Cite4Me: Semantic Retrieval and Analysis of Scientific Publications

Bernardo Pereira Nunes
PUC-Rio
Rio de Janeiro, Brazil
bnunes@inf.puc-rio.br

Besnik Fetahu
L3S Research Center
Appelstrasse 9a
Hannover, Germany
fetahu@l3s.de

Marco Antonio Casanova
PUC-Rio
Rio de Janeiro, Brazil
casanova@inf.puc-rio.br

ABSTRACT

This paper presents the Cite4Me Web application and its features created for the LAK Challenge 2013. The Web application focuses on two main directions: (i) interlinking of the LAK dataset with related data sources from the Linked Open Data cloud; and (ii) providing innovative search, visualization, retrieval and recommendation of scientific publications from the LAK dataset and related interlinked resources. Our approach is based on semantic and co-occurrence relations to provide new browsing experiences to Web users and an overview of scientific data available. Furthermore, we present a detailed analysis of the LAK dataset along with applications which contributes to the development of the learning analytics field.

1. INTRODUCTION

The volume of information on the Web has been growing steadily over the last decade and has doubled every two years. The vast amount of data available on the Web along with new means of communications have transformed our society, including the way we work, live, relate to each other and learn.

In the midst of change, the *Learning Analytics* emerges to make sense of the produced educational data reported by learners, professors, institutions and so on. Analyzing and understanding the changes along the past years help us to understand the current state and be aware of the forthcoming trends, enabling a new outlook of the future of learning.

A recent challenge initiative of SOLAR[1] and LinkedUp[2] project arises to leverage the creation of tools that enables the analysis, visualization, browsing and recommendation of scientific and educational data.

Although the scientific field has fostered the creation of new applications in several areas, such as medical, biology, physics, amongst others, the information access is based mostly on free text search and on hierarchical classification system of the publications[3]. However, current approaches by main digital library providers, such as ACM Digital Library[4] and Elsevier[5], do not represent the current state of research on exploring resources using approaches from Information Retrieval, Information Extraction and Semantic Web. Thus, get an overview of research topics, find publications and discover new nomenclatures are an arduous and laborious task that are not always successful.

In this paper, we introduce *Cite4Me* a novel application for exploratory search, retrieval and visualization of scientific publications. *Cite4Me* intends to provide to the end users a single point for accessing papers and hence reducing efforts of searching in several data sources. Our system takes advantage of reference datasets, such as DBpedia[6], to explore semantic relationships between scientific papers and user queries. Additionally, an analysis of topic coverage and shared concepts from related educational datasets, extracted from the Linked Open Data cloud, will be introduced.

The remaining of the paper is organized as follows. Section 2 presents the approach used for searching, retrieving and recommending papers. Section 3 describes the process of dataset discovery and interlinking and Section 4 shows a brief result analysis of the data discovery. Finally, Section 5 presents related work and Section 6 presents some concluding remarks.

2. CITE4ME

As one of the main goals of the field of "Learning Analytics" is to support students in their learning process, we developed a Web application called *Cite4Me*[7] that assists students in making decisions to find scientific publications and identify relevant research topics.

Cite4Me implements semantic and co-occurrence methods to (a) search and retrieve scientific publications; and (b) recommend scientific publications. Moreover, it provides a Web interface that facilitates the search for publications and may help users on discovering related terms to a given query.

In this section, we provide an overview of the major features of the Web application and its Web interface that assist users to explore scientific data on the Web.

2.1 Search and Retrieval

Cite4Me relies on search functionalities to meet the users needs. Briefly, we implemented standard Information Retrieval (**IR**) and Semantic Web (**SW**) approaches to retrieve and recommend scientific papers to the users. We divided this subsection into (i) free text

[1]Society for Learning Analytics Research - http://www.solaresearch.org
[2]http://linkedup-project.eu/

[3]http://www.acm.org/about/class/
[4]http://dl.acm.org
[5]http://www.elsevier.com
[6]http://dbpedia.org
[7]http://www.cite4me.com/

LAK-Data Challenge '13 Leuven, Belgium

search; (ii) exploratory search; and (iii) semantic search.

