RecLAK: Analysis and Recommendation of Interlinking Datasets

Giseli Rabello Lopes
Departament of Informatics, PUC-Rio
Rio de Janeiro/RJ, Brazil
grlopes@inf.puc-rio.br

Luiz André P. Paes Leme
Computer Science Institute, UFF
Niterói/RJ, Brazil
lapaesleme@ic.uff.br

Bernardo Pereira Nunes
Departament of Informatics, PUC-Rio
Rio de Janeiro/RJ, Brazil
bnunes@inf.puc-rio.br

Marco A. Casanova
Departament of Informatics, PUC-Rio
Rio de Janeiro/RJ, Brazil
casanova@inf.puc-rio.br

ABSTRACT

This paper presents the *RecLAK*, a Web application developed for the LAK Challenge 2014. *RecLAK* focuses on the analysis of the LAK dataset metadata and provides recommendations of potential candidate datasets to be interlinked with the LAK dataset. *RecLAK* follows an approach to generate recommendations based on Bayesian classifiers and on Social Networks Analysis measures. Furthermore, *RecLAK* generates graph visualizations that explore the LAK dataset over other datasets in the Linked Open Data cloud. The results of the experiments contribute to the understanding and improvement of the LAK dataset. Furthermore, it can also help researchers of the fields covered by LAK dataset, such as learning analytics and educational data mining.

1. INTRODUCTION

The effort of publishing Linked Data has been accompanied by the creation of catalogs of Linked Data datasets, such as the *DataHub*[1], to make data findable and reusable. However, despite the fact that extensive lists of open datasets are available in these catalogs, most of the data publishers typically link their datasets only to popular ones, such as DBpedia[2], Freebase[3] and Geonames[4]. Although the linkage to popular datasets allows the exploration of external resources, it fails to cover more specialized data.

As a practical example of this scenario, we may highlight the LinkedUp project[5], which is an initiative that aims at providing educational organizations and institutions with a collection of open data available on the Web. One of the datasets covered by the LinkedUp project is the *Learning Analytics and Knowledge* (LAK) dataset. The LAK dataset, referred to as *lak*, provides access to structured fulltext and metadata from key research publications in the field of learning analytics and educational data mining[6]. *lak* is regularly updated with data, for instance, from the LAK (Learning Analytics and Knowledge) and EDM (Educational Data Mining) conference series. According to the DataHub metadata, *lak* was not linked to other datasets, except DBpedia. However, an exploratory search in the DataHub in fact revealed related datasets that *lak* could be linked to, such as other bibliographic datasets.

This scenario is very common. Most of the published datasets are still awaiting to be linked and, therefore, they do not fulfill the requirements to be considered 5-star [1] and fail to take advantage of other data. Basically, as argued in [10], the linkage to popular datasets is favoured for two main reasons: the difficulty of finding related open datasets; and the strenuous task of discovering instance mappings between different datasets.

In this sense, *lak* will be explored as a case study. The recommendation challenge associated to the interlinking of *lak* in the LOD can be posed by considering two main questions:

Q1. For a dataset d, published in the LOD, is it interesting for the publisher of d to try to link it to *lak*?

Q2. For a dataset d, published in the LOD, is it interesting for the *lak* administrator to try to link his dataset to d?

In more detail, let t and d_i be two datasets. A *link* from t to d_i is a triple of the form (s, p, o) such that s is defined in t and o is defined in d_i. We say that t is *linked to* d_i, or that d_i is *linked from* t, iff there is at least a link from t to d_i. We also say that d_i is *relevant* for t iff there is at least a resource defined in d_i that can be linked from a resource defined in t.

[1] `http://datahub.io/`
[2] `http://dbpedia.org`
[3] `http://www.freebase.com`
[4] `http://www.geonames.org/`
[5] `http://linkedup-project.eu`
[6] `http://lak.linkededucation.org`

Questions Q1 and Q2 are special cases of the *dataset interlinking recommendation problem* posed as follows:

> *Given a finite set of datasets $\boldsymbol{D}$ and a dataset t, compute a rank score for each dataset $d_i \in \boldsymbol{D}$ such that the rank score of d_i increases with the chances of d_i being relevant for t.*

In this paper, we first introduce two rank score functions to address the dataset interlinking recommendation problem. Then, we apply the functions to answer question Q2.

