Research Journal of Applied Sciences, Engineering and Technology 7(22): 4749-4753, 2014
DOI:10.19026/rjaset.7.861
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: January 16, 2014 Accepted: February 15, 2014 Published: June 10, 2014

Research Article

Data Mining-an Evolutionary Arena

[1]Priya Govindarajan and [2]K.S. Ravichandran
[1]Sastra University, Kumbakonam, India
[2]Sastra University, Thanjavur, India

Abstract: This study presents a survey of information retrieval and its various methodologies. In today's escalating world, tracking of information should be done with ease. Keeping that as a constraint, most of the qualms can be deciphered with the aid of Machine Learning (ML). ML can be envisioned as a tool, which identifies and disseminates all information through computerized systems, which can be integrated in the respective domains, in order to get a better and more proficient retrieval of content. This study summarizes the well-known methods used in feature extraction and for classification of text. ML can be portrayed as a major tracker for surveillance, with the aid of some trained ML algorithms. In order to strengthen the response policies for any queries, which is being surrounded with two main issues like policy matching and policy administration, can be prevailed over Joint Threshold Administration Model, JTAM (i.e., Principle of separation of duty). This study gives an overall review about tracking of information with respective to semantic as well as syntactic perspective. It revolves around some of the application as well as administrative mechanism involved in Information Retrieval for mining the data. Data mining techniques in various arenas has been explored; this survey explores the various techniques and evolution of mining in detail.

Keywords: Data mining feature extraction, information retrieval, machine learning, syntactic and semantic perspective

INTRODUCTION

This hectic life has made people to think and to be more cautious about health than ever. Which led to an ambiguous situation that exist between disease and treatment and there is a need of some recording system, for accessing the information and to gain potential benefits out of it. The Machine Learning (ML) field has achieved its force in almost any domain of research from medical to surveillance. ML is visualized as a tool by which healthcare information can be integrated through computerized systems in order to get a superior, more proficient medical care. ML-based methodology is directed to identify and disseminate information (Oana *et al*., 2011) based on semantic as well as syntactic perspective.

ML techniques are used for tracking words (that can be substituted) for the purpose of surveillance (Fong *et al*., 2008) such words has to be further scrutinized, using prevailing mining algorithms. Beyond nouns, the verbs have to be tested and substituted, while exploring group of words use parsers to locate key relations between the words based on which the corresponding substitutions are brought about (for example, "The marriage is going to take place" instead of "the bomb is going to rupture") and by incorporating variety of test sets, these are considered to be some of the specific issues in the areas of research.

The intrusion for a relational database faces two major issues such as policy matching and policy administration. A matching for an anomalous request, with perspective to policies is carried off with algorithms associated with policy matching difficulties (Kamra and Bertino, 2011). The alteration of malicious policy objects from valid users can be thwarted, with perspective to administration response policies. Any alteration made to the data object will be invalid, until it is being jointly approved by DBA (Database Administrator); this incorporates the mechanism of Joint Threshold Administration Model (JTAM).

Detection of personal name aliases without any ambiguity can be carried off with the aid of lexical pattern frequency, co-occurrence analysis and with the threshold frequency on the web. Ranking support vector machine, this has been multiplexed with the source of different ranking scores which is integrated into a stout and effective system (Bollegala *et al*., 2011). The system can be scrutinized with different sorts of data sets. Pattern frequency and matching are considered as the hub in the Machine Learning realm.

From the area of research (in any domain) till surveillance, where information can be retrieved, pattern can be matched for processing and for shielding.

Corresponding Author: Priya Govindarajan, Sastra University, Kumbakonam, India

Finally the collected data can be conceded with the assist of Data Mining.

Finally this retrieved Data can also be mined with the aid of the software, where users are allowed to scrutinize hefty databases to unravel all classes of problems. Tools for navigating data, envisage future behaviors and trends, allowing the end-users to make knowledge-driven, proactive decisions and can response all issues which tends to be time-consuming. Retrieval of information for mining doesn't come with an apt solution, it is just considered as a budding system with an emerging technology.

Data mining may serve as a guide to inherent and to disclose trends and tendencies for accessing the information. This accessed information is used for statistical groupings, predictions and classifications of data. For which, consequently tracking of data is crucial, which are foremost carried off by the retrieval process available within this domain (Data Mining).

