ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2015

Real Time Data Detecting Trend Process and Predictions using Living Analytics

Dr. G. Murugan

Professor and Research Analyst, Velsoft Technologies , Chennai, India

ABSTRACT: Real time system is a highly interdisciplinary research area, bringing together research insights from the fields of data mining, natural language processing, machine learning, and information retrieval. The amount of textual data available is too huge to be managed manually. An automatic system is needed to analyze and interpret the text. Some of the systems are semi automatic requiring user input to begin processing others are fully automatic producing output from the input corpus without guidance. The review literatures on trend detection indicates that much progress has been made toward automating the process of detecting emerging trends but there is room for improvement. In this work, we proposedetecting trend processand analytic predictions using living analytics to detect emerging trends from live data to cater the needs of various users irrespective of their domain. The system needs to serve as general purpose software that will help the users to identifyand visualize current happenings pertaining to any domain in an efficient and user friendly way. The paper also aims at forecasting the future of the trends obtained in helping the users to look forward and make quick decisions.

Keywords: mining, detecting trend process, analytic predictions.

I. INTRODUCTION

Text mining is the process of exploratory text analysis either by automatic or semi-automatic means that helps finding previously unknown information. Text mining is challenging mainly due to the characteristics of text. Text is not well structured, and text data could be noisy. It has high dimensionality. There is dependency in the text, that is, relevant information is a complex conjunction of words/phrases. Moreover, the text being analyzed will have word and semantic ambiguity.

Knowledge of emerging trends is particularly important to individuals and companies who are charged with monitoring a particular field or business. Figure 1 shows the major phases involved in text mining and each phase is being discussed below. Raw data which is unstructured in nature and of varied form is inputted to the process and undergoes the following phases to obtain patterns that are meaningful and useful.

Analytic predictions encompass a variety of techniques from statistics and data mining that process current and historical data in order to make "predictions" about future events. Such predictions rarely take the form of absolute statements, and are more likely to be expressed as values that correspond to the odds of a particular event or
behavior taking place in the future.

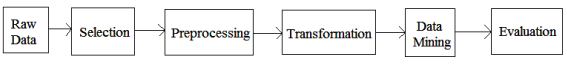

Figure 1 Phases of Mining

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2015

Analytic prediction is widely used in making customer decisions. One of the most well-known applications is credit scoring, which is used throughout financial services.Scoring models process a customer"s credit history, loan application,customer data, etc., in order to rank-order individuals by their likelihood of making future credit payments on time. Predictive analytics are also used in insurance, telecommunications, retail, travel, healthcare, pharmaceuticals and other fields.

A. Existing Systems that Detect Trends

An emerging trend is a topic area that is growing in interest and utility over time. A detailed description of several semi automatic and fully automatic emerging trend detection systems used for research purposes or educational purposes is as follows.

Technology Opportunities Analysis (TOA) - TOA is a semi automatic trend detection system for technology opportunities analysis. Input to the system would be abstracts from technical database such as INSPEC, COMPENDEX, US patents. Potential keywords from the abstracts are extracted manually by domain experts. These keywords are then combined into queries using appropriate Boolean operators to generate comprehensive and accurate searches. Technology Opportunities Analysis Knowbot (TOAK) extracts the relevant documents abstracts and provides analysis of the data by using information such as word counts, date information, word co-occurrence information, citation information and publication information to track activity in a subject area.

Constructive Collaborative Inquiry based Multimedia ELearning (CIMEL) - CIMEL is a multimedia framework for constructive and collaborative inquiry based learning. It is a multimedia tutorial that guides students through the process of emerging trend detection. Through the detection of incipient emerging trends students see the role that current topics play in course related research areas. The methodology relies on web resources to identify candidate emerging trends. Classroom knowledge along withautomated assistants helps students to evaluate the identified candidate trends.

TIME MINES - This takes free text data with explicit date tags and develops an overview timeline of statistically significant topics covered by the corpus. Time Mines relies on Information Extraction (IE) and Natural Language Processing (NLP) techniques to gather the data. This begins processing with a default model that assumes the distribution of a feature depends only on a base rate of occurren ce that does not vary with time. Each feature in a document is compared to the default model. The reduced set of features that is developed using the first round of hypothesis testing is then input into a second processing phase which groups related features together. The grouping again relies on probabilistic techniques that combine terms that tend to appear in the same timeframes into a single topic. Finally a threshold is used to determine which topics are most important and these are displayed via the timeline interface. The threshold is set manually and is determined empirically.

THEME RIVER - This is yet another trend detection tool that summarizes the main topics in a corpus and presents a summary of the importance of each topic via a graphical user interface. This is made up of multiple streams, stream represents a topic and topic represents color and maintains its place in the river relative to other topics.

