A relevance-focused search application for personalised ranking model

AL SHARJI, Safiya, BEER, Martin <http://orcid.org/0000-0001-5368-6550> and URUCHURTU, Elizabeth <http://orcid.org/0000-0003-1385-9060>

Published version

AL SHARJI, Safiya, BEER, Martin and URUCHURTU, Elizabeth (2016). A relevance-focused search application for personalised ranking model. In: HARTMANM, Sven and MA, Hui, (eds.) Database and expert systems applications : 27th International Conference, DEXA 2016, Porto, Portugal, September 5-8, 2016, Proceedings. Lecture Notes in Computer Science, 2 (9828). Switzerland, Springer, 244-253.

A Relevance-Focused Search Application for Personalised Ranking Model

Al Sharji Safiya, Martin Beer and Elizabeth Uruchurtu

Communication & Computing Research Institute,
Sheffield Hallam University
153 Arundel Street, S1 2NU, Sheffield, UK

Abstract. The assumption that users' profiles can be exploited by employing their implicit feedback for query expansion through a conceptual search to index documents has been proven in previous research. Several successful approaches leading to an improvement in the accuracy of personalised search results have been proposed. This paper extends existing approaches and combines the keyword-based and semantic-based features in order to provide further evidence of relevance-focused search application for Personalised Ranking Model (PRM). A description of the hybridisation of these approaches is provided and various issues arising in the context of computing the similarity between users' profiles are discussed. As compared to any traditional search system, the superiority of our approach lies in pushing significantly relevant documents to the top of the ranked lists. The results were empirically confirmed through human subjects who conducted several real-life Web searches.

Keywords: User Profile, Keyword-Based Features, Semantic-Based Features.

1 INTRODUCTION

The use of Implicit Feedback (IF) is proven to improve the performance of retrieval systems [4]. In this paper, it was empirically demonstrated that users' intentions can be learnt by implicitly mining their interaction data. Consequently, relevant documents matching both the user's inputted keywords (i.e. queries) and particular needs can be retrieved. We build upon these ideas to construct users' interest profiles which are used to infer relevant documents. Ranking functions are then crafted based on both the relevance and interest scores of these documents leading to the generation of a relevance-focused personalised search. Query expansion technique is employed through WordNet[1] ontology to integrate terms which are not directly expressed in the users' queries.

The requirements for personalised search models include a learning process to extract users' information (i.e. interaction activities) meeting their individual information needs. We employ users' clicked documents to build and maintain their interest profiles. The rank algorithm takes into account the learned patterns together with the active users' profiles [10] to develop a PRM based on which search results are ranked to represent the users' interests [10]. The main argument is that IR can be em-

[1] https://wordnet.princeton.edu.

ployed by the PRM to provide ranked lists of the documents based on individual user's interests. It is thus investigated whether users' interests can be identified through implicit interactions in digital web documents. The main challenge addressed is how query keywords and their related concepts can be used to identify users' individual interests (i.e. relevant documents); and how acquired feedback is preserved over time in order to include representation of both the users' interests and modelling.

2 RELATED WORK

Personalised searches differ in the type of data and approaches used to build the user profile [10] both of which play a major role in personalised search approaches. A recent study [13] uses spreading mechanism through ontology to provide inherent relationships between terms/concepts appearing in their respective bag-of-word representation in order to extend the semantic similarity concept between two entities. However, it is still an open research question whether a mechanism could be devised to control and correct the integration of ontology terms in the query expansion. This would match the users' information needs thereby guaranteeing that recall is improved during the phase without degrading precision as a result of this process. A technical report by William [14] presented the idea of indexing material at the sentence and phrase level to support improved information access so that the content of an individual sentence or phrase could be located in response to a specific description of need. To identify appropriate concepts within annotated audio text, Khan [5] has also presented an automatic disambiguation algorithm which could prune as many irrelevant concepts as possible while at the same time retaining the largest possible number of relevant concepts. While these studies provide the techniques adopted to improve the performance of Information Retrieval (IR) systems in terms of precision or recall or both, they do not however detail the effects of such integration with regards to different levels of keyword mixtures of the terms in both queries and ontology during the matching process. Following on from [1], this paper presents such effects.

