Volume: 07 Issue: 05 Pages: 2882-2886 (2016) ISSN: 0975-0290

A Comparative Study on Serial and Parallel Web Content Mining

Binayak Panda
Dept. of CSE, GIET, Gunupur, Odisha, India
Email: binayak.panda@gmail.com
Dr. Satya Narayan Tripathy
PG Dept. of Computer Science, Berhampur University, Odisha, India
Email:snt.cs@buodisha.edu.in
Dr. Nilambar Sethi
Dept. of CSE, GIET, Gunupur, Odisha, India
Email:dr.nilambarsethi@gmail.com
Om Prakash Samantray
Dept. of CSE, Narasaraopeta Engineering College,Narasaraopet, Andhra Pradesh
Email:om.prakash02420@gmail.com

ABSTRACT

World Wide Web (WWW) is such a repository which serves every individuals need starting with the context of education to entertainment etc. But from users point of view getting relevant information with respect to one particular context is time consuming and also not so easy. It is because of the volume of data which is unstructured, distributed and dynamic in nature. There can be automation to extract relevant information with respect to one particular context, which is named as Web Content Mining. The efficiency of automation depends on validity of expected outcome as well as amount of processing time. The acceptability of outcome depends on user or user's policy. But the amount of processing time depends on the methodology of Web Content Mining. In this work a study has been carried out between Serial Web Content Mining and Parallel Web Content Mining. This work also focuses on the frame work of implementation of parallelism in Web Content Mining.

Key Words**: WWW, World Wide Web, Web Content Mining, Serial Web Content Mining, Parallel Web Content Mining, Scalability, Cost Optimality**

Date of Submission: March 14, 2016 Date of Acceptance: March 24, 2016

1. INTRODUCTION

The concept Web mining came from the concept of Data mining. Data mining is a process of extracting predictive information from large quantities of data, and hence it is data driven. It is also a process of discovering knowledge from a huge data set [1]. The volume of data available in World Wide Web is huge, unstructured or unorganized and also dynamic. It is dynamic because the volume grows day by day. The process of collecting and integrating relevant data with respect to a particular context can be named as Web Mining.

Web mining as a topic offers an unprecedented opportunity and challenge for data mining. It is so due to the following characteristics of the Web [2]:

1. Web data is open to access.
2. The behavior of data is dynamic.
3. The data volume is huge and still growing rapidly.
4. One can find information about almost anything on the Web. Hence it is wide and diverse.
5. Availability of existing data on the Web varies from structured tables to texts, multimedia data (e.g., images and movies), etc.
6. Nature of data in Web is heterogeneous. Redundant information spreads over multiple Web pages. The challenge is collection and integration of irredundant data which may be present at various sources with completely different formats or syntaxes.
7. Information on Web is nested in structure.
8. Web information is linked.
9. Much of the Web information is redundant. This is explored in many Web data mining tasks.
10. Web a virtual society for information and service sharing. Also it is about interactions among peoples or organizations.
11. Dynamic behavior of web is a challenging task in many cases because of the complexity in keeping track of changes.

We can see why the Web is such a fascinating place and why it offers so many opportunities for web data mining.

Volume: 07 Issue: 05 Pages: 2882-2886 (2016) ISSN: 0975-0290

Web Mining is that area of Data Mining which deals with the extraction of interesting knowledge from the World Wide Web. More precisely , Web Con-tent Mining is that part of Web Mining which focuses on the raw information available in web pages; source data mainly consist of textual data in web pages (e.g., words, but also tags); [4].

Web mining can be defined as mining of the World Wide Web (WWW) to find useful knowledge about user behavior, content, and structure of the web. It involves application of data mining techniques on the contents of WWW but is not limited to it [5].
From the Figure 1.1 classification of Web Mining as follows:

Web Structure Mining: is the technique to analyze and explain the links between different web pages and web sites. It mainly focuses on developing web crawlers. It works on hyperlinks and mines the topology of their arrangement.

