Increasing search engine's accuracy using links clustering

Ahmed bahaa-al deen abdul-wahab

technical college of management

Abstract :

This study is devoted for searching free e-books problem because of its importance with spreading the e-libraries. When the web client asking for a free e-books the search engine returns selling e-books sites, this produce a conclusion of inaccuracy about this search engine. According to this case this study suggest the philosophy of clustering a sample of web sites on the number of links which lead to the downloadable e-books and rank this sites higher than the others with matching the entered keywords. Three clusters have been noticed (sites with high rank, sites with medium ranking, and low ranked sites).

The proposed system (Ico.com) has been built form many components; the first is the web crawler which is a software Program for fetching data of web pages (like page address, subject, number of links…etc), the second component is the clustering program written using (VB6.0) language to cluster the web pages databases according to the number of links to real existing downloadable books, and the last component is the search engine interface which is built using HTML and VB script under Active server pages technology. This work reached to increasing the search engine accuracy using factors like analyzing the number and the type of links by mining the web site's database to improve accuracy for search engines.

1- Introduction

Currently the World Wide Web contains billions of documents and it is still growing rapidly [1], continues to grow at a rate of one million document / day [2]. A web search engine is the primary tool used to locate web page based on content [4], while finding the relevant document to satisfy user's information need is very important and challenging task. [1]. to accomplish this, search engine developer can cluster the web documents based on the number of the existing hyperlinks (links) in documents. This way can be effectively used to enhance clustering effectiveness to increase search engine's results accuracy [1],while the other search engines like Google depends on word clustering and page rank system that depending on the number of pages that link to a specific page[3]. The objective of this research is to Build a search engine system using Active server pages technology (HTML and VB Script) to provide accurate result for finding especially free e-books depending on the number of the real existing links to the downloadable e-books with a combined clustering software using Visual Basic 6.0 that apply the known (K-means clustering method) to cluster the web database into classes in order to increase the search engine sensitivity about different client demands.

1-1 the problem description

Suppose a web search engine's user are looking for useful document about "free computer e-books", the user can observe that most of the search results don't lead to free e-books site, mostly these pages take the user to web sites for selling e-books, and this yield a bad conclusion about most used search engines (like yahoo and Google) that this search engine are not accurate

1-2 the search engine

The search engine is the primary tool used to locate web pages based on content [4].

A web search engine must present a user with a set of results, given the input. There are Five logical tasks a search engine performs for each search. Each task corresponds with a specific Component of the architecture presented in Figure (1).

Five tasks every search engine performs for each search:

1_ Accept user input

2_ Process user input

3_ Apply database query

4_ Process results

5_ Display results

Figure (1) describes the four required components of a web search engine and the crawler for populating the database. The first component is the *user interface*, which is responsible for accepting user input and presenting the output. Second is the *query processor*, which generates a reasonable database query from the user input. Third is the *database*, which is the component that stores the knowledge about each page. In addition to these components, most web search engines have a *crawler*, which is used to populate and maintain their database [4].

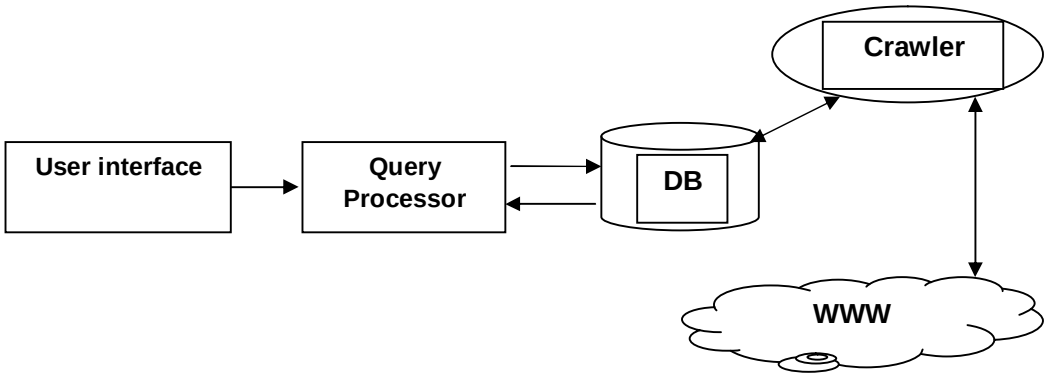

Figure (1) the traditional common component of any search engine

1-3 Theoretical aspects

Relevance is often used to describe the agreement of a document with user's needs of information [4]; Budzik also describes a small study confirming that user judgment of what is useful are distinct to similarity to the query. This understanding to the document relevancy developed in Cox. Description in the notion of indexing documents based on their expected usefulness for a given query.

