# System for Information Extracting from Web Pages

*Rajaa kadem*
*Al-rafidain university college*

**Abstract:**

Internet contains information that is readily available. Internet is always available to you. So the great problem with the Internet is how to extract what you need efficiently. Information extraction here can be done by build a specific strategy for a search in Internet, that consist of the combination of data mining with advanced search engine for efficient information extraction. Also apply data mining techniques to the web server to support good information extraction from the web to the visitors.

## 1- Introduction:

With so much data on the internet, it can be difficult, frustrating, and seemingly impossible to find the exact information you need. To assist you in your researches, there are several different ways to navigate the internet. The web is enormous and growing at an incredibly fast pace. It has been said that if you spent only one minute per page, 10 hours a day, it would take four-and-a-half years to explore only 1 million web pages. Thus, a real need exists for some way to search this huge resource [1].

There are many powerful search utilities on the web such as Yahoo!, AltaVista, Lycos, InfoSeek, Excite, and WebCrawler. In the Internet world, these search utilities are called ***search engine [2].***

Search engines are composed of large databases. These databases contain information about Web pages that have registered with a particular search engine, such as Yahoo! At Yahoo!, registrations are entered by humans, who categories entries by subject.

Through keywords, you can find information on any subject that you need to investigate. A search engine database typically contains information such as the title of the page, the Uniform Resource Locators URL, a short abstract of the contents, and keywords to help the search engine. The URL is added to the search engine database.

Not every web page is registered with every search engine. Some web site managers register their site only with particular search engines, neglecting others. Because of this, and the vast array of search engines, you may receive completely different results from each search engine [3, 4].

## 2- Advanced Search Technique in Search Engines:

For general searches, you typically enter a single keyword or a short phrase within double quotation marks. Standard searches are useful when an easy and

distinguishable topic is requested. With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making.

*Data Mining*, also popularly known as *Knowledge Discovery in Databases* (KDD). The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge [5, 6].

The properties of Web link structures have led researchers to consider another important category of Web pages called a **hub**. A hub is one or a set of Web pages that provides collections of links to authorities. Hub pages may not be prominent themselves, or there may exist few links pointing to them; however, they provide links to a collection of prominent sites on a common topic.

**"So, how can we use hub pages to find authoritative pages?"** An algorithm using hubs, called HITS (Hyperlink- lnduced Topic Search), was developed as follows:

**First**. HITS uses the query terms to collect a starting set of, say, 200 pages from an index-based search engine. Therefore, the root set can be expanded into a base set by including all of the pages that the root set pages link to, and all of the pages that link to a page in the root set, up to a designated size cutoff, such as 1000 to 5000 pages (to be included in the base set).

**Second**, a weight-propagation phase is initiated. This is an iterative process that determines numerical estimates of hub and authority weights.

We first associate a nonnegative authority weight **ap**, and a nonnegative hub weight **hp**, with each page **p** in the base set, and initialize all **a** and **h** values to a uniform constant. The weights are normalized and art invariant is maintained that the squares of all weights sum to 1. The authority and hub weights are updated based on the following equations:

$$\mathbf{a}p = \sum_{(q \text{ such } q \rightarrow p)} \mathbf{h}q \qquad (1)$$

$$\mathbf{h}p = \sum_{(q \text{ such } q \; p)} \mathbf{a}q \qquad (2)$$

Equation (1) implies that if a page is pointed to by many good hubs, its authority weight should increase. Equation (2) implies that if a page is pointing to many good authorities, its hub weight should increase.

**Finally**, the HITS algorithm outputs a short list of the pages with large hub weights, and the pages with large authority weights for the given search topic. Many experiments have shown that HITS provides surprisingly good search results for a wide range of queries.

