Personalization of Search Profile Using Ant Foraging Approach

Pattira Phinitkar and Peraphon Sophatsathit

Advanced Virtual and Intelligent Computing (AVIC) Research Center
Department of Mathematics, Faculty of Science
Chulalongkorn University, Bangkok 10330, Thailand
pattirap@gmail.com, peraphon.s@chula.ac.th

Abstract. This paper proposes a three-stage analysis of web navigation that yields search results being relevant to the user's interests and preferences. The approach is inspired by ant foraging behavior. The first stage focuses on a user's profile based on the web pages visited to be proportional with the amount of pheromone deposited by the ants. The level of pheromone denotes scores of user's interest. The second stage classifies the user's profile data. The final stage personalizes the search results based on the user's profile. Search results, which may span across a wide range of document archives and scatter over the Internet, will then be logically grouped by category for easy access and meaningful use. The experiments mainly consider the search results with reference to the user's profile in presenting the most relevant information to the user.

Keywords: search profile, search personalization, word similarity, Ant Colony Foraging.

1 Introduction

As the volume of information grows rapidly on the Internet, more investment on web search engines follows suit. Unfortunately, the number of search results usually turns out to be unsatisfactory. Often times, users must go through a long listing of documents to look for a few relevant ones. Determining the relevance of search results mostly relies on the user's own background and search context, e.g., the search result of "palm" yields more information on the Personal Digital Assistant (PDA) than a palm tree. Bearing such issues of user's interests and preferences (hereafter will be referred to as "user's profile") in mind, this paper provides a straightforward approach to build a user's profile based on interest scores which are derived from pheromone deposited by the ants. This profile reflects the user's behavior as pheromone being accumulated or evaporated. In the mean time, the content keywords of user's profile are classified in a reference concept hierarchy. A set of experiments was devised to carry out personalized search according to the proposed approach, yielding satisfactorily results.

D. Taniar et al. (Eds.): ICCSA 2010, Part IV, LNCS 6019, pp. 209–224, 2010.
© Springer-Verlag Berlin Heidelberg 2010

The paper is organized as follows. Section 2 briefly recounts some related work. The proposed approach is procedurally elucidated in Section 3, along with the supporting experiments in Section 4. Some final thoughts and future challenges are given in Section 5.

2 Related Work

There have been numerous improvement challenges to personalize web mechanisms offered by search engines over the last few years. This is because people are naturally unwilling to spend extra efforts on specifying their intention. One approach to personalization is to have the users describe their own interests. Other approaches to automatic characterization use the user's profile derived from their interests and preferences. All pertinent information being extracted is then used to create a personal profile for setting designated queries on web page. Jaime, et al [1] explores rich models of user's interests built from both search-related and personal information about the user. We focus on personalization search results without having to be over expressive on user's interest.

One essential step in the improvement procedure is classification of personalized information. The classification process is to organize disordered information in a methodical arrangement. Emily, et al [2] proposes a method to classify queries by intent (CQI). Knowing the type of query intent greatly affects the returning relevant search results.

A prevalent shortcoming of search process is that many personal search approaches often return search results by focusing on the user's interests and preferences rather than on user's queries. The underlying principle utilizes personal profiles in the search context to re-rank the search results for furnishing more relevant outcomes to the users. The search process and ranking of relevant documents are achieved by contextual search based on the user's profile. Amiya, et al [3] presents a system architecture that would work as a subordinate to a normal search engine by taking its results, calculating the relevance of these results with respect to the user's profile, and displaying the results along with its relevance to the user. Ahu, et al [4] demonstrates that re-ranking the search results based on user's interest is effective in presenting the most relevant results to the user.

3 Proposed Approach

The focus of this work is to provide the most relevant search results to a user by personalization search according to his profile. Our approach will adopt ant colony foraging algorithm to perform both gathering the user's interests and updating the user's profile.

The proposed approach consists of three main steps, namely, building the user's profile, classifying the user's profile data, and personalization the search results by means of the user's profile as illustrated in Figure 1. The analysis will proceed in two systematic processes, namely, coarse-grained overview and fine-grained scrutiny.

