IMPACT: International Journal of Research in Engineering & Technology (IMPACT: IJRET)
ISSN(E): 2321-8843; ISSN(P): 2347-4599
Vol. 2, Issue 8, Aug 2014, 45-52

PERSONALIZING RESULTS IN SEARCH ENGINES USING WORDS CORRELATION

MOHAMED IBRAHIM SHUJAA

Lecturer, Technical Department of Information Technology, College of Management, Baghdad, Iraq

ABSTRACT

World Wide Web is a huge informational environment is still growing, so the searching problem is still arise, for this the user need to develop a search engine to get closer result to the user's request. this work take the principle of personalization which means make web page content closer to a specific client and use this principle in a new area, by finding sites close to the requested term according to the correlation with other keywords specifying a special area, this work took a word (sport) as a sample, the proposed search engine (closer) built using active server pages technique.

KEYWORDS: Search Engines, Computer Web Search, Semantic Correlation

INTRODUCTION

Since the creation of the web until now, the internet has become the greatest resource of information available in the world [1].

On the web there are no standards or style rules; the contents are created by set of very heterogeneous people in the autonomous way [2], providing the people with access to the information is not the problem, the problem is that people with varying needs and preferences navigate through large web structure missing the goal of their inquiry so web personalization is one of the most promising approaches for alleviating this information overload. Tailoring the presentation of a website's content to match a specific user's instructions or preferences. This custom tailoring is accomplished either by the user choosing from a menu of available alternatives or by tracking his or her behavior (such as which pages are accessed and how often) on the site [3]. Personalization provides users with that they want or need without having to ask it explicitly, and it is a multidiscipline area deploying techniques from various scientific fields for putting together data and providing personalized output for individual users. These fields are like Information retrieval, users modeling, and artificial intelligent [2].

THE PROBLEM

The method of finding information using search engines works well when users want to retrieve home pages web sites related to corporations , institutions or specific events , however when users want to explore related information's from several pages , this way has some deficiencies , the ranked list are not conceptually ordered and information in different sources is not related , so this work introduce word correlation to personalize the web search result and gathering related topics closely so the user can find the related sites and topics easily.

THE SERACH 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

- Accept user input
- Process user input
- Apply database query
- Process results
- 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 [3].

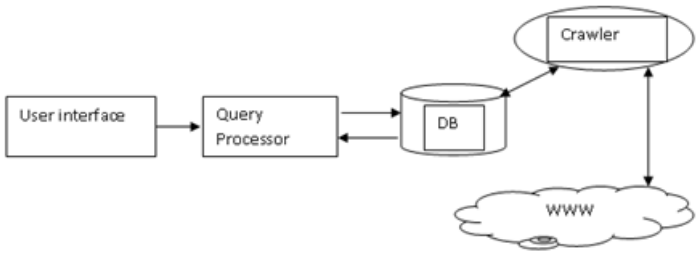

Figure 1: The Traditional Common Component of Any Search Engine

WORDS CORRELATION AND PERSONALIZATION

The first related work is "ontology –based distance measure for text clustering" ,this paper propose a new method which fully uses the existing learning ontology method and the well known lexical database (wordnet) to find term mutual information combining this mutual information matrix and the traditional vector space model to design a new data model (considering the correlation between terms) on which the Euclidian distance can be used then the k-means method can be implemented with new ontology based distance measure[5] .

The second more development paper by shaoux song and chunping li propose asymmetric similarities measure in their paper, by construct a semantic correlation network by asymmetric similarity between documents shaoux notice that some specialized article may dedicated to one topic (i.e. basketball) in single document, while some summarized article may include several topics (i.e. sports, including basketball and football), this paper providing a more accurate method than the pervious one because it can tell the difference between the summarized article and the specialized one, this asymmetry feature induces that the summarized documents have more connections (correlations) while the specialized one have less [6]. Web personalization is a strategy, a marketing tool, and an art. Personalization requires implicitly or explicitly collecting visitor information and leveraging that knowledge in your content delivery framework to manipulate what information you present to your users and how you present it[6]. This research employ the thought of words correlation in web search engine's personalizing results in order to get related results closer to searched topics using a matrix of word's

correlations by adding an additional component to the search engine called "word correlation calculator" in order to improve personalizing searching results according to the user's entered term

