WEB DATA EXTRACTION SYSTEM USING BOYER MOORE STRING PATTERN MATCHING ALGORITHM

B. Umamageswari[1], Dr. R. Kalpana[2], V. Archana[3]
[1]PhD Scholar Dept. of CSE PEC, [2]Professor Dept. of CSE PEC, [3]M.Tech Dept. of CSE PEC

Abstract
Web data extraction is one of the active research areas in the recent years. There are voluminous number of online-data items which can be extracted effectively and efficiently using web data extraction techniques. The extraction consists of many phases namely, web crawling, storing the organized data, choosing data structure for data storage pattern search algorithm and retrieve the data. In this paper, a new string pattern matching algorithm is used for web data extraction to extract the data efficiently. It is evident from the results that the proposed Boyer Moore string pattern matching algorithm outperforms the existing pattern matching algorithm in terms of the metrics namely precision and recall.
Keywords: Web Data Extraction; wrapper; string pattern matching; templates

I. INTRODUCTION

The quantity of information available in World Wide Web is beyond our imagination. The information is present diverse formats such as text, video, image etc. [1]. It is difficult to extract and edit data from the web manually. So the concept of web data extraction system was introduced. Web data extractor is a software system that automatically extracts data from a web site. After extracting data from the web pages, it is stored in database for use in other applications. Web data extractor contains wide range of applications such as business intelligence, product intelligence, data analytics, data mash ups, meta- search, meta query etc.

There are two types of data extraction techniques available in the literature. They are manual and automatic web data extraction techniques. In manual technique, user manually writes the programs called wrappers to extract data from the web pages. Manual extraction technique uses some built in rules to extract the data. This technique is working based on some prior knowledge of the format of the web page. The examples of manual data extraction techniques are TSIMMIS[2], Minerva[3], Web-OQL[4], W4F[5] and XWRAP[6]. They are many shortcomings associated with this technique which led to the era of automatic web data extraction techniques. The automatic web data extraction techniques[7] are classified into supervised techniques, semi-supervised techniques and unsupervised techniques. In Supervised techniques, wrapper construction output and extraction rules are based on the training sample provided by the designer of the wrapper. Some of the supervised techniques are WIEN[8], SoftMealy[9] and Stalker[10]. IEPAD[11] and OLERA[12] are examples of semi-supervised techniques. Unsupervised techniques learn rules and extract as much as prospective data as they can, and the user then gathers the relevant data from the output.

Some of the unsupervised techniques are RoadRunner[13], EXALG[14], FivaTech[15] and Trinity[16]. RoadRunner uses AMCE (Align, Collapse, Match and Extract) technique in order to induce wrapper from a set of web pages by identifying similarities and difference between them. EXALG technique is used for extracting the structured data from a collection of web pages generated using common template. It consists of two stages such as Equivalent Class Generation stage (ECGM) and analysis stage. This technique performs page by page data extraction. FivaTech first identifies node in the input DOM tree that have a similar structure and then aligns their children and mines repetitive and optional pattern to create the extraction rule.

Trinity performs record level data extraction using Knuth-Pratt-Morris[17] pattern matching algorithm. This pattern matching algorithm uses word by word matching of the text, which takes longer time to find the text.

In this paper, Boyer Moore string pattern matching algorithm[18] is used for finding pattern in the extracted data. This algorithm successively aligns pattern with text and then checks whether pattern matches the opposing character of text. After the check is complete, pattern is shifted right relative to the text. Boyer Moore algorithm outperforms the other string pattern matching algorithm with respect to efficiency.

Section II gives more details about web data extraction techniques, Section III gives The framework for web data extraction based on Boyer Moore string pattern matching algorithm, Section IV gives Experimental study and Result Analysis, Section V gives concludes and reference of the paper.

II. RELATED WORKS

RoadRunner[13] technique uses ACME (Align, Collapse, Match and Extract) techniques in order to induce wrapper from a set of web pages by identifying similarities and differences between them. The outcome of the step is a wrapper which is used as a template during data extraction. The earliest approaches are manual techniques in that users have to write the wrapper program which is laborious task.

