Application of Web Mining in Medical and Health Care: A Survey

J.S.Raikwal[1], Dr. Kanak Saxena[2]

[1]Department of Information Technology,

I.E.T, D.A.V.V, Indore

[2]Department of Computer Application,

S.A.T.I., Vidisha

jraikwaliet@gmai.com, Kanak.saxena@gmail.com

Abstract: *Web mining is the field having applications in variety of domains like e-commerce, e-learning, digital libraries etc. As the popularity of website related to medical and health car continues to grow, so web mining also derived its applications in medical and health care domain. This paper provides a survey of web mining applications in medical and health care domain and focuses on some research work in this domain.*

1. Introduction

Web is one of the central largest sources of data in the world. However, the extraction of information from web is a difficult task due to its unstructured definition, its un-trusted sources and its dynamically changing nature. The number of health information websites and online services is increasing day by day. Some of the services provided by these sites are: information about symptoms and cause of any disease, information regarding the cure and medicine for the particular disease, information about the practitioners and specialists for a particular disease and information about the health care centers and hospitals in particular area. Much more information may also be provided by these websites in addition to above.

These medical and health care websites are helping and providing information to their users (patients, general public and physicians), but also raising some issues. Some of them are: it is difficult for health information consumers, such as the patients and the general public, to assess by themselves the quality of the information because they are not always familiar with the medical domains and vocabularies [1], it is necessary to implement control measures that give the consumers adequate guarantee that the health web sites they are visiting, meet a minimum level of quality standards and that the professionals offering the information on the web site are responsible for its contents [2], automatic and autonomous methodology to discover taxonomies of terms from the Web and represent retrieved web documents into a meaningful organization, speed up detection of potential health threats and allow timely response (Epidemic Intelligence), public health messaging, deriving new medical hypothesis, etc. This paper is a survey for the application of web mining techniques in resolving or proposing methodologies for above mentioned issues regarding the medical and health care domain.

This paper is organized as follows: section 2 gives the overview of web mining, section 3 is focused on the importance of web mining in medical and health care domain, section 4 presents some related work in the domain and finally section 5 contains the conclusions and proposes lines of future work

2. WEB MINING

2.1 Overview:

The web is a means for accessing a vast variety of information stored in various parts of the world. Information is mostly in the form of unstructured data [3]. As the data on the web grows rapidly, web users face the following problems:

a) Finding relevant information.

b) Creating new knowledge out of the information available on the Web.

c) Personalization of the information.

d) Learning about consumers or individual users.

These problems are briefly described in [4]. Web mining is an emerging research area focused on resolving these problems [5].

2.2 Definition:

Web mining is the application of data mining techniques to discover and retrieve useful information (knowledge) from the web document and services [6, 7].

2.3 Taxonomy [8]:

Web Mining can be broadly divided into three distinct categories, according to the kinds of data to be mined:

1. **Web Content Mining:** the process of extracting useful information from the contents of Web documents. It may consist of text, images, audio, video, or structured records such as lists and tables.
2. **Web Structure Mining:** The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages. Web Structure Mining can be regarded as the process of discovering structure information from the Web.
3. **Web Usage Mining:** Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a website

2.4 Techniques:

1. **Document Clustering:** Document Clustering is unsupervised technique for web mining [8]. Documents are divided into groups based on a similarity metric
2. **Document Classification:** According to [8] classification is supervised technique for web mining. Some categories are defined and documents are identified according to those categories. A document may belong to more than one category and many documents may belong to only one category.

3. **Association Rules [9]:** The association rule can be seen as the identification of actions or facts that, being initially independent, they happen in a combined or associate way. The considered facts can be characteristics or behaviors observed in the individuals.
4. **Sequential Patterns [9]:** A history of transactions and web pages visited by user is stored in web server for a period of time. This discover of patterns is based on identifying the group of more frequent accesses in a group of transactions or visits in a given time.
5. **Cubes [9]:** A cube of data is a type of multidimensional array that allows the users to explore and to analyze a collection of data from different perspectives. From a structural perspective, the cubes of data are composed of two elements: dimensions and measures. The dimensions are categories that describe the studied factors for their analysis, and the measures are the values of the data stored in that structure.