This research adopts the Graft description to the relevancy (usefulness) which says the system that considers many factors in addition to the topic for retrieving document [4].

These many factors in web documents may include (text, page description, photo, and hyperlink)while reputable search engines like Google depends on text indexing and page rank system on the number of income links to a specified page and this factor may be trick for the search engine in some cases[3]. The search engine ought to product clusters that group document relevant to the user's query separately from irrelevant ones [5], the Information retrieval Community has explored document clustering as an alternative method of organizing retrieval results [5],

so clustering using content and hyperlink information shall increase sensitivity and accuracy of the search engine about the useful documents, especially in topics like free downloadable books, and free downloadable software [1]. Clustering is another setting for learning, which does not use labeled objects and therefore is *unsupervised*. The objective of clustering is finding common patterns, Grouping similar objects, or organizing them in hierarchies. In the context of the Web, the objects manipulated by the learning system are web documents, and the class labels are usually topics or user preferences. Thus, supervised web learning system would build a mapping between documents and topics, while a clustering system would group web documents or organize them in hierarchies according to their Topics [6].

A clustering system can be useful in web search for grouping search results into closely related sets of documents. Clustering can improve similarity search by focusing on sets of relevant documents [7]. This work adopt the known K-means clustering algorithm [6]

I. arbitrarily choose K objects as the initial cluster centers

II.repeat

III. (re)assign each object to the cluster to which the object is the most similar ,based on it's distance to the mean value of the object in the cluster;

IV.Update the cluster means, i.e.; calculate the mean value of the objects for each cluster ;

V. Until the new calculated mean for each cluster equal previous mean.

Notice that the Manhattan distance has been used to calculate the distance between records and the mean of the clusters , For a set1 (x_1, $x_{2...}$ xn) and a set2 (y_1, y_2, ...,y_n), the Manhattan distance between points calculated by the equation [8]:-

$$d(Xi, Yi) = \sum_{I=1}^{n} |Xi - Yi| \qquad \text{---------- (1)}$$

Where d (Xi, Yi) is the distance between points Xi and Yi, and Xi and Yi are elements in set 1 and set2

2- The proposed system

2-1 the proposed search engine

The proposed search engine named as (Ico.com) and its Purpose is special search engine to help researchers to find free e-books they needed. This search engine built using Languages HTML and Visual basic script (Visual basic edition for internet purposes), and finally the reader can see figure (2) to understand the search engines parts.

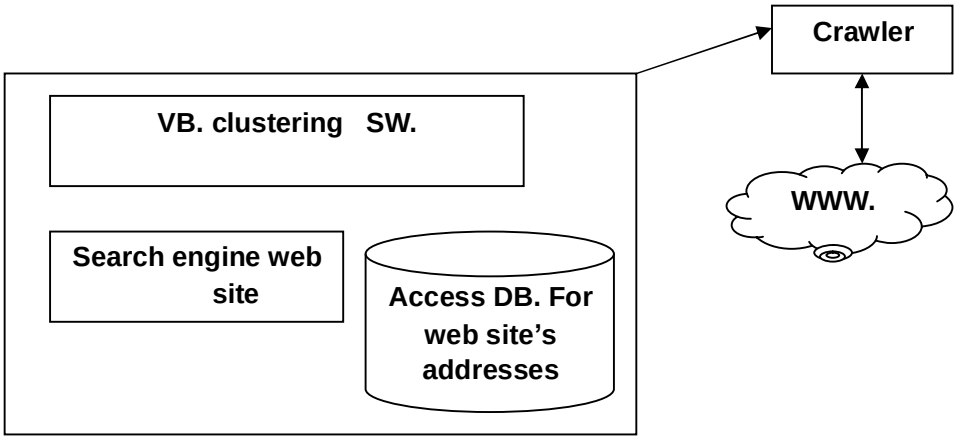

Figure (2) the proposed search engine's architecture

The search engines parts are:-

1.VB. clustering software:- this K-means clustering software is built using visual basic 6.0 see figure(3)

