A Personalized Graph-Based Document Ranking Model Using a Semantic User Profile

Mariam Daoud, Lynda Tamine, and Mohand Boughanem

IRIT, Paul Sabatier University,
118 Route Narbonne F-31062, Toulouse cedex 9, France
{daoud,tamine,bougha}@irit.fr

Abstract. The overload of the information available on the web, held with the diversity of the user information needs and the ambiguity of their queries have led the researchers to develop personalized search tools that return only documents that meet the user profile representing his main interests and needs. We present in this paper a personalized document ranking model based on an extended graph-based distance measure that exploits a semantic user profile derived from a predefined web ontology (ODP). The measure is based on combining Minimum Common Supergraph (MCS) and Maximum Common Subgraph (mcs) between graphs representing respectively the document and the user profile. We extend this measure in order to take into account a semantic recovery between the document and the user profile through common concepts and cross links connecting the two graphs. Results show the effectiveness of our personalized graph-based ranking model compared to Yahoo[1] search results.

1 Introduction

A major limitation of most existing search engines is that they are based on a content-based query-document matching pattern. Retrieving the most relevant documents for short queries in a large scale document collection, where the user needs are diverse, is a limitation of traditional IR strategies. Indeed, these latter consider that the query is the only key that represents the user information need. Personalized search aims at tackling this problem by considering the user profile that describes the main user interests and preferences, in the search process. The main challenging task in the field is how to represent, infer, and exploit the user profile so as to improve the search performance. User profile models are arranged from very simple representations to complex representations based on semantic resources used to describe the user interests with a rich variety of interrelations among them. The user profile representation model could be based on a bag of words [1], graph of terms [2, 3] defined usually by term co-occurrence, or conceptual representation based on a list of concepts [4, 5] or an instance of the ontology [6, 7]. Improving the search accuracy is then achieved by using

[1] http://www.yahoo.com

P. De Bra, A. Kobsa, and D. Chin (Eds.): UMAP 2010, LNCS 6075, pp. 171–182, 2010.
© Springer-Verlag Berlin Heidelberg 2010

the user profile in a personalized document ranking such as query reformulation techniques, query-document matching models or result re-ranking.

We present in this paper a personalized graph-based document ranking model using a semantic user profile. We have already proposed in previous works [8, 9] a semantic representation of the user profile based on a graph of semantically related concepts issued from a predefined web ontology, namely the ODP ontology[2]. Personalization is based on reranking the search results by calculating for each retrieved document, a personalized score using the cosine similarity measure between the document and the top weighted concepts of the user profile. In this paper, we focus on a personalized document ranking model that represents both the document and the user profile in a graph-based model and uses a graph-based distance measure to calculate a document-profile semantic recovery. It is based on a semantic extension of a graph-based distance measure combining Maximum common subgraph (mcs) and minimum common super-graph (MCS). We extend this measure in order to increase the personalized score of the most related documents to the user profile by taking into account not only common concepts but also cross links connecting the two graphs.

The rest of this paper is organized as follows. In Sect. 2, we present related works in personalized document ranking models and then highlight our contribution. In Sect. 3, our personalized document ranking model is detailed. In Sect. 4, we present the experimental evaluation and results by comparing the performance of our personalized search ranking to Yahoo ranking results. In the last section, we present our conclusion and plan for future work.

2 Related Work

Personalized document ranking is usually achieved by integrating the user profile in the query reformulation process [4, 2], query-document matching [10] or document ranking [7, 6, 5].

Most of query reformulation techniques are based on adding or reweighting terms using *Rocchio* algorithm [4] or by using rewriting rules [2] depending on the user profile representation. In [4], the query is matched with the most appropriate pair of concepts in the user profile; the first concept of the pair is the most similar to the user query, it is used to add or enhance the weight of relevant terms while the second one is the less similar one used to eliminate non relevant terms in the search. In [2], a more elaborated user profile based on terms connected by different edge relations (negation, substitution, etc.) allows to add, eliminate or substitute the query terms with relevant terms.

