# Profile-Based Semantic Method using Heuristics for Web Search Personalization

Hikmat A. M. Abdeljaber
Department of Computer Science
College of Computer Engineering and Sciences
Prince Sattam bin Abdulaziz University
AlKharj, Saudi Arabia

***Abstract*****—User profiles play a critical role in personalizing user search. It assists search systems in retrieving relevant information that is searched on the web considering the user needs. Researchers presented a vast number of profile-based approaches that aims to improve the effectiveness of information retrieval. However, these approaches are syntactic-based which fail to achieve the user satisfaction. By the means that the search results do not meet user preferences, due to the fact that the search is keyword-based rather than semantic-based. Exploiting user profiles with the application of semantic web technology into personalization might produce a step forward in future retrieval systems. By adopting profiling approach and using ontology base characteristics, a semantic-based method using heuristics and KNN algorithm is proposed. It engages searching ontology base domains *horizontally* and *vertically* to discover and extract the closest *concept* to the meaning of the query keyword. The extracted *concept* is used to expand the user query to personalize the search result and present the customized information for individuals.**

***Keywords—Semantic search method; user profile; heuristics; web search personalization; information retrieval***

## I. Introduction

Information Retrieval (IR) is a highly active research field; it depends significantly on the web as the main source of information. It involves assisting users to find information from vast amount of information resources on the Web. However, finding relevant information that satisfies users' query is a vital problem [1, 2]. Search engines perform "one size fits the all", in which the submitted query of keywords returns the same result to users of different interests. Many query keywords may have the same syntax but different semantics (homonyms). These keywords could be located at different horizontal and vertical domains of ontology base of semantic web. They could have different meanings under different concepts within the same domain.

Therefore, the main research focus of this paper is how to find the domain that reflects the closest meaning of the query keyword from various domains of ontology base. Then how to find the *concept* that reflects the closest meaning of that query, from different *concepts* within the same domain. For instance, the query keyword 'cell' may be found across various *horizontal* domains such as computer, biology and politics. In addition, the same query 'cell' may be found under various *vertical concepts* within the computer domain including processors, excel and company name. Accordingly, finding which particular *horizontal* domain and particular *vertical concept* within that specific domain for query keyword 'cell' at the same search session. Once the *concept* and domain of the query keyword are found, then it can take an advantage of both in the query expansion to retrieve better search results. This paper presented a novel approach to address these two issues.

Therefore, it is important to optimize the means of personalizing the web search and locating relevant documents tailored for individual users. Web search personalization approaches and techniques have been reviewed in [3-7]. Generally, there are two types of approaches to personalizing the search results, first is by the user query modification and secondly is by the search results re-ranking. On the other hand, there are many techniques that are based on the web contents, web link structure, browsing history, user profiles and user queries. These techniques have been widely used to implement personalized web search models [3, 6]. These models include, but not limited to, the use of hybrid of fuzzy set and ant colony optimization as described in [8]. Firefly algorithm is used to create and choose the cluster based optimal ranked clicked URLs to recommend a set of terms that expand the query search [9]. Ant colony optimization with a genetic algorithm were used by [10] to rank web pages. [11] has used a hybrid of genetic algorithm and back propagation neural network to classify user queries to clusters for web page recommendations. However, the proposed model is implemented using a hybrid of artificial intelligence (AI) heuristics and K-nearest neighbour (KNN) algorithm to extract semantic *concepts* from the ontology base using the user profile to expand the query.

User profile is the major method used for personalizing web search as presented in [12-14]. For capturing user's current information need, [15] have represented user's activities in the form of time-sensitive profile. It integrates both the current and the recurrent interactions with the search engines such as submitted queries, reformulated queries, and clicked results within a session search. User interactions are taken into account under the assumption that recent performed ones are more related to the current needs than to the foregoing ones. Authors in [16] have proposed a Funnel Mesh-5 algorithm. It constructs a search string by taking into account the context of information need and the user intention. This information is identified by the user profile then it is used to generate a personalized disambiguated search string for query expansion. In [17], the user profile is created based on the user search behaviour using the web search logs and the eye tracking. It measures the user behaviour during the query session. Their system keeps on updating the user profile to build and enhance the user profile to suggest more relevant web pages to the user. However, my proposed approach maps the user profile content onto the ontology base of the semantic web technology to extract the closest *concepts* to the meaning of query keywords. Researchers have broadly studied the personalized techniques to represent users' information needs in the user models. These techniques are used not only for monolingual IR systems [5] but also for multilingual IR systems for re-ranking of web search results [18], browsing and searching the behaviour of polyglots [19], and personalizing the query expansion [20].

The idea of the semantic web aims at making the semantics of the web content machine understandable [21]. Semantic web standards and technologies can be used to enable the semantic search [22]. Semantic search, as an application of semantic web in the field of IR, has shown a significant potential in the function of improving the performance of retrieval. Compared with the traditional search that focuses on the frequency of word appearance, the semantic search attempts to understand the meanings hidden in the retrieved documents and users' queries. It works by adding the semantic tags into texts to structuralize and conceptualize the objects within documents [23]. Therefore, many semantic-based methods have been proposed in personalizing the web search. [24] used ontologies to improve the reliability of personalization through exploiting the formal semantics of query-based relevance processing, user preference representation, preference update, and result ranking. In [25], a semantic mapper is used to map the user query terms with personalized ontology, that were created from user web log file to identify semantic relation between user's queries.

However, this research combines the use of both the user profile and the characteristics of ontology base to develop an effective semantic search method. It purposes to deliver better and close results compared to the conventional method currently used in IR. Thus, the main objective of this paper is to implement the semantic-based search method. By acquiring the user intent implicitly by exploiting ontology base and the user profile to personalize the web search and enhance the contextual IR of web documents.