EXALG[14] technique is used for extracting the structured data from a collection of web pages generated using common template. EXALG consists of two stages such as equivalent class generation stage (ECGM) and analysis stage. ECGM stage finds the sets of tokens having the same frequency of occurrence in every page which are known as Equivalence Classes. EXALG[14] retains only the equivalence classes that are large and whose tokens occur in a large number of input pages. Such type of equivalence classes are known as LFEQs which represents the template.

FivaTech[15] is a page - level web data extraction technique, which automatically detects the schema of a website. FivaTech[15] proposes a new structure, called fixed/variant pattern tree. The fixed or variant pattern tree is used for to identify the template and find the database schema. This technique is the combination several techniques such as alignment and pattern mining. FivaTech[15] consists of two stages. First stage is merging input DOM trees to construct the fixed/variant pattern tree. Second stage corresponds to identification of schema and pattern.

Trinity[16] technique for data extraction is based on the hypothesis that templates are shared across web pages and when templates are removed, we get the needed data. The algorithm for construction of trinary tree requires at least two web pages. It compares the string representation of web pages to find the longest matching substring. The portion of the string before the matching pattern is considered as prefix. The portion of the string in between the matching pattern and its next occurrence is considered as separator and the portion of the string after the matching pattern is considered as suffix. At each step, a node is expanded to the above three nodes. Once the trinary tree is constructed, regular expression is deduced and then, it is used for extracting data.

The comparison of various web data extraction systems is shown in table 1.

Sl. no	Methods	Human Intervention	Limitation	Level of extraction
1	SoftMealy[9]	Supervised	SoftMealy is it not able generalize inspect separator	Record-level
2	IEPAD[11]	Semi-supervised	IEPAD is that could not handle complex and nested structure data	Record-level
3	OLERA[12]	Semi-supervised	It is sensitive to the ordering of the input information	Record-level
4	Road Runner[13]	Un-supervised	Number of error in the input document affects it effectiveness	Page- level
5	FivaTech[15]	Un-supervised	Searching the longest pattern is time consuming process	Page-level
6	EXALG[14]	Un-supervised	It is not clear whether EXALG can work on malformed input document or not	Page- level
7	Trinity[16]	Un-supervised	It does not work well with templates having alternating formatting for same data	Record-level

Table1. Comparison of Web Data Extraction Techniques

III. WEB DATA EXTRACTION SYSTEM USING STRING PATTERN MATCHING ALGORITHM

The web data extraction system (WDES) consists of Trinity centralized data extractor, splitting module and database. Web pages are given as input to the data extractor. Formatting Tags are removed and It is passed on to the splitting module where the text will be segregated into three parts namely prefix node, separator node and suffix node. The architecture diagram of web data extraction system is presented in Fig1.shows the modules of web data extraction system namely,

a. Extracting website using web crawling
b. Generating trinary tree
c. Store the extracted information in database.
d. Search and retrieve the data records.

Web pages are considered as sequence of string and string matching algorithm is applied to find the shared pattern. Using this shared pattern, trinary tree is created consisting of prefix node, separator node and suffix node. In Trinity[16], the authors have used Knuth-Pratt-Morris pattern searching algorithm [KPM] for finding shared pattern which is time consuming. In the proposed work, Boyer Moore algorithm is considered and experimental results prove that it is much more efficient compared to Knuth Morris Pattern matching algorithm.

