Implicit Evaluation of User's Expertise in Scientific Domains

Alessandro Bonifacio, Claudio Biancalana, Fabio Gasparetti,
Alessandro Micarelli, and Giuseppe Sansonetti(✉)

Department of Engineering, Roma Tre University,
Via della Vasca Navale 79, 00146 Rome, Italy
ailab@ing.uniroma3.it

Abstract. In this article, we propose a system able to implicitly assess a user's expertise in a particular topic based on her publications (e.g., scientific papers) on it and available through online bibliographic databases. This task is performed through two different approaches, both of them based on a graph-based model. The first approach (content-based) considers the text content, the second one (collaborative) analyzes the relationships in the same content in terms of co-citations. Preliminary experimental results are encouraging and raise several interesting considerations. In particular, they show that the best solution is obtained by integrating the two approaches above, in which each of them allows the system to overcome the limitations of the other one.

Keywords: Expertise retrieval · User profile · Graph model

1 Introduzione

Among the various information that a user profile in adaptive systems may include, there is also her competence in a specific knowledge domain. In this article, we propose a system able to implicitly assess the user's expertise in a particular topic based on her publications (e.g., scientific papers) on it and available through online bibliographic databases, such as *Scopus*[1], *Google Scholar*[2], and *ResearchGate*[3]. The proposed system takes in input a candidate user u and a specific knowledge area ka and returns a $score(u, ka)$ expressing the level of competence of u in ka. This task is performed through two different approaches, both of them based on a graph-based model. The first approach (content-based) considers the text content, the second one (collaborative) analyzes the relationships in the same content in terms of co-citations. Specifically, the content-based approach retrieves the most relevant documents for a given knowledge area ka, extracts the most significant entities and stores them in a graph database. Then, it performs the same operations on the documents generated by u on ka and

[1] https://www.scopus.com.

[2] https://scholar.google.com.

[3] https://www.researchgate.net.

C. Stephanidis (Ed.): HCII Posters 2017, Part I, CCIS 713, pp. 420–427, 2017.
DOI: 10.1007/978-3-319-58750-9_58

builds a second graph. Finally, the similarity between the two graphs is computed in order to estimate $score(u, ka)$. The collaborative approach always involves the collection of documents related to the topic ka, but takes into account only the co-citations among them and, therefore, their authors. The evaluation of $score(u, ka)$ is performed through a version of the well-known *Hyperlink Induced Topic Search (HITS)* algorithm [16], which considers the incoming and outgoing edges among nodes.

2 The Proposed System

Nowadays, the increasing availability of online material has led to the need for adaptive systems for its personalized selection [6,7,10], based on the target user's characteristics. Those systems can take into account the personality [8,17], the context [4,5], as well as the effective nature [11–15] and the temporal dynamics [1,3,9] of users' interests. Some adaptive systems also consider the information on the user's expertise in specific knowledge areas. Such information may be obtained through the so-called *expertise retrieval* systems [2]. Approaches to expertise retrieval can be categorized in two main classes, inherited from the Information Retrieval techniques: the first one based on the information content (*content-based*) and the second one independent of it (*collaborative*). The former ones take advantage of the information extracted from the domain of the individual's knowledge to create a profile of her experiences, where the relevance of her documents to the specific field is evaluated. Differently, in collaborative systems user's expertise is assessed based on the authority inferred by analyzing her social network. Both of these approaches have been implemented within the proposed system.

Content-Based Approach. In Fig. 1 the diagram of the overall content-based approach is depicted. Specifically, the first step consists in extracting a set of

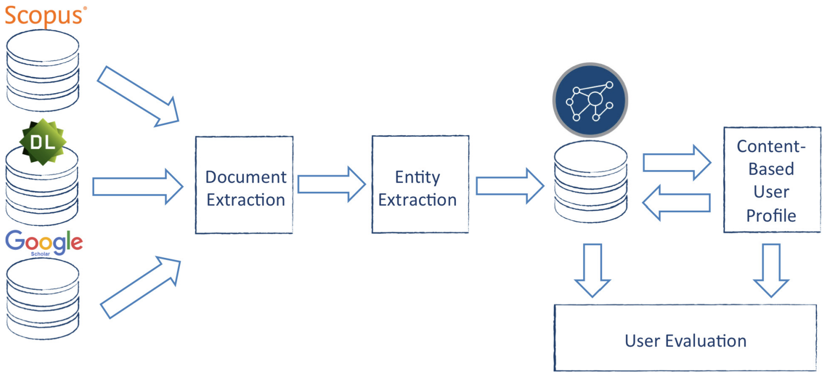

Fig. 1. Content-based approach schema.