THE PROPOSED SYSTEM

The proposed search engine is called (closer) and it is consist of word correlation calculator in addition to the basic search engine's component maintained above see diagram below

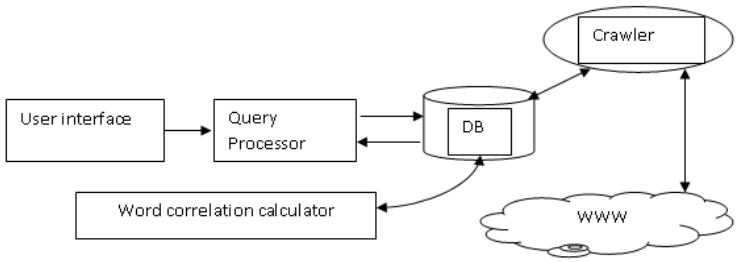

Figure 2: The Improved Search Engine with Word Correlation Calculator

This calculator is building a word correlation matrix from word database taken from search engine's database crated from the web

Let Tin be the term entered by the user of the search engine; propose the following set of documents is defined by X={X1, X2, …,Xn }

Terms in the vocabulary collection is represented by the set T={T1,T2,…,Tm}

So Xji is the weight of the term Ti in the document Xj usually determined by the number of the times Ti appears in the document Xj

Let Ct1t2 be the correlation between T1 and T2, so the software is building the correlation matrix:

$$M = \begin{matrix} c11 & c12 & c13 & c1n \\ c21 & & & \\ c31 & & & \\ cnm & & & \end{matrix}$$

The software which is building maintained matrix created using visual basic language With Microsoft access database see below.

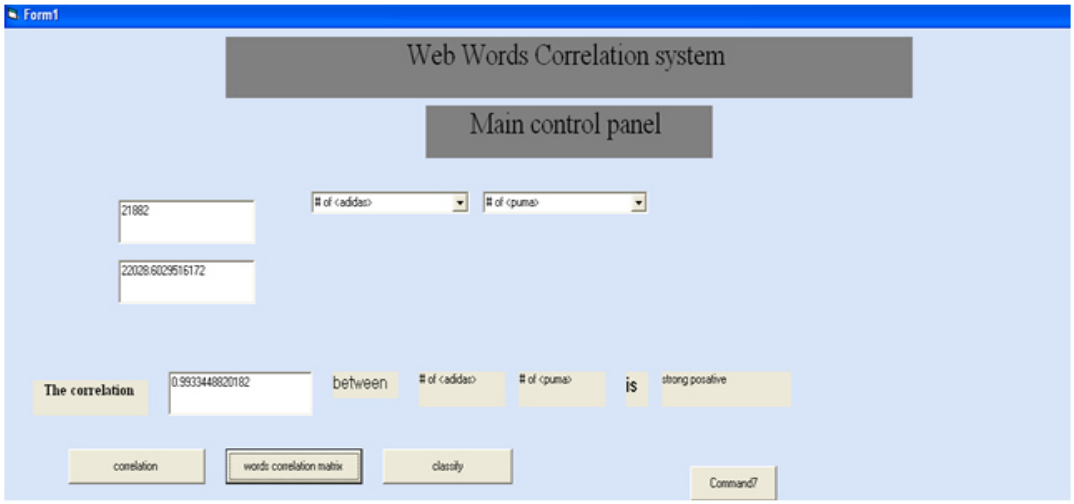

Figure 3: The Web Word Correlation System and Matrix Generator Interface

This program can calculate the correlation between essential words in the web database and classify the relation between words into (strong positive correlation, strong negative, weak positive, and weak negative correlations) and

saving these classes in the web database to create the matrix of correlation between words .so when the user input term (T) the search engine shall go to matrix M, if T is member in the matrix, the search engine program shall return all related topics and the closer one according to it's correlation to the term T and other terms, the reader can see the following flowchart

