An Immune System Approach To Personalize Search Results

Hamid Rastegari, Siti Mariyam Shamsuddin
Faculty of Computer Science and Information System
University Technology Malaysia
E-mail: hamidrst@gmail.com, mariyam@utm.my

Abstract— **Since dawn of the World Wide Web, the retrieval of relevant information has been a major problem for search engines. Current Web search engines use search algorithms to generate results that are suitable for discovering relevant pages to a query, but in doing so they do not consider the user who requested the query. A personalization system helps users to find interesting documents based on their preferences. The issue in this field is low accuracy in retrieved relevant information for particular user. This article proposed a novel technique to personalize the search results. We utilize artificial immune system (AIS) as a successful intelligent technique to find relevant pages in the search results based on user preferences. According to the obtain result, the proposed algorithm based on the AIS improved accuracy in retrieval of relevant information in the Web search results.**

Keywords- Personalization of Web Search, Artificial Immune System, Clonal Selection, Affinity Function, Search Result.

I. INTRODUCTION

World Wide Web (WWW) is the largest and most accessible source of information. Usually, Web structures are large and sophisticated and users often miss the goal of their inquiry, or receive ambiguous results when they try to navigate through them. Users seek a subject that they need information about it. Common search engines investigate Web and find relevant pages according to user query. Most of the times, they find a lot information about each subject through the Web. One of the issues is to discover helpful information from search results. Therefore, there are many studies on investigating the importance of personalization in Web search engine [1][2],[3].

Personalization of search results is defined as any action to find more relevant pages in search results for particular user or a set of users. The objective of a personalization system is to "provide information that users want or need exactly, without expecting from them to ask for it explicitly" [4]. Search personalization can reduce waste time to obtain needed information on the Web.

When the same query is submitted by different users, most search engines return the same results regardless of who submits the query. In general, users have different requests for their query. Gauch [5] mentioned that more than half of the documents returned by search engines are irrelevant information. There are several aspects to the problem [6]. First reason is the problem in synonyms and homonyms terms. Synonyms are words with same meaning and different spelt. Homonyms are words that are spelt the same but have different meanings. Without prior knowledge, there is no way for the search engine to predict user interest from simple text based queries. Secondly, search engines should be deterministic in that it should return the same set of documents to all users with the same query at a certain time. Therefore it is natural that search engines are not designed to adapt to personal preferences.

In the Google personalization system[1], Google makes a user account to keep the search history of the user. Also, it save the link of pages in search results that user clicked and viewed. In the next search, the visited pages in search results are placed on top of page rank. In the most situations this technique can help to user, but some researcher mentioned the best parameters to find user interested keywords are calculating the time spent and number of clicks on the Web pages [7][8],[9].

This paper is provided an overview of the topic of analyzing user behavior in Web search personalization. In Section II we classified the various techniques used to generate a personalized Web search in previous studies while in Section III we proposed a new approach with association of artificial immune system and Web search personalization. Experimental evaluation is explained in Section IV and finally, the paper concludes in Section V with a discussion on the current state and future direction.

II. RELATED WORK

This section reviews the most significant research in the search personalization area. We classified the related work to the three classes as follow:

A. Collaborative Filtering

One of the successful techniques in recommendation systems is Collaborative filtering (CF). Goldberg et al. proposed CF in the first time [10]. CF means that collaborate other people to recommend Web content by analyzing their transactions. The GroupLens research system [11], which recommends news for Usenet Website, was a first automated system using the k-nearest neighbor-based algorithm. In this algorithm, a subset of appropriate k users is chosen based on their similarity to the active user, and a weighted aggregate of their rating is used to generate predictions for the active user. Now, automated CF has been used at the E-commerce sites like Amazon.com, CDnow.com, YouTube.com and Movie-Finder.com, with considerable success.

B. Query Expansion

Submitted user queries to search engines are usually short and many times they are inadequate for an effective

[1] https://www.google.com/psearch

information retrieval on the Web. Query Expansion (QE) is used in the search engines to increase the quality of user search results. QE is assumed that users do not always submit the best terms in the search engine. QE helps Web searcher to better formulate the user query in order to enhance information retrieval. It exploits the best additional term to append to the user query before retrieval information. Cui et.al [12] proposed a QE technique where probabilistic correlations between query terms and document terms are used to expand given query. This approach is proven to outperform many other techniques which use only document space to expand the query than using both query space and document space. Palleti improved QE technique to provide personalized results to user by proposing a method to consider similar users in account to find the probability of a document term[13].

