2010 International Conference on Web Information Systems and Mining

My Internet

Ben Choi
Computer Science
Louisiana Tech University
USA
pro@BenChoi.org

Abstract — **This paper describes a new type of search engine that allows the users to have more interaction and control of their Internet. Using the new search engine, we can treat Web pages and files on the Internet like files on our own disk directory. We can browse, organize, search, and even edit Internet files the same way as we do on our files. The proposed approach is to present the search results in a hierarchical structure much like a directory tree structure. This new approach not only allows the users to click on links, but also allows them to move a link from one folder to another. The proposed approach has four advantages: providing new ways for people to use the Internet, giving people more control of their search engine and their Internet, providing new ways to organize and classify Web pages found on the Internet, and providing new ways to capture the preferences of the Internet communities. The proposed approach to build a new type of search engine has been patented, implemented, and tested.**

Keywords - search engine; Web page classification; Web mining; usage mining; Web ontology

I. Introduction

We are facing information overloaded. Finding information relevant to what we are seeking is becoming more important as the Web is growing in explosive speed. Nowadays, most people try to find whatever information on the Web by using search engines. Given a few search keywords, most search engines today will retrieve more than a few thousand Web pages. The problem now is that we need to scan pages after pages, manually and time consumedly, to find what we need or often give up without getting the needed information.

We need to address the problem of helping Web users to find the information that they need. There are several approaches to address the problem. The currently most popular method to address the problem is by ordering the search results and presenting to the users the most relevant pages first. This method is called page ranking, which is one of the important factors that makes Google currently the most successful search engine. Google uses over 100 factors in their methods to rank the search results [22]. Their methods seem to help Web users find the needed information quicker than their competitors. Even with the help of page ranking, we are facing the problem of manually performing sequential search through Web pages after Web pages.

Another approach to help Web users to find the information that they need is by presenting the search results in a hierarchical structure much like a directory tree structure. Using the tree structure, the Web users can browse from one group of Web pages to another group, much like browsing the computer files on a directory tree. For example, if a Web user searches the word “apple”. He can focus his search on the computer group if he is looking for Apple computers. She can browse the fruit group if she is looking for healthy foods. Using this approach could largely reduce the manual search time for the Web users. To make this approach possible, Web pages need to be first classified or clustered into groups forming hierarchical structures [10, 25].

In this paper, we will focus on the approaches to present search results in hierarchical structures. We addressed some other approaches in relating to building better search engines in other publications [5, 6, 9].

This paper focuses on the research direction to create new type of search engine that allows the users to have more interaction and control of their Internet [11, 12]. The proposed approach is to present the search results in a hierarchical structure much like a directory tree structure. The approach not only allows the users to click on links, but also allows them to move a link from one directory to another.

When we use the proposed new search engine for the first time, we get an Internet directory that contains folders and subfolders of categorized Web pages and files found on the Internet. We use the Internet directory the same way as we use our own disk directory. We can move files from a folder to another, in doing so we reclassify Web pages or Internet files. Whatever we do to the Internet directory is saved for us. We have more control of our Internet. Our changes to our Internet directory and other people’s changes to their Internet directories are gathered and analyzed to produce the next generation of Internet directory that is given to the first-time users. The next generation of Internet directory also contains new Web pages found on the Internet, which is also used to update our Internet directory. We are a member of the Internet community and we help reshape our Internet.

The new search engine described in this paper provides the following four advantages:

- Providing new ways for people to use the Internet,
- Giving people more control of their search engine and their Internet,
- Providing news way to organize and classify Web pages found on the Internet, and
- Providing a new way to capture the preferences of the Internet communities.

978-0-7695-4224-9/10 $26.00 © 2010 IEEE
DOI 10.1109/WISM.2010.69

The remaining of this paper is organized as follows. Section II outlines related work. Section III describes the proposed approach for building a new type of search engine. Section IV provides some details on implementation and testing of the new search engine and outlines further developments. And, Section V gives the conclusion and outlines the future work.

II. Related Work

Since the success of search engine depends on its ranking methods, many researches on search engine has been focused on Web page ranking [3], which arranges Web pages into sequence order based on their relevancies to the search key words. An effective ranking method has great commercial values thus many of the research results have been patented. One the most famous work is the PageRank, which was developed in Stanford University, patented [19], and licensed exclusively to Google. The key idea of the method is to view all the Web pages forming a weighed graph, having each Web page as a vertex and the links between Web pages as the edges. Each Web page is assigned a weight to measure its importance, that is, a Web page, having a large number of other Web pages linking into it, will become more important.

