Literature Survey on Extraction of Top-k Lists from Web Pages

Darshana Dabhi[1] Ms. Jasmine Jha[2]
[1]M.E Student [2]Assistant Professor
[1,2]Department of Computer Engineering
[1,2]L. J. Institute of Engineering & Technology, Ahmedabad, Gujarat, India

Abstract— The web contains data in huge amounts. This data is a large source of information. All this information is in the form of structured or unstructured data. List is a crucial source of structured data on the web. Ranking the list data is generously important for information retrieval. Tremendous efforts have been done for extracting information from the structured data, especially from web tables, which contain quality information. Instead of focusing on context- free structured data, we aim to focus on context that we can spot, and then using the context to render less controlled information and proceed to its extraction. Here we highlight expensive as well as, rich source of information on the web, those are top-k web pages. Top-k web pages contain rich and quality information. They aim to identity the top attribute values for the entities of interest. Extraction of such lists can help answering engines to generate different fact and can act as a pre-processing step.

Key words: Rank Search, Time Sensitive Queries, binning

I. INTRODUCTION

The World Wide Web contains huge amount of information. For extracting the structured information from the web, which can be in the form of tags like <table>, <li>, <tr>, etc. Understanding the context of the structured data is very important in order to utilize that data. Understanding the context means the relation between the listed items must be known like, why these items are listed together and what is their common feature. The context is mostly present in the natural language but machine cannot easily interpret it. The parameters relating the list items to each other need to be identified.[5]

Whenever a user wants to find any top-k data, like Top 5 mobile phones in India, Top 10 Bollywood hits, 10 Most Influential Books of 2014, etc, the user fires a query in search engine. The search engine processes the query and gives various links as results [1]. But these links may or may not give direct results to the users. The user has to visit each link and find out if the desired information is present in that web page. If the information is found the search terminates, otherwise the user can visit next links and the process repeats till the user gets the desired results.

Hence a top-k list helps us to solve the above issue. Top-k lists have rich information of high quality. Also, top-k data on web is large and have interesting semantics.

For extracting the top-k data, top-k titles need to be identified first. The top-k titles always have a value k, which is an integer number, like 5, 10, 15. Top-k titles also specify a concept which defines which kind of items need to be retrieved. And a criterion in a top-k list tells us on what basis the items need to be filtered or ranked. Place and time information can also be given in a top-k title, but they are optional

II. LITERATURE SURVEY

A. *Title Extraction from the Bodies of Html Documents [1]*

This paper takes up machine learning approach to address the problem of automatically extracting titles from HTML documents (web pages). It has 2 phases:

Training and extraction The input is a document for pre-processing. It parses the body of the HTML document and constructs a DOM tree. And then it extracts all the leaf nodes as units from the DOM tree [10].

In learning, the input is a sequence of units from one document, and each sequence corresponds to one document. Labelled units are taken as training data and a model is constructed for identifying whether a unit is a title. In extraction, the input is a sequence of units from one document. The model is used to identify each unit in the sequence to find whether it is a title and it assigns a score to each unit.

The output is the extracted titles of the document. The consecutive units with the highest scores are chosen first for title and then consecutive units with second highest scores as second titles[1]. This paper tackles the issue of extracting the titles from the bodies of HTML documents. The main drawback being that, it only uses HTML pages and uses the titles extracted from the bodies of the HTML pages.

B. *Popularity Guided Top-K Extraction [2]*

This paper aims to return the top-k values of the attribute for the entity according to a scoring function for extracted attribute values. This scoring function depends on extraction confidence and importance. More often each document is accessed by users when searching for information related to an entity, the more likely it contains important information[2].By analysing query click-through data, search engines can identify the web documents that people refer to for information. For each entity in dataset, a frequency measure is computed on the basis of how many users have searched for the entity and how many pages matching a particular pattern have been clicked as a result of the search [2].

It follows the following algorithm:

- Document Selection: Select a batch of unprocessed documents
- Extraction : Process each document in batch with extraction system
- Top-k Calculation : Update rank of extracted attribute values for each entity
- Stopping Condition : If top-k values for each entity have been identified, stop, otherwise go to step 1[2]

This paper addresses both quality and efficiency challenges and gives more popular documents in results by focusing on the importance of data. But this method may ignore the new and fresh web pages, which may be containing important data. Popular data may get more and

more popular and new web pages will take some time to come into the result set.

C. Extracting General Lists from Web Documents: Hybrid Approach [3]

The paper introduces a Hybrid approach for automatic list discovery and extraction on the web (HyLiEn). HyLiEn uses the CSS2 visual box model to segment a web page into a number of boxes, each having a position and size. It recursively considers inner boxes and then extracts list boxes which are visually aligned and structurally similar to other boxes. Visual clues in the web page are utilised to generate candidate lists, which are subsequently pruned with a test for structural similarity in the DOM tree [3].

