

Data Warehousing and Application of Data mining in Whether Forecasting

Chavan Hemlata P.¹ Patel Bhumi D.²

^{1,2}Department of Information Technology

Abstract— In that process briefly included what the exact use of data mining & data warehousing. “A one way we can explain the data warehouse is the single, complete and consistent store of data”. “It is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision-making process”. A variety of different sources made available to end users in a what they can understand and use in a business context. “Data mining is the automatic discovery of relationship in typically large databases and, in some instances, the use of discovery results in predicting relationships”. Data mining system allowing the user to interact with system by specifying a data mining query or task providing information. There are many applications of data warehousing and data mining. Just like Marketing, Finance, Health care, Manufacturing, Commerce and trade, planning, Resource planning, Public Administration, Data Mining on the Web, Fraud detection, wheather forecasting. Out of them in our process we can studied weather forecasting. Most importantly, it adapts readily to produce the long-horizon forecasts of relevance in weather derivatives contexts. We produce and evaluate both point and distributional forecasts of average temperature, with some success. We conclude that additional inquiry into nonstructural weather forecasting methods, as relevant for weather derivatives, will likely prove useful.

Key words: Whether Forecasting, Data mining

I. INTRODUCTION

A. Data mining

Data mining, is the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

B. Data Warehouse

Data warehouse system are valuable tools in today's competitive,fast evolving world.In the last several years,many firms have spent millions of dollrs in building enterprise wide data warehouse. BrainTech, LLC differs from typical data mining and demand forecasting companies by considering and trend as small pieces of a larger puzzle. Specifically, we combine time-related, seasonal, environmental, and psychological factors to build models that describe and quantify influences of past behavior. Using these models, we are able to:

A decision support database that is maintained separately from the organization’s operational database Support information processing by providing a solid platform of consolidated, historical data for analysis.

A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision-making process.

In summary, we use multiple sources of data to develop models that identify and quantify market-related and behavioral variables that influence demand. For companies that deal with perishable inventory, we are able to generate predictive models that estimate short-term demand with a high degree of accuracy, allowing our clients to efficiently manage inventory, reducing the number of products that expire unsold, while still ensuring that potential sales are not lost due to inability to meet demand.

C. Data Mining Steps

There are various steps that are involved in mining data as shown in the figure1.

- (1) Data Integration: First of all the data are collected and integrated from all the different sources.
- (2) Data Selection: We may not all the data we have collected in the first step. So in this step we select only those data which we think useful for data mining.
- (3) Data Cleaning: The data we have collected are not clean and may contain errors, missing values, noisy or inconsistent data. So we need to apply different techniques to get rid of such anomalies.
- (4) Data Transformation: The data even after cleaning are not ready for mining as we need to transform them into forms appropriate for mining. The techniques used to accomplish this are smoothing, aggregation, normalization etc.
- (5) Data Mining: Now we are ready to apply data mining techniques on the data to discover the interesting patterns. Techniques like clustering and association analysis are among the many different techniques used for data mining.
- (6) Pattern Evaluation and Knowledge Presentation: This step involves visualization, transformation, removing redundant patterns etc from the patterns we generated.
- (7) Decisions / Use of Discovered Knowledge: This step helps user to make use of the knowledge acquired to take better decision.

II. WEATHER FORECASTING – INTRODUCTION

Weather affects nearly everyone nearly every day

- Weather forecasts are issued:
 - To save lives
 - Reduce property damage
 - Reduce crop damage
 - To let the general public know what to expect

- Forecasts are often utilized to make many Important decisions on a daily basis
- So, how is it done, and how is it done correctly?
 - Weather forecast is a complex multi-disciplinary problem which requires a cascade of different scientific tools, from differential equation solvers to high-dimensional statistical and data-mining algorithms. The demand for high-resolution predictions is continuously increasing due to the multiple applications in hydrology, agronomy, etc., which require regional meteorological inputs. To fill the gap between the coarse-resolution lattices used by global weather models and the regional needs of applications, a number of statistical downscaling techniques have been proposed. In this paper we describe a Web portal which integrates the necessary tools with Grid middleware allowing for distributed data access and computing. The portal is part of the ENSEMBLES EU-funded project and allows end users to interactively downscale weather predictions using a web browser. Both the architecture and the usage of the portal are described in this paper.