Semantic user profile is created in [26] for capturing scholar's interests, tasks, and competences in different research topics across different projects and publications in the scientific domain. The semantic user profile is modelled through an automated text mining pipeline approach of NLP for using it in semantic publishing applications as personalized web applications. [27] have developed a model to capture user context by generating the query context and the user context. Also, it introduced a forgetting factor to merge the independent user context in the user session for maintaining the evolution of user preferences. In contrast, my proposed model has applied AI heuristics techniques and KNN algorithm that implements a search method pluggable in the semantic web search applications. Therefore, this paper contributes mainly in the area of the semantic-based search methods in the semantic web technology. The development of the new semantic-based search method, as a web search service for the semantic web technology, characterized by its ability to interact with an ontology base, reason, infer, and consequently find and extract the closest meaning of the users' query keywords. It assists users to provide them with the results that are close to accurate, when they work with searching the search engines for specific information.

Query expansion is the process of adding more terms to an original query to attempt to refine the information search and improve the retrieval effectiveness [28-30]. The proposed model used the query expansion to improve the results by including *concepts*, extracted semantically from the ontology base, which lead to retrieving more relevant documents.

The remainder of this paper is structured as follows: Section II describes the structure of the system prototype. Section III illustrates the proposed semantic method for web search personalization. Section IV shows the experimental results and evaluation that were performed to validate the proposed approach. The paper is concluded and future work is presented in Section V.

## II. System Prototype

Information retrieval systems (IRSs) are text-based prototypes that use traditional methods to retrieve information. IRSs perform limited personalization for individual users and consequently provide irrelevant documents in terms of search precision. Combining the profile-based sematic search approaches and the AI techniques together to retrieve information is challenging. Therefore, the user profiles should use the semantic web technologies in IR process to improve the search results. This vision needs prototype to incorporate the semantic web tools and the user profile with the search system. Fig. 1 illustrates a high-level picture of the system prototype.

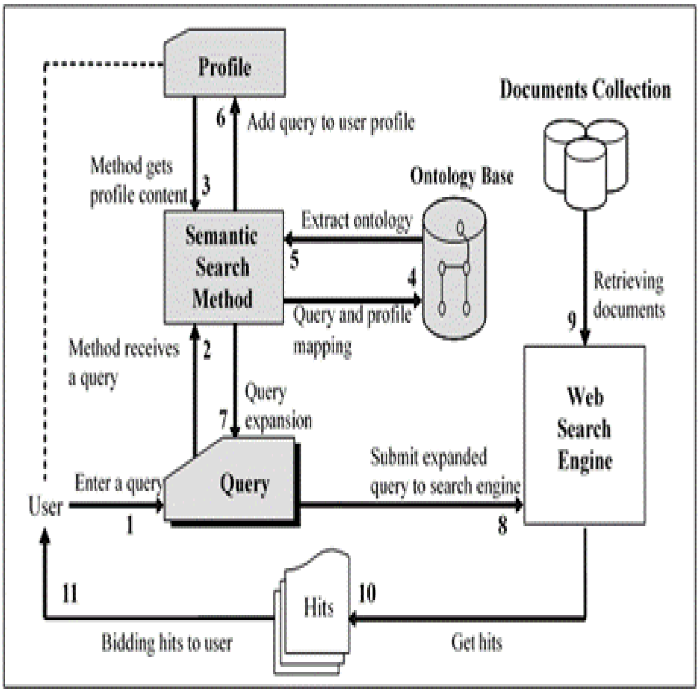

Fig. 1. Architecture of the system prototype.

The architecture of the system prototype consists of user query, semantic search method, user profile, and ontology base. The system prototype is described as follows: First, the user types a query. Second, the semantic search method receives the query. Third, the method obtains the user profile content. Fourth, the method maps the query and the user profile content onto the ontology base. Fifth, the method extracts the closest *concept* to the meaning of the query from the ontology base. Sixth, the method adds the query keywords to the user profile. Seventh, the method expands the query. Query expansion is performed to disambiguate the query by adding the extracted *concept* to the initial query automatically. Since this added information is originally acquired from the ontology base, it reflects the user needs. Eighth, the expanded query is submitted to the search engine. Ninth, the search engine retrieves the documents from the documents collection. Tenth, the search engine gets the hits and finally these hits are provided to the user. However, the proposed system prototype can be built and plugged into the typical IRS of web documents, such as search engines or metasearch engines, without effecting its standard operations.

## III. A Semantic Method for Web Search Personalization

The adopted approach needs a combination of five basic related components as follows: the user query, the user profile to maintain user query keywords, the semantic-based search method to extract from the ontology base the closest *concepts* to the meaning of query keywords, an ontology base to provide these *concepts* as contextual information for query expansion, and the web search engine to search the expanded query. The keywords of user profile and query are mapped onto the ontology base as shown in Fig. 2. The top level nodes of Fig. 2 are root domains of ontology base. The user profile keywords are marked with gray color in the solid circles and the query keyword is marked with a gray color in the dashed circle.

