Extract The Target List With High Accuracy From Top-K Web Pages

[1]Mr.Naveen Kumar.N[2]Usha Rani.G

[1]AssistantProfessor, Department of CSE,School of Information Technology JNTUH, Kukatpally, Hyderabad, Telangana, India

[2]M.TechStudent, Software Engineering, School of Information Technology JNTUH, Kukatpally, Hyderabad, Telangana, India

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 anduseful as a preprocessing step to produce facts for a fact answering engine.

1.INTRODUCTION

The World Wide Web is by far the largest source of information today. Much of that information contains structured data such as tables and lists which are very valuable for knowledge discover age and data mining. This structured data is valuable not only because of the relational values it contains, but also because it is relatively easier to unlock information from data with some regular patterns than free text which makes up most of the web content. However, when encoded in HTML, structured data becomes semi-structured. And because HTML is designed for rendering in a browser, different HTML code segments can give the same visual effect at least to the human eye. As a result, HTML coding is much less stringent than XML, and inconsistencies and errors are abundant in HTML documents. All these pose significant challenges in the extraction of structured data from the web . In this demo, we focus on list data in web pages. In particular, we are interested in extracting from a kind of web pages which present a list of k instances of a topic or a concept. Examples of such topic include “20 Most Influential Scientists Alive Today”, “Ten Hollywood Classics You Shouldn’t Miss”, “50 Tallest Persons in the World”. Figure 1.(a) shows a snapshot of one such web page [2]. Informally, our problem is: given a web page with a title that contains a integer number k (Figure1.(b)), check whether the page contains a list of k items as its main content, and if it does, extract these k items. We call such lists top-k lists and pages that contain the entirety of a top-k list top-k pages. There are also lists that span multiple pages but are connected by hyperlinked “Next” button, but these are not considered in this work. A typical scenario we consider each list item is more than an instance of the topic in the title, but instead contain additional information such as a textual description and images. Our objective is to extract the actual instance whenever possible, but in the worst case, produce list items that at least contain the wanted instances. The work described in this work is an important step in our bigger effort of building an effective fact answer engine With such an engine, we can answer queries such as “Who are the 10 tallest persons in the world”, or “What are 50 bestselling books in 2010” directly, instead of referring the users to a set of ranked pages like all search engines do today. Because of the special relation between the page title and the main list contained in the page, the semantics of the list items are more specific and hence it’s much easier for us to integrate the extracted lists into a general knowledge base that empowers the fact answer engine. There were many previous attempts to extract lists or tables from the web. None of them targets the top-k list extraction that is studied in this work. In fact, most of the methods are based on either very specific list-related tags [4], [5] such as <ul>,<li> and <table>or the similarity between DOM trees [6], [7] and completely ignore the visual aspect of HTML documents. These approaches are likely to be brittle because of the dynamic and inconsistent nature of web pages. More recently several groups have attempted to leverage visual information in HTML in information extraction. Most notably, Ventex [8] and HyLiEn [9] are designed to correlate the rendered visual model or features with the corresponding DOM structure and achieved remarkable improvements in performance. However, these techniques indiscriminatingly extract all elements of all lists or tables from a web page, therefore the objective is different from that of this work which is to extract one specific list from a page while purging all other lists (e.g. (d) and (e) in Figure 1) as noise. The latter poses different challenges such as distinguishing ambiguous list boundaries and identifying noise and unwanted lists. The main approach of this work goes along the line of analyzing similar tag paths in the DOM tree with the help of visual features to identify the main list in a page. The key contributions of this demo are:

- ➢ We defined a novel top-k list extractproblem which is useful in knowledge discovery and fact answering;
- ➢ We designed an unsupervised general-purpose algorithm along with a number of key optimizations that is capable of extracting top-k lists from any web pages ;
- ➢ Our evaluation shows that ouralgorithm scales with the data size and achieves significantly better accuracy than competing methods.

2. RESEARCH

1. An Automatic Extraction of Top-k Lists from the Web

Important source of structured information on the web is links. This paper is concerned with “top-k list” pages, which are web pages that specify a list of k instances of a particular query. Examples include “the 10 tallest building in the world” and “the top 20 best cricket players in India”. We present an efficient algorithm that fetch the target lists with great accuracy even when the input pages contain other non-useful data of the same size or errors. The extraction of such lists can help develop existing knowledge bases about general consideration.