METHODOLOGY

Rule based approach: Co-occurrences analysis is based on Lexical knowledge, where the precision is considered to be Low, as Substantiation to the above analysis can be, the work called Service Retrieval of high-precision. Another is the Rule-based approach, used for relation extraction tasks. It stands on either semantic (OR) syntactic information-which contain Trigger words in the form of fixed patterns, which requires human-experts effort and this is considered to be one of the major snags within this approach, but it fetches good precision (Oana *et al*., 2011).

Table 1, contains description about data set, by Rosario and Hearst (2004)-the table is based on Co-occurrence analysis as well as Rule-based approach, which predicts the training as well as testing data sets in the brackets.

Some researchers combine both Syntactic (for better flexibility) and Semantic rules (for good Precision), to fetch the best result. Statistical Methods are used for relation extraction and the most used technique is Bag-Of-Words. All these measures are for identifying any kind of relations in short texts.

For allocating and tracing corresponding aliases of a specified person name is helpful in a variety of web related tasks such as sentiment analysis, information retrieval, relation extraction and personal name disambiguation. Information regarding aliases are dragged using lexical patterns (Bollegala *et al*., 2011). The patterns which are extracted are helpful in tracing candidate aliases for a specified name. Lexical knowledge and co-occurrence analysis plays a vital role in extracting, tracing and allocating the perfect match of the corresponding aliases name in perspective to the corresponding queries.

Sentence oddity-A tracking measure: Sentence oddity (EO), is a methodology where tracked sentences are queried for a catalog of words. Sentence oddity can be calculated by dividing bag of words and its frequency, removing target word with the entire bag of words and its frequency. Fong *et al*. (2008) ESO, Enhanced Sentence Oddity is the extended version of EO:

$$EO = \frac{\text{Bag of words and its frequency,removing target word}}{\text{Entire bag of words and its frequency}}$$

In the calculation of EO, while encountering some sentences for numerator may also being counted for denominator as well; this issue is being eliminated in the extended version of EO, i.e., ESO. ESO is calculated by dividing bag of words and its frequency, excluding target word with the entire bag of words and its frequency. Through these methods words can be probed on a watch list in an effortless mode:

$$ESO = \frac{\text{Bag of words and its frequency,excluding target word}}{\text{Entire bag of words and its frequency}}$$

Table 1: Data set by Rosario and Hearst (2004)

Relationship	Definition and example
Cure	Treat cures DIS
810 (648, 162)	Intravenous immune globulin for recurrent spontaneous abortion
Only DIS	Treat not mentioned
616 (492, 124)	Social ties and susceptibility to the common cold
Only treat	DIS not mentioned
166 (132, 34)	Flucticasome propionate is safe in recommended doses
Prevent	Treat prevents the DIS
63 (50, 13)	Statins for prevention of stroke
Vague	Very unclear relationship
36 (28, 8)	Phenylbutazone and leukemia
Side effect	DIS is a result of treat
29 (24, 5)	Malignant mesodermal mixed tumor of the uterus following irradiation
No cure	Treat does not cure DIS
4 (3, 1)	Evidence for double resistance to permethrin and malathion in head lice
Total relevant: 1724 (1377, 347)	
Irrelevant	Treat and DIS not present
1771 (1416, 355)	Patient were followed up for 6 months
Total: 3495 (2793, 702)	

Table 2: Calculation of frequency (using EO, ESO)

Bag of words	Frequency
The expected event may occur today	f = 1.05 M
The expected campaign may occur today	f = 1.43 M
The expected attack may occur today	f = 2.38 M
The expected operation may occur today	f = 1.52 M

Using the above measures (EO, ESO) a sentences has been dragged from the source and the frequency is calculated for its due course, Table 2.

Knowledge base: The pedestal information must be accurate, to build a Knowledge Base (KB) without fake information. The terms which are grasped using the corresponding functions, must not depend on a patterns. For better precision, the method can make use of Word Net's glossary or else the word net type axioms can be utilized by that method (Hwang *et al*., 2011). A fit KB helps in tracking words, with a related match and promptly word and its meanings can be searched for further usage.

These methods are used to generate glossary noun list, through which concept list is created which is conceded to build integrated glossary noun list for manipulation with word sense. Most of the relation includes transitive property, expansion of which is possible by including transitive axioms, while expressing chain-like facts.