The remainder of this paper is organized as follows. Section 2 presents related work. Section 3 briefly describes the recommendation approaches. Section 4 shows the result analysis of the metadata exploration and the generated recommendations. Finally, Section 5 presents some final remarks.

2. RELATED WORK

In this paper, we use an extended version [5] of previous work [3, 4], that introduced the rank score functions based on the Bayesian and the Social Network approaches. The extended version also explores different sets of features related to the metadata of the datasets, such as properties, classes and vocabularies, to compute the rank score functions.

Nikolov et al. [9, 10] propose an approach to identify relevant datasets for interlinking applying keywords searches and ontology matching techniques. Kuznetsov [2] describes a linking system, which is responsible for discovering relevant datasets for a given dataset and for creating instance level linkage. When compared with these approaches, the rank score functions applied in this paper use only metadata and are, therefore, much simpler to compute and yet achieve a good performance [5].

Lóscio et al. [6] and Wagner et al. [15] propose techniques to find relevant datasets for user queries. The first approach is based on information quality criteria of correctness, schema completeness and data completeness while the second one is based on the overlapping of sets of instances of datasets. Oliveira et al. [13] use application queries and user feedback to discover relevant datasets. These papers aim at recommending datasets with respect to user queries, which is a problem close, but not identical to the problem discussed in this paper.

Nunes et al. [11, 12] performed several analysis on *lak* but their focus was mainly in the dataset content. They also proposed other datasets to be interlinked with *lak* considering their links with DBpedia. By contrast, this paper focuses on analyzing the metadata for creating rankings of candidate datasets to be interlinked with *lak* using different recommendation techniques.

3. RECOMMENDATION APPROACHES

3.1 Bayesian ranking

A rank score function, inspired on conditional probabilities, that induces the ranking of the datasets in $\boldsymbol{D}$ (from the largest to the smallest score), can be defined as follows:

$$score(d_i, t) = \left(\sum_{j=1..n} log(P(F_j|D_i)) \right) + log(P(D_i)) \quad (1)$$

Based on the maximum likelihood estimate of the probabilities [8] in a training set of datasets, the above probabilities can be estimated as follows:

$$P(F_j|D_i) = \frac{count(f_j, d_i)}{\sum_{j=1}^{n} count(f_j, d_i)} \;\; ; \;\; P(D_i) = \frac{count(d_i)}{\sum_{i=1}^{m} count(d_i)}$$

where $count(f_j, d_i)$ is the number of datasets in the training set that have feature f_j and that are linked to d_i, and $count(d_i)$ is the number of datasets in the training set that are linked to d_i, disregarding the feature set.

For the score function computation, some auxiliary functions help to avoid computing $log(0)$ replacing this value by c, which is a constant small enough to penalize the datasets d_i that do not have datasets with features F_j linked to them or that do not have links from other datasets [5]. Thus, the idea is that, if the set of features of t is very often correlated with datasets that are linked to d_i and t is not already linked to d_i, then it is recommended to try to link t to d_i.

3.2 Social Network-based ranking

We propose to analyze the dataset interlinking recommendation problem in much the same way as the link prediction problem in Social Networks [7]. Analogously, the *Linked Data network* for $\boldsymbol{D}$ is a directed graph such that the nodes are the datasets in $\boldsymbol{D}$ and there is an edge between datasets u and v in $\boldsymbol{D}$ iff there is a link from u to v. To obtain more accurate results, we combine two measures, Preferential Attachment (pa) and Resource Allocation (ra), into a single score [5], defined as follows:

$$score(t, d_i) = ra(t, d_i) + \frac{pa(t, d_i)}{|\boldsymbol{D}|} \quad (2)$$

$$pa(t, d_i) = |P_{d_i}| \;\; ; \;\; ra(t, d_i) = \sum_{d_j \in S_t \cap P_{d_i}} \frac{1}{|P_{d_j}|}$$

where P_{d_i} is the *popularity set* of a dataset $d_i \in \boldsymbol{D}$, that is, the set of all datasets in $\boldsymbol{D}$ that have links to d_i, and S_t is the *similarity set* of a dataset t, that is, the set of all datasets in $\boldsymbol{D}$ that have features in common with t.

