International Journal of Engineering & Technology, 7 (1.2) (2018) 130-134

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Design and development of text extraction and retrieval using style of documents in web searching

S. Balan [1] *, P. Ponmuthuramalingam [2]

[1] *Post Graduate & Research Department of Computer Science, Government Arts College (Autonomous), Coimbatore, Tamilnadu, India*
[2] *Controller of Examiniations & Associate Professor Post Graduate & Research Department of Computer Science, Government Arts College (Autonomous), Coimbatore, Tamilnadu, India*
**Corresponding author E-mail: Balan.sethuramalingam@gmail.com*

Abstract

This research focuses on study and extraction of web pages and documents are returned from goggle search engine. The useful task of web is to exactly match the accurate information. That information are categorized into many ways such as manual, structured, semi-structured texts and images. Query Result Records (QRR's) is used to extract the text information from the different type of documents. Data region is used to identify the actual segmentation step and the domain of documents contains suffix and prefix. Time compared to the existing pruning and other techniques are more efficient in manner. We analyze the different type of alignments in this paper and propose a new technique for alignment retrieval to find precision and recall evaluating the retrieval performance.

Keywords: *Web Search; Text Extraction; Data Alignment; Data Retrieval.*

1. Introduction

Data mining refers to process of extracting hidden knowledge from the large amount of raw data. Web mining is defined as data mining techniques that automatically extract the useful information from the World Wide Web. Web contains huge amount of information's are stored in millions and billions of pages. Web content mining contains image, audio, text and video are stored in tables, structured and unstructured data to extract the useful information to generate patterns various issues arises.

There are various existing techniques are available to download the web pages and stored in database. One of the common methods is document object model. Some of the classifications of web data extraction techniques are namely as follows. Manual extraction techniques are Minerva (Storage Systems), TSIMMIS (The Stanford-IBM Manager of Multiple Information Sources) and WebOQL. Supervised techniques are Rapier, SRV (Learner Rule), and WHISK (Highly structured to free text).

Semi-Supervised techniques are IEPAD (Information Extraction Based on Pattern Discovery) and OLERA (Online Extraction Rule Analysis). Un-Supervised techniques are Roadrunner, EXALG (Extracting Structured data from Web Pages), FIVATECH (Automatically detect scheme of website) and Trinity [11].

In recent years, the web data extraction problem has more attention and mainly focused on HTML (Hyper-Text markup Language) pages [7]. The page blocks are identified by using html parser, Dom Tree, support vector and vision based page segmentation. The vision based data extractor is classified in to vision based data record extractor and vision based data item extractor. The visual page of web page is mainly focused on font, position feature, Layout Feature, Appearance feature and content feature [10].

The three components of web page information extraction are pattern matching, rule generator and extractor. Two data extraction problems are extraction given a single list page and extraction given multiple pages. String edit distance and tree edit distance are two main techniques used to fine the template patterns in the HTML encoding strings. Web page automatic annotation consists of three phases alignment phase, annotation phase and annotation wrapper generation phase [1].

The four steps involved for web page extraction in this proposed method given as:

1) Enter a text and extract the information from the web and stored as HTML file.
2) Extract the HTML file to Text file and stored in separate domains
3) Text item separation and data alignment of same semantic information together.
4) Generate the result from database stored in different text domains based on alignment techniques to find the precision and recall value.

The main contribution of paper are categorized as follows 1) A detailed study on web information extraction techniques 2) Discussed about the existing techniques of information extraction and alignment methods 3) proposed methods for extracting information via alignment methods 4) experimental results and discussion of performance evaluation 5) future directions are mentioned for alignment techniques. At last summarizes this paper.

2. Literature review : survey

Template Extract from Web Pages (TEW) is used to extract templates from web pages. Multiple web pages are constructed by using DOM (Document Object Model) tree and it is divided in to multiple blocks to perform hierarchical clustering to extract templates from the database to identify non-content blocks. Vision Based Page Segmentation Algorithm (VIPS) is used to find a given input document, DOM tree is constructed, Extract Blocks, de-

Copyright © 2018 S. Balan, Dr. P. Ponmuthuramalingam. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

tect separator, weight calculation and construct visual content structure.

