MOMEMI: Modern Methods of Data Mining

Special Session along with
ICCGI 2017: The Eleventh International Multi-Conference on Computing in the Global Information Technology
November 13 - 17, 2016 - Barcelona, Spain
http://www.iaria.org/conferences2016/ICCGI16.html

Elena V. Ravve
Software Engineering Department
Ort Braude College
Karmiel, Israel
Email: cselena@braude.ac.il

Abstract— **Modern data mining is used in order to classify and to discover relationships in big data sets. The papers, presented in the framework of the MOMEMI, deals with the most important fields of modern data mining: determining and use of patterns and templates, incremental reasoning, geometrical associations as well as text mining.**

Keywords-data mining; classification; forecast; cluster; association; pattern; incremental and distributed reasoning.

I. Introduction

Data mining is the process of extracting nontrivial and potentially useful information, or knowledge, from the big data sets. The main aims of data mining are classification of the data, its interpretation and further forecasting. In fact, every day, more than 2 quintillion bytes of data are created and 90% of the data in the world today was created within the past two years [1]. It means that the traditionally used hands-on data approaches of the data proceeding like standard statistical methods are not more applicable. Automatic data analysis is ultimately needed.

Modern data mining techniques are: association rules, decision trees, Gaussian mixture models, regression algorithms, neural networks, support vector machines, Bayesian networks, etc.

Moreover, the modern data mining machinery is used in order to discover relationships and patterns in the available information. The main types of such relationships are: classes (partition into predefined groups); clusters (partition into logically related groups); associations; behavior patterns and trends.

Data mining is widely used in practice in many civilian applications: financial, retail, communication and marketing companies. Moreover, it has also security and military applications. Among the civilian applications, the most popular are marketing oriented. WalMart is an example of pioneering massive data mining. WalMart allows more than 3,500 suppliers, to access data on their products and perform data analyses. These suppliers use this data to identify customer buying patterns.

More surprisingly, the National Basketball Association (NBA) uses a data mining application in order to analyze the movements of players to help coaches orchestrate plays and strategies.

Text mining and authorship analysis is one of branches of the modern data mining. It is connected to many real world problems appearing in various areas such as plagiarism detection, identification of authors of threats, verification of suicide notes, computer forensics and other. Classically, the authorship identification is the process aiming at determining the author of a text seemed as a closed-set classification task where all possible author candidates are acknowledged. The writing style of the doubtful document is compared with the known ones planning to discover the most similar fashion susceptible matches to the questionable texts.

Writing Style conveys a writer's outline of attendance and is an individual embodiment of the general writing process composed from many fuzzy and attaching phases, which are commonly recognized as Pre-writing, Drafting and Writing, Sharing and Responding, Revising and Editing and Publishing. This attitude naturally leads to writing styles evaluation by means of a comparison of the writing process implementations affected by the authors' distinctive characteristic.

A text mining methodology was applied to analysis of editorial texts published in the Egyptian "Al-Ahraam" newspaper and successes to indicate several important events connected to the "Arab Spring".

Finally, the maintenance and proceeding of big data requires "massively parallel software running on tens, hundreds, or even thousands of servers" [2]. That is why one of the most recent trends in data mining is incremental and distributed analysis and reasoning

II. PRESENTED CONTRIBUTIONS

A. *Template Based Automatic Generation of Runsets*

One of the new fields of data mining applications is determining of patterns and templates for layout verification of the modern electronic devices (ships). Layout of the modern electronic devices consists of billions of polygons for chemical layers. There exist hundreds of design rules, defining how the polygons may be drowning.

Design rule checkers (DRC) guarantee that the chip may be manufactured. Moreover, any manufacturing process allows a finite set of supported legal devices. Layout Versus Schematic (LVS) comparison determines one-to-one equivalency between a circuit schematic and its layout. The correctness of the DRC and LVS runsets is verified using test cases, which contain shapes, representing failing and passing conditions. Creation and maintenance of the complete set of runsets and the corresponding test cases is complicated and time consuming process that should be automatized.

Usually almost all design rules may be divided into a set of categories: width, space/distance, enclosure, extension, coverage, etc. Moreover, the set of legal devices for any process may be divided into a set of technology independent categories: transistors, capacitors, resistors, diodes and so on.

In this paper, these categories are used in order to define reusable patterns. The integrator will use the pre-defined patterns in order to compose the design rule manuscript (DRM) rather than to write it. DRC and LVS runsets are then automatically generated using the DRM. Moreover, the patterns may be used in order to automatically create the corresponding test cases.

B. *Incremental Reasoning on Strongly Distributed Fuzzy Systems*

In this paper, the notion of strongly distributed fuzzy systems and present a uniform approach to incremental problem solving on them. The approach is based on the systematic use of two logical reduction techniques: Feferman-Vaught reductions and syntactically defined translation schemes. The fuzzy systems are presented as logical structures. The problems are presented as fuzzy formulae on them. A uniform template for methods is proposed. It allows (for a certain cost) evaluation of formulae of fuzzy logic over the structure from values of formulae over its components and values of formulae over the index structure.

C. *Modelling behavior patterns in cellular networks*

In this paper, customer behavior in cellular networks is explored. A novel model of the fundamental user profiles is developed. The study is based on investigation of activities of millions of customers of Orange, France. A way of decomposition of the observed distributions according to certain external criteria is proposed. The distribution of customers having the same number of calls during a fixed period is analyzed.

A segmentation of the population is provided by an approximation of the considered distribution by means of a mixture of several more ”basic” distributions presenting the ”granularity” of the user’s activity. In order to examine the meaning of the found approximation, a clustering of the customers is provided using their daily activity, and a new clustering procedure is constructed. The optimal number of clusters turned out to be three.

The approximation is the reduced in the optimal partition to a single-exponential one in one of the clusters and to two double exponential in others. This fact confirms that the proposed partition corresponds to reliable consequential social groups.

D. *Improvement of Identifying Join Candidates in the Cairo Genizah Collection*

The Cairo Genizah is a famous collection containing more than 350,000 different size pieces of ancient handwritten texts. The originals of these texts are stored in different libraries and private collections. International project, which is aimed to digitize the Cairo Genizah, faces an algorithmic challenge of partitioning and classifying of the scanned texts. "Text fragment extensions" is the task of "hinting" for all scanned text images, related to the same handwritten text. Previous methods were based on learning algorithms and probability of each given pair of documents to part of the same original. The proposed idea is rather based on a geometrical approach, which examines geometrical assignment between the paper sheets of the documents.

III. SUMMARY

Data mining is a very rapidly developing branch of computer science. Modern Methods of Data Mining (MOMEMI) special section of The Eleventh International Multi-Conference on Computing in the Global Information Technology (ICCGI, 2016) provides the recent achievements in the most important fields of data mining: determining and use of patterns and templates, incremental reasoning, geometrical associations as well as text mining. We hope that the session will be useful for all the participants of the conferences independently of the particular scope of their expertise.

REFERENCES

[1] IBM. Big data and information management. http://www-01.ibm.com/software/data/bigdata/, 2014

[2] A. Jacobs. The pathologies of Big Data. Commun. ACM, 52(8), 2009, pp.36-44