Besides mining web contents and web linkage structures, another important task for web mining is web usage mining, which mines Web log

records to discover user access patterns of Web pages. Analyzing and exploring regularities in Web log records identify potential customers for electronic commerce, enhance the quality and delivery of Internet information services to the end user, and improve Web server system performance. A web server usually registers a (Web) log entry, or Web log entry, for every access of a Web page. It includes the URL requested, the IP address from which the request originated, and a timestamp. For Web-based c-commerce servers, a huge number of Web access log records are being collected. Popular Web sites may register the Web log records in the order of hundreds of megabytes every day [7, 8].

## 3- The proposed system:

The aim is to extract the information from the internet efficiently by the employees of the large organization and to provide it to the visiting users of these organization efficiently. This research proposes building a web site to this organization represented by a main web page to control all the functions that provided by this web site. In addition to the common functions of any web there are many proposed functions to provide efficient search to the users and visitors within the organization, by reduce the time and produce good search results. Each function does its work when it is activated by the users or the web administrators. The proposed functions will be explained in details.

### 3.1- The main web page:

This web page represents the home page of the web showing the proposed system, see figure (1). This web page displays all the functions of the proposed system, which will be presented in the following sections.

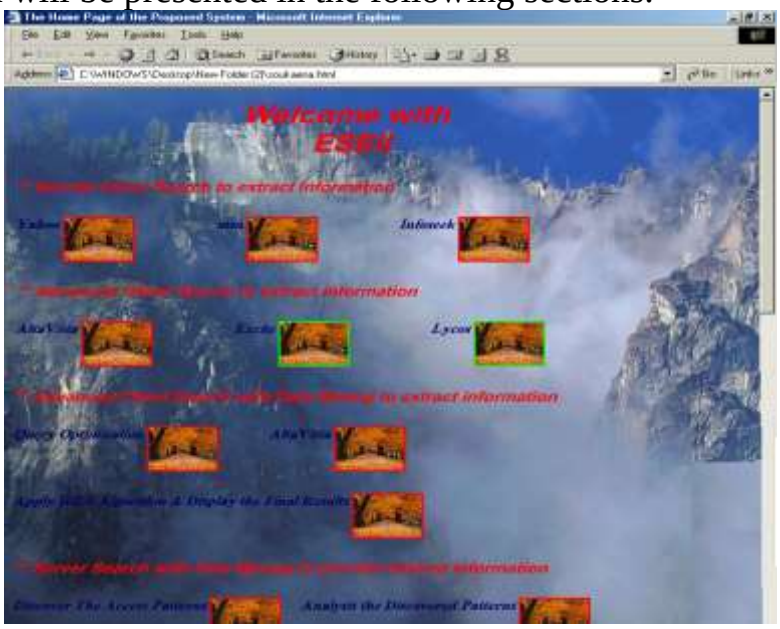


**Figure (1) The home page of the proposed system.**

**3.2- The normal client search function (used by the user):**

This is the first function of the proposed system. When the user wants to extract information about some topic generally without specific knowledge about what is wanted, the proposed system suggests to work with one of the three common search engines: Yahoo, InfoSeek and msn, directly by click the buttons relate to their names, without writing their URL.

**3.3- The advanced client search function (used by the user):**

This is the second function of the proposed system. When the user wants to extract specific information about some topic with certain knowledge about what is wanted, the proposed system suggests to work with one of the three advanced search engines: AltaVista, Lycos, Excite, directly by click the buttons relate to their names without writing their URL.

**3.4- The advanced client search with data mining (used by users) :**

This is the third function of the proposed system. When the user wants to extract specific information about some topic with certain knowledge about what is wanted and the hope to get authorities information.

The proposed system would apply the HITS algorithm with the following improvements:

***First*** built the root set by collect the results of the search by the advanced search engine, AltaVista, after optimize the query by delete the stop list from the query where the stop list are {a, an, for, the,…..}. continue with activate the hyperlink in the collected pages to obtain all possible page to built the base set.

***Second*** apply the second step in that algorithm by giving each page authority weight and hub weight, to determine the best pages in that search.

***Third***, this is the last step in the algorithm, which would display the final result pages of the search.

**3.5- The server search function (used by the administrator):**

This is the fourth function of the proposed system. This function focused on web usage mining to provide efficient information providing for the visitors.

