Mining the Web in Bengali Language: Opportunities and Future

Tushar Kanti Saha [#1], Dr. Ahsan-Ul-Ambia [*2]

[#]*Lecturer, Department of computer science and engineering, Jatiya Kabi Kazi Nazrul Islam University*
Trishal, Mymensingh, Bangladesh
[1]tusharcsebd@gmail.com

[*]*Associate Professor, Department of computer science and engineering, Islamic University*
Kushtia, Bangladesh
[2]ambiaiu@yahoo.com

Abstract-**In this paper, we have focused the opportunity and future of mining the Web pages in Bengali language which will help to convert unstructured data in the Web to structured knowledge. Researches of Web mining have been advanced in other languages except Bengali. Knowledge discovery from Web is now interesting area of research. We can accomplish this goal by using Web mining technique. In the context of Natural Language Processing (NLP), Web mining in Bengali language is a challenging job because most of the Webs, where Bengali text is used, are not developed using unique font and style. Moreover, working with other language text rather than English is a challenging one for the researchers.**

Keywords-Web Mining, Web content mining, Knowledge Discovery, Bengali, Language, Opportunity, Future.

I. INTRODUCTION

With the advancement of Internet technology, Web has become popular among Bengali people all over the world. Web services are not only used by Government, organizations and researchers but also used by businessmen, professionals and students widely. There have been developed a lot of Websites in Bengali language for government organizations, business, online media, social blogs, information of particular area, etc. Due to rapid growth of using Bengali text in the Web, very soon we will be overwhelmed with huge Bengali Web data. Data in the Web can be categorized as unstructured (e.g pdf, document, excel files, etc) [1] and semi-structured data (e.g. mail message with attached document) [2]. In case of information retrieval in Bengali from the Web, search engines like Google, Yahoo, Bing, etc. have done some important jobs that will inspire our research. Now the time is to develop tools to transform these data into knowledge in a well organized way. As a part of data mining, Web mining in Bengali language is a new arena of research. A lot of works has already been done in other languages where a few works has been done in Bengali. There are several fields to work with Web mining in Bengali language. Research is needed in the field of online educational and sports data mining, online news mining, news comments mining, blog reviews mining and opinion mining, mining products reviews in e-commerce sites, etc.

This paper aims to explain the opportunities and future of Bengali web text mining in the mentioned phases. The rest of the paper is ordered as follows. Section 2 discusses literary survey of web mining in Bengali and other languages. Section 3 shows the use of Bengali text in the web. In section 4 and 5, we have given an overview of data mining and web mining. Research Opportunities in Bengali Web was conducted in section 6. Some research challenges and recovery is shown in section 7. Future direction for researchers is explained in section 8. We have concluded in section 9 as a summary of our work.

II. Literature Survey

In 1997, researchers at university of Minnesota provide an overview of tools, techniques, and problems associated with Web content mining and Web usage mining. They present taxonomy of Web mining, and place various aspects of Web mining in their proper context [3]. In 1999, researchers at Nayang Technological University, Singapore showed some research issues in Web structure, Web data and Web usage mining as part of their WHOEDA project [4]. In 2009, Zubi [5] has shown the uses of some Web content mining techniques for classifying Arabic text documents. As a part of Web mining research and natural language processing, Bengali blog mining researchers (Das and Bandyopadhyay, 2010) worked on finding emotion in Bengali blog text. For this research, they have developed Bengali WordNet Affect for analyzing emotion [6]. Moreover for opinion mining in Bengali, SentiWordNet for Bangla has been developed in 2010 [7]. As a part Web content mining, researchers of Brazil works on online news data mining where they describe the development and implementation methodology for building an environment of knowledge extraction and seeking information on news sites in Portuguese Language [8]. Recently Indian researcher has worked on online Hindi text mining as part of group research of NI Systems (India) Pvt. Ltd., Bangalore and Microsoft Research Labs, India. They have described a method to mine Hindi-English transliteration pairs from online Hindi song lyrics [9].

