Research Article

Spatial, Temporal and Text Mining - Security, Privacy and Ethical Issues

Dr. Manish Pandey
Professor, Himalayan University, Dehradun

Abstract

The World Wide Web (or www) has impacted on almost each aspect of our lives. It is the main widely known information source that is easily available and searchable. It consists of billions of interconnected Web pages) which are uploaded and written by millions of people. Web has dramatically changed our information seeking behavior and lifestyle. With the World Wide Web information is only a few clicks away from us. We can find needed information on the Web and also we can easily share information and knowledge with others. The WWW is officially an Internet-based computer network which allows user to access information stored on another computer through the world-wide network called the Internet. This world-wide network follows a standard client-server model. In this model, a client relies on a program on one end and connects to a remote machine called the server where the data is stored. Web browsers like Netscape, Internet Explorer, Firefox work by sending requests to remote servers for information retrieval and interpret the returned documents in HTML in the form of text and graphics on the computer screen on the client side. Hypertext link on a Web Page allows author to link their documents to other related documents residing on computers anyplace in the world. To view these documents, one simply follows the hyperlinks.

Keywords*:* Spatial Mining, Temporal Mining, Text Mining, Structured Data, Cluster Analysis

Introduction:

The concept of Web warehousing originated with the initiation of data warehousing. In 1992 W.H. Inmon defined data warehousing as a subject-oriented, integrated, non-volatile and time variant collection of data in an organization. The only difference between data and Web warehousing is that the underlying database in Web warehousing is the entire World Wide Web. As a readily accessible resource, the Web is a very large data warehouse that contains huge volatile and time variant information that is gathered and extracted into valuable information for use in an organization. Using traditional data mining techniques the Web mining is the process of extracting data from the Web and arranging them into an identifiable patterns and relationships. The following are the various data techniques to analyze the data:

- Association: A pattern in the data exists in which one event is connected to another.
- Sequence or path analysis: A pattern exists in the data in which one event leads to another event in a sequence.
- Classification: A totally new pattern exists which develops a completely new structure.
- Clustering: Within the data related facts (not previously known), grouping developed by finding and visually inspecting.

- Forecasting: Predicting future conditions based on existing data patterns or clusters

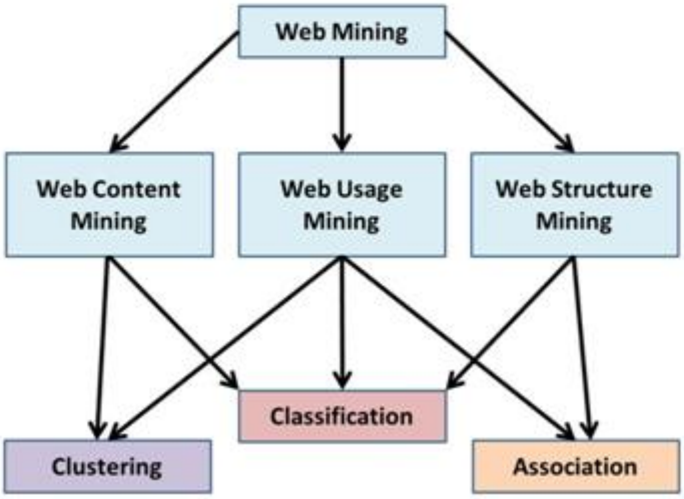

Fig: Web Data Mining

Since data, information, and knowledge are so critical to overall operational success, Web mining, warehousing and knowledge management are logical extensions of present operational activities. To get timely and accurate decisions, an essential element is to obtain the best possible way to produce appropriate and effective courses of action. Web has many unique characteristics to make mining useful. In Web mining, information and knowledge are fascinating and challenging tasks, some of the important characteristics are:

- Hyperlinks exist to connect Web pages within a site and across different sites. Within a website, hyperlinks provide information organization mechanisms. Across different sites, hyperlinks provide conveyance of authority to the target pages.
- The information on the World Wide Web is noisy. A typical Web page contains many pieces of data and information, e.g., the information of the page, links, advertisements, copyright policies, privacy policies, contact information etc. Only part of the information is useful for a particular application, the rest is considered noise for that application. A large amount of information on the Web is of low quality, incorrect, or even misleading.
- The amount of data and information on the Web is huge, time variant and still growing. The exposure of the information is also very miscellaneous. We can find any information about almost anything on the Web.
- The commercial Web sites allow people to carry out useful operations at their sites, e.g., to select, to store, to purchase products, to pay bills, and to fill in forms.
- The Web is dynamic and time variant constantly. Monitoring these change are important issues for many applications.
- The Web is not only about data, information and services, but it is a virtual society to interact among people, organizations and automated systems. We can communicate with people anywhere in the world easily and instantly. We can also express our views on anything in Internet forums, blogs and review sites.

