Recommending Multimedia Educational Resources on the MOVING Platform

Iacopo Vagliano[1[0000−0002−3066−9464]] and Sibgha Nazir[2]

[1] ZBW – Leibniz Information Centre for Economics, Germany
i.vagliano@zbw.eu
[2] Technical University of Darmstadt, Germany
sibgha.nazir@stud.tu-darmstadt.de

Abstract. The MOVING platform includes a huge amount of heterogeneous educational resources, such as documents, videos, and social media posts. We show how the MOVING recommender system can support users in dealing with such a massive information flow by leveraging semantic profiling. The HCF-IDF model exploits a thesaurus or ontology to represents users and documents and it is used to recommend educational resources based on users' search history. We describe how the recommender is implemented how it is applied to the MOVING platform to deal with the huge amount of resources stored in the platform, their variety and the increasing number of users.

Keywords: Recommender systems · Semantic profiling · Technology-enhanced learning · Multimedia content recommendation

1 Introduction

Nowadays, much more information is produced that what we can actually consume. This issue is known as information overload and affects all information professional, such as researchers and students, which daily deal with an enormous amount of information. Recommender systems are tools to suggest interesting items to users, such as songs, movies, products, etc., that can address the information overload by enabling users to shift from searching to discovering.

The MOVING platform[3] enables its users to improve their information literacy by training how to exploit data and text mining methods in their daily research tasks [23, 24]. It integrates a vast amount of educational resources which are of various kinds, such as documents, videos, and social media data. Some of these resources are automatically harvested from the Web and social networks. Through the platform, users can search these resources and display the search results in different ways thanks to the advanced visualizations available. One of its components is a recommender system which suggests possibly interesting educational resources. It takes into account all the various kinds of resources in the MOVING platform, including videos, documents, and social media posts.

[3] http://platform.moving-project.eu

In this paper, we focus on the MOVING recommender system. We show how it exploits semantic profiling of users and documents to provide useful suggestions through the HCF-IDF model, an approach that exploits a thesaurus or ontology to represents users and documents. We also describe how the recommender is implemented in the MOVING platform. While the HCF-IDF model was previously presented, in this paper we show how it is applied to the MOVING platform to deal with the huge amount of resources stored in the platform, their variety and the increasing number of users.

The rest of the paper is organized as follows: in Section 2 we briefly review the state of the art in educational and semantic-aware recommender systems; in Section 4 we describe the semantic profiling method used; in Section 5 we outline how this method is applied in the MOVING platform; we conclude in Section 7.

2 Related Work

2.1 Educational recommender systems

In the MOVING platform, we recommend multimedia resources for educational purposes. We briefly recall the main studies in the area. Manouselis et al. presented an extensive discussion of research educational recommender systems [16]. In this field, a lot of works focus on the recommendation of research papers; these studies have been discussed by Beel et al. [2]. As an example of work in the educational domain, Docear [3] provides various features for scientists including a recommender system. Another popular educational recommender system is BibTip [11], while CiteSeer [4] is a well-known recommender system for scientific papers. More recently, works that rely on deep learning to suggest citations or subject labels are emerging [10]. While these works usually rely on publications or clicks for the user profiles and can be broadly classified as collaborative filtering approaches [22], the MOVING recommender system exploits the users' search history and it belongs to the content-based techniques [22]. While the latter can provide less diverse recommendations, the first is more subjected to the cold-start problem: recommending resources is challenging in case of a new user or a new item (no clicks available). The MOVING recommender system still has the new user problem as no or too few searches are available, but is not affected by the new item since only the content is used not its clicks. In addition, the diversity of recommendations is increased by the use of semantics [25, 26].

2.2 Semantic-aware recommender systems

The MOVING recommender system relies on HCF-IDF [18], which can be considered a semantic-aware recommender system. A survey on similar systems is available in the literature [9]. Typically, these recommender systems consider the relationships among resources by taking into account the semantic similarity of the resources. Below, we summarize the main works.

