Volume : 2 | Issue : 11 | Nov 2013 • ISSN No 2277 - 8160

Research Paper | Engineering

Query Based Effective Data Extraction and Alignment Method

Mrs. Sahunthala. S	M.E(Asst Professor),AIHT,Chennai,Tamil Nadu,India
Ms. Bharathi. R	B.Tech(Student),AIHT,Chennai,Tamil Nadu,India
Ms. Brindha. R	B.Tech(Student),AIHT,Chennai,Tamil Nadu,India

ABSTRACT

Web databases generate query result pages based on a user's query. Automatically extracting the data from these query result pages is very important for many applications, such as data integration, which need to cooperate with multiple web databases. We present a novel data extraction and alignment method called CTVS that combines both tag and value similarity. CTVS automatically extracts data from query result pages by first identifying and segmenting the query result records (QRRs) in the query result pages and then aligning the segmented QRRs into a table, in which the data values from the same attribute are put into the same column. The drawback in the existing system is CTVS does not handle the case where multiple data values from more than one attribute are clustered inside one leaf node of the tag tree. In this paper, we overcome this limitation. We also design a new record alignment algorithm that aligns the attributes in a record, first pairwise and then holistically, by combining the tag and data value similarity information.

KEYWORDS : Page Ranking, Query Result Records, Clustering Methods, QRR extraction

1 Introduction

Online databases, called web databases, comprise the deep web. Compared with webpages in the surface web, which can be accessed by the unique URL, pages in the deep web are dynamically generated in response to a user query submitted through the query interface of a web database.

Many web applications such as Meta querying, data integration and comparison shopping, need the data from multiple web databases. The goal of wed database data extraction is to remove any irrelevant information from the query result pages, extract the query result records from the page, and align the extracted QRRs into a table such that the data values belonging to the same attribute are placed into same table column.

1.Recordextractionprocess identifies the QRRs in p and involves two sub steps: Data region Identification and the actual segmentation step.

2. Record alignment process aligns the data values of the QRRs in pinto a table so that data values for the same attribute are aligned into the same table column.

CTVS improves data extraction accuracy in three ways.

1. Efficiency of the new techniques are proposed to handle the case when the QRRs are not contiguous in p, which may be due to the presence of auxiliary information. Assume that the QRRs are presented contiguously in only one data region in a page.

a. An adapted data region identification method is proposed to identify the non- contiguous QRRs that have the same parents according to their tag similarities.

b. A merge method is proposed to combine different data regions that contain the QRRs (with or without the same parent) into a single data region. Our experimental results show that the two techniques are effective for addressing the non- contiguous data region problem.

2. A contemporary alignment algorithm has been introduces in order that it provides an efficient aligning process, first pair wise then holistically, so that they can be put into a table with the data values belonging to the same attribute arranged into the same table column.

3. A new nested-structure processing algorithm is proposed to handle any nested structure in the QRRs after the holistic alignment. Unlike existingnested- structure processing algorithms that rely on only tag information, the updated CTVS method uses both tag and data value similarity information to improve nested- structure processing accuracy.

In this paper the problem is focused on extracting the data from web pages and alignment of data.

PAGE RANKING

Page Rank is an important part of search engine optimization (SEO) past and present history. Page Rank is a link analysis algorithm that assigns a number or rank to each hyperlinked web page within the World Wide Web. The basic purpose of Page Rank is to list web pages from the most important to the least important, reflecting on a search engine results page when a keyword search occurs. The basic process involves Page Rank evaluating all of the links to a particular web page. If a web page has a lot of links from large websites that also rank well, then the original web page is given a high ranking. Higher ranking in Page Rank equates to a greater probability of the site being reached because they are not only quantitatively linked to a number of times, but are also linked to other popular web pages. We have also used page ranking method for speeding up the search engine in order to extract the data records rapidly in order to reduce the extraction time of the data records.

Inktomi, Altavista, et al. use link popularity in their ranking algorithms, but to a much lesser extent than Google which has feature called Web Rank.

CLUSTERING METHODS

Finding the group of objects such that the objects in a group will be similar or related to one another and different groups from the objects will be put into another group. In an existing CTVS method Page ranking and clustering algorithms have not been introduced. Enhancement of these algorithms will give an extra value to the web pages as per the page ranking algorithm and similar pages will be clustered in order to extract the needed information immediately which reduces the time consumption.

