.cz`.

References

1. Freitag, D., McCallum, A.K.: Information extraction with hmms and shrinkage. In: Proceedings of the AAAI Workshop on Machine Learining for Information Extraction (1999)
2. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the ICML, pp. 282–289 (2001); Text processing: interobserver agreement among linquists at 70
3. Dorigo, M., Stutzle, T.: Ant Colony Optimization. MIT Press, Cambridge (2004)
4. Deneubourg, J.L., Goss, S., Franks, N., Sendova-Franks, A., Detrain, C., Chretien, L.: The dynamics of collective sorting robot-like ants and ant-like robots. In: Proceedings of the First International Conference on Simulation of Adaptive Behavior on From Animals to Animats, pp. 356–363. MIT Press, Cambridge (1990)
5. Lumer, E.D., Faieta, B.: Diversity and adaptation in populations of clustering ants. In: From Animals to Animats: Proceedings of the 3th International Conference on the Simulation of Adaptive Behaviour, vol. 3, pp. 501–508 (1994)
6. Vizine, A.L., de Castro, N.L., Hruschka, E.R., Gudwin, R.R.: Towards improving clustering ants: An adaptive ant clustering algorithm. Informatica 29, 143–154 (2005)
7. Handl, J., Knowles, J., Dorigo, M.: Ant-based clustering and topographic mapping. Artificial Life 12(1), 35–61 (2006)
8. Bursa, M., Lhotska, L., Macas, M.: Hybridized swarm metaheuristics for evolutionary random forest generation. In: Proceedings of the 7th International Conference on Hybrid Intelligent Systems 2007 (IEEE CSP), pp. 150–155 (2007)
9. Bursa, M., Lhotska, L.: Ant colony cooperative strategy in biomedical data clustering. SCI. Springer, Heidelberg (2007)
10. Bursa, M., Huptych, M., Lhotska, L.: Ant colony inspired metaheuristics in biological signal processing: Hybrid ant colony and evolutionary approach. In: Biosignals 2008-II, vol. 2, pp. 90–95. INSTICC Press, Setubal (2008)