III. Uses of Bengali Texts in the Web

Now Bengali language is extensively used in Website development in Bangladesh and West Bengal. Most of the Web developers use various strategies to show Bengali text in their Websites. One strategy is to use Bengali font based and another is using Unicode [10]. However some of the Websites are developed using image, pdf file containing Bengali text. Examples are online newspaper like ‘The Daily e-prothomalo’ [11], ‘The Daily Inqilab’ [12], and ‘Sangbad Pratidin ePaper’ [13], etc. Beside this, in Bangladesh there are several Government regional Websites like www.dcrajbari.gov.bd, blog sites like www.prothom-aloblog.com where Bengali language is used for showing their content. Moreover, there are several websites related to Bengali literature like www.bn.wikisource.org/wiki and www.banglapoems.wordpress.com. Their contents may be helpful to language and literature researchers. Besides these, there are several Websites like www.anandabazar.com, www.somewhereinblog.net, etc. contain bilingual data [14]. Here bilingual data means web text that contains both Bengali and English text.

IV. Data Mining

Before entering into the brief discussion of Web mining we have to know a little about data mining. Data mining is technique of extracting knowledge from structured database. For example, in online shopping database of an electronics company, we can find out the types of customers who are buying a particular product. This information can help marketing people of that company for the advertisement of the product to increase the selling. Mining Object, Spatial, Multimedia, Text, and Web Data are some practical applications of data mining. Due to advancement of data mining Technology, various data mining applications are available in the world market. There are about 53 data mining tools vendor according to Open Directory Project [15]. Tools like ‘BLIASoft Knowledge Discovery software’, ‘11Ants Model Builder’, AdvancedMiner, ‘Angoss Knowledge Studio’, KEEL, etc. are commercially available data mining tools. Some examples of the free and open source data mining software are Orange, RapidMiner, Weka, JHepWork, Rattle, etc.

V. Web Mining

Web data mining refers to extracting knowledge from unstructured Web data. Here knowledge can be different types depending on user’s query. In data mining, knowledge is gathered from database which contains structured data. But in Web data mining information retrieval techniques that are used by different search engines are used to collect the unstructured data in the Web pages and Web logs. Various commercial and open source web mining tools are available in the world market. Bixolabs, Extractiv, Ficstar, FMiner, iWebScraping, etc are the examples of commercial web data mining tool. Some of the open source web mining tools are Bixo, DEiXTo, ScraperWiki, WebSundew, etc.

A. Types of Web Mining

We can classify Web mining into three categories depending on its task. They are Web structure mining, Web content mining and Web usage mining [16]. We will discuss these types with the respect to Web in Bengali language.

1) Web Structure Mining: The mining of the structure of the hyperlinks within the Web itself is called Web structure mining [17]. Actually Web structure is related link analysis of a Web page. A Web page may contain separate links which entitled in Bengali. We will analysis the links and delete the repetition. We can also find out most visited link in the website using search engine data analysis for some query result. We can also find out most luminous [4] website or document for reference in the Bengali web. This may help to generate site map from the links. Web structure mining may also help us to compare two web pages structure using Document Object Model [18].

2) *Web Content Mining:* The process of extracting useful information from the contents of Web documents is known as Web content mining. Content data corresponds to the collection of facts by which a Web page was designed to convey to the users [19]. It may consist of Bangla text, images, Bengali Song, Bengali video, documents in Bengali and so on. Bengali text may include news, news comments, blog, reviews, opinion, sports data, educational data which may be research centric data. We can also work on video and audio data to stop video and audio piracy.

3) *Web Usage Mining:* The discovery of user access patterns from Web usage logs, which record every click made by each user, is known as Web usage mining [16]. Many data mining algorithms are used in Web usage mining. One of the key issues in Web usage mining is the pre-processing of click-stream data in usage logs in order to produce the right data for mining [14]. Here our domain of work is log of all Bengali Websites where Bengali text has been used.

VI. Opportunities of Research in Bengali Web

Throughout Web mining research for the Bengali Language dependent Websites, researchers can acquire useful knowledge. We can specify their opportunity to work by the following subsections.