Web Content Mining: focuses on extracting knowledge from the contents or their descriptions. It involves techniques for summarizing, classification and clustering of the web contents. It can provide useful and interesting patterns about user needs and contribution behavior.

Web Usage Mining: It focuses on digging the usage of web contents from the logs maintained on web servers, cookies logs, application server logs etc. It works on how and when user moves from one type of content to other. Thus, it can provide association between different contents.

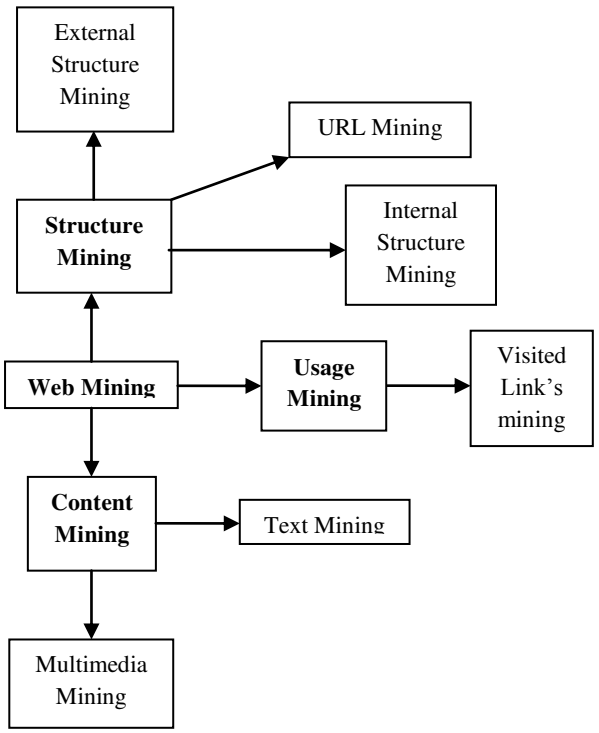

Figure1.1 Classification of Web Mining

2 STUDIES ON APPROACHES OF WEB CONTENT MINING

Web Content Mining is the process of extracting useful information from the contents of Web documents. It may consist of text, images, audio, video information which is used to convey to the users about that documents [3]. Text mining and its application to Web content has been the most widely researched. Some of the research issues addressed in text mining are, topic discovery, extracting association patterns, clustering of web documents and classification of Web Pages. Web content mining issues in term of Information Retrieval (IR) and Database (DB) view verses data representation, method and application categories is discuss and summarized in . While extracting the knowledge from images - in the fields of image processing and computer vision - the application of these techniques to Web content mining has not been very rapid.

Web Content Mining can be carried out in any of following approaches:

- Serial Web Content Mining
- Parallel Web Content Mining

Figure 2.1 shows a model of Serial Web Content Mining and Figure 2.2 shows a model of Parallel Web Content Mining.