3 RELEVANCE-FOCUSED SEARCH

This section outlines our two models representing users' interests and preferences in a formal way, such that both approaches can be checked for validity to form customised views of a relevance-focused search application for personalised search.

3.1 Keyword-based features

Users' profiles are often defined by storing the content of documents clicked after being collected over time [10]. Given a set of users' Web search logs, any search documents clicked are archived for each user whose representations are determined based on these documents. For our purpose, a feature can be considered as an attribute of text content (i.e. document or query content) which is used to make decisions related

to it. Thus, to determine a relevant document means to extract its important features that can determine factors which are important to a user searching for such a document. These features are then used to craft the ranking predictors which are often combined together to improve the retrieval process.

Assuming there is a set m of users represented by $U = \{u_1, u_2, ...u_m\}$ and a set n of documents represented by $D = \{d_1, d_2, ...d_n\}$, a profile for user $u \in U$ can be represented as an ordered pair of n-dimensional vectors using equation 3.1 [10].

$$u^{(n)} = \langle (d_1, s_u(d_1)), (d_2, s_u(d_2)), ... (d_n, s_u(d_n)) \rangle \quad (1)$$

where each $d_j \in D$ and s_u is the function for user u which assigns interest scores (i.e. interest score) to documents.

Since each document $d_j \in D$ can represent an HTML document in the context where the focus is to capture the implicit feedback related to the document clicked, equation 3.1 might be used to represent the user's profile. Each document d_j can then be represented as an attribute vector of k-dimensional features where k is the total number of features extracted [10]; and the feature weight associated with the document is represented by its corresponding dimension in a feature vector which is given by: $d_j = \langle fw_j(f_1), fw_j(f_2), ..., fw_j(f_k) \rangle$, where $fw_j(f_p)$ is the weight of the pth feature in $d_j \in D$, for $1 \le p \le k$. Since the features extracted are the textual content of pages represented in Bag-of-Words (BOW, i.e. a set of pairs, denoted as $\{t_i, w_i\}$, where t_i is a term describing the content of the page (i.e. document) such that $t_i \in d_j$, and w_i is its weight found by using the normalised $tf \bullet idf$ term values [9], each document can thus be represented by sets of term-score pairs (e.g., sport (cricket; 0:54); (baseball; 0:39); (soccer; 0:45)[2]) leading to the user profile represented as a feature vector using the terms of documents as features.

Given a user profile UP containing v interest vectors for a user u_k, an overall interest vector is often determined by combining all interest vectors for that user [9]. Assuming T_i is the set terms in the i^{th} ($i \in [1, v]$) interest vector, the set of terms of the overall interest vector T can be found as $T = \bigcup_{i=1}^{v} T_i$. For every term $t \in T$, its overall interest vector can be calculated as $s_u(t) = \sum_{i=1}^{v} s_i(t) \bullet w_i$, where $s_i(t)$ is the score (relevance score) of t in the i^{th} interest vector ($s_i(t) = 0$, if $t \notin T_i$) and w_i is the actual weight of the i^{th} interest vector.

[2] Figures based on a different experiment and given here solely for illustrative purposes.

3.2 Semantic-based features

The spreading approach can be adopted [13] in order to perform the automatic query expansion [13] by appending terms that are conceptually related to the original set of terms in documents. We build on this earlier work and provide a conclusive empirical analysis when related terms are considered and the degree of their contribution to improve the performance of IR systems. Although there are many overlaps between the current research and the latter approach aimed at providing semantic similarities through ontologies, in terms of classification technique employed to create users' profiles to describe the contents of Web documents clicked, this project applies both term weight (i.e. term frequency factor) and dwell weight[3] directly as a dimensional feature to enrich the users' models [1,2]. For instance, not only was it shown in these surveys that the performance of the PRM improved, but it was also demonstrated that it could be used as a complementary feature for the system to rely on when the keyword feature proves unsuccessful in identifying the relevance of documents.