III. FORECAST CHARACTERISTICS

- Many sources of weather forecast information
- What are the distinguishing and relevant Features for the user
- Geographical coverage
- Time scales
- Format
- Cost

Weather derivatives are also different from insurance. First, there is no need to file a claim or prove damages. Second, there is little moral hazard, although there is some, as when someone with a long precipitation position attempts to seed the clouds. (Don't laugh – it has happened!) Third, unlike insurance, weather derivatives allow one to hedge against comparatively good weather in other locations, which may be bad for local business (e.g., a bumper crop of California oranges may lower the prices received by Florida growers).

Weather forecasting is crucial to both the demand and supply sides of the weather derivatives market. Consider first the demand side, consisting of obvious players such as energy companies, utilities and insurance companies, and less obvious players such as ski resorts, grain millers, cities facing snow.

Data mining is the process of extracting hidden patterns from data. As more data is gathered, with the amount of data doubling every three years, data mining is becoming an increasingly important tool to transform this data into information. It is commonly used in a wide range of applications, such as marketing, fraud detection and scientific discovery. Data mining can be applied to data sets of any size, and while it can be used to uncover hidden patterns, it cannot uncover patterns which are not already present in the data set.

IV. ARCHITECTURE OF DATA MINING

A. Architecture: Typical Data Mining System

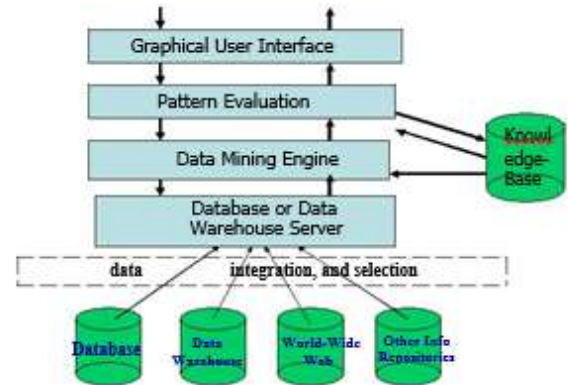


Fig. 1: Architecture of Typical Data Mining System

V. DATA MINING TECHNIQUE

A. Substance:

- Similarity/Clustering
- K-nearest neighbor
- Statistical e.g.”Ratio Score”

B. Text:

- similarity/Clustering
- ”Expert” System
- Statistical, concept heading frequencies

VI. DATA MINING FUNCTIONALITY

- Concept description: Characterization and discrimination Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions
- Association(correlation and causality)
- Multi-dimensional vs. single-dimensional association
- $\text{age}(X, \text{“20..29”}) \wedge \text{income}(X, \text{“20..29K”}) \Rightarrow \text{buys}(X, \text{“PC”})$ [support = 2%, confidence = 60%]
- $\text{contains}(T, \text{“computer”}) \Rightarrow \text{contains}(T, \text{“software”})$ [1%, 75%]
- Classification and Prediction
- Finding models (functions) that describe and distinguish classes or concepts for future prediction
- E.g., classify countries based on climate, or classify cars based on gas mileage
- Presentation: decision-tree, classification rule, neural network
- Prediction: Predict some unknown or missing numerical values
- Cluster analysis
- Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
- Clustering based on the principle: maximizing the intra-class similarity and minimizing the interclass similarity
- Outlier analysis
- Outlier: a data object that does not comply with the general behavior of the data

- It can be considered as noise or exception but is quite useful in fraud detection, rare events analysis
- Trend and evolution analysis
- Trend and deviation: regression analysis
- Sequential pattern mining, periodicity analysis
- Similarity-based analysis
- Other pattern-directed or statistical analyses

VII. KNOWLEDGE DISCOVERY IN DATABASES (KDD)

Knowledge discovery in databases (KDD) is the process of discovering useful knowledge from a collection of data. This widely used data mining technique is a process that includes data preparation and selection, data cleansing, incorporating prior knowledge on data sets and interpreting accurate solutions from the observed results. Major KDD application areas include marketing, fraud detection, telecommunication and manufacturing. Techopedia explains Knowledge Discovery in Databases (KDD) Traditionally, data mining and knowledge discovery was performed manually. As time passed, the amount of data in many systems grew to larger than terabyte size, and could no longer be maintained manually. Moreover, for the successful existence of any business, discovering underlying patterns in data is considered essential. As a result, several software tools were developed to discover hidden data and make assumptions, which formed a part of artificial intelligence. The KDD process has reached its peak in the last 10 years. It now houses many different approaches to discovery, which includes inductive learning, Bayesian statistics, semantic query optimization, knowledge acquisition for expert systems and information theory. The ultimate goal is to extract high-level knowledge from low-level data. KDD includes multidisciplinary activities. This encompasses data storage and access, scaling algorithms to massive data sets and interpreting results. The data cleansing and data access process included in data warehousing facilitate the KDD process. Artificial intelligence also supports KDD by discovering empirical laws from experimentation and observations. The patterns recognized in the data must be valid on new data, and possess some degree of certainty. These patterns are considered new knowledge. Steps involved in the entire KDD process are:

- (1) Identify the goal of the KDD process from the customer's perspective.
- (2) Understand application domains involved and the knowledge that's required
- (3) Select a target data set or subset of data samples on which discovery is to be performed.
- (4) Cleanse and preprocess data by deciding strategies to handle missing fields and alter the data as per the requirements.
- (5) Simplify the data sets by removing unwanted variables. Then, analyze useful features that can be used to represent the data, depending on the goal or task.
- (6) Match KDD goals with data mining methods to suggest hidden patterns.
- (7) Choose data mining algorithms to discover hidden patterns. This process includes deciding which models and parameters might be appropriate for the overall KDD process.

- (8) Search for patterns of interest in a particular representational form, which include classification rules or trees, regression and clustering.
- (9) Interpret essential knowledge from the mined patterns.
- (10) Use the knowledge and incorporate it into another system for further action.
- (11) Document it and make reports for interested parties.

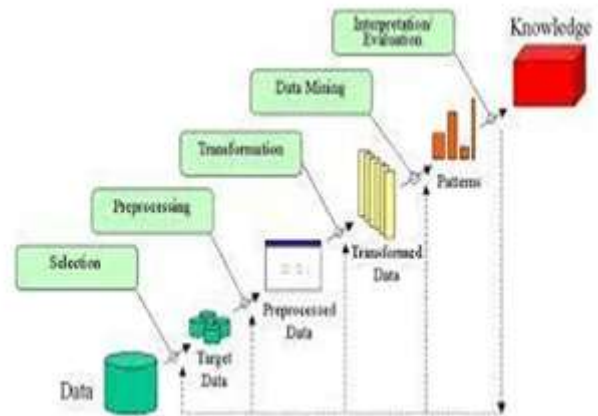


Fig. 2: Typical Architecture Of KDD Process

Knowledge extraction from the World Wide Web has become an important and challenging task as enormous amount of data in form of semi-structured nature is available. Web mining approaches such as web content mining, web structure mining and web usage mining are available which help to extract and produce useful knowledge from the web. Web usage mining is used for applications such as customer shopping sequences, web clicks, biological sequences, creation of dynamic user profiles, etc. In this paper, the discovery of frequent patterns from web log data is being considered. A web server usually registers a log entry for every access of a web page. First, raw web log data needs to be cleaned, condensed and transformed in order to retrieve and analyze significant and useful information. Second, pattern mining can be performed on log records to find association patterns, sequential patterns and trends of web accessing. With the use of such patterns, studies have been conducted on analyzing system performance, improving system design by web caching, web page pre-fetching, understanding the nature of web traffic, etc. The process of web usage mining consists of three major steps (1) data pre-processing, (2) pattern discovery, and (3) pattern analysis stages. Log files are stored on the server side, on the client side and on the proxy servers. In this work, only the server side web log data is being used which is the Click stream data set.

The data pre-processing phase includes the data cleaning, user identification, session identification and data transformation respectively. The pattern discovery phase involves the discovery of frequent sequences. The pattern analysis phase involves the analysis of the frequent patterns generated by the pattern discovery phase. Many different techniques for mining frequent sequential patterns from the log data have been proposed in the recent past. The authors in discuss that the candidate generation-and-test approach outperforms the pattern-growth approach on mining short

patterns, while pattern-growth approach is better on mining long patterns. The authors in discuss the process of web log mining. Sequence mining is accomplished in, where a so-called WAP-tree is used for storing the patterns efficiently. Tree-like topology patterns and frequent path traversals are searched by.

VIII. DATA MINING FOR CUSTOMER MODELING

- Customer Tasks:
- attrition prediction
- targeted marketing:
- cross-sell, customer acquisition
- credit-risk
- fraud detection
- Industries
- banking, telecom, retail sales.

IX. MAJOR DATA MINING METHODS

- Classification: predicting an item class
- Clustering: finding clusters in data
- Associations:e.g. A & B & C occur frequently
- Visualization: to facilitate human discovery
- Summarization: describing a group
- Deviation Detection: finding changes
- Estimation: predicting a continuous value
- Link Analysis: finding relationships

X. DATA MINING APPLICATION AREAS

- Science:
Astronomy, bioinformatics, drug discovery
- Business:
Advertising, CRM (Customer Relationship management), investments, manufacturing, sports/entertainment, telecom, e-Commerce, targeted marketing, health care.
- Web:
Search engines, bots, ...
- Government
Law enforcement, profiling tax cheaters, anti-terror

XI. CONCLUSION

Data mining: Discovering interesting patterns from large amounts of data

- A natural evolution of database technology, in great demand, with wide applications
- A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
- Mining can be performed in a variety of information repositories
- Data mining functionalities: characterization, discrimination, association, classification, clustering etc.
- Data mining systems and architectures.