Web usage mining is the automatic discovery of the user access patterns from the web servers. Organization collects large volumes of data in their daily operations, generated automatically by web servers and collected in server access log. Analyzing such data can help organization to determine the life time value of customers, cross marketing strategies across products, and effectiveness of promotional campaigns, among other things. It can also provide information on how to restructure a web site to create more effective organizational presence, and shed light on more effective management of workgroup communications and organizational infrastructure.

## 4- Experimental Work:

To illustrate the proposed system clearly we would trace the functions of the system as follow:

**In the first two functions** for searching the Internet ( normal client search and advanced client search ) the clients have only select normal or advanced search engine needed to extract the desired information.

In normal client search the client may choose the Yahoo search engine to search about what he/she want see figure (2) and figure (3).


**Figure (2) Yahoo: one of the most important normal client search engines, at most the search accomplished by writing the keywords represent what the client search about.**

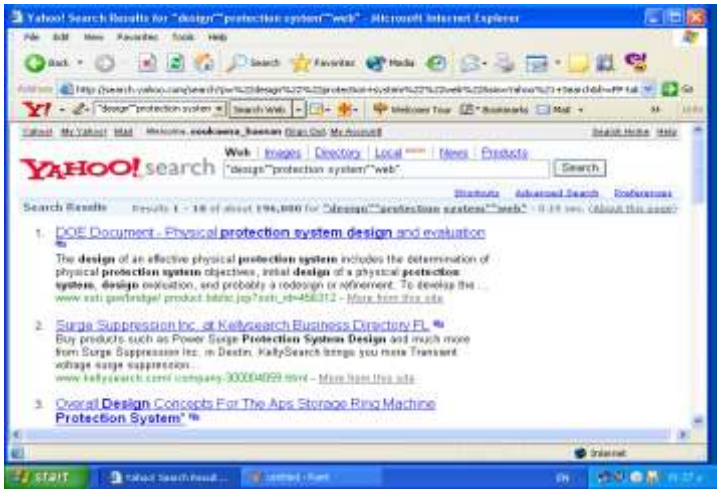


**Figure (3) all the web pages obtained by Yahoo search engine and linked web pages for the client keywords.**

In advanced client search the client may choose Excite search engine to search what he/she want see figure (4) and figure (5).

**Figure (4) Excite: one of the most important advanced client search engines, at most the search accomplished by writing the keywords represent what the client search about.**

**Figure (5) all the web pages obtained by Excite search engine and linked web pages for the client keywords.**

The third and fourth functions are completely different because they have sequence of procedures, this will be explained in the following.

**The third function** has three procedures, the first one is the query optimization when it activated this would display a query interface to make the client writes query which contain the keywords that specify what is wanted , then this query would be optimized by displaying some suggested keywords to have the same meaning, and then explained how to determine exactly the optimal query. This shown in figure (6):

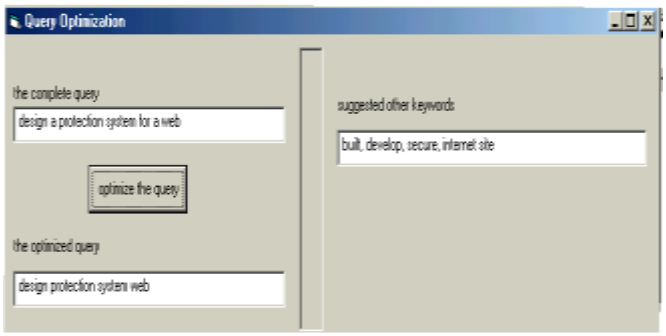


**Figure (6) the query optimization and some keywords suggestion.**

The second procedure is to apply the resulted keywords from the first procedure in the AltaVista search engine, collects the results of this search engine and activates all the links to collect all pages referred by these links. The result will be the input set to the third procedure which represents the HITS algorithm to produce the final set of the pages related efficiently to the subject interested by the client see figures (7-9).