Many researches on search engine have focused on extending and improving PageRank method [3]. For instance, the research in Gianna et al [14] focused on improving the processing speed. The research in Xing and Ghorhani [24] focused on taking into account the importance of the in-links and out-links and the popularity of Web pages. The research in Shi et al [21] focused on performing distributed page ranking on top of peer-to-peer networks. Many new and innovative ideas have been proposed for ranking Web pages. For instance, Diligenti et al [13] proposed a unified probabilistic framework for ranking Web page. Wu and Aberer [23] related the behavior of Web surfing to Swarm Intelligent and ranked Web pages based on the interactions of the Web surfers and the search engine.

Another research direction on search engine is to arrange search results into hierarchical structure much like a directory tree structure. This technique for managing vast amounts of information on the Internet is to arrange Web pages into categories. From the point of view of Web page classification, search engines can be grouped into three classes: manual classification, non-classification, or automatic classification. The current status is that there is one search engine providing manual classification (Yahoo), one providing automatic classification (Northern Light), and the rest of them either provide not classification or rudimental manual classification (Google, Bing, A9, AltaVista, Excite, Go, DirectHit, and Lycos).

The automated categorization of Web documents has been investigated for many years. Northern Light received a United States patent on July 13, 1999 for their classification mechanisms [17]. Mladenic [18] has investigated the automatic construction of Web directories. Pazzani and Billsus [20] apply Bayesian classifiers to the creation and revision of user profiles. WebWather [16] performs as a learning apprentice that perceives user's actions when browsing on the Internet, and learns to rate links on the base of current page and the user's interests. For the techniques of construction of Web page classifiers, several solutions have been proposed in the literature, such as Bayesian classifiers [20], decision trees [1], and adaptations of Rocchio's algorithm to text categorization [15]. For more detail and recent research on automatically organizing Web pages into categories, see the two book chapters by the author [7, 8].

This paper proposes another new research direction on search engine, which takes advantage of the abilities to automatically organize Web pages into categories, and which also introduces new methods for allowing Web users to perform manual re-classification of the Web pages.

III. Building a New Type of Search Engine

Unlike a regular search engine that focuses on searching, the proposed new search engine also focuses on providing more control to the Web users. In addition to searching, our new search engine provides methods for allowing multiple users to arrange and modify one existing classification that can be category of files on the Internet or Intranet.

Initially a global Internet directory is created (existing classification or directory structure can also be used). The global Internet directory is presented to a user in a directory structure that has similar looks and feels as user’s own directory. In addition to searching, the users can browse the Internet directory, and can even manipulate the files in the Internet directory in the same way as manipulating their own files. For example, they can move files from one sub-directory to another, rename sub-directory or files, open sub-directory or files, and even edit files.

The results of each user’s changes to the global Internet directory are stored in a file called delta file. The delta file records the user’s preferred arrangement of the Internet files. For a new user a default global Internet directory is provided. The delta file is used to update the global Internet directory to create a customized Internet directory for each user.

All the users’ changes to the Internet directory are collected and analyzed. The collection of the usage histories of all users records the overall preferred arrangement of the Internet files. It is used to update the global Internet directory to reflect the needs and preferences of the entire user community.

The overall working of the new search engine system is highlighted as shown in Figure 1. The system requires six major processes as shown in six circles in the figure. For clarity, the figure describes the interaction of one Web user and the search engine, which is designed to handle millions of users. The functions of each of the six processes are described below.