3. Importance of Web Mining in Medical and Health Care

Web mining techniques helps:

- To discover knowledge for understanding the cause of any disease and its treatment.
- To track error done by hospital staff and enable them t correct the error and prohibit them to repeat the same in future.
- To identify some patterns to set the policies for health care centers and hospitals.
- In early detection and/or prevention of diseases. It helps in Non-invasive diagnosis and decision support. It helps to identify Adverse Drug Events (ADE, Some drugs and chemicals that have been approved as non.harmful to humans are later discovered to have harmful effects after long-term public use.).
- Practitioners and medical researchers to generate some new theory regarding any disease.

4. Survey of Research Work in Medical and Health Care Domain

In [11], authors proposed a system for Epidemic Surveillance. The system comprises of three stages: (1) identification of articles relevant to Epidemic Surveillance using keyword web search, (2) use of information extraction (IE) for determining the content of the article and find out information like, who was infected by which disease, what are the symptoms of the disease, what are the conditions for the illness, where and when illness occurs, etc. This information can be used to create a structured record that is stored in the database. Articles that do not trigger creation of a database record are discarded, (3) determines the relevance of the selected articles and cases that they describe to the domain of Public Health.

A concept for an augmented medical information management (AMIM) is presented in [12]. They worked on content-based medical image retrieval based on similarity of both the global image and local regions, capturing human expert knowledge with visual similarity metrics, retrieval from EHR archives and the Internet based on both visual and textual patterns, and a cheap and effective paradigm for home-based telemedicine. With their work the found that currently all indexing and searching of web is based only on the textual data often available with images and very small number of medical images database is available on the web. They presented an example of Google, that it results only 120 images from the keyword “lung CT” but groups like “http://www.csimage.com” and “http://www.healthcentral/org” contains at least 1000 lungCTs. According to their AMIM approach, available medical web sources for images are indexed in textual and visual form to make the wealth of knowledge available to practitioners for teaching, research and also diagnostic aid. Authors used the web content mining techniques for image extraction from different web resource related to medical domain also machine learning methods are used to capture the notion of similarity.

In [14] authors describe a system, Proteus-BIO, automatically creates a table of outbreaks, with each table entry linked to the document describing that outbreak. Their system consists of three basic components: (1) Web Crawler: gathers relevant documents from web, (2) Extraction Engine: converts the individual outbreak events to a tabular database, and (3) Database Browser: provides access to the events and, through them, to the documents. A web crawler is one of the web usage mining techniques and traverses portions of the web, looking for new, relevant web pages. It searches the web trees starting from the root nodes of selected general and medical news sources, looking for new web pages. The Web crawler also finds the text body within the Web page. A typical Web page has lots of information besides the actual text of a story: headlines, links to other stories, sponsorship information or advertisements, etc. For most Web pages, the crawler uses the HTML markup to locate the relevant text. For the ProMed Web pages, which contain primarily text without HTML markup, the crawler uses specific text tags and other layout indicators (blank lines, capitalized lines, etc.).

In [15] authors presented the Global Health Monitor, which is a Web-based system for detecting and mapping infectious diseases from Web. The system collects news from news feed providers, analyzes news and plots disease relevant data onto a Google map. The monitor comprising of three modules

In [16] presents a freely accessible system designed to monitor disease epidemics by analyzing textual reports, mostly in the form of news gathered from the Web. They describe the IR and early-warning functionality of MedISys, and how it inter-operates with the information extraction (IE) system PULS, which analyses the documents identiûed by MedISys, retrieves from them events , or structured facts about outbreaks of communicable disease, aggregates the events into a database, and highlights the extracted information in the text. The Medical Information System, MedISys, automatically gathers reports concerning Public Health in various languages from many Internet sources world-wide, classiûes them according to hundreds of categories, detects trends across categories and languages, and notiûes users. PULS, the Pattern-based Understanding and Learning System, is developed at the University of Helsinki to extract factual information from plain text. PULS has been adapted to analyze texts in the epidemiological domain, for processing documents that trigger MedISys notifications [17]. The aggregation of reports into larger units. MedISys and PULS use different approaches to aggregation. MedISys use clustering and PULSE group disease events into outbreaks.