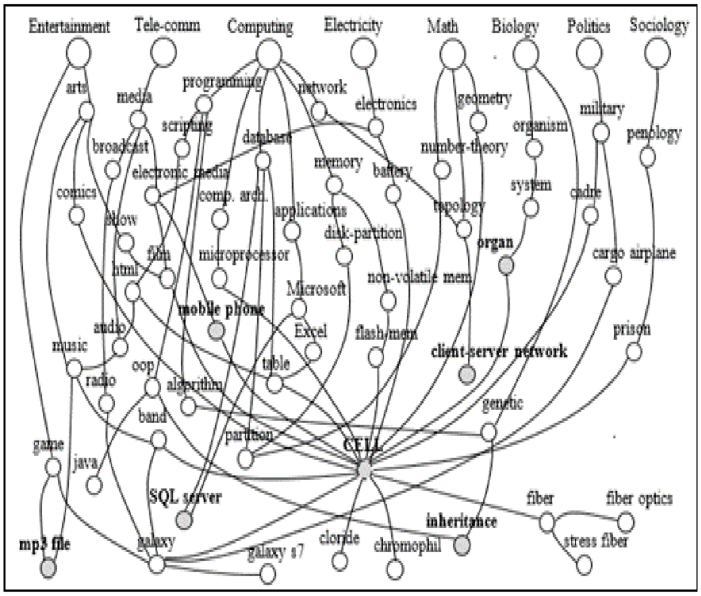

Fig. 2. An ontology base marked with user profile and query terms.

The proposed approach aims to find the closest *horizontal* root domain of the ontology base to the meaning of the given query and then search that domain *vertically* to determine the closest *concept* to the meaning of that query. To find the closest *horizontal* root domain, the information of long-term browsing history and short-term current browsing are utilized. The browsing history concerns the content of user profile whereas the current browsing concerns the user query itself. The heuristics technique is employed to make use of the user profile and query keywords to find the closest *horizontal* root domain. Heuristics provide many algorithms designed to traverse graph paths in order to discover the minimum cost path from a starting node to the goal. This work uses A* heuristic algorithm to find the closest *horizontal* root domain of the ontology base to the meaning of the query keyword. However, application of A* algorithm requires calculating two values for each node $n$ of the ontology base. The first value is known as $g(n)$ which is the distance value from the user profile or query keywords to the node $n$ and the second value is known as $h(n)$ which is the distance value from the root domains to the node $n$. Note that the node $n$ can be an ontology base node (concept), user profile keyword, or query keyword. Furthermore, keywords indicate the user profile and query keywords unless stated otherwise.

The distance from the keywords to the node $n$, $g\ (n)$, is calculated as follows:

$$g(n) = \underset{\forall\ child\ i}{MIN\ \{\ distance_i\ \} + 1} \qquad (1)$$

Where $\underset{\forall \text{ child } i}{\{\ distance_i\ \}}$ is the set of distances from the keywords to all children of node $n$. Since the calculation of $g(n)$ starts from the user profile and query keywords, the initial value of $g(n)$ for each user profile keyword and query keyword is 0. It is important to note that the nodes which have paths with smaller values of $g(n)$ are closer to root domains.

The distance from the root domains to the node $n$, $h(n)$, is calculated as follows:

$$h(n) = \underset{\forall\ parent\ j}{MIN\left\{distance_j\right\} + 1} \qquad (2)$$

Where $\underset{\forall\ parent\ j}{\left\{distance_j\right\}}$ is the set of distances from the root domains to all parents of node $n$. Note that the root domains which have path to the query keyword are only involved in this step and the rest domains are excluded. Since the calculation of $h(n)$ starts from the root domains, the initial value of $h(n)$ for each root domain is 0. It is also important to note that the nodes which have paths with smaller values of $h(n)$ are closer to root domains.

A* algorithm is defined on the basis of (1) and (2) as follows:

$$f(n) = \ g(n) + h(n) \qquad (3)$$

Where $f(n)$ is the evaluation function value of node $n$ that maintains its distance from the keywords to the root domains, $g(n)$ is the distance from the keywords to node $n$; and $h(n)$ is the distance from the root domains to node $n$. The value of $f(n)$ is computed based on the calculations of $g(n)$ and $h(n)$. There could be many paths from one keyword to one root domain in the ontology base. Each path may have different evaluation function value. However, this approach emphasizes on the *nearest* (smallest) of these values. Equation (4) is used to compute the *nearest evaluation function value* from each keyword to each root domain and the results are represented in a tabular form.

$$Nf(K_i)_{D_j} = \underset{\forall\, ancestor\ a}{MIN} \{f((K_i)_a \rightarrow D_j\}, \quad \forall i \forall j \tag{4}$$

Where $K_i$ is a keyword s.t $K_i \in \{K_1, K_2, \ldots, K_l\}$ and $D_j$ is a root domain s.t $D_j \in \{D_1, D_2, \ldots, D_m\}$. $MIN\ \{f((K_i)_a \rightarrow D_j)\}$ is the smallest value of all evaluation function values that are ancestors of $K_i$ $((K_i)_a, \forall\ ancestor\ a)$ and in the path from $K_i$ to $D_j$. This smallest value is assigned to $Nf(K_i)_{D_j}$ as the nearest heuristic evaluation function value from $K_i$ to $D_j$. Equation (4) repetitively computes the nearest values from all keywords corresponding to all root domains.

Frequencies of the user profile keywords are another source of heuristic information that help to find the closest *horizontal* root domain of the ontology base for a given query. The keyword with higher frequency has more weight than the one with lower frequency. To involve this information in the process, each $Nf(K_i)_{D_j}$ resulted from (4) is multiplied by the frequency of its corresponding user profile keyword ($K_i$) as given in (5).

$$(Nf(K_i)_{D_j})_{freq} = Nf(K_i)_{D_j} \times (K_i)_{freq} \tag{5}$$

Where $(Nf(K_i)_{D_j})_{freq}$ is the nearest value from $K_i$ to $D_j$ taking into consideration the frequency value of $K_i$, $Nf(K_i)_{D_j}$ is the nearest value from $K_i$ to $D_j$, and $(K_i)_{freq}$ is the frequency of user profile keyword $K_i$. The nearest values that are generated by (5) are then represented in a tabular form as depicted in Table I.