Given ontology O and term t_i, spreading process might employ the ontology O to spread document d_j, to determine the terms that are related to t_i, so that any terms related to the original terms of the document can be included. Denoting these terms as $\mathrm{Re}\,lO(t_i)$, the results of spreading the document d_j, is an expanded document $\hat{d}_j$ such that the set of terms $\hat{d}_j = \{t_1,...,t_n,t_{11},...,t_{mn}\}$ and $d_j \subseteq \hat{d}_j$ where $\forall t_{ij} / t_{ij} \in \mathrm{Re}\,lO(t_i)$ and a path exists from t_i to t_j.

This spreading process is an iterative process; and the terms from the previous iterations that are related to the original terms are joined to the document at the end of the iteration. The spreading process terminates when there are no related terms to spread the document with, or simply when $\forall t_i \in d_j / \mathrm{Re}\,lO(t_i) = \theta$.

3.3 Cosine Similarity Measure

For the purpose of this work, in order to compute the vector similarities determining the user's interest in a particular document, the cosine similarity measure is adopted [9] as the technique to represent the user model.

Given a user profile $UP = s_{u_k}(d_j)$ and a document $d_j = \{t_1,...,t_n,t_{11},...,t_{mn}\}$ for a given search (document containing a set of texts where each t_i is a k-dimensional vector in the space of content features), the binary cosine similarity [9] denoted as $Sim(UP, d_j)$ can be determined using equation 3.2. Such similarity between the two sets of texts clearly indicates the relevance of the document in the keyword-based approach which can be applied to the respective vectors.

[3] This dual technique was thoroughly explained previously and authors do not claim this contribution in the current paper.

$$Sim(UP, d_j) = \frac{|UP \cap d_j|}{|UP| \times |d_j|} \tag{2}$$

where $|UP \cap d_j|$ represents the number of keywords in both UP and d_j, and $|UP|$ and $|d_j|$ are respectively the number of keywords in UP and d_j.

3.4 Semantic Similarity Measure

Similarity can be determined to be equal to the inverse of distance in its simplest form or some other mathematical function based on ontological distance. Semantic similarity can thus be inversely proportional to the distance between concepts whereby the closer two concepts are in the ontological representation; the higher the similarity score between them is [6]. Any similarity between two concepts can then be determined by taking the cosine angle between the two corresponding vectors [8]. Mathematically, semantic similarity is determined here by employing a fuzzy ontology value [7], whereby increasing distance between two consecutive terms is inversely proportional to an increase in semantic similarity. Here it is important to recall that words which have been integrated are not directly related to the keyword queries, thus, it is not feasible to apply the cosine similarity measure directly. The application of fuzzy ontology values as shown in equation 3.4 [2] addresses this problem. Thus, based upon this similarity measure (i.e. fuzzy ontology values) a set of relevant documents are obtained. However, expanded documents are still those documents matching the users' queries at first place as demonstrated elsewhere; therefore, after constructing the semantic document vectors in this way, the normal binary cosine similarity measures are applied to refine the ranking function.

Given a user profile with a set of texts $UP = s_{u_k}(d_j)$ and a document $\hat{d}_j = \{t_1,...,t_n,t_{11},...,t_{mn}\}$ for search (expanded document containing a set of texts where each t_{ij} is a k-dimensional vector in the space of content features); cosine similarity denoted as $Sim(UP, \hat{d}_j)$ which is determined following equation 3.2 can be applied to represent the user's interests.

$$F_{jk} = \frac{c_{jk}}{X^2} \tag{3}$$

where c_{jk} is the distance between keywords t_{ij} and t_{ik} or the frequency of the keywords/concepts appearing consecutively in the keyword list, and X is the total number of t_{mn} terms (i.e. keywords) in that document.