REFERENCES

[1] Qiankun Zhao, Sourav S. Bhowmick, Association Rule Mining: A Survey, Technical Report, Center for Advanced Information Systems (CAIS),

Nanyang Technological University, Singapore, 2003.

- [2] W.W. Chapman, J.N. Dowling, and M.M. Wagner, "Fever detection from free-text clinical records for biosurveillance.", *Journal of Biomedical Informatics*, Volume 37, Issue 2(April 2004), pp.120-127.
- [3] D.T. Heinze, M.L. Morsch, and J. Holbrook, "Mining Free-Text Medical Reports.", *Proceedings of the AMIA Annual Symposium, 2002*, pp. 254-258.
- [4] B.W. Mamlin, D.T. Heinze, and C.J. McDonald, "Automated Extraction and Normalization of Findings from Cancer-Related Free-Text Radiology Reports.", *Proceedings of the AMIA 2003 Annual Symposium 2003*, pp. 420-424.
- [5] Y. Shahar, O. Young, E. Shalom, A. Mayaffit, R. Moskovitch, A. Hessian, and M. Galperin, "DEGEL: A Hybrid, multiple-ontology framework for specification and retrieval of clinical guidelines." *Proceedings the Ninth Conference on Artificial Intelligence in Medicine Europe (AIME-03)*, Protaras, Cyprus, 2003, pp. 122-131.
- [6] S. Mukherjea, B. Bamba, P. Kankar, "Information Retrieval and Knowledge Discovery Utilizing a BioMedical Patent Semantic Web." *Knowledge and Data Engineering, IEEE Transactions on Volume 17, Issue 8, Aug. 2005*, pp. 1099 – 1110
- [7] T.C. Rindflesch, M. Fiszman, "The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic Pages to Enhance Web Information Retrieval.", "The propositions in biomedical text.", *Journal of Biomedical Informatics* 36(6),2003 , pp. 462-477
- [8] G. Leroy and H. Chen, "Meeting medical terminology needs-the ontology-enhanced Medical Concept Mapper." *IEEE Transactions on Information Technology in Biomedicine* 5(4), 2001, pp. 261-270.
- [9] H. Muller, E.E. Kenny, and P.W. Sternberg, "Textpresso: An ontology-based information retrieval and extraction system for biological literature", *PLoS Biol* 2(11), 2004
- [10] D.P. Corney, B.F. Buxton, W.B. Langdon, and D.T. Jones, "BioRAT: Extracting Biological Information from Full-Length Papers.", *Bioinformatics*, Vol. 20, No. 17, 2004, pp. 3206–3213.
- [11] Hettne K M, Mos M, Bruijn AG., et al. Applied information retrieval and multidiscipl research: new mechanistic hypotheses in complex regional pain syndrome *J Biomed D Collab.* 2007;2:2
- [12] Rebholz-Schuhman D, Cameron G, Clark D, et al. SYMBIOmatics: synergies in Medi Informatics and Bioinformatics–exploring current scientific literature for emerging topic Bioinformatics. 2007;8 Suppl 1:S18

- [13] Cerrito P. Inside text mining. Text mining provides a powerful diagnosis of hospital q rankings Health Manag Technol. 2004;25:28-3.
- [14] Ananiadou S, Kell D B, Tsujii J. Text mining and its potential applications in systems Trends Biotechnol. 2006;24:571-9.
- [15] D. Magdalene Delighta Angeline, I. Samuel Peter James. Association Rule Generation Using Apriori Mend Algorithm for Student's Placement", Int. j. emerg. sci. 2006. 2(1): 78-86
- [16] Roberts P M. Mining literature for systems biology Brief Bioinform. 2006;7:399-406.
- [17] Kareem, S., Bajwa, I.S. A Virtual Telehealth Framework: Applications and Technical Considerations. IEEE International Conference on Emerging Technologies 2011 (ICET 2011) NUST Pakistan
- [18] Kareem, S., Bajwa, I.S. Clinical Decision Support System based Virtual Telemedicine In: 3rd International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC 2011) pp:16-21 Hangzhou, China