The combined score induces the ranking of the datasets in $\boldsymbol{D}$ (from the largest to the smallest score) and gives priority to the ra score; the pa score, normalized by the total number of datasets to be ranked ($|\boldsymbol{D}|$), will play a role when there is a tie or when the ra value is zero.

4. RESULT ANALYSIS

4.1 Data used in the experiments

We selected a subset of the datasets indexed by the DataHub, using the *Learning Analytics and Knowledge* dataset [14] as the target of the recommendation. From the DataHub catalog, we managed to obtain 295 datasets with at least one

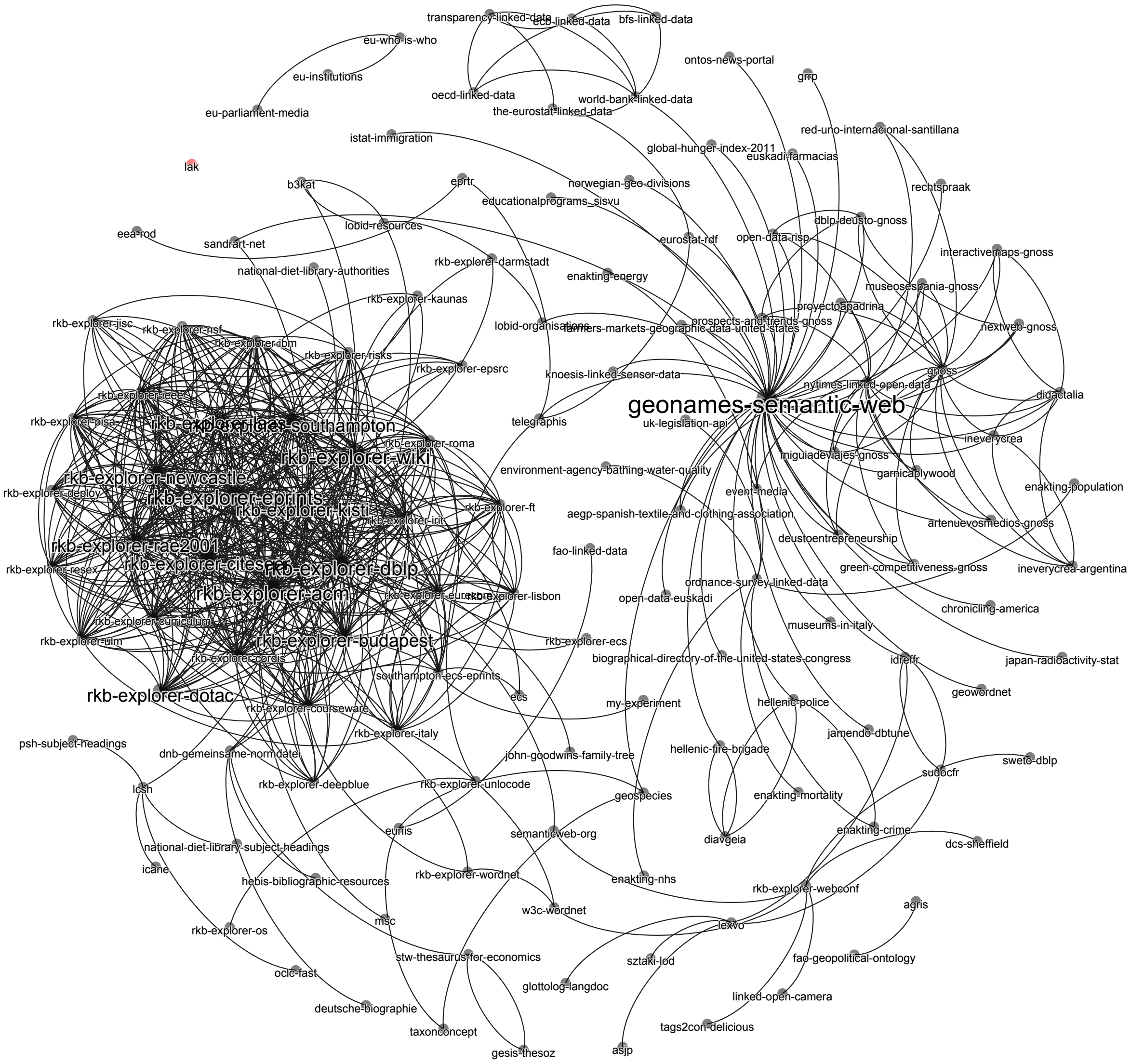


Figure 1: The datasets and their links.

feature (class, property or vocabulary). Among the datasets with links defined, there are 139 datasets with 697 known links. Figure 1 presents a graph representing the datasets and their known links. In this graph, the size of a dataset node is proportional to the number of datasets linked to it (*in-degree*).

The number of distinct features between classes and properties was 11,868. The number of relations between datasets and classes or properties was 16,750, where 6,447 were references to classes and 10,303 were references to properties. For the details on how we extracted metadata from the DataHub catalog, see [5].

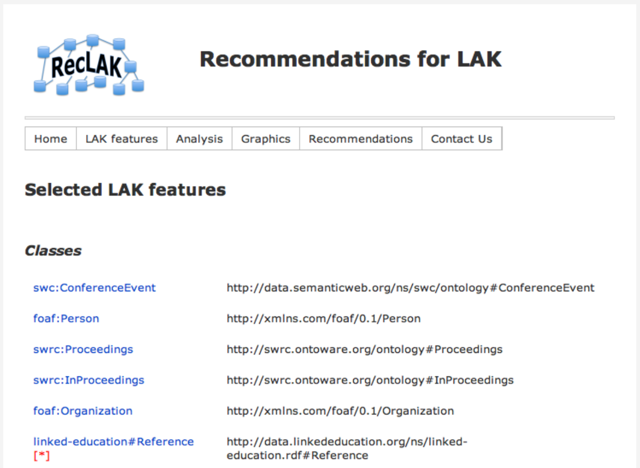
4.2 LAK features