Due to increasing number of health information related websites, it is difficult to assess the quality of information provided by them. In [2] authors proposed a quality labeling technique (or Medical Website Certification) for those websites which provide medical and health care information to consumers. For this they use semantic web technologies that enable the creation of machine-processable labels and also automate the labeling process. Web mining techniques like web crawling or web spidering with IE is used for continuous monitoring of labeled resources alerting the labeling agency in case some changes occur against the labeling criteria. The AQUA (Assisting QUality Assessment) system [19], developed within the MedIEQ project [20], aims to provide the infrastructure and the means to organize and support various aspects of the daily work of labeling experts by making them computer-assisted. AQUA consists of five major components (each, in turn, incorporating several specialized tools): Web Content Collection (WCC), Information Extraction (IET), Multilingual Resources Management (MRM), Label Management environment (LAM), and Monitor Update Alert (MUA).

Authors in [21] find out that an internet that provides medical information which is important for researchers, health care providers and public is a powerful tool for public health messaging. Optimal use of the Internet for public health messaging requires an understanding of user characteristics, needs, and interests. Websites function as bidirectional communication channels, whereby Web content is the message sent to users and Web usage data from interactions between users and the website represents the visitor feedback. Web usage data reflects user's contextual interests, geographic locations, and navigation patterns; appropriate analysis provides insight to better understand and serve user's needs [6, 22]. In their work they examine the utilization of the chronic fatigue syndrome (CFS) website at the Centers for Disease Control and Prevention (CDC) by analyzing the web usage by consumers and proposed an efficient health messaging.

In [13] authors suggest a methodology to extract knowledge from the Web to build automatically taxonomy of concepts and web resources for the discovery of medical knowledge. Named entities for a discovered concept are found and are considered as instances. During the building process, the most representative web sites for each subclass or instance are retrieved and categorized according to the specific topic covered. With the final hierarchy, an algorithm for discovering different lexicalizations and synonyms of the domain keywords is also performed to find lexicalizations or synonyms, which can be used to widen the search.

Authors in [23] proposed that everyday, medically-oriented Web content is a valuable and viable data source for medical hypothesis generation and testing, despite its being noisy. They have constructed a corpus comprising news articles relating to the drugs Vioxx, Naproxen and Ibuprofen that were published between 1998 and 2002. Using this corpus, they showed that there was a signiûcant link between Vioxx and the concept “Myocardial Infarction” well before the drug was withdrawn from the market in 2004. Authors have used web mining techniques to derive this hypothesis.

Conclusion

With this survey we conclude that there is a vast application of web mining in medical and health care domain. Many of the techniques like classification, web crawling, and keyword based searching, constructing a knowledge base from web data, text mining etc, have been used successfully in the domain. Techniques are used to ensure the quality of information used by the consumers, to make new policies and hypothesis, to understand the criteria of using medical and health care related websites and many more applications. There is a scope of proposing a future research in the domain to min the WWW for providing quality, trusted and quick information to the patients and physicians to educate themselves for the domain.

References

[1] Soualmia LF, Darmoni SJ, Douyère M, Thirion B. Modelisation of Consumer Health Information in a Quality-Controled gateway. In: Baud R et al. (ed.). The New Navigators: from Professionals to Patients. Proc of MIE2003 (2003), 701-706.

[2] V. Karkaletsis, Stamatakis, K., Karampiperis, P., Labský, M., Ruzicka , M., Svátek, V., Cabrera, E. A., Pöllä, M., Mayer, M. A., Villaroel Gonzales, D., “Management of Medical Website Quality Labels via Web Mining”., http://www.medieq.org/publications.