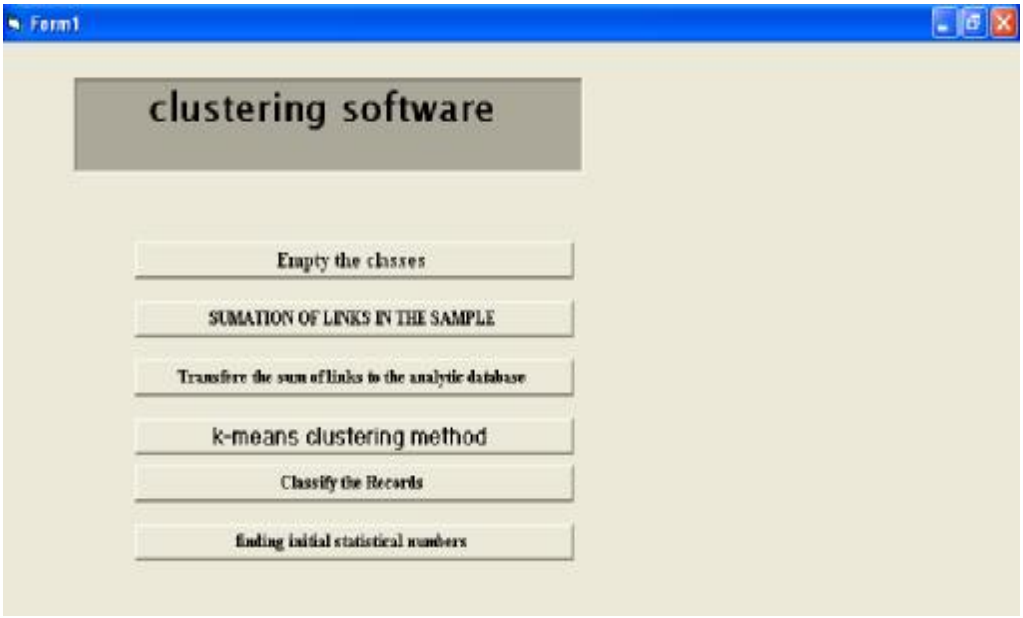

Figure (3) the clustering visual basic program Interface

This clustering software is used to cluster a sample database of (70 records) according to the sum of links (PDF. +RAR. +ZIP) which are absolutely exist in these sites see figure (4)

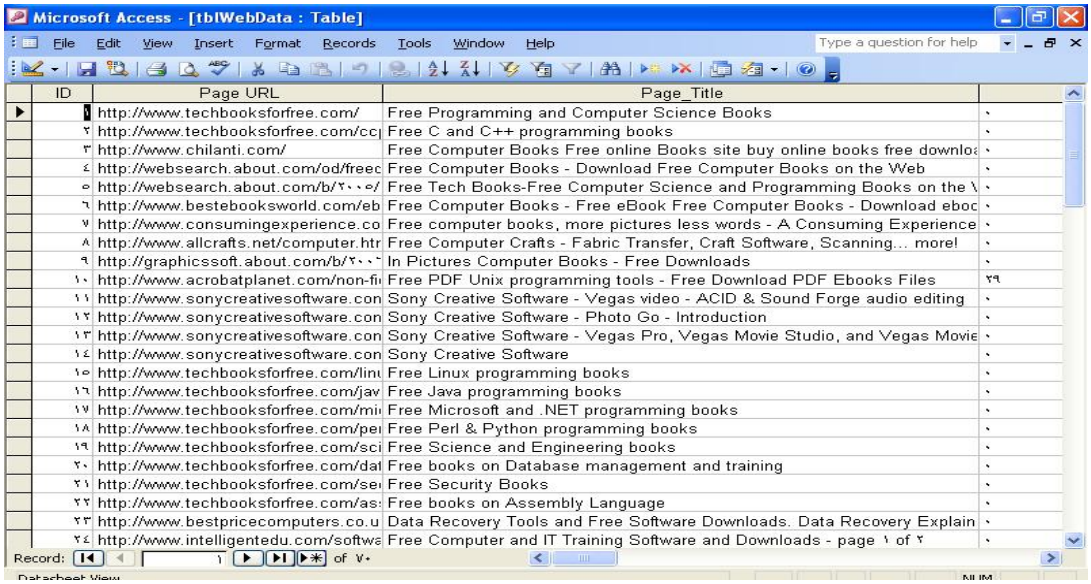

Figure (4) the search engines database about 70 web pages

2.The search engine's database: - a database consists of a sample of 70 records representing 70 different web sites see figure (4). These web sites are special in free e-books.

3.Web site: - or the (search engine's interface) this web site is built using HTML. And Visual basic script. This web site is operating using IIS server (Internet Information server) under Microsoft windows. , this web site consists of two pages as shown in figure (5).

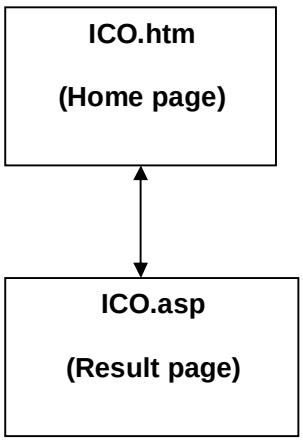

Figure (5) view of the proposed search engine's pages

- **ICO.com :- this web page represent the client interface for this search engine and accept client input for the desired query See figure (6)**