Some studies rely on the interlinking of resources. Damljanovic et al. [7] suggested domain experts in an open innovation scenario by discovering related resources through hierarchical or transversal relationships. Passant [21] presented a music recommender system, which relies on the number of direct and indirect links between two resources. ReDyAl [25] exploits existing relationships between resources by dynamically analyzing both their categories and their explicit references to other resources. SemRevRec [26] combines semantic annotation of user reviews with additional information from the Web. Musto et al. [17] studied the impact of the knowledge available in the Web on the overall performance of a graph-based recommendation algorithm. Karpus et al. [15] presented a context-aware recommender system based on a semantic representation of the user context.

Other works combine semantic relationships with machine learning. Heitmann and Hayes [13] proposed a semantic-aware recommender system to mitigate the new-user, new-item and sparsity problems of collaborative recommender systems. SPrank [19] extracts semantic path-based features and computes recommendations using Learning to Rank. Techniques that combine semantics with deep learning are also emerging [20]. CF-IDF is as an extension of TF-IDF that counts frequencies of concepts instead of terms [12]. HCF-IDF [18] extends CF-IDF by combining its statistical strength with semantics provided by a thesaurus. Specifically, it can exploit the hierarchical relationships among the concepts in the thesaurus. This model is described more in detail in Section 4.

HCF-IDF has been selected as a reference method for the MOVING platform because has proved to be effective also when the full-text of the resources to recommend is not available, as described in Section 6. In the platform, this may happen because of legal reasons or due to the type of data, e. g. for videos sometimes the transcript is available, sometimes not. In addition, the use of a thesaurus can increase the diversity of recommendations, which can be an issue for content-based methods. In this paper, we show how HCF-IDF is implemented in the MOVING platform to deal with the huge amount of resources stored in the platform, their variety and the increasing number of users.

3 The MOVING platform

The MOVING platform provides access to a great variety and amount of educational resources, such as documents, videos, and social media data. Some of these resources are automatically harvested from the Web and social networks, while others are manually added by the administrators from domain-specific sources, e.g. VideoLectures.NET[4], EconBiz[5], and the Social Science Open Access Repository (SSOAR)[6]. Through the platform, users can search these resources and display the search results in different ways thanks to the advanced visualizations available.

[4] http://videolectures.net/
[5] https://www.econbiz.de/
[6] https://www.gesis.org/ssoar/home/

The architecture of the platform is depicted in Figure 1. The crawlers automatically ingest data from the Web. Data processing techniques, including author disambiguation, automatic concepts annotation, data deduplication, and entity extraction, generate additional information for the index. The search engine allows users to efficiently retrieve the indexed data. WevQuery [1] tracks the users' behavior on the platform by capturing UI events, while the Adaptive Training Support (ATS) [8] analyses the logged data to support users to improve their use of the platform and progress in the selected curriculum based on their usage patterns.

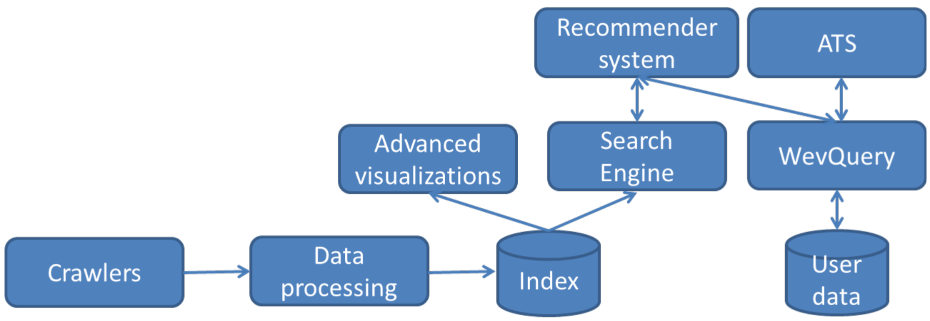

Fig. 1. The MOVING platform at a glance [24].

The recommender system interacts with both the search engine and WevQuery. To build users' profiles based on their search history, it obtains the search history from the user data previously logged through WevQuery, and then it retrieves in the index the documents to suggest depending on the user's profile.