The frequency values given in Table I are assumed values for the purpose of clarification. The root domains $D_7$ and $D_8$ are included in Table I but they are excluded from Table II because they have no nearest evaluation function values (i.e. they have no paths to the user profile keywords). In addition, the query keyword 'CELL' which is included in Table I is also excluded from Table II because it is assumed that it is being inquired for the first time and it is not included in the user profile and hence it's $(Nf('CELL')_{D_j})_{freq}$ is 0, for all $D_j$, since its frequency value is 0.

TABLE I. THE NEAREST HEURISTIC EVALUATION FUNCTION VALUES OF KEYWORDS CORRESPONDING TO ONTOLOGY BASE ROOT DOMAINS.

| **keywords** / **root domains** | $K_1$ | $K_2$ | $K_3$ | $K_4$ | $K_5$ | $K_6$ | $K_7$ |
|---|---|---|---|---|---|---|---|
| | (mp3 file)<br>freq. 2 | (mobile phone)<br>freq. 1 | (SQL server)<br>freq. 2 | (CELL)<br>freq. 0 | (inheritance)<br>freq. 1 | (client- server)<br>freq. 3 | (organ)<br>freq. 1 |
| | $(Nf(K_1)_{D_j})_{freq}$ | $(Nf(K_2)_{D_j})_{freq}$ | $(Nf(K_3)_{D_j})_{freq}$ | $(Nf(K_4)_{D_j})_{freq}$ | $(Nf(K_5)_{D_j})_{freq}$ | $(Nf(K_6)_{D_j})_{freq}$ | $(Nf(K_7)_{D_j})_{freq}$ |
| $\boldsymbol{D_1}$(Entertainment) | 4 | | | – | | | |
| $\boldsymbol{D_2}$(Tele-Comm.) | 6 | 3 | | – | | | |
| $\boldsymbol{D_3}$(Computing) | 6 | | 4 | – | 3 | 9 | |
| $\boldsymbol{D_4}$(Electricity) | | 3 | | – | | | |
| $\boldsymbol{D_5}$(Mathematics) | | | | – | | 6 | |
| $\boldsymbol{D_6}$(Biology) | | | | – | 2 | | 3 |
| $\boldsymbol{D_7}$(Politics) | – | – | – | – | – | – | – |
| $\boldsymbol{D_8}$(Sociology) | – | – | – | – | – | – | – |

Since the smallest values are the closest in meaning between keywords and root domains, they are better than the highest values. Therefore, a *normalization* process as given in (6) is needed to convert the smaller values to become more valued than higher values.

$$NNf(K_i)_{D_j} = \underset{\forall\, domains\ d}{MAX}(Nf(K_i)_d) - (Nf(K_i)_{D_j})_{freq} + \underset{\forall\, domains\ d}{MIN}(Nf(K_i)_d) \tag{6}$$

Where $NNf(K_i)_{D_j}$ is the *normalized nearest value* of the keyword $K_i$ for the domain $D_j$. The $MAX(Nf(K_i)_d)$ and $MIN(Nf(K_i)_d)$ are the maximum and minimum nearest values of the keyword $K_i$ associated with all the root domains, respectively. The $(Nf(K_i)_{D_j})_{freq}$ is the current nearest value of the keyword $K_i$ for the domain $D_j$. The normalized values are then maintained in Table II.

TABLE II. THE NORMALIZED NEAREST HEURISTIC EVALUATION FUNCTION VALUES OF KEYWORDS CORRESPONDING TO ONTOLOGY BASE ROOT DOMAINS.

| **keywords** / **root domains** | $NNf(K_1)_{D_j}$ | $NNf(K_2)_{D_j}$ | $NNf(K_3)_{D_j}$ | $NNf(K_5)_{D_j}$ | $NNf(K_6)_{D_j}$ | $NNf(K_7)_{D_j}$ |
|---|---|---|---|---|---|---|
| $D_1$(Entertainment) | 6 | | | | | |
| $D_2$(Tele-Comm.) | 4 | 3 | | | | |
| $D_3$(Computing) | 4 | | 4 | 2 | 6 | |
| $D_4$(Electricity) | | 3 | | | | |
| $D_5$(Mathematics) | | | | | 9 | |
| $D_6$(Biology) | | | | 3 | | 3 |

Each $NNf(K_i)_{D_j}$ in Table II may have different values for different root domains. Equation (7) is used to calculate the *degree of closeness* of *each* keyword $K_i$ for *each* connected root domain $D_j$ with respect to all root domains.

$$(NNf(K_i)_{D_j})_{close} = \frac{NNf(K_i)_{D_j}}{\sum_{j=1}^{m} NNf(K_i)_{D_j}} \ , \ \forall i \forall j \qquad (7)$$

Where $(NNf(K_i)_{D_j})_{close}$ is the degree of closeness of each $K_i$ for each connected $D_j$, $NNf(K_i)_{D_j}$ is the normalized nearest value from $K_i$ to $D_j$, and $\sum_{j=1}^{m} NNf\ (K_i)_{D_j}$ is the summation of all normalized nearest values from $K_i$ to all its associated root domains $D_j$ , s.t $i = \{1, 2, \dots, \text{l}\}$ and $j = \{1, 2, \dots, m\}$ . The results of processing (7) are maintained in Table III.