Support Vector Machine (SVM) is used to identify the classifying of non-content blocks. Hierarchical clustering is used to find the similarity measure value of DOM tree [9]. There are two kinds of automatic page annotation search results namely tabling annotator and query based annotator [13]. The various existing methods of page annotation result is vide: A Vision based approach for Deep-Web Based extraction, on Deep Annotation, Annotating Structure data from the deep web, Data Unit Similarity, Data Content Similarity, Presentation Style Similarity, Data Type Similarity, Tag Path Similarity, Adjacency Similarity and annotation Wrapper [15].

Text Entity Alignment for methodology of entity linking is used to find surface similarity, candidate attribute and contextual similarity [14]. Information extraction of web page article model is based on web article model, visual; feature, DOM Tree structure and characteristic, page block algorithm from top to bottom. Body identification method of web article is review stage, text block, block node based on page block algorithm.

The design for identifying the web text is analyzing the characteristic of text, single pass, and web article text extraction [12]. Automatic annotation of query result from web database is classified into automatic annotation, decorative tag detector, cluster based shifting and Simple probabilistic , Tag Path(TP), Combining tag and vale Similarity (CTV'S), Local interface Schema (LIS), New Combining tag and vale Similarity (NCTV'S), Integrated Interface Schema(IIS) [8], [9] and [12].

Automatic annotation and wrapper generation is based on six annotators namely schema value, table, query based, frequency based, in-text prefix/suffix, common knowledge annotator. Wrapper generation, frequent item set generation [16]. The process flow of data extraction and annotation is extract data from web, alignment process for web data, finding similar data based on alignment, group the similar data into cluster and display the result in annotation wrapper [6].

Some of the data extraction tools are 2semi-automatic approach, W4F, XWRAP, World Wide Web Wrapper Factory, DEPTA [2], [4] and [5].

3. Methods

Retrieval Information with format and style of documents method, to extract the query result records (QRRs) from a query result page p.

1) Record extraction identifies the QRRs in p and involves two sub steps: data region identification and the actual segmentation step.
2) Record alignment aligns the data values of the QRRs in p into a table so that the data values for the same attribute are aligned into the same table column.

Given a query result page, the Tag Tree Construction module first constructs a tag tree for the page rooted in the <HTML> tag. Each node represents a tag in the HTML page and its children are tags enclosed inside it. Every internal node n of the tag tree has a tag string ts_n, which includes the tags of n and all tags of n's descendants and a tag path tp_n, which contains the tags from the root to n.

Next, the Data Region Identification module identifies all possible data regions, which usually have dynamically generated data, approach of top down starting from the root node. Record Segmentation module is segments the identified data regions into data records according to the tag patterns in the data regions. The Data Region Merge module merges the data regions containing similar records in given the segmented data records. At last, the Query Result Section Identification module selects one of the merged data regions as the one that contains the QRRs.

Starting from the root of the query result page tag tree, data region identification algorithm is applied to a node n and recursively to its children n_i, i = 1 . . . m as follows:

1. Compute the similarity sim_{ij} of each pair of nodes n_i and n_j, i, j = 1 . . . m and i ≠ j, using the node similarity calculation method. The data region identification algorithm is recursively applied to the children of n_i only if it does not have any similar siblings. The recognized similar nodes with the same parent form a data region. In this step, multiple data regions may be identified.
2. Segment the data region into data records using the record segmentation algorithm.

A novel three-step data alignment method is used for QRR alignment which is performed by that combines tag and value similarity.

1) Pair wise data value alignment aligns the data values in a pair of QRRs to provide the evidence for how the data values should be aligned among all QRRs.
2) Comprehensive data value alignment aligns the data values in all the QRRs.
3) Combination of data value alignment identifies the nested structures that exist in the QRRs.