As features of *lak*, we used a selected set of classes and properties obtained from the *lak* and from the LinkedUp project Web site. We filtered out, from 51 initial features, those that were not related to the content of the dataset and that are used in many datasets, such as *owl:sameAs*, *rdf:Property*, *rdfs:Resource*, among others. The core of the selected set comes from the SWC ontology[7] (*Semantic Web Conference*), which describes academic conferences and establishes a convention on how to use classes and properties from other ontologies, mostly FOAF (*Friend of a Friend*), for people and organizations, and SWRC (*Semantic Web for Research Communities*), for papers. It also includes metadata from other ontologies, such as SIOC (*Semantically-Interlinked Online Communities*) and DC (*Dublin Core*). The selected *lak* features added to 37, where 31 of them are shared by other datasets in our set of data. A preview of the *RecLAK* interface showing the selected *lak* classes is presented in Figure 2.

4.3 Datasets with LAK features

The set of datasets (represented by their *id* in DataHub) that have at least one feature in common with *lak* consists

[7] `http://data.semanticweb.org/ns/swc/ontology`



Recommendations for LAK

Selected LAK features

Classes

swc:ConferenceEvent	http://data.semanticweb.org/ns/swc/ontology#ConferenceEvent
foaf:Person	http://xmlns.com/foaf/0.1/Person
swrc:Proceedings	http://swrc.ontoware.org/ontology#Proceedings
swrc:InProceedings	http://swrc.ontoware.org/ontology#InProceedings
foaf:Organization	http://xmlns.com/foaf/0.1/Organization
linked-education#Reference [*]	http://data.linkededucation.org/ns/linked-education.rdf#Reference

Figure 2: Preview of the *RecLAK* interface showing the selected *lak* classes.