Spatial Mining

Spatial data mining is the process of discovering information and knowledge from interesting and previously unknown patterns from spatial databases. In this analyst uses geographical or spatial information to produce business intelligence or patterns. Challenges involved in spatial data mining include finding objects or identifying patterns that are relevant to data analysis. The extensive usage of spatial databases has led to spatial knowledge discovery.

For example, if you want to get information about every location you have visited in the past week, you have to capture your destination's coordinates and list a number of attributes such as place name, duration of visit, and more. Afterword you have to create a shapefile in Quantum GIS or similar software with this information. Spatial data must have latitude or longitude, some other coordinates denoting a point's location in space. Spatial data can contain any number of attributes pertaining to a place. You may choose any number of attributes you want to describe a place.

Following are the tasks of spatial data mining:

- **Classification:** This determines a set of rules to find the class of the specified object as per its attributes of scientific data, engineering data, astronomical data, multimedia data, genomic and web data.
- **Association rules:** Association rules mining determine rules from the data sets, and it describes patterns that are usually in the Spatial database.
- **Characteristic rules:** Characteristic rules can describe some parts of the data set.
- **Discriminate rules:** Discriminate rules describe the differences between any two kind of data of the database, such as calculating the difference between two cities as per employment rate.

Temporal Mining

Temporal data mining defines the method of extraction of non-trivial, implicit, and potentially important data from large sets of temporal data. Temporal data is a series of primary data types and numerical values. It deals with gathering required knowledge from temporal data. The objective of temporal data mining is to find time variant temporal data patterns, unexpected trends and several hidden relationships in the higher sequential data. Temporal data may contain a sequence of nominal symbols from the alphabet referred to as a temporal sequence. A sequence of continuous time series, real-valued components utilize a set of approaches from machine learning, statistics, and database technologies.

Temporal Data Mining includes processing time series data mining, which composed of three major tasks such as the description of temporal data, representation of similar measures and mining services. Temporal Data Mining consists of sequences of data, which compute values of same attribute at a sequence of multifold time points.

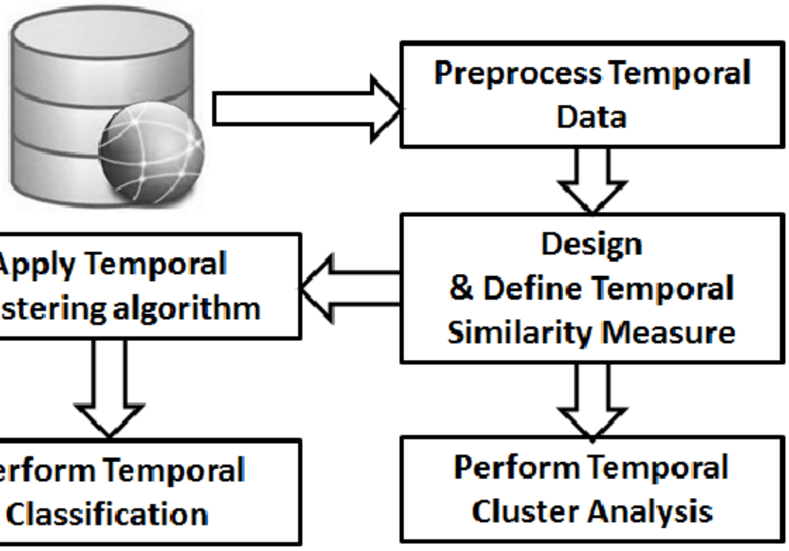
Preprocess Temporal Data
Design & Define Temporal Similarity Measure
Apply Temporal Clustering algorithm
Perform Temporal Classification
Perform Temporal Cluster Analysis

Fig: Temporal Mining

Temporal classification is to predict temporally related fields in a temporal database. The problem is deciding the general value of the temporal variable being predicted to give the different fields. Temporal classification in the training data targets the variable each given observation, and a set of assumptions representing prior knowledge of the problem. Temporal classification technique is associated with the complex problem of density estimation of a temporal data set.