3.1. Pair wise data value alignment

The Pair wise data value alignment algorithm is based on the observation that the data values belonging to the same attribute usually have the same data type and may contain similar strings, particularly since the QRRs are for the same query. For example, both QRR 1 and 2 have a real number for the Price attribute and both contain the value "Queen" for the Size attribute. Given two QRRS $r_1 = \{f_{11}. \ . \ . f_{1m}\}$, where f1i refers to the ith data value of r1, and $r_2 = \{f_{21}. \ . \ . f_{2m}\}$, we first calculate the data value similarity, s_{ij}, between every pair of data values in r_1 and r_2 using the method.

Then, a pair wise data value alignment of the two QRRs is performed to determine whether the paired data values belong to the same attribute related to the calculated data value similarities. That means a pair wise alignment of r_1 and r_2 is composed of a set of data value alignments and each of which assumes that the corresponding data values from r_1 and r_2 belong to the same attribute. The input to the pair wise alignment algorithm is a pair of QRRs r_1 and r_2 and an m × n two-dimensional similarity array S in which each element s_{ij} contains the similarity value between f_{1i} and f_{2j}. Each QRR includes two kinds of information: the text string for the ith value and the tag path for the ith value.

If f_{1i} and f_{2j} have the same tag path means only one of the following three data value alignments is possible.

1) The first (i - 1) values of r1 can be aligned with the first (j - 1) values of r2 plus the data value alignment between f1i and f2j, which has the summing similarity score $L_{(i-1)(j-1)} + s_{ij}$.
2) f_{1i} can be ignored and the first (i - 1) values of r_1 can be aligned with the first j values of r_2, which has the summing similarity score $L_{(i-1)j}$.
3) f_{2j} can be ignored and the first i values of r_1 can be aligned with the first j - 1 values of r_2, which has the summing similarity score $L_{i(j-1)}$.

3.2. Comprehensive data value alignment

Given the pair wise data value alignments between every pair of QRRs, the comprehensive alignment steps performs the alignment globally among all QRRs to construct a table in which all data values of the same attribute are aligned in the same table column. Instantly, if we view each data value in the QRRs as a vertex and each pair wise alignment between two data values as an edge, viewed the pair wise alignment set as an undirected graph.

As a result our comprehensive alignment problem is equivalent to that of finding connected components in an undirected graph. Each and Every connected component of the graph represents a table column inside which the connected data values from different records are aligned vertically. Even there are many efficient algorithms for finding connected components in the Graph Theory

literature; we need to consider two application constraints that are specific to our comprehensive alignment problem.

1) Vertices from the same record are not allowed to be included in the same connected component as they are considered to come from two different attributes of the record. If two vertices from the same record breach this constraint, then a path must exist between the two, which called as breach path.
2) Connected components are not allowed to intersect each other. If C1 and C2 are two connected components, then vertices in C1 should be either all on the left side of C2 or all on the right side of C2, and vice versa (i.e., no edge in C1 cuts across C2, and no edge in C2 cuts across C1).

Accordingly, we design a 3-step algorithm for the comprehensive alignment problem. First step, we traverse the graph once by a depth-first search to discover the preliminary connected components. We also mark those components containing breach paths at the same time. Then we traverse the components containing breach paths to remove some edges so as to break the breach paths (i.e., enforcing the first constraint). At last, we use a divide-and-conquer method to identify and split up the intersecting components to enforce the second constraint.

3.3. Combination of data value alignment

Comprehensive data value alignment constrains a data value in a QRR to be aligned to at most one data value from another QRR. When a QRR includes a nested structure such that an attribute has multiple values, some of the values may not be aligned to any other values. And so, nested structure processing identifies the data values of a QRR that are generated by nested structures (i.e., the repetitive parts of a generating template).

Depending only on HTML tags to identify nested structures, such as is done by almost all existing methods, may wrongly identify a plain structure as a nested one. To solve this problem, our approach uses both the HTML tags and the data values to identify the nested structures. The nested column set C is initialized to be an empty set. For each QRR with record root node t in T, the procedure nest_column_identify is invoked to identify any nested columns in the QRR.