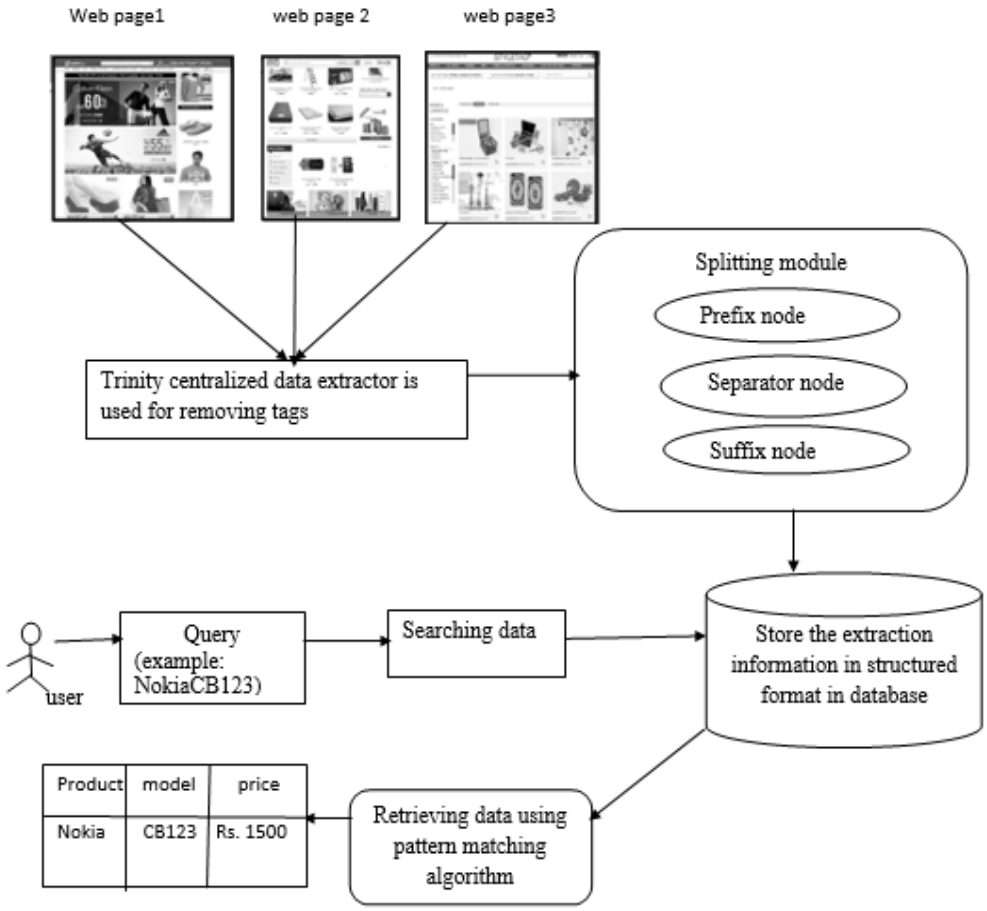

Fig.1. Web Data Extraction System Using String pattern matching algorithm.

The Boyer Moore string searching algorithm is an efficient algorithm which is the standard benchmark for string search in literature. It aligns P(pattern) with T(text) and then check whether P matches the opposing characters of T. After the checking operation is complete, P is shifted to right relative to T. This algorithm performs the comparisons from right to left.

BOYER_MOORE_MATCHER (T, P)

1. Sigma = alphabet in use
2. T = Search string (text)
3. P = Pattern
4. N = length[T]
5. M = length[P]
6. Compute_Last_Occurrence_Function (P, M, Sigma);
7. For each character a in sigma
8. initialize L[a] = 0
9. for each j until end pattern
10. pattern = j
11. return L
12. Compute_Good_Suffix_Function (P, M);
13. Compute_Prefix_Function(P)
14. Reverse(P)
15. Compute_Prefix_Function(P')
16. For each i until end pattern

17. suffix_distance = M – Compute_Prefix_Function(M)
18. for each i until end pattern
19. i = M - Compute_Prefix_Function(M)
20. if suffix_distance> j - Compute_Prefix_Function(M)
21. suff_distance= j – Compute_Prefix_Function(M)
22. end

The data records which we get after tree generation are stored in database which can be retrieved using user query.

IV. EXPERIMENTAL RESULTS

The result is completed on an Intel core i3 processor with clock speed 1.80GHZ and 4GB RAM running on Windows 8.1. The program is written in Java and the results were analyzed. Real world websites such as Flipkart, Amazon, Shop Clues and Snapdeal are used as input for carrying out extraction task. WDE based on KMP and BM algorithm are compared based on precision, recall values and extraction times.

The chart in fig.2 represents the extraction time occurred in implementing the algorithm. The X axis represented the websites considered for web data extraction namely Flipkart, Amazon, Shopcules and Snapdeal. The Y axis gives the details of extraction time occurred in millisecond.