Table 1: Top 10 datasets sharing features with *lak*.

Dataset *id*	# shared features
rkb-explorer-webconf	31
linked-open-vocabularies-lov	8
krystian-pietruszka	7
aksworg	7
dcs-sheffield	6
southampton-ac-uk-profile	6
jamendo-dbtune	6
sudocfr	6
rkb-explorer-webscience	6
msc	6

of 132 datasets, with 376 associations between datasets and *lak* features. Figure 3 presents a graph representing the datasets and their associated *lak* features. In this graph, the size of a feature node is proportional to the number of datasets having it.

Among the *lak* features, the most popular are from DC: *dc:title*, shared by 60 datasets, and *dc:creator*, with 56 datasets references, and from FOAF: *foaf:name* and *foaf:homepage* with, respectively, 41 and 36 other datasets beyond *lak* referring to them. The least popular features are metadata directly from SWC and SWRC ontologies (some of them used by only 1 dataset other than *lak*).

The datasets with more than 5 features shared with *lak* are shown in Table 1. The more expressive result is obtained by the *rkb-explorer-webconf* dataset which shares 31 features with *lak*. This was the most correlated dataset with the selected classes and properties of *lak*. The *rkb-explorer-webconf* is a semantic repository that publishes RDF linked data and co-reference information from the RKB Explorer initiative. This dataset includes information about authors and publications in several conferences, such as ESWC.

4.4 Dataset Interlinking recommendations

Using the score functions, briefly described in Section 3, we generated recommendations for *lak*. A preview of the *RecLAK* interface presenting the recommendations for LAK is presented in Figure 4.

The top 10 recommendations generated by each of the two approaches (Bayesian and Social Network-based rankings) and the respective score values estimated for each recommended dataset are presented in Table 2. The top 10 ranked datasets for each approach will be briefly described below.

Bayesian ranking. The topmost-ranked is a generic dataset with concepts from the Semantic Web community. Dataset #2 is a well-known lexical database of English. Datasets from #3 to #6 positions of the Bayesian ranking presented tied scores. Dataset #3 is a dataset with concepts from tags generated by human annotators. Dataset #4 describes people, research groups and publications of the members of the Computer Science Department at the University of Sheffield. Dataset #5 is maintained by the chamber of deputies in Italy, which is working to publish quality linked data in several domains, including research. Dataset #6 describes the DBLP digital library, which provides bibliographic information on major computer science journals and proceedings. *dblp* also indexes the papers published in the LAK and EDM conferences. Dataset #7 is the Geonames dataset, which contains information about geographical locations. Dataset #8 contains information about languages, words, characters, and other human language-related entities to the Linked Data Web and Semantic Web. *lexvo* has links to WordNet and *thesauris*. Dataset #9 is a Linked Data version of the Association for Computing Machinery (ACM) digital library. Finally, dataset #10 is a dataset of the Library of Congress Subject Headings (LCSH), which catalogs materials stored by the Library of Congress and other libraries around the United States.

Social Network-based ranking. Since, there is some overlap between the top 10 recommendations of Social Network-based (SN-based) and Bayesian ranking, we will comment the top 10 datasets ranked only by the SN-based approach. Dataset #2 publishes the news vocabularies used by *The New York Times* as Linked Open Data. It covers data and resources about people, locations and organizations. Dataset #3 covers topics related to innovation, technology, business and education. Dataset #6 has links catalogued in the DataHub for other bibliographic datasets such as Citeseer, DBLP, ACM, IEEE and EPrints. Dataset #7 was created with the objective of being capable of networking the wide range of resources and information held by libraries and other cultural institutions in German-speaking countries. This dataset uses established vocabularies, such as FOAF. Dataset #9 describes e-prints and has links catalogued in the DataHub for other bibliographic datasets such as Citeseer, DBLP, ACM and IEEE. Dataset #10 is also a Linked Data version of publications information of the DBLP digital library, similar to *sweto-dblp*.