4 The HCF-IDF model

The MOVING recommender system is based on HCF-IDF [18], a novel semantic profiling approach which can exploit a thesaurus or ontology to provide better recommendations. In this section, we recall its main features, while in Section 5 we explain how it is employed in the MOVING platform. The HCF-IDF method extends CF-IDF [12], which in turn is an extension of the classical TF-IDF model. After formalizing the recommendation problem (Section 4.1) , we describe the CF-IDF model (Section 4.2) , then HCF-IDF (Section 4.3).

4.1 Problem statement

Given a set of m documents $\mathbb{D}$ and a set of n users $\mathbb{U}$, the typical recommendation task is to model the spanned space, $\mathbb{U} \times \mathbb{D}$. With documents, we intend the multimedia resources available in the MOVING platform, i. e. textual documents (e. g.

articles, books, regulations), videos and social media data. We model our recommendation problem as the Top-N recommendation problem [5]. Specifically, the goal is returning the set of Top-N documents which have the highest similarity with a user u_i, for each user $u_i \in U$. Typically users and documents represented with user and document profiles, respectively. In our case, user profiles are sets of terms previously searched by the user ordered by time and frequency of search (more details are provided in Section 5.2), while documents profiles consist of concepts preassigned to the documents. As the platform is integrating documents from various sources these concepts can be automatically generated or manually assigned by domain experts (e.g. in the case of data from EconBiz[7]).

4.2 CF-IDF

In contrast to TF-IDF, the CF-IDF model [12] substitutes the term frequency with the frequency of semantic concepts. Each concept is uniquely identified by a URI and has one or more labels to describe it.[8] For instance, the concept *Innovation management* is represented in the STW thesaurus[9], which describes the Economics domain. The concept has various labels that indicates synonyms, such as *Innovation strategy* and *Technology management*. The advantage of CF-IDF is exploiting concepts to handle such synonyms. For example, if in a text the terms *Innovation management* and *Innovation strategy* are used once TF-IDF considers them different and assign each a frequency equal to 1, while CF-IDF refers to the concept frequency of 2 by computing the sum of the label frequencies.

More formally, the weight assigned to concepts by CF-IDF is described in Equation 1 [12], where $n_{i,j}$ is the occurrence of a concept c_i in a document d_j, and $\sum_k n_{k,j}$ is the total number of occurrences of all concepts in the document d_j. $|D|$ is the total number of documents, while $|\{d : c_i \in d\}|$ counts the documents in which the concept c_i appears.

$$w_{cf-idf} = \frac{n_{i,j}}{\sum_{c_k \in d_j} n_{k,j}} \cdot \log \frac{|D|}{|\{d : c_i \in d\}|} \tag{1}$$

4.3 HCF-IDF

HCF-IDF [18] further improves CF-IDF by taking into account the hierarchy of concepts. This enables the model to consider related concepts not directly mentioned in a text. To do so, it applies spreading activation [6] over a given concept tree and through the IDF component it prevents very generic concepts accounts for high weights.

As an example, if a user profile includes the concept *Open innovation* and assuming the latter is a sub concept of *Innovation management* which in turn

[7] https://www.econbiz.de/
[8] https://www.w3.org/DesignIssues/LinkedData.html
[9] http://zbw.eu/stw/

is a sub concept of *Management*, then HCF-IDF assigns non-zero weights to the concepts *Innovation management* and *Management*, even if they are not directly mentioned in the document. In this way, if *Innovation management* is part of the user profile, then also the documents related to *Open innovation* can be recommended. Similarly, if *Open innovation* is part of the user profile, then also the documents concerning *Innovation management* can be suggested. This helps to the system to generate more diverse recommendations since documents not directly related to the user profile but still relevant to it are considered.

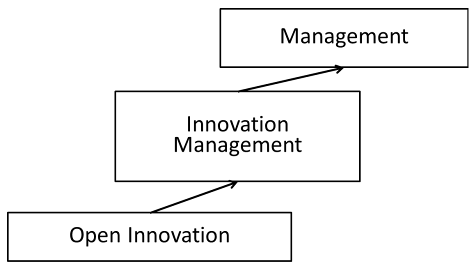

Fig. 2. A concept tree.

