A Linked-Data-Driven Web Portal for Learning Analytics: Data Enrichment, Interactive Visualization, and Knowledge Discovery

Yingjie Hu
STKO Lab
University of California, Santa Barbara
yingjiehu@geog.ucsb.edu

Grant McKenzie
STKO Lab
University of California, Santa Barbara
grant.mckenzie@geog.ucsb.edu

Jiue-An Yang
STKO Lab
San Diego State University / University of California, Santa Barbara
jiueanyang@geog.ucsb.edu

Song Gao
STKO Lab
University of California, Santa Barbara
sgao@geog.ucsb.edu

Amin Abdalla
STKO Lab
Vienna University of Technology
abdalla@geoinfo.tuwien.ac.at

Krzysztof Janowicz
STKO Lab
University of California, Santa Barbara
jano@geog.ucsb.edu

ABSTRACT

This paper presents a Linked-Data-driven Web portal for the field of learning analytics. The portal allows users to browse the linked datasets and explore data about researchers, conferences, and publications. Additionally, users can interact with various dynamic visualization applications and perform analysis, e.g., study temporal change of research trends. Based on the provided datasets on Learning Analytics and Knowledge (LAK) and Educational Data Mining (EDM), we enriched the data with geospatial locations of research institutes, topics extracted from papers, and the expertise of researchers. The interactive modules of the Web portal are then designed and implemented using the enriched RDF data. The implemented modules can be divided into two groups. The first group is concerned with providing dynamic and interactive visualization of the data, such as the modules of *Conference Participants* and *Reference Map*. The modules in the second group are designed for more advanced analysis and discovery of new knowledge, such as the modules of *Scholar Similarity* and *Reviewer Recommendation*. The modules have been designed following a loosely coupled, modular infrastructure, and can be easily migrated and reused in other projects.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Scientific Databases—*bibliography database*; H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Management, Design, Measurement

Keywords

Linked scientometrics, semantics, topic modeling, interactive visualization, Linked Data, learning analytics, data mining

1. INTRODUCTION AND MOTIVATION

The Learning Analytics and Knowledge (LAK) dataset provides a rich collection of data extracted from publications in the field of learning analytics and structured in a machine-readable manner [7]. It provides a great opportunity for scientometrics research. For instance, one can investigate the topic trends in this emerging research field, study the network of collaborative researchers, examine the topical similarity among researchers, explore the spatio-temporal spread of LAK-related ideas, and check the relation between conference locations and affiliations of the contributing researchers. In this paper, we present a Linked-Data-driven Web portal for scientometrics based on the LAK dataset. It is designed to facilitate the exploration, enrichment, visualization, and analysis of the data. While part of the portal is based on a previous project which implements a semantically-enabled and Linked-Data-driven platform [5], we have designed and developed 11 new scientometric modules for the LAK challenge. We are especially interested in the novel area of spatio-temporal scientometrics [2, 3]. While this includes geographic space, we are also interested in more general spaces. All modules are based on a combination of various technologies, such as Linked Data, Semantic Web reasoning, geocoding, D3 (Data-Driven Documents) library, GeoJSON, and a variety of KDD techniques such as Latent Dirichlet Allocation-based topic modeling, Multi-Dimensional Scaling, and so forth. Specifically, the Web portal provides the following features:

1. Enriched data. The existing LAK dataset has been enriched with extracted research topics, key concepts,

institute locations, citations, and researcher expertise. We also integrated the data from two prominent scientometric services, Microsoft Academic Search[1] and ArnetMiner[2], to provide more comprehensive analysis.

2. An intuitive user interface. We provide four accesses for the data: *conferences*, *researchers*, *publications* and *analytics*. Users can start browsing the data from any of these access points using either keyword-based search or *follow your nose* exploration. From a particular data item (e.g, a conference), users can also check its related items (e.g., researchers and publications in this conference) by following the hyperlinks. Users can also use the analytic modules to perform analysis based on the entire LAK dataset instead of exploring subsets sequentially.

3. A set of dynamic and interactive (geospatial) visualizations and animations. Modules, such as the *Coauthor Treemap* and the *Academic Network*, enable users to interact with the visualization and explore details. This portal especially features a group of geospatial visualizations, such as the *Reference Map*, *Citation Map*, and *Participants Map*, which present the geospatial distribution of researchers, institutions, citations, and references.

4. Data mining and knowledge discovery. The *Active Scholar* module, for instance, can help identify the most active researchers in a conference, while the *Scholar Similarity* module can detect the similarity among scholars based on their publications. The *Topic Trending* module explores the popular topics for the EDM conference from 2008 to 2013[3], while the *Key Concept* module enables the user to grasp the gist of a paper.

5. Linked-Data-driven analytic tools for the LAK community. Based on the LAK dataset, two useful tools have been designed. The *Reviewer Recommendation* tool recommends reviewers for a paper (related to the topics presented in LAK) based on the existing publications of researchers in this community while at the same time excluding co-authors based on data provided from the network module. The *Potential Collaborator* tool can help researchers find potential collaborators who have very similar research interests but may have not collaborated before.

