Development of Information Agent Reranking By Using Weights Measurement

Aliaa A.Youssif aliaay@helwan.edu.eg
Computer Science and Information Technology/ Computer Science/Helwan University
Helwan, Egypt

Ashraf A. Darwish amodarwish@yahoo.com
Computer Science/Mathematics/Helwan University
Helwan, Egypt

Ahmed Roshdy ahmed.mind@hotmail.com
Computer Science/Mathematics/Helwan University
Helwan, Egypt

Abstract

Web search is one of the most challenging problems of the Internet today, seeking to provide users with search results most relevant to their information needs. The new improvements of search engines technologies have made available to the internet users an enormous amount of knowledge that can be accessed in many different ways. However, there are some problems that face the search engines. In this paper proposes an agent system which parses information sources and extract weights that determine the powerful of relevant information sources and prove that the word positions scores may affect on reduction the relevance of those information sources. Moreover, it will show that the user profile plays an important role in effectiveness of re-ranking and updating ranking relevant web pages where agent learns user behavior by observing user browsing for the interested result pages. In experimental work section shows how weights algorithm gets more relevant web pages than other algorithms that use word position which may reduce the value of relevance of web pages.

Keywords: Information Agent, Knowledge-base, Weights, and Re-ranking.

1. INTRODUCTION

The vast amount information source in web today rendered the intelligent information agent subtending to challenge re-ranking web pages on the fly. Given the constantly increasing information overflow of the digital age, the importance of information retrieval has become critical. Since the documents source of information retrieval have two types unstructured records and semi-structured records depending on natural language text, and also kinds of data that can be unstructured, e.g., photographic images, audio, video, etc [1,2].

The famous search engines on the marketing are now providing search facilities for databases containing billions of web pages, where queries are executed instantly [3]. But there are some problems that face search engines as following:

- The crawler-based search engines like Google can catalogues web pages and the documents automatically where it crawls the web, then the users searches through what they have found and sometimes they lead to poor queries and increase the gap between information need and request [2].
- On the other side most of search engines beside Google crawls the web pages by tracking hyperlinks to find authoritative pages using HITS algorithms that indeed get satisfied relevant pages for users queries, but they suffer from another problem that some authoritative pages have subjects and/or information sources and didn't satisfy the relevance of the user request.

Though feedback is one approach that deduces information about the user and his/her search context [4], beside that the techniques that depend on relevance feedback re-rank the search results by re-calculating the relative importance of keywords in the query.

In this paper it concentrates on the processing of information sources and those contents to extract words weights that are important to determine the content of the page subject more powerful and can re-rank web pages more efficiently [4].

The paper is organized as follows: the next section is devoted to the related work that shows the reasons of word weight importance more than its position that is used by the other agent's techniques. Section 3 contains detailed description of the proposed information agent system, showing its several components. The effectiveness of user information profile for learning agent and effecting on re-ranking web pages in sections 4 and 5, are introduced respectively. The report on an implementation and describe some experimental results in section 6 and conclusion and future work have been presented in section 7.

2. RELATED WORK

The information agents developed and have addressed many tasks, such as assisting users in searching and browsing the Web, finding, filtering and accessing large amounts of information on behalf of users, and present a reduced and potentially relevant part of this information to them [5]. Information extraction from semi-structured documents has received more attention recently. Agent in most cases is called *PersonalSearcher* [6] where it helps users to identify filtered documents with high probability of being relevant to them by transforming the search process according to users' information preferences and learns about a user's interests from the observation of user browsing on the Web.

Some techniques focus on the structure of web page like CASA [1] and Textual Case-Base Reasoning (TCBR) [7]. These algorithms depend on the position of the word either in text, line or paragraph and give a degree for its position increasing from text till paragraph which in more cases web pages puts words in text or low positions that refer to hyperlinks linked to addition information that maybe more relevant to user request. Also the advertisements web pages interest in the photos of goods and prices and little care about words positions.