TABLE III. THE DEGREE OF CLOSENESS VALUES OF KEYWORDS CORRESPONDING TO ONTOLOGY BASE ROOT DOMAINSNS.

| **keywords**<br>**root domains** | $(NNf(K_1)_{D_j})_{close}$ | $(NNf(K_2)_{D_j})_{close}$ | $(NNf(K_3)_{D_j})_{close}$ | $(NNf(K_5)_{D_j})_{close}$ | $(NNf(K_6)_{D_j})_{close}$ | $(NNf(K_7)_{D_j})_{close}$ |
|---|---|---|---|---|---|---|
| $D_1$(Entertainment ) | 6/14 | | | | | |
| $D_2$(Tele-Comm.) | 4/14 | 3/6 | | | | |
| $D_3$(Computing) | 4/14 | | 4/4 | 2/5 | 6/15 | |
| $D_4$(Electricity) | | 3/6 | | | | |
| $D_5$(Mathematics) | | | | | 9/15 | |
| $D_6$(Biology) | | | | 3/5 | | 3/3 |

While (7) calculates the degree of closeness of each keyword $K_i$ for each connected root domain $D_j$, (8) given below, calculates the *degree of participation* of *each* keyword $K_i$ for *all* connected root domains with respect to *all* user profile keywords. This process is performed by dividing the total count number of non-zero normalized nearest values of a keyword $K_i$ for all root domains by the total count number of non-zero normalized nearest values of all keywords for all root domains. The result values are then maintained in Table IV.

$$\left(NNf(K_i)\right)_{part} = \frac{\left|NNf(K_i)_{D_j} \neq 0\right|_{\forall j}}{\left|NNf(K_i)_{D_j} \neq 0\right|_{\forall i \forall j}} \qquad (8)$$

$$i = \{1, \dots, l\}, \qquad j = \{1, \dots, m\}$$

Where $\left(NNf(K_i)\right)_{part}$ is the degree of participation of a keyword $K_i$ for all connected root domains, $\left|NNf(K_i)_{D_j} \neq 0\right|_{\forall j}$ counts the non-zero values of a keyword $K_i$ for all root domains, and $\left|NNf(K_i)_{D_j} \neq 0\right|_{\forall i \forall j}$ counts the non-zero values of all keywords for all root domains.

TABLE IV. THE DEGREE OF PARTICIPATION VALUES OF KEYWORDS IN ONTOLOGY BASE ROOT DOMAINS.

| **keywords**<br>**root domains** | $(NNf(K_1))_{part}$ | $(NNf(K_2))_{part}$ | $(NNf(K_3))_{part}$ | $(NNf(K_5))_{part}$ | $(NNf(K_6))_{part}$ | $(NNf(K_7))_{part}$ |
|---|---|---|---|---|---|---|
| $D_1$(Entertainment) | 1/11 | | | | | |
| $D_2$(Tele-Comm.) | 1/11 | 1/11 | | | | |
| $D_3$(Computing) | 1/11 | | 1/11 | 1/11 | 1/11 | |
| $D_4$(Electricity) | | 1/11 | | | | |
| $D_5$(Mathematics) | | | | | 1/11 | |
| $D_6$(Biology) | | | | 1/11 | | 1/11 |
| **For all root domains** | **3/11** | **2/11** | **1/11** | **2/11** | **2/11** | **1/11** |

Finally, the *confidence* values that represent the closest meaning of the user profile keywords corresponding to the root domains are computed by multiplying (7) by (8) as shown in (9) and the results are maintained in Table V.

$$(NNf(K_i)_{D_j})_{conf} = (NNf(K_i)_{D_j})_{close} \times (NNf(K_i))_{part} \qquad (9)$$

Where $(NNf(K_i)_{D_j})_{conf}$ is the confidence value of the keyword $K_i$ corresponding to the root domain $D_j$ . The confidence values of all keywords for every root domain are summed up and the root domain that has higher value is the closest to the meaning of the query keyword. Table V clearly shows that 'Computing' is the closest root domain to the query keyword 'CELL' since its confidence value (0.3118) is the highest.

TABLE V. CONFIDENCE VALUES OF THE KEYWORDS CORRESPONDING TO THE ONTOLOGY BASE ROOT DOMAINS.

| **keywords**<br>**root domains** | $\left(NNf(K_1)_{D_j}\right)_{conf}$ | $\left(NNf(K_2)_{D_j}\right)_{conf}$ | $\left(NNf(K_3)_{D_j}\right)_{conf}$ | $\left(NNf(K_5)_{D_j}\right)_{conf}$ | $\left(NNf(K_6)_{D_j}\right)_{conf}$ | $\left(NNf(K_7)_{D_j}\right)_{conf}$ | **sum of confidence values for each root domain** |
|---|---|---|---|---|---|---|---|
| $D_1$(Entertainment) | 0.116 | | | | | | **0.116** |
| $D_2$(Tele-Comm.) | 0.077 | 0.09 | | | | | **0.167** |
| $D_3$(Computing) | 0.077 | | 0.09 | 0.07 | 0.07 | | **0.311** |
| $D_4$(Electricity) | | 0.09 | | | | | **0.09** |
| $D_5$(Mathematics) | | | | | 0.10 | | **0.108** |
| $D_6$(Biology) | | | | 0.10 | | 0.09 | **0.198** |

Once the closest root domain (e.g. 'Computing') is determined, the process turns to search that domain *vertically* to identify the closest *concept* to the meaning of the query keyword (e.g. 'CELL'). For this purpose, KNN algorithm is employed. Fig. 2 shows some *concepts* that could indicate the meaning of the query keyword 'CELL' under 'Computing' domain including 'band', 'microprocessor', 'table', and 'flash-memory'. Calculating similarity between the query 'CELL' and each of these *concepts* might be beneficial to find the closest *concept* to the meaning of that query. The similarity for K=2 is calculated by using Euclidean distance as follows:

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \qquad (10)$$

Where $d(p,q)$ is the distance between the *concept* p and the query q.