2.1.1 Free Text Search

The purpose of the *free text search* functionality is to offer users the abilities to search for mentions, titles and authors of academic publications contained in the LAK dataset. Even though, this functionality is similar to existing digital libraries, we agree that this is a basic functionality that must be provided by our application. Therefore, we use standard vector space models (*tf-idf*) for indexing and retrieving documents.

The *tf-idf* scores were computed for each term extracted from the publication content after applying stemming [14]. Furthermore, the searching functionality offers boolean queries with standard operators, such as 'OR', 'AND', and also a ranking of the matching publications based on the sum of *tf-idf* scores from the individual query terms.

In summary, our free text search provides to the users publications (P) that match query terms and non-matching publications P', which are related to P according to a degree of similarity (see Eq. 1), but does not contain the query terms.

The similarity between a matching publication P and other non-matching publication P' in the LAK dataset is measured by the standard *cosine similarity measure*, which is built on top of the computed *tf-idf* scores.

$$Sim(P, P') = \frac{P \cdot P'}{|P||P'|} \tag{1}$$

where P and P' represent the *tf-idf* scores for the terms in two distinct publications.

2.1.2 Exploratory Search

In this section, we provide detailed insights on the *exploratory search* functionality of our application. As a preliminary step to provide analytics and information about the actual content and topics coverage, all the scientific publications contained in the LAK dataset are previously enriched. The enrichment process was performed using DBpedia Spotlight API[8], where entities, entity types and their respective categories were extracted.

After the enrichment process, we cluster the publications according to entities and its categories found in each document. The publications are clustered in a tree-based structure over the enrichments. Note that, each node of the tree represents a topic in which a publication under this node covers. Thus, the exploratory search is performed through the topics covered by each publication.

The process of linking publications, categories and extra resources is mediated by DBpedia knowledge graph, where we use the *dcterms:subject* property to match the resources.

Thus, as a result, the exploratory search provides a way to explore resources through the connections between their topics, which facilitates the search for topically related resources. Figure 1 shows the exploratory search.

2.1.3 Semantic Search

Cite4Me provides also a semantic search engine that assists users to find publications semantically related to the query terms. Analogously to *explicit semantic analysis* (ESA) technique [5], the relatedness score, is computed between the enriched concepts found in the publications' content.

Basically, the semantic search is an adaptation of the *free text search* presented in the Section 2.1.1. Instead of computing the *tf-idf* scores for the words in a text, it computes the *tf-idf* score for the entities contained in a publication. Finally, the ranking of the results is based on the sum of the *tf-idf* scores of the matching concepts.

Figure 2 illustrates the semantic search functionality. It also generates a tag cloud from matching publications, showing the most prominent terms for a given query. Specifically, the tag cloug based on the results helps the users to have an insight about the topics and may assist in finding related terms previously unknown by them.

2.2 Paper recommendation

Another key feature of our system is the paper recommendation based on semantic relationships extracted from reference datasets. The recommendation is based on a previous work [12, 11], where we exploit the number of paths and the distance (length of a path) between given entities to compute a relatedness score between extracted entities and associated documents. The first step to measure the relatedness between documents is to compute the *semantic connectivity score* (SCS_e) of the entities found in each text (see Eq. 2).

$$SCS_e(a, b) = \sum_{l=1}^{\tau} \beta^l \cdot |paths^{<l>}_{(a,b)}| \tag{2}$$

where $|paths^{<l>}_{(a,b)}|$ is the number of paths between a and b of length l and $0 < \beta \leq 1$ is a positive damping factor. As in [12, 11], we used $\beta = 0.5$ as our damping factor. Furthermore, we also constrained the length of a path to $\tau = 4$.