Another technique in its algorithm calculate number of links that are linked to that page, and checks position level of query keywords if it is either highest, middle, or lowest in web page as in automated fuzzy agent [8]. Such this technique faces another problem that some pages have titles and keywords of user query request and have high position but the information source is not relevant for user query. Beside that some web pages have more relevant information sources and titles either in medium or low position and add some commercial advertisements in upper layer. In another hand the technique repeats itself for each page tracked by hyperlinks which make overhead in processing.

The research focuses on extracting words weights to determine the rich information, density of these words in the information source and how much they are more relevant to user query, and they effect on re-ranking web pages neither depending on them position nor subject title position.

3. THE PROPOSED AGENT ARCHITECTURE

Since the agent is an integrated part of the end-program, there are two major paradigms for the architecture. The first approach is automated treating with the results to get knowledge of certain domain and learns from user or other agent for getting feedback about user interest. The second approach is knowledge-based approach where the agent has extensive information of certain domain and learns from user profile about his/her behavior and interest to expand the feedback of the agent. The paper interests for the second approach and explain below of the architecture.

Figure 1 presents the high level view of the proposed architecture for information agent depending on words weights and information profile. The knowledge-base system which is the feedback of information agent has four major parts for processing documents as following:

(1) Entering information sources by downloading URLs and web pages that are retrieved from the search engine or more search engines and indexing them for removing redundant URLs.
(2) Parsing web pages by parser for extracting semi-structured data, Meta tags and link tags then determining words weights.
(3) Re-ranking web pages by calculating relevance equation ratio.
(4) Learning information agent from information profile about user browsing result web pages.

The user, by means of a graphical interface, submits a query to the parser by the keywords. The parser extract query to its words and find word weights for each web page which is downloaded in the knowledge-base. After that the knowledge-base calculate the relevance word weight ratio to re-ranking web pages by descending and represent re-ranked result to user interface.

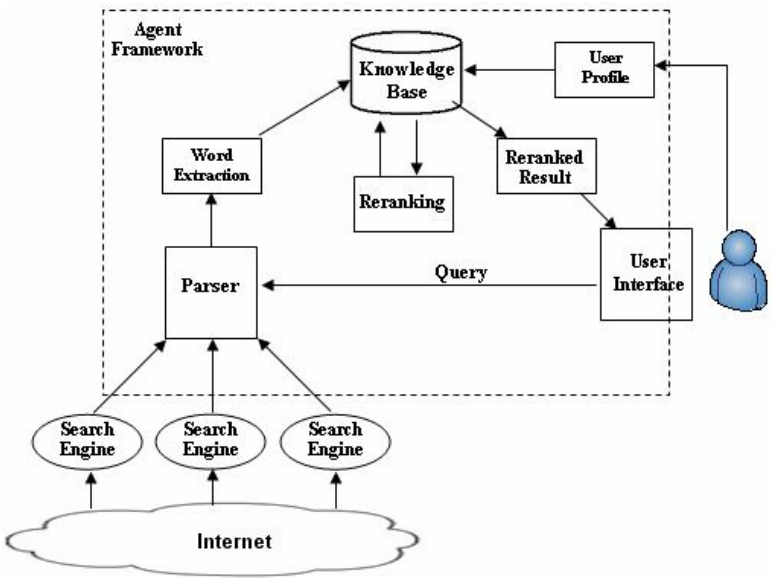

FIGURE 1: Show the high level of agent architecture

3.1. Knowledge-base

As it is explained in section 2 the importance of knowledge-base for information agent feedback in certain domain. The architecture contains on the attributes of the domain, interactions, and filtered results. The content of the base shown in Figure 2 is described as following:

DATA_EXTRACTION entity stores URL_ID (for not replication of URLs), URLs of result web pages that are downloaded from the search engine, HTML source code, meta tags, hyperlink tags, title and header tags, the number of queries that request a web page, and textfix that contains paragraphs of data source and be processed from the parser.

LOCAL_WORDS entity stores the words of each page which is extracted by the parser and those calculated weights. GENERAL_WORDS entity stores calculated word weight for all web pages that are stored in the knowledge-base.