Joint Threshold Administration Model (JTAM): To maintain the records in the database in a faultless way and to prevent malevolent changes from authenticated users, a policy has been put forth called Joint Threshold Administration Model (JTAM). The mechanism is that, a policy operation's validity depends on the authorization of at least k DBAs (Kamra and Bertino, 2011). There is no changes to the present mechanism of access control and internal threats can be averted using man-power, which are considered as a major enhancement for this model. After the policy is created, it is being scrutinized by at least k-1 administrators and then the policy's state is altered as activated:

Authorized Response Policy (Policy ID) Create

The security parameters are scrutinized, once and all to use the response policies, which are being done by public-private keys, secret keys and with various algorithms.

Machine learning technique: Onboard medical systems predicts/prevents health related problems in a timely manner using Machine Learning methodologies, which lessens the risk of direct communication between the crew and the ground support medical specialists for a space mission (Wang *et al*., 2013). With the aid of Support Vector Machine (SVM) classifier and ML techniques, the death certificates of the patients due to cancer have been classified (Butt *et al*., 2013).

Information retrieval for surveillance: Wireless Sensor Networks (WSN's) is used to detect the presence of intruders near the line of control (Bellazreg *et al*., 2013). Intrusion detection and firing-unit, which are tend to be computerized beneath the authoritative personnel for detection (Vittal *et al*., 2010) Suicide Bombing (SB) forecaster, which predicts the patterns of bombing in sensitive places with warnings (Usmani *et al*., 2012), for any kind of surveillance database and its manipulations projects a vivacious role.

"Attack", is a word as well as it's a major issue which is increasing day by day and year by year. Cyber terrorism and hackers are acting as the major area of concern, through which major information are being communicated as well as tracked and already the previous year has been the 'year of the hack' due to a kind of internet blackout as well as the recent cyber attack on DRDO.. In order to avert this many advanced security monitoring tools are being in cooperated with firewalls, gateways and antivirus and Security Information and Event-Management [SIEM] system. Beyond India's cyber law, India is in a position, to implement a huge knowledge management system which can aid the defense force along with DRDO, NTRO. Internally cyber attack preparedness can be increased with the assist of knowledge management system.

One of the main operational preventing mechanism and which loads many records as its rows and columns is, RASSI (Road Accident Sampling System-India) It records on-site crash investigations and the respective reasons for the crash. It has a team of trained automotive engineers and injury coding experts who work in collaboration with the corresponding state officials. This not only improves the safety of highway but also projects the area of concern (vehicle modeling) for many automobile industries.

Many objectives along with many different methodologies for some of the structured, semi structured as well as un-structured issues have been sorted out. The solution along with its most achievable precision has been dragged in most of the domain of research.

RESULT ANALYSIS

Some of the methods for mining the data, in various domains, have been summarized. Machine learning a major methodology used for identifying diseases and its treatment in short text (Oana *et al*., 2011); it also predicts the health problems in space mission (Wang *et al*., 2013). This technique is inducted for mining business data through Rule Induction (RI), Neural Network (NNs), Case-Based Reasoning (CBR). Interactive timeline can be generated through classical hypothesis testing (Swan and Allan, 2000). Granular Neural Networks (GNN) are used for predicting credit

card detection (Syed *et al*., 2002) by training data sets, fuzzy rules and learning algorithms. Without the need of priori technical knowledge, data mining algorithms (Zaidi *et al*., 2004) are used for dragging knowledge in a distributed environment.

Oddity measure and K-gram techniques Fong *et al*. (2008) are integrated for searching words which can be substituted on a watch list. To avoid disambiguation in a semantic relation and to enrich the network (Hwang *et al*., 2011), rule based approach is used. Joint Threshold Administration Model (JTAM) prevents intrusion in relational databases (Kamra and Bertino, 2011). A data mining technique called Marketing Support System (MSS) (Zhang *et al*., 2008) is integrated to improve customer relationship management. Lexical and co-occurrence analysis (Bollegala *et al*., 2011) are used for accurate identification of aliases (name) for an individual. One can extract knowledge or automatic pattern through large database for supporting crucial decisions (Cheng *et al*., 2006) either through data-driven or through knowledge-driven. To explore the relationship between drug effects and genome, Pharmacogenomics, a database (Regnstrom and Burgess, 2005) was created which can be accessed by data mining techniques. SVM (Support Vector Machine) and machine learning techniques classifies cancer-related death certificates (Butt *et al*., 2013). Finally in a database, the impact of data being missed during the mining process can be fetched by concatenating multiple databases.