documents related to the subject from the knowledge database. A topic annotator[4] is used to extract the entities that characterize those documents. Such entities are stored in a graph database along with information about authors, abstract, affiliations, tags, and categories. When a user has to be profiled, the system performs steps similar to the previous ones but only comprising information regarding her content. Figure 2 illustrates a snapshot of the graph database with regard to the content-based approach. Note the different types of node, such as authors, papers, abstracts, entities, and categories.

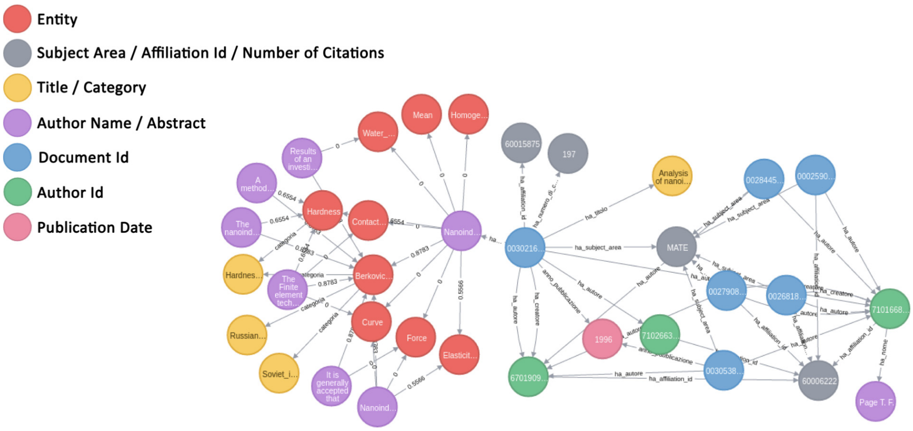

Fig. 2. Snapshot of the graph database in the content-based approach.

Once the domain is defined, different strategies can be applied to evaluate a user's expertise. More specifically, the following four strategies have been implemented in the proposed system:

- *Occurrences.* The first method performs the analysis of occurrences by comparing the keywords extracted from the user's profile with those extracted the stored domain within the graph. The ratio is then between the absolute value of their intersection set and the set of keywords that characterize the domain, as expressed in the Eq. 1, where KW_{ka} identifies the set of keywords describing the knowledge area, KW_u denotes the set of keywords related to the topic used by the user. Such ratio gives a *score*, which expresses the user u's expertise level in that specific knowledge area.

$$score(u, ka) = \frac{|KW_u \bigcap KW_{ka}|}{|KW_{ka}|} \tag{1}$$

- *Weighed Occurrences.* The second method is a variant of the first one, in which it is also considered the weight that each identified entity within the

[4] https://tagme.d4science.org/tagme/.

domain and the user's profile has associated according to how much that entity is relevant to the topic under examination. Such weight is calculated by estimating the distance, namely, the number of levels between pages and categories, between the Wikipedia page associated with the extracted term and the page related to the domain of interest. The weight is also stored inside the edge that links the tag to its abstract within the graph. The method can be described through the following equation:

$$score_{weight}(u, ka) = \frac{|WeightedKW_u \bigcap WeightedKW_{ka}|}{|WeightedKW_{ka}|} \tag{2}$$

- *Log-Entity.* It relies on the comparison through the cosine-similarity metric between the vector representing the candidate user and the one representing the topic. For the weighting function a version of the *TF-IDF* model, well-known in Information Retrieval, has been employed. In particular, the equation for weighing the user is as follows:

$$u = \left\langle \left(e_1, \log\left(\frac{|D_u|}{|d : e_1 \in d|}\right) \cdot w_{e_1,t}\right), \ldots, \left(e_n, \log\left(\frac{|D_u|}{|d : e_n \in d|}\right) \cdot w_{e_n,t}\right) \right\rangle \tag{3}$$

while the equation for weighing the domain is as follows:

$$ka = \left\langle \left(e_1, \log\left(\frac{|D|}{|d : e_1 \in d|}\right) \cdot w_{e_1,t}\right), \ldots, \left(e_n, \log\left(\frac{|D|}{|d : e_n \in d|}\right) \cdot w_{e_n,t}\right) \right\rangle \tag{4}$$

The vectors so obtained are then compared using the cosine-similarity metric. The obtained results, comprised between 0 and 1, describe the user's expertise level in that specific knowledge area.