The KNN is an effective classifier method [31] used to find out the distance (closeness) between the query 'CELL' and its direct upper *concepts*, where K is determined by the heuristics or features of these *concepts*. Two features are identified to calculate the distance: first, the number of root domains connected to the *concept* and second, the number of user profile keywords connected to the *concept*. Therefore, K here equals to 2. Table VI shows the direct upper *concepts* of the query 'CELL' that are connected to the 'Computing' domain (column 1) and their identified features (column 2 and column 3). In addition, Table VI shows the ranks (column 5) of the *concepts* based on their distances (column 4) from the query.

TABLE VI. THE FEATURES OF CONCEPTS AND THEIR RANKS BASED ON DISTANCES OF CONCEPTS FROM THE QUERY

| **The concept** | **# Root domains connected to the concept** | **# User profile keywords connected to the concept** | **Distance** | **Ran k** |
|---|---|---|---|---|
| Band | 2 | 1 | $\sqrt{5}$=2.2 36 | 2 |
| microprocess or | 1 | 0 | $\sqrt{9}$=3 | 3 |
| Table | 1 | 2 | $\sqrt{1}$=1 | 1 |
| flash-memory | 1 | 0 | $\sqrt{9}$=3 | 3 |

The Euclidean distance of KNN emphasize that the smallest distance value is the closest *concept* to the meaning of the query. Table VI, clearly shows that the *concept* 'table' is the closest *concept* to the meaning of the query 'CELL'. Extending the query 'CELL' by adding the *concept* 'table' to the query would personalize the user search and therefore, improve the search effectiveness.

## IV. EXPERIMENTAL RESULTS AND EVALUATION

A semantic-based web search method is proposed. This method adopts profiles and uses an ontology base to personalize users search and improve the accuracy of their search results. Essentially, a conceptual hierarchy is used in the experiments as ontology base for providing a shared understanding of the searchable domains. It is significant in discovering the closest meaning of the query keywords. Query keywords entered by the users should be selected from the experimental concept hierarchy terms (i.e. concepts) since the proposed method based on the ontology base. Moreover, the method employs Google search engine in the evaluation process to show the effectiveness of sematic-based search approach over keyword-based search approach. Such evaluation needs users to enter queries and to judge the relevancy of returned hits. Furthermore, the method implicitly extracts the closest *concept* to the meaning of the query from the ontology base and appends it to the initial query, to form the expanded query. The expanded query forms the final query, which is entered in the Google search engine to retrieve the desired documents.

The experimental trials had been conducted to evaluate the effectiveness of the proposed method in retrieving the relevant information. The evaluation investigates the degree to which the stated objectives are achieved. Recall and precision are two evaluation measures identified for IR systems [32]. Where recall measures the ability of the system to present all the relevant items in the collection, and precision measures the ability of the system to present only those items from the collection that are relevant. The documents collection of Google search engine is used intensively in the experiment tests. Users enter their queries into Google's search text box and consequently, Google searches its collection and returns a list of hits to users. Using Google collection restricts the evaluation process to use cut-off/precision measure rather than recall/precision measure, because it cannot calculate the normal recall points since the number of relevant documents in Google collection is unknown. However, precision is defined formally as follows:

$$precision = \frac{Number\ of\ relevant\ items\ retrieved}{Total\ number\ of\ items\ retrieved} \times 100 \qquad (11)$$

Cut-off points are made in the experiments for the first 100 documents of the search engine hits. Precision values are calculated at cut-off (the first) 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100 documents. The user should count the number of documents that are relevant to his needs and divides this number by 10 to obtain the precision value of the first group (10 cut-off points). The last step is repeated for all other groups as well. The user should count the number of relevant documents and divides this number by 20 to obtain the precision value of the second group (20 cut-off points), and so on. Thereby, the cut-off points and precision are for a single query. However, to evaluate the retrieval method accurately, we run it for several distinct queries; and an average is used for the cut-off and precision values. It is important to note that the average here means the precision value of all distinct queries at the corresponding cut-off point. In order to find the precision value at all cut-off points, we take the rate of all of its averages.

This paper compares the effectiveness of the semantic-based search approach with the keyword-based search approach. In particular, it compares the effectiveness of the proposed semantic-based search approach with a well-known and standard text-based IRS such as Google search engine. The comparison is achieved by searching the web twice. The first search uses Google search engine without employing the proposed method in the search process and the overall average of results is computed. The second search uses Google search engine with employing the proposed method in the search process and the overall average of results is computed. The difference between these two rates shows how happened to be an improvement of one over the other.

The averages of precision values of all entered queries at cut-off (the first group of documents) are calculated for 10 groups of sizes ranging from 10 to 100 documents with increments of 10. This calculation is done for both the proposed profiling semantic-based search method and the text-based search method (Google) as shown in Table VII.

TABLE VII. THE PRECISION VALUES AT CUT-OFF POINTS FOR THE PROPOSED METHOD AND GOOGLE AND THEIR AVERAGES.

| Cut-off | The proposed method (Semantic-based search) | Google (Text-based search) |
|---|---|---|
| 10 | 0.87 | 0.47 |
| 20 | 0.84 | 0.45 |
| 30 | 0.82 | 0.42 |
| 40 | 0.81 | 0.41 |
| 50 | 0.77 | 0.40 |
| 60 | 0.75 | 0.38 |
| 70 | 0.71 | 0.36 |
| 80 | 0.69 | 0.35 |
| 90 | 0.66 | 0.35 |
| 100 | 0.63 | 0.33 |
| **Average** | **0.755** | **0.392** |

The average of precision values of all entered queries for each of the 10 groups employing the proposed method is calculated. Then, an identical process is repeated for Google without employing the proposed method. As shown in Table VII, the first column (cut-off) denotes the number of documents taken as cut-off points, the second column shows the average precision values at these cut-off points using Google employing the proposed method, and the third column (Google) shows the average precision values at these cut-off points using Google without employing the proposed method.