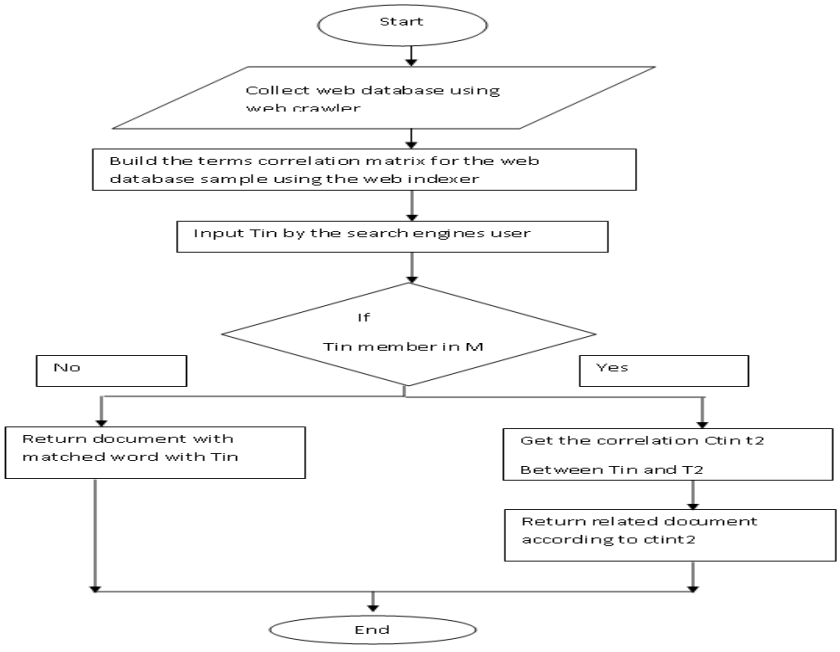

Figure 4: Flowchart of the Proposal System Work Flow

RESULTS & DISCUSSIONS

This research took a sample of (17) sport sites and build a correlation matrix between essential keywords in the sport area keywords like (sport, fifa, clubs, balls, football, soccer, puma, adidas, clothes, shoes) using the correlation builder See figure below

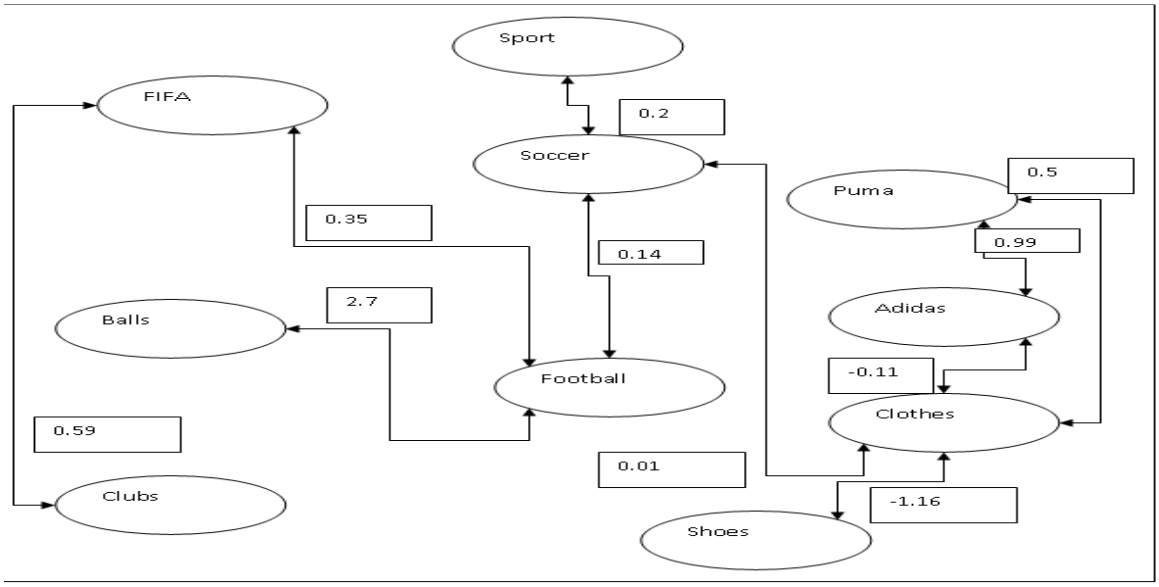

Figure 5: Word Correlation Sample Network

The reader can notice that a strong correlation between:-

Puma & adidas

Clothes & shoes

Fifa & clubs

Puma & clothes

Adidas & clothes

This results depending on the taken sample are logically acceptable result because there may be a convenient relation between (for example puma and clothes) or between (puma and adidas), this relation can improve gathering the related topics for web search engines users.