Discussion. Based on the top 10 rankings of both approaches, we identified three main groups of candidate datasets that were recommended to be interlinked with *lak*:

- **generic:** *semanticweb-org*, *w3c-wordnet*, *tags2con-*

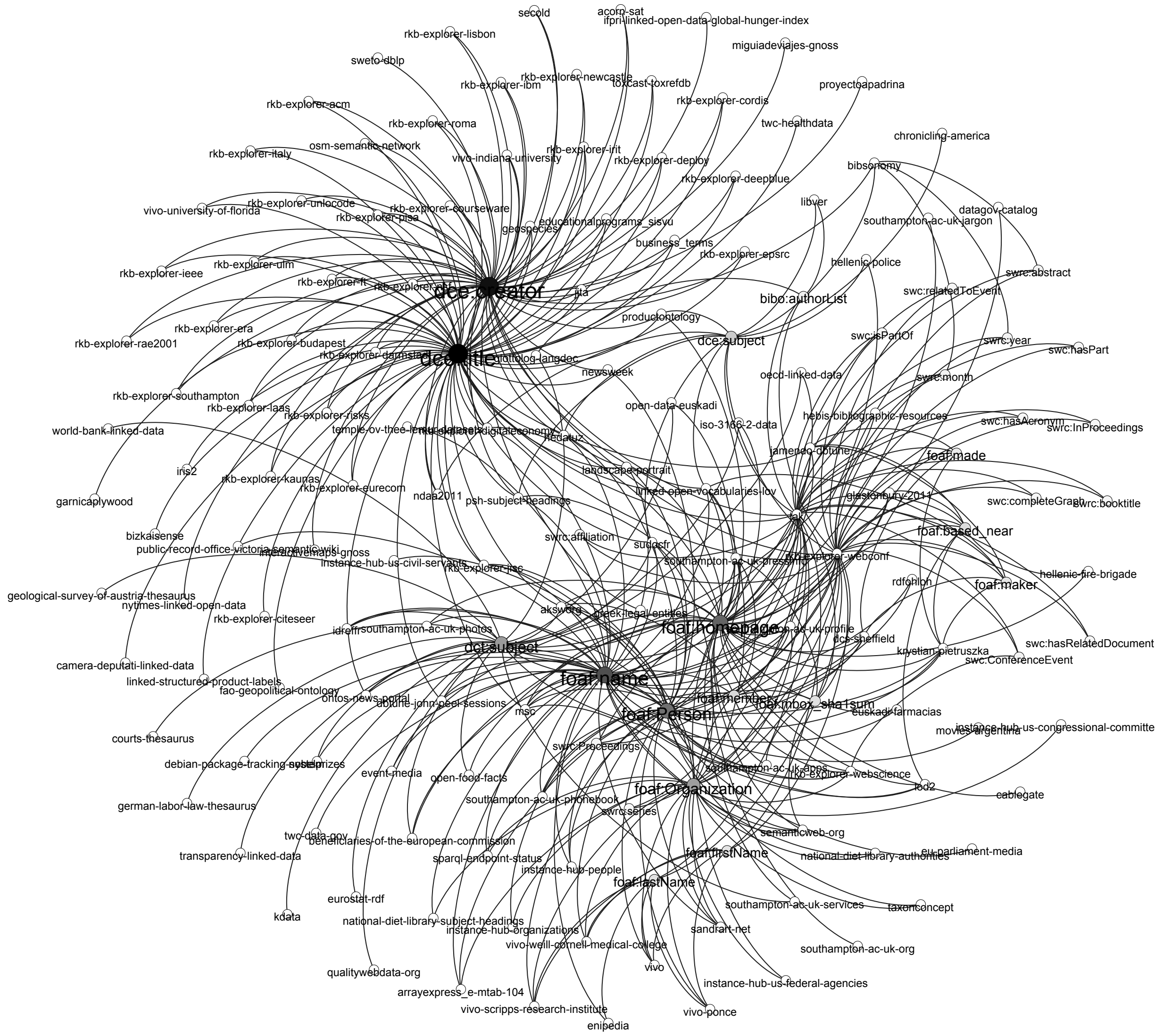


Figure 3: The datasets and their associated *lak* features.