3.5 Query Processing and Ranking

Users' queries expressed in keywords to represent their information needs can be considered as short documents. Thus, for each user u_k, a BOW representation for each query issued by the user in a particular session must also be created and compared with its set of corresponding documents. This comparison is based on the similarity between both the query and the targeted documents. Thus, equation 3.4 is applied to calculate the cosine similarity measure between the query vector, the vectors of the matching documents and the vectors of the matching user profile respectively.

$$Sim(q_i, d_j) = \frac{\left|q \cap UP \cap d_j\right|}{\left|q\right| \times \left|UP\right| \times \left|d_j\right|} \tag{4}$$

where $\left|q \cap UP \cap d_j\right|$ represent the number of keywords in q, UP and d_j, and $\left|q\right|$, $\left|UP\right|$ and $\left|d_j\right|$ are respectively the number of keywords in q, UP and d_j.

The highest similarity values are used to establish our relevance-focused search application. They are provided by equation 3.4 when considering the keyword-based features as well as the semantic-based features when the document is integrated with ontology terms. Thus they represent the most similar documents between the query, the user profile and the available documents.

3.6 Search Result Personalisation

The personalisation of search results to a large degree lies in merging the models that provide them. A description of the linear combination adopted in the current research can be found in [2]. Here, as outlined in the following section, the aim is to test the system' models on a deeper level and to investigate their real world problems as closely as possible. A set of the experiments performed by using human subjects (i.e. 729 query keywords[4]) while conducting real-life Web searches is thus presented to validate each model individually. Such evaluation enabled us not only to obtain the system's performance based on each model, but also to evaluate real collections based on different terms integration with different terms of query keywords.

4 EVALUATION

The experimental results are presented in this subsection. For simplicity, the proposed personalised search approach is referred to as experimental system while the search approach which is not personalised is referred to as Baseline. There are two main sets of experiments: (1) Implicit Feedback vs. No-Feedback. Its relative experi-

[4] A detail description of this data set can be found in [2].

mental results are presented in Table 4-1 and visualised in Figure 4-1. (2) Keyword-Based vs. Semantic-Based. Its relative experimental results are shown in Figure 4-2.

4.1 Experimental Set up

Assuming a given user $u_k \in U$ clicked the document d_j after issuing a query containing the word t, then the document d_j is considered useful and relevant to t for user u_k, and documents that are not retrieved, are judged as non-relevant by the user [12]. To evaluate the search accuracy of the two models, sets of documents $d_j \in D$ containing the word t selected by u_k were checked whether they are highly ranked in the ranked list generated by the personalised search solution.

Implicit Feedback vs. No-Feedback.

In this experiment, it was investigated how a sample of real data collected during interaction between users and the system can affect the performance of the personalised search. This includes investigating how useful the acquired feedback is when preserved over time in the form of user profiles [11] to include the representation of their interests. If the experimental system generates accurate ranked lists in terms of higher precision in the lower ranks, then it can be considered to perform better.

A system's performance is often assessed in terms of search results and by its ability to push relevant documents to the lower ranks. Thus, to compare the performances of two systems - here, experimental and baseline systems - ranked lists of search results obtained by the user need to be considered for both systems. The one that is better able to *push* relevant documents to the top of the ranked lists of search results is the more efficient. Table 4-1 gives the overall precision obtained at rank 5 and 10 of both systems. It is important to recall that precision is obtained by dividing the number of relevant documents - for each user - among the top 5/10 documents by 5 or 10 accordingly. Here, results to the first page (i.e. 10 documents) are considered.