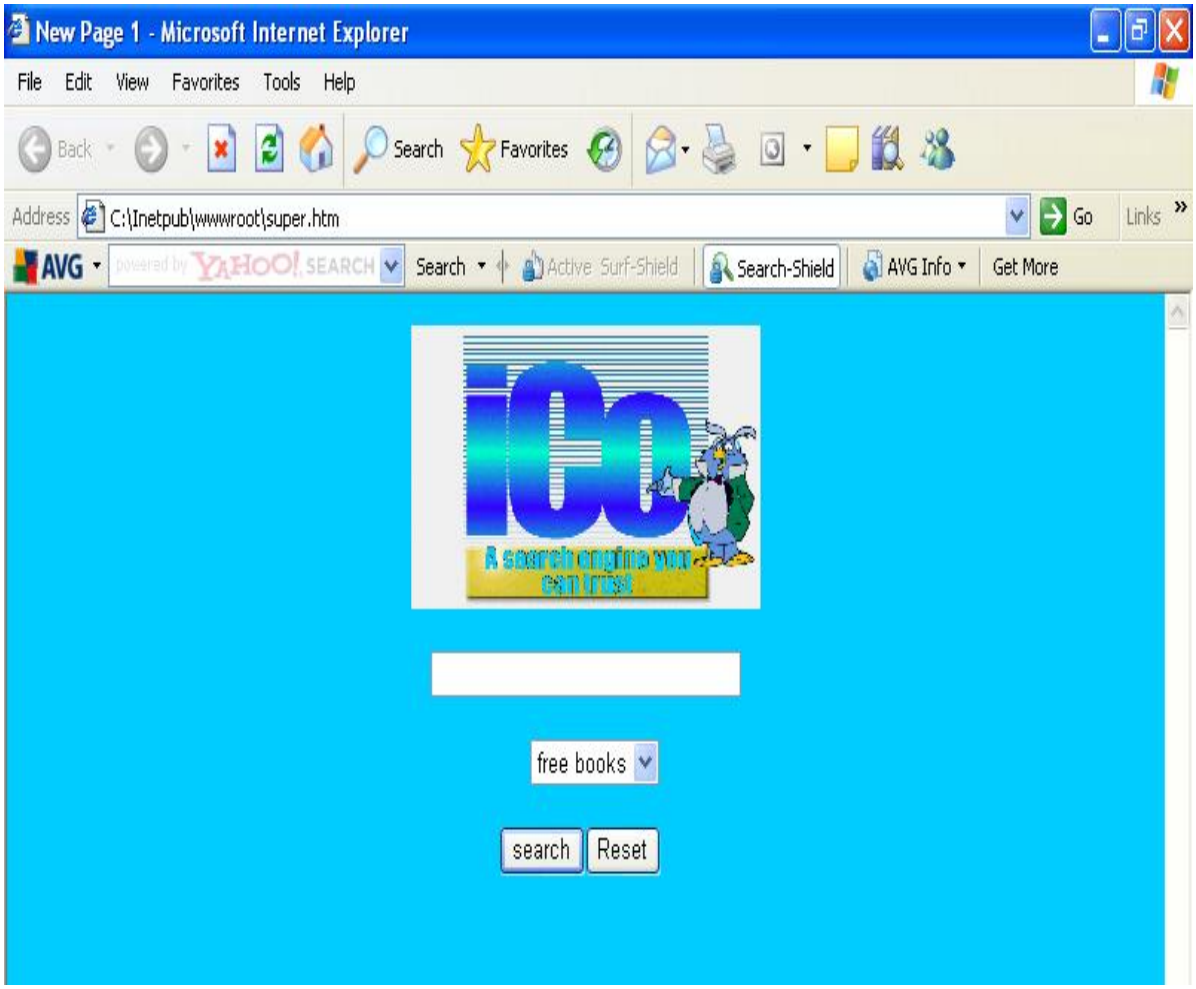

Figure (6) the Ico.com home page

- **Ico.asp:- this web page built using VB script and HTML. This page receive user input keywords and returning the useful web pages.**

4.Web crawler: - a web crawler used to collect the data of web pages from the web and save these data in the Access database.

2-2 Search engine's work's flowchart

This flowchart illustrate how the search engines components work together to accomplish the search work with clustered database as shown in figure (7).

