International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2020): 7.803

Investigation of Automatic Data Extraction Method from Complex Web Pages

Nitin More[1], Rupali A. Mangrule[2]

[1]MIT Aurangabab

[2]Professor, MIT Aurangabab

Abstract: ***The Internet presents great deal of helpful info that is sometimes formatted for its users, that makes it laborious to extract relevant knowledge from numerous sources. Therefore, there's a big would like of strong, versatile info Extraction systems that remodel the net pages into program friendly structures like a computer database can become essential. The projected system focuses on info extraction from websites. We tend to cluster the net documents supported the common example structures so the example for every cluster is extracted at the same time. The planet wide net could be a huge and speedily growing supply of helpful info that is employed to publish and access the knowledge on the net. It uses totally different templates with contents for providing quick access for readers. This is often wont to extract info from example websites.***

Keywords: Information Extraction, Clustering, Minimum Description Length Principle, MinHash, Template extraction, Clustering web pages

1. Introduction

World Wide net may be an approach of accessing information over the medium of net. It's the foremost helpful supply of knowledge that is growing at a speedy rate in range of web sites. Web site may be a set of connected sites that contain contents like text, images, audio and video. Web content may be a document which could be accessed through an internet browser. Thus, sites contain a mixture of distinctive content and template, that is conferred across multiple pages but the unknown templates are additional harmful, since they contain an outsized range of inapplicable terms within the net documents and so templates degrade the accuracy of knowledge extraction from sites. Info Extraction, is that the task of mechanically extracting the relevant information from net documents. There's no thanks to differentiate between the text which will be a district of template and text that's in addition a district of knowledge. Second, the schema of knowledge in sites isn't typically a flat set of attributes, but its additional difficult and semi structured. The planet wide net may be a large and quickly growing supply of helpful info that is employed to publish and access the data on the web. During which the foremost of this info is within the variety of unstructured text that makes laborious to question the data needed. Totally different templates are used for visualizing the content of web content. Server ought to discover the template of an internet page before commercial enterprise it once requested by the user, that has given plenty of attention to template detection in recent times. Many algorithms are developed to discover this template structures mechanically so as to spot and extract contents of a documents. The sudden growth and recognition of World Wide net has resulted in a very vast quantity of knowledge sources on the web. However, as a result of the no uniformity and lack of structure of net info sources, access to the present vast assortment of knowledge has been restricted to browsing and looking. To automate the interpretation of input sites into structured information, plenty of efforts are devoted within the space of knowledge extraction. This paper studies the matter of mechanically extracting structured information encoded in a very given assortment of pages. With none human input likes manually generated rules or coaching sets. A vital characteristic of pages happiness to an equivalent web site and secret writing information of an equivalent schema is that the information secret writing is finished in a very consistent manner across all the pages. A site-level template detection methodology has some limitations. First, site-level templates represent solely tiny low Fraction of all templates on the net. Second, these ways are error prone once the quantity of pages analyzed from a web site is statistically insignificant. Clump of net documents during which all the documents associated with a cluster contains common template methods, the correctness of template structure depends on the clump.

2. Literature Review

Many similarity based measures are there for clustering of web documents based on the similarity between trees. But tree related distance measures are very expensive for clustering. The motivation of our work is to extract the most meaningful data from XML web documents. In this section we survey the previously proposed approaches for the accurate extraction of data from web pages

Zhao et al. proposed a method for Fully Automatic Wrapper Generation for Web Pages. This paper presents a technique for automatically producing wrappers which can be used to extract the data from dynamically generated result pages which is returned by search engines. Automatic data extraction is very important for wide range of applications that need to interact with search engines such as Meta search engines and web crawling

Zheng et al. proposed a Joint Optimization Technique which combines Wrapper Generation and Template Detection. It includes template detection and wrapper generation in a single step. It utilizes similarity between pages or any other external features to detect templates. It separates pages with notable inner differences and then generates wrappers for

the web pages. The approach is more stable as it does not rely on URLs to detect templates. Also, tree related similarity measures used for clustering are very expensive.

Layaida et al. developed an article about an Impact of XML Schema Evolution. This aims to cover the most general issue of schema evolution by taking into account the impact on the validity of documents. It presents a framework for checking those criteria with the schemas by specifying the main standard document formats used on the web. It also provides a unifying framework that allows for the automatic verification of properties related to XML schema evolution and its impact on the validity of documents and queries.

We are using hierarchal clustering approach of the text based web document clustering. The text-based web document clustering approaches characterize each document according to its content in it. The basic idea is that if two documents contain many common words then it is likely that the two documents are very similar. In the proposed system we are using agglomerative hierarchal clustering algorithm for clustering the documents which produces a sequence of nested partitions.
Reis et al. presented a method in which a small number of sampled documents are clustered first, and then, the other documents are classified to the closest clusters. In both clustering and classifying, a restricted tree-edit distance is used to measure the similarity between documents.