This research can improve these results using Google trends service, this Google trends service enable the user to monitor the searching trends of Google users in any period of time, for example if the user want to measure the correlation between words (football, fifa, soccer, sport, and shoes) for Google users between (2004 and 2010) using Google trends see the diagram below.

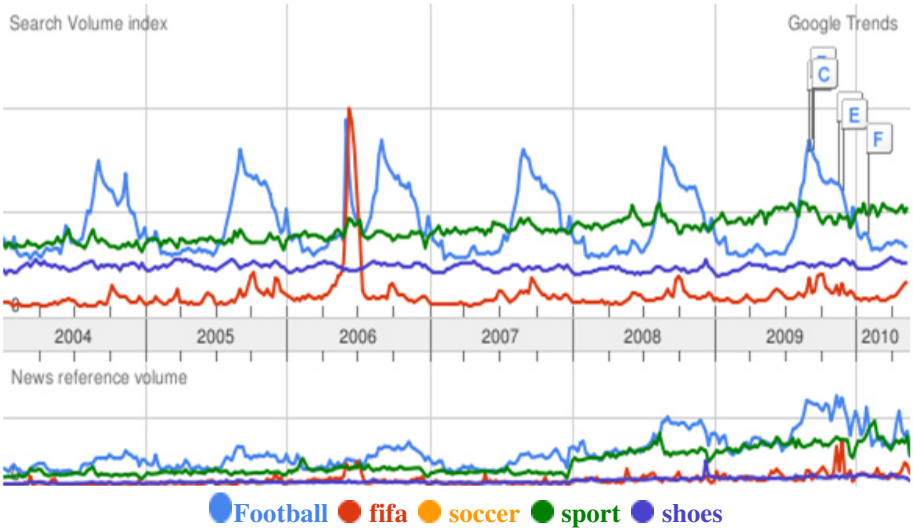

Figure 6: Diagram of Terms Requested from Google Trends

The reader can notice that there is some relation between words like football and fifa requests especially in some period like the worlds cup in 2006 also the reader can observe a relation between word like (sport) and (fifa and football).and this improve the idea of using words correlation to personalize search results according to the user's entered term.

So when the user input for example word like (soccer) to the search engine's front page like below

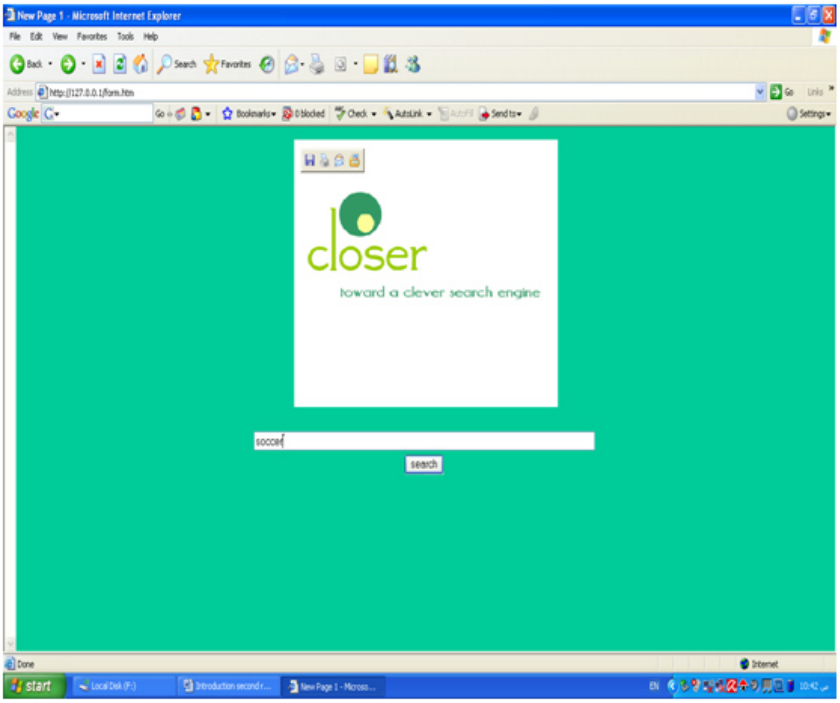

Figure 7: Closer Search Engine Front Page

The result will be all pages that carrying the requested word (soccer) and related pages with related topic that have a correlation (strong or weak correlation) with word (soccer) in the related web sites section, see picture below