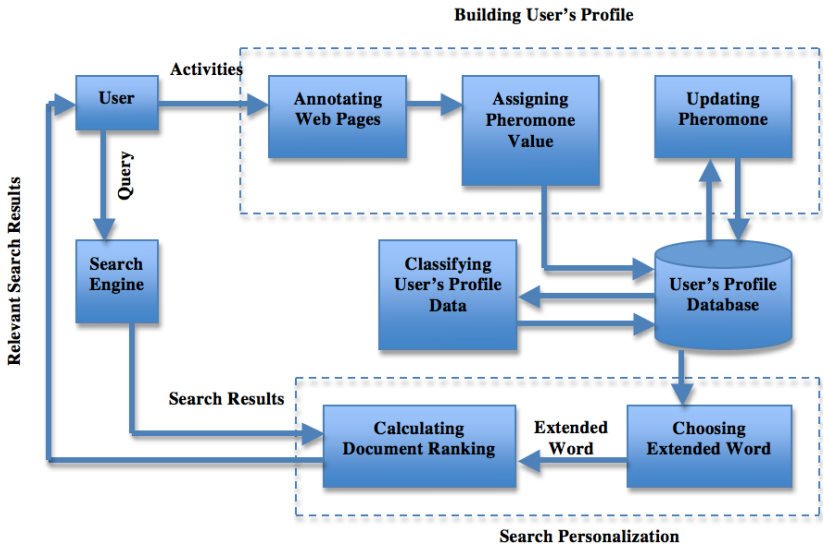

Fig. 1. Architecture of the proposed approach

3.1 Building User's Profile

A user's profile represents the user's interests and preferences in order to deduce user's intention for queries. In our approach, a user's profile consists of a set of categories, each of which encompasses a set of elements and its corresponding weight. Each category denotes the user's interest in that category. The weight or score of user's interest in an element represents the significance of that element with respect to the category.

The fundamental principle of user's profile creation is inspired by the nature of ant colony foraging behavior [5]. The ant leaves pheromone chemical as a communication means. Ants will typically choose to lay pheromone depending on the quality and quantity of food found at a source when foraging for food. Consequently, a strong pheromone path is created as soon as a profitably high value of food source is found. In general, the stronger value of pheromone it produces, the less pheromone evaporates. As food sources become depleted, it is to the advantage of the colony for the pheromone to evaporate over time. This eliminates the possibility of ants following a strong pheromone trail to a food source that has already been diminished. It follows then that, in a situation where there is a certain probability of food randomly appearing, the colony could find new food sources.

By the same token, the user's profile is changing periodically. Thus, to solve the problem by optimizing of the pheromone level of concentration, a pheromone update is required to keep the user's profile up to date at all time. Bearing this notion in mind, assuming that web page destinations portray the food sources, the system adds the interest scores to the user's profile as pheromone gets deposited when the user visits the destination web pages. As such, the amount of pheromone being deposited depends on the user's interest of the destination web page, particularly for the pages that are located deep under the home page links of interest.

The creation of user's profile serves as a coarse-grained process that exploits pheromone deposit technique to arrive at a user's interest summary and efficient search results.

3.1.1 Annotating Web Page

When a user visits the destination web page, information must be extracted to annotate web page contents. Using full text documents to annotate web page takes considerable amount of time. Hence, our approach extracts information from parts of HTML document. Since a web page is a semi-structured document, annotation of web page contents is determined by the structure of the web pages. To confine the size of search space, the proposed approach considers only three kinds of tags and attributes from HTML pages, namely, URL, tag<title>, and tag<meta name="description">. URL is selected because not only navigation path can be traced from the URL, but also web page contents are usually related to their source. The URL strings accurately describe what is contained in each folder by means of descriptive words to enhance intuitive meaning to the user as shown in Figure 2. The tag<title> and tag<meta name="description"> provide descriptions of the web pages. The tag<title> gives a brief definition of the web page, whereas the tag <meta name="description"> provides a concise explanation of the content of web page. The proposed approach looks for the most redundant words to apply annotation of the page. For example, if the most redundant word is football, one criterion on page annotation is to choose the web page annotated with football. Another criterion is to employ page type classification that is determined by the pheromone count. The procedure will be described in subsequent sections. At any rate, a systematic procedure for participating candidate word consideration that is produced by tags and attributes is the root word, disregarding all derivatives thereof.