(1) Arranging files in the Internet to form a global Internet directory: Web pages (or files) stored in the Internet (networked computers located all over the world) are collected, analyzed, and classified to build a global Internet directory. The directory contains links to the files and it groups the files into meaningful categories (folders). This process also periodically updates the global Internet directory to add new files and delete obsolete links. The

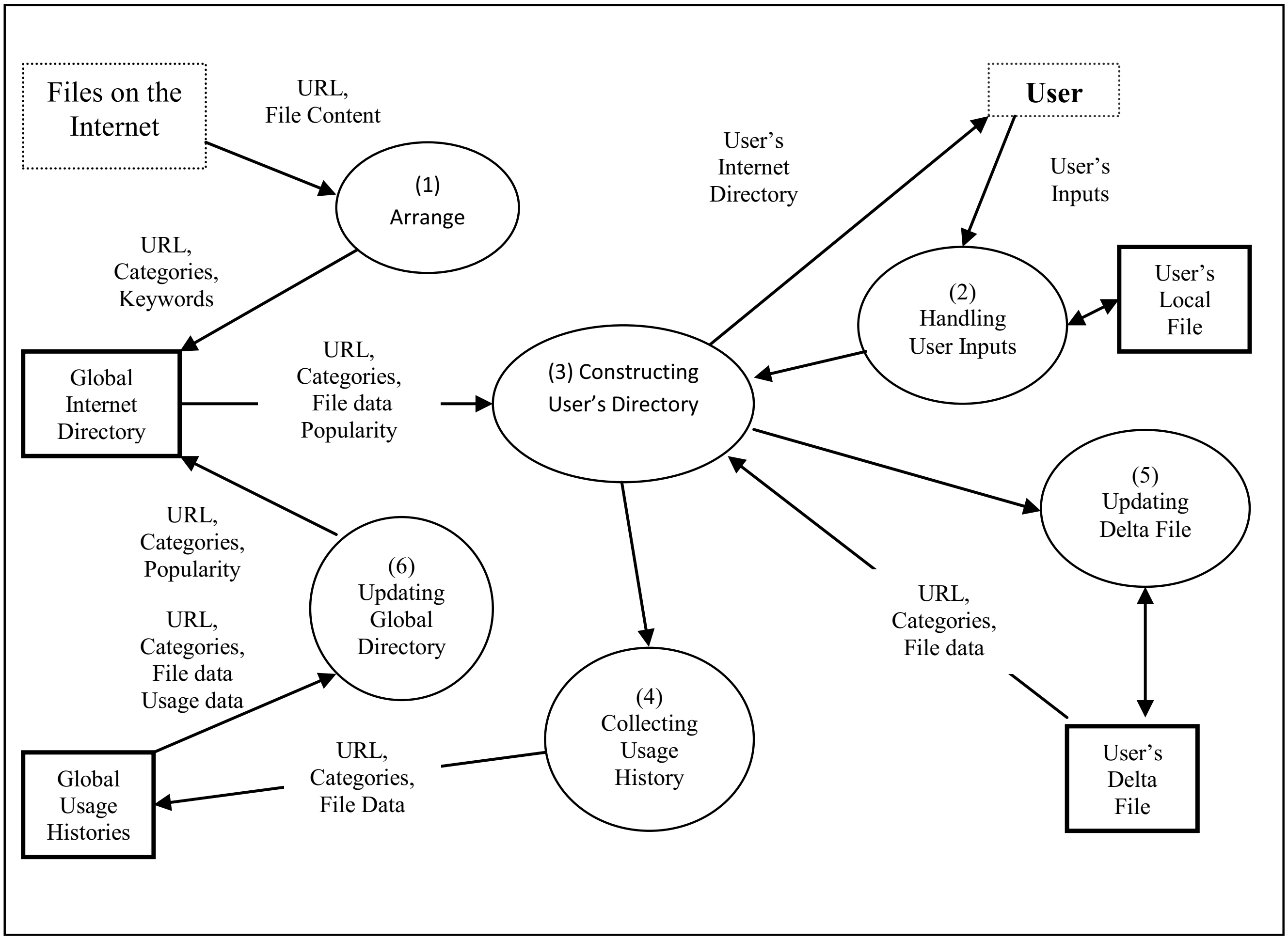

Files on the Internet
URL, File Content
(1) Arrange
User
User's Internet Directory
User's Inputs
URL, Categories, Keywords
(2) Handling User Inputs
User's Local File
Global Internet Directory
URL, Categories, File data Popularity
(3) Constructing User's Directory
(5) Updating Delta File
URL, Categories, Popularity
(6) Updating Global Directory
URL, Categories, File data
URL, Categories, File data Usage data
(4) Collecting Usage History
User's Delta File
URL, Categories, File Data
Global Usage Histories

Figure 1. New Search Engine System [12]