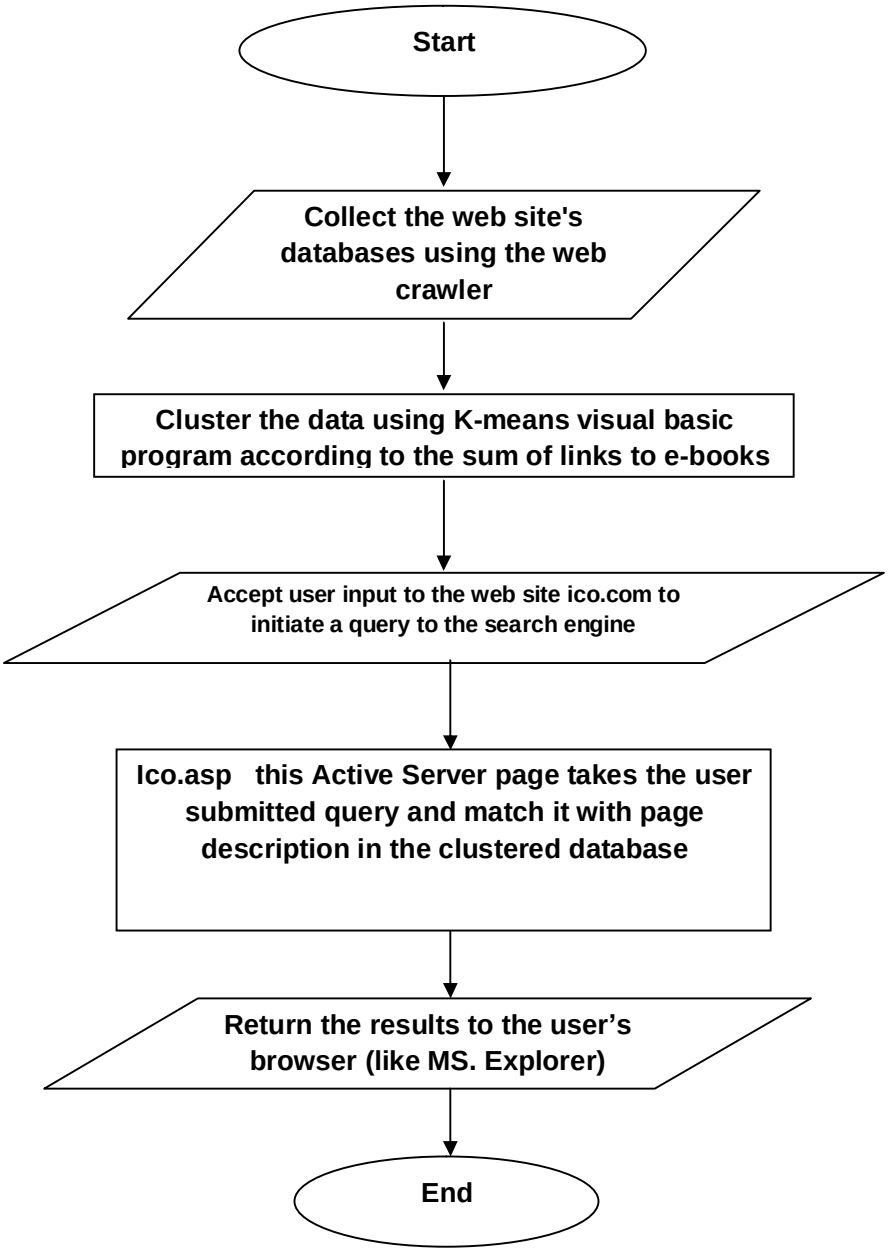

Figure (7) search engine flowchart

3- Interfaces

3-1 proposed search engine interfaces

Suppose the client searching for (computer) free books , all what the user need is inserting the word (computer) into the text box and specify the category for his search using the choosing list below (for example : free books) see the figure(8) below

