# Exploring the Use of Linked Open Data for User Research Interest Modeling

Rubén Manrique, Omar Herazo, and Olga Mariño(✉)

Systems and Computing Engineering Department, School of Engineering, Universidad de los Andes, Bogotá, Colombia
{rf.manrique,oa.herazo3009,olmarino}@uniandes.edu.co

**Abstract.** In the context of the Social Web, user' profiles reflecting an individual's interests are being modeled using semantic techniques that consider the users posts' and take advantage of the rich background knowledge in a Linked Open Dataset (LOD). To enrich the user profile, expansion strategies are applied. While these strategies are useful in Social Network posts, their suitability for modeling users' interests with larger documents as input has not yet been validated. Thus, we built a profile of user's research interests to recommend academic documents of possible interest. Contrary to the results obtained in the Social Web, the expansion techniques are inadequate for the academic texts scenario when all of text in the documents are used as input. Our results show a new filtering strategy performs better in such a scenario. An additional contribution was our creation of a DBpedia annotated dataset for academic document recommendation, which was built from a corpus of open access papers available through Core and Arxiv. Findings suggest the need to further explore new strategies to construct semantic models that are able to operate in different domains.

**Keywords:** User modeling · Linked Open Data · Semantic web

## 1 Introduction

With the advent of the Semantic Web and the Linked Open Data (LOD) cloud, new ways of semantically representing users have been proposed [1,8,14,16]. In these new modeling approaches, the user profile is created using a set of entities in the LOD cloud that are discovered from information collected about the user. The advantage of using such representations is the additional knowledge that can be gathered about the entities and the relationships between them (i.e. background knowledge). Using this information, it is possible to extend the user profile and to infer previously unknown interests. Additionally, the LOD cloud provides a set of comprehensive datasets with domain-independent capabilities like DBPedia, which support user modeling in different contexts.

Although considerable research has been devoted to evaluating different LOD user modeling strategies from Twitter and Facebook content [2,14,17], this Social Web based approach has some limitations. On one hand, it is difficult to access


A. Solano and H. Ordoñez (Eds.): CCC 2017, CCIS 735, pp. 3–16, 2017.
DOI: 10.1007/978-3-319-66562-7_1

the information produced by users in social networks, either because the accounts are no longer active or the user refuses to give access to their accounts. On the other hand, since most Social Web contents produced by a user express interests associated with short term events like news, natural disasters, sports matches, political debates, etc., user profiles based on this information may only reflect short-term or fleeting interests, thus overlooking more lasting interests such as research and work.

This paper addresses the problem of building an accurate LOD user profile nurtured by the user's consumption and production of digital documents, namely academic documents. This document-based approach presents a major challenge due to increased length and complexity compared with social networks. On the Social Networks, published content is limited in the number of words or characters per post, so it is more likely that the user will express only information of interest [5]. That is to say, due to concision of the message and the need for it to be self-contained, it is more likely that entities discovered in posts are of interest to the user. Thus, it is expected that extending the profile using the rich knowledge of LOD datasets will enhance the user profile without creating significant drift from the user's interests. However, in longer documents like academic texts, it is likely that not all the entities found will represent the real interests of the user. Extending the profile through entities in less succinct documents could generate important noise.

In this paper, we are interested in creating LOD research interests profiles based on a set of academic documents, and we present a more reliable strategy for building these user profiles. This began by constructing the base semantic user profile starting with concepts identified in the publication list of each user. Building on this, strategies that have been successfully used in Social Networks to expand the profile by using the background knowledge found in a linked dataset were applied. We evaluated these strategies in recommending relevant academic documents of potential interest to the user. Recommender systems have the ability to predict whether a user would prefer an item or not based on a user profile. In order to evaluate the effectiveness of different user profiles at predicting users' research interests, we built an academic document recommendation engine. Our evaluation suggests that expansion strategies do not necessarily improve the user profile; indeed, some of them degrade the user profile quality. Hence, instead of an expansion strategy, a filtering strategy that prunes the resources and leaves only those that are related to each other in the KB was proposed. This strategy proved to be better suited than expansion for modeling the user's lasting interests in the context of academic research.

In the process of arriving to this final filtering strategy, we also made the following contributions to the work on this topic. These began with creating a DBpedia annotated dataset for academic document recommendation, which was built from a corpus of open access papers available through Core[1] and Arxiv[2] services. To the best of our knowledge, no other academic recommendation dataset

[1] https://core.ac.uk/.
[2] https://arxiv.org/.

with these characteristics exists. The second contribution is the evaluation of semantic modeling techniques in a non Social Web scenario. Other discourse communities would likely benefit from the filtering strategy provided.