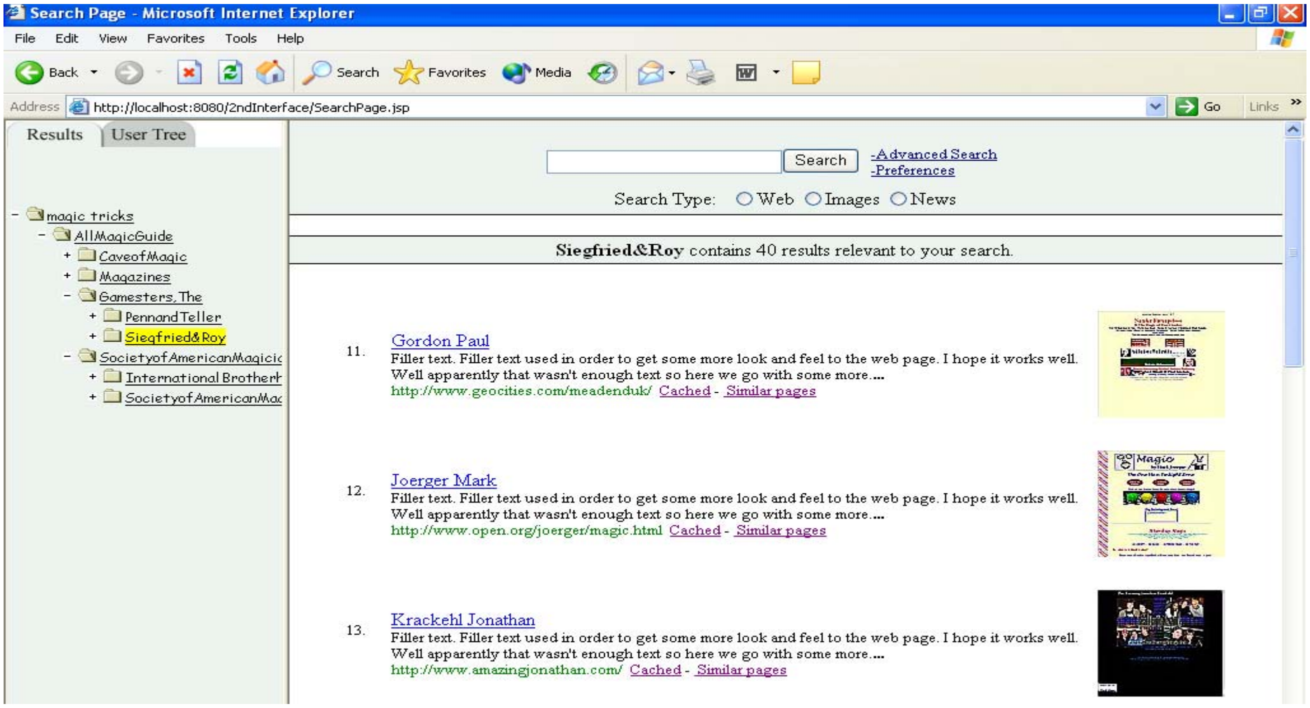
process further consists of Web crawler that automatically locates and retrieves Web pages. It also consists of automatic Web page classification engines that automatically analyzes Web pages and puts them into the appropriate categories (folders).

(2) Handling a user's inputs to arrange and modify files in the user's Internet directory: The user can treat the user's Internet directory as user's own directory. Functions such as Move, Copy, Rename, Delete, and Create can be applied to a folder or a file. This process receives user's inputs and sends requests to the servers. The server will handle the requests and send back an updated directory.

(3) Constructing a user's Internet directory based on the global Internet directory and a user's delta file: A user's Internet directory will be constructed if the user's request is not a Search, otherwise the search request will be processed and the search results will be sent back to the client. The user's delta file contains the user's preference settings and customization data. Based on the delta file the global Internet directory is customized to produce the user's Internet directory. The delta file contains a history of users' actions preformed on the global Internet directory. The history records the difference between the global Internet directory and the user's preferred Internet directory. The data stored in the user's delta file is sent to the server. The delta file may also be stored in the server under the user's account. Based on the data the server amends the global Internet directory and creates a customized Internet directory for the user.

(4) Collecting user's arrangement and usage histories: The data on all the requests received in the server are collected and summarized. These data are stored in the global Usage Histories database. The data include the attributes such as the number of open function preformed on a particular file, the number of times a file has been move from one folder to another folder, the number of times a file has been renamed, and other statistical data. These attributes are updated based on each user request.