PATENT MINER – This was developed to discover trends in patent data using a dynamically generated SQL query based upon selection criteria input by the user. The system is connected to an IBM DB2 database containing all granted United States patents. There are two major components to the system phrase identification using sequential pattern mining and trend detection using shape queries.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2015

Several semi automatic and fully automatic ETD systems providing detailed information relating to linguistic and statistical features training and test set generation learning algorithms has been discussed above. It indicates that much progress has been made towards automating the process of detecting emerging trends but there remains room for improvement. All of the systems rely on human domain expertise to separate emerging trends from noise in the system. In addition few systems whether research or commercial in nature have employed formal evaluation metrics and methodologies to determine effectiveness.

B. Existing Systems for Predictive Analytics

Predictive Analytics is a branch of business intelligence category, uses data mining and statistics to make predictions on future happenings. The predictions tell you what are the odds that a certain event will be taking place or not, under what circumstances, or following trends. Description of few open source Predictive Analytics is discussed below.

RAPID MINER - This is an environment for machine learning and data mining experiments. It allows experiments to be made up of a large number of arbitrarily nestable operators, described in XML files which are created with Rapid Miner's graphical user interface. This is used for both research and real-world data mining tasks. This provides a GUI to design an analytical pipeline. This file is then read by Rapid Miner to run the analyses automatically.

Waikato Environment for Knowledge Analysis (WEKA) - is a popular suite of machine learning software written in Java, developed at the University of Waikato. This workbench contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality.

II. SYSTEM DESIGN

The design phase begins with the requirement specification document for the system to be made available. While the requirement activity is entirely a problem domain, design is the first step in moving from the problem domain towards the solution domain.

A. Input Sources

The system under development is not domain specific and requires data that is both historic as well as up to date i.e. Live Data. Owing to the above constraints data cannot be stored in Databases for use. Instead input data has to flow into the system from the enormous digital resources that are readily available from sources such as WWW.

WWW - The World Wide Web is a very large distributed digital information space. From its origins in 1991 as an organization -wide collaborative environment at CERN for sharing research documents in nuclear physics, the Web has grown to encompass diverse information resources: personal home pages, online digital libraries, virtual museums, product and service catalogs, government information for public dissemination, research publications, FTP, Usenet news, and mail servers. The ability to search and retrieve information from the Web efficiently and effectively is an enabling technology for realizing its full potential.

News Wires - Major news agencies generally prepare hard news stories and feature articles that can be used by other news organizations with little or no modification, and then sell them to other news organizations. They provide these articles in bulk electronically through wire service by means of Internet. Corporations, individuals, analysts and intelligence agencies may also subscribe.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2015

BLOGS - A blog is a type of website, usually maintained by an individual with regular entries of commentary, descriptions of events, or other material such as graphics or video. Many blogs provide commentary or news on a particular subject whereas others function as more personal online diaries. There are many different types of blogs, differing not only in the type of content, but also in the way that content is delivered or written.

Wiki - Wiki is a website that allows easy creation and editing of any number of interlinked web pages via a web browser using a simplified markup language. Wikis are typically powered by wiki software and are often used to create collaborative websites, to power community websites, for personal note taking, in corporate intranets, and in knowledge management systems.

B. Overall Design

On determining the purpose and specification of the project, the design of the project is accomplished to develop plan for the obtained solution.

Considering various aspects of the software such as compatibility, extensibility, fault - tolerance, maintainability, modularity, reusability and usability the following design has been constructed.

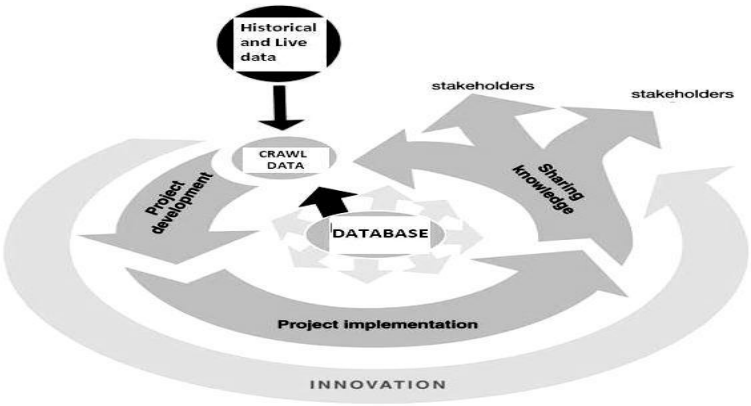

Figure 2 Overall Design

Figure 2 shows the overall design (Higher-level) of the system to be implemented. Figure 3 shows various phases (Functional level design) that have been analyzed and identified. Trend Detection module consists of a Back-End and a Front-End layer as shown in figure 4. From Back-End the Carrot[2] Clustering engine pulls data from various live sources like W3, Wiki, News, and Blogs via respective APIs according to the keyword inputted by the end user using the GUI.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2015