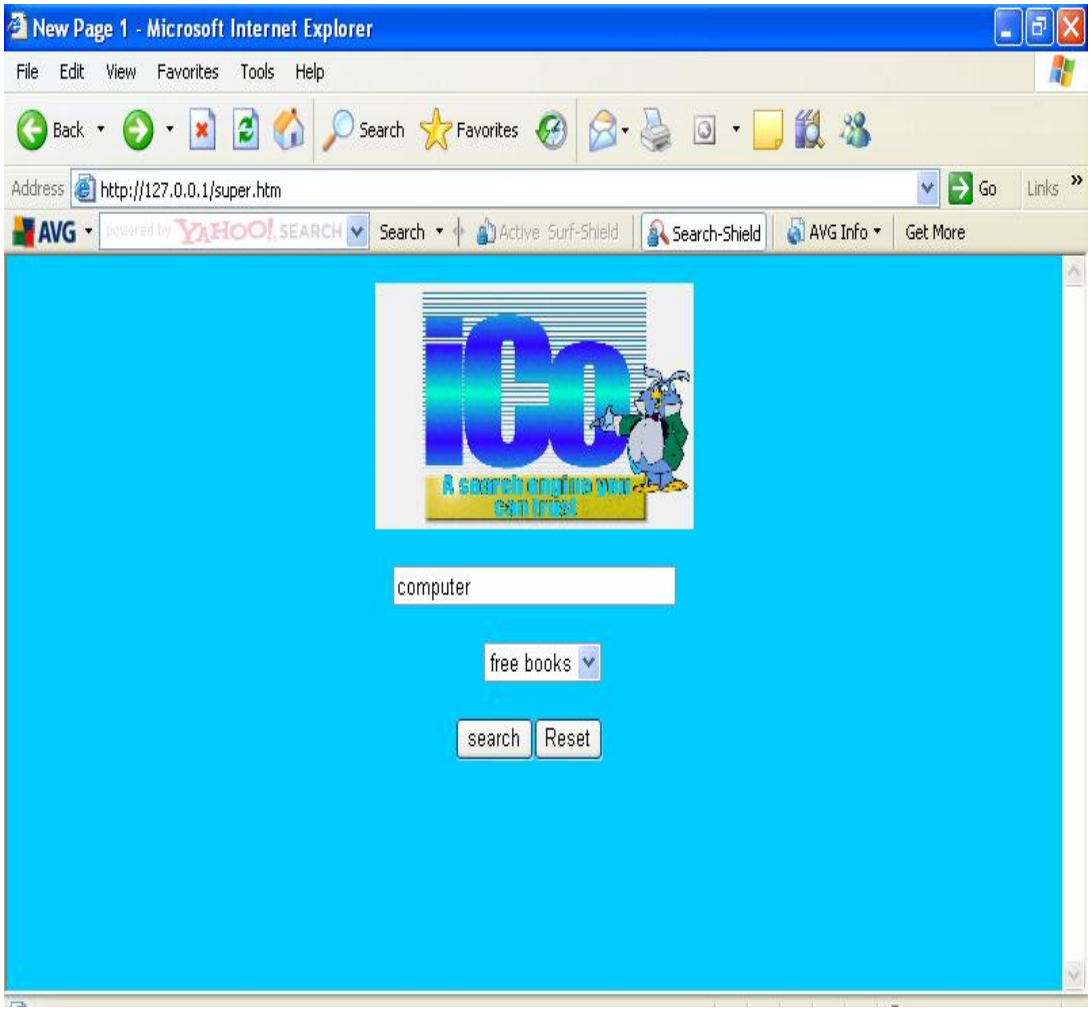

Figure (8) the user entered a keyword "computer"

Then click (search) button waiting for the result, the user shall see the result as below in figure (9)

Figure (9) the result page containing strong and weak results

The user can observe the most possible pages that contain e-books and the less one based on the clustered database for these sites which already has been clustered using the clustering visual basic program.

4- The results discussion

4-1 Reason behind the clustering

If the researchers take the raw data sample which consist of free e-books web sites data and calculate the summation of real links to e-books using equation:-

Sum of links = number of PDF links + number of RAR links + number of ZIP links ----- (2)

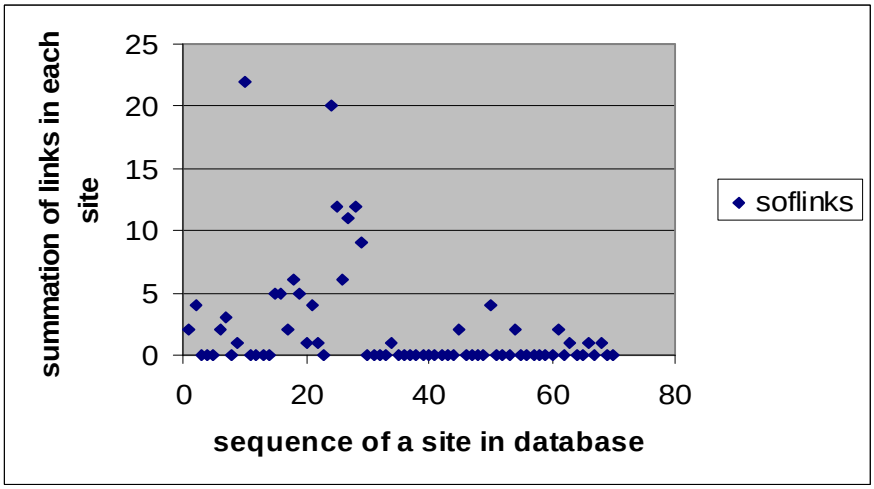

Figure (10) the number of the links to real e-books in the web database

The reader can observe from the figure above(10) that most of the web pages about free e-books actually don't contain any real links (most of the sites on zero line) although these pages classified as free e-books according to the text contained within it, and little between (5-15) and only two pages contain a real 20 to 25 links to e-books (page number 11 and page number 25 in the database). So there is a need to cluster the web pages according to the number of the existing real links can improve the accuracy for the search engines result.

