ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

Automatic Interpretation of Search results from Search engines for Machine Processing

Shaik Haseena
M.Tech, Student,
Computer Science Engineering Department,
Rao & Naidu Engineering College, Ongole.

N.Venkateswararao
Asst Professor, HoD,
Computer Science Engineering Department,
Rao & Naidu Engineering College, Ongole.

Abstract:

The word "Search engine" consists of two words: Search means to find something and engine means the procedures that find the specified information. So it's meaning can be clearly understood from its name. i.e. a search engine is a utility that provides the uses to find any information on the World Wide Web within a few seconds. Essentially search engines provide easy access to large databases of information. Essentially, the Internet is one very large database. It is not possible to scroll through or alphabetize every web page on the Internet. For this reason, dynamic search engines provide relevant results to search queries.

An increasing number of databases have become web accessible through HTML form-based search interfaces. The data units returned from the underlying database are usually encoded into the result pages dynamically for human browsing. For the encoded data units to be machine process able, which is essential for many applications such as deep web data collection and Internet comparison shopping, they need to be extracted out and assigned meaningful labels. In this paper, we explored a dynamic Interpretation method that initially arranges the search data units on a result page into array of groups.

The data with the same meaning is placed in the same group. Then, we try to understand each group from various perspectives and criteria, before giving a final label to it. An interpretation wrapper for the search site is dynamically assembled and can be utilized to interpret updated result pages from the same web database. Our survey shows that the projected method is extremely required in the current scenario of Internet shopping boom in India. Our experiments indicate that the proposed approach is highly effective.

Keywords:

web database, Data interpretation, Search, dynamic search, Data arrangement.

Introduction: Internet search engines are special sites on the Web that are designed to help people find information stored on other sites. There are differences in the ways various search engines work, but they all perform three basic tasks:

1.They search the Internet -- or select pieces of the Internet -- based on important words.

2.They keep an index of the words they find, and where they find them.

3.They allow users to look for words or combinations of words found in that index.

Early search engines held an index of a few hundred thousand pages and documents, and received maybe one or two thousand inquiries each day. Today, a top search engine will index hundreds of millions of pages, and respond to tens of millions of queries per day.

A web database is a system for storing information that can then be accessed via a website. For example, an online community may have a database that stores the username, password, and other details of all its members. The most commonly used database system for the internet is MySQL due to its integration with PHP — one of the most widely used server side programming languages.

At its most simple level, a web database is a set of one or more tables that contain data. Each table has different fields for storing information of various types.

These tables can then be linked together in order to manipulate data in useful or interesting ways. In many cases, a table will use a primary key, which must be unique for each entry and allows for unambiguous selection of data.

Now lets observe what happens when these search engines come across deep Web databases. Web search engines use Web crawling or spidering software to update their web content or indexes of others sites' web content. Web crawlers can copy all the pages they visit for later processing by a search engine that indexes the downloaded pages so that users can search them much more quickly.

The search results are usually displayed in a column of results frequently termed to as search results record (SRR). Web database has numerous search result records. Each search results record (SRR) refers to a specific entity or group. Search results records (SRR) from web database have numerous data units. Data units are texts that correspond to the single group having similar meaning. Here the interpretation of data is done on the basis of data units. The data units are interpreted by allocating labels to them.

Dynamic Interpretation solution of search results record (SRR) is carried out in three phases.

Alignment/Arrangement phase: In this phase we first recognize all the data units and categorize them in to array of groups. Grouping data units depending on their meanings helps to make out the frequent patterns among these data units which is the basis for interpretation.

Interpreter Phase/Annotator phase: Each basic annotator/interpreter is used to label the units of same group. It is also used for recognizing best suitable label for each specific group.

Wrapper generation phase: In this phase an interpretation rule is created for every recognized idea which demonstrates how to take out data units of same group. The collective interpretation rule for associated groups is identified as interpretation wrapper for the matching web database. This annotation/ interpretation wrapper is utilized to interpret the data units for diverse queries without creating alignment and interpretation phase.

As a result interpretation is done a lot quicker.

Existing System:

In this existing system, a data unit is a piece of text that semantically represents one concept of an entity. It corresponds to the value of a record under an attribute. It is different from a text node which refers to a sequence of text surrounded by a pair of HTML tags.

It describes the relationships between text nodes and data units in detail. In this paper, we perform data unit level annotation.

There is a high demand for collecting data of interest from multiple WDBs. For example, once a book comparison shopping system collects multiple result records from different book sites, it needs to determine whether any two SRRs refer to the same book.

Disadvantages Of Existing System:

If ISBNs are not available, their titles and authors could be compared. The system also needs to list the prices offered by each site. Thus, the system needs to know the semantic of each data unit. Unfortunately, the semantic labels of data units are often not provided in result pages.

