IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

A Review on Extracting Top-k Lists from the web

Shreya U. Wadkar[1], Prof. N. G. Pardeshi[2]

PG Student, Computer Engineering Department, SRESCOE, Kopargaon, India [1]

Assistant Professor, Computer Engineering Department, SRESCOE, Kopargaon, India[2]

Abstract: The web contains a huge amount of data and this data results into a large amount of information. The information on the web is in two forms i) Structured data and ii) Unstructured data. In this paper, we focus on structured data. List data is one of the most important source of structured data for information retrieval on the web. This paper deals with the "Top-k Lists", web pages that describe a list of *k* instances of a particular topic or concept. Examples are, "the 5 top most cars in the world", "the 10 richest businessman in the world" etc. Top-k lists are a richer, larger and of high quality source of information. Therefore, top-k lists are highly valuable. This paper reviews a various traditional methods for extracting the top-k lists. After studying these, we present an efficient method that obtains the target lists from web pages with high accuracy. Extraction of such lists can help enrich existing knowledge bases about general concepts and useful as a preprocessing step to produce facts for a fact answering engine.

Keywords: Web information extraction, Top-k lists, List extraction, Web mining.

I. INTRODUCTION

Currently, web becomes the largest source of information. Web information is unstructured text in natural languages and extracting knowledge from such natural language text is not easy. But still, some of the web information exists in structured or semi-structured forms. Lists or web tables coded with specific tags such as <ul>, <li>, and <table> on html pages are the examples of these forms. Structured information is very useful to extract the knowledge easily. As a result, recently, many researchers have focused on acquiring knowledge from structured information on the web, specifically, from web tables [1], [2], [3], [4], [5], [6], [7].

Still, it is doubtful about how much useful and valuable information we can extract from lists and web tables. It is sure that the total number of web tables are huge in the whole corpus, but only a very little percentage of them contains useful information. A smaller percentage of them contains information interpretable without context. Particularly, based on our knowledge, more than 90% of the tables are used for content arrangement on the web. Additionally, a lot of the remaining tables are not "relational." We are only concerned about relational tables because they are interpretable, with rows as entities, and columns as attributes of those entities. As per Cafarella et al. [2], 1.1% of all web tables that are relational, numerous are worthless without context. For instance, suppose we derived a table having 5 rows and 2 columns, also have the 2 columns labelled "Companies" and "Revenue" respectively. But, it is not clear why these 5 companies are combined together (e.g., are they the most profitable, most innovative, or most employee favourable companies of a specific industry, or in a particular area?), and how we should know their revenues (e.g., in which year and in what currency).

We are unknown of the extract situations under which the extracted information is useful. So, understanding the context is very essential for extracting the information. Sometimes, the context is represented in unstructured text format that machines are unable to interpret. Instead of focusing on structured data (such as tables) and ignoring context, this paper focuses on the context that we understands, and then we use the context to interpret less structured or almost free-text information, and guide its extraction.

Proposed system has been invented to find out top-k lists from a world wide web that contains millions of pages. Top k list is linked with very high quality and key information, particularly evaluate with web tables, it contain large amount of high quality information. Furthermore, top k lists are associated with the context which is more useful and accurate in quality analysis, search and other systems.

The rest of the paper is organized as follows: Section II presents the literature survey over the related work. In section III, proposed system is presented. Finally, the section IV concludes the review paper.

II. LITERATURE SURVEY

Z. Zhang, K. Q. Zhu, and H. Wang [8] discover a story record extraction problem, It aims at realizing, extracting and knowledge "top-k" lists from web pages. The thing is different from discrete knowledge mining jobs, since, in comparison to different ordered knowledge, "top-k" lists are clearer, easier to know and well interesting for readers. Apart from these, "top-k" lists are important in knowledge discovery and truth addressing merely because there are a magnificent number of "top-k" lists around on the web.

Along with the bulky knowledge located in those lists, we can intensify the example place of a general purpose knowledge bottom such as Professional base. It is also possible to build a research engine for "top-k" lists as a

DOI 10.17148/IJARCCE.2015.41275

strong truth addressing machine. The proposed 4-stage extraction construction has shown its ability to access large number of "top-k" lists at a really large precision.