- *Entity Frequency.* This method, as the previous one, relies on the computation of the cosine-similarity between vectors, but differs from the previous one for the weighing of the vector. In this case, the vector describing the user's profile is constituted by elements which, for each entity belonging to the user's profile, have associated the number of user's documents that contain that entity.

$$u = \left\langle \left(e_1, \left(\frac{|d : e_1 \in d_u|}{|D_u|}\right)\right), \ldots, \left(e_n, \left(\frac{|d : e_n \in d_u|}{|D_u|}\right)\right) \right\rangle \tag{5}$$

The weighing of the vector related to the knowledge area takes place analogously and is described as follows:

$$ka = \left\langle \left(e_1, \left(\frac{|d : e_1 \in d|}{|D|}\right)\right), \ldots, \left(e_n, \left(\frac{|d : e_n \in d|}{|D|}\right)\right) \right\rangle \tag{6}$$

The two vectors are then compared through the cosine-similarity technique, which returns a score expressing the user's expertise in that specific subject.

Collaborative Approach. The system developed according to the collaborative approach analyzes information concerning co-citations among documents related to a particular topic. More specifically, a graph containing documents and their co-citations is built. Such a graph is then analyzed via the HITS algorithm, which for each entity p within the graph calculates the *authority score* $A(p)$ and the *hub score* $H(p)$. Once the ranking of documents is obtained, sorted by their authority value, the ranking of the authors corresponding to those documents is generated. Assuming the possibility that several documents can be written by the same author, it was decided to assign the authority value to the user according to the Eq. 7, which allows us to modify how much weight to assign to the sum of all the authority values of the documents produced by the author or the maximum authority value among the user's documents:

$$Authority(u) = A \cdot \lambda + B \cdot (1 - \lambda) \tag{7}$$

The λ parameter identifies a value between 0 and 1. A and B are respectively the values given by the sum of authority values and the maximum authority value among documents written by the candidate user. In Fig. 3, the diagram of the overall collaborative approach is shown. In this approach, unlike the previous one, the edges of the graph database are only related to the co-citations among documents.

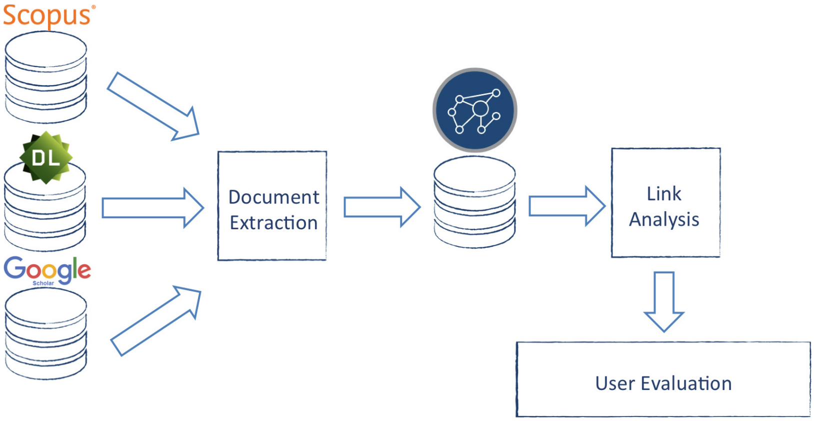

Fig. 3. Collaborative approach schema.