As the templates degrade the accuracy of data extraction, this proposed work mainly focuses on the extraction of most meaningful and relevant data from XML web documents to enhance the performance of web applications. Semantic searching using XML can be achieved since it is used to get the most meaningful information compared to HTML. As Minimum Description Length is used for clustering, the scalability is also achieved.

In recent years, many researchers have tried to improve performance of template detection methodology.

Data Extraction from HTML document
In Automatic Web News Extraction Using Tree Edit Distance, presents a domain oriented approach to web data extraction. This approach is based on highly efficient tree structure analysis.HTML document can be represented with a Document Object Model (DOM) tree, web documents are considered as trees.

Structure of web page can be described by a tree. Tree- edit distance is used to evaluate the structural similarity between pages. Restrictive Top down Mapping (RTDM) algorithm is used to identify relevant text. But the worst case complexity of RTDM algorithm is O(n1n2) where n1 and n2 are sizes of the two trees, but it performs much better than traditional top down mapping. Here, they evaluate structural similarity between HTML pages and based on that, grouping of pages is done to form cluster and find generic representation of structure of pages within a cluster.

In Extracting structured data from web pages, extraction of data is done in 2 steps: 1. formally define a template and propose a model that describes how values are encoded into pages using a template Present algorithm that takes as input a set of template generated pages; deduce the unknown template used to generate pages and extracts as output. However, if this approach is used for crawling, indexing, then there will be problem of automatically locating collection of structured pages.

Roadrunner: Towards Automatic Data Extraction from Large Web Sites [3], introduced extracting data from HTML sites through the use of automatically generated wrappers. This paper develops a novel technique to compare HTML pages and generate a wrapper based on similarity and differences. Goal is automatic generation of wrapper that is without any prior knowledge of target pages and human interaction. Matching technique is used to compare the HTML codes of two pages and to infer a common structure and a wrapper.

A Fast and Robust Method for Web Page Template Detection and Removal [4], RTDM-TD algorithm is used to find optimal mappings between the Document Object Model (DOM) trees of web pages. This algorithm is based on a restricted formulation of top down mapping between two trees, which is particularly suitable for detecting structural similarities among web pages. But the operations related to trees are expensive.

Data extraction from XML document
Extract provides a system for extracting Document Type Descriptor (DTD) from XML documents. XML document can be accompanied by a Document Type Descriptor (DTD) which plays the role of a schema for an XML data collection. DTD contains valuable information on the structure of document. Extract method solved the problem of DTD extraction from multiple XML documents.

Clustering method
In Automatic Web News Extraction Using Tree Edit Distance, presented a method, in which small number of sampled documents are clustered first, and then the other documents are classified to the closest clusters. In this approach selecting proper training data is not easy task. In Joint Optimization of Wrapper Generation and Template Detection, labeled training data is used for clustering.

3. Proposed Algorithm Essential paths of document

Define path set PW for document W. PW is set of all paths in W. Support of path(SP) is the number of documents in W, which contain the path. For all wi, there exists twi. twi is decided by taking mode of support values. If a path is contained by a document wi and support of path is at least the given twi then the path is essential path of wi . Set of such essential paths are EP. These essential paths are used in extracting template. Set of paths and HTML documents are denoted by matrix of size |PW|×|W|

<html>
<head>
<>title> abcd </title>
</head>
<body>
<h1> pqrs</h1>
** tuvw **
</body>
</html> HTML Document

Template of document
A guide is about of common layout and format feature that seem in an exceedingly set of HTML pages that's made by one program or script that dynamically generates the HTML page content. Guide of a document cluster could be a set of ways that usually seem within the documents of the cluster.

Representation of Clustering
Cluster ci-> (Ti, Wi), Ti is set of paths representing the template of ci and Wi is set of documents belonging to ci . Successful condition for clustering is WifiWj=Ø and U1<=i<=m Wi= W

Minimum Description Length (MDL) principles
MDL principle is used to manage unknown number of clusters and to select good partitioning from all possible partitions of HTML documents. The MDL principle states that the best model inferred from a given set of data is the one which minimizes the sum of 1) the length of the model, in bits, and 2) the length of encoding of the data, in bits, when described with the help of the model.