The paper is organized as follows: In Sect. 2, we review related work in user interest modeling and recent research in content based academic paper recommendation. In Sect. 3, we present the semantic profiling process and the diverse expansion and filtering strategies. In Sect. 4, we describe the protocol used to build the dataset and the evaluation framework. Results and conclusions are discussed in Sects. 5 and 6 respectively.

## 2 Related Work

A user model or profile[3] is a representation of information about an individual [9] used to personalize applications. Different kinds of information about the individual could be part of the profile; however, most user profiles in retrieval and recommendation systems are based on the user's interests [18]. The most common representation of the user's interests is the keyword-based representation [18]. In this type of profile, users are represented as a weighted vector of words. The weights signify the importance of the term for the user, and they are implicitly calculated from the input content (i.e. documents or posts from which it is possible to infer the user's interest). Weighting schemas such as the word frequency and the TF-IDF (term frequency/inverse document frequency) have been extensively used [4,10,19]. The disadvantage of this representation is that it cannot provide additional information about the semantic relationships of the entities or concepts present in the text.

More recent approaches [2,16,17] have focused on representing the user as a bag of concepts where a concept is any kind of entity that has an explicit representation in a Knowledge Base (KB). In this context, LOD can be used as KB. Indeed, the current web of data offers a large set of linked semantic datasets encoded in machine-understandable RDF standard. They provide excellent modeling capabilities thanks to their cross-domain vocabularies. DBpedia, one of the main datasets in the LOD initiative, for example, currently describes 6 million entities, 1.7 million categories and close to 5 billion facts in its English version alone.

Different research has been evaluated in the context of the Social Web. Abel et al. [2] explore the use of LOD to build user profiles that improve link recommendation on Twitter. They also explore expansion strategies, which they call "indirect mentions", that take advantage of the rich background knowledge in DBpedia. Expansion strategies are better at recommendation. Orlandi et al. [14] follow a similar approach and compute user profiles by harvesting the user's posts in Twitter and Facebook. They propose two representation models based on DBpedia: one based on the entities found in the text, and the other on these

[3] Although some authors distinguish between a user model and a user profile [12], we will use both terms interchangeably.

entities and their categorical information. No significant differences between the two were reported.

Recent work by Piao and Breslin [16,17], compare diverse semantic modeling and expansion strategies for Twitter link recommendation. First, they show the superior behaviour of the Inverse Document Frequency (IDF) strategy as a weighting scheme for concept-based profiles. Second, they compare three strategies for user profile expansion using DBpedia: class-based, category-based and property-based. Class and category-based strategies incorporate the DBpedia classes and categories of the initial entities found in the user's Tweets into the profiles. Property expansion includes other entities connected to primitive interests through properties in DBpedia Ontology. According to the results obtained, categorical and property expansion have superior expansion capabilities.

We did not find any studies evaluating LOD user profiling techniques outside of the Social Web even though these profiles are frequently used in LOD recommender systems. Our literature review reveals that recent contributions to LOD recommender systems are more focused on the recommendation algorithm than on the user profile [7,8,13]. Moreover, the current LOD recommenders work almost exclusively on domains where there is already a direct map between the recommendation object and a concept in the KB. For example, the ESWC 2014 Challenge [6] worked on books that were mapped to their corresponding DBpedia resource. Similarly, DiNoia et al. [8] reduce the MoviLens dataset to those movies that have a corresponding DBpedia resource. Our recommendation approach is different from those because we are not limited to a candidate set in which each item has a direct resource that represents it in the LOD. Instead, we address the problem by taking the textual information of documents as input and identifying the set of concepts present.

## 3 Semantic Profiling Process

This section presents the process for building the different semantic user profiles, and takes into account semantic information recovered from a KB in the Linked Open Data Cloud. The process involves four different modules (Fig. 1). The first one, called the Semantic Document Annotator, receives a document and identifies entities of the KB in the text. We will use the word annotations to refer to these entities. The set of annotations found constitutes a initial semantic user profile ($ISUP$). Then, the Expansion Module receives the initial profile and expands it through the rich number of relationships in the KB. In this module, new expanded concepts that are not found in the text, but are related with the annotation, are incorporated into the user profile. In the Filtering Module, we apply our proposed filtering technique to select concepts that are highly connected. The strategy looks for connection paths between annotations as it uses these to select and assign the concept's initial weight in the user profile representation. Finally, the Weighting Module checks the importance of each concept in the interest profile and assigns a weight accordingly. We use an IDF approach to assign the annotation weights [17], and different discount strategies to assign the weights to the expanded concepts.