After all the nested columns are identified, a new row is generated by copying the remaining parts as well as the repetitive data values. The repetitive data values copy is removed from the original row. Given a node t in tag tree T, the comprehensive alignment columns and the nested column set C as input, the procedure nest_column_identify identifies all repetitive parts under t in T. This procedure is called recursively until it reaches a node that contains only one data value. Hence, nested column identification is performed from leaf nodes of T to the root.

We identify the repetitive tag pattern in its children for each node. Suppose there is a repetitive tag pattern found in t's children, each of which contains a data value of the record. For each tag repetition p that contains data value f_1. . . f_n, c_p is defined to be the columns in the comprehensive alignment that contain $f_1, \ldots, f_n$. We now need to decide, related to the data value similarity in the columns c_p, whether the repetitive tag is generated from a nested structure or it is actually a flat structure. If it is generated by a nested structure, c_p is added to the nested column set C.

4. Experimental results and discussion

A new web search system is designed to search the information around World Wide Web. Here the given query is "A Guided Tour to Approximate String Matching". It is IEEE paper topic based on this web search is done and crawl the information's to stored in the web folder. Figure 1 shows the design of web search system and Figure 2 shows the design of web pages extraction download to web folder.

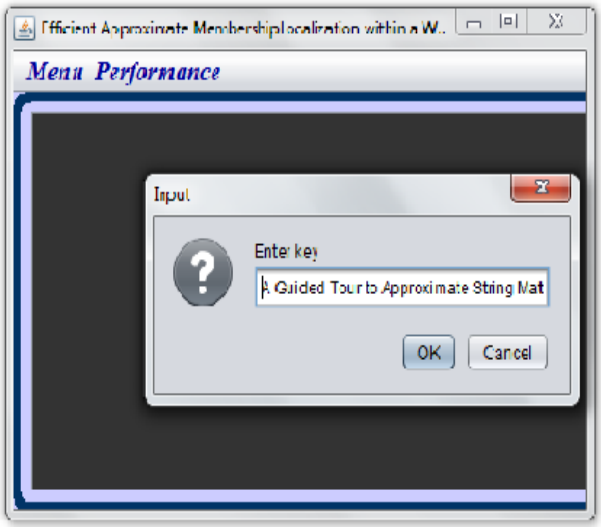

Fig. 1: Web Search System.

There are three categories to cover web search system namely informational queries, navigational queries, transactional queries and connectivity queries. Web crawler visits the URL (Uniform Resource Locator) and identifies the hyperlink of the page and crawls to the crawl frontier. Recursively this process is done to repeat the web pages to download until the user needs based on the input given query.

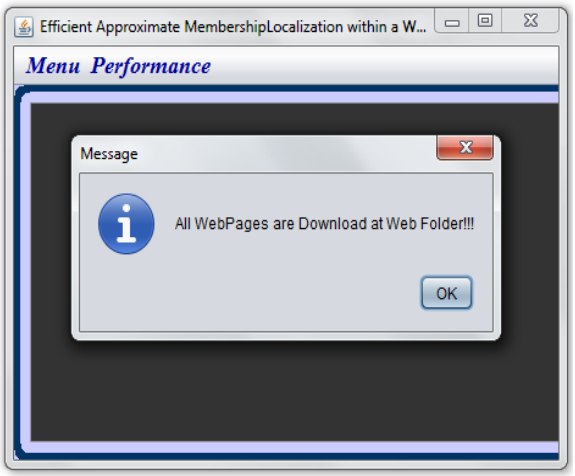

Fig. 2: Web Page Extractions.

Using the HTML parsers to convert the web page into text file and stored in a separate folder. Figure 3 shows the text extraction from web folder.

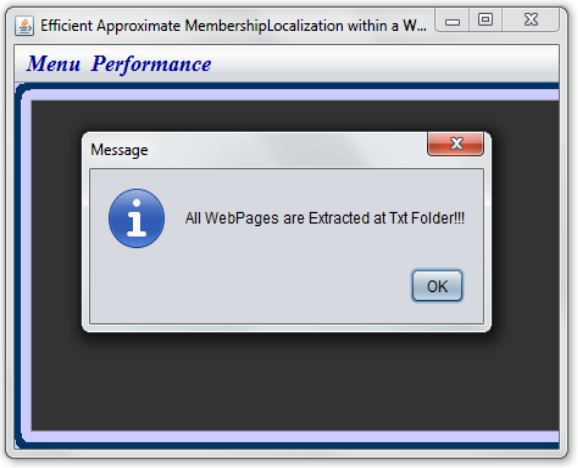

Fig. 3: Web Pages to Text Extraction.