[14] proposed a QE model to personalize search result using ant foraging. They investigated the URL and titles of all visited Web documents to construct and classify a user profile for extended user queries. The user may open some Web pages and close those pages immediately as the information in the Web page is not of interest.

C. *Reranking search results*

Page et al. [15] proposed the first personalized Web search by modifying the global PageRank algorithm with the input of bookmarks or homepages of a user. Their work mainly focuses on global "importance" by taking advantage of the link structure of the Web. Bharat and Mihaila [16] suggested an approach called Hilltop, which generates a query-specific authority score by detecting and indexing pages that appear to be good experts for certain keywords, based on their links. Hilltop was designed to improve results for popular queries; however, query terms for which experts were not found will not be handled by the Hilltop algorithm. Haveliwala [17] used personalized PageRank scores to enable "topic sensitive" Web search. They concluded that the use of personalized PageRank scores can improve Web search, but the number of hub vectors (e.g., number of interesting Web pages used in a bookmark) used was limited to 16 due to the computational requirements. Kamvar [18] determined that PageRank could be computed for very large subgraphs of the Web on machines with limited main memory.

In the other way to personalize Web search based browsing history, Dou [19] proposed a framework that enables large-scale evaluation of personalized search. In their framework, user clicks on the Web pages is used to recording in the search engine logs as a simulate user experiences. In fact, after issue a query to search engine, the user usually checks documents in a result list from top to down. The user clicks one or more documents that look relevant and skips those documents that the user is not interested in. If a specific personalization method can re-rank relevant documents for a user higher in results list, the user would be more satisfied. Hence, it can be a judgment to evaluate search accuracy. Since click-through data can be done at low cost, it is possible to do large-scale evaluation.

III. Proposed Approach

When a user submits a query to a search engine through a Web browser, the search engine returns search results corresponding to the query. In these results, the user may select number Web pages according to request information. Also, the user may access more Web pages by following the hyperlinks on the selected Web pages and continue to browse. In the proposed approach, the system monitors user's browsing history and constructs a profile according to the user behavior. When the user submits a query in the next time, personalization system re-ranks the search results based on the user profile. The proposed approach has been shown in Figure 1.

A. *User Profile Construction*

The first step in the Web personalization process is extracting of the user preferences through the visited pages, which will be analyzed to provide useful information about the users' behavior. There are two main sources of data: data on the Web server side and data on the client side.

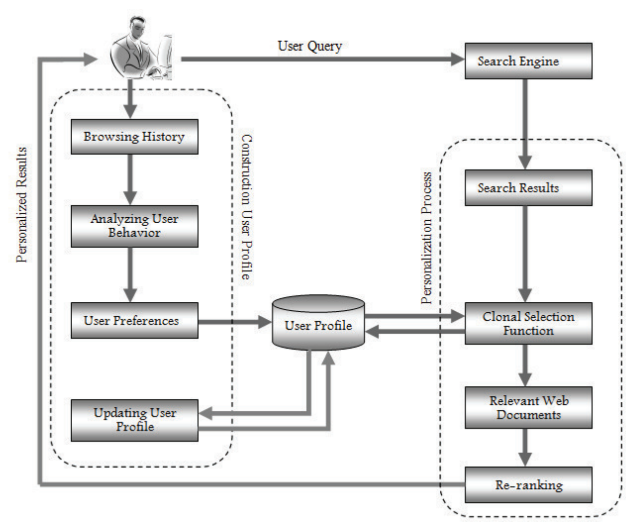

Figure 1. Personalized Web Search Process

Web server side collects data and stores in a server log files. This data consists primarily of various types of logs generated by the Web server. These logs record the Web pages accessed by the visitors of the site. Web mining tools use Web server log files as the main data source for discovering usage patterns. While, in the client side data are collected to construct user profile in implicit or explicit approaches. The simplest method for collection of data is to collect the user's preferences explicitly through forms, rating of search results, questioners, value elicitation or preference feedback [20].

Although, explicit information has high quality and prepares a reliable information about the users, but studies have shown that reluctance and lack of motivation on the user's part to provide information makes the explicit collection of sufficient data for the profile difficult [21].

