An Architecture to Enhance a Reference Management System with Recommendations from Open Linked Data

María Hallo¹ and Sergio Luján-Mora²

¹Department of Informatics and Computer Science, Escuela Politécnica Nacional, Quito, Ecuador
²Department of Software and Computing Systems, University of Alicante, Alicante, Spain
maria.hallo@epn.edu.ec, sergio.lujan@ua.es

Keywords: Content-based Recommender Systems, Open Linked Data, Reference Management Software.

Abstract: Reference management software helps students and researchers to store and to cite publications in different citing formats. Common features of reference management software include advanced searching, reference libraries and generation of citations. Some implementations help users to be connected to digital libraries to get the article metadata of interest. Some researchers need to manage references of publications from open access journals because of the elimination of barriers such as the price to get publications. However, the number of publications grows each year and the researchers devote so much time to the retrieval, analysis and management of bibliographic information. To solve this problem, in this work, we present a framework to support the search, download and management of bibliographic information. A content-based recommender module based on Open Linked Data is included into the framework. The metadata of the research publications and the corresponding PDF files links are extracted using the recommender module and the Application Program Interface from the Directory of Open Access Journals (DOAJ). The results are presented to the user for the selection process. The metadata of the selected publications are stored in a local database integrated in a bibliographic management system. A prototype was developed and was tested with information from open access journals managed by the DOAJ.

1 INTRODUCTION

Reference management is a hard task for researchers. At present, the researchers have access to more information than they can consume and most of the retrieved publications are not so relevant. The search for related work is a hard consuming task for researchers. The number of publications grows each year and the open access initiative helps to have better diffusion to some of those publications.

Some programs have been used to collect research literature helping to organize communities for sharing research literature (Ayers & Priedhorsky, 2011), such as: Wikindx, Zotero, Mendeley, Reworks, Referencer, JabRef, Kibtex, EndNote, Readcube, etc. (Kaur & Dhindsa, 2016). However, most of these tools do not have an associated recommender system.

Recommender systems for research publications are useful applications to help researchers to know the state of the research in a specific topic. A good recommender system is one that recommends the most relevant items considering the user preferences and goals (Beel, Langer & Genzmehr, 2013).

There are several types of recommender systems. Content-based filtering has been used in several recommender systems for reference management such as: Dblib, Rec4LRW and TechLens (Torres et al, 2004). Mendeley Suggest combine content based recommendation with collaborative filtering (Jack et al, 2016). The main problem in those solutions is matching domain specific schemas. In collaborative recommendations the main problem is the lack of standards for data portability to integrate data from different sources (Passant & Vojta, 2010).

Generally, recommender systems have been built based on user preferences, interactions and resource descriptions. A new approach of recommender systems based on interlinked data has been reported in several studies, enabling better integration of resources between applications (Peska, 2013). The common representation formats for Linked data, typically is Resource Description Framework (RDF), a language to represent machine-readable statements in the form of triples <subject> <predicate> <object>. The common semantics of this data is represented using ontologies and related languages, i.e. RDFS.
(W3C, 2014), OWL (W3C, 2004); and SPARQL (W3C, 2008). Some research has been reported using Linked Data on content-based recommender systems (Di Noia, 2012). Linked Data is a set of best practices to publish an interlinked data on the Web, and it is the basis of an interconnected global data space where data providers publish their content publicly (Berners-Lee, 2006).

According to (Figueroa et al, 2015) the main research contributions in content-based recommender systems using Linked Data are resumed on: a) the definition or extension of a similarity measures, b) the definition or extension of an ontology, c) the definition or extension of recommender algorithms, d) the information enrichment.

In this paper we present a framework for a reference management system including a content-based filtering recommender approach based on Open Linked Data for open access to scientific publications. Using this proposed framework, a prototype system has been developed. This system has been tested with information from the Directory of Open Access Journals (DOAJ) for the initial search combined with information from ACM publication metadata for the recommendations.

The rest of the paper is structured as follow: In Section 2, we introduce related work. In Section 3, we present the architecture of the proposed solution. In Section 4, the results of a system implementation using the proposed framework is described. Finally, in Section 5, we show the conclusions and future work.