For example, consider research paper title as keyword searching from various journals such as ACM, Springer, IEEE, Elsevier, VLDB, SIGMOD, Science Direct, ICDE, IJCA, ICDE, and JCS. Based on the given keyword the text is searched and displayed the probability value of given string. We assume the probability value is 5. Compared to the different alignment methods the result is identified and achieves the probability value is 4.9. Time taken to search the string is 156 mill seconds and memory taken is 159744 bytes. Figure 4 shows the result of data extraction based on tag value similarity measure using different alignment methods and the result is identified to found successful.

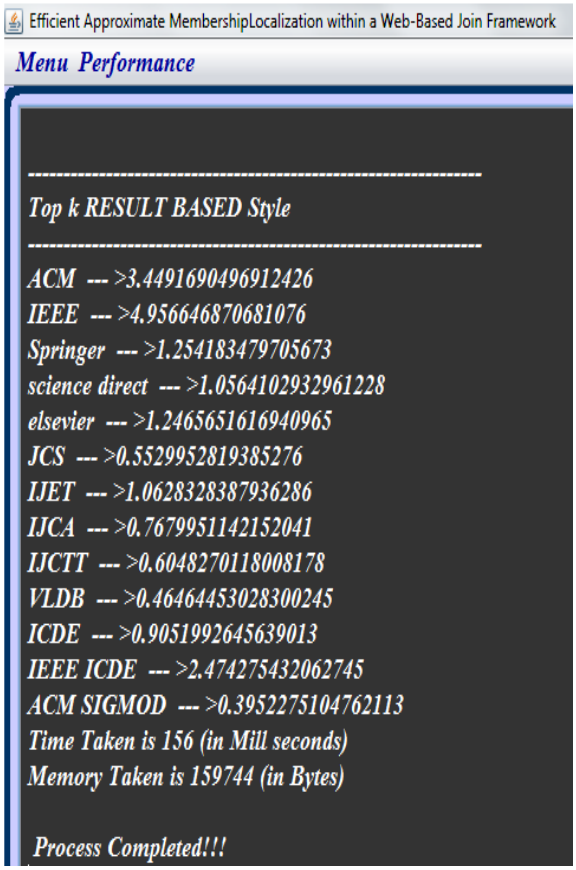

Fig. 4: Retrieval of Document Similarity Measure Using Alignment Methods.

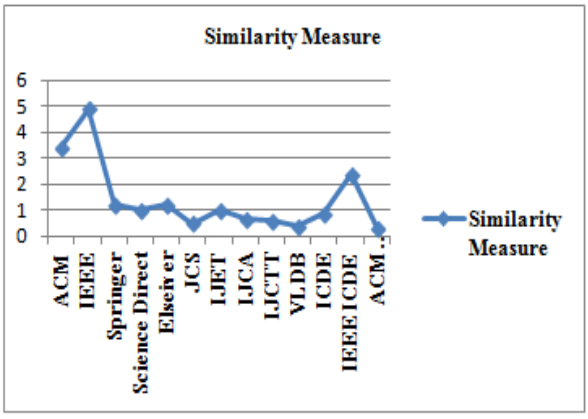

Fig. 5: Similarity Measure.

The approximate string matching is identified by using alignment method for the given string "A Guided tour to Approximate String Matching " paper is published in IEEE and the similarity value is 4.9. Figure 5 shows the graph Similarity Measure Value.

5. Summarization & future scope

This research is concerned with study and analysis of data extraction, data alignment methods. The retrieval method is based on data alignment techniques. The algorithm for data retrieval is based on style of documents and then developed software prototype allows retrieving the string similarity matching from the given input query. The prototype is tested with research paper title and found successful. This research work can be further extended in the following directions: journal logo matching to retrieve the documents or web pages from web.