Using the user profile in the query-document matching model consists of calculating the relevance score of the document relatively to both the user query and the user profile. A bayesian model integrating a document relevance score function is proposed in [10] in order to increase the score of the document when the document vocabulary matches the user profile one.

[2] http://www.dmoz.org

Personalization based on result re-ranking consists of combining the content-based document score with the personalized document score [7, 6, 5] or with the personalized *PageRank* of the document [11]. The personalized score is computed in [7, 6] using the cosine similarity measure between each returned document and the most similar concepts of the user profile. In [5], personalization consists of combining personalized categorization and result re-ranking using a voting-based merging scheme. Personalized result-reranking based on *PageRank* is described in [11]. The user selects his preferred pages from a set of hub pages and one personalized PageRank vector is computed for each user interest used to redirect the returned web pages to the preferred ones.

In most of other related works based on an ontological user profile, the document is represented by a vector of weighted terms and the personalized score of the document is computed according to a term-based similarity measure (cosine) [7, 6] between the document and some concepts of the user profile. The main distinctive feature of our work is to use a semantic document-profile matching model that represents both the user profile and the document by graphs derived from the ODP ontology and computes a personalized relevance score of the document based on an extended semantic graph-based distance measure.

3 A Personalized Graph-Based Document Ranking Model

In our approach, we make use of a semantic user profile model proposed in our previous work [8], which holds the user interest built across a search session. This latter is defined by a sequence of queries related to the same user information need. The user profile is represented as a graph of interrelated concepts of the ODP ontology [3] considered as a highly expressive ground to describe a semantic user profile model. It is built by mapping the user's documents of interests on the ODP and selecting the highly weighted group of concepts that are semantically linked with different edge types in the ontology. Formally, the user profile is represented by a hierarchical (tree) component composed of "is-a" links, and a non hierarchical component composed of cross links of different types predefined in the ontology. It is defined as a directed graph G=(V,E) where:

- V is a set of weighted nodes, representing concepts of interest,
- E is a set of edges between nodes in V, partitioned into three subsets T, S and R, such that: T corresponds to edges made of "is-a" links, S corresponds to edges made of "symbolic" cross links and R corresponds to edges made of "related" cross links.

We exploit the user profile in a result reranking process using a personalized graph-based document ranking model. The model consists of calculating a personalized relevance score of a document with respect to the user profile according to a graph-based distance measure. This latter is based on a semantic extension

[3] http://www.dmoz.org

of the combined distance measure using minimum common supergraph (MCS) and maximum common subgraph (mcs). We review in this section the most common used graph-based distance measures and then present our extended semantic graph-based measure used in the personalized document ranking model.

3.1 Background: Graph-Based Distance Measures

The most common known graph-based distance measures use the maximum common subgraph (mcs) [12], the minimum common supergraph (MCS) [12], the combined measure using MCS and mcs [13] or the edit distance [14].

The maximum common subgraph (mcs) of two graphs g_1 and g_2, is a subgraph g of both g_1 and g_2 based on common concepts and has among all the subgraphs, the maximum number of nodes [15]. The graph-based distance measure based on the subgraph mcs [12] is given by the following formula:

$$d(g_1, g_2) = 1 - \frac{|mcs(g_1, g_2)|}{max(|g_1|, |g_2|)} \tag{1}$$

$|g_1|$ (resp. $|g_2|$) is the the number of nodes in g_1 (resp. in g_2). This formula gives lower distance for graphs having large mcs.

The minimum common supergraph (MCS) of two graphs, g_1 and g_2, is a graph g that contains both g_1 and g_2 as subgraphs and that has the minimum number of nodes and edges [15]. The distance measure based on only the supergraph (MCS) [15] is given in the following formula:

$$d(g_1, g_2) = 1 - \frac{|g_1| + |g_2| - |MCS(g_1, g_2)|}{max(|g_1|, |g_1|)} \tag{2}$$

The distance measure based on combining MCS and mcs is given in [13] as follows:

$$d_{MMCS}(g_1, g_2) = |MCS(g_1, g_2)| - |mcs(g_1, g_2)| \tag{3}$$

According to this measure, the distance between graphs is lower when the size of the supergraph is smaller and the size of the subgraph is larger.