A numerous web applications has necessity to automatically extract data from multiple databases. The obtained query result pages includes some low adjacent QRR. This unrelated data is removed by two stage method known as QRR extraction. In this, record extraction first finds the creatively repeating data on a website and then extracts the info record using tag course clustering [9].

The idea of visible signal is proposed to simply the web site representation as set of binary visible signal vectors on behalf of a normal DOM tree. To extract information quickly from query result page, record positioning is done in CTVS approach. First, set sensible and then arrange the info in the QRRs. Hence, CTVS dehumanize the data extraction from multiple databases which supports many web applications. Also CTVS ejects the nested structure using nested structure processing for proper alignment.

Yogesh W. Wanjari [10] and his colleagues, have examined various data extraction techniques as well as automatic annotation method using many annotators from different web data bases. They also described that how a data extraction from the various web and the traditional methods are having many drawbacks like human disturbance, the inexactness in effect and poor scalability. Some methods used different feature extraction techniques for example series based pine edit range, DOM tree and HTML draw structure. The pleasing data extraction approach is language independent. This view largely aware about the demonstration design of and get the data successfully from the template. But nonetheless there is need to find the best strategy for knowledge annotation problems.

W. Gatterbauer [5], represents an abnormal and promising approach towards organized information extraction from the Web; particularly, from web tables. This approach uses a model of the aesthetic illustration of web pages as made by a web browser and, as a result, changes the problem of information extraction from the lower degree of rule model like HTML tag structure, CSS, JavaScript rule, etc. to the top degree of aesthetic functions like 2-D topology and typography.

Proposed approach works to execute successfully even without focusing for particular request domains including the model of various solutions. A varied check collection of web tables is provided to show this. Applying an aesthetic paradigm so as to approach automatic information extraction from web tables is encouraging, specifically given the rising hurdle in the encoding of web pages on the source rule level. Particularly, very powerful pages which tend to obtain more favoured by the rise of Web 2.0 is unable to prepare without composite model of the source code.

The possible usage of the extracted top-k lists is to serve as background knowledge for a Q/A system [11] to answer top-k related queries. To provide such knowledge, we require techniques to mingle a number of similar or connected provides within a more detailed one, which is nearer to top-k query processing. One of the most well known algorithms there is TA (threshold algorithm) [12], [13]. TA uses aggregation features to mix the results of objects in different lists and calculates the top-k objects on the basis of the mixed score. Further, Chakrabarti et al. [14] introduced the OF (object finder) query, which positions top-k objects in a search query exploring the connection between TOs (Target Objects, like, writers, products) and SOs (Search Objects, like, documents, reviewers). Bansal et al [15]. utilizes a similar platform but raises terms at a growing level by taking advantage of taxonomy, to be able to compute precise rankings. Angel et al [16]. Think about the EPF (entity Packet finder) issue which is worried with associations, relations between divergent forms of TOs. A part of these techniques can serve as the basis for detailed integration of top-k lists.

III.PROPOSED SYSTEM

The proposed system aims to find out top-k lists from web that contains millions of pages. Top k list is linked with very high quality and key information, particularly evaluate with web tables, it contain large amount of high quality information.

Following are the modules of the proposed system.

A. *Title Classifier:*

The title of a web page helps us identify a "top-*k*" page. This title is enclosed in <title> tag. The goal of the classifier is to identify "top-*k* like" titles, the probable name of a "top-*k*" page.

B. Candidate Picker:

Using HTML page body and the number *k*, the candidate picker gathers a set of lists as candidates. Each list item is a text node in the page body.

C. Top-k Ranker:

Top-K Ranker positions the candidate set and picks the top ranked list as the top-k list . This is to be done by using a scoring function, a weighted sum of two feature scores below:

P-Score: P-Score measures the correlation between the list and title.

V -Score: V-Score determines the visual area occupied by a list, as the main list of the page inclines to be enormous and more important compared to other minor lists.

D. Content Processor:

The content processor takes as input a "top-*k*" list and extracts the main entities as well as their attributes.

