Forecasting from Identifying Events and Temporal Relationships from Wikipedia Mining

Rajani Singh[1] and Dr. S.B. Kishor[2]

[1]Researcher, RTMNU, Nagpur and Working in S.P. College, Chandrapur
Email: rajanisingh13@gmail.com
[2] HOD, Department of Computer Science, Sardar Patel Mahavidyalaya, Chandrapur
Email: s.b.kishor.spc@gmail.com

ABSTRACT

In this real world people needs the greater opportunities in their life which enable them to grow beyond the expectation. Wikipedia is a web-based free encyclopedia that can be assessed by the whole netizen. This paper draws the Events and Temporal Relationship from the Wikipedia Mining. Identification of events and temporal Relationships uses the Timelines from Wikipedia Articles which help to foretell about the near developments and events in the different Areas. This Paper trends to forecast either from qualitative computation or from quantitative output from simulations based on historical data.

Keywords**:** Wikipedia Mining, Web Uses Mining, Web Content Mining, Web Structure Mining, Web Corpus, Timelines, Time Series Method, Qualitative Computation, Quantitative Outputs, Smoothing Method.

INTRODUCTION

Recently most of the research has been devoted to the Web Mining. Web mining is mainly categorized in to three types: *Web usage mining, Web content mining* and *Web structure mining*. “Web Uses mining” provides the useful information via extracting the data stored in the server logs i.e. from user’s history. “Web content mining” provides the process to use the Graph Theory to examine the node and structure of connection of any site where as “Web Structure mining” extraction and integration of useful data, information and knowledge from the contents of Web page. Data mining which also called as knowledge discovery is the process of analysis of data from different linear position and summarize it into useful information. Data mining software is the analytical tools that allow users to examine data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, Data mining is the extraction of hidden predictive information from large databases, is a powerful new technology.

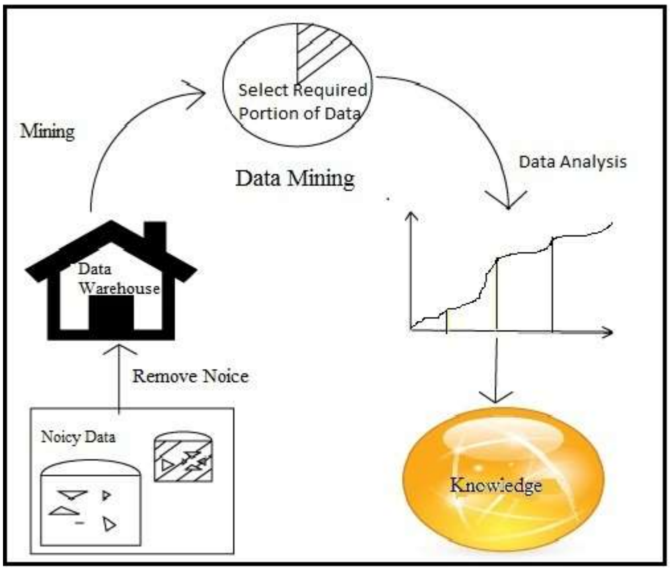

Fig 1: Deriving Knowledge from Noisy Data.

Some Issues encounters while knowledge retrieval in Data Mining technique are as follows:

- Handling noisy or incomplete data Pattern evaluation—the interestingness problem
- Incorporation of background knowledge
- Data mining query languages and ad hoc data mining
- Mining different kinds of knowledge in databases
- Interactive mining of knowledge at multiple levels of abstraction
- Presentation and visualization of data mining results

Wikipedia Mining

"Wikipedia mining" is the novel research area and it has various impressive characteristics like a huge amount of articles, live updates, a dense link structure, brief link texts and URL identification. The given statistics binds various aspects of Wikipedia, as an encyclopedia, a website, or a community. Some provide current snapshots and others track growth and development over time. Following is the manually created chart of English-language Wikipedia Article Count in the month of January 2001 to January 2012 published by en.wikipedia.org.