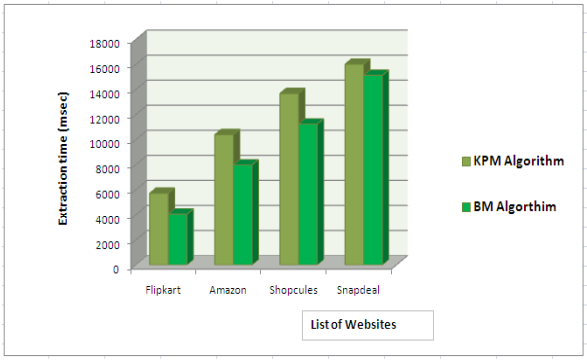

Fig.2. Comparison of KPM and BM string pattern matching algorithms

The results show that the extraction time for Boyer Moore string pattern algorithm is less than that of Knuth-Pratt-Morris algorithm for all input websites. The reason for reduced extraction time is it works the fastest when the alphabet is moderately sized and the pattern is relatively strong.

The web data system is also evaluated using the metrics namely precision, recall and F1-measure. Precision means (P) is defined as the number of records extracted correctly to the total number records extracted. Recall(R) is defined as the number of records extracted correctly to the total of records available. F1-measure is defined as weighted harmonic mean of precision and recall.

$$\text{Precision } (P) = \frac{tp}{tp+fp}$$

$$\text{Recall } (R) = \frac{tp}{tp+fn}$$

$$\text{F-Measure } (F1) = \frac{2(p*R)}{P+R}$$

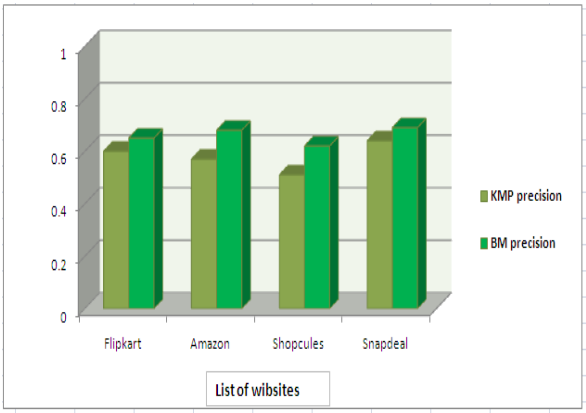

Fig.3. Comparison of KMP and BM precision values

Fig.3. shows comparison of KMP and BM based on precision values. When compared to the existing system precision value obtained is improved in the proposed work.

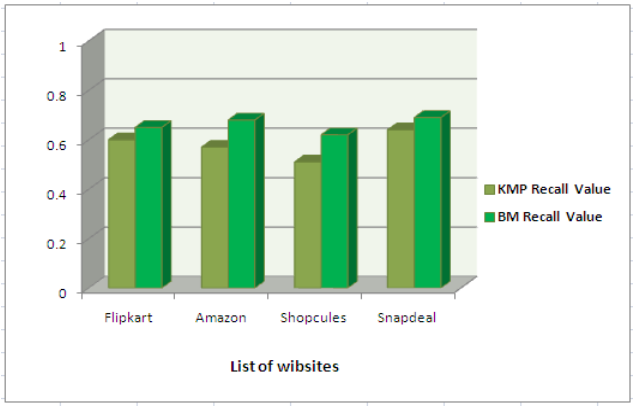

Fig.4. Comparison of KMP and BM Recall values

Fig.4. shows comparison of KMP and BM based on recall values. When compared to the existing system recall value is better for the proposed work.

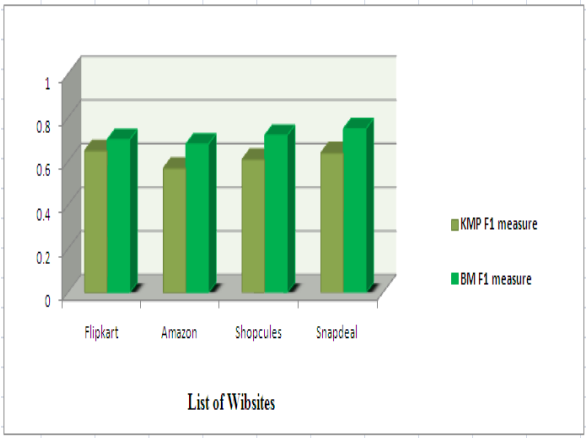

Fig.5. Comparison of KMP and BM F1 measure

Fig.5. shows the graph of F1 measure for KPM and BM techniques. The comparison is done based on the four websites and their results shown in graph. The proposed system is found to be effective and efficient compared to the existing system

V. CONCLUSION

This paper compares web data extraction system using Boyer Moore with String Pattern Matching algorithm. The modified framework of web data extraction system is experimented using real world websites. Existing and proposed systems are compared using Precision, recall and F measure performance indicators. The results show that the proposed system gives better result with respect to the performance indicators.