Table 1. Average of Precision at Rank 5 and Rank 10

Precision	Baseline	Experimental System			
		Keyword-Based	P(paired t-test)	Semantic-Based	P(paired t-test)
System @ Rank 5	0.79	0.83	0.006%	0.94	0.005
System @ Rank 10	0.56	0.75	0.50%	0.85	0.78%

From table 4-1, the overall averages of the precision at rank 5 and at rank 10 for the experimental system when employing the semantic-based approach, clearly indicate that out of 5 documents, the system can rank more than 4 documents based on their relevancy to the query (0.94*5 = 4.70 and 0.85*5 = 4.25). While the perfor-

mance of the system is more or less constant at rank 5 by employing the keyword-based approach, it is poorer at rank 10, since out of 5 documents, it can only rank 3.75 (0.75*5 = 3.75) documents. The worst performance can be observed from the baseline, as its overall averages of the precision at rank 5 and at rank 10 indicate that having 5 documents, the system is able to rank, based on their relevancy to the query, less than 4 documents (0.79*5 = 3.95) and less than 3 documents (0.56*5 = 2.80) respectively.

Overall, the experiments showed that the personalised system outperforms the baseline with a statistically significant (paired t-test) difference between them

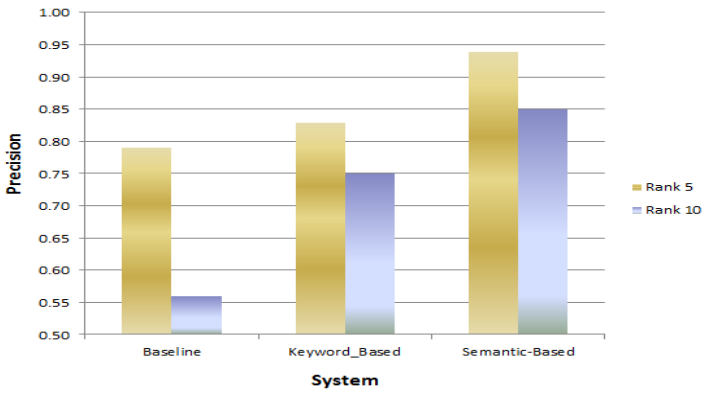

Fig. 1. System Performance

Keyword-Based vs. Semantic-Based.

The goal of this experiment was to use the same idea with the same data set to study whether the semantic-based approach is superior to relying on the keyword-based approach with regards to a personalised search. Here, it should be recalled that in the semantic-based approach, the spreading mechanism was used to incorporate the concept terms into the documents, however, the same statistics were used in both models. Therefore, the semantic-based approach is the expansion of the keyword-based approach with the integration of content semantics expressed in ontology terms so that an enriched user model (i.e. user profile) is generated. This experiment will test the effects of combinations of keywords from the ontology terms with the keywords from the query to enrich the user model, so that the effect of mixing different keywords to generate ranked lists can be investigated.

Each of the participant collections was thus indexed individually into document vector files. Figure 4-2 shows a representation of the distribution of document indices (here, the values of interest vector - denoted as $s_u(t)$) according to different combinations of the query keywords[5] with its related concepts[6] mixtures. Here, $kxny$ means

[5] According to [3], on the average, a query contains 2.21 terms.

x keyword(s) and y concept(s) or ontology terms are employed in the user model. For example, $k2n2$ and $k2n3$ are respectively the keywords employed in the iterations in which two and three ontology terms are integrated into the user model for the second keyword of the query. The threshold interest vector values are the values represented by $kxn\theta$, meaning that only keyword-based is employed and no ontology terms have yet been added to the documents.

As can be seen from Figure 4-2, the semantic-based layout showed the best results when a document is integrated with 3 and 4 keywords (at $kxn3$ and $kxn4$) regardless of the original number of terms (i.e. keywords) contained in the query. The presentation given here is related to only one query, but statistical evidence (ANOVA p value = 6.80%) indicated that out of 729 keyword queries, this observation is consistent across more than 650 keyword queries.

However, expanding the document with 1 or 2, 5 and 7 keyword(s) showed some slight improvements for most documents. On the other hand, integrating the document with 6 and 8 keywords showed worse performance (represented by $kxn6$ and $kxn8$ in Figure 4-2), which might be due to the inclusion of keywords not related to the original term meaning.