2 RELATED WORK

Recommender systems are used to increase the user satisfaction and precision of the retrieved information. They help users to find their items of interest (Ricci, 2011). There are several classifications of recommender systems, one of them (Lops, De Gemmis & Semeraro, 2011) presents three groups of recommender systems: Content-based, collaborative filtering and knowledge-based recommender systems.

Content-based recommender systems match up the attributes of a user profile, constructed by user’s past interest, with the information extracted from the item in order to recommend to the user new interesting items. Collaborative filtering (CF) recommender systems makes recommendations to a user based on items that other users liked in the past (Goldberg et al, 1999). The main problem is to obtain sufficient number of ratings to have a useful system.

Knowledge recommender systems are based on specific domain knowledge about item features and their corresponding matching with user preferences (Ricci et al, 2011). Ontologies and Linked Data can be used to improve the search based on domain concepts and open data structure as Linked Data (Hallo et al, 2016). There are also hybrid recommender systems which use the combination of the advantages of the previous cited techniques based on the better features of each one. Additional methods such as citation analysis (Vellino, 2015), taxonomic topic expansion (Zarrinkalam & Kahani, 2012), query expansion (Lüke et al, 2012) or term recommender (Lüke et al, 2013) were also studied to enhance CF algorithms.

There are reported recommender systems for reference management enabling researchers to control their research paper metadata, annotations and PDF files. A Research Paper Recommender System (RPRS) suggests research papers to the users according to their personal preferences.

Most of the RPRS use content-based recommender systems; few of them use collaborative filtering or knowledge-based recommendations (Pohl et al, 2007). Open Linked Data has been proposed to improve the relevance of the retrieved items.

In Linked Open Data cloud there are millions of RDF triples distributed in several datasets (Yang, 2010). This data, in RDF format, could be used to enhance research paper recommender systems. Following, we describe some of the studies related to Linked Data-enable recommender systems, dataset used and algorithms to produce recommendations.

In 2017, a SKOS Recommender Prototype which produce scalable recommendations through SPARQL-like query from Linked Data repositories was proposed (Wenige & Rughland, 2017). This system offers a combination of similar resource retrieval and graph pattern matching. In other study, Amnami et al (2016) developed a Latent Dirichlet allocation (LDA)-based approach to scientific paper recommendation. In this proposal the core idea is to exploit the topics related to the authored articles to formally define the author’s profile using topic modelling and language modelling to represent the recommender papers. The recommendation technique is based on the closeness of the language used in the research papers and the one used in the recommender papers. This proposal alleviates the cold start problem typical of collaborative filtering techniques. In other study, Meymandpour et al (2013) report a hybrid recommender system based on Linked Open Data which could be also used for scientific paper recommendation. In this study the
recommendation method is based on the measure of the semantic similarity of the resources combined with user ratings. The features of each item are extracted based on the relations with their neighbours. Each link has a weight which could be assigned by the experts. This system also work in absence of ratings for items.

Beel et al. (2013) recommend to consider the impact of the labelling in research paper recommendations. In that work the nature of the recommendation was analysed finding that the institutional recommendations have better acceptance than the sponsored recommendations. For all the proposals it was necessary to have open RDF data sets to improve the recommendations.

There are some bibliographic data sets available on the Linked Open Data. RKBexplorer is a Semantic Web application containing information from several sources such as: DBLP, ACM, IEEE, Citeseer. (Glaser et al., 2007). The datasets are available through an SQL endpoints and resolvable URIs. Semantic Web Dog Food is other large structured dataset that focuses on the publications from the Semantic Web community (Hu et al., 2015).

For the recommendations, Passant (2010) proposed a semantic distance measure and collaborative filtering approach based on Linked Data.

In relation to similarity measures, the studies selected applied a variety of similarity measures (Cheniki et al., 2016). These include pairwise cosine function for vector similarity computation between items, feature-based similarity to evaluate semantic distance on different datasets, rating-based similarity to compute the popularity of items among users.

Semantic similarity is also used based on diverse relationships that can be found between concepts of Linked Data datasets. Such relationships can be paths, links or shared topics among a set of items. In addition topic-based similarity has been used capturing the relatedness between items based on the categories they belong to (Figueroa et al., 2015). Recommendations of a document x are based on the most related to x in descending order (Hajra et al, 2014).