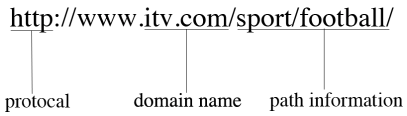

Fig. 2. A dichotomy of a URL

In this approach, URL, tag<title>, and tag<meta name="description"> are segmented into tokens. The procedure breaks non-alphanumeric characters, conjunction words, and stop words to create smaller tokens, and applies Porter stemming algorithm [6] to transform each token to a root word. An indicative statistics, that is, word density is computed from these tokens to gather web page annotation statistics.

Word density can be computed from the equation:

$$\text{density} = ((\text{Nkr} * \text{Nwp}) / \text{Tkn}) * 100 . \quad (1)$$

where Nkr denotes word frequency, Nwp is number of word occurrences in a phrase, and Tkn is the total number of words.

The sample web page annotation is shown in Table 1. Each URL, tag, and meta tag are segmented into tokens and assigned weight derived from the previous density statistics as shown in Table 2.

Table 1. Sample web page with URL, tag, and META tag

URL	Tag<title>	Tag<meta name="description">
http://news.bbc.co.uk/sport2/hi/football/default.stm	BBC SPORT | Football	The latest BBC Football news plus live scores, fixtures, results, tables, video, audio, blogs and analysis for all major UK and international leagues.

Table 2. Weight and density of individual token

Tokens	Weight	Density
news	1,1	10.526
sport	1,1	10.526
hi	1	5.263
football	1,1,1	15.789
:	:	:
league	1	2.263

3.1.2 Assigning Pheromone Value

The analogy of quality of the food source that affects the amount of pheromone deposit gives rise to the pheromone value of user's interest in the web page. Typically, most users prefer a direct link to the web page that they are interest in. However, if they cannot find sufficient information required to reach the designated web page, they will surf through the web pages to find the desired information. Consequently, counting the path along the URL measures the user's interest in the web page. The frequency of visiting the same type of web pages can also be used to evaluate user's concentration. These two factors constitute the pheromone value of the designated web page.

One essential ant foraging behavior occurs when ants found a good quality food source. They will congregate at the food source to acquire as much food as they can. Thus, heavy pheromone will be deposited along the path to food source. By this analogy, we also consider the selected web pages acquired from search results as an additional factor of pheromone value calculation.

In creating a new user's profile, the above information so obtained is inadequate to infer what the user real interests are. Fortunately, the selected web pages can make up "short-term" user's interests. By assigning higher weight to increase the amount of pheromone deposit, the level of information in the user's profile will quickly become steady for inference of user's interest.

As the user's interest diverts over a period of time, these short-term interest surges will gradually subside. The corresponding assigned weight will also decrease (or evaporate). This is called a "long-term" user's interest. Switching between short-term and long-term user's interest determined by the rate of pheromone evaporation, which will be further elaborated in next section. The rationale behind this observation is to

render a newly created user's profile reaching steady state as soon as possible in proportional to the selected web pages (or pheromone deposit). When the surge subsides so does pheromone deposit amount as they evaporate.

Table 3 shows sample frequencies of visiting web pages. Each node represents the web page annotated from the previous step. The amount of pheromone deposit denotes the frequency of visit, which includes the same type of web page. The most visited node is technology node as shown in the table, reflecting higher user's interest in technology topic than the rest of the topics under investigation.

Table 3. Pheromone deposition of visit

No.	Node	Amount of pheromone deposit
1	Technology	37
2	Football	29
3	System Analysis	20
4	Car	8
5	Camera	5
6	Game	3

3.1.3 Updating Pheromone

As user's interests and preferences always change over time, the value of pheromone must be updated, preferably in real-time, to keep the user's profile up-to-date. According to ant colony behavior, deposit and evaporation of pheromone must be proportionated. If the destination has abundant food source, many ants will go there and lay pheromone which results in strong pheromone deposit and lower evaporation rate along the path. On the other hand, for a low quantity of food source, the path will make a weak pheromone path having high evaporation rate because few ants will visit the area.

When the rate of pheromone evaporation for the node becomes 1, or the highest of the rate of pheromone evaporation, that node will be deleted from the user's profile. The fact is that the user has lost interest in that topic.