An alternative of the distance measure based on mcs is the edit distance [14]. It expresses the shortest sequence of edit operations that transform a graph g_1 into another graph g_2. An edit operation is either the deletion, insertion or substitution applied to nodes and edges. This measure is defined as follows:

$$d(g_1, g_2) = min\left\{C(\xi)\right\} \tag{4}$$

where ξ is the sequence of edit operations to transform graph g_1 into graph g_2.

3.2 A Semantic Graph-Based Matching Model

We propose a personalized document ranking model based on a graph-based distance measure for calculating the personalized score of a document. This latter reflects a semantic recovery with the user profile based on the following hypothesis: "*a document is ranked higher if it recovers the maximum of concepts of the user profile at both specific and general levels*". Based on this hypothesis, personalization is achieved as follows:

- Map each retrieved document's content d_k represented by a term-based vector on the ODP ontology using the cosine similarity measure and extract the document's graph by connecting the top 20 weighted concepts using different edge types of the ontology.
- calculate the personalized score $S_p(d_k, G_u)$ of each retrieved document d_k using an extended semantic graph-based distance measure combining MCS and mcs with respect to the reference ontology (ODP),
- re-rerank the search results according to the final score $S_f(d_k)$ of each retrieved document d_k calculated by combining, its initial score S_i returned by the system using a content-based ranking model and its personalized score S_p as follows:

$$S_f(d_k) = \gamma * S_i(q, d_k) + (1 - \gamma) * S_p(d_k, G_u) \tag{5}$$

  $0 \leq \gamma \leq 1$. When γ has a value of 1, personalized score is not given any weight. If γ has a value of 0, the original score is ignored and pure personalized score is considered.

In the next sections, we present our motivations behind using a graph-based distance measure and the way of calculating the personalized score of a document using the semantic extension of this measure over the ODP ontology.

Toward a Semantic Graph-Based Measure. We propose a graph-based distance measure based on the combination of MCS and mcs. Our choice is based on the following reasons:

- The Minimum Common Supergraph (MCS) allows measuring the similarity between the document and the user profile at general levels even if the graph do not have common concepts. Indeed, the supergraph of two graphs is smaller when the root nodes of both graphs are close in the ontology.
- The Maximum Common Subgraph (mcs) allows measuring the similarity between the document and the user profile at specific levels. Indeed, the subgraph is larger when the two graphs have common concepts at specific levels, and consequently this enlarge the subgraph with common concepts at general levels.

However, this measure can't deal with approximate matching between the document and the user profile. Indeed, this measure give similar distance for documents that have cross links with the user profile and others that dont have cross links neither common concepts. This is due to the exact recovery assumption using the common concepts to build the subgraph. That's why we argue to extend semantically the subgraph of two graphs by taking into account relative document-profile semantic recovery. Our intuition at this level is to consider two types of recovery in the semantic distance measure:

1. *Exact recovery*: refers to an exact similarity between the document and the user profile. It is calculated by the number of common concepts.
2. *Relative recovery*: refers to a relative similarity between the document and the user profile. It is calculated by the number of common related concepts connecting the two graphs, in other terms, those linked with cross links.

Calculating the Personalized Document Rank. We calculate the personalized document rank by extending first each of the document's graph and the user-profile's graph and then by computing the distance measure based on the combination of MCS and mcs of the two extended graphs.