Overall, employing keyword-based features alone showed poorer performance than employing semantic-based features if the spreading or query expansion integrates 3 or to 4 keywords into the document.

5 CONCLUSIONS

Derived from several existing techniques, this paper has presented an effective personalised search model that exploits users' profiles by employing their implicit feedback for query expansion through a conceptual search to index documents. Empirical validation confirmed the reliability of our system. A combination of the keyword-based and semantic-based features to provide further evidence of relevance-focused search application for each individual user was validated by using human subjects conducting real-life Web searches. The findings of the experiments demonstrated that, compared to any traditional search system, our approach can push significantly higher number of relevant documents to the top of the ranked lists.

A series of two different web search experiments was performed using different keywords from real users. For each search session, a list of personalised webpage re-ranking over the search results returned by Google was generated. Both the evaluation metric parameters of precision and recall were adopted to measure the ranking quality of the personalised search engine in order to determine the relevance of the results according to their order of relevance.

---

[6] It was demonstrated in [13] that the computation process of terms' weight during document expansion turns monotonic after the third iteration. In current work, this computation turns monotonic after the document is expanded with the eighth term concept.

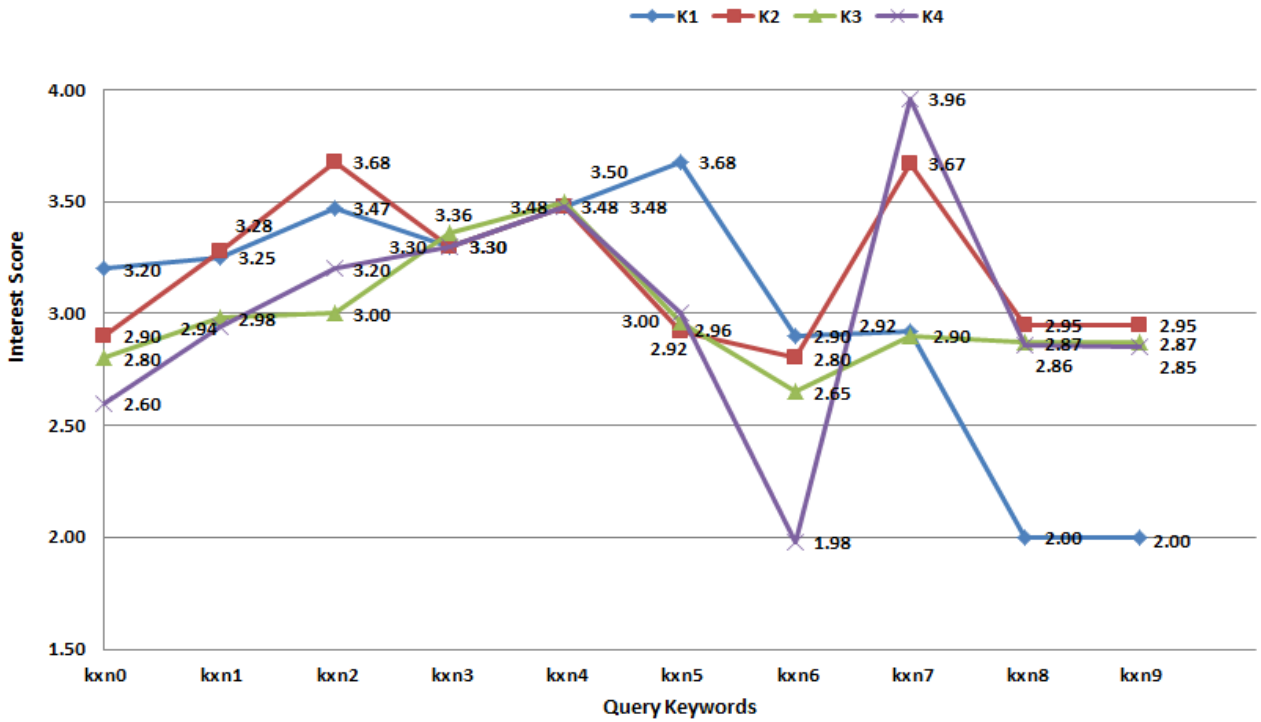

Fig. 2. Comparisons of Mixtures of Query Keywords with Ontology Terms

ACKNOWLEDGEMENT

This research was supported by the Ministry of Manpower in Oman which has granted the funding for the survey of this research.