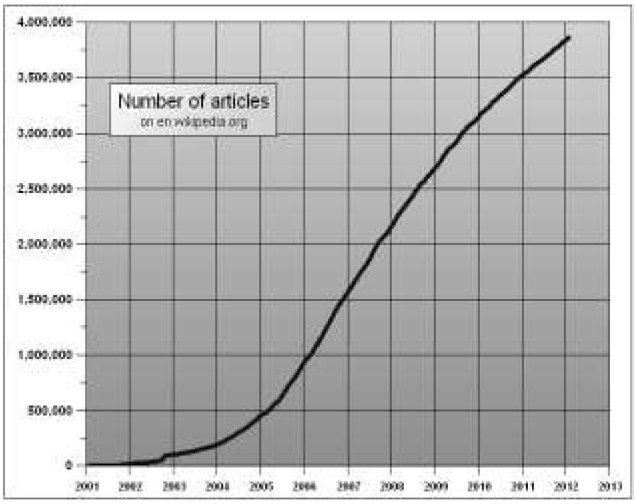

Fig 2: Manually created chart of English-language Wikipedia Article Count - January 2001-January 2012

Wikipedia is hosted by the Wikimedia Foundation, which is a non-profit organization that also contain the number of Wikipedia's Sister projects like Commons (Free media repository), Wiki books (Free textbooks and manuals), Wikinews (Free-content news), Wikiquote (Collection of quotations), Wikisource Free-content library, Wiki species Directory of species, Wikiversity (Free learning materials and activities), Wiktionary(Dictionary and thesaurus) and Meta-Wiki (Wikimedia project coordination).

Wikipedia is the famous Web Corpus uses for the extraction of knowledge. Web Corpus can either gaining Web content and processing it into a static corpus or as a dynamic corpus. As the Web uses is constantly expanding, so its size is unknowable. In the year 2008 Google noted that it had identified approximate over a trillion (1012) distinct URLs (Web addresses), and that several billion (109) new web pages appear daily. In this educated world, internet accessing is fast and inexpensive as well as the growing demand from the developing countries improve service and reduce costs. With a home broadband connection one can compile and process a multimillion word corpus in minutes.

1. Dense Link Structure

One of the most interesting characteristics of Wikipedia is its "dense link structure", "Dense" means it has a lot of "inner links," links from one Wikipedia page to another Wikipedia Page. It's means that all the articles are tightly connected with many hyperlinks. Dense link structure shows the potential of Wikipedia mining and it is possible to extract valuable knowledge by analyzing the link structure. As compared to the other average Web-site Wikipedia provides the clear, simple and brief structure of data via the feature called "Link Text". The facility like "Click here for more information" is also provided by the thesaurus visualize based on Web mining but sometimes it does not contain any information about the link page, whereas Link Text provides the clear solution over this problem.

2. URL Identification

Another characteristic of Wikipedia is URL identification where almost every page or article contains their own URL as an identifier i.e. each page/concept can be find by its own URL. The Wikipedia follows the Bottom-Up approach for word searching plus it provides several entries for the single term with detailed information and Picture. It provides the facility to easily and timely edit the content by User by providing edit option.

Wikipedia Thesaurus Visualizer

Wikipedia Thesaurus Visualizer is a application for managing the huge scale thesaurus of data extracted from Wikipedia. It allows the users to know about the relation between the different concepts visually and easily. Wikipedia Thesaurus Visualizer is developed with the help of Web based technology known as Rich Interface Application, supported by almost all modern desktop environments like:

- Operating System: Windows XP/Vista, Mac OS X
- Browser: Firefox, Safari, Chrome etc.

For clear and effective visualization of data this system provides some network analysis function described as follow:

- Power Node Detection - It highlights nodes having many relations to other concepts in the network.
- Minor Node Detection - It hides nodes that have few (relations) edges.

Finding Events and Temporal Relationships

Once the data is extracted from the Wikipedia using Wikipedia Mining, relation among different data can be easily identified using the Temporal Relationships. By finding the temporal relation between the events one can check the simultaneity or ordering in time of events or states. To identifying the relationship between two we need to calculate the timeline between the different linear events. Timeline is typically a graphic design which shows a long bar which is labeled with dates and events labeled at points where they would have happened. It can use any time scale, from minutes to years to millions of years. Timeline identifies the two successive events and it is beneficial to forecast the near developments and events in the field of Technology, Science and in Social Area. Forecasting helps the policy making process used in the Government and Commercial Area. We can foretell either from qualitative computation or from quantitative output from simulations depends on the Data (Historic Data).

