ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

News Media Analysis

[I]Neha Rajadnya, [II]Sayali Pendharkar, [III]Anisha Dhekne
[I,II,III]Pune University, India

Abstract
The analysis of media content has been central in social sciences, due to the key role that media plays in shaping public opinion. News analysis refers to the measurement of the various qualitative and quantitative attributes of textual (unstructured data) news stories. Newly evolving technologies in Big Data domain are becoming key basis of competition, underpinning new waves of productivity growth, innovation, and consumer surplus.
This paper proposes a methodology that will allow user to fire queries in simple format i.e. by only specifying the names of objects, person etc that he wants to analyze & provide the user with analysis in the form of graphical representation (charts/graphs) to his query. Similar to Google Trends, this service will also use aggregated data stored on Hadoop, however data provided by Google trends is updated daily whereas this service will analyze the query on the basis of data available on news websites in the past one year. This analysis will include the crawling of WebPages of news articles using a web crawler and further processing of the data.

Keywords
Big Data, Hadoop, parsing, crawling, News Analysis, Graph, web Crawler

I. Introduction

The data available on the various news websites will be in the unstructured format. This service will help in presenting the unstructured information in an ultra precise and summarized format. News analysis refers to the measurement of the various qualitative and quantitative attributes of textual (unstructured data) news data.The textual data will be very tedious to handle if the user wants to analyze a particular thing. In order to overcome this challenge, the paper emphasis is to provide processed content which will be presented to the user in a more readable pictorial format which also reduces the time taken by the user to obtain desired information.

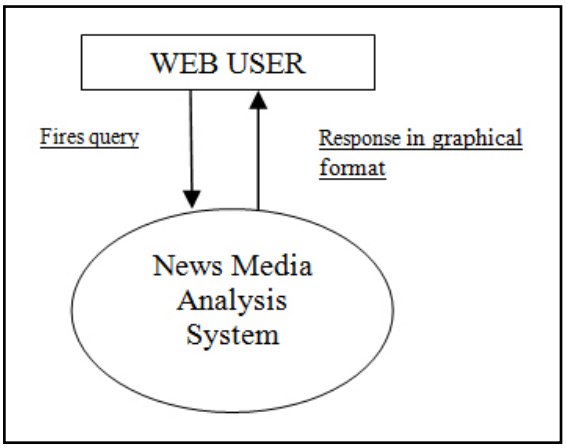

Fig.1: News Media Analysis

The key factors that make news media analysis such an important topic in today's world are as follows:

1) The analysis of media content has been central in social sciences, due to the key role that media plays in shaping public opinion.
2) The idea that data and the ability to filter it and make sense out of it can be a powerful tool for digital journalism.
3) The data or more precisely the news that is available on various news websites will give its readers information in the textual form.
4) It would be more helpful for the public to take decisions or analyze particular issues if the data was presented to them in pictorial format.

II. Literature Review

Google Trends is a public web facility of Google Inc., based on Google Search, which shows how often a particular search-term is entered relative to the total search-volume across various regions of the world, and in various languages [3]. Google trends will allow user to fire queries in simple format .i.e. by only specifying the names of objects, person etc that he wants to analyze. Some of these attributes include relevance, sentiment and novelty. The data provided by Google Trends is updated daily. Similar to Google Trends, our service will also use aggregated data; however data provided by them is updated daily whereas we are going to analyze the query on the basis of data available on news websites in the past one year.

However, regarding Google trends, it has been found that no algorithm is perfect and anomalies in the data may be found on rare occasions. Currently Google Trends data is computed by a sampling method and varies somewhat from day to day. This sampling error adds some additional noise to the data. A more accurate estimation of the Trends query share indices is expected.

While Google Trends and Twitter have already been recognized as a valuable source of trend information, efforts in the field of trend detection over Facebook public posts too emerged recently. A system for trend detection based on the characteristics of the posts shared on Facebook has been evaluated [2].Based on the results, three categories of trending topics: 'disruptive events', 'popular topics' and 'daily routines' were proposed. Analysis and comparison of the characteristics of the proposed categories in terms of distribution and information diffusion in order to increase the understanding of emerging trends on Facebook has been done. Finally conclusions from the findings in terms of challenges and opportunities for future work in this direction have been done. It can be concluded from the above project that the analysis performs well only on certain topic groups and results in problems with topics where there is a little overlap between separate terms belonging to the same topic group and an existing overlap with the more dominant topic group. Therefore, this algorithm needs to be further improved to achieve optimal results.

