ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 1, pp: (374-376), Month: January - March 2015, Available at: www.researchpublish.com

Enrichment in Extraction of Top-K List

[1]Kalyani Sonawane, [2]Nayan Sonawane, [3]Vedika Suryawanshi, [4]Dhwanit Vibhandik

[1,2,3,4] B.E. Computer, K. K. Wagh Institute of Engineering Education and Research, Nasik, India

Abstract: **This paper is concerned with enhancement in top k list in which top k pages are extracted from the web pages. In today's fast world everybody needs quick and accurate results. In this paper, the main focus is on the results required by the user when he tries to search some contents on the web. Web contains very large amount of data and large data is available in Top-k web pages. The information in top-k lists is larger, richer and of higher quality. This system is capable of handling all web pages including structured, non-structured or semi-structured. This system provides more accuracy and has less time complexity than previous system. This system also supports Dynamic Web system which is a web application which is useful for the online forums which provide information about given area.**

Keywords: **web information extraction, top-k list, web mining and dynamic web.**

1. INTRODUCTION

In today's fast world everybody needs quick and accurate results. Most of the data on the web is not concerned about the topic of user's interest. Therefore extraction the meaningful information from the web is difficult as the data on the web page might be in structured, semi-structured or non-structured form. Hence there is requirement of top-k pages for information extraction for following reasons:

1) Top k data on the web is wide.

2) Top k data is of high quality.

3) Top k data is ranked.

This system also supports Dynamic Web system is a web application which is useful for the online forums which provide information about given area. This system provides location base information. User only views information of the area from where he/she login or open's the website. System use Service provider (SP) system to access the area information.

This system for dynamic web is totally automated and does not require any human interaction to update content of the system. System contains artificial bots which provide live update to system. Information of web system is automatically updated when the owner of the client website update its website contents. System use different web mining technology to extract data from remote website at run time. System contain different client like city malls, news and hospitals. User can easily access all information from this application by just login to system.

2. PROBLEM STATEMENT

In this section the formal problem definition of extraction of top k lists from page is given.

Let the web page be a pair (title, body) where 'title' is the title of page and 'body' is the HTML body of page. The page (title, body) is a top-k page if:

1)From a 'title' we can extract a 5-tuple (k ,c ,m ,t ,l) where 'k' is a natural number, 'c' is a concept or the information, 'm' is the ranking criterion, 't' is temporal information and 'l' is location information.

2) From the 'body' we can extract k and only k items such that:

a) Each tuple represents an entity that is an instance of the concept c.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 1, pp: (374-376), Month: January - March 2015, Available at: www.researchpublish.com

b) The pair wise syntactic similarly of the k items is greater than threshold.

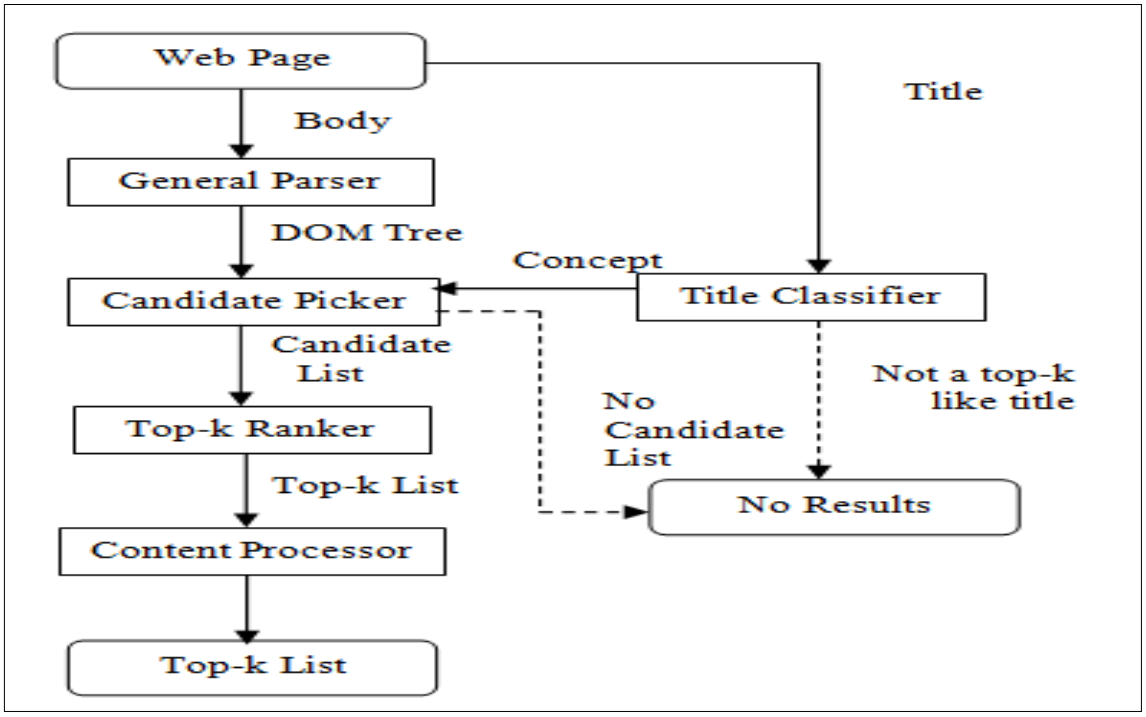
The top-k extraction problem can be defined as three sub problems:

1) Title recognition titlerec: (title, body) -> (k, c, m, t, l)

2) List extractor listext: (k, c, d) -> ***I*** where ***I*** is the set of terms which are instance of c and | ***I*** |=k.

3) Content extractor contentext: (c, d, ***I***) -> (T, schema) where T is a table attribute values for the element in ***I*** and s is its schema.

3. PROPOSED SYSTEM



System outline

4. IMPLEMENTATION DETAILS

1) Title classifier: The title of a web page (string enclosed in <title> tag) helps us identify a top-k page. There are several reasons for us to utilize the page title to recognize a top-k page.

2) Candidate picker: This step extracts one or more list structures which appear to be top-k lists from a given page.

3) Top k Ranker [1]: Top-K Ranker ranks the candidate set and picks the top ranked list as the top-k list by a scoring function which is a weighted sum of two feature scores below:

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

$$\textbf{P-score}=\frac{1}{k}\sum_{n\in L}\frac{LMI(n)}{Len(n)}$$

b) V-score: V -Score calculates the visual area consumed by a list, since the main list of the page tends to be larger and more prominent than other minor lists. The V –Score of a list is the sum of the visual area of each node and is computed by:

$$\textbf{Area}(\textbf{L})=\sum_{\textbf{n}\in\textbf{L}}\textbf{TextLength}(\textbf{n})\times\textbf{Fontsize}(\textbf{n}^2)$$

4) Content processor: After getting top-k list, we extract the attributes and their corresponding values from the given concept. The goal is to obtain structured information for each item.

5. RELATED WORK

The problem of obtaining top-k list, presented in this system, is a part of the general area of web structured data extraction, where many techniques have been developed and improved recently. Google Tables and Web Sets extract web lists or tables based on very specific list-related tags, such as <UL>, <DL>,<TABLE>. The data is extracted based on the similarity between DOM, which is measured by edit distance. [1]

Although this system is inspired by some of approaches but it has several major differences:

1) Different goals: The previous approaches is to extract all the lists and tables from the web page but this system extract specific lists which are of the users interest as well as it provides the information about the given area.

2) Use of number k: Due to number k, it is easy to extract the information in limited fashion that means it will only give the output of list of k items.

6. CONCLUSION

We are trying to build a system which will provide top ranked list of k items. In our system we are going to use N gram model and clustering which will reduce the time complexity and enhance the performance of the system as compared to the previous one.

REFERENCES

[1] Zhixian Zhang, Kenny Q. Zhu , Haixun Wang , Hongsong Li ,"Automatic Extraction of Top-k Lists from the Web",, IEEE , ICDE Conference, 2013, 978-1-4673-4910-9.

[2] G. Miao, J. Tatemura, W.-P. Hsiung, A. Sawires, and L. E. Moser,"Extracting data records from the web using tag path clustering".

[3] Z. Zhang, K. Q. Zhu, and H. Wang, "A system for extracting top-k lists from the web," in KDD, 2012

[4] C.-H. Chang and S.-C. Lui, "Iepad: information extraction based on pattern discovery," in WWW, 2001, pp. 681–688.

[5] W. Wu, H. Li, H. Wang, and K. Q. Zhu, "Probase: A probabilistic taxonomy for text understanding," in SIGMOD, 2012.

[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] 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.