– *A semantic extension of the mcs:* Formally, let g_1 and g_2 the graphs representing respectively the user profile and the document. As shown in Fig. 1, the set of concepts of graph g_2 connected to graph g_1 with cross links represent the extension of graph g_1 (in Fig. 1, concepts c_{11} and c_{14} forms the extension of graph g_1). Formally, we define the extended graph g_1^{2*} of graph g_1 with respect to g_2, as follows:

$$g_1^{2*} = g_1 \cup \{c_i \in g_2 / \exists c_j \in g_1 \wedge e_{ij} \in S \cup R\} \quad (6)$$

e_{ij} is the edge linking concept c_i to concept c_j, $S \cup R$ is the set of symbolic and related concepts of the ODP ontology (cf. Sect. 3). We create the extended graph g_2^{1*} by the same manner as the graph g_1^{2*}. We obtain two extended graphs g_1^{2*} and g_2^{1*} that will be used in the personalized document ranking model.

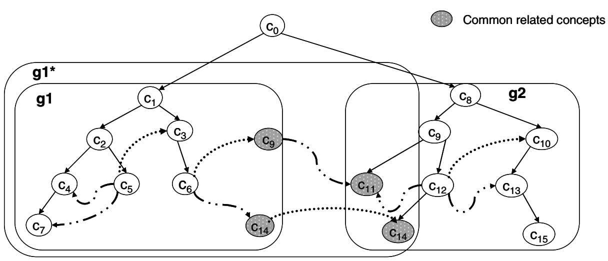

Fig. 1. A semantic extended graph through cross links

– *Calculating the personalized relevance score:* We use the extended graphs g_1^{2*} and g_2^{1*} to calculate the personalized relevance score of the document based on a semantic distance combining the MCS and mcs between the document and the user profile. The mcs of the two extended graphs contains initially common concepts called mcs_{cc} and the activated concepts issued from the graph extension called mcs_{ca} that are either the concepts linking the two graphs through cross links or the inner concepts linking the concepts of the subgraph together.
In order to distinguish the role of the related concepts connected through the cross links relatively to the direct common concepts between graphs, we

used a decay factor f_{ca}. This factor is calculated automatically based on the following assumption: *"the number of activated concepts must be reduced as more as we have symbolic or related edges connecting the graphs"*. f_{ca} is given as follows: $f_{ca} = \frac{L_R}{1+L_R}$, L_R is the set of cross edges linking concepts of the two graphs. Finally, in order to take into account the graph size difference between the documents compared to the user profile, we normalized the semantic distance measure between graphs by dividing over MCS as follows:

$$d(g_1^{2*}, g_2^{1*}) = \frac{\left|MCS(g_1^{2*}, g_2^{1*})\right| - (\left|mcs_{cc}(g_1^{2*}, g_2^{1*})\right| + f_{ca} * \left|mcs_{ca}(g_1^{2*}, g_2^{1*})\right|)}{|MCS(g_1^{2*}, g_2^{1*})|} \tag{7}$$

4 Experimental Evaluation

In this section, we present an experimental evaluation undertaken through a user study in order to compare the effectiveness of our personalized search in front of Yahoo search, and to evaluate the impact of different document-profile similarity measures.

4.1 Dataset

We exploited a search log of a commercial web search engine, namely *Exalead*, and we extracted the search history of 10 users collected along three months. As our approach is based on personalizing search across sessions defined by a sequence of related queries, we have selected 25 user search sessions for all the 10 users as follows:

- Each session contains three queries related to the same user information need and are submitted by the same user in a chronological order.
- Each query has at least one clicked document as it is the only source of evidence to build the user profile in a search session. We consider obviously that a document is relevant if it is clicked by the user.

In order to test the personalized search effectiveness along the user search session, we have divided the query set per session into a training query set to learn the user profile and a testing query set to evaluate the retrieval performance. The first two queries of each search session are part of the training query set, which contains a total of 50 queries. The last formulated user query of each session is part of the testing query set, which contains a total of 25 queries. Testing query terms vary between 1 and 4 and the user intent behind these queries is mostly informational ("*Risques auditifs*") or transactional ("*Le bourg d'oisans hotel*").