4. Conclusion and Future Work

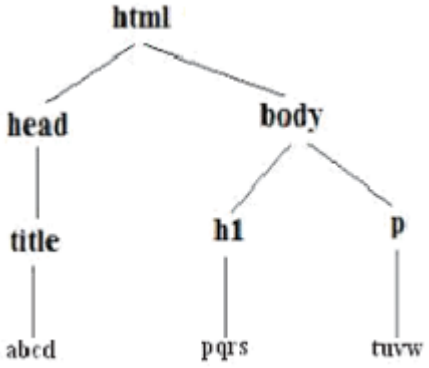

Figure: DOM Tree

Clustering using MDL COST
Clustering is a process of forming groups (clusters) of similar objects from a given set of inputs. Good clusters have the characteristic that objects belonging to the same cluster are "similar" to each other, while objects from two different clusters are "dissimilar". Input to clustering is set of web document W= {w1,w2, w3, ..} and output is set of cluster C= {c1, c2, ..}. Cluster ci-> (Ti, Wi), as explained in section Clustering of web document can be done by using agglomerative hierarchical clustering algorithm as explained in . Initially each document is an Individual cluster. When a pair of cluster is merged, the MDL cost of the clustering model can be Increased or decreased. Find a pair of cluster whose reduction of MDL cost is increased in each step of merging and the pair is repeatedly merged until any reduction is not possible. Proposed system includes above six steps. System will take input as HTML document and automatically extract template within very short period of time as compared to manual method. Here, we describe psedo code for our system:

1) Procedure: START(W)
2) While W Ç Ø
3) CreateDOMtree(W)
4) FindPath(DOM)
5) FindSupporOfPath(DOM)
6) FindEssentialPath(PW, SP)
7) ExtractTemplate(EP)
8) ClusterDocuments(W,T)
9) GetMDLcost(ci , cj, C)
10) GetBestPair(C)
11) End while
12) End procedure

Template extraction from heterogeneous sites will be done by constructing Document Object Model (DOM) tree of HTML document and finding essential methods of document. Clump of internet Document will be done on the premise of template structure to manage unknown range of template. Most of the internet sites contain large set of web documents that square measure generated exploitation common templates. The connation in many internet applications is affected since the templates square measure wide used on the online. Also, templates scale back the performance of internet applications and ends up in wastage of resources. So to avoid those problems, associate approach for knowledge extraction from heterogeneous internet documents has been enforced during this planned work.

References

[1] H. Zhao, W. Meng, Z. Wu, V. Raghavan, and C. Yu, Fully Automatic Wrapper Generation for Search Engines, Proc. 14th Intl Conf. World Wide Web (WWW), 2005.
[2] Arasu and H. Garcia-Molina, "Extracting Structured Data from Web Pages," Proc. ACM SIGMOD, 2003.
[3] Crescenzi, G. Mecca, and P. Merialdo, "Roadrunner: Towards Automatic Data Extraction from Large Web ites," Proc. 27th Int'l Conf.Very Large Data Bases (VLDB), 2001
[4] K. Vieira, A.S. da Silva, N. Pinto, E.S. de Moura, J.M.B. Cavalcanti, and J. Freire, "A Fast and Robust Method for Web Page Template Detection and Removal," Proc. 15th ACM Int'l Conf. Information and Knowledge Management (CIKM), 2006.
[5] M. de Castro Reis, P. B. Golgher, A.S. da Silva, and A. H. Laender, "Automatic Web News Extraction Using Tree Edit Distance," Proc. 13th Int'l Conf. World Wide Web (WWW), 2004.
[6] Arasu, and H. Garcia-Molina, “Extracting Structured Data from Web Pages”, in Proceedings of the ACM SIGMOD International Conference on Management of Data, San Diego, USA, 2003, pp. 337- 348
[7] hulyun kim, and kyuseok shim, “TEXT: Automatic Template Extraction from Heterogeneous Web Pages”, IEEE Transactions on Knowledge and Data Engineering, Vol. 23, No. 4, 2011, pp. 612-626.

[8] V. Crescenzi, P. Merialdo, and P. Missier, "Clustering Web Pages Based on Their Structure", IEEE Transactions on Data and Knowledge Engineering, Vol. 54, No. 3, 2005, pp. 279-299.
[9] M. de Castro Reis, P. B. Golgher, A. S. da Silva, and A. H. F. Laender, "Automatic Web News Extraction Using Tree Edit Distance", in Proceedings of the 13th International Conference on World Wide Web, New York, USA, 2004, pp. 502-511.
[10] Z. Bar-Yossef and S. Rajagopalan, "Template Detection via Data Mining and Its Applications," Proc. 11th Int'l Conf. World Wide Web (WWW), 2002.
[11] N. Garofalakis, A. Gionis, R. Rastogi, S. Seshadri, and K. Shim, "Xtract: A System for Extracting Document Type Descriptors from Xml Documents," Proc. ACM SIGMOD, 2000.
[12] Crescenzi, G. Mecca, and P. Merialdo, "Roadrunner: Towards Automatic Data Extraction from Large Web Sites," Proc. 27th Int'l Conf. Very Large Data Bases (VLDB), 2001.
[13] Bar-Yossef and S. Rajagopalan, "Template Detection via Data Mining and Its Applications," Proc.11th Int'l Conf. World Wide Web (WWW), 2002.
[14] S. Abiteboul, R. Hull, and V. Vianu. Foundations of Databases. Addison Wesley, Reading, Massachussetts, 1995.
[15] Brin. Extracting patterns and relations from the world wide web. In WebDB Workshop at 6th Intl. Conf. on Extending Database Technology, 1998.