A. Automatic classification of Bengali news

A lot of Bengali online newspaper is available in Bangladesh and West Bengal. They are publishing a lot of news every day. Government, administration and other related people are anxious about that published news. They have a little time to read all online newspapers and check for their news. By analyzing the semantics of online news, we can find out good news and bad news for them. We can also classify them news according to their types of semantics. Then this news will be helpful to everyone who needs them. Also this classification may also helpful for statistical analysis.

B. Advertise Placement

Online advertisement is now popular at different websites including search engines like Google, Yahoo, Bing, etc. and also in social engine, online newspapers, job, and ecommerce sites. These companies may introduce web text's language dependent advertisement. That means if the web text is in Bengali then advertise text will be in Bengali, if Spanish then advertise text will be Spanish and so on. If someone praises for a product in a social engine or forum or blog or discussion board in Bengali then place a related advertisement in Bengali. If someone criticizes a product in an e-commerce site then place an advertisement of that product's competitor. So here researchers have a great opportunity to work in Web data mining with NLP.

C. Opinion Mining in Bengali

Opinion mining in Bengali that is a part of web data mining and NLP is now ongoing research at Jadappur University, Kokata, India [6][7]. Generally in a forum discussion section reply given by a user wait for the administrator review for checking the text. Because this text may contain slang words which is very much odd looking for publish. So, contrastive opinion on political Bengali text, slang and urban opinion words and phrases from blog and news comments can also derived using opinion mining. This will help to filter the opinion in a forum automatically and thus minimize the administrator's tasks.

D. Customer Reviews of Various Products

In a Bengali e-commerce site, customer has the option to post review for their buying products. Mining customer reviews of a product we can find out different types of reviews like subjective, positive, negative and objective reviews. This can help the respective company to improve the quality of product and promote a better marketing.

E. *Mining Educational Data*

Generally online Bengali newspapers publish different types of preparatory articles for examinees of different public examination. Published articles are not always useful for all students. Students have to sort out their necessary articles by reading newspaper. If there is an automated online tool by which registered students will get their necessary data in their mail. Here researcher may develop a tool which will give Web Mining tools the educational data from online newspaper help to find out educational information for various categories of students. Then these tools can easily be applied to mining the articles of different newspaper.

F. Finding News of Celebrity

Celebrities are always anxious about what news are publishing everyday in newspaper and magazines about them. By reading all newspapers and magazines, it is quite impossible for them to find out all news due to lack of time. So they have to employ other person to search news. Here researchers may develop a tool using web mining technique which will be able to send a text message to the particular celebrity if there is online news about them. This tool will send the message to their registered users only.

G. Finding Popularity of Celebrity

Various newspaper and magazines conduct manual survey to find the popularity of different celebrity. Through web mining technique, researchers may help them to find out a popularity of celebrity by analyzing the news and comments which are published in online newspaper and magazines. Here researchers have to apply knowledge of NLP also.

VII. Challenges and Recovery

During working on web mining for Bengali text, researchers have to face some problems as follows:

- Bilingual nature of web pages
- Use of urban or slang word in text
- Unique font and style problem

To minimize the bilingual nature of we have to apply some tricks to avoid one language text. For avoiding urban or slang words, we have to sort out all the words used in web text and develop a database. To minimize the unique font and style problem we have categorize the web pages according to unique font and style.

VIII. Future Directions

Researchers should come forward to work with Web mining for Bengali language. Online daily newspaper and blogs are very much popular among Bengali people. They publish a lot of article along with news and discussions. It is good news that most of the universities in Bangladesh and India (especially in West Bengal) where technical departments exist are conducting Bengali language research. We think that they will also take a look on this vast area of web mining research in Bengali language. Future researchers have the option to develop new algorithms for web mining which will strengthen our Bengali language resources.