Based on the score for entities, we then define the *semantic connectivity score* (SCS_w) between two documents W_1 and W_2 as follows:

$$SCS_w(W_1, W_2) = \left(\sum_{\substack{e_1 \in E_1 \\ e_2 \in E_2 \\ e_1 \neq e_2}} SCS_e(e_1, e_2) + \frac{|E_1 \cap E_2|}{2} \right) * \frac{1}{|E_1| * |E_2|} \tag{3}$$

where E_i is the set of entities associated with W_i, for $i = 1, 2$. Note that documents that contain the same entities receive an extra bonus (the second term on the right-hand side of Eq. 3).

Thus, a list of documents pairs is generated and ranked according the score and suggested to the user. Figure 3 illustrates the paper recommendation process computed based on SCS_w.

3. DATASET DISCOVERY AND INTERLINKING

This section briefly describes the datasets used on automatic related data discovery from DataHub[9] and future steps on dataset discovery and interlinking.

3.1 LAK Dataset

The LAK dataset contains the metadata of the papers published in the proceedings of LAK conference 2011-12, a special issue of Learning and Knowledge Analytics: Educational Technology & Society, the proceedings of the International Conference on Educational Data Mining (2008-12) and the Journal of Educational Data Mining (2008-12). In total, 315 descriptions of papers containing detailed information about authors, institutions, conference venues and the full content of the paper were available.

3.2 Data Analysis

The goal of the data analysis procedure is to align the various publications in the LAK dataset based on mutual information, such

[8] http://dbpedia.org/spotlight

[9] http://www.datahub.io

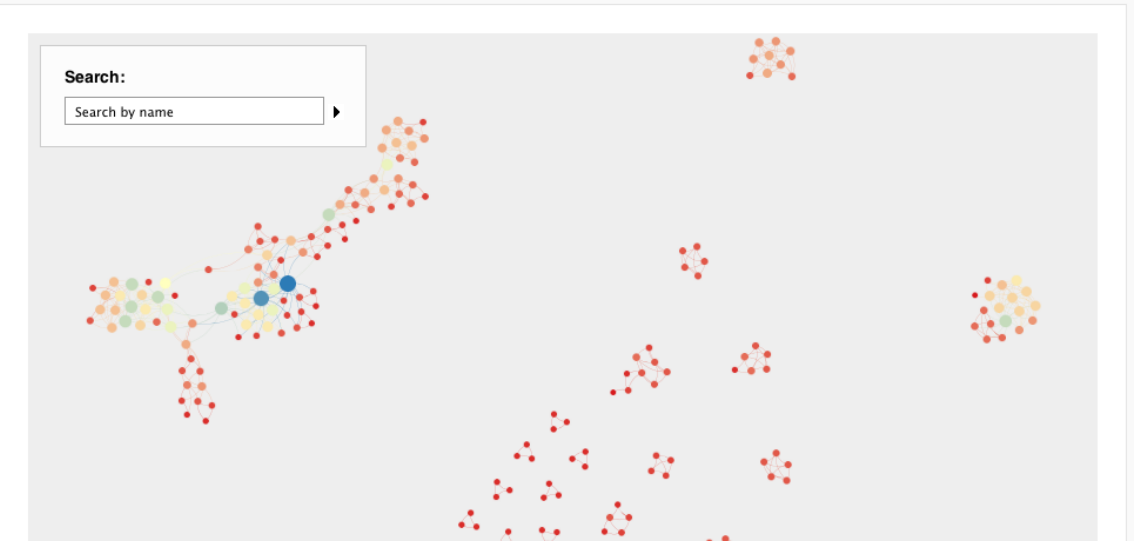

Figure 1: Preview of the exploratory search funcionality.