(5) Updating the user's delta file based on the user's inputs: The user's arrangement and modification of the user's Internet directory are recorded in the delta file. The delta file will be updated each time the user makes any change to the Internet directory. The delta file will also be updated after the global Internet directory has been updated.



http://localhost:8080/2ndInterface/SearchPage.jsp

Search Type: ○ Web ○ Images ○ News

- magic tricks
  - AllMagicGuide
    + CaveofMagic
    + Magazines
    - Gamesters, The
      + PennandTeller
      + Siegfried&Roy
    - SocietyofAmericanMagicic
      + International Brotherh
      + SocietyofAmericanMac

Siegfried&Roy contains 40 results relevant to your search.

11. Gordon Paul
Filler text. Filler text used in order to get some more look and feel to the web page. I hope it works well. Well apparently that wasn't enough text so here we go with some more....
http://www.geocities.com/meadenduk/ Cached - Similar pages

12. Joerger Mark
Filler text. Filler text used in order to get some more look and feel to the web page. I hope it works well. Well apparently that wasn't enough text so here we go with some more....
http://www.open.org/joerger/magic.html Cached - Similar pages

13. Krackehl Jonathan
Filler text. Filler text used in order to get some more look and feel to the web page. I hope it works well. Well apparently that wasn't enough text so here we go with some more....
http://www.amazingjonathan.com/ Cached - Similar pages

Figure 2. Search Result User Interface: Showing Categories, Summaries, and Page Previews.

In this case, those changes that have been incorporated into the global Internet directory will be removed from the user's delta file.

(6) Updating the global Internet directory based on the collected histories: Periodically the collected global usage histories are analyzed. The global Internet directory is updated to reflect the common preference of all users. The update includes changing the popularity attribute of the file, moving a file from one folder to another, renaming a file or a folder, and deleting a file from a folder. All these update are based on the statistical data stored in the global usage histories database. If majority of the users prefers a new arrangement, then this new arrangement will be incorporated into the global Internet directory.

For being a search engine, the proposed new search engine also handles search requests. Figure 2 shows one of our designs to represent the search results. In additional to present the search results in a list format ordered based on relevancy, this search engine also presents the search results in categories ordered based on relevancy, and also provides the page previews, as shown in the figure. The Web users can focus and narrow down their search by selecting the folders (categories) that they are interested in. Some folders also contain sub-folders that allow users to further homing into what they are looking for.

IV. Implementation and Testing

The proposed approach to build a new type of search engine has been patented [12], implemented, and tested. Figure 2 shows search results retrieved from our working prototype new search engine. By using Java, we were able to implement drag-and-drop functionality that allows Web users to move a Web page from one folder to another. Our search engine utilizes object database, called ObjectStore, which allows our search engine to effectively organize Web pages into categories and to effectively retrieve search results through Java interfaces.

Over twenty students have been working to develop various aspects of building new and more advanced search engine. Some students work on creating new indexing methods to speed up keyword search [2], and some work on ranking Web pages relevant to search keywords [6]. Some students work on classification and clustering of Web pages [7, 10, 25], and some work on create new parallel and distributed computer systems to support millions of search users [5]. Some students work on even more advanced feature, including automatic Web page summarization [9]. Readers are encouraged to see the related publications for details on the various aspects of building advanced search engine.

V. Conclusion and Future Work

This paper outlined new approaches to build a new type of search engine that allows the users to have more interaction and control of their Internet. The proposed approaches organize Web pages into hierarchical structures of categories. They allow the Web users to interact the Web pages found in the Internet much like interacting on files stored their own disk directory tree. Search results are also presented to the users in hierarchical structures of categories, which allow the users to quickly home in to what they are looking for. Building a completely new search engine is a very large project and various aspects have been developed and reported elsewhere.

More research can be done on developing new and more advance search engine. While this paper proposed one new way for people to interact with Web pages found on the Internet, future research can provide other innovative methods. Future research can also be done to provide many other aspects of building advanced search engine, including for example, advanced methods for Web page summarization.

REFERENCES

[1] Apte C., Damerau F., and Weiss S. M., "Automated learning of decision rules for text categorization," *ACM Trans. on Information Systems*, 12(3), pp. 223-251, 1994.