Hence, user preferences can be collected implicitly by inference of the user's activities in the Web browsing history. This can be implemented using techniques like [22]:

- Monitoring the user's past search queries.
- Selecting links in search results by the user. Number of mouse click or spending time on a Web document confirms its relevance to the user.
- Monitoring the user's browsing patterns, like book marking, downloading, and printing a Web document.
- Background information about the user: The IP address of the user gives an idea of his geographical location.

Implicit and explicit data collection methods can be used in conjunction with one another, potentially giving the best of both the worlds [21]. For implicit data, time spent on each page can be calculated. The longer a user spent on a page, the likelier the user is interested in the page. If a page is not interesting, a user usually jumps to another page quickly. Experimental studies in [9] confirm this observation. However, a quick jump might be caused by the short length of the page; hence the user's interest might be more appropriately approximated by the time spent on a page normalized by the page's length.

Therefore, this research considers time spent and number of clicks on visited Web pages for profiling the user preferences. We consider term frequency and word density to extract important keywords in a Web document based on the equation (1).

$$\mathrm{TF}_{i,d} = \frac{n_{i,d}}{\sum_k n_{k,d}} \qquad (1)$$

Where $n_{i,d}$ donates the number of occurrences of the term i in the Web document d, and the denominator is the sum of number of occurrences of all terms in the Web document d. Equation (2) shows the user interest score for a keywords in the visited Web documents:

$$S_i = t_d \times C_d \times \mathrm{TF}_{i,d} \qquad (2)$$

Where S_i is the initial score for term i, and t_d is time spent and C_d is number of clicks on the Web document d. We apply this formula for keywords that have more frequency in the visited Web documents.

B. Personalized Results

The second phase of this approach is finding more relevant links in search results and re-ranking the results based on user preferences that were made in the first phase. This section is inspired by immune system. Artificial Immune System (AIS) refers to a group of computational intelligence techniques that are inspired by and attempt to emulate the information processing capabilities of the biological immune system [23]. AIS intend to solve complex computational or engineering problems and it is valuable to utilize AIS in the search personalization in order to discover relevant Web documents and improve accuracy in compare with pervious study.

In this section, a new algorithm based on clonal selection is introduced in detail. Clonal selection is an immune system hypothesis that has been proven to be an appropriate algorithm for Web mining and for finding similarity between Web pages [24]. Clonal selection consist of a set of immune cells, named antibodies and antigens, and a set of functions called affinity functions, which can clone, mutate and need a memory to save the best antibodies for the next antigens.

The proposed algorithm has been shown in the Figure 2. The first step in this algorithm is making an initial population for antibodies. We have selected n terms with high score in the user profile as the initial population. In the next step, we obtain k ULRs in search results. In this case, k equals 50, which means we evaluate 50 URLs in search results to be considered as antigens called the AG set. We want to calculate the similarity between antibodies (terms in user profile) and antigens (URLs in search results).

In this method, we activate the first antigen from AG and calculate term frequency for all terms in this antigen. The affinity between the antigen and all antibodies in P is calculated by affinity function. The affinity value determines similarity between a Web documents and the user profile. In the end of this approach, we sort all Web pages based on the similarity value in descending order and select top-ten pages to display to the user as a system result.

```
P = select n term of top-score keywords in the user profile
    % P is initial population
AG = k URLs in search results %set of Antigens
while (|AG|>0)
   Ag = the first URL in AG
   open this URL and calculate the term frequency for all terms
   for each antibody like Ab[i] in population P do
      aff[i] = affinity (Ab[i], Ag)
      if (aff[i]>= σ)
         clone and mutate the selected antibody by
            finding synonyms and homonyms the word
         add new words to population P
      end
      replace antibodies with low aff in P and other terms in
      the user profile
   end
   re-rank P
   calculate the similarity between P and Ag by
      summation affinities for all antibodies
   save the similarity
   remove Ag
   save best new antibodies in the user profile
end
```

Figure 2. Clonal Selection Algorithm for finding relevant Web pages

C. Affinity Function

The affinity function is a mathematical formulation for measurement of similarity between interesting keywords in the user profile and URLs in search result pages. In fact, this function is the rate of antibody for matching with antigen and removes it. In this case, this function helps to discover interest Web pages in the search results for particular user.