IX. Conclusions

Throughout this article, our goal is to clarify the vast area of web data mining research for Bengali language webs. We hope that researchers will contribute here more and make our knowledge database in Bengali strong. Here we have elaborately described the opportunity of different research areas of Web data mining in Bengali language. We also expect that future directions will inspire the researchers a lot thus help to develop a lot of open source and commercial web mining tools for Bengali language in future.

References

[1] L. Ma, N. Goharian, and A. Chowdhury. Extracting Unstructured Data from Tem-plate Generated Web Document. In Proceedings of the ACM International. Conference on Information and Knowledge Management(CIKM’03), pp. 512–515, 2003

[2] C.N. Hsu and M.T. Dung. Generating Finite-State Transducers for Semi-Structured Data Extraction from the Web. Inf. System, 23(9), pp. 521–538, 1998.

[3] Sanjay Kumar Madria , Sourav S. Bhowmick , Wee Keong Ng , Ee-Peng Lim, Research Issues in Web Data Mining, Proceedings of the First International Conference on Data Warehousing and Knowledge Discovery, p.303-312, September 01, 1999.

[4] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. Web mining: In formation and pattern discovery on the World Wide Web. In International Conference on Tools with Artificial Intelligence, pages 558-567, Newport Beach, 1997. IEEE.

[5] Z. S. Zubi, "Using Some Web Content Mining Techniques for Arabic Text Classification", RECENT ADVANCES on DATA NETWORKS, COMMUNICATIONS, COMPUTERS, 2009.

[6] Das Dipankar and Sivaji Bandyopadhyay. Developing Bengali WordNet Affect for Analyzing Emotion. International Conference on the Computer Processing of Oriental Languages-International Conference on Software Engineer-ing and Knowledge Engineering-2010, USA.

[7] Das, A. and Bandyopadhyay, S. (2010a). SentiWord-Net for Bangla. In Knowledge Sharing Event-4: Task 2: Building Electronic Dictionary, February, 23th to 24th, 2010, Mysore.

[8] Mauricio Onada, Valeria M. Bastos, Cristian K. Santos, Marcello P. A. Fonseca, Victor S. Bursztyn, Alexandre G. Evsukoff, Nelson F. F. Ebecken, Text Mining Applied to Online News, Mecánica Computacional, Volume XXIX. Number 96. Computational Intelligence Techniques for Optimization and Data Modeling (D). pp-9425-9438 November, 15-18, 2010.

[9] Kanika Gupta, Monojit Choudhury, Kalika Bali, Mining Hindi-English Transliteration Pairs from Online Hindi Lyrics, Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), 23-25 May, Istanbul, Turkey.

[10] Unicode Bengali Code Chart - Unicode Consortium, Available: http://unicode.org/charts/PDF/U0980.pdf

[11] The daily eProthomalo, Available: www.eprothomalo.com.

[12] The Daily Inqilab, Available: www.dailyinqilab.com.

[13] Sangbad Pratidin ePaper, Available: www.epratidin.in

[14] Jiang, Long, Shiquan Yang, Ming Zhou, Xiaohua Liu, and Qingsheng Zhu. 2009. Mining Bilingual Data from the Web with Adaptively Learnt Patterns. In Proceedings of ACL-IJCNLP. pp. 870-878.

[15] Open Directory Project, Available: www.dmoz.org

[16] B. Liu, Web Data Mining-Exploring Hyperlinks, Contents, and Usage Data, Springer Series on Data-Centric Systems and Applications, 2007. Pg-7

[17] Miguel Gomes da Costa Júnior, Zhiguo Gong, “Web Structure Mining: An introduction”. Proceedings of the IEEE International Conference on Information Acquisition, 2005, China.

[18] C. Brabrand, Moller, A. M. Ricky, and Schwartzbach, “Powerforms: Declarative client-side form field validation”, World Wide Web 3(4), 205-214 (2000).

[19] J. Srivastava, P. Desikan, and V. Kumar, “Web Mining: Accomplishments and Future Directions,” Proc. US National Science Foundation Workshop on Next-Generation Data Mining (NGDM), National Science Foundation, 2002.