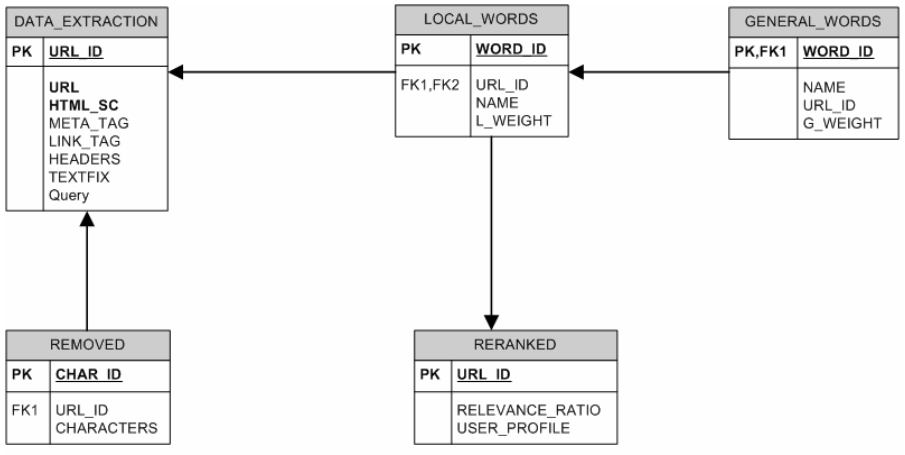

FIGURE 2: Show ER diagram of the knowledge-base

RERANKED_PAGE entity stores the calculated relevance word weight ratio for each page and counting of user browsing for web pages result. REMOVED entity stores the characters, letters, Preposition, etc that are used in paragraphs.

3.2. Parsing

The mission of the parser is extracting information sources of the web pages and determines the kind of data source according to the requirements of search engine or information agent technique [1, 9, 10, 11, 12]. The Paper focuses on the frequency of words in web pages to determine those weights and find most relevant web pages depending on the concentration of words query on those pages.

Therefore the more increasing of words query weights on a web page, the more determination of relevant web page. Example for what to discuss about, when user submit specific query request, the parser begins extract these words and compares them with those weights in knowledge-base and then finds the highly weights that refer to highly relevant web pages after calculating relevance word weight ratio.

4. USER PROFILE

Since user-profiling involves information about users of behavior-based information about them, for example their actions on browsing some pages the which learns the information agent about their interesting, besides increasing the information of the knowledge-base that are considered a feedback of user interesting [13,14].

User behavior is the most deserving sources of information for getting profiles and, from their successful interpretation; the information agent will be able to follow the user actions [15]. Therefore our agent monitors the user browsing for result pages and begins counting the user behaviors by accessing web pages; the results will be automatically updated when unobserved

voting is highly score in some web pages than others. Implicit interest indicators like time consumed in reading a Web page (considering its length), the amount of scrolling in a page, and whether it was added to the list of bookmarks or not are considered a strong correlation with explicit interest feedback, as opposed to the number of mouse clicks, which does not result in a good indicator [16,17].

The dynamic part reflects the path of user behavior information in the way of user's search. The static part lets user to show his/her desire when agent offers some options to the user. In many cases user ignores static part [18].

5. RE-RANKING

Many search engines hide the mechanism used for the document ranking this is the reason of the merging problem for the results becomes even more difficult. In addition, these kinds of approaches suffer from ignoring or knowing nothing about the user conducting the search, nor the context of the search [4].

Our evaluation for ranking of relative documents is depending on some weights using equation of relevance which is discussed in [4]. For each document d in the response to query q, the document rating is evaluated as follows (the adapted cosine function):

$$relevance_{q,d} = \frac{\sum w_d \, X \, w_{prof} \, X \, w_q}{W_d} \quad (1)$$

Where w_d is the weight of word in the document *d*, w_{prof} is the weight of chosen document d (i.e. accessing web page from different users as mentioned in section 3), we supposed its score in the first re-ranked iteration by 1, w_q is the weight of query for chosen document from agent result, and W_d is evaluated as following:

$$W_d = \sqrt{\sum (w_{prof} \; w_d)^2} \quad (2)$$

6. METHODOLOGY

The sequence of proposed algorithm can be obtained as following:

1) The user submits his/her search query.
2) The agent searches several other search engines with user's query.
3) The agent downloads the resulted web pages
4) The agent extracts the information source from web pages using the Parser and saves the information in the knowledge-base.
5) The agent calculates words weights of the query keywords.
6) The agent calculates relevance score for each web page and re-ranks them in a descending order.
7) The agent keeps track of the re-ranked web pages. Once a user browses any of these web pages, the w_{prof} score of such page increases.
8) Steps 1-7 will be repeated for each search.
9) If the search hits any of the re-ranked web pages, it will increase its w_q score.
10) The agent recalculates the relevance score and updates re-ranked web pages.

7. EXPERIMENTAL RESULTS

The experiment was constructed in three stages: training, retrieval and learning. For the former, the knowledge-base was built using the method proposed in section 3.1 and determined the domain to download web pages that was cars advertisements. Also parser was built to deal with

downloaded web pages for extracting information and re-ranking pages. In the retrieval stage, knowledge-base was retrieved with each first few web pages were downloaded from five search engines (Google, AltaVista, Yahoo, MSN, and Excite) and indexed these pages to prevent redundant pages. In the program we generated 20 different queries and calculated summation average of number of relevant pages and irrelevant pages. In learning stage we browsed in re-ranked pages to learn information agent of interested pages by accessing them and calculating period time of browsing for each page; the longest time taken page get score in w_{prof} and be affected for re-ranking again. The stages are described briefly as following:

7.1. Training Stage:

(1) Build a knowledge-base and determine the domain for downloading pages.
(2) Build parser.

7.2. Retrieval Stage:

(1) Download first 200 web pages from each search engine.
(2) Index web pages to prevent redundant web pages.
(3) Generate different queries and calculated summation average.

7.3. Learning Stage:

For each term, Browsed in re-ranked pages to give interested pages scores in w_{prof} for next re-ranking.

The Figure 4 represents the most common words of user query that insure the relevance of web page.

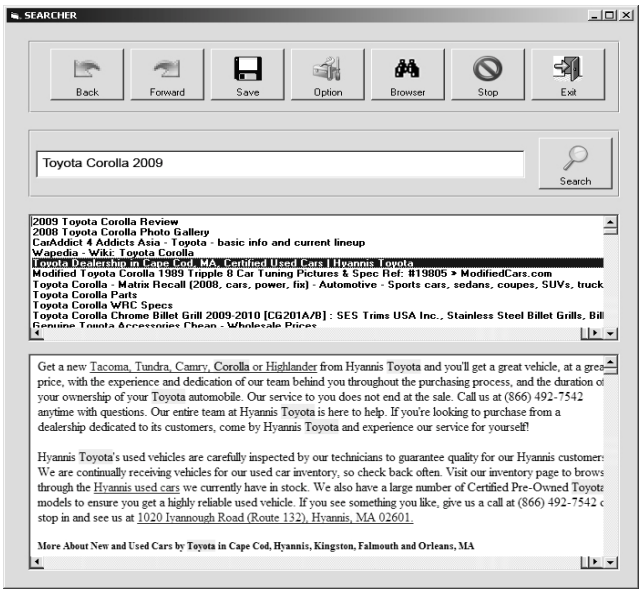

FIGURE 3: Show Agent Search Result

7.4. Comparative Analysis

In order to evaluate our proposed algorithm, we compared our findings with the score resulted from TCBR [7], CASA [1] Automated fuzzy algorithm [8]. Table 1 shows the comparative evaluation of common queries that have effect of word position in average of first 20 relevant web pages.