The document collection consists of collecting the top 50 results retrieved from the publicly available Yahoo API[4] for each testing query. In our evaluation setting, these documents are used only for reranking the search results using

[4] `search.cpan.org/perldoc?Yahoo::Search`

the user profile. In order to evaluate the retrieval effectiveness, the relevance assessments for the testing queries were given through a user study. To do, 5 computer science students of our lab were presented with the set of top 50 results retrieved from Yahoo. Each participant was considered the user who has formulated the query and asked to judge whether each document was relevant or not according to the subject of a subset of testing queries.

4.2 Evaluation Protocol

The evaluation protocol consists of a training step and a testing step.

1. *Training step:* This step consists of learning the user profiles for each testing query using the clicked documents of the corresponding training queries belonging to the same user.
2. *Testing step:* This step consists of evaluating the personalized retrieval effectiveness for each testing query using the user profile compared to the baseline search performed by Yahoo search using only the testing query. Personalized search is based on reranking the top 50 results of Yahoo for each testing query using the appropriate user profile and by combining for each document its original rank (sorted by Yahoo) and its personalized rank calculated using our graph-based distance measure. We use the precision at top 10 and top 20 documents (P@10, P@20) as an evaluation metric which measures the system performance for documents that are most viewed. We pool together the queries and judgments of all the ten users, so that the evaluation result will be an average over the whole testing queries.

4.3 Experimental Results

In this section, we present the evaluation results by comparing our personalized retrieval effectiveness to the baseline search. Results concern the following objectives: (1) Evaluating the effect of the combination parameter γ on the retrieval effectiveness, (2) Evaluating our personalized search to the baseline search performed by Yahoo and to different similarity measures used to calculate the personalized score of the document.

Effect of γ Combination Parameter on the Retrieval Performance. In this experiment, we study the effect of combining the original document's rank of Yahoo (corresponding to the original document score in formula 5) and the personalized document rank on the retrieval effectiveness using a combination parameter γ (formula 3). Figure 2 shows the improvement of our personalized search with varying γ in the interval [0 1]. Results show that the best performance is obtained when γ is 0, i.e., when the original search engine rankings are ignored altogether. This is likely due to the fact that all the results on the top 50 match the query well and thus the distinguishing feature is how well they match the user profile.

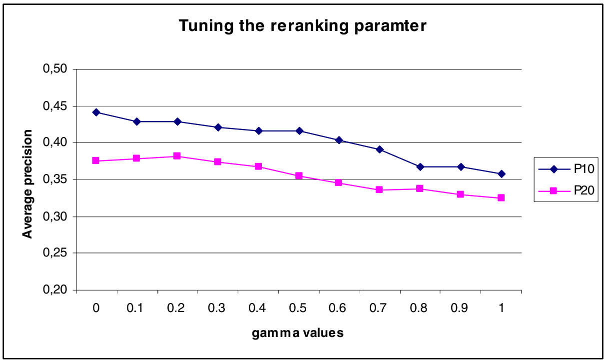

Fig. 2. Effect of γ on the final rank

Evaluating the Personalized Ranking Model Effectiveness. In this experiment, we compare our personalized retrieval effectiveness to the baseline search and to other different measures. These measures are (1) the basic combined measure using MCS and mcs, normalized without using a decay factor (formula 3), (2) the distance measure based on mcs (formula 1), (3) the distance measure using MCS (formula 2) and (4) the cosine similarity measure proposed in our previous works [8, 9]. This latter calculates the personalized score of the document using the most highly weighted concepts of the user profile. In order to set a reliable comparison between the different measures, we have conducted a preliminary experiment in order to identify the best number of concepts used to calculate the personalized score of the document according to the cosine similarity measure. We outline that earlier experiments [8] have shown the effectiveness of personalizing search using the basic cosine similarity measure compared to a typical search and also by comparison to the personalized search approach described in [6].