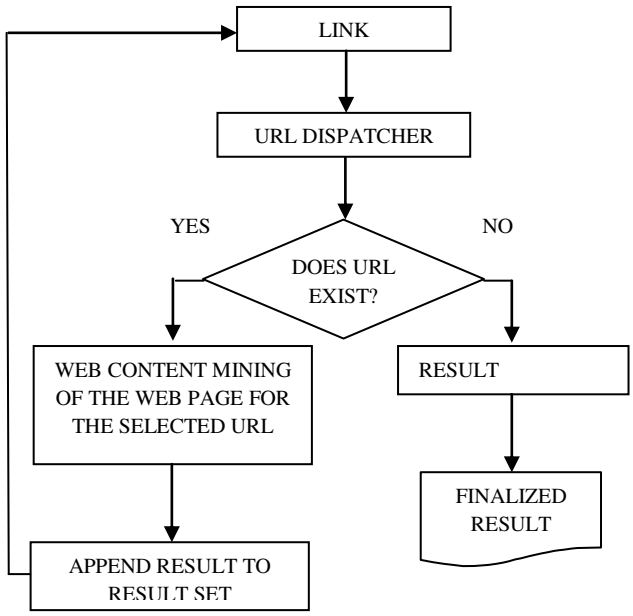

Figure2.1. Model of Serial Web Content Mining

Volume: 07 Issue: 05 Pages: 2882-2886 (2016) ISSN: 0975-0290

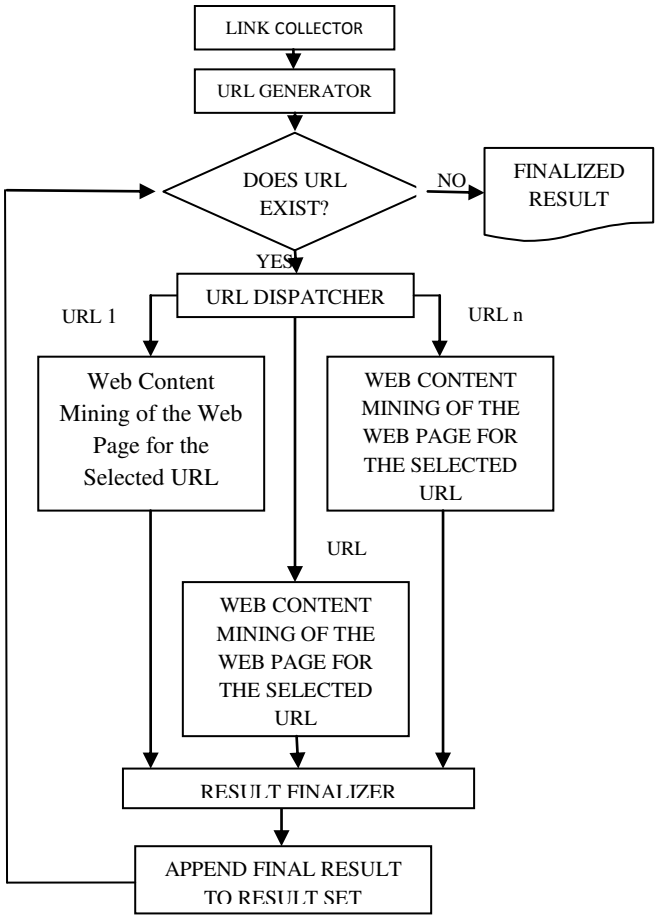

Figure2.2. Model of Parallel Web Content Mining

3. PROCESSING MECHANISM

3.1 SERIAL PROCESSING

With reference to figure 2.1 the processing node will process on a link dispatched by the url dispatcher. It is observed that dispatch time and processing time will be constant.

For one link: Assume D(t) the dispatch time and P(t) the processing time.

For n links: Total time W=n*(D(t) + P(t))=n*c=O(n).
c is a positive constant.

3.2 PARALLEL APPROACHES

With reference to figure 2.2 several interconnection networks for processing nodes like linear array, star, mesh and hypercube. With consideration of communication cost and topology overhead, the authors have chosen hypercube interconnection network for the processing node representation and further study [9].

The k-dimensional hypercube, or k-cube, is a general purpose interconnection network in parallel processing and has been widely used. It has 2k nodes. Two nodes are neighbors iff their k-bit addresses differ in a single bit. The k-cube has small diameter, equal to k. In each steps each node communicate with neighbors for the message passing and receiving which require log n step to broadcast the own message to all other nodes [6][7][8].

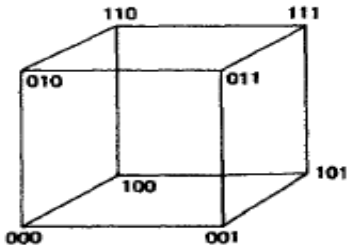

Figure3.1. Hypercube Representation of 8 Processing Nodes

3.2.1 Working Principle of DISPATCHER in Model of Parallel Web Content Mining

The DISPATCHER is responsible for dispatching the links to available processing elements to achieve parallelism.