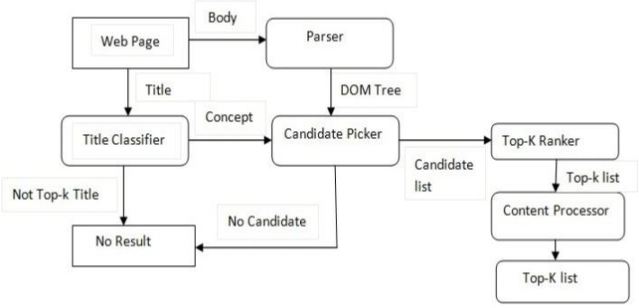

Fig.1 Top-k System Overview

DOI 10.17148/IJARCCE.2015.41275

IV. CONCLUSION

This paper reviews various methods for extracting top-k lists from the web. This paper represents a novel and exciting approach of extracting top-k provides from the web. Compared to other structured data, top-k lists are clearer, easier to understand and more exciting for human use, and therefore are significant source for knowledge mining and information finding.

The framework accomplishes an interesting issue of extracting top-k list from web, which goes for perceiving, extracting and comprehension top-k list from web pages. The extracted top-k list is of ranked and high quality. This top-k information is to a great extent accessible and has interesting semantic. Client can easily get results of top-k query utilizing above framework executed to concentrate top-k list from the web.

ACKNOWLEDGMENT

We are glad to express our sentiments of gratitude to all who rendered their valuable guidance to us. We would like to express our appreciation and thanks to the Principal of our college. We are also thankful to the Head of Department. We thank to the anonymous reviewers for their valuable comments.

REFERENCES

[1] "Google sets," http://labs.google.com/sets.
[2] M. J. Cafarella, E. Wu, A. Halevy, Y. Zhang, and D. Z. Wang, "Webtables: Exploring the power of tables on the web," in *VLDB*, 2008.
[3] B. Liu, R. L. Grossman, and Y. Zhai, "Mining data records in web pages," in *KDD*, 2003, pp. 601–606.
[4] G. Miao, J. Tatemura, W.-P. Hsiung, A. Sawires, and L. E. Moser, "Extracting data records from the web using tag path clustering," in *WWW*, 2009, pp. 981–990.
[5] W. Gatterbauer, P. Bohunsky, M. Herzog, B. Krupl, and B. Pollak, "Towards domain-independent information extraction from web tables," in *WWW*. ACM Press, 2007, pp. 71–80.
[6] F. Fumarola, T. Weninger, R. Barber, D. Malerba, and J. Han, "Extracting general lists from web documents: A hybrid approach," in *IEA/AIE (1)*, 2011, pp. 285–294.
[7] J. Wang, H. Wang, Z. Wang, and K. Q. Zhu, "Understanding tables on the web," in *ER*, 2012, pp. 141–155.
[8] Z. Zhang, K. Q. Zhu, and H. Wang, "A system for extracting top-k lists from the web," in KDD, 2012.
[9] J.Kowsalya, K.Deepa, "Extracting and Aligning the Data Using Tag Path Clustering and CTVS Method" International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 4, April 2013.
[10] Yogesh W. Wanjari, Dipali B. Gaikwad, Vivek D. Mohod, Sachin N. Deshmukh, "Data Extraction and Annotation for Web Databases using Multiple Annotators Approach- A Review", International Journal of Computer Applications (0975 – 8887) Volume 88 – No.18, February 2014.
[11] X. Cao, G. Cong, B. Cui, C. Jensen, and Q. Yuan, "Approaches to exploring category information for question retrieval in community question-answer archives," TOIS, vol. 30, no. 2, p. 7, 2012.
[12] R. Fagin, A. Lotem, and M. Naor, "Optimal aggregation algorithms for middleware," in PODS, 2001, pp. 102–113.
[13] U. Guntzer, W. Balke, and W. Kießling, "Optimizing multi-feature queries for image databases," in VLDB, 2000, pp. 419–428.
[14] K. Chakrabarti, V. Ganti, J. Han, and D. Xin, "Ranking objects based on relationships," in SIGMOD, 2006, pp. 371–382.
[15] N. Bansal, S. Guha, and N. Koudas, "Ad-hoc aggregations of ranked lists in the presence of hierarchies," in SIGMOD, 2008, pp. 67–78.
[16] A. Angel, S. Chaudhuri, G. Das, and N. Koudas, "Ranking objects based on relationships and fix associations," in EDBT, 2009, pp. 910–921.

DOI 10.17148/IJARCCE.2015.41275