Temporal data mining techniques divided into following classes:

- **Temporal data clustering:** It targets separating the temporal data into subsets as data clusters that are similar to each other. There are two problems of temporal clustering: one is to define a meaningful similar measure, and, other is to choose the number of temporal clusters.
- **Temporal data prediction:** The goal of temporal prediction is to predict patterns from some fields based on other temporal fields. Temporal data predictions also involve prior temporal patterns of models and knowledge to find data attributes relevant to the attribute of interest.
- **Temporal data summarization:** Temporal data summarization is a process to describe a subset of temporal data by representing extracted information in a model in a pattern. Summarization provides a compact description for a temporal dataset which may involve a logical language such as temporal logic or fuzzy logic.
- **Temporal data dependency:** Temporal dependency modeling describes time dependencies among the datasets or in temporal attributes of data. There are two dependency models involve in temporal data dependency, which are qualitative and quantitative. The qualitative dependency model specifies temporal variables such as time gap that are locally dependent on a given space. And the quantitative dependency models specify the value dependencies such as using numerical scale in a statistical space.

Text Mining

Text mining is an automatic process to extract valuable information from unstructured text with the use of natural language processing. Text mining transforms data into information that a machines can understand. Text mining automates the method of classifying texts by response, topic, and intent. Text mining is a process of examining large collection of data in the form of documents to discover new information. Text mining identifies facts and relationships into a structured form that can be analyzed by using clustered HTML tables, mind maps and charts. Text mining is the process of transforming unstructured text into a structured format by identifying meaningful patterns. By applying advanced analytical techniques using Natural Language Processing and other deep learning algorithms, you are able to explore and discover hidden relationships within their unstructured data.

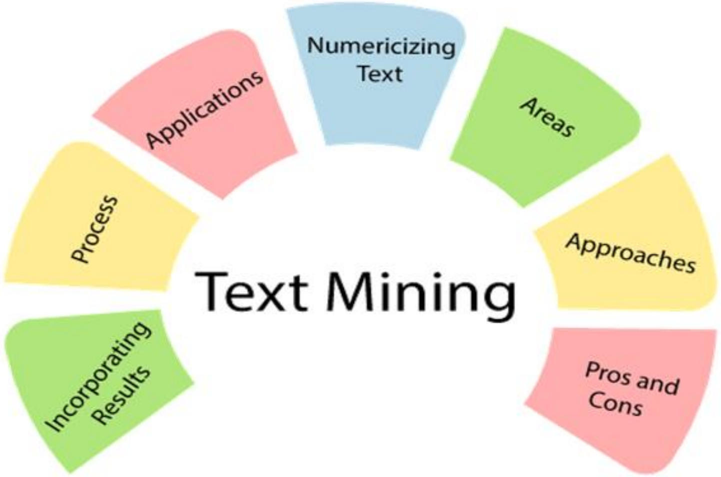

Fig: Text Mining

Natural language processing is a form of computational linguistics that uses methods from various disciplines, such as artificial intelligence, linguistics and data sciences, to understand human language in both written and verbal forms. Common sub-tasks of Natural Language Processing include:

- **Summarization:** Summarization provides a synopsis of long text to create a concise, coherent summary of the main points of documents.
- **Part-of-Speech (PoS) tagging:** PoS assign a tag to every token in a document. This token is based on its part of speech, for example denoting nouns, verbs, adjectives, etc. PoS tagging enables semantic analysis on unstructured text.
- **Text categorization**: This task is also known as text classification that is responsible for analyzing text documents, classification based on predefined topics or categories such as categorizing synonyms and abbreviations.
- **Sentiment analysis**: Sentiment analysis detects positive or negative sentiment from internal or external data sources. Sentiment analysis allows tracking the changes in customer attitudes over the time. This technique is commonly used to provide information about perceptions of brands, products, and services.