During the parallel evaluation we have three cases:

Case 1: Number of URL and processing node are same. Each processing node will get one URL which can be updated the constant time. Hence Parallel time complexity is Θ(1).

Case 2: Number of URL less then processing node
Here some processing node will get one URL where as some will be unoccupied hence Parallel time complexity is also Θ(1).

Case 3: Number of URL more than processing node
Here each processing node will get n/p URL, where n is number of URL and p is the number of processing node.

If a[i] is the array of URL the processor p_i will fetch a[i] %p URL. Hence each p_i will execute maximum n/p URL.

Hence parallel time = Θ(n/p).

For all above cases, If any processing element finds the required result, the other processing element should not compute further, hence the communication among required processing elements is required which will take O (log n) time.

Hence the total parallel time is $T_p = (\frac{n}{p} + 2\log p)$.

Volume: 07 Issue: 05 Pages: 2882-2886 (2016) ISSN: 0975-0290

3.2.2 Working Principle of RESULT FINALIZER in Model of Parallel Web Content Mining

The RESULT FINALIZER is responsible for recognizing the expected result and communicating all other processing elements.

All_to_all_broadcast (my_id, my_msg, k, result) [9]
//my_id: Unique Id(k bit) of Processing element
//my_msg: The Message to be broad casted
//k: dimension of the Hyper Cube for 2^k Processing elements
//result: Own result with received result

1. result=my_msg
2. for i:=0 to k-1
   a. partner:=my_id XOR 2^i
   b. send result to partner
   c. receive msg from partner
   d. result= result U msg

4. PERFORMANCE EVALUATION OF BOTH APPROACHES

Sequential time W= Θ(n).

Parallel time $T_p = (\frac{n}{p} + 2\log p)$

Hence speed up ratio is $S = \frac{W}{T_p} = \frac{n}{\frac{n}{p} + 2\log p}$.

Cost for parallel computation is $P*T_p$=

$(\frac{n}{p} + 2\log p)p$.

The Overhead value $(p.T_p - W)$ is

$(\frac{n}{p} + 2\log p)p - W = 2p\log p = O(p\log p)$

as long as $n = \Omega(p\log p)$

The cost Θ (n) is the serial time complexity so the parallel evaluation is cost optimal as the overhead function does not asymptotically exceed the problem size. As we can say the parallel system is cost optimal if the product of number of processing element and the parallel execution time is proportional to the execution time of fastest known sequential algorithms on a single processing element.

4. CONCLUSION

Very frequently, programs are intended and tested for lesser problem size and fewer processing node. Though the real problems these programs are intended to solve are much larger and the machine surround large processing node. Whereas the code development is simplified by using scaled down version of machine and the problem, their accuracy and performance is much more difficult to establish based on scale down system. We investigated keeping processing element fix, if problem size augmented the overhead function T_0 grow sub linearly with respect to Ts, hence increasing efficiency. It is possible to keeping the efficiency fixed by increasing both problem size and processing unit. Efficiency can be measured by

$$E = \frac{S}{p} = \frac{T_s}{pT_{p\ s}}$$

$$= \frac{1}{1 + \frac{T_0}{T_s}} = \frac{1}{1 + \frac{2p\log p}{n}}$$

Efficiency as a function with respect to n number of URL and p number of processing node

n	p=1	P=4	P=8	P=16	P=32
64	1	**0.8**	0.57	0.333	0.2
192	1	0.92	**0.8**	0.6	0.4
512	1	0.97	0.91	**0.8**	0.6
1280	1	0.97	0.96	0.95	**0.8**

Table-1: Efficiency as a function with respect to URL and processing elements.

From the table-1, the efficiency of adding 64 numbers by using 4 processing unit is 0.80. if the number of processing unit increase to 8 and the size of problem scaled to 192 the efficiency remain 0.80. This ability to maintain efficiency at a fixed value by simultaneously increasing the number of processing element and size of the problem is called scalable. Hence this parallel approach is also scalable.