**Figure (7) entering the optimal query mixed with suggested keywords for the AltaVista search engine.**


**Figure (8) set of all the displayed web pages and linked web pages.**


**Figure (9) final result of the algorithm HITS.**

**In the forth function** there are two procedures, the first one is the pattern discovery stage and here we must have a DM task to be applied on the web logging data to extract the access pattern to study the visitor behavior. Then reconfigure the web server information and attentions according to what is preferred by those visitors. This step of the web usage would use the web logging data see figure (10), as input for the association analysis DM algorithm, so the output would be the discovered hidden pattern in the web server data such as the following instance:

| **Local IP** | **Remote IP** | **Local port** | **Remote IP** | **State** | ***Type of service*** | **Time stamp** |
|---|---|---|---|---|---|---|
| **200.12.33.0** | **120.21.44.23** | **80** | **137** | **Listen** | **Tcp** | **2:30** |
| **200.12.33.0** | **22.34.210.11** | **80** | **94** | **Listen** | **Udp** | **10:20** |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |

**Figure (10) flat web logging data file.**

The pattern discovered is:

- 65% of client left the site after four or less page references.
- 40% of client who accessed the web page with URL/ company/ product1, also accessed/ company/ product2; or
- 30% of client who accessed/ company/ special, placed an online order in / company/ product1.

The second procedure is, all the discovered patterns would be displayed for the web usage miner to analysis this set of patterns, understand them very well and by the importance miner factors detect which patterns would be visualized for the web administrators, to decide how to reconfigure the web server according to user interests.

**5- Conclusion:**

In context with the results of the present study it can be concluded that:

1. In addition to improve the technique of extracting the information efficiently from the internet, the research tends to improve the technique of providing efficient information with high level of accuracy to the web visitors.
3. Since Web log data provide information about what kind of users will access what kind of Web pages, Web log information can be integrated with Web content and Web linkage structure mining to help Web page ranking, Web document classification, and the construction of a multilayered Web information base as well.

**References:**


[1] Crumlish .C ., Al_sherman .R .,and their Group , " Internet Complete" , SYBEX, Inc.,2000.

[2] Grauer .R .T ., Marx .G . , " Exploring The Internet with Microsoft Internet Explorer 4.0 ", Prentice _ Hall , Inc . ,1998.

[3] Comer .D .E ., Steven .D .L ., " Internetworking with TCP/IP Vol II: Design, Implementation, and internals", Second Edition, Prentice-Hall, Inc., 1998.

[4] Russell .T .," Telecommunication Protocols" , Second Edition , Mc-Graw Hill Companies ,Inc.,2000.

[5] Arron Ceglar, John Roddick; "*Association Mining"*; ACM Computing Surveys, Vol. 38, No. 2, Article 5, pp. 1-42, July 2006.

[6] Kantardzic M.; *"DM concepts, models, methods and algorithms",* jhon wiley &Sons, 2003.

[ 7] Hutchinson .S .E . , Sawyer .S .C . , with Contribution by Coulthard .G .J ., " Computers, Communications , and Information", Revised Edition, Mc-Graw Hill Company (Irwin), 1998.

[6] He .J ., " Internet Resource For Engineers : A practical Handbook For Engineers and Students", Reed International Books Australia PtyLtd,1998.

# نظام لاستخلاص المعلومات

*رجاء كاظم حسون*
*كلية الرافدين الجامعة*

**الخلاصة**

**يحتوي الانترنت على كم من المعلومات حيث انه دائما متوفرللمستخدمين، لدلك فان المشكلة الكبيرة مع الانترنت هو كيفية استخلاص مانحتاجه بكفاءة .استخلاص المعلومات هنا يمكن ان نعمله عن طريق بناء ستراتيجية خاصة للبحث في الانترنت.هده الستراتيجية تتضمن تشكيلة من تنقيب البيانات مع مكائن بحث متقدمة من اجل الحصول على استخلاص معلومات كفوء وايضا يتم تطبيق تقنية تعدين البيانات على خادمات الويب**