REFERENCES

1. Al-Sharji Safiya, Beer Martin and Uruchurtu Elizabeth: A Dwell Time-Based Technique for Personalised Ranking Model. In: Database and Expert Systems Applications, Springer International Publishing, pp. 205-214, LNCS (2015).
2. Al-Sharji Safiya, Beer Martin and Uruchurtu Elizabeth: Enhancing the Degree of Personalisation through Vector Space Model and Profile Ontology. In: IEEE RIVF International Conference Computing and Communication Technologies, Research, Innovation and Vision for Future (RIVF), pp. 248-252, IEEE (2013).
3. Jansen Bernard J., Spink Amanda and Saracevic Tefko: Real Life, Real Users, and Real Needs: a Study and Analysis of User Queries on the Web. Information Processing & Management, Vol.36 (2), pp. 207-227 (2000).
4. Kelly Diane and Teevan Jaime.: Implicit Feedback For Inferring User Preference: A Bibliography. In: ACM SIGIR Forum, Vol. 37(2), pp. 18-28, ACM (2003).
5. Khan Latifur and Dennis McLeod: Disambiguation of Annotated Text of Audio Using Onologies. In: Proceeding of ACM SIGKDD Workshop on Text Mining (2000).
6. Leacock Claudia and Martin Chodorow: Combining Local Context and WordNet Similarity for Word Sense Identification. In: Christiane Fellbaum (Ed.), WordNet: An electronic lexical database, MIT Press, Vol. 49(2), pp. 265-283, MIT (1998).
7. Lucarella Dario and Morara R.: First: Fuzzy Information Retrieval System. In: Journal of Information Science, Vol. 17(2), pp.81-91, (1991).
8. Mabotuwana Thusitha, Michael C. Lee and Eric V. Cohen-Solal: An Ontology-Based Similarity Measure for Biomedical Data–Application to Radiology Reports. In: Journal of biomedical informatics, Vol. 46(5), pp. 857-868, (2013).
9. Manning Christopher D., Raghavan Prabhakar and Schütze Hinrich: Introduction to Information Retrieval. Cambridge University Press Cambridge (2008).
10. Mobasher Bamshad: Data Mining for Web Personalization. In: Peter Brusilovski, Alfred Kobsa, Wolfgang Nejdl (Ed.). The Adaptive Web, Methods and Strategies of Web Personalization. Springer-Verlag Ed., LNCS 4321, pp. 90–135, LNCS (2007).
11. Susan Gauch, Mirco Speretta, Aravind Chandramouli and Alessandro Micarelli: User Profiles for Personalized Information Access. In: Peter Brusilovski, Alfred Kobsa, Wolfgang Nejdl (Ed.). The Adaptive Web, Methods and Strategies of Web Personalization. Springer-Verlag Ed., LNCS 4321, pp. 54-90, LNCS (2007),
12. Teevan Jaime, Susan T. Dumais, and Eric Horvitz: Characterizing the Value of Personalizing Search." Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 757-758, ACM, (2007).
13. Thiagarajan Rajesh, Geetha Manjunath and Markus Stumptner: Computing Semantic Similarity Using Ontologies. Hewlett-Packard (HP) Development Company, L.P, Labs Technical Report HPL-2008-87, (2008).
14. Woods William A.: Conceptual Indexing: A Better Way to Organize Knowledge. A technical report of Sun Microsystems, Inc. (1997).