(A) Effet of the number of concepts used in the cosine similarity measure on the retrieval performance. According to our previous work [8], the personalized score of the document using the cosine similarity measure is computed as follows:

$$S_p(d_k, G_u) = \frac{1}{h} \cdot \sum_{j=1..h} score(c_j) * \cos(\overrightarrow{d_k}, \overrightarrow{c_j}) \qquad (8)$$

where $\overrightarrow{d_k}$ and $\overrightarrow{c_j}$ are term-based vectors representing respectively document d_k and concept c_j, $score(c_j)$ is the weight of concept c_j in the user profile G_u and h is the number of concepts considered in the personalized search.

In this experiment, we varied the number of concepts used to calculate the personalized score of the document using the cosine similarity measure. Figure 3 shows P@10 and P@20 of the personalized search. We can see that the best improvement ($17,24\%$ at P@10 and $8,07\%$ at P@20) is obtained when using 7 concepts of the user profile. We retain this value for comparing the retrieval effectiveness using the cosine measure with other graph-based measures.

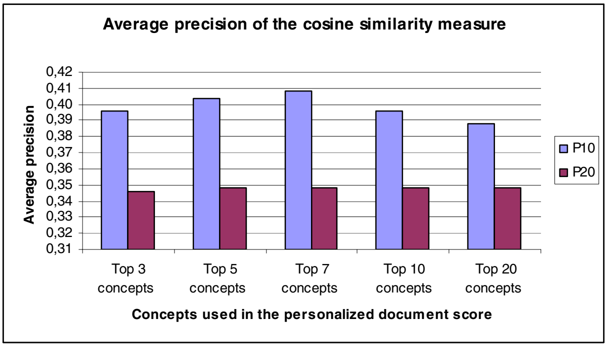

Fig. 3. Effect of the number of concepts using the Cosine measure

(B) Evaluating the effect of the personalized graph-based ranking. In this experiment, we compare the effectiveness of our personalized search to Yahoo search results as well as to personalized search results performed using the measures mentioned above. All graph-based measures calculate the distance between the extended graphs of both the document and the user profile. We recall that for each testing query, the personalized search is based on reranking the top 50 results in descending order according to the graph-based distance measure and ascending order according to the cosine similarity measure.

Table 1 shows the improvement of each measure compared to Yahoo search. Results show a significant improvement of our personalized search at both P@10 and P@20. This indicates the effectiveness of ranking semantically the documents with respect to the user profile using our semantic graph-based distance measure. Indeed, our measure gives higher ranks for documents that are semantically related to the user profile by bringing them to the top 10 and the top 20 results presented to the user. We can confirm also that information from the clicked web pages of training queries can be used to provide an effective personalized search. In order to get a more detailed understanding of the effects of our personalized search, we examined the results on a query by query basis at P@10. For the 25 testing queries reranked using our graph-based measure, 12 (48%) showed an improvement, 8 (32%) were unchanged, and only 5 (20%) were negatively impacted. Thus, the personalized reranking helped 2 times as many queries as it hurts. This is probably due to the difference of clarity degree between the queries measured by the number of query terms.

By comparison to the basic combined distance measure, the decay factor f_{ca} of our measure has a positive effect of increasing the P@20 documents. This proves that the direct common concepts should be given more weights than common related ones. Using the distance measure based only on the subgraph (mcs) or the supergraph (MCS) gives lower improvement compared to the combined measure especially at P@10. This proves that combining semantic distance at general and specific levels allows performing better personalized search improvement. The cosine similarity measure performs a reasonable improvement at P@10 documents, which proves that the user profile is effective even using only

Table 1. Comparison of different document-profile similarity measures

	P@10	P@20
Yahoo search	0,3480	0,3220
Our semantic extended measure ($MCS+mcs+f_{ca}$)	0,4280	0,3660
Improvement	**22,99%**	**13,66%**
Classic combination measure ($MCS+mcs$)	0,4280	0,3620
Improvement	22,99%	12,42%
mcs	0,4160	0,3660
Improvement	19,54%	13,66%
MCS	0,3960	0,3600
Improvement	13,79%	11,80%
Cosine	0,4080	0,3480
Improvement	17,24%	8,07%

few concepts. On the other hand, it performs the lowest improvement at P@20 documents which proves the effectiveness of using a semantic distance measure between the document and the user profile.