The formula for pheromone value update, or equivalently the user's interest score, can be determined as follows:

$$\tau_d = (1 - \rho)\, \tau_d \,. \tag{2}$$

where τ_d denotes the amount of pheromone on a given destination web page and ρ denotes the rate of pheromone evaporation. The equation of the rate of pheromone evaporation (ρ) is

$$\rho = 1 - (\tau_d / \Sigma\tau_d) \,. \tag{3}$$

In this paper, we adjust the pheromone differences to accommodate subsequent computations by the equation:

$$\tau_d = (1 - \rho)\, \tau_d^{\alpha} \,. \tag{4}$$

3.2 Classifying User's Profile Data

After annotating the web pages and collecting user's interest to be archived in the user's profile, some of this information may be similar by category, others may be different. Organizing this information for fine-grained scrutiny is a necessary requirement to keep track of the objects of similar properties, as well as their relationships if they exist. In other words, it is an issue of how this information should appropriately be classified. In so doing, performance of the personalization process will improve. The proposed approach employs WordNet [7] to establish a user's preference word-list, and Leacock-Chodorow measure [8] for semantic similarity in the classification process.

Leacock-Chodorow measure deals with semantic similarity by only considering the IS-A relation. To determine semantic similarity of two synsets, the shortest path between the two synsets in the taxonomy is determined and scaled by the depth of the taxonomy. The following formula computes semantic similarity:

$$\text{Sim}_{\text{LCH}}(a,b) = -\log(\text{length}(a,b) / (2 * D)) . \quad (5)$$

where length denotes the length of the shortest path between synset *a* and synset *b*, and D denotes the maximum depth of the taxonomy.

Leacock-Chodorow measure assumes a virtual top node dominating all nodes and will always return a value greater than zero, as long as the two synsets compared can be found in WordNet. Leacock-Chodorow measure gives a score of 3.583 for maximum similarity that is the similarity of a concept and itself.

Table 4. Similarity value of word-list from user's profile

	football	cruise	tour	tennis	travel	resort	sport	game	seafood
football	-	1.1239	1.2040	1.8971	1.3863	1.7430	*2.5903*	2.3026	1.1239
cruise	1.1239	-	1.9459	1.1239	*2.6391*	2.2336	1.7228	1.3863	0.8557
tour	1.2040	1.9459	-	1.2040	*2.6391*	1.5404	1.5404	1.6094	1.2910
tennis	1.8971	1.1239	1.2040	-	1.3863	1.7430	2.3026	2.3026	0.9808
travel	1.3863	*2.6391*	*2.6391*	1.3863	-	*2.6391*	1.9459	1.743	1.2040
resort	1.7430	2.2336	1.5404	1.7430	*2.6391*	-	2.0794	2.3026	1.3863
sport	*2.5903*	1.7228	1.5404	*2.3026*	1.9459	2.0794	-	*2.5903*	1.6094
game	2.3026	1.3863	1.6094	2.0794	1.7430	2.3026	*2.5903*	-	*2.3026*
seafood	1.1239	0.8557	1.2910	0.9808	1.2040	1.3863	1.6094	2.3026	-

The first step of classifying the user's profile is to compute similarity value from Equation (5) by setting up a word-list matrix created from the user's profile. A word-pair similarity value so computed indicates the closeness between the designated word-pair. The results are shown in Table 4.

The second step creates a bipartite graph designating the classification of word relations that build from the most similarity value of each designated word-pair. Each designated word-pair, which is selected to build a bipartite graph, is called word-list. The upper hierarchy represents category name which is derived from the nodes having common characteristics. The lower hierarchy of category name represents the corresponding elements.

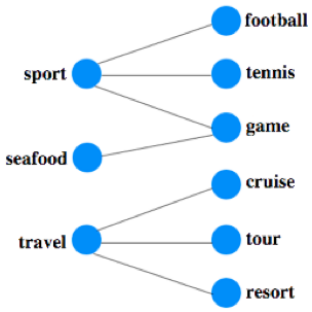

Fig. 3. An example of a bipartite graph with category name and its elements

Figure 3 depicts a sample bipartite graph of category name and its elements. To check whether an element belongs to the right category, the relationship between category name and its element must exist. If any element does not have a word-list relation, the element does not belong to its category and will be removed from the category and placed on unclassified category as shown in Table 5. For example, in Figure 3, sport– football, sport – tennis, sport – game, and game – seafood are element-category pairs. It appears that sport – game – seafood are related. However, sport – seafood does not contain in word-list relationship pair, thus seafood does not map to sport category.