[3] A.Mendez-Torreblanca, M.Monte,”A Trend Discovery for Dynamic Web Content Mining”, IEEE, Inteligence System, Vol 14, pages.20-22, 2002.

[4] Raymond Kosala and Hendrik Blockeel, "Web Mining Research: A Survey", SIGKDD Explorations. ACM SIGKDD, July 2000.

[5] Kshitija Pol, Nita Patil, Shreya Patankar and Chhaya Das, "A Survey on Web Content Mining and extraction of Structured and Semi structured data", First International Conference on Emerging Trends in Engineering and Technology, IEEE, 2008

[6] R.Cooley, B.Mobasher and J.Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web". In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), 1997.

[7] Companion slides for the text by Dr. M. H. Dunham, "Data Mining: Introductory and Advanced Topics", Prentice Hall, 2002.

[8] J. Srivastava, P. Desikan, and V. Kumar, "Web Mining: Accomplishments and Future Directions," *Proc. US Nat'l Science Foundation Workshop on Next-Generation Data Mining* (NGDM), Nat'l Science Foundation, 2002.

[9] Jose Aguilar, "A Web Mining System", WSEAS TRANSACTIONS on INFORMATION SCIENCE and APPLICATIONS, Issue 9, Volume 6, page number 1523-1532, September 2009

[10] http://www.medieq.org/about.

[11] Peter von Etter, Silja Huttunen, Arto Vihavainen, Matti Vuorinen and Roman Yangarber, "Assessment of Utility in Web Mining for the Domain of Public Health", 2010

[12] Thomas M. Lehmann, Henning Müller , Qi Tian , Nikolas P. Galatsanos, and Daniel Mlynek, "Augmented Medical Image Management", 2005

[13] David Sánchez, Antonio Moreno, "Web mining techniques for automatic discovery of medical knowledge "

[14] Ralph Grishman, Silja Huttunen, and Roman Yangarber, "Information Extraction for Enhanced Access to Disease Outbreak Reports"

[15] S. Doan, Q. H-N., A. Kawazoe, and N. Collier. 2008. Global Health Monitor—a web-based system for de-tecting and mapping infectious diseases. In Proc. In-ternational Joint Conf. on NLP (IJCNLP).

[16] Ralf STEINBERGER , Flavio FUART , Erik van der GOOT , Clive BEST , Peter von ETTER , and Roman YANGARBER , "Text Mining from the Web for Medical Intelligence ", Mining Massive Data Sets for Security F. Fogelman-Soulié et al. (Eds.) IOS Press, 2008

[17] C. Freifeld, K. Mandl, B. Reis, and J. Brownstein, "HealthMap: Global infectious disease monitoring through automated classiûcation and visualization of internet media reports," J Am Med Inform Assoc, vol. 15, pp. 150–157, 2008.

[18] Ulrika Josefsson & Ole Hanseth, "Patient's Use of Medical Information on the Internet: Opportunities and Challenges", Proceedings of IRIS 23. Laboratorium for Interaction Technology, University of Trollhättan Uddevalla, 2000.

[19] http://www.medieq.org/aqua/welcome.seam

[20] http://www.medieq.org

[21] Jason Bonander Hao Tian, PhD,[1] Dana J Brimmer, PhD, MPH,[1] Jin-Mann S Lin, PhD,[1] Abbigail J Tumpey, MPH,[2] and William C Reeves, MD, MSc[1] , "Web Usage Data as a Means of Evaluating Public Health Messaging and Outreach ", JMIR, 2009, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2802568/

[22] Srivastava J, Cooley R, Deshpande M, Tan P. Web usage mining: discovery and applications of usage patterns from web data. SIGKDD Explorations. 2000; 1(2):12–23. doi: 10.1145/846183.846188.

[23] Diana Maclean†, Margo Seltzer‡, "MINING THE WEB FOR MEDICAL HYPOTHESES- A Proof-of-Concept System ", 2010