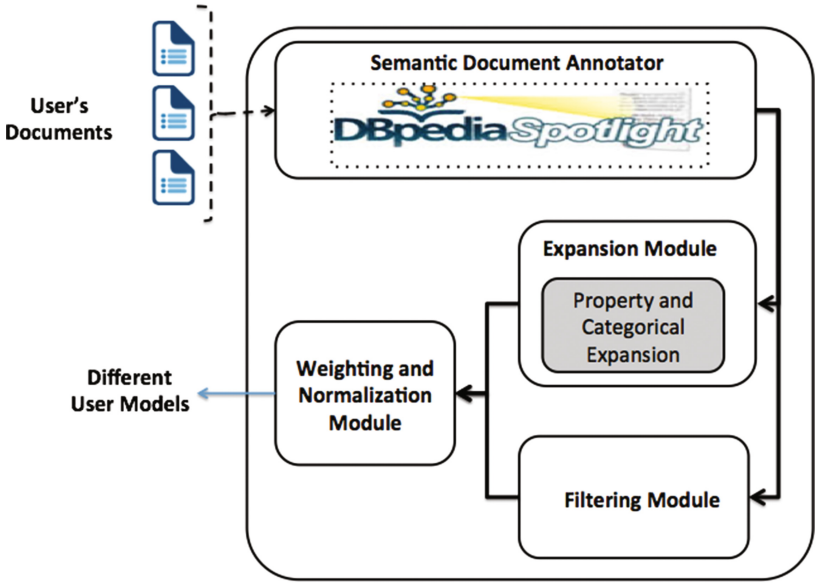


**Fig. 1.** Semantic profiling process

To implement the aforementioned semantic profiling process, we rely on DBpedia as the KB. Its comprehensive vocabularies, extensive relationships between concepts and continuous updating enable cross-domain modeling capabilities. DBpedia consists of a set of resources $R$ (i.e. we will use the word "resources" to refer to DBpedia entities such as "http://dbpedia.org/resouce/Colombia" and categories such as "http://dbpedia.org/resource/Category:Republics" ) and literals $L$ interrelated through a set of properties $P$, which are used to denote specific relationships. Under a RDF model, the DBpedia data consist of a set of statements $E \subset R \times P \times (R \setminus L)$. Each $e \in E$ is a triplet composed of a subject, a predicate and an object/literal. We define a user profile as the set of pairs $(r_i, w_i)$, where $r_i$ is a DBpedia resource and $w_i$ is the associated weight that represents the resource importance in the user profile:

$$UP = \{(r_1, w_1), (r_2, w_2) \dots (r_i, w_i) \mid r_j \in R\} \tag{1}$$

Since there are multiple users, this process is repeated multiple times, and for each user we obtain multiple profiles according to the different strategies. English and Spanish versions of DBpedia 2016 are used in a instance in Virtuoso. In the following sub-sections, each step in the process is described in detail.

### 3.1 Semantic Document Annotator

In this step, we were interested in finding annotations of DBpedia resources in the text. In order to find these annotations, DBpedia Spotlight service with a

JAVA heap memory of 16 GB was employed. The outcome of this step is an initial semantic user interest profile $ISUP_{u_i} = \{(r_1, w_{ini_1}), (r_2, w_{ini_2}) \ldots (r_i, w_{ini_i}) | r_j \in R\}$ for the user $u_i$. The initial resource weight $w_{ini_i}$ is calculated as the number of occurrences in the user document collection (i.e. users publications). It is important to mention that we do not perform any additional verification on the annotations discovered. As was mentioned by [22], there is no guarantee of a correct identification of annotations, so a manual cleaning is suggested. However, in a realistic scenario, a manual correction process is not feasible.

### 3.2 Model Expansion

In this step, we expand ISUP by relating its resources to new ones employing the rich background knowledge in DBpedia. We follow two different expansion approaches proposed by [2,14] which were later expanded upon in [16,17]:
Categorical expansion (CE): We add the DBpedia categories of each resource in $ISUP$. Then, we find such categorical resources through the Dublin Core dct:subject property and calculate their initial weight as the number of resources belonging to that category in the user document collection.
Property expansion (PE): The $ISUP$ is enriched with the set of resources recovered by following the set of properties $p : p \in DBpediaOntology$ of all the resources $r \in ISUP$.