For instance, no semantic labels for the values of title, author, publisher, etc., are given. Having semantic labels for data units is not only important for the above record linkage task, but also for storing collected SRRs into a database table.

Proposed System:

We put forward a system to arrange data units into various groups, Groups are formed such that units with similar meaning are placed in the same group.

Replacing the existing system of allocating labels to every HTML text node, we propose to take into account additional significant characteristics common to data units, which are: data types (DT), data contents (DC), presentation styles (PS), Tag Path and adjacency (AD) information.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

Data type:

Data types are predefined features that have their own meaning. Fundamentally used data types are date, time, currency, integer, decimal etc.
Data content: Data unit or text node of similar idea shares certain keywords which are utilized to search for the information swiftly. For e.g., keyword "Oxygen" will return the data that are relevant to word Oxygen.

Presentation style:

Presentation characteristics illustrate how a data unit is shown on a web page. Example:font face, font size, colour, text decoration etc.

Tag path:

A Tag path is a series of tags that range from the very root of the search results record (SRR) to the matching node in the tree. Every node has two parts a tag name and a direction signifying whether the subsequent node is a sibling or the first child node.

Adjacency:

Adjacency refers to the data units that are immediately before and after in the search results record (SRR). They are termed as preceding and succeeding data unit. For example: Andhra Pradesh and Assam are both states in alphabetical order but do not share content keywords.

Advantages Of Proposed System:

This paper has the following contributions:

• While most existing approaches simply assign labels to each HTML text node, we thoroughly analyze the relationships between text nodes and data units. We perform data unit level annotation.

• We propose a clustering-based shifting technique to align data units into different groups so that the data units inside the same group have the same semantic. Instead of using only the DOM tree or other HTML tag tree structures of the SRRs to align the data units (like most current methods do), our approach also considers other important features shared among data units, such as their data types (DT), data contents (DC), presentation styles (PS), and adjacency (AD) information.

• We utilize the integrated interface schema (IIS) over multiple WDBs in the same domain to enhance data unit annotation. To the best of our knowledge, we are the first to utilize IIS for annotating SRRs.

• We employ six basic annotators; each annotator can independently assign labels to data units based on certain features of the data units. We also employ a probabilistic model to combine the results from different annotators into a single label.

This model is highly flexible so that the existing basic annotators may be modified and new annotators may be added easily without affecting the operation of other annotators.

• We construct an annotation wrapper for any given WDB. The wrapper can be applied to efficiently annotating the SRRs retrieved from the same WDB with new queries.

Problem Statement:

Basically in every search engines just shows the web content and web links related to our input in the search box. It is just a text node which refers to a sequence of text surrounded by a pair of HTML tags. There is no the relationship between text nodes and data units. In this paper, we perform data unit level annotation.

Scope:

The scope of the project is when we search any content in a search engine, it will group the content into different category related to what we are searching about and also provides data unit level annotation which means order or group the content which belongs to our wish.

Alignment Algorithm:

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

```
ALIGN(SRRs)
j ← 1;
while true
    //create alignment groups
    for i ← 1 to number of SRRs
        Gj ← SRR[i][j];      //jth element in SRR[i]
    if Gj is empty
         exit; //break the loop
    V ← CLUSTERING(G);
    if |V| > 1
         //collect all data units in groups following j
       S ← Ø;
       for x ← 1 to number of SRRs
           for y ← j+1 to SRR[i].length
               S ← SRR[x][y];
       //find cluster c least similar to following groups
       V[c] = min_{k=1 to |V|} (sim(V[k], S));
       //shifting
       for k ← 1 to |V| and k ≠ c
           foreach SRR[x][j] in V[k]
               insert NIL at position j in SRR[x];
    j ← j+1;          //move to next group

CLUSTERING(G)
V ← all data units in G;
while |V| > 1
    best ← 0;
    L ← NIL; R ← NIL;
    foreach A in V
        foreach B in V
            if ((A != B) and (sim(A, B) > best))
                best ← sim(A,B);
                L ← A;
                R ← B;
    If best > T
        remove L from V;
        remove R from V;
        add L ∪ R to V;
    else break loop;
return V;
```

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

Alignment algorithm has following four steps:

Step 1:

Merge text nodes: This step detects and removes decorative tags from each SRR to allow the text nodes corresponding to the same attribute merge into a single one.

Step 2:

Align text nodes: After the merging aligns text nodes into different groups. So that same group has the same concepts.