Table 5. Category and its element

Category name	Element
sport	football, tennis, game
travel	cruise, tour, resort
-	seafood

To test the proposed approach classification capability, we compared our directory with Yahoo [9] and Google directories [10]. The results are close to both Yahoo directory and Google directory searches as shown in Table 6a and 6b, but are smaller and less complex than those of Yahoo and Google. Our directory contains all relevant category names and their elements that are easy to comprehend.

Table 6a. Sport category comparison of Yahoo, Google, and our directory

Sport directory		
Yahoo directory	Google directory	Our directory
recreation> sport>football	sport> football	sport> football
recreation> sport>tennis	sport> tennis	sport> tennis
recreation> game	game	sport> game

Table 6b. Travel category comparison of Yahoo, Google, and our directory

Travel directory		
Yahoo directory	Google directory	Our directory
recreation> travel>cruise	recreation> travel>specialty travel>cruise	travel> cruise
recreation> travel>tour	recreation> travel>tour	travel> tour
recreation> travel>resort	recreation> travel>loading> resort	travel> resort

3.3 Search Personalization

A typical search query often ends up with plentiful results that contain few relevant ones. To reduce unwanted search results, the above user's profile classification can be exploited to establish a search personalization mechanism. Search personalization is based on user's interest subjects (described by single word) having the highest amount of pheromone deposit. By reordering the pheromone deposit, all subjects (words) can be arranged according to their relevance to suit the user's personal preference.

The proposed approach supports word query between nouns, verbs, adjectives, and adverbs. Word query can be a word or collocation of words in the form of a sequence of words that go concurrently for a specific meaning such as "system analysis and design". However, the proposed approach does not support sentence query. One may contend that the more query words used, the clearer the (meaning of) query. Nevertheless, most users do not like to enter too many words just to look for a piece of information, typically about 3 words [13]. As such, adding one or two "key" words to form an extended word (to be described subsequently) entails a keyword search approach that yields high performance and useful results. This is because the extended word will help narrow down search theme which enables the search engine to recognize the user's interests and preferences.

Search personalization is then carried out in two steps. The first step is to choose an extended word based on the user's profile to make a new query which is more relevant to the user. The second step is to rank the search results obtained from the first step. The results are sorted in descending order. The procedural details are described below.

3.3.1 Choosing Extended Word for Making a New Query Keyword

To analyze if a word in the user's profile can be used as an extended word, all words are first classified and formed a bipartite graph. Each node of the bipartite graph is assigned a weight which is the pheromone deposit of user's interest. All words in the bipartite graph hierarchy form an extended word list to match the input query search results, irrespective of individual word position. This is the first step of the search personalization process. Figure 4 illustrates a user's profile classification from the bipartite graph and the corresponding pheromone value. There are three matching scenarios to consider:

1. Matching category name with the query. The query matches a category name having the highest pheromone deposit. The element becomes the extended word.
2. Matching element with the query. The query matches an element that belongs to one or more categories. If the element belongs to one category, the element will match with its own category. Otherwise, the element will match with the category that has the highest pheromone deposit.
3. No matching word between the query and user's profile word-list. This scenario will choose a word having the highest pheromone deposit in the user's profile to be an extended word.

The above matching scenarios are exemplified by the following examples.

Scenario	Query word	Extended word
1	sport	sport, football
2	palm	palm, technology
3	news	news, technology

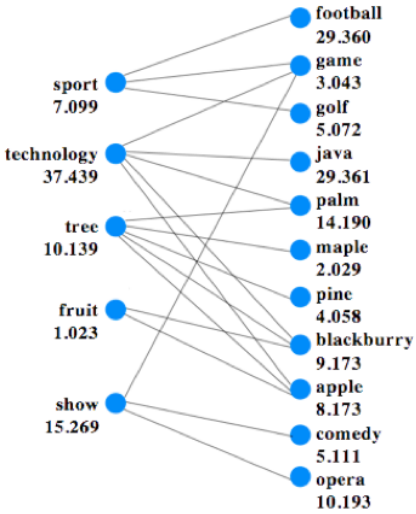

Fig. 4. A bipartite graph of category name and its elements with pheromone value

3.3.2 Calculating Document Ranking

The new extended query will yield search results that provide more relevant information to the users. Relevancy is obtained from re-ranking all words in the pertinent document, whereby the most likely related document will be retrieved.