REFERENCES

[1] V. Crescenzi and G. Mecca, "Automatic information extraction from large websites," J. ACM, vol. 51, no. 5, pp. 731–779, Sept. 2004.

[2] Hammer, J., McHugh, J., and Gracia-Molina, H., Semi structured data: The TSIMMIS experience in proceedings of the First East-Europen Symposium on Advances in Databases and Information Systems (St. Petersburg, Russia, 1997), pp. 1-8.

[3] Crescenzi, V., Mecca, G. Grammers have Exceptions. Information Systems 23, 8 (1998), 539-565.

[4] B Motik, PF Patel-Schneider, B Parsia, C Bock, A Fokoue, P Haase,OWL 2 web ontology language: Structural specification and functional-style syntax W3C recommendation 27 (65), 159

[5] Sahuguet, A., Azavant, F., Building Intelligent Web Applications using Lightweight Wrappers, Data and Knowledge Engineering, Volume 36, Issue 3, 2001, pages 283-316.

[6] XWRAP: An XML-enabled wrapper construction system for web information sources. L Liu, C Pu, W Han. Data Engineering, 2000. Proceedings. 16th International Conference on, 611-621

[7] C.-H. Chang, M. Kayed, M. R. Girgis, and K. F. Shaalan, "A survey of web information extraction systems," IEEE Trans. Knowl. Data Eng., vol. 18, no. 10, pp. 1411–1428, Oct. 2006.

[8] Kushmerick, N.,Weld, D. and Doorenbos, R., Wrapper Induction for Information Extraction, Proceedings of the Fifteenth International Conference on Artificial Intelligence (IJCAI), pp. 729-735, 1997.

[9] C.-N. Hsu and M.-T. Dung, "Generating finite-state transducers for semi-structured data extraction from the web," Inform. Syst.,vol. 23, no. 8, pp. 521–538, Dec. 1998.

[10] V. Kovalev, S. Bhowmick, S. Madria, HW-STALKER: a machine learning-based system for transforming QURE-Pagelets to XML, Data and Knowledge Engineering Journal 54 (2) (2005) 241–276

[11] C.-H.Chang and S.-C. Lui, "IEPAD: Information extraction based on pattern discovery," in Proc. 10th Int. Conf. WWW, Hong Kong, China, 2001, pp. 681– 688.

[12] C.-H. Chang and S.-C.Kuo, "OLERA: Semi supervised web-data extraction with visual support," IEEE Intell. Syst., vol. 19, no. 6,pp. 56–64, Nov./Dec. 2004.

[13] V. Crescenzi, G. Mecca, and P. Merialdo, "Road runner: Towards automatic data extraction from large web sites," in Proc. 27th Int. Conf. VLDB, Rome, Italy, 2001, pp. 109–118.

[14] A. Arasu and H. Garcia-Molina, "Extracting structured data from web pages," in Proc. 2003 ACM SIGMOD, San Diego, CA, USA, pp. 337–348.

[15] M. Kayed and C.-H. Chang, "FiVaTech: Page-level web data extraction from template pages," IEEE Trans. Knowl. Data Eng., vol. 22, no. 2, pp. 249–263, Feb. 2010.

[16] Hassan A. Sleiman and Rafael Corchuelo," Trinity: On Using Trinary Trees for Unsupervised Web Data Extraction", IEEE trans. Knowl, Data Eng., vol.26, no. 2, pp.1544-1556, June 2014.

[17] Knuth, Donald E., James H. Morris, Jr, and Vaughan R. Pratt. "Fast pattern matching in strings." SIAM journal on computing 6.2 (1977): 323-350.

[18] Tarhio J., Ukkonen E. (1990) Boyer-Moore approach to approximate string matching. In: Gilbert J.R., Karlsson R. (eds) SWAT 90. SWAT 1990. Lecture Notes in Computer Science, vol 447. Springer, Berlin, Heidelberg.