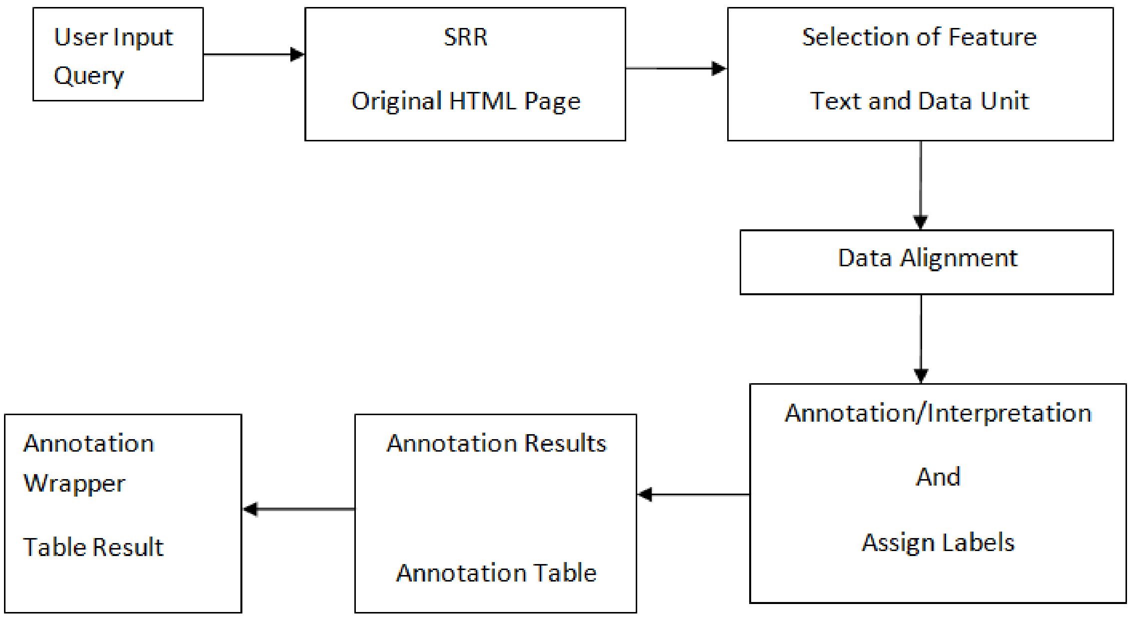
Step 3:

Split text nodes: In this step split the composite text nodes into separate data unit.

Step 4:

Align data units: This is the last step for alignment in which separates each composite group into multiple aligned groups with each containing the data units of the same concept

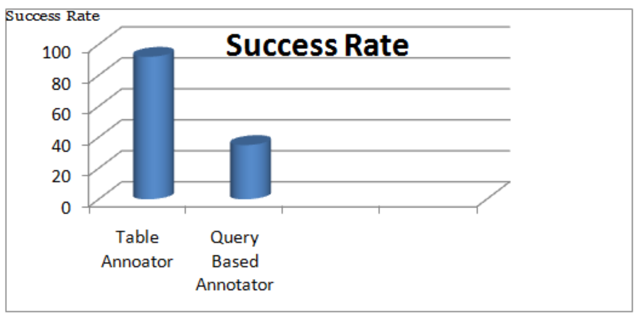
Fig: Interpretation/Annotation Architecture for results from SRR



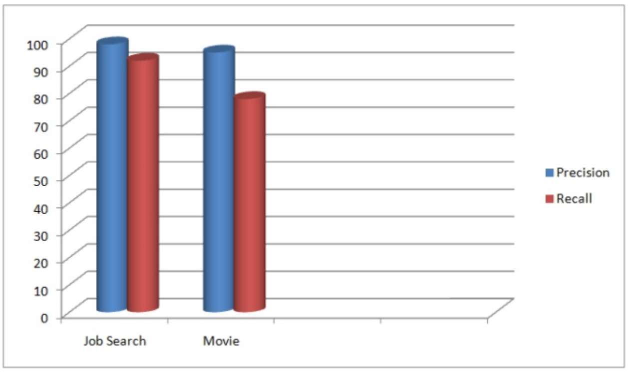
Survey Conducted:

We have conducted a web survey and observed that there is great need for this kind of system in ever increasing ecommerce industry in India. Applications for this system can be used for searches for Hotel rooms, Holiday Packages, online shopping for merchandise, books and Magazines etc.

This can also be used in online sales for cars, vehicles, motor insurance etc. We can also use this system for comparing similar products from various web databases.



ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal



CONCLUSION:

In this paper we addressed on the problem of annotating/Interpretation of search results. The search results of search engines form web databases which can be utilized for additional processing in order to use them in different applications like content evaluation, data mining etc.

We developed a software application that enables users to give a query, and then the query is dynamically submitted to search engine.

The results of search engine are processed in the three phases. The phases are alignment phase, annotation phase and wrapper generation phase.

Then, the application gives results which are nothing but the annotated/interpreted documents. HTML tags are employed to process the web pages while annotating them. The interpreted results are useful in real world applications. Accurate alignment is critical to achieving holistic and accurate annotation.

Our method is a clustering based shifting method utilizing richer yet automatically obtainable features. This method is capable of handling a variety of relationships between HTML text nodes and data units, including one-to-one, one-to-many, many-to-one, and one-to-nothing.