[2] Baberwal S. & Choi B., "Speeding up Keyword Search for Search Engines," *The 3rd IASTED International Conference on Communications, Internet, and Information Technology*, pp. 255-260, 2004.

[3] Berkhin P., "A Survey on PageRank Computing," *Internet Mathematics*, Vol. 2, No. 1: 73-120, 2005.

[4] Choi B. & Huang X., "Web Page Summarization by Using Concept Hierarchies," *International Conference on Agents and Artificial Intelligence* (ICAART), pp.281-286, Proto, Portugal, January, 2009.

[5] Choi B. & Dhawan R., "Agent Space Architecture for Search Engines," *IEEE/WIC/ACM International Conference on Intelligent Agent Technology*, pp. 521-525, 2004.

[6] Choi B. & Tyagi S, "Ranking Web Pages Relevant to Search Keywords," *The IADIS International Conference on WWW/Internet*, pp.200-205, November 2009.

[7] Choi B. & Yao Z, "Web Mining by Automatically Organizing Web Pages into Categories," *Distributed Artificial Intelligence, Agent Technology, and Collaborative Applications*, Idea Group Inc (IGI), chapter XII, pp. 214-231, 2008.

[8] Choi B. & Yao Z., "Web Page Classification," *Foundations and Advances in Data Mining*, Springer-Verag, pp. 221 - 274, 2005.

[9] Choi B. and Huang X., "Web Page Summarization by Using Concept Hierarchies," *International Conference on Agents and Artificial Intelligence* (ICAART 2009), pp.281-286, January 2009.

[10] Choi B. and Peng X., "Dynamic and hierarchical classification of web pages," *Online Information Review*, 28(2) 139-147, 2004.

[11] Choi B., "Making Sense of Search Results by Automatic Web-page Classifications," *WebNet 2001 -- World Conference on the WWW and Internet*, pp.184-186, 2001.

[12] Choi B., "Method and Apparatus for Individualizing and Updating a Directory of Computer Files," *United States Patent* # 7,134,082, November 7, 2006.

[13] Diligenti M., Gori M., and Maggini M., "A Unified Probabilistic Framework for Web page scoring Systems," *IEEE Transactions on Knowledge and Data Engineering, Volume16-Issue 1*, pp. 4-16, 2004.

[14] Gianna M., Corso, D., Gullí, A, and Romani, F., "Fast PageRank Computation via a Sparse Linear System," *Internet Mathematics,* Vol. 2, No. 3: 251-273, 2006.

[15] Ittner D., Lewis D., and Ahn D., "Text categorization of low quality images," *Symposium on Document Analysis and Information Retrieval*, pp. 301-515, 1995.

[16] Joachims T., "A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization," *Proc. of the 14th International Conference on Machine Learning*, pp. 143-151, 1997.

[17] Krellenstein M. F., Method and apparatus for searching a database of records, US Patent & Trademark Office, United States Patent 5,924,090, July 13, 1999.

[18] Mladenic D., Turning Yahoo into an automatic web-page classifier. In H. Prade (Ed.), *Proc. 13th European Conference on Artificial Intelligence*, pp.471-474, 1998.

[19] Page L., "Method for node ranking in a linked database," US Patent 6,285,999, 2001.

[20] Pazzani M. & Billsus D, "Learning and revising user profiles: The identification of interesting web sites," *Machine Learning Journal*, 23, pp. 313-331, 1997.

[21] Shi S.M., Yu J., Yang G.W., and Wang D.X., "Distributed Page Ranking in structured P2P networks," *International Conference on Parallel Processing* (ICPP'03), pp. 179-186, 2003.

[22] Vaughn, 2008. "Google search engine optimization information," http://www.vaughns-1-pagers.com/internet/googlerankingfactors.htm.

[23] Wu J., and Aberer K., "Swarm intelligence surfing in the web," *J.M. Cueva Lovelle et al. (Eds.): ICWE* 2003, *LNCS* 2722, pp. 431–440, 2003.

[24] Xing W., and Ghorbani A., "Weighted PageRank Algorithm," *Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR'04),* pp 305-314, 2004.

[25] Yao Z. & Choi B., "Clustering Web Pages into Hierarchical Categories," *International Journal of Intelligent Information Technologies*, Vol. 3, No. 2, pp.17-35, April-June, 2007.