2.1 Experimental Evaluation

To evaluate the performance of our system, we carried out some experimental tests on six candidate users using both approaches. Those candidates were selected so that u_2, u_3, and u_5 were to be considered actually *experts* on the knowledge area of interest, while the other candidates were less experienced. As to the content-based approach, we obtained the results shown in Table 1. In particular, the first two columns show data when the candidates were evaluated through the co-occurrence of terms and those occurrences were subsequently multiplied by the weight that entity obtains related to the subject, based on the ontology extracted from Wikipedia. The third column shows the results by comparing by means of the cosine-similarity the vectors weighed through a weighing based on the Log-Entity. The vector is weighed by the occurrences of the entities within the user's production and the product with the relevance value that given entity obtains with respect to the topic under consideration. The last columns show the results obtained with the Entity Frequency method while varying the reference domain, that is, taking into account the first n elements of the list of entities in descending order of frequency within the graph.

Table 1. Experimental results of the content-based approach

User	Occurr.	Weighed occurr.	Log-Entity	Domain elements				
				Total	Top40	Top30	Top20	Top10
u_1	0.63	0.66	0.72	0.58	0.73	0.73	0.71	0.67
u_2	0.59	0.61	0.68	0.80	0.87	0.87	0.90	0.93
u_3	0.59	0.62	0.73	0.67	0.78	0.77	0.81	0.88
u_4	0.63	0.64	0.69	0.50	0.58	0.57	0.55	0.47
u_5	0.43	0.46	0.62	0.67	0.78	0.77	0.79	0.87
u_6	0.29	0.30	0.21	0.41	0.53	0.52	0.57	0.68

Table 2 shows the results obtained for the same candidate users through the collaborative approach. Notice the maximum authority value obtained by a document produced by the candidate user, the sum of the authority values related to each document of the graph associated with the candidate user u, and the value given by Eq. 7 with $\lambda = 2$.

Table 2. Experimental results of the collaborative approach

User	$\lambda = 0.2$	Maximum authority value	Sum of authority values
u_1	0	0	0
u_2	0.16	0.10	0.39
u_3	0.12	0.10	0.20
u_4	0.02	0.01	0.03
u_5	0.01	0.01	0.02
u_6	0.02	0.01	0.06

The obtained data allow us to make some interesting observations. It can be noted that the content-based method considering the occurrences, whether not weighed or weighed by the relevance of the entities within the context, does not seem to produce results as expected. The Entity Frequency method, especially in its filtered version (i.e., based on the extraction of the top-n entities belonging to the domain), instead shows satisfactory results. The candidate users, which were assessed based on their generated content, were evaluated on their experience so to obtain positive values but differentiated, and the score gap between the expert users known to us and the other candidates is a faithful picture of the supposed accuracy of this method. Especially in the version with $n = 10$, the results show reliable values. Finally, the scores obtained through the collaborative approach show that the algorithm built through the HITS implementation performs rather trustworthy evaluations of expert candidates, but only if within the dataset (i.e., the graph built on the co-citations among the different documents) the expert candidate u's documents were found. For instance, the collaborative approach was not able to assign a value to the candidate u_1's expertise, which is therefore set equal to 0.

3 Conclusions

In this article, we have described a system for the implicit assessment of a user's expertise in a specific knowledge area. The development of two main approaches allows us to choose between one or both of them, thus enabling the system to overcome their individual weaknesses. The experimental results show that in some situations the content-based approach can be better, in others the collaborative one is to be preferred. Hence, the best results may come from an integrated solution. The heterogeneous structure of the graph database chosen for the system implementation actually enables complex queries to be satisfied based on the different stored information.

Among the possible future developments, we would like to increase the number of knowledge bases (i.e., available documents) to enhance the reliability of the system output. As for the experimental evaluation, we plan to test our system on other domains and allow testers to provide explicit feedbacks on the received results.

References

1. Arru, G., Feltoni Gurini, D., Gasparetti, F., Micarelli, A., Sansonetti, G.: Signal-based user recommendation on Twitter. In: Proceedings of 22nd International Conference on World Wide Web, WWW 2013 Companion, pp. 941–944. ACM, New York (2013). http://doi.acm.org/10.1145/2487788.2488088
2. Balog, K., Fang, Y., de Rijke, M., Serdyukov, P., Si, L.: Expertise retrieval. Found. Trends Inf. Retr. **6**(2–3), 127–256 (2012). http://dx.doi.org/10.1561/1500000024
3. Biancalana, C., Gasparetti, F., Micarelli, A., Miola, A., Sansonetti, G.: Wavelet-based music recommendation. In: WEBIST 2012 - Proceedings of 8th International Conference on Web Information Systems and Technologies, pp. 399–402 (2012)