delicious, *geonames-semantic-web*, *lexvo*, *nytimes-linked-open-data*, *rkb-explorer-wiki*

- **bibliographic:** *dcs-sheffiedl*, *linked-open-camera*, *sweto-dblp*, *rkb-explorer-acm*, *lcsh*, *dnb-gemeinsame-normdatei*, *rkb-explorer-eprints*, *rkb-explorer-dblp*

- **educational area:** *gnoss.*

The top 10 recommendations of the rankings differ in some aspects. Considering the groups identified above, the Bayesian ranking contains a higher number of generic datasets, while the Social Network-based ranking contains a higher number of bibliographic datasets. This probably happens because Bayesian ranking prioritizes recommendations for *lak* of datasets linked from the larger number of other datasets having the larger number of *lak* features. On the other hand, the Social Network-based ranking prioritizes the datasets pointed by the larger number of other datasets with smaller popularity and having at least one feature of *lak*.

The results also indicate that the selection of the feature set is very important because it directly influences the generated rankings and can lead to recommendations of datasets which are more as well as less generic. In our experiments with *lak*, we filtered out some generic features (e.g., *owl:sameAs*), but included DC and FOAF elements. Thus, we expected that both generic and specific datasets from our set of datasets were recommended. As the metadata used to triplify *lak* were not using classes and properties specifically related to the application domain, this characteristic was not evidenced in the recommendation results.

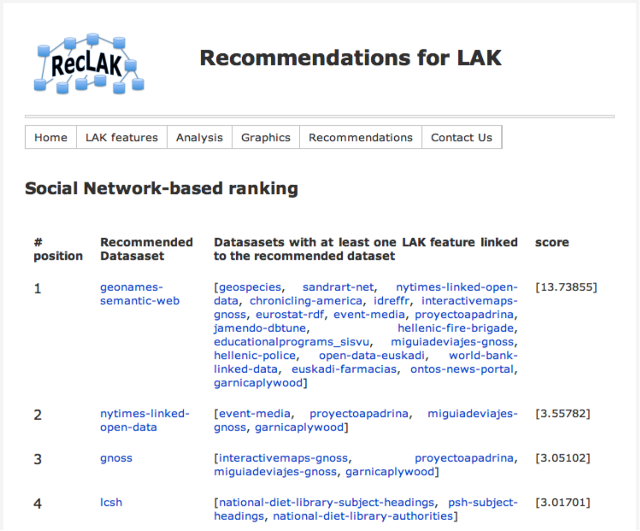
5. CONCLUSIONS

This paper presented a detailed analysis, based on Bayesian classifiers and on Social Network Analysis techniques, to address the dataset interlinking recommendation problem for *lak*, using only metadata. Thus, the rank score functions are

Table 2: Top 10 ranked recommendations for *lak*.

#	Bayesian ranking	score*	#	SN-based ranking	score
1	semanticweb-org	-162.025	1	geonames-semantic-web	13.738
2	w3c-wordnet	-162.236	2	nytimes-linked-open-data	3.558
3	tags2con-delicious	-163.025	3	gnoss	3.051
4	dcs-sheffield	-163.025	4	lcsh	3.017
5	linked-open-camera	-163.025	5	rkb-explorer-acm	2.430
6	sweto-dblp	-163.025	6	rkb-explorer-wiki	2.408
7	geonames-semantic-web	-3281.339	7	dnb-gemeinsame-normdatei	2.020
8	lexvo	-4107.754	8	lexvo	2.017
9	rkb-explorer-acm	-4114.493	9	rkb-explorer-eprints	1.632
10	lcsh	-4273.558	10	rkb-explorer-dblp	1.466

* Estimated using log_2, c=-170 and considering only *lak* features shared with at least one dataset.