To calculate ranking of each document, the cosine similarity between the extended query and the user's profile is computed. The cosine of two vectors is a measure of how similar two vectors will be on the (0,1) scale, where 1 means completely related (or similar) and 0 means completely unrelated (or dissimilar). The cosine similarity of two vectors *a1* and *a2* is defined as follows:

$$\mathrm{Sim}_{\cos}(a1, a2) = \cos(a1, a2) . \tag{6}$$

$$\cos(a1, a2) = \mathrm{dot}(a1, a2) / \|a1\| \, \|a2\| . \tag{7}$$

where *a1* denotes the extended query, *a2* denotes the term frequency of document, and dot(a1, a2) denotes the dot product of *a1* and *a2*. Term frequency, which measures how often an extended query is found in a document, is defined as follows:

$$\mathrm{tf}_{\mathrm{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}} . \tag{8}$$

where $n_{i,j}$ denotes the number of occurrences of the considered term (t_i) in document d_j and $\Sigma_k n_{k,j}$ denotes total occurrences of all terms in document d_j.

Table 7 depicts a new ranking of search results from the extended query "technology news" ordered by the most similarity to the least similarity.

Table 7. Using cosine similarity for re-ranking search results

Rank	Pervious rank	URL	Cosine similarity
1	2	http://edition.cnn.com/TECH/	0.99228
1	9	http://www.techweb.com/home	0.99228
2	7	http://news.zdnet.com/	0.99160
3	5	http://www.nytimes.com/pages/technology/index.html	0.85749
3	10	http://www.t3.com/	0.85749
4	6	http://www.businessweek.com/technology/	0.83957
5	8	http://www.physorg.com/technology-news/	0.80717
6	1	http://news.cnet.com/	0.79262
7	3	http://www.technewsworld.com/	0.70711
7	4	http://news.yahoo.com/technology	0.70711

4 Experiments

The experiments were carried out in different stages, namely, experimental setup, annotating web page, assigning, updating pheromone value, and re-ranking. The outcomes were measured by their precision and tested against Yahoo [11] and Yahoo Motif [12].

4.1 Experimental Setup

Four sets of extensive experiments were conducted based on the proposed procedures described. The first set, involving a basic word "Technology" will be elucidated in the sections that follow. The remaining three sets, i.e., Zoology, Botany, and Finance were carried out in the same manner.

4.2 Annotating Web Page

Since evaluation of the proposed approach is based primarily on the frequency of user's web access, we annotated the web pages which the users visited and found that 81% of the annotated results was similar to the page title. Figure 5 shows the density of each token from sample web pages. From Figure 5, football is the selected annotated web page because it has the highest density.

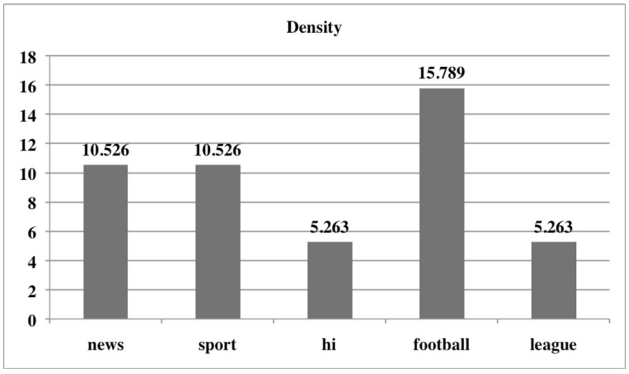

Fig. 5. Comparison of token density

4.3 Assigning and Updating Pheromone Value

The user's interests and preferences were determined from the amount of pheromone deposit, while the user's profile was kept up-to-date by the rate of pheromone evaporation. Thus, the updated pheromone value of each node in the bipartite graph would reflect the current degree of user's interest.

Table 8 summarizes the amount of pheromone deposit, rate of pheromone evaporation, and pheromone of each node. The higher the amount of pheromone deposit, the higher the degree of user's interest. The rate of pheromone evaporation is used to update the user's profile. Lower pheromone rate of evaporation reflects the intense of current user's interest, whilst higher pheromone rate of evaporation signifies the topics of interest currently being faded away. The pheromone represents the actual degree of user's interest.