CONCLUSION

Mining information for different perspectives and for various research purposes are being highlighted with its functionality along with its area of improvement. The same principle can also be implemented with Big Data, where vast amount of data has to be processed, captured and analyzed. Big data is also considered to be an efficient game changer, since it provides big opportunities, which catalyzes the revenues indirectly. To define 3Vs-Volume, Variety, Velocity while accessing big data, secure methods along with traditional data mining tools can be in co operated for fetching better precision value while tracing or searching the data. From health-care till surveillance, the entire SPACE is being tracked, recorded for assessment and for future reference, Hence the name Evolutionary Arena.

REFERENCES

Bellazreg, R., N. Boudriga, K. Trimeche and S. An, 2013. Border surveillance: A dynamic deployment scheme for WSN-based solutions. Proceeding of 6th Joint IFIP Wireless and Mobile Networking Conference (WMNC, 2013), pp: 1-8.

Bollegala, D., Y. Matsuo and M. Ishizuka, 2011. Automatic discovery of personal name aliases from the web. IEEE T. Knowl. Data En., 23(6): 831-844.

Butt, L., G. Zuccon, A. Nguyen, A. Bergheim and N. Grayson, 2013. Classification of cancer-related death certificates using machine learning. Australas. Med. J., 6(5): 292-300.

Cheng, T.H., C.P. Wei and V.S. Tseng, 2006. Feature selection for medical data mining: Comparisons of expert judgment and automatic approaches. Proceeding of 19th IEEE Symposium on Computer-Based Medical System (CBMC, 2006), pp: 165-170.

Fong, S.W., D. Roussinov and D.B. Skillicorn, 2008. Detecting word substitutions in text. IEEE T. Knowl. Data En., 20(8): 1067-1076.

Hwang, M., C. Choi and P. Kim, 2011. Automatic enrichment of semantic relation network and its application to word sense disambiguation. IEEE T. Knowl. Data En., 23(6): 845-858.

Kamra, A. and E. Bertino, 2011. Design and implementation of an intrusion response system for relational databases. IEEE T. Knowl. Data En., 23(6): 875-888.

Oana, F., D. Inkpen and T. Tran, 2011. Learning approach a machine for identifying disease-treatment relations in short texts. IEEE T. Knowl. Data En., 23(6): 801-814.

Regnstrom, K. and D.J. Burgess, 2005. Pharmacogenomics and its potential impact on drug and formulation development. Crit. Rev. Ther. Drug, 22(5): 465-492.

Rosario, B. and M.A. Hearst, 2004. Classifying semantic relations in bioscience text. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, pp: 430.

Swan, R. and J. Allan, 2000. Automatic generation of overview timelines. Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR, 2000), pp: 49-56.

Syed, M., Y.Q. Zhang and Y. Pan, 2002. Parallel granular networks for credit card fraud detection. Proceeding of IEEE International Conference on Fuzzy System, 1: 572-577.

Usmani, Z.U.H., S. Irum, S. Qadeer and T. Qureshi, 2012. Suicide bombing forecaster-novel techniques to predict patterns of suicide bombing in Pakistan. Simul. Series, 44(17): 36-43.

Vittal, K.P., P.P. Ajay, S.B. Ajay and C.H.S. Rao, 2010. Computer controlled intrusion-detector and automatic firing-unit for border security. Proceeding of 2nd International Conference on Computer and Network Technology (ICCNT, 2010), pp: 289-293.

Wang, N., M.R. Lyu and C. Yang, 2013. A Machine learning framework for space medicine predictive diagnostics with physiological signals. Proceeding of the IEEE Aerospace Conference. Big Sky, MT, USA, pp: 1-12.

Zaidi, S.Z.H., S.S.R. Abidi, S. Manikam and C. Yu-N, 2004. ADMI: A multi-agent architecture to autonomously generate data mining services. Proceeding of the 2nd International IEEE Conference on Intelligent Systems, pp: 273-279.

Zhang, X.H., X.C. Yang, W.H. Shi and T.J. Lu, 2008. Data mining-based marketing support system for telecom operators. Proceeding of the 4th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM '08), pp: 1-6.