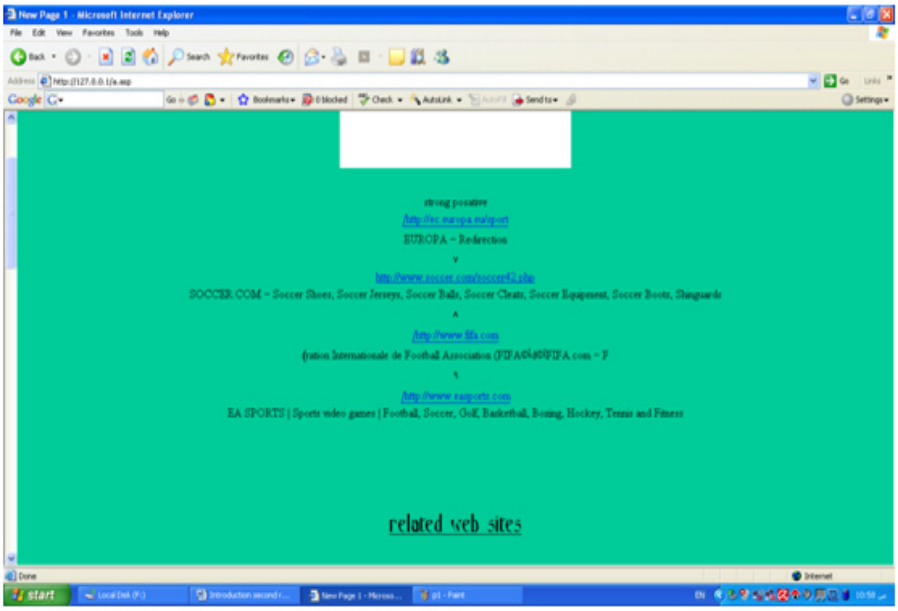

Figure 8: Search Engines Result for Word SOCC

The reader now can see the related topics section in the picture below.

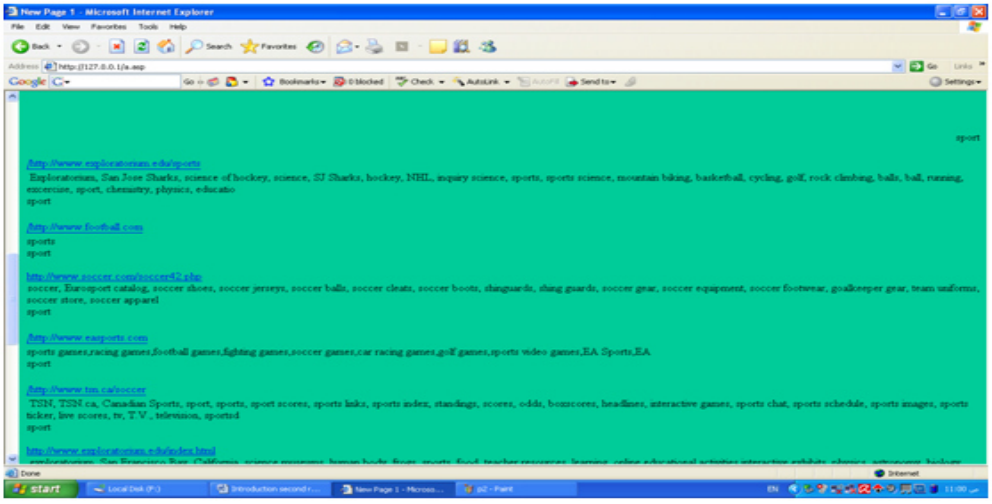

Figure 9: Related Pages for the Result According to the Entered Word Correlation

SUGGESTION FOR FUTURE WORK

- Use xml technologies like semantic web and data mining to build clever search engines.
- Develop more expand word correlation network in order to specify relation between words and topics.

REFERENCES

1. John Wang, ***"encyclopedia of data warehousing and mining"***, idea group publishing, 2006
2. Anthony scime, ***"web mining application and techniques"***, idea group publishing, 2005
3. **Business dictionary.com, "personalization definition",** http://www.businessdictionary.com/definition/web-personalization.html, 2010

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. Liping jing, lixin Zhou, Michael k., Joshua huang,**"** ***ontology based distance measure for text clustering*****",** www.siam.org/meetings/sdmo6/workproceed/text%20mining/jing.pdf