Equation (3) denotes the affinity function for calculating affinity value between search results and user profiles.

$$Affinity = \frac{\sum_{i=1}^{|P|} \delta_i . score_{Ab_i} . TF_{Ag,i}}{\sum_{j=1}^{|R|} \left(\sum_{i=1}^{|P|} \delta_i . Ab_i . TF_{Ag,i} \right)}$$

$$where\ \delta_i = \begin{cases} 1 & if\ Ab_i \in Ag \\ 0 & Otherwise \end{cases} \quad (3)$$

Where Ab_i is the ith term in initial population P and $TF_{Ag,i}$ is term frequency of ith term in the Web page named Ag. In this equation, if the keyword found in a Web page ($Ab_i \in Ag$) then $\delta_i = 1$ and we consider the interest score and term frequency of the keyword for calculation of the affinity value. This process repeats for all term in the user profile. Affinity function is used to find relevant and interesting Web documents that match the user profile.

IV. Experimental Results

For evaluation of the proposed approach, we consider the relevancy of retrieved results. The effectiveness of personalized search is measured by precision of relevancy in the retrieved documents [25]. Consequently, precision is defined as follows:

$$precision = \frac{relevant\ retrival\ documents}{total\ retrived\ documents} \quad (4)$$

The accuracy of the predicted interesting Web pages was evaluated by mean square error (MSE) function. The MSE is a statistical accuracy metric and evaluates proposed methods in terms of quality. MSE measures the average absolute between the first 10 URLs in Google search engine and the output of this work. MSE function defined in follow equation:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (R_i - R_i')^2 \quad (5)$$

Where n is number of pages that are investigated in this formula, R_i denotes user-rated relevancy degree for a Web document delivered by the search engine and R_i' is relevancy degree in the new re-rank list.

A. Dataset

One of the available datasets for query log files is AOL dataset. This collection consists of more than 21,500,000 Web queries collected from 65,000 users over three months. However, AOL dataset does not consider the time spent on each Web page. As we need to incorporate time spent for evaluation of the method, notably regarding time stamps on each click, we have calculated time spent approximately.

B. Experimental Evaluation

The accuracy of the predictions was evaluated by MSE function and the system performance was assessed by the relevancy of the retrieval information in these experiments. To compare of the proposed method for construction of the user model with pervious work, we selected research that has been done in (F. Li 2008) that used six users to construct the user model. Hence we also selected six users in the AOL dataset with more activities because profiling all users in this dataset is impossible and some of users just submit a query and we cannot analyze their behavior.

This part of dataset is divided into two partitions. The first partition includes submitted queries that user selected a link from first 10 search results in the first results page. This partition is used to construct users' profile. The second partition of dataset includes user activities that they selected a link in the second page or others pages in the search results. This partition is applied for testing the output. The expectation of the proposed approach is the selected link in the testing dataset appears to top-ten links after re-ranking process. Table 1 shows the number of visited pages by each user and the function values. The MSE helped us to compare this work with previous work. According to equation (5), R_i is position of the selected link by the user in the baseline results and R_i' is the same link position in the personalized results.

TABLE I. NUMBER OF VISITED PAGES BY 6 USERS

User ID	# of page	MSE
1337	119	0.00294
1326	89	0.00336
2178	105	0.00250
1334	74	0.00321
2421	131	0.00070
2465	128	0.00047

In the second phase of the experiments, we collect a user transaction in Web browser. All visited Web pages with time spent and number of clicks on each page is considered. According to the proposed method we analyzed the user behavior and made a user profile. In the next user query, we saved fifty URLs in the search results. The user profile is used for re-ranking the search results. In this experiment, all terms with a high degree are selected as an initial population in clonal selection, and the affinity value between the antibodies and each antigen are calculated. If the value is equal or greater than threshold, the antibody will be cloned and mutated. Cloning and mutation of antibodies have been done by finding synonyms and homonyms words. Ultimately, the affinity between all antibodies in P and each antigen will be calculated as the measurement of similarity between the user profiles and URLs in the search results. We re-rank the URLs in the search results based on the similarity and return to the user. Table 2 shows the result of the re-ranking process with the affinity value compared with the original rank by Google.

For evaluation of the second phase, we ask the user to rate the Google results and personalized results. In fact, we calculated the affinity of the 50 URLs in search results and re-rank them based on the user profile, selecting the ten-top URLs. Figure 3 denotes the relevancy of the new list in comparison with the Google results.