Recommendations for LAK

Social Network-based ranking

# position	Recommended Datasaset	Datasasets with at least one LAK feature linked to the recommended dataset	score
1	geonames-semantic-web	[geospecies, sandrart-net, nytimes-linked-open-data, chronicling-america, idreffr, interactivemaps-gnoss, eurostat-rdf, event-media, proyectoapadrina, jamendo-dbtune, hellenic-fire-brigade, educationalprograms_sisvu, miguiadeviajes-gnoss, hellenic-police, open-data-euskadi, world-bank-linked-data, euskadi-farmacias, ontos-news-portal, garnicaplywood]	[13.73855]
2	nytimes-linked-open-data	[event-media, proyectoapadrina, miguiadeviajes-gnoss, garnicaplywood]	[3.55782]
3	gnoss	[interactivemaps-gnoss, proyectoapadrina, miguiadeviajes-gnoss, garnicaplywood]	[3.05102]
4	lcsh	[national-diet-library-subject-headings, psh-subject-headings, national-diet-library-authorities]	[3.01701]

Figure 4: Preview of the *RecLAK* recommendation interface.

potentially useful to reduce the cost of dataset interlinking. For more information, including the full set of data used in the experiments, graphical visualizations and detailed results, we refer to the *RecLAK* Web application, avaliable at `http://www.inf.puc-rio.br/~grlopes/RecLAK`.

6. ACKNOWLEDGMENTS

This work was partly funded by CNPq, under grants 160326/2012-5, 301497/2006-0 and 57128/2009-9, and by FAPERJ, under grants E-26/170028/2008 and E-26/103.070/2011.

7. REFERENCES

[1] T. Berners-Lee. Linked Data. In *Design Issues*. W3C, July 2006.

[2] K. A. Kuznetsov. Scientific data integration system in the linked open data space. *Programming and Computer Software*, 39(1):43–48, Jan. 2013.

[3] L. A. P. P. Leme, G. R. Lopes, B. P. Nunes, M. A. Casanova, and S. Dietze. Identifying candidate datasets for data interlinking. In *ICWE'13*, pages 354–366, 2013.

[4] G. R. Lopes, L. A. P. P. Leme, B. P. Nunes, M. A. Casanova, and S. Dietze. Recommending tripleset interlinking through a social network approach. In *WISE'13*, pages 149–161, 2013.

[5] G. R. Lopes, L. A. P. Paes, B. P. Nunes, M. A. Casanova, and S. Dietze. Comparing recommendation approaches for dataset interlinking. Technical report, Department of Informatics, PUC-Rio, 2013.

[6] B. F. Lóscio, M. Batista, and D. Souza. Using information quality for the identification of relevant web data sources. In *IIWAS'12*, pages 36–44, New York, NY, USA, 2012. ACM.

[7] L. Lü, C.-H. Jin, and T. Zhou. Similarity index based on local paths for link prediction of complex networks. *Physical Review E*, 80(4):046122, 2009.

[8] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 2002.

[9] A. Nikolov and M. d'Aquin. Identifying Relevant Sources for Data Linking using a Semantic Web Index. In *LDOW'11*, 2011.

[10] A. Nikolov, M. d'Aquin, and E. Motta. What Should I Link to? Identifying Relevant Sources and Classes for Data Linking. In *JIST'12*, pages 284–299. Springer Berlin Heidelberg, 2012.

[11] B. P. Nunes, B. Fetahu, and M. A. Casanova. Cite4me: Semantic retrieval and analysis of scientific publications. In *LAK (Data Challenge)*, volume 974 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013.

[12] B. P. Nunes, B. Fetahu, S. Dietze, and M. A. Casanova. Cite4me: A semantic search and retrieval web application for scientific publications. In *ISWC (Posters & Demos)*, volume 1035 of *CEUR Workshop Proceedings*, pages 25–28. CEUR-WS.org, 2013.

[13] H. R. d. Oliveira, A. T. Tavares, and B. F. Lóscio. Feedback-based data set recommendation for building linked data applications. In *I-SEMANTICS'12*, pages 49–55, 2012.

[14] D. Taibi and S. Dietze. Fostering analytics on learning analytics research: the lak dataset. In *LAK (Data Challenge)*, volume 974 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013.

[15] A. Wagner, P. Haase, A. Rettinger, and H. Lamm. Discovering related data sources in data-portals. In *SemStats workshop, ISWC'13*, 2013.