Cite4Me: A Semantic Search and Retrieval Web Application for Scientific Publications

Bernardo Pereira Nunes[1,2], Besnik Fetahu[1], Stefan Dietze[1], and Marco A. Casanova[2]

[1]L3S Research Center, Leibniz University Hannover, Appelstr. 9a, 30167 Hannover, Germany
{nunes, fetahu, dietze}@L3S.de
[2]Department of Informatics, Pontifical Catholic University of Rio de Janeiro,
Rio de Janeiro/RJ – Brazil, CEP 22451-900
{bnunes, casanova}@inf.puc-rio.br

Abstract. Cite4Me is a Web application that leverages Semantic Web technologies to provide a new perspective on search and retrieval of bibliographical data. The Web application presented in this work focuses on: (i) semantic recommendation of papers; (ii) novel semantic search & retrieval of papers; (iii) data interlinking of bibliographical data with related data sources from LOD; (iv) innovative user interface design; and (v) sentiment analysis of extracted paper citations. Finally, as this work also targets some educational aspects, our application provides an in-depth analysis of the data that guides a user on his research field.

1 Introduction

The huge amount of Web data and resources, particularly in the academic area, calls for strategies to analyse and explore resources and data.

While scientific disciplines are very data- and knowledge-intensive, the lack of semantic tools hampers information management and decision making. This includes scientific data as well as unstructured academic publications as one of the key outcome of scientific work. This is due to information access offered by digital library providers such as ACM Digital Library[1] and Elsevier[2] being mostly based on free text search and hierarchical classification[3].

Thus, we present a novel Web application for exploratory search, retrieval and visualization of scientific publications. *Cite4Me* aims at providing a single access point for accessing papers and, therefore, assisting searchers on finding relevant topics, papers, and unveiling new nomenclature more efficiently. For this, we use reference datasets such as DBpedia[4] to explore semantic relationships between scientific papers and user queries. We also perform a topic coverage analysis to provide an overview of different bibliographic datasets. *Cite4Me* is a Web application which exploits results of previous research works [2–5].

[1] `http://dl.acm.org`
[2] `http://www.elsevier.com`
[3] `http://www.acm.org/about/class/`
[4] `http://dbpedia.org`

2 Cite4Me - The Application

Cite4Me implements semantic and co-occurrence-based methods to search and retrieve academic papers and suggest related work in a user-friendly interface that assists users in exploring relationships between authors, institutions, papers and query terms. Due to space restrictions, we present in this paper the most relevant features of *Cite4Me* to the Semantic Web field.

2.1 Search and Retrieval

Cite4Me implements standard techniques, such as free text search, to search and retrieve scientific publications. In this section, we emphasize the semantic and exploratory search mechanisms.

Exploratory Search. The *exploratory search* or *graph search* component assists users to discover related work, people and institutions that are working on a specific topic. A crucial step to provide this type of search is the annotation of the publications' content. For this, we used DBpedia Spotlight API[5] for extracting entities, entity types and their categories. For instance, the categories of the extracted concepts are used to interlink publications through the topics they cover. In cases where two publications share the same category (*dcterms:subject* property), then a link between both publications is created. Figure 1 shows an example of topically related publications.

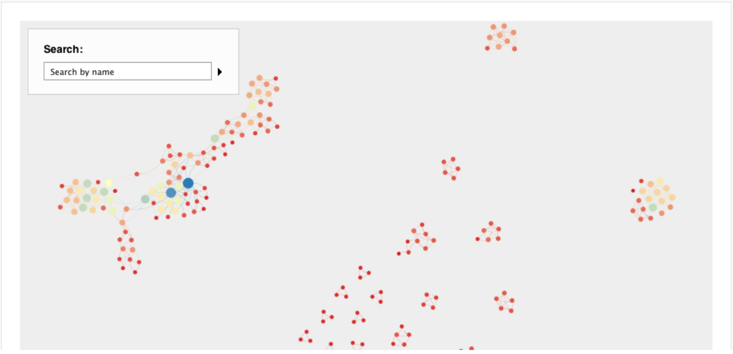


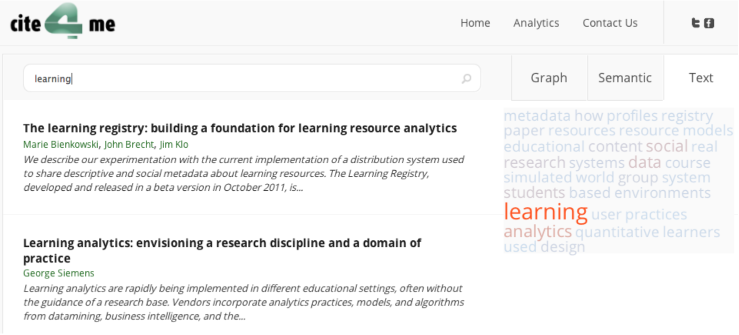
Fig. 1. Preview of the exploratory search funcionality.

Semantic Search. The *semantic search* component of *Cite4Me* is similar to the *explicit semantic analysis* (ESA) technique [1]. After running the annotation process aforementioned, the relatedness score between the enriched concepts (DBpedia entities) found

[5] `http://dbpedia.org/spotlight`

in the user query terms and the publications' content are computed and ranked. The relatedness score is computed based on the *tf-idf* score for the entities found in the publications' content. The ranking of the retrieved documents is based on the sum of the *tf-idf* scores of the matching concepts.