### 3.3 Weighting and Normalization Module

**Annotations Weighting.** The final weight of each resource is calculated taking into account its presence in the total set of user profiles. Thus, the final weight of a resource $r_i$ is calculated as the initial weight multiplied by its IDF as follows [17]:

$$w_{idf_i} = w_{ini_i} \times log \frac{M}{m_{r_i}} \tag{2}$$

where $M$ is the total number of users and $m_{r_i}$ is the number of users interested in a resource $r_i$. The IDF strategy penalizes annotations that appear in multiple user profiles. We will refer to the user profile without the extension, but with IDF weighting strategy as follows:

$$SUP = \{(r_1, w_1), (r_2, w_2) \ldots (r_i, w_i) \mid r_j \in R\} \tag{3}$$

**Expanded Resource Weighting.** The weights of the expanded resources incorporated through CE and PE follow the discount formulas proposed in [14,17]. For an CE extended resource $cat_{e_i}$ and an PE extended resource $pro_{e_i}$ obtained from the resource $r_i \in SUP$, the weights with which they are incorporated into the user profile are calculated as:

$$w_{cat_{e_i}} = w_{idf_{cat_{e_i}}} \times \frac{1}{log(SP)} \times \frac{1}{log(SC)} \tag{4}$$

$$w_{pro_{e_i}} = w_{idf_{pro_{e_i}}} \times \frac{1}{log(P)} \tag{5}$$

where SP is the set of resources belonging to the category, SC is the set of sub-categories in the DBpedia categorical hierarchical and P is the number of occurrences of a property in the whole DBpedia graph. Only categories in the hierarchical structure processed by [11] were used to avoid disconnected categories and cycles. Finally, $w_{idf_{pro_{e_i}}}$ and $w_{idf_{cate_i}}$ are calculated in the same way as the weights for the annotations were.

### 3.4 Model Filtering

We argue that the weighting strategies explained above may not be enough to avoid the drift from the real user interests given that many resources found in a long document may be unrelated to the main topic. In academic papers, for example, multiple concepts could be found in the references sections that are not necessary related with the academic research interests of the user (universities, people, years, etc.). Actually, noise could increase if these two conditions occur: the same reference appears in multiple publications of the same user profile and it does not appear in other users' profiles. In this case, the IDF raises the noise for this user. The hypothesis is that the expansion strategies explained before could actually make the situation worse given that they could reinforce the noise in the user profile through the incorporation of other irrelevant resources. Consider the following subset of resources of a user profiles built from academic publications as an example:

*SUP={(Self-esteem,3.71), (Education,1.96), (Université du Québec, 4.11), (Undergraduate education, 2.72), (Higher education, 3.3), (Aquaculture, 3.12)}*

Université du Québec and Aquaculture are two resources identified in the text, but they drift from the real user interests. Université du Québec appears in the reference section in multiple publications by the same author. Aquaculture appears in the middle of the results section as part of an example that shows some of the author's findings. Université du Québec has a high IDF because it does not appear in any other user profiles, yet neither of these reflect the main interests of the user. Consequently, these resources lead to a poor representation of the real user interests. In order to address this problem, it is possible to analyze how connected the resources are by taking advantage of the graph representation of LOD datasets as DBpedia. Analyzing connecting paths[4] of length 1 (i.e. a direct relationship between two resources through a DBpedia property) for each possible pair of resources, it is possible to find the following connections:

*(Self-esteem, Education), (Education, Undergraduate education), (Education, Undergraduate education), (Education, Higher education), (Education, Higher education), (Undergraduate education, Higher education), (Undergraduate education, Higher education), (Higher education, Undergraduate education)*

In some cases there are multiple connections for the same pair of annotations; for example, there are two connections between Education and Undergraduate

[4] https://www.w3.org/TR/sparql11-property-paths/.

education through properties seeAlso and wikiPageLink. In contrast, no connections were found for Université du Québec and Aquaculture resources. Based on these findings, we think that analyzing the connections between annotations could be a useful approach to reducing noisy resources in $SUP$. Additionally, the number of times that a resource appears in connection paths can be used as an indicator of the importance of the resource in the interest profile, so we use this frequency as the basis for the weighting strategy.

We also take advantage of the analysis of connection path lengths greater than one. For those paths, there is the possibility of finding new resources that are related to two annotations in the connection path, but are not part of $SUP$. Additionally, we evaluate the incorporation of such resources into the user representation. The intuition behind this expansion is that resources that connect two annotations are more likely to reinforce the real user interests than those that are incorporated from the annotation properties or categories.