4. Biancalana, C., Flamini, A., Gasparetti, F., Micarelli, A., Millevolte, S., Sansonetti, G.: Enhancing traditional local search recommendations with context-awareness. In: Konstan, J.A., Conejo, R., Marzo, J.L., Oliver, N. (eds.) UMAP 2011. LNCS, vol. 6787, pp. 335–340. Springer, Heidelberg (2011). doi:10.1007/978-3-642-22362-4_29. http://dl.acm.org/citation.cfm?id=2021855.2021886
5. Biancalana, C., Gasparetti, F., Micarelli, A., Miola, A., Sansonetti, G.: Context-aware movie recommendation based on signal processing and machine learning. In: Proceedings of 2nd Challenge on Context-Aware Movie Recommendation, CAMRa 2011, pp. 5–10. ACM, New York (2011). http://doi.acm.org/10.1145/2096112.2096114
6. Biancalana, C., Gasparetti, F., Micarelli, A., Sansonetti, G.: An approach to social recommendation for context-aware mobile services. ACM Trans. Intell. Syst. Technol. **4**(1), 10:1–10:31 (2013). http://doi.acm.org/10.1145/2414425.2414435
7. Biancalana, C., Gasparetti, F., Micarelli, A., Sansonetti, G.: Social semantic query expansion. ACM Trans. Intell. Syst. Technol. **4**(4), 60:1–60:43 (2013). http://doi.acm.org/10.1145/2508037.2508041
8. Bologna, C., De Rosa, A., De Vivo, A., Gaeta, M., Sansonetti, G., Viserta, V.: Personality-based recommendation in e-commerce. In: CEUR Workshop Proceedings, vol. 997 (2013). Cited by 6
9. Caldarelli, S., Gurini, D., Micarelli, A., Sansonetti, G.: A signal-based approach to news recommendation. In: CEUR Workshop Proceedings, vol. 1618 (2016)
10. D'Agostino, D., Gasparetti, F., Micarelli, A., Sansonetti, G.: A social context-aware recommender of itineraries between relevant points of interest. In: Stephanidis, C. (ed.) HCI 2016. CCIS, vol. 618, pp. 354–359. Springer, Cham (2016). doi:10.1007/978-3-319-40542-1_58
11. Gurini, D.F., Gasparetti, F., Micarelli, A., Sansonetti, G.: iSCUR: interest and sentiment-based community detection for user recommendation on Twitter. In: Dimitrova, V., Kuflik, T., Chin, D., Ricci, F., Dolog, P., Houben, G.-J. (eds.) UMAP 2014. LNCS, vol. 8538, pp. 314–319. Springer, Cham (2014). doi:10.1007/978-3-319-08786-3_27
12. Gurini, D.F., Gasparetti, F., Micarelli, A., Sansonetti, G.: Temporal people-to-people recommendation on social networks with sentiment-based matrix factorization. Future Gener. Comput. Syst. (2017). http://www.sciencedirect.com/science/article/pii/S0167739X17304077
13. Gurini, D., Gasparetti, F., Micarelli, A., Sansonetti, G.: A sentiment-based approach to twitter user recommendation. In: CEUR Workshop Proceedings, vol. 1066 (2013)
14. Gurini, D., Gasparetti, F., Micarelli, A., Sansonetti, G.: Analysis of sentiment communities in online networks. In: CEUR Workshop Proceedings, vol. 1421, pp. 17–20 (2015)
15. Feltoni Gurini, D., Gasparetti, F., Micarelli, A., Sansonetti, G.: Enhancing social recommendation with sentiment communities. In: Wang, J., Cellary, W., Wang, D., Wang, H., Chen, S.-C., Li, T., Zhang, Y. (eds.) WISE 2015. LNCS (LNAI and LNBI), vol. 9419, pp. 308–315. Springer, Cham (2015). doi:10.1007/978-3-319-26187-4_28
16. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM **46**(5), 604–632 (1999). http://doi.acm.org/10.1145/324133.324140
17. Onori, M., Micarelli, A., Sansonetti, G.: A comparative analysis of personality-based music recommender systems. In: CEUR Workshop Proceedings, vol. 1680, pp. 55–59 (2016)