Table 8. Pheromone deposit, rate of pheromone evaporation, and pheromone of each node

No.	Node	Amount of pheromone deposit	Rate of pheromone evaporation	Pheromone
1	Technology	37	0.561	37.439
2	Football	29	0.639	29.361
3	System Analysis	20	0.738	20.262
4	Car	8	0.887	8.113
5	Camera	5	0.928	5.072
6	Game	3	0.957	3.043

4.4 Re-ranking

Re-ranking process makes use of extended word notation based on user's interest and input query word. Using as few input keywords as possible, the user's profile is searched to retrieve the word having highest interest score to be combined with the input query word, or extended word in our context. Experiments were tested to compare simple query words with extended query words, and to compare Yahoo Motif with extended query words. For instance, the input query word "palm", along

with the highest interest scored word "technology", form an extended word "palm technology" for use in the search query. As a consequence, the search results yield different documents to be retrieved from a new list of URLs. This process is called re-ranking.

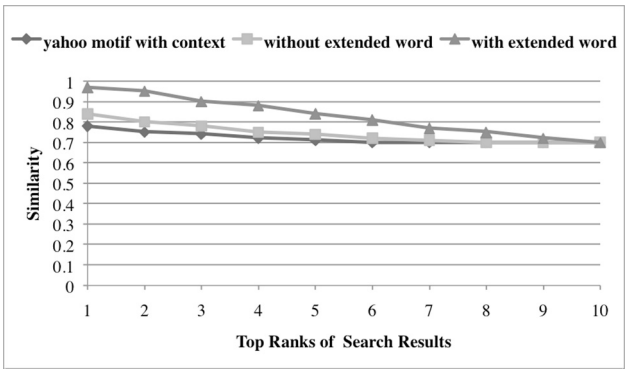

Fig. 6. Input (simple) query: palm, Yahoo motif query: palm, extended query: palm technology

Figure 6 shows the result comparison between query with simple query words, yahoo motif, and extended query words. The results of the three queries show that the last query yields more relevant documents to the user than the other two queries. A closer look at the cosine similarity value of the results reveals that the new ranking from extended word is closer to 1 than the ranking without extended word and Yahoo motif with context.

4.5 Experimental Results

Effectiveness of the proposed approach relates directly to the relevancy of retrieved results. The effectiveness of personalized search is measured by precision of the ability to retrieve top-ranked results that are mostly relevant to the user's interest. The precision is defined as follows:

$$\text{precision} = \frac{\text{number of relevant documents retrieved}}{\text{total number of documents retrieved}}. \quad (9)$$

For personalized search evaluation, we used four different user's profiles in each set of experiment. Some entries could appear in more than one profile as search went on. For example, "python" could fit either technology profile or zoology profile, "palm" could be in technology profile or botany profile, or "portfolio" could be in technology profile or finance profile. Table 9 compares the top-10 ranked of relevant search results based on user's profile. Table 10 shows the precision of the overall search results at different time. Those that are not relevant to the user's interest exhibit low precision values. However, as activities increase, search results improve since sufficient information is accumulated. This fact is depicted in Figure 7.

Table 9. Comparison of top-10 ranked relevant search results between Yahoo, Yahoo Motif, and our approach

Input query	Profile	Extended query	Amount of relevant results in top-10 ranked		
			Yahoo	Yahoo Motif	Our approach
palm	technology	palm technology	9	2	9
python	zoology	python snake	0	9	10
palm	botany	palm tree	0	2	8
portfolio	finance	portfolio finance	6	10	9

Table 10. The precision of user's profiles at different time

		Precision			
		Technology	Zoology	Botany	Finance
Time	T1	0.5	0.0	0.0	0.6
	T2	0.6	0.2	0.0	0.6
	T3	0.7	0.3	0.1	0.6
	T4	0.8	0.4	0.2	0.7
	T5	0.8	0.5	0.3	0.7
	T6	0.8	0.7	0.4	0.8
	T7	0.9	0.8	0.6	0.8
	T8	0.9	0.9	0.7	0.8
	T9	0.9	1.0	0.8	0.9
	T10	0.9	1.0	0.8	0.9