Figure 2 illustrates the semantic search functionality. Alongside the results of the semantic search a tag cloud shows the most prominent terms for a given user query. The tag cloud is updated while browsing through the list of results. The tags are selected based on the *tf-idf* score for the entities found in the abstract of the retrieved papers.


cite 4 me
Home
Analytics
Contact Us
learning
Graph
Semantic
Text
metadata how profiles registry
paper resources resource models
educational content social real
research systems data course
simulated world group system
students based environments
learning user practices
analytics quantitative learners
used design

The learning registry: building a foundation for learning resource analytics
Marie Bienkowski, John Brecht, Jim Klo
We describe our experimentation with the current implementation of a distribution system used to share descriptive and social metadata about learning resources. The Learning Registry, developed and released in a beta version in October 2011, is...

Learning analytics: envisioning a research discipline and a domain of practice
George Siemens
Learning analytics are rapidly being implemented in different educational settings, often without the guidance of a research base. Vendors incorporate analytics practices, models, and algorithms from datamining, business intelligence, and the...

Fig. 2. Preview of the semantic search funcionality.

Paper recommendation. Another important feature of *Cite4Me* and which differentiates it from similar tools is the *semantic paper recommendation*. Given a scientific publication, the tool recommends a related paper based on a score calculated according to direct and lateral relationships between the publication of interest and the remaining papers in our corpus.

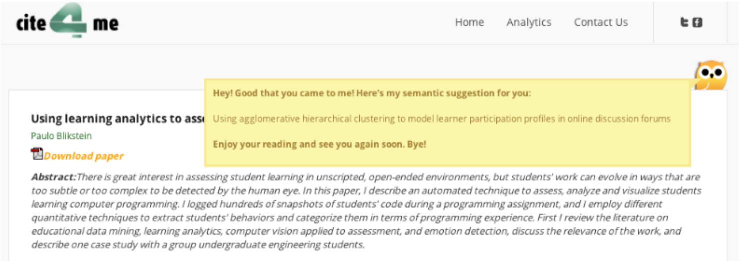
To compute the relatedness score, we rely on previous work by Nunes et al. [2, 4], where the paths connecting two enriched concepts in the scientific publications are analysed using a variation of the Katz index, a measure based on Social Network Theory, and quantifying the weight of the connectivity between two concepts given a knowledge graph (in our case DBpedia graph).

After computing the relatedness scored between enriched concepts, the paper recommendation relies on an aggregated measure that takes into account the relatedness inter-documents. Finally, we generate a ranked list of pairwise publications according to the overall score (see [5] for more details). Thus, the top-ranked publication is recommended to the user, as shown in Figure 3.

3 Datasets

Currently, *Cite4Me* is linked to a dataset (*LAK Dataset*[6]) which contains semi-structured research publications from the ACM Digital Library (under a special license)

[6] http://www.solaresearch.org/resources/lak-dataset/



Hey! Good that you came to me! Here's my semantic suggestion for you:

Using agglomerative hierarchical clustering to model learner participation profiles in online discussion forums

Enjoy your reading and see you again soon. Bye!

Using learning analytics to ass

Paulo Blikstein

Download paper

Abstract: *There is great interest in assessing student learning in unscripted, open-ended environments, but students' work can evolve in ways that are too subtle or too complex to be detected by the human eye. In this paper, I describe an automated technique to assess, analyze and visualize students learning computer programming. I logged hundreds of snapshots of students' code during a programming assignment, and I employ different quantitative techniques to extract students' behaviors and categorize them in terms of programming experience. First I review the literature on educational data mining, learning analytics, computer vision applied to assessment, and emotion detection, discuss the relevance of the work, and describe one case study with a group undergraduate engineering students.*

Fig. 3. An example of paper recommendation based on SCS_w.

and other public datasets (see also [6] for details). The dataset contains 315 full papers along with their descriptive metadata while new publications are added continuously. Metadata as well as the full text body are freely available in a variety of formats, including RDF accessible via a public SPARQL endpoint. We are currently working on expanding the number of papers available in *Cite4Me*. However, due to copyright reasons, the process to expose scientific publications from publishers is still under discussion.

4 Conclusion

This paper presented the application of previous works in the Semantic Web field within *Cite4Me*, a Web application that assists users in finding relevant scientific papers by exploring semantic relationships between them. For more information about the *Cite4Me* Web application please refer to `http://www.cite4me.com`.

References

1. E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of IJCAI'07*, pages 1606–1611, San Francisco, CA, USA, 2007.
2. B. Pereira Nunes, S. Dietze, M. A. Casanova, R. Kawase, B. Fetahu, and W. Nejdl. Combining a co-occurrence-based and a semantic measure for entity linking. In *ESWC*, 2013 (to appear).
3. B. Pereira Nunes, B. Fetahu, and M. A. Casanova. Cite4me: Semantic retrieval and analysis of scientific publications. In M. d'Aquin, S. Dietze, H. Drachsler, E. Herder, and D. Taibi, editors, *LAK (Data Challenge)*, volume 974 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013.
4. B. Pereira Nunes, R. Kawase, S. Dietze, D. Taibi, M. A. Casanova, and W. Nejdl. Can entities be friends? In *Proceedings of WOLE, in conjuction with the ISWC'12*, volume 906 of *CEUR-WS.org*, pages 45–57, Nov. 2012.
5. B. Pereira Nunes, R. Kawase, B. Fetahu, S. Dietze, M. A. Casanova, and D. Maynard. Interlinking documents based on semantic graphs. In *Proceedings of KES'13*, 2013 (to appear).
6. D. Taibi and S. Dietze. Fostering analytics on learning analytics research: the lak dataset. In *Proceedings of the LAK Data Challenge, held at LAK2013*, April 2013.