Acknowledgement

First and foremost my whole-hearted thanks to the Lord Almighty, for his abundant blessings. I am very glad to express my deep sense of gratitude and profound thanks to my guide Dr. P. Ponmuthuramalingam, Controller of examinations and Associate Professor, Post Graduate and Research Department of Computer science, Government Arts College (Autonomous), Coimbatore, for initiating me into this field of research and for providing me with the necessary guidance, great encouragement throughout the preparation of this paper. I take this opportunity to express my gratitude to the Staff members, Non-Teaching Staff members and Research Scholars of the Graduate and Research Department of Computer science, Government Arts College (Autonomous), Coimbatore, for their timely help and encouragement. I express my heart full thanks to my beloved father and mother without whose support achieving the success in the paper would be impossible. The amount of encouragement received especially from my friends requires a special mention. I record my deep indebtedness to them for their support.

References

[1] Bhosale, C (2015). Automatic Annotation of Query Results from Deep Web Database. International Journal of Engineering Sciences & Research Technology, 1(4), pp. 239-246.
[2] Crescenzi, G. Mecca, and P. Merialdo (2003), "Road Runner: Towards Automatic Data extraction from Large Web Sites," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 109-118, 2001 Web Conf. (APWeb), pp.406-417.
[3] Hai He, Hongkun Zhao, Y. Yiyao Lu, Weiyi Meng (mar. 2013), Annotating Search Result Records from web databases, IEEE Transaction on Knowledge and Data Engg., 25(3), pp. 239-246.
[4] Hammer, J. McHugh, and H. Garcia-Molina, (1997) "Semi structured Data: The TSIMMIS Experience," Proc .East-European Workshop Advances in Databases and Information Systems (ADBIS), pp. 1-8.
[5] http://db.cis.upenn.edu/DL/www8.pdf (accessed on 12th Nov 16)
[6] Jadhav, T., & Chobe, S. (2015). Data Extraction and Alignment of Search Results by Combining Tag Value Structure. IJETT, 2(2). Pp. 381-384.
[7] Liu, W., Meng, X., & Meng, W. (2006). Vision-based web data records extraction. In Proc. 9th international workshop on the web and databases (pp. 20-25).
[8] Lu.Y, H. He, H. Zhao, W. Meng, and C. Yu (2007), Annotating Structured Data of the Deep Web, Procedure IEEE 23rd Intl Conference Data Eng. (ICDE). Pp. 1-18.
[9] Manjula, R., & Chilambu chelvan, A. (2013). Hauling Templates from Web Pages Using Clustering Techniques. International Journal of Engineering Sciences & Emerging Technologies, 5(2), pp. 119-126.
[10] Muneeswari, G. (2014). Agent based Authentication for Deep Web Data Extraction. International Journal of Innovative Research in Information Security (IJIRIS), 2(4), pp. 44-52.
[11] Patel, D., & Thakkar, A. (2015). A Survey of Unsupervised technique for web data extraction. International Journal of Computer Science, 6(2), pp. 1-5.
[12] Shen, W., & Zou, X. (2015). An Algorithm on Web Article Automatic Extraction Based on DOM Structure. International Journal of Hybrid Information Technology, 8(3), 243-254. https://doi.org/10.14257/ijhit.2015.8.3.22.

[13] Sriramoju, S. B. (2014). An Application for Annotating Web Search Results. Proc. International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol, 2. Pp. 3306-3312.
[14] Stern, R., & Sagot, B. (2012, June). Population of a knowledge base for news metadata from unstructured text and web data. In Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction Association for Computational Linguistics. pp. 35-40.
[15] Thomas, S (2014). Clustering Based Annotation of Search Results. International Journal of Emerging Trends in Engineering and Development 4(3). Pp.123-130.
[16] Yogam, V., & Uma maheswari (2014), K. Automatic Annotation Wrapper Generation and Mining Web Database Search Result. International Journal of Innovative Research in Science, Engineering and Technology, 3(3). Pp 10562-10569.