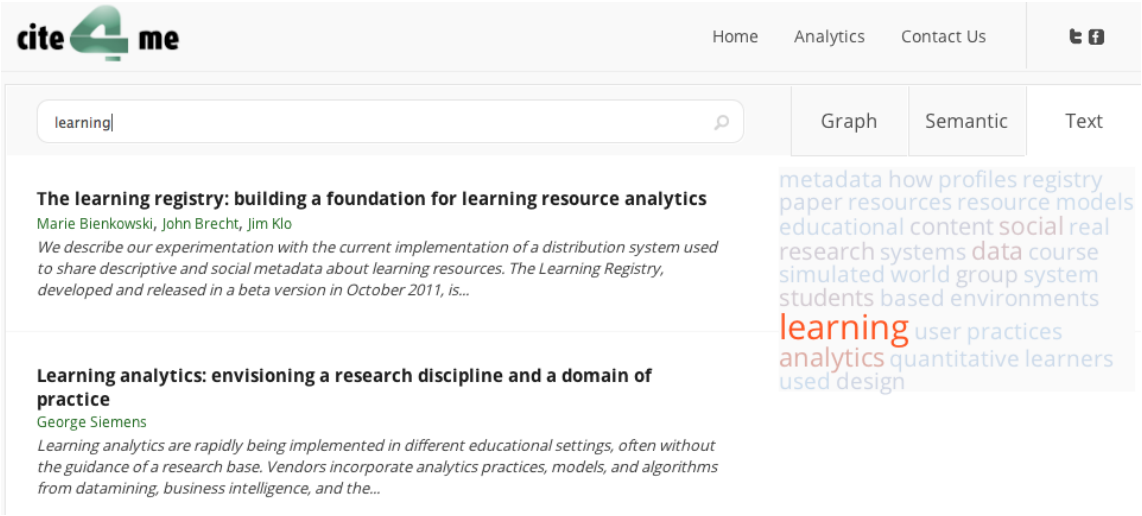

Figure 2: Preview of the semantic search funcionality.

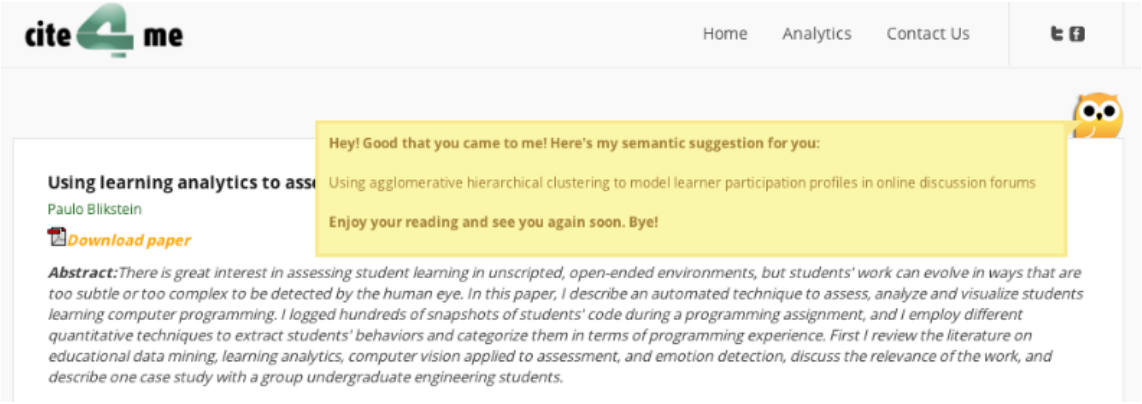

Figure 3: An example of paper recommendation based on SCS_w.