2 Clustering in Data Mining

DESIGN CONSIDERATION

Design consideration of this project includes a novel data extraction method, CTVS, to automatically extract QRRs from a query result page.

Tag tree construction module has been implemented in order to check from the parent root tag and traversing along the child node path and reaches the bottom most tag or node. By having similar tags and also the similarity between the values of two data regions have been obtained and arrangements of those values have implemented to perform effective alignment or aggregation method.

QUERY PAGES AND QRR EXTRACTION

Each node represents a tag in the HTML page and its children are tags

enclosed inside it. The Record Segmentation module then segments the identified data regions into data records according to the tag patterns in the data regions. Given the segmented data records, the Data Region Merge module merges the data regions containing similar records[1]. Overcoming the problems of existing system this method involves in combining the tags.

3 QRR Extraction Frame work

Final Aligned R esult for Fig. 1

QRR 1	Imag e 1	DNK Y	Siz e:	Que en	Colo r:	Blue	$55
QRR 2	Imag e 2	DNK Y	Siz e:	Que en	Colo r:	Red	$56
QRR 2	Imag e 2	DNK Y	Siz e:	Kin g	Colo r:	Gree n	$57

3.1 Data mining and Region Identification

The similar data records are typically clustered using several clustering techniques like distance based clustering and hierarchical clustering etc.

3.2 Identification of Data Region Merge

After the clustering methods have been introduced we need to determine whether any of the data regions should be merged. Given any two data regions, we treat them as similar if the segmented records they contain are similar. The similarity between any two records from two data regions is measured by the similarity of their tag strings. The similarity between two data regions is calculated as the average record similarity. Two data regions can be merged into a merged data region if the records in the two data regions have an average similarity greater or equal to 0.6, which is a threshold used to judge whether two records are similar in regions. The similar data records are typically clustered using several clustering techniques like distance based clustering and hierarchical clustering etc.

Data Extraction method

Method	Nested Structure Processing	Single Result Page	Non Contiguous Data Regions
CTVS	Yes	Yes	Yes
DeLa	Yes	Yes	No
ViNTs	No	No	No

A specific page's relevance ranking for a specific query currently depends on three factors:

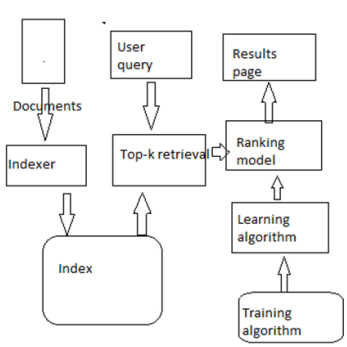
- Its relevance to the words and concepts in the query
- Its overall link popularity
- Whether or not it is being penalized for excessive search engine optimization (SEO).

Keyword Search

Most search engines handle words and simple phrases. In its simplest form, text search looks for pages with lots of occurrences of each of the words in a query, stopwords aside.

The more common a word is on a page, compared with its frequency in the overall language, the more likely that page will appear among the search results.

Good Search Engine Rank is extremely important for the success of your Internet Marketing Business.



Techniques to improve search engine ranking

1.Focus on keyword

Choose Your Keyword: The first and most important step is to choose the right keywords for which you wish to rank well. You have to focus on this because all your efforts are going to be centered around this. While doing this for your home page it is important to adopt the same procedure and target different keywords for different pages of your site. This is because search engines match each query with the most relevant page and not the site.

Target The Right Keyword: Optimizing your site for a popular keyword and securing a high page rank is not enough. You should optimize for the right keyword that is most relevant to your site otherwise your efforts will be in vain.

The keywords that you choose should be specific. Since there is fierce competition for popular single word keywords, it is better to choose two or three word phrases that specifically describe your site content.

Though you might obtain a good page rank with a very well optimized site and it appears high in the results, if no one searches for them in the search engines you are not going to enjoy the benefits of a high PR. so choose one that is reasonably popular and ones that people actually type.

2.Integrate keyword

Once the right keywords relevant to your site are chosen, you should now integrate these keywords into your site, placing very special attention to places where Search Engines look.

URL:It is very important that you have at least one of the important keywords in the URL.This will help the search engine to understand what your site is all about.