The average here means the precision value at the corresponding cut-off point. In order to find the precision value of the proposed semantic-based search method using the user profile at all cut-off points, we take the rate of all of its averages. The same process is repeated with Google. The rate (average) of the proposed method and the rate (average) of Google are calculated and presented in the last row of Table VII. The difference between these two rates shows the improvement of one over the other. Table VII shows that the profile-based semantic search method improves the search results about 36% over the text-based search method (Google).

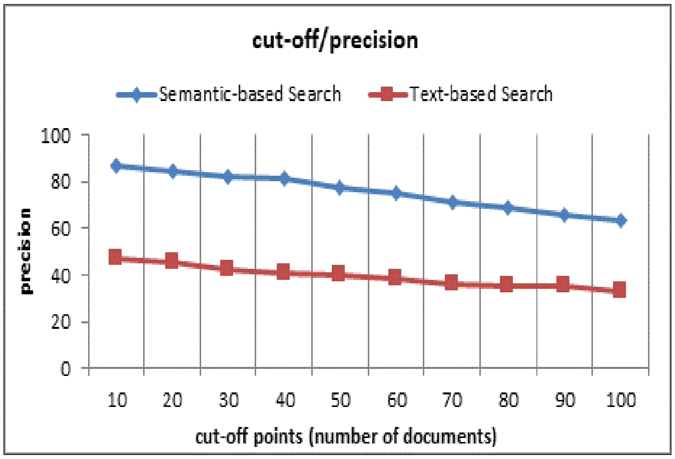

Fig. 3. Precision values at cut-off points for the proposed method and Google.

Fig. 3 is drawn based on Table VII. For each entered query, a cut-off/precision curve is drawn. These drawn curves are averaged to produce the final cut-off/precision shown in Fig. 3. The figure illustrates that the semantic-based IR is better than text-based IR. The figure shows 36% improvement on search results when employing the profiles and the heuristics in the retrieval process, especially when using ontologies for interpreting query terms, as recorded by the experiments. This aspect is promising to shift the web search engines from the text-based to the semantic-based information retrieval systems.

## V. CONCLUSION AND FUTURE WORK

Issuing a query to the web search system for retrieving relevant pages according to user preferences provides better result if the search method discover and extracts the *concept* from the ontology base that reflect the closest meaning of the query keyword. The proposed approach is significantly beneficial, especially when the query keyword found across several different *horizontal* root domains and under various different *concepts* of the same *vertical* root domain. The extracted *concept* is used to expand the query for personalizing the user's search. Incorporating the user profile and ontology base of the semantic web into the search process was the base of the proposed semantic-based search method. The heuristics and the KNN algorithm are applied to discover useful information in interpreting the query keywords. A profile-based personalized semantic search method shows a considerable improvement than text-based search in terms of search effectiveness, as recorded by the experiments. Despite the effectiveness and accuracy improvement of this approach, it has two limitations. First, the proposed search method is developed for handling only one-keyword size queries. Secondly, the effectiveness of the proposed approach is evaluated based on a comparison made with Google as a text-based search method. To get more accurate result, a comparison must be made with another semantic-based search method adopting different approach. For future work, identifying additional implicit features or heuristics and engaging knowledge management to discover knowledge from the heuristics information and representing them in user profile, may further improve the search results.

## ACKNOWLEDGMENT

This work was sponsored and supported financially by a research grant from the deanship of scientific research at Prince Sattam bin Abdulaziz University "REF 2014/01/2006".

## REFERENCES

[1] S. Chawla, "Search Engines: Information Retrieval on the Web," Everyman's Science, Indian Science Congress Association, Kolkata, India, vol. LI, no. 4, http://sciencecongress.nic.in/pdf/e-book/oct-nov-16.pdf (2016a). Accessed 2 Sep. 2017.

[2] H. A. Abdeljaber, "A profile-based semantic search strategy using an ontology base," Ph.D. Thesis, Universiti Kebangsaan Malaysia (UKM), Malaysia, 2009.

[3] J. Jayanthi and S. Rathi, "Personalized web search methods – a complete review," Journal of Theoretical and Applied Information Technology, vol. 62 no. 3, 2014.

[4] N. Borse, S. Patil, N. Agrawal, and R. Pachlor, "Survey on A Personalized Ontology Model for Web Information Gathering," International Journal of Emerging Technology and Advanced Engineering (IJETAE), vol. 3, no. 12, 2013.

[5] M. R. Ghorab, D. Zhou, A. O'connor, V. Wade, "Personalised information retrieval: Survey and classification," User Modeling and User-Adapted Interaction, vol. 23, no. 4, pp. 381-443, 2013. doi:http://dx.doi.org/10.1007/s11257-012-9124-1.

[6] B. Thomas, J. P. Jose, "A survey on web search results personalization," Compusoft, vol 4, no. 3, pp. 1582, 2015.

[7] V. Salonen and H. Karjaluoto, "Web personalization: The state of the art and future avenues for research and practice," Telematics and Informatics, vol. 33, no. 4, pp. 1088-1104, 2016. doi:10.1016/j.tele.2016.03.004.

[8] S. Chawla, "Use of hybrid of fuzzy set and ACO for effective personalized web search," International Journal of Advanced Research in Computer Science, vol. 8, no. 5, 2017a.

[9] V. Shakya and A. Sonker, "An optimal ranking approach for cluster based of clicked URLs using firefly algorithm for efficient personalized web search," International Journal of Advanced Research in Computer Science, vol. 8, no. 5, 2017.

[10] S. Chawla, "Web page ranking using ant colony optimisation and genetic algorithm for effective information retrieval," International Journal of Swarm Intelligence, vol. 3, no. 1, pp. 58, 2017b. doi:10.1504/IJSI.2017.082397.