2.Automatic Extraction Of Data From Deep Web Page

There is large amount of information accessible to be mined from the World Wide Web. The information on the Web is in the form of structured and unstructured objects, which is known as data records. Such data records are necessary because essential information are available in these pages, e.g.

lists of products and there detail information. It is important to extract such data records to provide proper information to user as per their concern. Manual approach, supervised learning, and automatic techniques are used to solve these problems. The manual method is not relevant for huge number of pages. It is a challenging work to retrieve appropriate and beneficial information from Web pages. Presently, numbers of web retrieval systems called web wrappers, web crawler have been invented. In this paper, some current techniques are inspected, then our work on web information extraction is presented. Experimental analysis on large number of real input web URL address selections indicates that our algorithm properly extracts data in most cases.

3.Survey on web mining techniques for Extraction of top k list

Today finding proper result within less time is important need but one more problem is that very poor percentage information available on web is useful and interpretable and which consumes lot of time to extract. The method for extracting information from top k web pages which contains top k instances of interested topic needed to deals with system. In contrast with other structured data like web tables Information in top-k lists contains valuable and exact information of rich, and interesting. Therefore top-k list are of higher quality as it can help to develop open domain knowledge bases to applications such as search for truth result.

4.Extracting general from web document:

In this paper, author proposed a new different technique for extraction of general lists from the web. Method uses basic premises on visual rendering of list and structural arrangement of items. The aim of system was to minimize the restrictions of existing work which deals with the principle of extracted lists. Several visual and structural features were combined for obtaining goal.

EXISTING SYSTEM

Existing method is used to retrieve the general information from the web. It can be applied on the structured data. Method uses basic premises on visual rendering of list and structural arrangement of items. . The aim of system was to minimize the restrictions of existing work which deals with the principle of extracted lists. Several visual and structural features were combined for obtaining goal.

DISADVANTAGES:

It will retrieve the target from the structured data.

PROPOSED SYSTEM

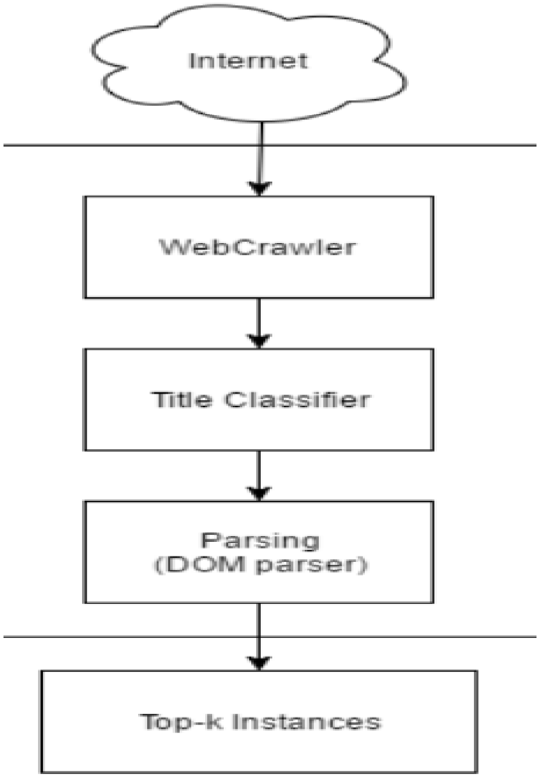
This paper is concerned with information extraction from top-k web pages, which are web pages that describe top k instances of a topic which is of general interest. We present an efficient method that extracts top-k lists from web pages with high performance.

ADVANTAGES:

1. It will retrieve the target data from all types of data.

2. Accurate result will come from top-k.

3.SYSTEM ARCHITECTURE:



4. TECHNICAL SPECIFICATION :

1.WebCrawler:

A **Web crawler**, sometimes called a **spider**, is a internet bot that systematically browses the world wide web, typically for the purpose of web indexing (*web spidering*).

Web search engines and some other sites use Web crawling or spidering software to update their web content or indices of others sites' web content. Web crawlers copy pages for processing by a search engine which indexes the downloaded pages so users can search more efficiently. Crawlers can validate hyperlinks and HTML code. They can also be used for web scraping.