The paper considers the visual and structural features of the web lists and produces a general list, but not a ranked one. This method is applicable to only those web pages where CSS box model can be applied and does not have a notion of element distance that could be used to separate aligned but separated lists.

D. Automatic Extraction of Top-K Lists from the Web [4]

This paper uses tag paths, which is a path from the root to the arbitrary node in the DOM tree. It improves the result quality by optimization heuristics:

- Visual Area: The total visual area of the candidate list versus the total area of the page is considered, by calculating the combination of image sizes, font sizes and potential white spaces
- Interleaving Lists: Top-k lists may have list items with alternate visual styles such as background colours or fonts. A special heuristic is used to detect such interleaving patterns and reconstruct the whole list
- K+1 problem: Top-k pages may have additional header or footer that looks almost same with same tag paths. In another case, the first or the last item of a top-k list may have slightly different style and gets excluded from the class. Special attention is paid to such items by analysing text content[4]

The paper has a basic algorithm running in four *steps:*

- Compute the tag path for every node in the DOM tree of the input page
- Group nodes with identical tag paths into one class & select those classes having exactly k items as candidate classes
- Merge the candidate classes on whom the grow-up operation can be applied, item components that belong to the same list item are grouped together
- The candidate list is ranked by their importance to the page and returned as result[4]

This paper gives improved result quality and better performance by using the optimization heuristics, and also gives ranked results. As the focus of the paper is on the visual area and patterns, smaller lists may not get noticed.

E. Enhancement In Extraction Of Top-K Lists [7]

The system introduced here consists of the following components:

- Title Classifiers : It attempts to recognize the page title of the web page
- Candidate Picker: It extracts all the candidate lists from the input page. It is structurally a list of HTML tag paths which are identical. A tag path is a sequence of tag names, from the root node to a certain tag node.
- Top-k Ranker: It scores the candidate list and picks the best one by scoring function which is weighted sum of two features: P-score and V score.
- P score measure the correlation between the list and title. V score calculates the visual area occupied by a list, because usually the main list of a web page tends to occupy larger area than other lists.
- Content Processor : Processes the extracted list to produce attribute value pairs by inferring the structure of text nodes, conceptualizing the list attributes, using the tables heads or the attribute/value pairs.[5]

This method gives improved performance by providing domain-specific lists and focussing more on the content. It doesn't focus only on the visual area of the lists. But a list, if divided into more than one pages, may not get included completely.

III. CONCLUSION

The paper presents a survey on different aspects of the work done till now in the field of extraction of data from web pages. The traditional systems focused on retrieving tabular data and producing general lists. Mostly the description is in natural language which is not machine interpretable. Later the research continued with topic mining contiguous and non-contiguous data records. Next the research expanded to extracting general lists from the web more efficiently. Next the evolution was done in retrieving top-k list data from web pages, which gives the ranked results. Hence, top-k list data is of high importance. By, understanding the issues faced by the current systems, more improvements can be done in the field of extracting top-k lists from the web pages. Hence, top-k data is of high superiority and has cleaner data than other forms of data on the web.

IV. REFERENCES

[1] Yunhua hu, Guomao Xin, Ruihua Song, Guoping Hu, Shuming Shi, Yunbo Cao, Hang Li, "Ttitle Extraction From the Bodies of HTML Documents and its Application to Web Page Retrieval", Microsoft Research Asia, SIGIR '05, August 15-19, ACM, 2005.

[2] Mathew Solomon, Cong Yu, Luis Gravano,"Popularity Guided Top-k Extraction of Entity Attributes", Columbia University, Yahoo! Research, WebDB '10, ACM, 2010.

[3] Fabio Fumarola, Tim Weninger, Rick Barber, Donato Malerba, Jiawei Han, IEA/AIE Part 1 LNAI 6703, pp. 285-294, 2011.

[4] Zhixian Zhang, Zheng Wang, Haixun Wang, Kenny O. Zhu, Shanghai Jiao Tong University, "Automatic Extraction of Top-k Lists from the Web" Microsoft Research Asia.

[5] Dipali Patil, Nitin Dhawas, "Enhancement in Extraction of Top-k List", Singhad Institute of

Technology, IOSR-JCE, Volume 16 Issue 3, ver III (May-Jun), PP 129-133, 2014.

[6] Michael J. Cafarella, Alon Halevy, Daisy Zhe Wang, "Web Tables: Exploring the Power of Tables on the Web", PVLDB '0, August 23-28, ACM, 2008.

[7] Gengxin Miao, Junichi Tatemura, Wang-Pin Hsiung, Arsany Sawires, Louise E. Moser, "Extracting Data Records from the Web Using Tag Path Clustering", April 20-24, ACM, 2009.