Our experimental results show that the precision and recall of this method are both above 98 percent. There is still room for improvement in several areas as mentioned. For example, we need to enhance our method to split composite text node when there are no explicit separators. We would also like to try using different machine learning techniques and using more sample pages from each training site to obtain the feature weights so that we can identify the best technique to the data alignment problem.

REFERENCES:

[1] Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng and Clement Yu, (2013). Annotating Search Results from Web Databases. IEEE Transactions On Knowledge And Data Engineering, Vol. 25, NO. 3.p1-14.

[2] N. Krushmerick, D. Weld, and R. Doorenbos, "Wrapper Induction for Information Extraction," Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI), 1997.

[3] L. Liu, C. Pu, and W. Han, "XWRAP: An XML-Enabled Wrapper Construction System for Web Information Sources," Proc. IEEE16th Int'l Conf. Data Eng. (ICDE), 2001.

[4] Z. Wu et al., "Towards Automatic Incorporation of Search Engines into a Large-Scale Meta search Engine," Proc. IEEE/WIC Int'l Conf.Web Intelligence (WI '03), 2003.

[5] W. Meng, C. Yu, and K. Liu, "Building Efficient and Effective Meta search Engines," ACM Computing Surveys, vol. 34, no. 1,pp. 48-89, 2002.

[6] S. Mukherjee, I.V. Ramakrishnan, and A. Singh, "Boot strapping Semantic Annotation for Content-Rich HTML Documents," Proc.IEEE Int'l Conf. Data Eng. (ICDE), 2005.

[7] D. Embley, D. Campbell, Y. Jiang, S. Liddle, D. Lonsdale, Y. Ng, and R. Smith, "Conceptual-Model-Based Data Extraction from Multiple-Record Web Pages," Data and Knowledge Eng., vol. 31,no. 3, pp. 227-251, 1999.

[8] J. Wang and F.H. Lochovsky, "Data Extraction and Labe Assignment for Web Databases," Proc. 12th Int'l Conf. World WideWeb (WWW),2003.

[9] W. Su, J. Wang, and F.H. Lochovsky, "ODE: Ontology-Assisted Data Extraction," ACM Trans. Database Systems, vol. 34, no. 2,article 12, June 2009.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

[10] L. Arlotta, V. Crescenzi, G. Mecca, and P. Merialdo, "Automatic Annotation of Data Extracted from Large Web Sites," Proc. SixthInt'l Workshop the Web and Databases (WebDB), 2003.

[11] J. Zhu, Z. Nie, J. Wen, B. Zhang, and W.-Y. Ma, "Simultaneous Record Detection and Attribute Labeling in Web Data Extraction,"Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 2006.

[12] Y. Zhai and B. Liu, "Web Data Extraction Based on Partial Tree Alignment," Proc. 14th Int'l Conf. World Wide Web (WWW '05),2005.

[13] W. Liu, X. Meng, and W. Meng, "ViDE: A Vision-Based Approach for Deep Web Data Extraction," IEEE Trans. Knowledge and DataEng., vol. 22, no. 3, pp. 447-460, Mar. 2010.

[14] H. Elmeleegy, J. Madhavan, and A. Halevy, "HarvestingRelational Tables from Lists on the Web," Proc. Very LargeDatabases (VLDB) Conf., 2009.

[15] H. Zhao, W. Meng, Z. Wu, V. Raghavan, and C. Yu, "Fully Automatic Wrapper Generation for Search Engines," Proc. Int'l Conf. World Wide Web (WWW), 2005.

[16] Y. Lu, H. He, H. Zhao, W. Meng, and C. Yu, "Annotating Structured Data of the Deep Web," Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE),2007.

[17] S. Handschuh, S. Staab, and R. Volz, "On Deep Annotation," Proc. 12th Int'l Conf. World Wide Web (WWW), 2003.

[18] S. Handschuh and S. Staab, "Authoring and Annotation of Web Pages in CREAM," Proc. 11th Int'l Conf. World Wide Web (WWW), 2003.

[19] B. He and K. Chang, "Statistical Schema Matching Across Web Query Interfaces," Proc. SIGMOD Int'l Conf. Management of Data, 2003.

[20] H. He, W. Meng, C. Yu, and Z. Wu, "Automatic Integration of Web Search Interfaces with WISE Integrator," VLDB J., vol. 13, no. 3, pp. 256-273, Sept. 2004.

[21] H. He, W. Meng, C. Yu, and Z. Wu, "Constructing Interface Schemas for Search Interfaces of Web Databases," Proc. Web Information Systems Eng. (WISE) Conf., 2005.

[22] J. Heflin and J. Hendler, "Searching the Web with SHOE," Proc. AAAI Workshop, 2000.

[23] L. Kaufman and P. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, 1990.