Page Title:To improve page rank the page title is another important place to integrate your keywords. Since this too gives an indication to the search engine useful information about your site. Though most sites use the same title for all the pages, it is better to have different titles with specific keywords integrated into it, reflecting the content in that particular page.

Page Content:It is a well known fact that the page content or text is most important to search engines. So to improve page rank you should focus all your attention to ensure that the targeted keywords are well represented in your content. Unless these keywords appear in the text your chances of obtaining a good page rank will be remote.

In addition to having your keywords in the text of the page three very important criteria should also be adhered to, which is very often overlooked by most webmasters.

a. Use keywords frequently but do not stuff your page with keywords. Search Engines do not approve keyword stuffing.
b. Place your keywords close to each other. This helps the the search engines.
c. Place the keywords near the top of your page in the page content.

3.Obtain link from sites.

To improve search engine rank another important thing you should do is to get other sites to link to you, because the number of links, linking to your site is a measure of popularity. Most search engines consider

link popularity as the most important factor in deciding the relevance of a site.

It becomes obvious that you should put in a lot of effort to get as many sites as possible to link to you.If it is a new site get a link from a website that has already been listed in search engines.

Search Engines look at not only the number of sites that link to you but also the quality of the sites that link to you. In other words a link from a popular and relevenant P7 site is much better than several links from smaller sites. Hence quality is considered more important than quantity. A judicious mix of quality and quantity will help improve search engine rank.

4.Meta data.
It is an embedded HTML tag that contains descriptive information, keywords and other information about a web page and does not appear when the web page is displayed in a browser.

There was a time when search engines relied on Meta tags to assess the site¿s relevance and rank them. Most Search Engines no longer rely on meta tags since they have developed new systems for this purpose. Nevertheless it is advisable to have them to improve search engine rank.

5.Alt tag.
Another important aspect that is often overlooked is the importance of ALT tags. Search Engines don't index images. They can't read them; so to fix this problem there are ALT tags which are descriptions of images that are recognized by search engines. So to get the search engines to recognize the images on your site, ALT tags with keywords can be used which will in turn help you to improve your rank in search engines.

Finally a word of advice, do not try to trick the search engines by using unacceptable techniques because it will not pay in the long run and you could get penalised. By following these methods you will succeed in your attempt to improve search engine rank yourself and be happy and proud of your achievement. Advice from a Specialist will certainly be of help if your intention is to reach the top.

Concepts in page ranking
Scoring: The major search engines use hundreds of factors nestled into many algorithms. Think about it like an onion and it's layers. All too often, people say things like "the Google algorithm" when in fact, there are many. The scoring over all of them makes up the initial rankings. matched with the user query and to discard the auxiliary information.

Scoring: The major searchengines use hundreds of factors nestled into many algorithms. Think about it like an onion and it's layers. All too often, people say things like "the Google algorithm" when in fact, there are many. The scoring over all of them makes up the initial rankings.

Boosting: this is another element or signal that might raise a page's position in the rankings. One example is a statement Google made that fast mobile sites are given a boost in mobile search. Various forms of personalization also use a boosting element to re-rank results.

Dampening: Not to be confused with penalties, a dampening factor is an element that would lower the rankings of a web page after the initial scoring process. One example is the now infamous Google Penguin or Panda algorithms. While it may seem like a penalty, it is in fact a dampening element.

4 Conclusion
All the information related to the user queries are retrieved based on the page ranking of the data containing the page. The paper describes how the data can be matched with the user query and to discard the auxiliary information.

REFERENCES

[1] R.Baeza-Yates,"Algorithm for String Matching: A survey", ACM SIGIR Forum, vol. 23, nos. ¾, pp,34-58, 1989. | [2] B. Liu and Y.Zhai," NET- A system for Extracting Web Data from Flat and Nested Data Records", Proc. Sixth Int'l Conf. Web Information Systems Eng., pp. 487- | 495,2005. [3] Ivan Ivanov, Peter Vajda, Pavel Korshunov, and Touradj Ebrahimi, "Comparative Study of Trust Modelling for Automatic Landmark Tagging", vol.8,no.6, June 2013. | [4] J.Wang and F.H.Lochovsky, "Data Extraction and Label Assignment for Web | th | | Databases" Proc,12 | | World Wide Web | | | Conf., pp. 187-196,2003. |