The weights in HCF-IDF are computed as defined in Equation 2 [18], where $BL(c,d)$ is the BellLog spreading-activation function [14], which is described in Equation 3. The function $h(c)$ returns the level where a concept c is located in the concept tree, while *nodes* counts the concepts at a given level in the tree. For example, with the tree showed in Figure 2, $h(\textit{Innovation management})$ returns 2 and $nodes(h(\textit{Innovation management})+1)$ returns 1. C_l is the set of concepts located in one level lower than the concept c considered. Referring to Figure 2, C_l is equal to $\{\textit{Open innovation}\}$ for *Innovation management*.

$$w_{hcf-idf} = BL\,(c,d) \cdot \log \frac{|D|}{|\{d : c_i \in d\}|} \tag{2}$$

$$BL\,(c,d) = \frac{n_{i,j}}{\sum_{c_k \in d_j} n_{k,j}} + \frac{1}{\log_{10}(nodes(h(c)+1))} \cdot \sum_{c_j \in C_l} BL(c_j,d) \tag{3}$$

While CF-IDF has outperformed TF-IDF [12], previous work has shown that HCF-IDF can achieve similar results to CF-IDF by only relying on the titles of the publications and not on the full-texts [18]. We chose HCF-IDF as a reference method because the MOVING platform does not usually store the documents directly but only their metadata due to license issues. To access the full-texts not stored, users are redirected to the original data provider.

5 Recommending resources in the MOVING platform

5.1 Scenario

Andrea is a student in Computer Science at Kiel University, she is using the MOVING platform to find additional learning material for the courses she is currently attending: *Processing and transmission of multimedia information*, *Machine Learning*, and *Software Architectures*. She has previously searched for terms related to these courses, such as *Multimedia Analysis*, *Deep Learning*, *RESTful services*, etc.. When she logs in again into the platform, the MOVING recommender systems suggest her other resources which may be interesting for her, as shown in Figure 3. These suggestions depend on her previous search history, which can be an estimation of her interest. Additionally, different kinds of resources are recommended: a video, a book, and a Web page. In this way, she can find useful resources even before to type a search query. If she is not interested in the recommended items, she can search for other documents in the platform.

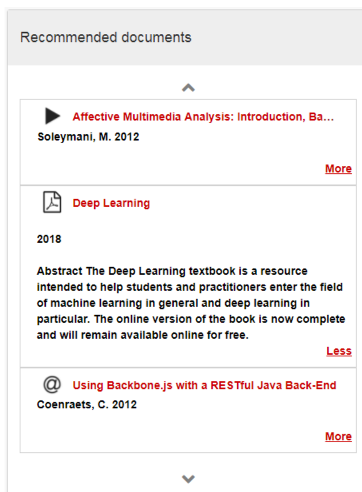

Fig. 3. Recommended resources of various types: a video, a book, and a Web page.

The recommender system widget, depicted in Figure 3, is part of the search page of the MOVING platform, as illustrated in Figure 4. Thanks to it, users can receive additional suggestions possibly discovering useful resources of which they were not aware.

5.2 Building user profiles

The user profile is a group of information that best describes a given user. In our case, it is the history of searches performed with the MOVING platform.