Our filtering strategy follows four steps:

- Step 1: Build a connection set with the highest $w_{idf_i}$ in $SUP$. For all our experiments we select the top 100 resources in $SUP$.
- Step 2: Find all paths of length $l$ ($1 <= l <= l_{maxpath}$) for each possible pair of resources $r_i$,$r_j$ in the connection set. Following previous experimental suggestions [15], we only consider outgoing edges from $r_i$ and $r_j$ in order to avoid the noise produced by highly indegree resources in DBpedia.
- Step 3: Build a filtered user profile incorporating all the resources found in the paths, and associating an initial weight representing the number of distinct paths in which the resource appears.
- Step 4: Apply the IDF weighting scheme explained in Sect. 3.3.1, but using as $w_{ini_i}$ the connection frequency of the above step.

For the rest of the document, we will refer to the filter user semantic model as $FSUP$.

## 4 Evaluation

We conducted our experimental setup to determine if: (i) semantic representations of the user perform better than classical user TF-IDF vector representation; (ii) expansion techniques improve the quality of the semantic user profile build from non-Social Web content; and (iii) our filter strategy outperforms expansion strategies in such cases.

Academic document recommendation is an appropriate scenario to address the above issues since academic documents are usually long; thus, it is possible to find many more annotations than those found in a Tweet or a Facebook post. Academic papers often use formal language and involve concepts that have complex relationships. Additionally, the task of recommending academic documents is a good scenario to evaluate content-based profiling techniques since textual data is the main (and sometimes the only) reliable source of information upon

which to base recommendations [19]. The full text of some of the user's publications will be used as input for the process based on the hypothesis that the user's academic interests are reflected in the documents they produce. This hypothesis is logical and has already been used in recent content-based recommendation systems [19–21].

### 4.1 Data Set

One of the main contributions of this work is the construction of a complete, semantically annotated dataset for academic document recommendation. We built our own dataset given that the Semantic Document Annotator process requires the text of the publication in order to correctly identify the annotations. Datasets used in previous research published the feature vector of the users and documents instead of the texts themselves (in a bag of words approach) mainly because they use sources that do not allow them to share the full text [19,20].

Our dataset contains the user profiles of 11 professors in the area of computer science. It was built from some of the most recent publications found on their Google Scholar web pages. At least a minimum of twelve of each professor's most recent publications were used as input for the semantic profiling process. The candidate set is a starting set of papers from which the recommended set of papers for the user is produced. In our case, it is a subset of Core and Arvix open corpora that was retrieved using different topic keywords in computer science as queries. After identifying and eliminating duplicate publications and unreadable pdf files, the final candidate set totaled 5710 different academic documents. The ground truth of papers is a subset of the candidate set in which the user expresses an explicit interest. Users interacted with a web-based search system to build this set. In order to reduce the possibility of selecting a paper unrelated to the main content, the web-based search system does not show the source of the academic document (journal or conference title). In the data set, we have at least 10 academic documents for which each user shows an explicit interest. The full dataset is available in https://github.com/Ruframapi. It contains the annotations found in the publications list of each user and the complete candidate set. The corresponding Arxiv and Core identifications are also shared in order to allow the access to the full text of the candidate set documents. It is important to signal that the process of building a ground truth for academic documents recommendation based on content is a challenging task, for it requires a time expensive participation of the users in order to analyze the text and explicitly assert its relevance. Hence, we also rely on a limited number of users as do previous studies.

The whole candidate set is annotated under the same representation as the user profile. However, the calculation of the IDF is made taking all the documents in the corpus into account. When the evaluation involved an expansion strategy, the corpora documents were also expanded.

### 4.2 Recommendation Algorithm

We follow a offline comparative framework that measures the quality of the recommendations for the different user profiles. Since the objective is to measure the influence of the user profile on the recommendation task, we must use a common, content-based recommendation algorithm. In other words, we want to measure the effects of the different user profiles as input for a common recommendation algorithm. For the purpose of this research, we select the documents with the highest cosine similarity with a given user profile.

### 4.3 Evaluation Metrics

We use the following typical metrics for the evaluation of Top-N recommender tasks [21]: MRR (Mean Reciprocal Rank), MAP@10 (Mean Average Precision), and NDCG@10 (Normalized Discounted Cumulative Gain). We select $N = 10$ as the recommendation objective since it is a common rank used in multiple applications [16,19] and it is not common to recommend a larger set of items. In our data set, the relevance measures are binaries (i.e. the recommended documents are relevant to the user or are not ), so we use a binary relevance scale for the calculation of NDCG. The final NDCG for each user strategy is calculated averaging the results for each user.