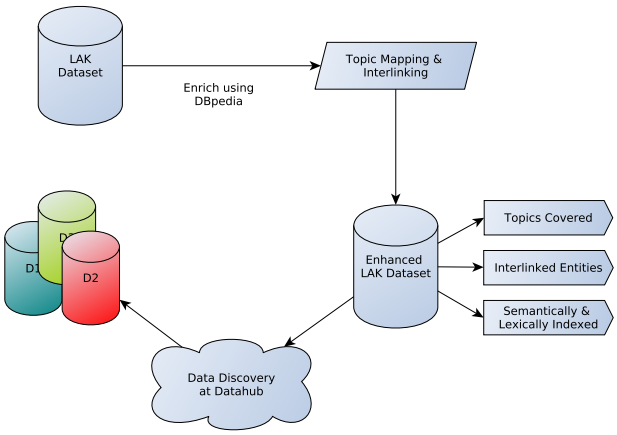

Figure 4: Relevant Dataset Discovery Framework based on the generated *feature set* used to query DataHub Linked Data provider.

as the topics covered by them. This is achieved using well established datasets like DBpedia[10] and Freebase[11], where a reference point for the unstructured textual content of publications is created through an enrichment process.

Again, the enrichment process is carried out using DBpedia Spotlight[12] [10] and addresses several issues of significant importance. For instance, it offers several advantages such as: (i) identification of (common) named entities, (ii) disambiguation; and (iii) expansion of the limited dataset and resource descriptions with additional background knowledge.

3.3 Data Discovery

Our Web application uses as its starting point the instances in the LAK dataset to automatically explore and recommend to users, datasets that covers similar topics. In order to query, detect and interlink related datasets, we chose the DataHub as a data provider. DataHub serves as a collecting point of datasets from various fields and currently it has over 5000 datasets. Note that, from the large number of datasets, only 300 datasets are provided as Linked Open Data. As the latter is the main focus of our work, the analysis and interlinking process is focused for such datasets.

Briefly, the data discovery is performed using CKAN[13] data management framework from DataHub, where based on data analysis and user interests (such as topics covered by a publication/resource) related datasets are suggested.

Additionally, we provide to the user a set of resources, amongst other data analytics, that enables the user to harvest and correlate new information from the discovered resources, considering the LAK dataset as a starting point of such discovery.

This approach presents several advantages such as the adoption and the widespread use of Linked Data principles for publishing scientific papers. Nowadays, many conferences make their proceedings and journals freely accessible, hence our approach would take advantage of such open data and offer users topically relevant papers for a particular resource in the LAK dataset.

[10] `http://dbpedia.org`
[11] `http://www.freebase.com/`
[12] `http://spotlight.dbpedia.org/`
[13] http://www.ckan.org

4. EVALUATION OF DATA ANALYSIS AND DATA DISCOVERY

This section presents an overview of the results obtained by analyzing the LAK dataset with respect to the constructed *feature set* that describes topics covered by individual publications. Moreover, based on the data analysis procedure and shared information, we show that the establishment of links between the different publications within the LAK dataset and from other datasets in DataHub is possible.

In the following subsections, we show the analysis of the LAK dataset and the discovery of relevant datasets and publications.

4.1 Data Analysis

The data analysis of the LAK dataset focuses mostly on assessing the individual publications for their *topic coverage*. In this manner, we build a connected data graph consisting of the individual publications and items from the *feature set*. This step is necessary to provide the *exploratory search* functionality, where based on the established edges between publications and *feature set* items, we can navigate through the publications or topics of interest. Therefore, the results obtained with respect to the constructed *feature set* and LAK dataset graph are shown in what follows.

Table 1 shows the top ranked items for each of the *feature sets*, along with the number of associations an item has with respect to all publications (*entity, category* and *type* items). Figure 5 shows the constructed data graph for the LAK dataset.

4.2 Data Discovery

After creating the *feature set* based on the information provided from reference datasets, we are able to query for relevant datasets in DataHub.

Thus, for the top ranked *feature set* items, the data discovery for relevant resources is considered. Table 2 shows the discovered resources and datasets for the top-10 *entity* items. Note that, we focus only on *bibliographic* datasets, since we aim at recommending topically related scientific publications. Due to the lack of *bibliographic* datasets, we were not able to find related publications for all *entities* considered. Table 2 summarizes the discovered resources. The dataset names are represented by their acronyms as follows: **b3kat** - *"Bayerische Staatsbibliothek"*, **hebis** - *"Hessisches Bibliotheks Informations System"* and **npg** - *"Nature Publishing Group - ALL"*.