We use here the subjective approach for the qualitative computation where this approach gives chance to different users to make decision based on their subjective feeling and ideas and participate in the forecasting. Here “Brainstorming sessions” are frequently used to develop the idea which helps to solve complex problems. This subjective approach is increasingly uses by many corporations in the United States to make a proper decision about the Business. Whereas the Time Series method is used to take the Quantitative output, in time series method measurements are taken from the successive points or from the successive period. These measurements can be taken hourly, weekly, monthly, yearly or at any regular or irregular interval of time. Time series data can displays some random fluctuations or it show gradual high or low shifts over the extended period of time this gradual high or low shifts may shows about the population size, changes in tastes and preference of customer whereas the Random fluctuation or variation shows irregularity and it cannot be predicted in advance that is any events cannot be forecast if there is any irregularity in time series method.

To overcome this problem smoothing method is used to smooth out the irregular components of the time series. In order to smooth out the time series this method uses the average of a number of adjoining data points or periods. This process uses the overlapping observation to generate average; it is explained with the given example:

Suppose we need to forecast the first two observations from the time series and aim is to calculate average using smoothing method. To calculate we need to drop the first observation and calculate the average of next two observations, this process continues until the two period averages are calculated. We need to move up or down in the time series to pick the observations to calculate an average from the fixed number of observations.

1. Forecasting Helps to Areas

After Wikipedia Mining events and the relation among the different Articles can be easily identified and to calculate timeline between the events is beneficial to the new development in the different area mentioned below:

- Artificial intelligence
- Biology medicine
- Communications
- Computing
- Economics
- Nanotechnology
- Politics
- Robotics
- Transportation
- Space etc.

Timeline from Wikipedia Articles

Timeline provides the way to display the events in the chronological order. It is a graph design which shows a long bar which is labeled with dates and the different events are labeled with points where they would have happens. It is generally used to study History as it shows the sense of Change over time. Timelines are often useful to show Wars, Social Movements and for biographies. It is helpful for the natural world and for the subjects of Science like Astrology, Biology, Geology, etc. A different kind of timeline is used in Project

Management where Team members use the Timeline to achieve their milestone under the time schedule.

CONCLUSION

In this paper we have presented the technique to identify events and temporal relationships from Wikipedia Mining. Once the process of Mining is completed we start with finding the timeline which foretell about the newer development, where forecasting helps for decision making. Definitely final result is based on Historic data but forecasting will be divided in to two types. First is the Qualitative calculation which depends on the current state and second is the Quantitative output which calculates from Models or from Simulations. Our technique of identifying events and temporal relationships helps some areas which involved in the process of invention through 2025 and some of them are: Nanomaterials, Distributed Energy, Personalized medicine, Pervasive Computing, Biomarkers for health, Bio fuels.

REFERENCES

1. Jaideep Srivastava, Prasanna Desikan, Vipin Kumar," Web Mining – Accomplishments & Future Directions", http://www.ieee.org.ar/downloads/Srivastava-tut-paper.pdf
2. James Pustejovsky, Jos´e Casta˜no, Robert Ingria, Roser Saur´, Robert Gaizauskas, Andrea Setzer, Graham Katz, "TimeML: Robust Specification of Event and Temporal Expressions in Text", http://timeml.org/site/publications/timeMLpubs/IWCS-v4.pdf
3. K. Nakayama, T. Hara, and S. Nishio: A Thesaurus Construction Method from Large Scale Web Dictionaries, Proc. of International Conference on Advanced Information Networking and Applications (IEEE AINA), pp. 932-939 (May 2007).
4. K. Nakayama, T. Hara, and S. Nishio: Wikipedia Mining for An Association Web Thesaurus Construction, Proc.of International Conference on Web Information Systems Engineering (WISE), pp, 322-334 (Dec. 2007).
5. M. Erdmann, K.Nakayama, T.Hara, and S.Nishio: An Approach for Extracting Bilingual Terminology from Wikipedia, Proc. of International Conference on Database Systems for Advanced Applications (DASFAA), (Mar. 2008).
6. William H. Fletcher," Corpus Analysis of the World Wide Web"

Webliography

- http://webascorpus.org/Corpus_Analysis_of_the_World_Wide_Web.pdf
- Forecasting http://www.referenceforbusiness.com/encyclopedia/Fa-For/Forecasting.html#b#ixzz1oXmHDULM