In order to using precision measurement, we assigned a user rating number to all links in the search results and then these numbers are normalized. Also, the search results are divided to 10 sections. Every section includes ten URLs and put in a different page. This segmentation has been done for both Google search results and personalized results in the proposed system. We applied the precision function and compare with the baseline result in the search engine. Figure 4 distinguished the effectiveness of the proposed approach in this chapter.

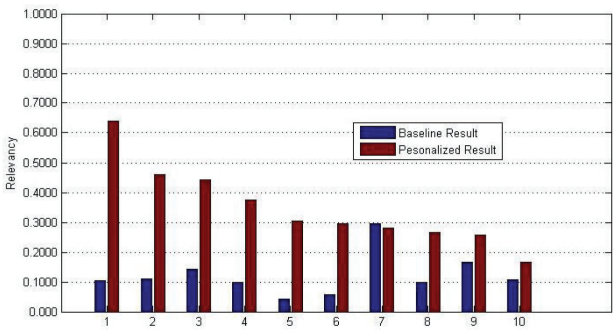

Figure 3. Relevancy of personalized result in comparison with Google

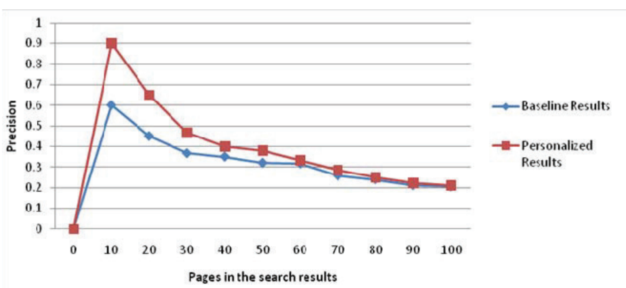

Figure 4. The effectiveness of the proposed approach for 100 results by precision measurement

As depicted in this figure, proposed technique gather more relevant Web pages of the baseline search results in the first page. Other pages in the results are almost similar with baseline.

V. CONCLUSION

In order to endow more relevant information in search results for particular user, this paper proposed new approach. We investigated Web browsing history to extract user's preferences and create a profile without involving the user. The user profile applied to adapting search results according to each user's information need. This work has utilized artificial immune system as a model to discover more relevant pages in the search results. AIS is a suitable inspiration for intelligent search personalization in order to mine the relevant documents in the search results based on user preferences. We defined affinity function for detection of user interesting page on Web.

There are some directions to improve in the future works. One direction is using other extra information about user such as location information. Also, the proposed method can be combined with other methods such as collaborative filtering and clustering methods. We can apply other method of artificial immune system for Web search personalization approach.

VI. ACKNOWLEDGMENT

This research has supported by Malaysian Ministry of Higher Education (MOHE) under the research grant, vote number: Q.J130000.7128.00H71. Also, the authors would like to thank of Soft Computing Research Group (SCRG) in Universiti Teknologi Malaysia (UTM) for the edits and comments on writing the paper.

TABLE II. PERSONALIZED RESULTS BY USING THE AFFINITY FUNCTION

Ranking List	URL	Affinity	Google Rank
1	www.prosoxi.gr/.../top-5-online-htaccess-mod-rewrite-rules-generator	0.6410	16
2	markmail.org/message/b557laxqne75qcnw	0.4609	37
3	stackoverflow.com/questions/3694740/rewrite-rule-generator	0.4417	39
4	ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1052348	0.3750	15
5	www.downv.com/Linux-software-download/iptables-rule-generator	0.3043	28
6	www.searchenginegenie.com/mod-rewrite-generator.php - United States	0.2963	7
7	www.ditii.com/2010/03/11/css3-cross-browser-rule-generator/	0.2804	19
8	www.downv.com/Linux-software-download/iptables-rule-generator	0.2667	21
9	vrt-sourcefire.blogspot.com/.../introduction-to-shared-object-rules.html	0.2581	14
10	www.ncbi.nlm.nih.gov/pubmed/21030735	0.1667	9

VII. REFERENCES

[1] K. Sugiyama, K. Hatano, M. Yoshikawa, Adaptive Web search based on user profile constructed without any effort from users, in: Proceedings of the 13th international conference on World Wide Web, ACM, New York, NY, USA, 2004.

[2] G. Jeh, J. Widom, Scaling personalized Web search, in: Proceedings of the 12th international conference on World Wide Web, ACM, Budapest, Hungary, 2003.

[3] J. Teevan, S.T. Dumais, E. Horvitz, 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, ACM, Salvador, Brazil, 2005.