5 Conclusion and Outlook

We presented in this paper an approach for personalizing search using a conceptual graph-based user profile. Personalization is achieved by reranking the search results based on a graph-based distance measure combining MCS and mcs and by considering cross links between graphs. Our experimental evaluation is carried out using real user queries issued from *Exalead* web search log. Results show that our model achieves higher performances compared to Yahoo search results and to other graph-based distance measures. In future work, we plan to improve the accuracy of both the document and the user profile graph-based representations and study the effect of tuning the importance of related concepts between graphs relatively to direct common concepts on the retrieval performance.

Acknowledgments. This work was in part supported by OSEO under the Quaero program.

References

1. Gowan, J.: A multiple model approach to personalised information access. Master thesis in computer science, University of College Dublin (2003)
2. Koutrika, G., Ioannidis, Y.: A unified user profile framework for query disambiguation and personalization. In: Proceedings of the workshop on New Technologies for Personalized Information Access, pp. 44–53 (2005)
3. Micarelli, A., Sciarrone, F.: Anatomy and empirical evaluation of an adaptive web-based information filtering system. User Modeling and User-Adapted Interaction 14(2-3), 159–200 (2004)

4. Sieg, A., Mobasher, B., Burke, R., Prabu, G., Lytinen, S.: Using concept hierarchies to enhance user queries in web-based information retrieval. In: AIA '04: Proceedings of the international Conference on Artificial Intelligence and Applications, Innsbruck, Austria, pp. 114–124 (2004)
5. Liu, F., Yu, C., Meng, W.: Personalized web search for improving retrieval effectiveness. IEEE Transactions on Knowledge and Data Engineering 16(1), 28–40 (2004)
6. Gauch, S., Chaffee, J., Pretschner, A.: Ontology-based personalized search and browsing. Web Intelligence and Agent Systems 1(3-4), 219–234 (2003)
7. Sieg, A., Mobasher, B., Burke, R.: Web search personalization with ontological user profiles. In: CIKM '07: Proceedings of the sixteenth ACM Conference on information and knowledge management, pp. 525–534. ACM, New York (2007)
8. Daoud, M., Tamine, L., Boughanem, M.: Towards a graph based user profile modeling for a session-based personalized search. Knowledge and Information Systems 21(3), 365–398 (2009)
9. Daoud, M., Tamine-Lechani, L., Boughanem, M., Chebaro, B.: A session based personalized search using an ontological user profile. In: SAC '09: Proceedings of the 2009 ACM symposium on Applied Computing, pp. 1732–1736. ACM, New York (2009)
10. Tamine-Lechani, L., Boughanem, M., Zemirli, N.: Personalized document ranking: exploiting evidence from multiple user interests for profiling and retrieval. Digital Information Management 6(5), 354–365 (2008)
11. Jeh, G., Widom, J.: Scaling personalized web search. In: WWW '03: Proceedings of the 12th international conference on World Wide Web, pp. 271–279. ACM, New York (2003)
12. Levi, G.: A note on the derivation of maximal common subgraphs of two directed or undirected graphs. Calcolo 9(4), 341–354 (1973)
13. Fernández, M.L., Valiente, G.: A graph distance metric combining maximum common subgraph and minimum common supergraph. Pattern Recognition Letters 22(6-7), 753–758 (2001)
14. El-Sonbaty, Y., Ismail, M.A.: A new error-correcting distance for attributed relational graph problems. In: Amin, A., Pudil, P., Ferri, F., Iñesta, J.M. (eds.) SPR 2000 and SSPR 2000. LNCS, vol. 1876, pp. 266–276. Springer, Heidelberg (2000)
15. Bunke, H., Jiang, X., Kandel, A.: On the minimum common supergraph of two graphs. Computing 65(1), 13–25 (2000)