Security Issue

The maintenance of financial, demographic, behavioral, monetary, and other value-based information may prompt the breakdown of common freedoms because of less security and individual self-rules. From security point of view, the test is to guarantee that information has supportable control over that information. In web mining it is less predictable to anticipate abuse and mishandling by information controllers, while saving information utility. The web poses the security issues and great challenges for knowledge discovery, based on the following observations:

- The web is too huge: The size of the web is very huge, rapidly increasing and the information available is endless. This shows that the data available on the web is too huge for data warehousing and data mining.
- Complexity of Web pages: The web pages do not have a defined structure. Information available on the web is very complex as compared to traditional database. There are huge amount of documents available in digital library of web. These libraries are not arranged in any particular sorted order to apply the mining algorithms.
- Web is a dynamic information source: The information on the web is time variant and rapidly updated. The data in the web such as news, stock markets, weather, shopping, etc., are regularly updated.
- Diversity of user communities: The user community on the web is expanding very fast. These users have different backgrounds and interests. More than 100 million workstations are connected to the Internet and still rapidly increasing.
- Relevancy of Information: A particular person is generally interested in only small portion of the information available on the web, while the rest of the information of the web is not relevant to the user and may swamp desired results.

Privacy Issue

When information is discovered through web data mining, privacy might be directly violated in the process. In web data mining the information is classified and clustered into different profiles, and then the knowledge is used for decision-making, people may feel violated in their privacy. The discovered information no longer links to individual persons because the data is made anonymous before producing the profile. There is no direct sense of privacy violation because the profiles do not contain the real data. The different kinds of customer web data will then be categorized and clustered to build detailed customer profiles to get behavior pattern. This helps companies to retain current customers by providing more personalized services as well as also contributes to search for potential customers. Web data mining is the basis for decision-making and formulating policies, but if profiles somehow become public, the individuality of people is threatened. This is harmful when profiles contain data of a sensitive nature. In web based clustering, people will be judged and treated as group members rather than individuals. People could be discriminated against and being labeled as a member of a group not as an individual with certain characteristics. Some criteria, like race and religion will be inappropriate in decision making.

Web mining is used to automatically discover and extract information from web documents and services. Web mining can be used in a business context and applied to some type of personal data to help companies to build the customer profile, and gain marketing intelligence to grow their business. Web mining is a threat to some important ethical values like privacy and individuality. Web structure mining is a cause for concern when data published on the web in a certain context and combined with other data for use in a totally different context. Problem arises in privacy when users are traced, and their actions are analyzed without their knowledge.

Data mining can be defined as the process of mining for hidden, previously unknown, and potentially useful information from huge databases, by efficient knowledge discovery algorithms. In web data mining new security concerns and problems are identified time to time. Finally rough set theory is discussed and some potential applications to security problems are designed and illustrated. Web mining is a technique, method or algorithms used to extract information and knowledge from the data originating from the web. This technique aims to analyze the behavior of users in order to improve the structure and content of visited web sites. There are a series of processing methodologies that operate at least from the point of view of the user's privacy. The use of powerful processing tools provided by web mining may threaten user's privacy. It has been observed that professionals, with the purpose of formulating practical rules on this matter, they have very narrow-minded concept of privacy. The aim of this chapter is to adopt an integrative approach based on the distinctive attributes in web mining in order to determine which techniques are harmful for the user's privacy.

5.1 Ethical Issue.

Web mining is a cause of concern when data on the web in a certain context is mined and combined with other data in a totally different context for the purpose of knowledge discovery. Web mining raises privacy concerns when web users are traced. Web mining is often used to create customer profiles with a strong tendency of judging and treating people on the basis of characteristics of a cluster instead of on their own individual characteristics and merits. The values of privacy and individuality must be protected and respected to make sure that people are judged and treated fairly. People must be aware of the seriousness of the dangers of their data to be used for web mining and continuously discuss these ethical issues. This must be a joint responsibility of web miners (both adopters and developers), web users, and the government.

A collective approach could be done by society, governments, business owners and developers to prevent web-data mining from causing any harm. Looking at the potential misuse of web content and personal data, efforts can be done to limit the dangers. Legal measures could provide a baseline level for improved ways to handle the problems. The monitoring of ethical matters on web mining should be done by an impartial organization, and it will only work if businesses co-operate and allow giving access into their web-data mining activities. There are several tools that can help to protect a web user's privacy while surfing the web. A privacy enhancing tool could enable the web user to make informed choices. Openness in web-data mining activities is required. Consumers have to be explicitly informed that their data will be used in data mining techniques for the business purposes, and that their data is currently being mined in ways, that they probably have not explicitly authorized. Therefore, consumers should be given three choices having:

- their data mined at all,
- their data mined to some extent (only within the company)
- their data mined without limits (for instance, shared to the third parties also).