## 5 Results

In this section, experimental results of different semantic user profiles are shown. Our first question was related to the performance of a classical TF-IDF vector space model representation in comparison to a semantic user profile (SUP). In order to answer this question, we built a representation of the user and corpus documents with a TF-IDF scheme. According to [3] TF-IDF is the most frequent weighting scheme in research paper recommender systems. We carried out typical text processing operations including tokenizer, stop word removal and stemming. For all the experiments, we used the entire text of the publications as input. Table 1 shows the results obtained using the evaluation metrics explain above. Based on these results, the semantic approach performed better than TF-IDF. We noticed that the classical TF-IDF approach has the problem of retrieving documents that are too similar to the final top 10 recommendations list. In order to solve the problem of retrieving too similar documents, diversification strategies could be implemented [3]. We do not use them as these strategies operate to improve the recommendation not the user model.

In our second experiment, the different expansion strategies were evaluated. We compared the semantic user profile with categorical expansion ($SUP+CE$), the semantic user profile with property expansion ($SUP+PE$), the semantic user profile with categorical and property expansion ($SUP+CE+PE$) and the profile filtering strategy for different path lengths ($FSUP_{l_{maxpath}=1}$, $FSUP_{l_{maxpath}=2}$, $FSUP_{l_{maxpath}=3}$). The average number of resources for each type of user profile

**Table 1.** Semantic User Profile (vs) TF-IDF vector space model

| | MRR | MAP | NDCG |
|---|---|---|---|
| SUP | **0.4015** | **0.3429** | **0.4731** |
| TF-IDF | 0.3720 | 0.2718 | 0.3961 |

**Table 2.** Average number of resources for each type of user profile

| | |
|---|---|
| SUP | 1911 |
| SUP+CE | 9889 |
| SUP+PE | 2024 |
| SUP+CE+PE | 10001 |
| $FSUP_{l_{maxpath}=1}$ | 71 |
| $FSUP_{l_{maxpath}=2}$ | 1369 |
| $FSUP_{l_{maxpath}=3}$ | >8000 |

is shown in Table 2. As was expected, the $SUP + CE$ and $SUP + CE + PE$ have a higher number of resources, which is five times the number of resources found in the profile without expansion. On the contrary, our filtering strategy for $l_{maxpath}$ of 1 and 2 reduced the number of resources in SUP. It is worth noting that the exact number for $FSUP_{l_{maxpath}=3}$ might be a bit higher than shown as restrictions on the infrastructure configuration SPARQL queries took a maximum of 700 seconds for the result; queries taking more time were discarded.

The results using the evaluation measures described in Sect. 4.3 are summarized in Fig. 2. Only for categorical expansion ($SUP + CE$) is there a slight increase in the quality of the recommendation in comparison with the profile without expansion ($SUP$). In contrast, the other expansion strategies ($SUP + PE, SUP + CE + PRO$) lead to a deterioration of the semantic profile. These results differ from those obtained in the Twitter link recommendation [17] where property-based expansion obtained similar results as categorical expansion.

The recommendation accuracy obtained using our proposed filtering approach outperformed the expansion strategies. Interestingly, the best performance was achieved with the filtering strategy containing paths of length 1 ($FSUP_{l_{maxpath}=1}$). This result validates our initial hypothesis about the need to filter the annotations found in the text to build the user profile. For paths of length 3 ($FSUP_{l_{maxpath}=3}$), the filtering strategy includes more resources than those in the original SUP representation, yet the filtering strategy with length of 3 performs better in comparison with expansion strategies. We argue that longer paths could lead to negative effects since resources in DBpedia tend to be highly connected if the full set of ontology properties are taken into account[5]. So, instead of a filter, we could add additional noise to the user profile.

[5] http://konect.uni-koblenz.de/networks/dbpedia-all.