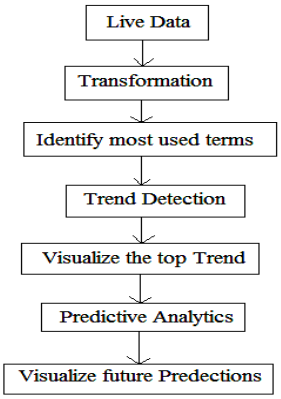

Figure 3 Phases of the Project

Every document is clustered using "Lingo" clustering algorithm. Then each cluster is being processed individually to obtain the most frequently used terms within each cluster.

The most frequent words is set to be 10% from the term that is used highest. Further analysis on the obtained trends is performed and top rated trends are sent to the GUI where the user visualizes various trends obtained. Figure 4 shows the various steps involved in detecting the trend for the keyword given. The proposed system is expected to perform predictive analytics as shown in figure 5, from a large input corpus which is predominantly textual in nature and hence it could also be stated as text analytics.

Like data mining, text analytics is an iterative process, and is most effective when it follows a proven methodology. This maximizes analyst productivity, supports comparability of results, allows findings from one analysis to be used to inform or guide others, and facilitates data-driven decision making. As with data mining, the two main steps in text analytics are data preparation and data understanding. Figure 5 shows various phases and concepts involved in Predictive Analytics Module.

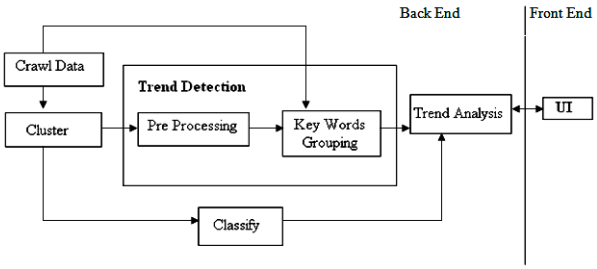

Figure 4 Trend Detection Architecture

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2015

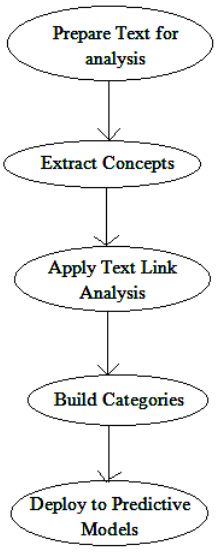

Figure 5 Predictive Analytics Process

III. IMPLEMENTATION AND RESULTS

The proposed algorithm has its base in text mining where live data from various sources like web, blogs, new wires etc., needs to be crawled to get only the relevant pages. Once the relevant pages of text are obtained, those pages should be mined for detection of topics. Various text mining preprocesses like stop word removal, stemming, POS tagging, disambiguation are to be performed to clean the text obtained.

Later text mining techniques like clustering, classification etc., and various other statistical methods are applied to categorize the text and obtain relevant topics. Trend detection algorithms are to be applied to the detected topics to identify the current trend in the topics obtained. Various predictive analytics methodologies need to be applied to the obtained trends to predict its future. The final output i.e. the predicted future trends needs to be presented to the user of the system via various visualization techniques. The proposed algorithm has five major steps:

- Crawling of Live data from various Websites
- Topic detection
- Trend Detection
- Predict the future of the trends obtained (Predictive Analytics)
- Visualization of the trends and their expected future

A. Crawling Data

The system under development is not domain specific and requires data that is both historic as well as up to date i.e. Live Data. Owing to the above constraints data cannot be stored in Databases for use. Instead

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2015

input data has to flow into the system from the enormous digital resources that are readily available from sources such as World Wide Web. A Web crawler is a computer program that browses the World Wide Web in a methodical, automated manner. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches. Crawlers can also be used for automating maintenance tasks on a Web site, such as checking links or validating HTML code.

Carrot^2 is being used for crawling of data and is not a search engine itself; it does not have a crawler and indexer. It calls the respective Website"s API which is inbuilt and uses the crawler that is supported by that website. Carrot^2 is open source.

Text Link Analysis - Once the extraction process is complete, Text Link Analysis has to be performed to describe relationships between concepts at the sentence level, as well as any opinions or qualifiers attached to these concepts.