Additionally, from the set of 96 *bibliographic* datasets available, only a few of them were offered as Linked Data, thus narrowing our search space for relevant resources.

Entity	b3kat	hebis	npg
Data	14	0	12
Learning	5	0	1
Data mining	0	0	0
Algorithm	4	0	0
Education	17	1	2
Analysis	42	1	6
Student	7	1	1
Knowledge	11	0	0
Methodology	4	0	1
Statistics	7	0	1

Table 2: Number of discovered resources from the *bibliographic* group for the top ranked items from the *entity feature set*, based on the LAK dataset.

Entity	Assoc.	Category	Assoc.	Type	Assoc.
Data	90	Educational_psychology	161	DBpedia:TopicalConcept	150
Learning	80	Data_analysis	150	Freebase:/book	142
Data_mining	67	Learning	139	Freebase:/book/book_subject	142
Algorithm	50	Scientific_method	137	Freebase:/media_common	138
Education	49	Neuropsychological_assessment	136	Freebase:/media_common/quotation_subject	136
Analysis	48	Greek_loanwords	135	Freebase:/computer	125
Student	46	Data	131	Freebase:/education	122
Knowledge	46	Evaluation_methods	129	Freebase:/education/field_of_study	120
Methodology	42	Computer_data	126	Freebase:/computer/software_genre	120
Statistics	41	Research_methods	124	Freebase:/internet	118
System	37	Systems_science	118	Freebase:/internet/website_category	118
Scientific_modelling	37	Formal_sciences	108	Freebase:/award	114
Prediction	36	Data_management	108	Freebase:/media_common/media_genre	105
Data_set	36	Cognitive_science	107	Freebase:/organization	103
Statistical_classification	30	Statistical_terminology	107	Freebase:/award/award_discipline	103
Evaluation	29	Developmental_psychology	101	Freebase:/business	102
Standard_deviation	29	Intelligence	93	Freebase:/organization/organization_sector	99
Probability	28	Data_mining	91	Freebase:/people	99
Behavior	26	Critical_thinking	87	Freebase:/film	94
Interaction	24	Thought	84	Freebase:/book/periodical_subject	93

Table 1: Top ranked items from the *feature set* for the LAK Dataset, from the dataset analysis.