4-2 conclusion

After executing the clustering program on the web database according to the real number of links to e-books the program divide the sites into three groups:-

- **Cluster A :- contains summation of links >= 20**
- **Cluster B :- contains summation of links >= 10**
- **Cluster C :- contains summation of links = 0**

So this clustering method increase the accuracy of the search engine to specify which page contain really links to free downloadable e-books .

In clustering software after deploying calculating the number of records in each cluster the program reach to the results

- **Cluster A:-5 records only.**
- **Cluster B: - 16 records.**
- **Cluster C :- 49 (see figure 11) below**

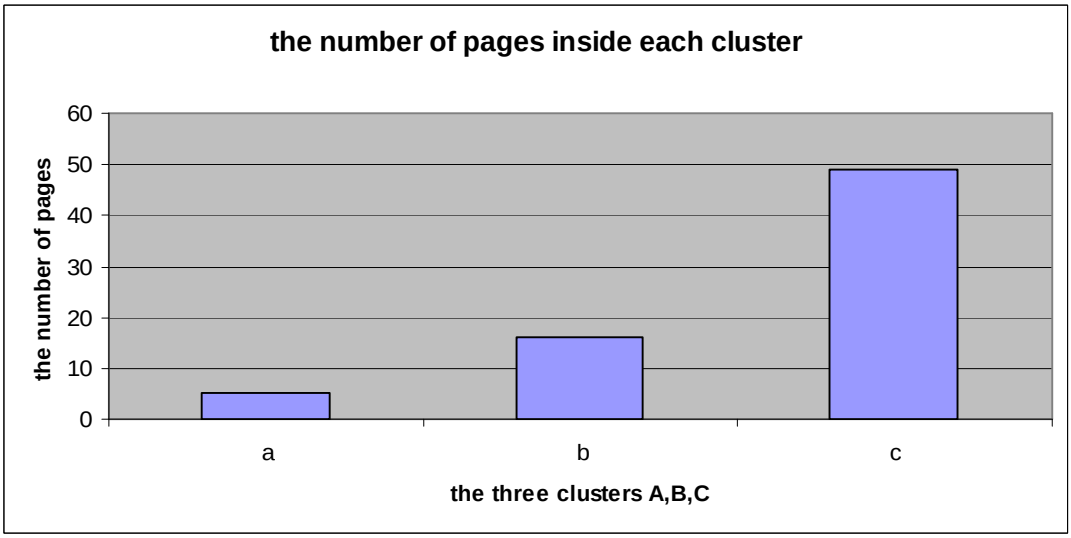

Figure (11) the clusters referring to number of sites in each cluster

This means that:-

1. **Only 7% of the web pages contain a real links to downloadable e-books.**
2. **23% of the web pages contain >= 10 and < 20 real link to e-books.**
3. **70% of the records which are described as free e-books sites do not contain real links.**

Note that these numbers are gotten from the visual basic software for k-means clustering see figure (12)

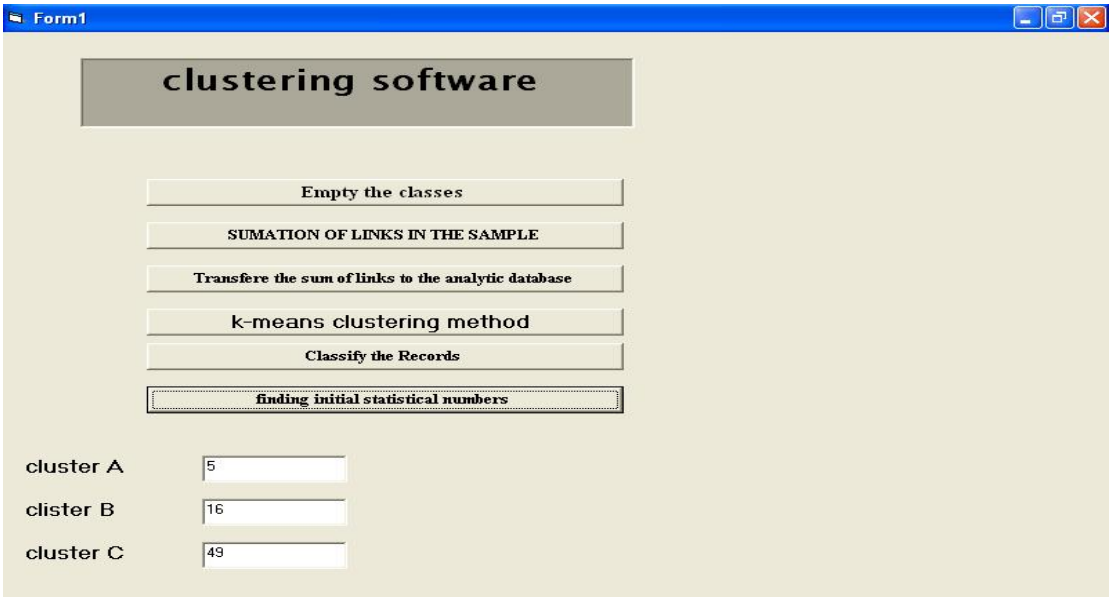

Figure (12) the number of sites in each cluster calculated by clustering software