Building categories - Building categories and categorizing documents are the next steps in analyzing text documents. Basically classification has to be performed for the concepts and links obtained.

Deploying results to predictive model - Deployment of text analytics results to predictive models is the step that will link text analytics to decision making.

The above mentioned steps and processes of predictive analytics is the proposed approach that has to be implemented into the system and visualized.

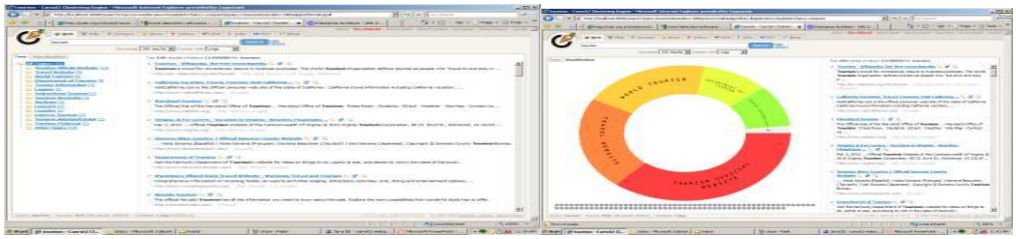

Figure 6 Clustered Document in Tree View and Visualization

Figure 6 shows snapshot of the system with keyword "tourism" inputted by the user. The display screen shows the clusters on the left pane and the documents in the right pane. Also, shows snapshot of the system with keyword "tourism" inputted by the user. The display screen shows the clusters in the left pane and the clicked document in the right pane. Thus the chapter gives the current state of implementation and also discusses on the proposed approaches of the modules to be implemented.

IV. CONCLUSIONS

This paper analysis on how trends easily detected from living data that is not present in specific domain has been accomplished fully. A discussion level of system design is formulated from the analysis. Many open source tools that will help in the implementation are also given a glimpse to. The approach towards predictive analytics and its visualization are under progress. Trend detection part is successfully

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2015

implemented and further enhancement would be identifying the association between various trends identified. The process of emerging trend detection can be made iterative. As of now Carrot^2 tool takes only an input of 150 documents at its maximum, thus steps can be taken to scale Carrot^2 for better precision in the trends obtained. Predictive Analytics has been analyzed and designed completely. Implementation of the same requires detailed study and research of the domain as it is an emerging and promising field. On successful implementation, it would serve as a greatmeans of decision making to all the stakeholders involved with the system. Efficient visualization of the same has to be implemented for better presentation and understanding.

REFERENCES

1. A. Berger, S. Della Pietra, and V. Della Pietra, "A maximum entropy approach to natural language processing", Computational Linguistics, 22 (1) : 39 –71, 1996.
2. April Kontostathis, Lars E. Holzman, and William M.Pottenger, "Use of Term Clusters for Emerging Trend Detection".
3. Charles Nyce, "Predictive Analytics White Paper",2007.
4. Christoph Morbitzer, Paul Strachan, and Catherine Simpson, "Application of Data Mining Techniques for BuildingSimulation Performance Prediction Analysis", Eighth International IBPSA Conference, Eindhoven, Netherlands, pp. 911-918, August 11-14,2003.
5. Debbie Mayville, "Using predictive Analytics to uncover root causes and solve problems Vs. Treatment symptoms", August2, 2006.
6. Han J. Kamber M., "Data Mining – Concepts and Techniques", Morgan Kaufmann Publishers, 2001.
7. http://carrot2.svn.sourceforge.net/
8. J. M. Ponte, and W.B. Croft, "Text Segmentation by Topic", Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries, pp. 120–129, 1997.
9. Kleinberg J. Bursty, "Hierarchical Structure in Streams", ISIGKDD'02, Edmonton, Alberta, Canada, 2002.
10. Michael Mathioudakis, and Nick Koudas, "Twitter Monitor: Trend Detection over the Twitter Stream", 2009.
11. S. Roy, D. Gevry, and W. M. Pottenger, "Methodologies for trend detection in textual data mining" Proceedings of the Textmine "02 Workshop, Second SIAM International Conference on Data Mining, April 2002.
12 . Soma Roy, David Gevry, and William M. Pottenger, "Methodologies for Trend Detection in Textual Data Mining", 2000.

BIOGRAPHY

Dr. G. Murugan was born on 15th May 1974 India. He finished his Ph.D in Computer Science Engineering in CMJ University, Meghalaya 2012, M.Tech in Computer Science Engineering in Sam Higginbottom Institute of Agriculture, Technology & Sciences, Allahabad 2006, M.C.A. in Bharathidasan University 1997. He has published three papers in International Journals in Image processing. His current research interests are image processing, data mining and computer network.