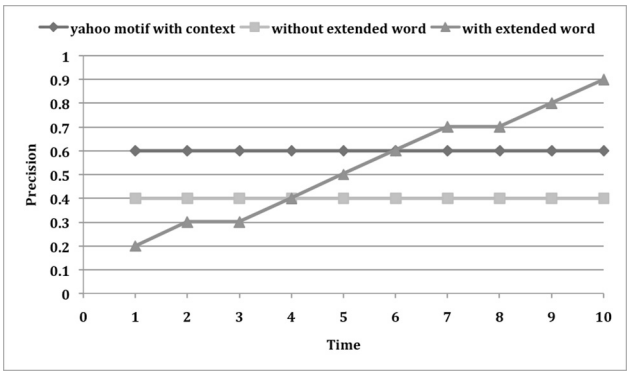

Fig. 7. Precision of personalized searches with extended word, general search without extended word and Yahoo Motif with context data

5 Conclusion and Future Work

We have presented the feasibility of personalized web search by means of an extended query that is automatically constructed from the most up-to-date user's

profile in accordance with the short-term and long-term interests. The short-term interest is induced by new events and vanishes quickly. On the contrary, the long-term interest generally reflects real user's interests. One way to realize this scheme is to set the short-term interest as the default search personalization and arrange the long-term interest as the secondary search. At a predetermined threshold period, short-term values are promoted to the long-term list. Older values in long-term list will eventually be discarded. Hence, search personalization will satisfy the user intent without having to resort to long strings of query.

The underlying principle was inspired by ant foraging behavior. When building a user's profile by extracting only the visited web pages and assigning a pheromone value as interest score, we found that the profiles converged to a stable set after approximately 350 visited web pages. On the other hand, if we incorporate the visited web pages and the selected web pages from the search results, the user profiles would converge to a stable set after approximately 100 visited web pages.

Our approach was tested on Yahoo and Yahoo Motif. The results yielded higher similarity score and precision score than those of Yahoo and Yahoo Motif, in particular, when comparisons were confined to the most relevant top-10 ranked results. However, a notable limitation of our approach is the performance which fell slightly as the profile contained less information, thereby the short-term compensation still fell short of what was anticipated.

There are no suitable personalization algorithms that fit all search queries. Different algorithms have different strengths and weaknesses. We will investigate in depth on extended words that exploit personalization algorithms to enhance the search results, whereby higher precision can be attained. Moreover, as the user gains more search experience, i.e., knowing how to select proper "search words", the precision score will increase. But this will take time to accumulate enough information before the steady state is reached. We envision that additional measures could be employed to shorten the profile accumulation cycle, namely, specificity, sensitivity, and accuracy, to see if the amount of information is sufficient for profile update, thereby user's experience will improve search profile personalization considerably.

References

1. Teevan, J., Dumais, T.S., Horvitz, E.: Personalizing search via automated analysis of interests and activities. In: Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2005), Salvador, Brazil, pp. 449–456 (2005)
2. Pitler, E., Church, K.: Using word-sense disambiguation methods to classify web queries by intent. In: Proceedings of the 2009 conference on empirical methods in natural language processing, Singapore, pp. 1428–1436 (2009)
3. Tripathy, K.A., Olivera, R.: UProRevs – user profile relevant results. In: Proceedings of the IEEE joint 10th international conference on information technology (ICIT 2007), pp. 271–276 (2007)
4. Sieg, A., Mobasher, B., Burke, R.: Web search personalization with ontological user profiles. In: Proceedings of the 16th ACM conference on information and knowledge management (CIKM 2007), Lisboa, Portugal, pp. 525–534 (2007)

5. Dorigo, M., Maniezzo, V., Colorni, A.: The ant system: optimization by a colony of cooperating agents. IEEE transactions on system, man, and cybernetics- part B 26(1), 29–41 (1996)
6. Porter, M.F.: An algorithm for suffix stripping. Program 14(3), 130–137 (1980)
7. WordNet, http://wordnet.princeton.edu/ (October 20, 2009)
8. Leacock, C., Chodorow, M.: Combining local context and WordNet similarity for word sense identification, ch. 11, pp. 265–283. MIT Press, Cambridge (1998)
9. Yahoo Directory, http://dir.yahoo.com/ (October 20, 2009)
10. Google Directory, http://directory.google.com/ (October 20, 2009)
11. Yahoo, http://www.yahoo.com/ (October 20, 2009)
12. Yahoo Motif, http://sandbox.yahoo.com/Motif/ (October 20, 2009)
13. Jansen, B.J., Spink, A., Bateman, J., Saracevic, T.: Real life information retrieval: a study of user queries on the web. ACM SIGIR Forum 32, 5–17 (1998)