5. REFERENCES

[1]. Samia Jones and OmPrakash K. Gupta "WEB DATA MINING: A CASE STUDY" – Communications of IIMA – 2006 Volume 6 Issue 4

[2]. Bing Liu and Kevin Chen-Chuan Chang "EDITORIAL: SPECIAL ISSUE ON WEB CONTENT MINING" - http://www.cs.uic.edu/~liub/publications/editorial.pdf

[3]. Pravin M. Kamde, Dr. Siddu. P. Algur "A SURVEY ON WEB MULTIMEDIA MINING" - The International Journal of Multimedia & Its Applications (IJMA) Vol.3, No.3, August 2011 PP 72 - 84

[4]. Federico Michele Facca and Pier Luca Lanzi "RECENT DEVELOPMENTS IN WEB USAGE MINING RESEARCH" - Springer-Verlag Berlin Heidelberg 2003 - DaWaK 2003, LNCS 2737, pp. 140–150, 2003

Volume: 07 Issue: 05 Pages: 2882-2886 (2016) ISSN: 0975-0290

[5]. Aarti Singh " AGENT BASED FRAMEWORK FOR SEMANTIC WEB CONTENT MINING" – International Journal of Advancements in Technology - Vol. 3 No.2 (April 2012) ISSN 0976-4860 Page 108 – 113

[6] Sotirios G. Ziavras *, Arup Mukherjee "Data broadcasting and reduction, prefix computation, and sorting on reduced hypercube parallel computers" ELSEVIER Parallel Computing 22 (1996) P:595-606

[7] D. P. Bertsekas, C. Ozveren, G. D. Stamoulis, P. Tseng, and J. N. Tsitsilkis "Optimal Communication Algorithms for Hypercubes" Journal of parallel and distributed computingI1, 263-275 (1991)

[8] E. Abuelrub, " Data Communication and Parallel Computing on Twisted Hypercube" Proceedings of the World Congress on Engineering 2007 Vol I WCE 2007, July 2 - 4, 2007, London, U.K.

[9] Introduction to parallel computation Ananth Grama, Anshul Gupta, George Karypis, Vipin Kumar Publisher: Pearson; 2 edition (January 26, 2003)

BIOGRAPHIES

Binayak Panda has received a bachelor's degree in Computer Science and Engineering from BPUT Odisha in the year 2005. In the year 2010 he has received a Master of Technology degree in Computer Science and Engineering from BPUT Odisha. Currently, he is pursuing Ph.D. in Computer Science from Berhampur University, Odisha. He has 2 years of industry experiences in the field of Software testing and maintenance. His interested areas of research are Malware Analysis, Data Mining, Software Engineering and Real Time System. He is a life time member of ISTE.

Dr. Satya Narayan Tripathy received his M.C.A. and Ph.D. degrees in Computer Science from Berhampur University, Berhampur, Odisha, India in the years 1998 and 2010, respectively. He has been teaching in the Department of Computer Science, Berhampur University since 2011. Currently, he is a Lecturer in the Department of Computer Science, Berhampur University. Dr. Tripathy serves on the advisory boards of several organizations and conferences. He is a Life Member of Computer Society of India (LMCSI), Life Member of Orissa Information Technology Society (LMOITS) and Member of several professional bodies. His research interests include computer network security, wireless ad hoc network, network security in wireless communication and data mining.

Dr. Nilambar Sethi has received a master of technology degree from Utkal University, Odisha in Computer Science and awarded Ph.D from Berhampur University, Odisha in Computer Science. Currently he is working as Assoc. Prof. in Dept. of CSE at GIET Gunupur. He has 12 years of teaching experience. His interested areas of research are fuzzy logic, Automata Theory and Data Mining.

Om Prakash Samantray got the M.Tech degree in Computer Science & Engineering from Biju Patnaik University of Technology, Odisha, India in 2010. Currently, he is pursuing Ph.D. in Computer Science from Berhampur University, Odisha, India. His research interests include information security, Computer network security, Data warehousing & mining and big data.