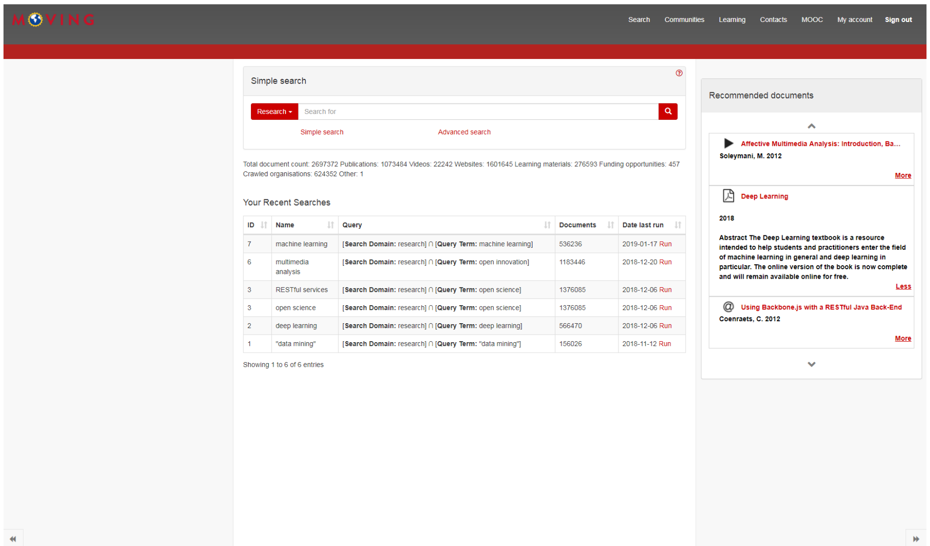

Fig. 4. The search Page of the MOVING platform. The recommender system widget is displayed on the right.

Every term has a weight associated, which depends on how many times and how recently the user has looked for a term. More formally, the weight w of a term k is defined as $w = \alpha_t \cdot f_t + \alpha_h \cdot f_h$. The time coefficient, α_t, and the hit coefficient, α_h, weight the time and frequency of each term in the profile. The time factor of a term, f_t, is the timestamp (t) of its last search, normalized by the current time (T): $f_t = \frac{t}{T}$. The hit factor of a term, f_h, is the number of times the term has been looked up by the user (h) divided by the total number of searches made by the user (H), $f_h = \frac{h}{H}$. The user profile is a set of pairs term-weight, $\langle k_i, w_i \rangle$, where k_i is a term and w_i a weight.

The HCF-IDF method has been tested in the Economics domain with a user study [18] (see Section 6). By means of an informal evaluation of the recommender system in the MOVING platform, we set both α_t and α_h to 0.5 and we decided to limit the user profile to the top 25 terms, as considering more terms does not significantly improve the recommendations while increases the response time. This last parameter can also be configured, similarly to α_t and α_h.

Additionally to the searches, further users' interactions could be taken into account when building the user profile. In our case, the problem is that, while the user profile is usually a collection of suggested items and corresponding explicit (e.g. ratings) or implicit (e.g. clicks or downloads) user feedback, HCF-IDF needs a collection of term-weight pairs. One possible solution to this problem would be adding to the user profile all the concepts of a suggested document when clicked. If a term of a clicked document already belongs to the user profile, its weight should be updated considering the new click on the corresponding document. However, this solution should be further investigated.

5.3 Implementation

In the MOVING platform, a search engine allows users to search the data indexed, while WevQuery [1] tracks the users behavior on the platform by capturing UI events. The recommender system interacts with both the search engine and WevQuery. To build users' profiles based on their search history, it obtains the search history from the user data previously logged through WevQuery, and then it retrieves in the index the documents to suggest based on the user's profile. After building the user profile, it sorts the terms based on their weights in descending order and appends them in a space-separated string to build a query to generate the list of recommendations through the search engine, using HCF-IDF. The search engine is based on Elasticsearch[10]. We have implemented HCF-IDF as an Elasticsearch plugin.

We implemented the recommender system as a RESTful web service. An HTTP `GET /recommendations` issues the execution of the `get_recommendations` method, which serves the request taking a `user_id` as an argument, and returns the list of recommendations in the JSON format. For building the user profile, we use the information stored in WevQuery: the recommender system retrieves all the searches made by the user with the specified `user_id` through the WevQuery web API. The user profile is sent to the HCF-IDF plugin, which generates the list of recommendations.

6 Evaluation

HCF-IDF has previously been evaluated with a user study with 123 participants [18]. The goal was to identify the best strategy for a recommender system along three factors: profiling method, decay function, and document content. It has been compared against eleven other methods, i. e. twelve approaches have been tested.