Conclusion:

The WWW is officially an Internet-based computer network which allows user to access information stored on another computer through the world-wide network called the Internet. This world-wide network follows a standard client-server model. The concept of Web warehousing originated with the initiation of data warehousing. In 1992 W.H. Inmon defined data warehousing as a subject-oriented, integrated, non-volatile and time variant collection of data in an organization. The only difference between data and Web warehousing is that the underlying database in Web warehousing is the entire World Wide Web. Web content mining is used to extract the data and information from the web content. Every HTML web page provides information that concerns not only the layout but also logical structure.

Spatial data mining is a process that determines some exciting and hypothetically valuable patterns from spatial databases. Spatial data mining is used to analyze scientific data, engineering data, astronomical data, multimedia data, genomic and web data.

Text mining creates structured data that can be integrated into databases, data warehouses or business intelligence dashboards. This dashboard is used for descriptive, prescriptive or predictive analysis.

The basic goal of temporal classification is to predict temporally related fields in a temporal database based on other fields. The problem in general is cast as determining the most likely value of the temporal variable being predicted given the other fields, the training data in which the target variable is given for each observation, and a set of assumptions representing one's prior knowledge of the problem. Temporal classification techniques are also related to the difficult problem of density estimation.

References:

1. *Manish Pandey,* Parul Sharma, Rakesh Kumar, "Analysis of Medical Data using Linear Regression Technique: Numerical and Graphical Application", JOURNAL OF CRITICAL REVIEWS ISSN- 2394-5125 VOL 07, ISSUE 01, 2020, Page No. – 1892-1902.
2. Manish Pandey, "Secondary Analysis of Clinical Data using Linear Regression and Time Series Technique", Suraj Punj Journal For Multidisciplinary Research, ISSN NO: 2394-2886, Volume 8, Issue 12, 2018, Page no. – 741-750.
3. D. Brillinger, editor. Time Series: Data Analysis and Theory. Holt, Rinehart and Winston, New York, 1975.
4. P. Cheeseman and J. Stutz. Bayesian classification (AUTOCLASS): Theory and results. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, Advances in Knowledge Discovery and Data Mining. AAAI Press / MIT Press, 1995.
5. T. Fulton, S. Salzberg, S. Kasif, and D. Waltz. Local induction of decision trees: Towards interactive data mining. In Simoudis et al. [21], page 14.

6. B. R. Gaines and P. Compton. Induction of metaknowledge about knowledge discovery. IEEE Trans. On Knowledge And Data Engineering, 5:990–992, 1993.
7. B. Padmanabhan and A. Tuzhilin. Pattern discovery in temporal databases: A temporal logic approach. In Simoudis et al. [21], page 351.
8. P.sprites, C.Glymour, and R.Scheines. Causation, Prediction and Search. Springer-Verlag, 1993.
9. J. Roddick and M. Spiliopoulou. A survey of temporal knowledge discovery paradigms and methods. IEEE Transactions on Knowledge and Data Engineering, 2002.
10. R.O.Duda and P. Hart. Pattern classification and scene analysis. John Wiley and Sons, 1973.
11. E. Simoudis, J. W. Han, and U. Fayyad, editors. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press, 1996.
12. P. Smyth. Clustering using monte carlo crossvalidation. In Simoudis et al., page 126.
13. T.Oates. Identifying distinctive subsequences in multivariate time series by clustering. In 5th International Conference on Knowledge Discovery Data Mining, pages 322–326, 1999.
14. J. D. Ullman and C. Zaniolo. Deductive databases: achievements and future directions. SIGMOD Record (ACM Special Interest Group on Management of Data), 19(4):75–82, Dec. 1990.
15. D. Urpani, X. Wu, and J. Sykes. RITIO - rule induction two in one. In Simoudis et al. [21], page 339.
16. P. Usama Fayyad and O.L.Mangasarian. Data mining: Overview and optimization opportunities. INFORMS, Special issue on Data Mining, 1998.
17. *Manish Pandey*, “Fuzzy Logic and a Time Series Approach for the Prediction of Atmospheric Temperature”, International Journal of Advanced in Management, Technology and Engineering Sciences, Volume VII, Issue XI, November 2017 ISSN No: 2249-7455, Page No. 526-532.