3 TECHNICAL ARCHITECTURE

In this section we describe the proposed technical architecture to integrate the open source reference management system Wikindex with a recommender module based on Open Linked Data. Open Linked Data from ACM metadata has been used for the test. In the Figure 1, we present the technical architecture, where each component and the information flow between them are shown. Following, we describe the components of the architecture: Integration Module, Extraction Module, Reference Management Module, and Recommender Module.

3.1 Integration Module

The integration module receives the query terms (keywords, title, abstract or author) from the user and it calls the extraction module to get the searched publications. This module also presents the retrieved publications to the user.
3.2 Extraction Module

This module extracts the searched publications from DOAJ database. The retrieved publications are selected to be stored in a Wikindx local database. The user could also add notes and additional files.

3.3 Recommender Module

An Open Linked Data content-based recommender module is used to retrieve the article metadata that better fits the search keywords from the DOAJ selected publications. A scientific research paper dataset in RDF is used to get the recommendations. This module, adapted from the proposal of Meymandpour (2013), has two components: semantic information retrieval, and resource ranking.

3.3.1 Semantic Information Retrieval

In this step, facts and relations related to the selected resources are extracted from several open linked data sources after checking the quality of them. In our proposal we consider for the recommendation the elements: Resource (R), Category (C) and the relations between them: R - R, R - C, C - C. (Figueroa et al., 2015).

a) R-R are relations between resources based on similar features. i.e. similar shared authors or key-words. We considered that two articles are similar in an RDF graph if they are the subjects of two RDF triples having the same property and the same object. (i.e. objects from the same author).

b) R-C are relationships between a resource and a category represented by the RDF property rdf:type and the SKOS properties skos:subject , skos:isSubjectOf or dcterms:subject from the Dublin Core vocabulary.

c) C-C are hierarchical relationships between categories. They can be represented by using RDFS property rdfs:subClassOf or the SKOS property skos:narrower or skos:broader.

Thesaurus terms are linked with each other with semantic relations such as “broader”, “narrower” or “related”. The Thesaurus is an instrument to index and retrieve subject-specific information.

Given an initial resource (or a set of initial ones) a set of candidate resources linked to them located at a predefined distance are generated. The resources are retrieved based on the links R-R, R-C and C-C.

Following we present some SPARQL queries used to find the resources for recommendation.

Find categories (?catRURI) of the resource <RURI>

```sparql
PREFIX dcterms: <http://purl.org/dc/terms>
SELECT ?catRURI WHERE {
  <RURI> dcterms:subject ?catRURI}.
```

- Find candidate resources (?cRURI) of the same category <catRURI>:

```sparql
PREFIX dcterms : <http://purl.org/dc/terms/>
SELECT DISTINCT ?cRURI WHERE {
  ?cRURI dcterms:subject <catRURI> .
}
```

- Find candidate resources ?cr1URI linked to the resource <RURI>:

```sparql
SELECT ?cr1URI WHERE {
  #output links
  {<RURI> ?link ?cr1URI. }
  UNION
  #input links
  {?cr1URI ?link <RURI>}.
}
```

3.3.2 Resources Ranking

For ranking the candidate resources, Passant propose to calculate the Linked Data semantic distance (LDSD) based on the number of direct input and output links between two resources. In this proposal the links came from a specific domain (Passant, 2010).

The SPARQL query that counts input and output direct links between an initial resource (<inUrl>) and a resource (<cRURI>) from the set of candidate resources is:

```sparql
SELECT DISTINCT count (?links) WHERE {
  # output links
  {<inURI> ?links ?cRURI .} 
  # input links
  {<cRURI> ?links <inURI> .}
}
```

The similarity of two resources (r1, r2) is measured by:

\[ LDSD(r1, r2) = \frac{1}{1+\text{C dout} + \text{C din}} \]
Cdout is the number of direct output links (from r1 to r2), Cdin is the number of direct input links. Using these SPARQL queries, the ranking algorithm calculates the LDSD for each pair of resources composed of an initial resource and each of the resources obtained from the previous step. The same author test also another measure of similarity considering direct and indirect links but the results are not so much different.