The results showed that HCF-IDF was the most effective profiling method. Overall, the best performing approach was CF-IDF relying on sliding windows and using both titles and full-texts. However, using only the titles of scientific publications this method achieved competitive recommendation results with full-texts. Thus, the spreading activation over the thesaurus enables HCF-IDF to extract a sufficient number of concepts from titles to compute competitive recommendations. This is an important result in the context of the MOVING platform as full-texts are not always available due to legal barriers or to the type of data (e.g. videos). That is why the HCF-IDF has been chosen as a reference method in the MOVING platform. A more detailed description of the experiments is available elsewhere [18].

Currently, an online evaluation is planned with the users of the MOVING MOOC on Science 2.0 and open research methods[11] where the main approaches compared in the user study (TDF-IDF, CF-IDF, and HCF-IDF) are going to

[10] https://www.elastic.co/products/elasticsearch
[11] https://moving.mz.tu-dresden.de/mooc

be evaluated in an online experiment to cross-check the outcome of the user study and also to consider further aspects related to the MOVING recommender system. These aspects include the placement of the recommender widget in the MOVING search page, the widget's user interface, and optimizing the HCF-IDF parameters. This online evaluation is possible because the MOVING platform tracks users interactions through WevQuery [1]. For instance, clicks and mouse-hovering events on recommended items displayed in the recommender widget are mapped to the corresponding resource represented.

7 Conclusions

We showed how the MOVING recommender system can help the MOVING platform's users in dealing with the huge amount of information stored. It enables users to discover useful resources by leveraging semantic user and document profiling through the HCF-IDF recommendation method. The recommender system is publicly available in the MOVING platform[12].

As future work, we intend to take into account in the user profile other users' interactions in addition to their searches, such as clicks on the suggested items.

Acknowledgments This work was supported by the EU's Horizon 2020 programme under grant agreement H2020-693092 MOVING.

References

1. Apaolaza, A., Vigo, M.: WevQuery: Testing hypotheses about web interaction patterns. Proc. ACM Hum.-Comput. Interact. **1**(EICS), 4:1–4:17 (Jun 2017)
2. Beel, J., Gipp, B., Langer, S., Breitinger, C.: Research-paper recommender systems: a literature survey. International Journal on Digital Libraries **17**(4), 305–338 (2016)
3. Beel, J., Langer, S., Gipp, B., Nürnberger, A.: The architecture and datasets of docear's research paper recommender system. D-Lib Magazine **20**(11/12) (2014)
4. Bollacker, K.D., Lawrence, S., Giles, C.L.: Discovering relevant scientific literature on the web. IEEE Int. Systems and their Applications **15**(2), 42–47 (March 2000)
5. Cremonesi, P., Koren, Y., Turrin, R.: Performance of recommender algorithms on top-n recommendation tasks. In: RecSys '10. pp. 39–46. ACM (2010)
6. Crestani, F.: Application of spreading activation techniques in information retrieval. Artificial Intelligence Review **11**(6), 453–482 (1997)
7. Damljanovic, D., Stankovic, M., Laublet, P.: Linked data-based concept recommendation: Comparison of different methods in open innovation scenario. In: ESWC. pp. 24–38. Springer (2012)
8. Fessl, A., Wertner, A., Pammer-Schindler, V.: Digging for gold: Motivating users to explore alternative search interfaces. In: Lifelong Technology-Enhanced Learning. pp. 636–639. Springer International Publishing, Cham (2018)
9. Figueroa, C., Vagliano, I., Rocha, O.R., Morisio, M.: A systematic literature review of linked data-based recommender systems. Concurr. Comput. : Pract. Exper. **27**(17), 4659–4684 (2015)