	Automated fuzzy score	TCBR	CASA score	Proposed score
MAZDA ph	3.016	2.255	0.918	4.565
FIAT Audio	3.927	3.573	1.641	5.126
BMW Engine	4.537	3.847	2.177	6.337
HONDA in Boston	2.472	1.501	0.861	3.691
Toyota Accessories	4.582	3.674	2.255	5.969
PROTON-GN2 Weight	3.415	2.293	1.106	4.846
PEUGEOT Spares	4.357	3.063	2.144	5.138
JEEP Cylinder	2.732	1.587	0.161	3.813
CHERRY	3.157	2.059	0.345	4.152
NISSAN Carpet	3.437	2.462	1.147	4.598
Used LANOS	3.843	2.826	1.164	5.326
Mercedes Brake	2.952	2.147	1.030	4.293
SEAT Filter	2.638	1.476	0.185	3.232
Hyundai Lights	2.431	1.630	0.289	3.813
SKODA Hatchback	4.362	3.326	2.577	6.429

TABLE 1: The difference relevant scores between algorithms

From Table 1, we noticed that our proposed algorithm gives higher word score than CASA, TCBR and fuzzy algorithm. In computing the relevance of a web page, our equation does not take into account the word location.

However, in some queries like “Toyota Accessories”, “Proton-GN2 Weight”, “Honda in Boston” the factor of word location in CASA algorithm and Δ_{ij} in TCBR equation for word weight $weight(w_j,d_i) = tf_{ij} + \Delta_{ij}$ decrease (where tf_{ij} is term frequency that determines the frequency of word appearance in document) because these words occurred more frequently in lines than in paragraphs and titles.

Moreover, in the automated fuzzy algorithm we noticed that the Position_Score of some words decreases to some extent. This is due to the fact that these words occurred in the middle of the web pages, which affected their score. The relevant web page score is determined by the equation:

Score = (2*Frequency_Score) + Position_Score + Links_Score

Summary:

Figure 3 shows the comparison results between 3 algorithms: CASA algorithm [1], TCBR algorithm [7], automated fuzzy algorithm [8] and weights measures algorithm of the first 200 re-ranked web pages from information agent and that describes the performance of agent can be more efficient depending on weights measurement.

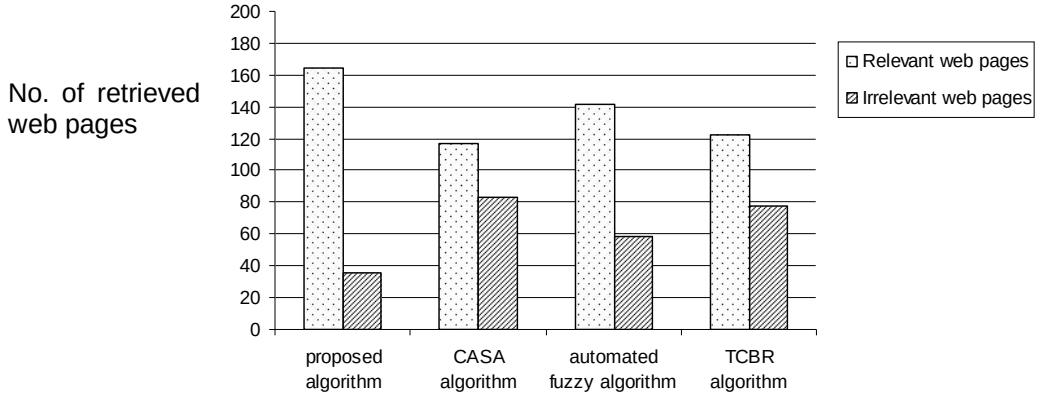

FIGURE 4: Comparison between four algorithms

The performance of information agent can be increased by improving the mechanism of downloading URLs and web pages using semantic web techniques as in [11,12,19] that determine word vectors of context meaning , and reaching to suitable solution to the retrieve relevant documents dynamically.

8. CONCLUSION AND FUTURE WORK

In this paper we scoped on filtering powerful information source by parsing the information to extract weights for finding more relevant pages and re-ranking web pages with learning information agent of user profile about interesting web pages that can be more efficient in re-ranking. The result of the experiment for proposed algorithm is 0.82% of relevant web pages compared to highly percentage of other algorithms about 0.71%. We used Microsoft Access to build knowledge-base and Visual Basic to build the other components of information agent.