After that; the reader can imagine the mass of inaccuracy can be occurred by the search engine if the user wants to reach to the e-books web sites depending on the traditional text matching because 70% of the pages do not contain links to e-books while the search engine can put this page at low rank result, so it is a good step for a search engine like AltaVista or Google to improve the text matching factor by taking the number of links within the page to give a good rank especially in the cases of searching for free software's or free e-books.

11-Suggestion for future work

This research can announce the following suggestions:-

1. **the researchers in search engines area can develop more clever search engines that depends on web site‘s content semantics and developing web site’s meta data to increase the search engine’s ability in connecting topical related sites.**

2. **The researchers can use data mining techniques to mine server log files or the web client's navigational series to predict web client needs in order to provide more accurate web pages.**
3. **Decide to write and building web sites with new semantic web technologies languages like (ontology describe language), these new languages will increase the search engine's ability to retrieve the useful segments from web sites.**
4. **this an Invitation to study the page rank system in reputable search engines like Google and AltaVista which are depends on text clustering and the number of web page that links to a web page and try to develop new methods to increase the result accuracy for this sites.**

References

1- Xiaofinghe, Hong yang Zha , Chris HQ. Ding and Harst Simon
"web document clustering using hyperlink structures " A research submitted to the department of computer science , the Pennsylvania state Universit,2002

2- K.bharat and A.bradr ."A technique for measuring the relative size and overlap of public web search engine ". Proc of the 7th WWW. Conference ,1998

3- "how does Google collect and rank result " , Google librarian center , http://www.google.com/librariancenter/articles/0512_01.html , 2009

4- Eric J.Glover "Using extra-topical preferences to improve web based meta search" .PhD dissertation .computer science and engineering dept. , university of Michigan, 2001

5- Oren Zamir and Oren Etizioni ;
"Web document clustering: a feasibility demonstration ", department of computer science & engineering, University of Washington,

http://www.cs.washington.edu ,1998

6- Jiawei Han & Micheline Kamber. "data mining concepts and technique", Simon Fraser university,1999 , http://www.4shared.com

7- Markov Z., and Larose Daniel ," data mining the web ",a john wiely & sons publications ,2007

8- Manhattan distance, http://www.wikipedia.com , 2009.

.

زيادة دقة محركات البحث بعنقدة روابط الانترنيت

م.م. احمد بهاء الدين عبد الوهاب عباس

الكلية التقنية الادارية / قسم تقنيات المعلوماتية

المستخلص:

تخصصت الدراسة في مشكلة البحث عن الكتب الالكترونية المجانية تحديدا لما لها من اهمية مع تزايد المكتبات الالكترونية وذلك عند طلب الباحث عن الكتب المجانية تقوم محركات البحث باعادة مواقع بيع الكتب ضمن النتائج مما يعطي انطباع بعدم دقة او وثوقية محركات البحث هذه ، لذا قام العمل على فلسفة عنقدة عينة من مواقع الانترنيت المختلفة على اساس عدد الروابط الموجودة فيها فعلا و التي تقود الى كتب الكترونية و منحها رتبة اعلى من بقية المواقع مع الاخذبنفر الاعتبار تشابه الكلمة المدخلة من مستخدم محرك البحث و على هذاالاساس تم استنتاج ثلاثة اصناف (مواقع ذات اسبقية عالية ، مواقع ذات اسبقية متوسطة ، مواقع ذات اسبقية قليلة). المنظومة المقترحة هي محرك البحث (Ico.com) و تم بناءه من عده مكونات ،المكون الاول عنكبوت الويب (web crawler) وهو برنامج لسحب المعلومات المتعلقة بصفحات الانترنت مثل(عنوان الصفحة ، موضوعها ، كلماتها المفتاحية و الروابط الموجودة فيها و عددها ... الخ) و الجزء الاخر هو برنامج العنقدة لعنقدة قاعدة بيانات صفحات الانترنيت حسب اعداد الروابط و اخيرا واجهة محرك البحث بنيت بلغتي HTML و VB Script و بتقنية صفحات الخادم الفعالة (Active server pages) .

ان هذا العمل توصل الى امكانية زيادة دقة محركات البحث بالاعتماد على عناصر اخرى غير تطابق الكلمات مثل تحليل عدد ونوع الروابط بتحليل قواعد بيانات محركات البحث .