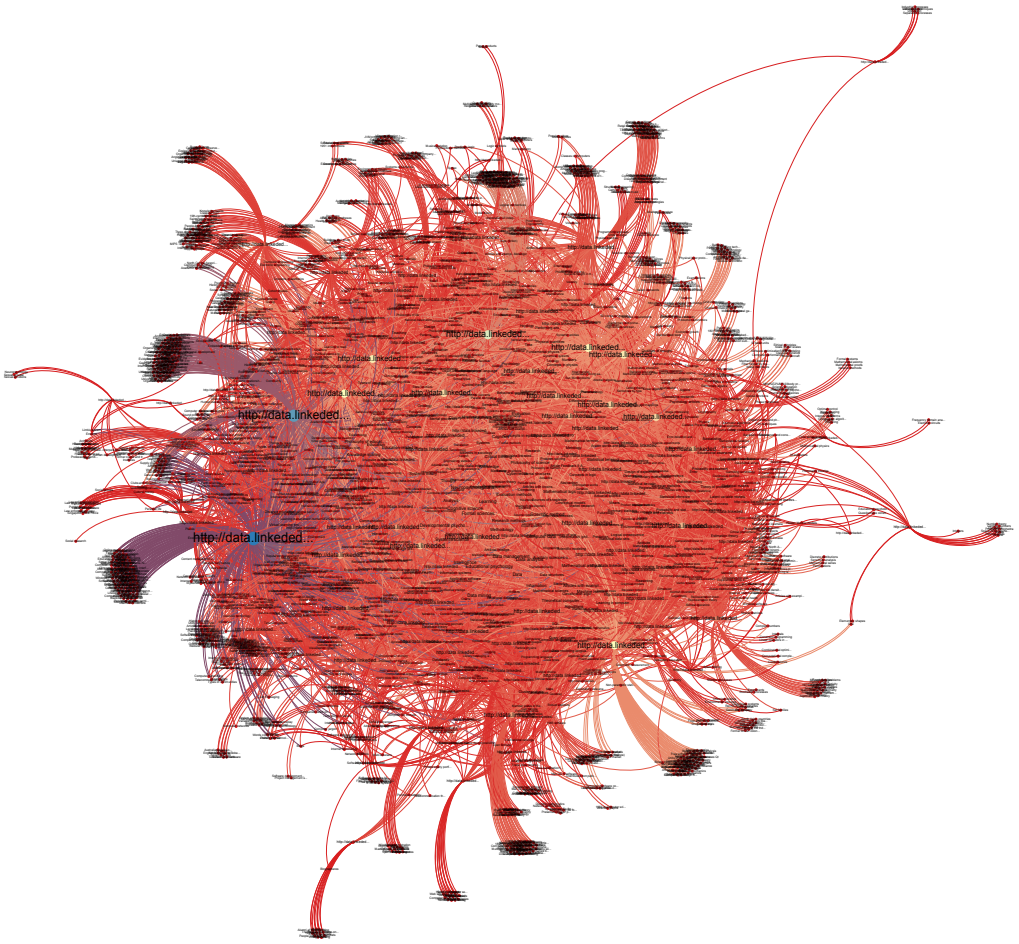

Figure 5: Topic coverage of LAK data graph for the individual resources.

5. RELATED WORK

Cobo et al.[3] presents an analysis of student participation in online discussion forums using an agglomerative hierarchical clustering algorithm, and explore the profiles to find relevant activity patterns and detect different student profiles. Barber et al. [1] uses a predictive analytic model to prevent students from failing in courses. They analyze several variables, such as grades, age, attendance and others, that can impede the student learning.Kahn et al. [7] present a long-term study using hierarchical cluster analysis, t-tests and Pearson correlation that identified seven behavior patterns of learners in online discussion forums based on their access. García-Solórzano et al. [6] introduce a new educational monitoring tool that helps tutors to monitor the development of the students. Unlike traditional monitoring systems, they propose a faceted browser visualization tool to facilitate the analysis of the student progress. Glass [8] provides a versatile visualization tool to enable the creation of additional visualizations of data collections.

Essa et al. [4] utilize predictive models to identify learners academically at-risk. They present the problem with an interesting analogy to the patient-doctor workflow, where first they identify the problem, analyze the situation and then prescribe courses that are indicated to help the student to succeed. Siadaty et al.[13] present the Learn-B environment, a hub system that captures information about the users usage in different softwares and learning activities in their workplace and present to the user feedback to support future decisions, planning and accompanies them in the learning process.

In the same way, McAuley et al. [9] propose a visual analytics to support organizational learning in online communities. They present their analysis through an adjacency matrix and an adjustable timeline that show the communication-actions of the users and is able to organize it into temporal patterns. Bramucci et al. [2] presents Sherpa an academic recommendation system to support students on making decisions. For instance, using the learner profiles they recommend courses or make interventions in case that students are at-risk.

In the related work, we showed how different perspectives and the necessity of new tools and methods to make data available and help decision-makers.

6. CONCLUSION

In this paper we presented the main features of the *Cite4Me* Web application. *Cite4Me* makes use of several data sources to provide information for users interested on scientific publications and its applications.

Additionally, we provided a general framework on data discovery and correlated resources based on a constructed *feature set*, consisting of items extracted from reference datasets. It made possible for users, to search and relate resources from a dataset with other resources offered as Linked Data.