Finding a wise way of extracting only the useful data for further analysis plays a significant role in promoting the efficient and effective use of the internet. A system which performs the analysis and visualization of the emerging consumer generated media

(CGM) posts and online news archives has been previously presented in a more user-friendly way [1]. In order to overcome the heavy time complexity incurred, an approach to extract only the useful data from the CGM by means of the Time Series Data Processing technique, namely, the Perceptual Important Point (PIP) has been used. By correlating the sorted out time series data with the online texts, further analysis could be done in a more effective and efficient way. Regarding this project, it would have been better to integrate something like the Natural Languages Processing (NLP) for automatic sentiment analysis and summarization for greater efficiency.

An approach to exploring twitter data which attempts to automatically analyze large volumes of twitter comments with respect to what was commented on positively or negatively has been presented [10]. To achieve the, a novel topic-based text stream analysis technique that automatically detects which attributes were frequently commented on in tweets, based on their density distribution, negativity, and influence characteristics was developed. Regarding this project, incorporating information about opinion associations to find related features and visualize them appropriately would have been an added feature. Also, using visual analysis tools (e.g., SAS JMP, Vivisimo, etc.) would have been better as they mainly provide feedback on reviews using yes/no questions, numeric ratings, and direct comments.

Our paper proposes the novel idea of developing the service that will help in analyzing the news stories since the information published in the news papers is legitimate and irrefutable, making us sure about the results of our analysis.

III. Methodology

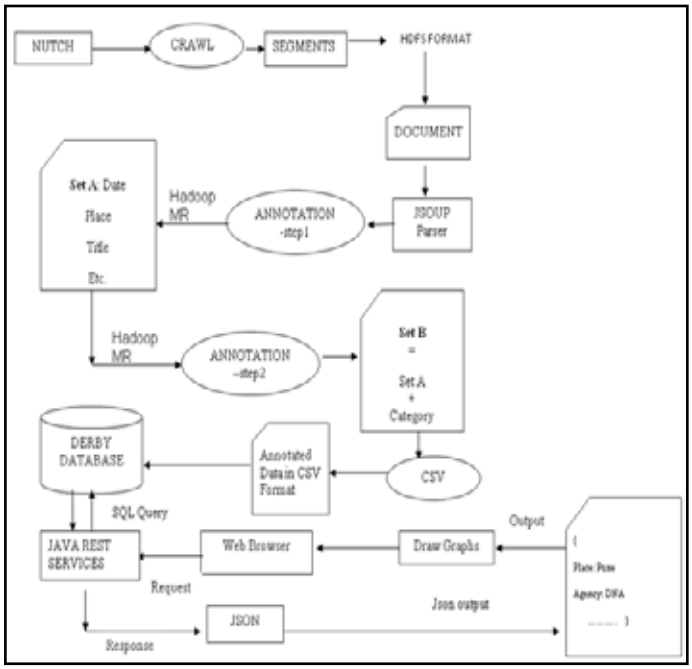

Fig. 2: Methodology

IV. Proposed System

Our paper proposes a system that finds an optimal way for conversion of unstructured data to structured format and presenting the user with the graphical representation for his query.

Fig. 2 illustrates the working of our system. Initially, the URLs of various news articles from news websites are fed to the crawler, Apache Nutch in the form of seeds. Crawling is the process that can copy all the pages web crawlers visit for later processing by a search engine that indexes the downloaded pages so that users can search them much more quickly. It is done with the help of a web crawler, a component that fetches the html page contents of the provided URL. Nutch is an effort to build an open source web search engine based on Lucene and Java for the search and index component.Nutch is coded entirely in the Java programming language and has a highly modular architecture, allowing developers to create plug-ins for media-type parsing, data retrieval, querying and clustering.The crawled data is then stored in the form of segments on Hadoop File System. Hadoop is an open source platform for scalable and distributed computing of large data sets across clusters of computers using a simple programming model.