[11] S. Chawla, "Application of genetic algorithm and back propagation neural network for effective personalize web search-based on clustered query sessions," International Journal of Applied Evolutionary Computation (IJAEC), vol. 7, no. 1, pp. 33-49, 2016b. doi:10.4018/IJAEC.2016010103.

[12] B. Mianowska and N. T. Nguyen, "Tuning user profiles based on analyzing dynamic preference in document retrieval systems," Multimedia Tools and Applications, vol. 65, no. 1, pp. 93-118, (2013;2012;). doi:10.1007/s11042-012-1145-6.

[13] A. Nanda, R. Omanwar, and B. Deshpande, "Implicitly learning a user interest profile for personalization of web search using collaborative filtering," Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), vol. 2, IEEE Computer Society; pp 54-62, 2014. doi:10.1109/WI-IAT.2014.80.

[14] S. Karpagam and P. N. Nesarajan, "Personalized web search protection using greedy algorithm," International Journal of Research in Computing Science, Engineering & Technology (IJRCSTE), vol. 1, no. 2, pp. 1-10, 2016.

[15] A. Kacem, M. Boughanem, and R. Faiz, "Emphasizing temporal-based user profile modeling in the context of session search," In: SAC, ACM, pp. 925–930, 2017.

[16] U. Gajendragadkar and S. Joshi, "Context sensitive search string composition algorithm using user intention to handle ambiguous keywords," International Journal of Electrical and Computer Engineering, vol. 7, no. 1, pp. 432-450, 2017.

[17] J. B. Baviskar, "Personalised web search using user behaviour," International Journal of Advanced Research in Computer Science, vol. 6, no. 1, pp. 185-187, 2015.

[18] M. R. Ghorab, D. Zhou, S. Lawless, and V. Wade, "Multilingual user modeling for personalized re-ranking of multilingual web search results, in Herder," E., Yacef, K., Chen, L. and Weibelzahl, S. (Eds), CEUR Workshop Proceeding of UMAP 2012, Montreal, Springer, Berlin, pp. 1-4, 2012.

[19] B. Steichen, M. R. Ghorab, A. O'Connor, S. Lawless, and V. Wade, "Towards Personalized Multilingual Information Access - Exploring the Browsing and Search Behavior of Multilingual Users," Proceedings from Conference on User Modelling, Adaptation and Personalization (UMAP 2014), Aalborg, Denmark, pp. 435–446, 2014. doi:10.1007/978--3--319-08786--3_39.

[20] D. Zhou, S. Lawless, X. Wu, W. Zhao, and J. Liu, "A study of user profile representation for personalized cross-language information retrieval," Aslib Journal of Information Management, vol. 68, no. 4, pp. 448-477, 2016. doi:10.1108/AJIM-06-2015-0091.

[21] R. D. Virgilio, F. Guerra, and Y. Velegrakis, Semantic search over the web, (1. Aufl.;2012;1; ed.). DE: Springer-Verlag (2012;2014;).

[22] A. Göker and J. Davies, Semantic search, Chichester, UK: John Wiley & Sons, Ltd. pp. 179-213. ch9, 2009. doi:10.1002/9780470033647.

[23] H. Dong, F. K. Hussain, and E. Chang, "A survey in semantic search technologies," 2nd IEEE International Conference on Digital Ecosystems and Technologies, pp. 403-408, 2008. doi:10.1109/DEST.2008.4635202.

[24] P. Castells, M. Fernández, D. Vallet, P. Mylonas, and Y. Avrithis, "Self-tuning personalized information retrieval in an ontology-based framework," vol. 3762, pp. 977-986, Berlin, Heidelberg: Springer Berlin Heidelberg, 2005. doi:10.1007/11575863_119.

[25] K. Makvana, P. Jay, P. Shah, and A. Thakkar, "An approach to identify semantic relations between user's queries in text retrieval," Proceedings of the Second International Conference on information and communication technology for competitive strategies, ACM, pp 1-6, 2016. doi:10.1145/2905055.2905271.

[26] B. Sateli, F. Löffler, B. König-Ries, and R. Witte, "Semantic user profiles: learning scholars' competences by analyzing their publications," International Workshop on Semantic, Analytics, Visualization. Enhancing Scholarly Data. Lecture Notes in Computer Science, Springer, Cham, vol. 9792, pp. 113-130, 2016.

[27] Z. Xu, H. Chen, and J. Yu, "Generating personalized web search using semantic context," The Scientific World Journal, vol. 2015, pp. 1-10, 2015. doi:10.1155/2015/462782.

[28] K. M. Fouad, A. R. Khalifa, N. M. Nagdy, and H. M. Harb, "Web-based semantic and personalized information retrieval," International Journal of Computer Science Issues (IJCSI), vol. 9, no. 3, pp. 266-276, 2012.

[29] K. Makvana, P. Shah, and P. Shah, "A novel approach to personalize web search through user profiling and query reformulation," International Conference on Data Mining and Intelligent Computing (ICDMIC), IEEE, pp. 1-10, 2014. doi: 10.1109/ICDMIC.2014.6954221.

[30] P. Jay, P. Shah, K. Makvana, and P. Shah, "An approach to identify user interest by reranking personalize web," Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies, ACM, pp. 1-8, 2016. doi:10.1145/2905055.2905270.

[31] J. Gou et al, "A generalized mean distance-based k-nearest neighbor classifier," Expert Systems with Applications, vol. 115, pp. 356-372, 2019.

[32] R. Baeza-Yates and B. Ribeiro-Neto, Modern information retrieval, volume 463, ACM press, New York, 1999.