For more information about the *Cite4Me* Web application refer to `http://www.cite4me.com`.

7. REFERENCES

[1] R. Barber and M. Sharkey. Course correction: using analytics to predict course success. In *Proc. of the 2nd International Conference on Learning Analytics and Knowledge*, LAK '12, pages 259–262, New York, NY, USA, 2012. ACM.

[2] R. Bramucci and J. Gaston. Sherpa: increasing student success with a recommendation engine. In *Proc. of the 2nd International Conference on Learning Analytics and Knowledge*, LAK '12, pages 82–83, New York, NY, USA, 2012. ACM.

[3] G. Cobo, D. García-Solórzano, J. A. Morán, E. Santamaría, C. Monzo, and J. Melenchón. Using agglomerative hierarchical clustering to model learner participation profiles in online discussion forums. In *Proc. of the 2nd International Conference on Learning Analytics and Knowledge*, LAK '12, pages 248–251, New York, NY, USA, 2012. ACM.

[4] A. Essa and H. Ayad. Student success system: risk analytics and data visualization using ensembles of predictive models. In *Proc. of the 2nd International Conference on Learning Analytics and Knowledge*, LAK '12, pages 158–161, New York, NY, USA, 2012. ACM.

[5] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proc. of the 20th international joint conference on Artifical intelligence*, IJCAI'07, pages 1606–1611, San Francisco, CA, USA, 2007. Morgan Kaufmann Pub. Inc.

[6] D. García-Solórzano, G. Cobo, E. Santamaría, J. A. Morán, C. Monzo, and J. Melenchón. Educational monitoring tool based on faceted browsing and data portraits. In *Proc. of the 2nd International Conference on Learning Analytics and Knowledge*, LAK '12, pages 170–178, New York, NY, USA, 2012. ACM.

[7] T. M. Khan, F. Clear, and S. S. Sajadi. The relationship between educational performance and online access routines: analysis of students' access to an online discussion forum. In *Proc. of the 2nd International Conference on Learning Analytics and Knowledge*, LAK '12, pages 226–229, New York, NY, USA, 2012. ACM.

[8] D. Leony, A. Pardo, L. de la Fuente Valentín, D. S. de Castro, and C. D. Kloos. Glass: a learning analytics visualization tool. In *Proc. of the 2nd International Conference on Learning Analytics and Knowledge*, LAK '12, pages 162–163, New York, NY, USA, 2012. ACM.

[9] J. McAuley, A. O'Connor, and D. Lewis. Exploring reflection in online communities. In *Proc. of the 2nd International Conference on Learning Analytics and Knowledge*, LAK '12, pages 102–110, New York, NY, USA, 2012. ACM.

[10] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. Dbpedia spotlight: shedding light on the web of documents. In *Proc. of the 7th International Conference on Semantic Systems*, I-Semantics '11, pages 1–8, New York, NY, USA, 2011. ACM.

[11] 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).

[12] B. Pereira Nunes, R. Kawase, S. Dietze, D. Taibi, M. A. Casanova, and W. Nejdl. Can entities be friends? In G. Rizzo, P. Mendes, E. Charton, S. Hellmann, and A. Kalyanpur, editors, *Proc. of the Web of Linked Entities Workshop in conjuction with the 11th International Semantic Web Conference*, volume 906 of *CEUR-WS.org*, pages 45–57, Nov. 2012.

[13] M. Siadaty, D. Gašević, J. Jovanović, N. Milikić, Z. Jeremić, L. Ali, A. Giljanović, and M. Hatala. Learn-b: a social analytics-enabled tool for self-regulated workplace learning. In *Proc. of the 2nd International Conference on Learning Analytics and Knowledge*, LAK '12, pages 115–119, New York, NY, USA, 2012. ACM.

[14] C. van Rijsbergen, S. Robertson, and M. Porter. New models in probabilistic information retrieval. 1980.