9. REFERENCES

[1] X. Gao, "A methodology for building information agents", in Yang, Y., Li, M., Ellis, A. (Eds), Web Technologies and Applications, International Academic Publishers, <http://www.cs.mu.z.au/~xga/apweb/index.htm >, pp.43-52. (1998)

[2] Ed Greengrass "Information Retrieval: A Survey", Available at: http://www. csee.umbc.edu. (2000)

[3] M. Caramia, G. Felici, and A. Pezzoli "Improving search results with data mining in a thematic search engine", Computers and Operations Research Volume 31, Pages: 2387 – 2404, ISSN: 0305-0548. (2004)

[4] B. Chidlovskii, N.S. Glance and M.A. Grasso, "Collaborative Re-Ranking of Search Results", Available at: http://citeseer.ist.psu.edu. (2000)

[5] Daniela Godoy, Silvia Schiaffino, Analia Amandi "Interface agents personalizing Web-based tasks", Cognitive Systems Research Volume 5, Issue 3, Pages 207-222. (2004)

[6] Godoy, D., & Amandi, A. "Personalsearcher: An intelligent agent for searching web pages" In International joint conference IBERAMIASBIA'2000 (pp. 43–52). Atibaia, Sao Paulo, Brazil: Springer. (2000)

[7] D. Lis Godoy, "Generating User Profiles for Information Agents", Available at: http://citeseer.ist.psu.edu. (2003)

[8] M. Mohammadian in the book "Intelligent Agents for Data Mining and Information Retrieval", Idea Group Publishing (an imprint of Idea Group Inc.), ISBN: 1-59140-277-8. (2004)

[9] L. Shen and A.K. Joshi, "An SVM Based Voting Algorithm with Application to Parse Reranking", the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4, Pages: 9 - 16. (2003)

[10] A. Penev and R. Wong, "Shallow NLP techniques for Internet Search", the 29th Australasian Computer Science Conference - Volume 48, Pages: 167 – 176, ISBN ~ ISSN: 1445-1336, 1-920682-30-9. (2006)

[11] C. Cesarano, A. d'Acierno and A. Picariello, "An Intelligent Search Agent System for Semantic Information Retrieval on the Internet", the 5th ACM international workshop on Web information and data management, Pages: 111 – 117, ISBN:1-58113-725-7. (2003)
[12] G. Wisniewski & P. Gallinari "From Layout to Semantic: A Reranking Model for Mapping Web Documents to Mediated XML Representations" proceeding of RIAOCID (2007) Conference, URL: http://dblp.uni-trier.de/db/conf/riao2007.html#WisniewskiG07. (2007)
[13] Bas van Gils and Eric D., "User-profiles for Information Retrieval", Methodologies for Intelligent Systems, volume 542 pages 102-111 ISBN: 978-3-540-54563-7. (2006)
[14] S.E. Middleton, "Capturing knowledge of user preferences with recommender systems", Available at: http://citeseer.ist.psu.edu. (2003)
[15] D. Godoy and A. Amandi "User profiling in personal information agents: a survey", The Knowledge Engineering Review, Volume 20, Pages: 329 - 361 ISSN: 0269-8889. (2005)
[16] Claypool, M., Le, P., Wased, M., & Brown, D. "Implicit interest indicators. Intelligent User Interfaces", Proceedings of the 6th international conference on Intelligent user interfaces Santa Fe, New Mexico, United States Pages: 33 – 40 ISBN:1-58113-325-1 (2001)
[17] Kelly, D., & Belkin, N. J. "Reading time, scrolling and interaction: Exploring implicit sources of user preferences for relevant feedback", proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval (pp. 408–409). New Orleans, LA, USA: ACM Press. (2001)
[18] C. Dharap "Context-based and user-profile driven information retrieval", Fremont CA (US), Philips Corporation New York (2001).
[19] G. Cheng, W. Ge and H. Wu, "Searching Semantic Web Objects Based on Class Hierarchies", proceeding of LDOW Conference, available at: http://iws.seu.edu.cn/publications/cgwq08 . (2008)