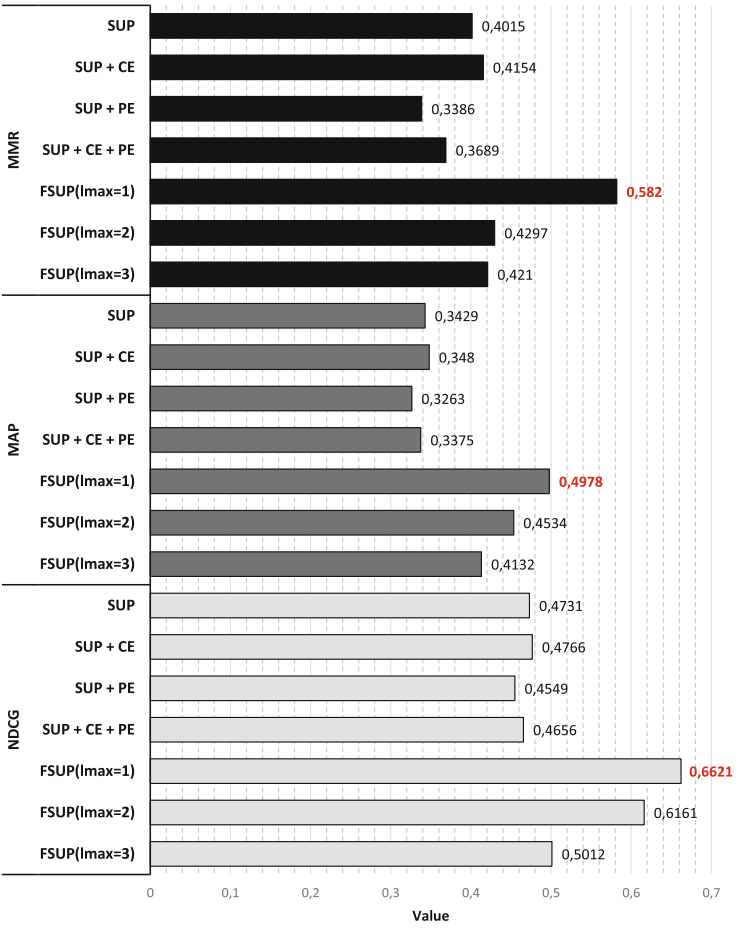


**Fig. 2.** Evaluation of expansion and filtering strategies.

In summary, our results suggest that: (i) there is a need for filtering techniques to refine the user profile and reduce the noise produced by the annotation process; (ii) expansion strategies applied directly to the annotations are not adequate in scenarios where the content input is not short texts; and (iii) it is possible to build better user interest profiles than those produced by the classical TF-IDF approach in the context of academic documents recommendation through a LOD knowledge base like DBpedia.

## 6 Conclusions and Future Work

In this paper, we presented the evaluation of different semantic user modeling strategies in an academic paper recommendation scenario. We showed that expansion strategies useful in the Social Web could increase the noise in the user

representation when dealing with longer documents. Because of this, in scenarios where the input content is a long document, filtering strategies outperform expansion strategies as this type of document involves multiple concepts that do not necessarily express the real interests of the user. Although our filtering strategy displayed superior performance than all the other models, further experimentation is needed to explore better ways to remove noise resources in the user representation. Reducing the set of possible properties in the paths or employing semantic similarity measures between resources are future routes to explore. Since the focus of our research was on the comparison of different user profile strategies and not on the recommendation itself, issues related to the recommendation process itself such as the temporality of the interests, which was addressed as a key issue in profile research interests [3], was not addressed. Future experimentation should be conducted in order to determine the ideal temporal frame to include the user's papers, and the weighting strategies to incorporate them in the user profile. Finally, we will continue working on increasing the size of the dataset to validate these findings with a larger number of examples. The objective is to build a common scenario to measure and compare future semantic user profile strategies.

**Acknowledgment.** This work was partially supported by COLCIENCIAS PhD scholarship (Call 647-2014).

## References

1. Abel, F., Gao, Q., Houben, G.-J., Tao, K.: Analyzing user modeling on Twitter for personalized news recommendations. In: Konstan, J.A., Conejo, R., Marzo, J.L., Oliver, N. (eds.) UMAP 2011. LNCS, vol. 6787, pp. 1–12. Springer, Heidelberg (2011). doi:10.1007/978-3-642-22362-4_1
2. Abel, F., Hauff, C., Houben, G.-J., Tao, K.: Leveraging user modeling on the social web with linked data. In: Brambilla, M., Tokuda, T., Tolksdorf, R. (eds.) ICWE 2012. LNCS, vol. 7387, pp. 378–385. Springer, Heidelberg (2012). doi:10.1007/978-3-642-31753-8_31
3. Beel, J., Gipp, B., Langer, S., Breitinger, C.: Research-paper recommender systems: a literature survey. Int. J. Digit. Libr. **17**(4), 305–338 (2016)
4. Beel, J., Langer, S., Gipp, B.: TF-IDuF: a novel term-weighting scheme for user modeling based on users' personal document collections. In: Proceedings of the iConference 2017 (2017)
5. Berrizbeita, F., Vidal, M.E.: Traversing the linking open data cloud to create news from Tweets. In: Meersman, R., et al. (eds.) On the Move to Meaningful Internet Systems: OTM 2014 Workshops. LNCS, vol. 8842, pp. 479–488. Springer, Heidelberg (2014). doi:10.1007/978-3-662-45550-0_48
6. Di Noia, T., Cantador, I., Ostuni, V.C.: Linked open data-enabled recommender systems: ESWC 2014 challenge on book recommendation. In: Presutti, V., Stankovic, M., Cambria, E., Cantador, I., Di Iorio, A., Di Noia, T., Lange, C., Recupero, D.R., Tordai, A. (eds.) Semantic Web Evaluation Challenge. Communications in Computer and Information Science, pp. 129–143. Springer International Publishing, Cham (2014)