The segments generated are converted into readable format and are given to the JSOUP parser for parsing. Jsoup is a Java library for working with real-world HTML. The data is annotated and stored in a CSV format. The resultant CSV is then loaded into the Derby database. JAVA REST services act as middleware and performs the job of creating the appropriate JSON from the data retrieved from the database according to the query fired by the user. This JSON is then given to various charts and the charts get modified accordingly.

A . Algorithm

/*Input: User query
Output: Pictorial Representation
Description: This project intends to crawl data only from news media website.
The user will receive the response to his request in the form of charts.*/

1. Start
2. seeds :=no_of_seeds
3. for i:=1 to no_of_seeds
4. segs[ ]=crawl seeds using Nutch
5. Parse the segments using JSOUP parser.
6. Get the Title, Heading, Category, Date, Topic for each article.
7. Store the annotated output into Derby.
8. Provide the relevant JSON to the UI using JAVA Rest services.
9. The required web page is obtained as the output.

V. Results and Discussion

Fig3. Illustrates a snapshot of the output generated on firing queries by the user.

Fig. 3 : Snapshot of Dashboard of our Proposed System

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

The dashboard comprises of five use cases represented by user interactive charts which are updated as per the query fired. The service limits query firing to two. The Line Chart displays comparison of the two queries for article count per month in the year 2013. The Donut Chart illustrates the total article count of that query in the system. The fired query extracts the category and analysis of the category-wise articles is plotted on the Bar Chart. Bubble Chart represents the Top 10 topics in the dataset which remains static irrespective of the query. Another use case which represents the category-wise distribution of the static dataset is represented by a Pie chart. The results of these interactive charts can be used by the user for his requirement.

VI. Conclusion and Future Scope

In this paper we presented a novel idea for establishing a static analysis engine. The focus of the present work is to make such analysis provide reliable understanding of likely outcomes and effects. Morover in future, the extension of the service can be made intelligent enough to capture the real-time news data and make the system dynamic. Also the system can be extended to provide sentiment analysis and relevance analysis for more communicative and descriptive outcomes.

VII. Acknowledgement

We would like to express the deepest appreciation to Persistent Systems Pvt. Ltd. for giving us a opportunity to get associated with them through BE project sponsorship. Also we would like to thank our mentor in Persistent Systems Pvt. Ltd. Rushikesh Pol. Without his supervision and constant help this dissertation would not have been possible.

We would like to thank our Institutional Head of Computer Department, MMCOE Pune, Professor Ram Joshi as well as our institutional project guide, Professor Rahul Dagade for their knowledgeable guidance and support.

References

[1] *"Analysis and Visualization of Time Series Data from Consumer-Generated Media and News Archives" Tak-chung Fu, Donahue C.M. Sze*

[2] *"Monitoring Trends on Facebook" Irena Pletikosa Cvijikj, Florian Michahelles, 2011*

[3] *"Predicting the Present with Google Trends" Hyunyoung Choi Hal Varian, 2009*

[4] *Big Data Analysis of News and Social Media Content" Ilias Flaounas, Saatviga Sudhahar, Thomas Lansdall-Welfare,Elena Hensiger. Intelligent Systems Laboratory, University of Bristol*

[5] *"A Novel Indexing Scheme for Efficient Handling of Small Files in Hadoop Distributed File System " Chandrasekar S, Dakshinamurthy R, Seshakumar P G, Prabavathy B, Chitra Babu,Department of Computer Science and Engineering,2013*

[6] *"Tibetan Web Information Collection System" Guixian Xu1,2, Dunhao Zhong, Xu Gao ,Yuan Lin, Xiaobing Zhao, Guosheng Yang, 2012*

[7] *"Designing and Implementing of the Webpage Information Extracting Model Based On Tags" Zhang Xu Zhang Xu, 2011*

[8] *"A Novel Indexing Scheme for Efficient Handling of Small Files in Hadoop Distributed File System" Chandrasekar S, Dakshinamurthy R, Seshakumar P G, Prabavathy B, Chitra Babu, 2013*

[9] *"Shared Disk Big Data Analytics with Apache Hadoop" Anirban Mukherjee, Joydip Datta, Raghavendra Jorapur, Ravi Singhvi, Saurav Haloi, Wasim Akram, 2012*

[10] *"Visual Sentiment Analysis on Twitter Data Streams" Ming Hao, Christian Rohrdantz, Halldór Janetzko, 2011*