[4] D.M. Maurice, S.A. Sarabjot, G.B. Alex, Personalization on the Net using Web mining: introduction, Commun. ACM, 43 (2000) 122-125.

[5] S. Gauch, J. Chaffee, A. Pretschner, Ontology-based personalized search and browsing, Web Intelligence and Agent Systems, 1 (2003) 219-234.

[6] F. Tanudjaja, L. Mui, Persona: A Contextualized and Personalized Web Search, in: Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS'02) - Volume 3, IEEE Computer Society, 2002.

[7] K. Jung, Modeling Web user interest with implicit indicators, in, Florida Institute of Technology, 2001.

[8] M. Claypool, P. Le, M. Wased, D. Brown, Implicit interest indicators, in: Proceedings of the 6th international conference on Intelligent user interfaces, ACM, Santa Fe, New Mexico, United States, 2001.

[9] M. Morita, Y. Shinoda, Information filtering based on user behavior analysis and best match text retrieval, in: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, Springer-Verlag New York, Inc., Dublin, Ireland, 1994.

[10] D. Goldberg, D. Nichols, B.M. Oki, D. Terry, Using collaborative filtering to weave an information tapestry, Commun. ACM, 35 (1992) 61-70.

[11] J.A. Konstan, B.N. Miller, D. Maltz, J.L. Herlocker, L.R. Gordon, J. Riedl, GroupLens: applying collaborative filtering to Usenet news, Commun. ACM, 40 (1997) 77-87.

[12] H. Cui, J.-R. Wen, J.-Y. Nie, W.-Y. Ma, Probabilistic query expansion using query logs, in: Proceedings of the 11th international conference on World Wide Web, ACM, Honolulu, Hawaii, USA, 2002, pp. 325-332.

[13] P. Palleti, H. Karnick, P. Mitra, Personalized Web Search Using Probabilistic Query Expansion, in: Web Intelligence and Intelligent Agent Technology Workshops, 2007 IEEE/WIC/ACM International Conferences on, 2007, pp. 83-86.

[14] P. Phinitkar, P. Sophatsathit, Personalization of Search Profile Using Ant Foraging Approach, in: D. Taniar, O. Gervasi, B. Murgante, E. Pardede, B. Apduhan (Eds.) Computational Science and Its Applications – ICCSA 2010, Springer Berlin / Heidelberg, 2010, pp. 209-224.

[15] L. Page, S. Brin, R. Motwani, The PageRank Citation Ranking: Bringing Order to the Web, in, en.scientificcommons.org, 1999.

[16] K. Bharat, G.A. Mihaila, When experts agree: using non-affiliated experts to rank popular topics, ACM Trans. Inf. Syst., 20 (2002) 47-58.

[17] T.H. Haveliwala, A. Gionis, D. Klein, P. Indyk, Evaluating strategies for similarity search on the Web, in: Proceedings of the 11th international conference on World Wide Web, ACM, Honolulu, Hawaii, USA, 2002.

[18] S. Kamvar, T. Haveliwala, G. Golub, Adaptive methods for the computation of PageRank, Linear Algebra and its Applications, 386 (2004) 51-65.

[19] Z. Dou, R. Song, J.R. Wen, Evaluating the Effectiveness of Personalized Web Search, IEEE Transactions on Knowledge and Data Engineering, (2009).

[20] S.S. Anand, B. Mobasher, Introduction to intelligent techniques for Web personalization, ACM Transactions on Internet Technology, 7 (2007).

[21] K. Keenoy, M. Levene, Web Search Personalization, in: Intelligent Techniques for Web Personalization, Lecture Notes in Computer Science, 2005.

[22] S. Gupta, B. Deshpande, S. Jain, M. Bedekar, M. Kazi, K. Kapoor, Personalization of Web search results based on user profiling, in: Proceedings - 1st International Conference on Emerging Trends in Engineering and Technology, ICETET 2008, 2008, pp. 1114-1119.

[23] D. Dasgupta, L.F. Nino, Immunological Computation, Theory and Application, CRC Prss, 2009.

[24] A. Secker, A.A. Freitas, J. Timmis, AISIID: An artificial immune system for interesting information discovery on the Web, Applied Soft Computing, 8 (2008) 885-905.

[25] R. Baeza Yates, B. Neto, Modern Information Retrieval, Addison-Wesley Longman Publishing Co., Inc., 1999.