7. Di Noia, T., Mirizzi, R., Ostuni, V.C., Romito, D.: Exploiting the web of data in model-based recommender systems. In: Proceedings of the Sixth ACM Conference on Recommender Systems, RecSys 2012, pp. 253–256. ACM, New York (2012)
8. Di Noia, T., Mirizzi, R., Ostuni, V.C., Romito, D., Zanker, M.: Linked open data to support content-based recommender systems. In: Proceedings of the 8th International Conference on Semantic Systems, I-SEMANTICS 2012, pp. 1–8. ACM, New York (2012)
9. Froschl, C.: User Modeling and User Profiling in Adaptive E-learning Systems. Master's thesis, Graz University of Technology (2005)
10. Godoy, D., Amandi, A.: A conceptual clustering approach for user profiling in personal information agents. AI Commun. **19**, 207–227 (2006)
11. Kapanipathi, P., Jain, P., Venkataramani, C.: Hierarchical interest graph. Technical report (2015)
12. Koch, N.: Software Engineering for Adaptive Hypermedia Systems: Reference Model, Modeling Techniques and Development Process. Ph.D. thesis, Ludwig-Maximilians-University (2000)
13. Meymandpour, R., Davis, J.: Enhancing recommender systems using linked open data-based semantic analysis of items. In: Davis, J.G., Bozzon, A. (eds.) 3rd Australasian Web Conference (AWC 2015). CRPIT, vol. 166, pp. 11–17. ACS, Sydney, Australia (2015)
14. Orlandi, F., Breslin, J., Passant, A.: Aggregated, interoperable and multi-domain user profiles for the social web. In: Proceedings of the 8th International Conference on Semantic Systems, I-SEMANTICS 2012, pp. 41–48. ACM, New York (2012)
15. Paul, C., Rettinger, A., Mogadala, A., Knoblock, C.A., Szekely, P.: Efficient graph-based document similarity. In: Sack, H., Blomqvist, E., d'Aquin, M., Ghidini, C., Ponzetto, S.P., Lange, C. (eds.) ESWC 2016. LNCS, vol. 9678, pp. 334–349. Springer, Cham (2016). doi:10.1007/978-3-319-34129-3_21
16. Piao, G., Breslin, J.G.: Analyzing aggregated semantics-enabled user modeling on Google+ and Twitter for personalized link recommendations. In: Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, UMAP 2016, pp. 105–109. ACM, New York (2016)
17. Piao, G., Breslin, J.G.: Exploring dynamics and semantics of user interests for user modeling on Twitter for link recommendations. In: Proceedings of the 12th International Conference on Semantic Systems, SEMANTiCS 2016, pp. 81–88. ACM, New York (2016)
18. Schiaffino, S., Amandi, A.: Intelligent user profiling. In: Bramer, M. (ed.) Artificial Intelligence An International Perspective. LNCS, vol. 5640, pp. 193–216. Springer, Heidelberg (2009). doi:10.1007/978-3-642-03226-4_11
19. Sugiyama, K., Kan, M.Y.: Scholarly paper recommendation via user's recent research interests. In: Proceedings of the 10th Annual Joint Conference on Digital Libraries, JCDL 2010, pp. 29–38. ACM, New York (2010). http://doi.acm.org/10.1145/1816123.1816129
20. Sugiyama, K., Kan, M.Y.: Exploiting potential citation papers in scholarly paper recommendation. In: Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL 2013, pp. 153–162. ACM, New York (2013)
21. Sugiyama, K., Kan, M.Y.: A comprehensive evaluation of scholarly paper recommendation using potential citation papers. Int. J. Digit. Libr. **16**(2), 91–109 (2015)
22. Waitelonis, J., Exeler, C., Sack, H.: Linked data enabled generalized vector space model to improve document retrieval. In: Proceedings of 3rd International Workshop on NLP & DBpedia 2015 (2015)