[12] http://platform.moving-project.eu

10. Galke, L., Mai, F., Vagliano, I., Scherp, A.: Multi-modal adversarial autoencoders for recommendations of citations and subject labels. In: UMAP '18. pp. 197–205. ACM (2018)
11. Geyer-Schulz, A., Hahsler, M., Jahn, M.: Recommendations for virtual universities from observed user behavior. In: Gaul, W., Ritter, G. (eds.) Classification, Automation, and New Media. pp. 273–280. Springer Berlin Heidelberg (2002)
12. Goossen, F., Ijntema, W., Frasincar, F., Hogenboom, F., Kaymak, U.: News personalization using the CF-IDF semantic recommender. In: WIMS '11. pp. 10:1–10:12. ACM (2011)
13. Heitmann, B., Hayes, C.: Using linked data to build open, collaborative recommender systems. In: AAAI Spring Symposium. pp. 76–81. AAAI (2010)
14. Kapanipathi, P., Jain, P., Venkataramani, C., Sheth, A.: User interests identification on Twitter using a hierarchical knowledge base. In: ESWC. Springer (2014)
15. Karpus, A., Vagliano, I., Goczya, K., Morisio, M.: An ontology-based contextual pre-filtering technique for recommender systems. In: 2016 Federated Conference on Computer Science and Information Systems (FedCSIS). pp. 411–420 (Sep 2016)
16. Manouselis, N., Drachsler, H., Vuorikari, R., Hummel, H., Koper, R.: Recommender Systems in Technology Enhanced Learning, pp. 387–415. Springer US (2011)
17. Musto, C., Lops, P., Basile, P., de Gemmis, M., Semeraro, G.: Semantics-aware graph-based recommender systems exploiting linked open data. In: UMAP '16. pp. 229–237. ACM (2016)
18. Nishioka, C., Scherp, A.: Profiling vs. time vs. content: What does matter for top-k publication recommendation based on twitter profiles? In: JCDL '16. pp. 171–180. ACM (2016)
19. Noia, T.D., Ostuni, V.C., Tomeo, P., Sciascio, E.D.: SPrank: Semantic path-based ranking for top-n recommendations using linked open data. ACM Trans. Intell. Syst. Technol. **8**(1), 9:1–9:34 (2016)
20. Palumbo, E., Rizzo, G., Troncy, R.: entity2rec: Learning user-item relatedness from knowledge graphs for top-n item recommendation. In: RecSys. pp. 32–36. ACM (2017)
21. Passant, A.: dbrec — music recommendations using DBpedia. In: ISWC '10. ISWC 2010, vol. 2, pp. 209–224. Springer Berlin Heidelberg (2010)
22. Ricci, F., Rokach, L., Shapira, B.: Recommender Systems: Introduction and Challenges, pp. 1–34. Springer US, Boston, MA (2015)
23. Vagliano, I., Günther, F., Heinz, M., Apaolaza, A., Bienia, I., Breitfuss, G., Blume, T., Collyda, C., Fessl, A., Gottfried, S., Hasitschka, P., Kellermann, J., Khler, T., Maas, A., Mezaris, V., Saleh, A., Skulimowski, A.M.J., Thalmann, S., Vigo, M., Wertner, A., Wiese, M., Scherp, A.: Open innovation in the big data era with the moving platform. IEEE MultiMedia **25**(3), 8–21 (July 2018)
24. Vagliano, I., Fessl, A., Günther, F., Köhler, T., Mezaris, V., Saleh, A., Scherp, A., Šimić, I.: Training researchers with the moving platform. In: Kompatsiaris, I., Huet, B., Mezaris, V., Gurrin, C., Cheng, W.H., Vrochidis, S. (eds.) MultiMedia Modeling. pp. 560–565. Springer International Publishing, Cham (2019)
25. Vagliano, I., Figueroa, C., Rodríguez Rocha, O., Torchiano, M., Faron Zucker, C., Morisio, M.: ReDyAl: A Dynamic Recommendation Algorithm based on Linked Data. In: CB-RecSys '16 co-located with ACM RecSys. vol. 1673. CEUR (2016)
26. Vagliano, I., Monti, D., Scherp, A., Morisio, M.: Content recommendation through semantic annotation of user reviews and linked data. In: K-CAP. pp. 32:1–32:4. ACM (2017)