3.4 Reference Management Module

Wikindx has been selected to be integrated with the extraction and the recommender modules. Wikindx is a free open source bibliographic management system. Its main characteristics are (Mucnjak, D. 2008):

- The system allows editing or creating bibliographic styles through a graphical interface.
- The articles can be reformatted to another citation style.
- The user can import and export other bibliographic formats including BibTeX or Endnote format.
- The users can export the bibliography in various bibliographic styles (APA, Chicago, IEEE).
- The system can be used by one or multiple users in a networked web server for a collaborative work.
- The controller manages the commands for the model or view.

4 RESULTS

A system prototype based on the proposed framework is been developed.

The user interface is developed applying the model, view, controller (MVC) software architecture pattern. It divides the application in three components:

- The model stores data to be presented in views.
- The view generates outputs to the user.
- The controller manages the commands for the model or view.

DOAJ was used to get the search metadata publications. The metadata were retrieved in a JSON file. Ajax and JQuery have been used to process the JSON metadata displaying them on the webpage.

The recommender module is developed based on information from Open Linked Data and tested with an ACM scientific research paper dataset, in RDF format, stored in RKBexplorer. The publication metadata obtained for the retrieval publications from DOAJ’s database are title, journal, abstract, keywords, publisher, category and authors. The users could record the metadata from the selected publications in their own local system.

Following we present some SPARQL query tests to get publications for recommendation:

- To find links to a publication.
- To find the authors of a publication.
- To find other publications of the same authors.
- To find other publications of the same area of interest.

![Figure 2: Prefix for SPARQL queries.](prefix.png)

All the queries use the prefix shown in the Figure 2 and the selected publication identified by the URI http://acm.rkbexplorer.com/id/100233.

- To find links to a publication.

In the Figure 3, we show the results of the indicated SPARQL query execution to present some links to the selected publication: http://acm.rkbexplorer.com/id/100233.

Some properties are found such as rdf:type, akt:has-title, akt:has-author, akt:addresses-generic-area-of-interest.

```sparql
PREFIX id: <http://acm.rkbexplorer.com/id/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX akt: <http://www.aktors.org/ontology/portal#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

SELECT DISTINCT ?links ?o WHERE
{<http://acm.rkbexplorer.com/id/100233> ?links ?o} LIMIT 20
```
To find the authors of a publication.

In the Figure 4, we show the results of the execution of the indicated SPARQL query to present the names of the authors of the selected publication and the results of the execution. The properties used in the query are akt:has-author and akt:full-name. Different names for the same author in the publications are also presented.

```sparql
SELECT DISTINCT ?name WHERE 
```

To find other publications of the same authors.

In the Figure 5, we show the results of the indicated SPARQL query execution to present the titles of the publications from the same authors of the selected publication. The properties used in the query are akt:has-author and akt:has-title.

```sparql
SELECT DISTINCT ?t WHERE 
  {?p akt:has-author ?a. ?a akt:has-title ?t}
```

To find other publications of the same area of interest.

In the Figure 6, we show the results of the execution of the SPARQL query to display the titles of the publications from the same area of interest of the selected publication. The properties used in the query are akt:addresses-generic-area-of-interest and akt:has-title.

```sparql
SELECT DISTINCT ?t WHERE 
```

5 CONCLUSIONS AND FUTURE WORK

We demonstrated how Linked Open Data can be used to get research publications for recommendations. The proposal architecture has been used to build a new recommender systems that can operate for a
particular domain data. We carried out a preliminary evaluation using the data and categories from Open ACM Linked Data resources. It is important to link the resources of different datasets considering that each one use different URIs for the resources i.e. publications or authors. In the future we will make relevance tests for the rankings and we will also study the possibilities to scale this solution to manage big quantities of information. Finally, we will work in: analyzing a term based recommender module, integrating other scientific databases and developing a method to control de quality of the data. We need also to define criteria for the final system evaluation of the recommender system.

ACKNOWLEDGEMENTS

This publication comes from research conducted in the project PII-16-06, with the financial support of Escuela Politécnica Nacional from Quito, Ecuador.

REFERENCES