2.Title Classifier

The title of a web page helps us identify a “top k” page. This title is enclosed in<title> tag in HTML page. The aim of the classifier is to identify “top-k like” titles, the probable name of a top-k page. This title represents the topic of “top-k” list.Fig.2 shows the flow of an title classifier.

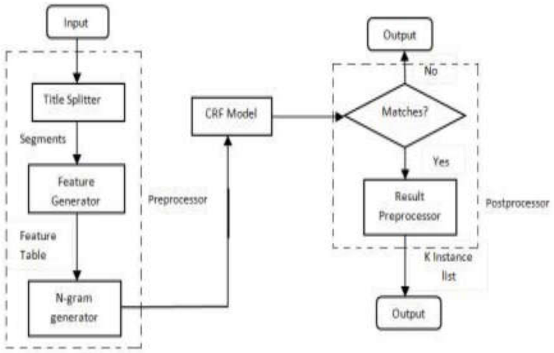


Fig :Flowchat of Title Classifier

A “top-k like” title contains more than one segments, which are separated by separator. From this segments only one segment describes topic of the page and rest of the segments shows the additional information. We therefore split the title in to number of segments. We trained a Conditional Random Fields (CRF) [17] model from both positive and Negative sample titles to recognize “top-k like” title. The classifier also transfers the cardinal digit word (word like ten or fifteen) into the number k, and outputs a set of concepts which are mentioned in the title.Fig.3 shows a sample of “top-k” title

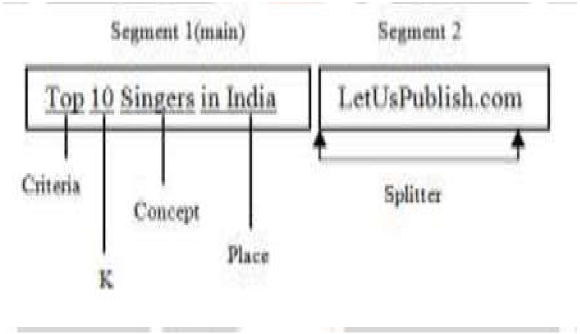


Fig:Sample of Top-K Title

4. IMPLEMENTATION

a. ADMIN MODULE:

Admin needs to login first with using his user name and password. He had the rights to delete the data from the server and he will monitor the user activities in searching. He can view the users public details such as name, user name, email id, location and date of registration.

b. USER MODULE:

User should register first in the site to access his webpage. He can search the content and images from the top-K list. He can upload the data.

C. LEVENSHTEIN DISTANCE ALGORITHM:

The Levenshtein distance algorithm is used to calculate distance between two sequences. Basically, the Levenshtein distance between two words is the less number of single-character edits needed to change one word into the other word. Levenshtein distance also is called as edit distance. It also denotes a larger family of distance metrics and it is closely related to pairwise string alignments.

5. CONCLUSION

Finally we would like to conclude that we have implemented the extraction of top-k list from the web. For finding out the top-k list. Also we would like to conclude that compared to other structure data top-k list are cleaner, easier to understand and more interesting for human consumption and therefore are an important source for data mining and knowledge discovery. Our project can be extended to find information from links present inside other links and try to reduce computational work.

6. REFERENCES

[1] T. Weninger, F. Fumarola, R. Barber, J. Han, and D. Malerba, "Unexpected results in automatic list extraction on the web," SIGKDD Explorations, vol. 12, no. 2, pp. 26–30, 2010. [2] "20 most influential scientists alive today," http://www.superscholar.org/ features/20-mostinfluential-scientists-alive-today/. [3] X. Yin, W. Tan, and C. Liu, "Facto: a fact lookup engine based on web tables," in WWW, 2011, pp. 507–516. [4] "Google sets," http://labs.google.com/sets. [5] 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. [6] B.Liu, R. L. Grossman and Y. Zhai, "Mining data records in web pages," in KDD, 2003, pp. 601– 606. [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. [8] W. Gatterbauer, P. Bohunsky, M. Herzog, B. Kr¨upl, and B. Pollak, "Towards domainindependent information extraction from web tables," in WWW. ACM Press, 2007, pp. 71–80. [9] 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. [10] Y. Yamada, N. Craswell, T. Nakatoh, and S.Hirokawa,

"Testbed for information extraction from deep web," in WWW, 2005.