6. A modular and self-contained design paradigm. All functionalities in the presented portal have been designed as self-contained modules. Therefore, such functionalities can be easily migrated and reused in future projects. Queries are based on RDF, SPARQL, and ontologies, which make storing part of the portal's *business logic* directly with the data possible.

[1] http://academic.research.microsoft.com/
[2] http://arnetminer.org/
[3] This is the only conference in the dataset that offers an extensive history worth analyzing. In the future, more conferences can be added.

The presented Web portal (called DEKDIV[4]) can be accessed at: http://stko-exp.geog.ucsb.edu/lak/.

In the following sections, we briefly discuss the data enrichment, visualization, and knowledge discovery from the LAK data.

2. DATASET DESCRIPTION AND ENRICHMENT

The LAK dataset contains information about publications and researchers in the Learning Analytics and Knowledge (LAK) conference series (from 2011 to 2013) and the Educational Data Mining (EDM) conference series (from 2008 to 2013). This dataset also contains a special issue from the *Journal of Educational Technology and Society* on learning analytics and knowledge. In contrast to other publication data hubs (e.g., DBLP[5] and CiteSeer[6]) on the Linked Open Data (LOD) cloud, the LAK dataset provides full texts and full references in addition to the typical bibliographic data. As a result, it is not only possible to explore the collaboration relations (e.g., via co-authorships), but also to mine key concepts, research topics, and citation networks. This facilitate a more holistic understanding of the learning analytics and educational data mining research fields.

In this work, we further enrich the provided LAK dataset with key concepts and topics extracted from papers, geographic locations of authors' affiliations, as well as the expertise of researchers and paper citations imported from Arnetminer and Microsoft Academic Search (MAS). While these additional data have been imported into our own triple store[7], they can also be merged with the existing LAK dataset. In addition, a customized daemon can be developed to dynamically enrich the data whenever new publications (e.g., papers from the LAK 2014 conference) are added. In fact, this is the strategy that has already been adopted in our previous Linked Data portal for the *Semantic Web Journal* to synchronize the information about new publications [5]. In the following subsections, we discuss the three types of data that have been added and linked to the existing LAK dataset.

2.1 Key Concepts and Research Topics

Based on the full text of each publication in the LAK dataset, we extracted important concepts using the Alchemy API[8]. Alchemy is a Web service that provides automatic natural language processing functions. Each of the extracted concepts (key phrases) is associated with a relevance value which indicates the term's importance in relation to the entire paper. The extracted concepts were used in the *Key Concepts* module to give users an overview of a paper's content. These concepts were also used to display the trend of research topics in the *Topic Trending* module. Additionally, we use a *Latent Dirichlet allocation (LDA)* model to extract

[4] Short for **D**ata **E**nrichment, **K**nowledge **D**iscovery and **I**nteractive **V**isualization.
[5] http://datahub.io/dataset/fu-berlin-dblp
[6] http://thedatahub.org/dataset/rkb-explorer-citeseer
[7] http://stko-exp.geog.ucsb.edu/pubby
[8] http://www.alchemyapi.com/

latent topics from the full text data. LDA is an unsupervised, generative probabilistic model used to infer the latent topics in a textual corpus [1]. In contrast to the key concepts extracted using Alchemy, LDA describes each topic using a mixture of keywords, and therefore can help discover hidden relations among key concepts. The topics presented by LDA were used for the functions in *Scholar Similarity* and *Reviewer Recommendation*.

2.2 Geospatial Reference Data

Scientific activities (e.g., co-publications, citations, and references) generally show geospatial patterns. However, such patterns have rarely been considered in traditional scientometric analysis which often focuses on numeric values (e.g., the H-index) [4]. Recently, researchers began to pay more attentions to spatial scientometrics [2] and a spatio-temporal framework for exploring the citation impact of publications and scientists has been proposed in a previous work [3]. Based on this motivation, we first identified the affiliation of each author in the LAK dataset and geolocate these institutes using a customized parser on top of the Google Maps Reverse Geocoding API (the algorithm can be found in [3]). We also geolocated the affiliation of the first author in each reference entry in order to reveal geospatial patterns in the reference data. These geospatial data are the foundation for the *Citation Map*, *Reference Map*, *Collaborative Institutes*, and *Conference Participants* modules.

2.3 Researcher Expertise and External Citation Data

In addition to the key phrases, topics, and geospatial locations, we also integrated the scientometric data from two major Web services: Arnetminer and Microsoft Academic Search. Among the 853 authors in the LAK dataset, Arnetminer contains data for 595 of them, and such data includes publications by these authors not only from the LAK dataset, but also from other journals and conferences. We imported the researcher expertise data (which were extracted based on all publications of a given author) from Arnetminer into our local triple store, and used these expertise data for the *Potential Collaborators* module. Microsoft Academic Search is another source whose data have been integrated into the presented Web portal. MAS contains a large amount of information about affiliations and citations. Such data have been used to find the affiliation of authors in the reference data. The citation data from MAS have also been integrated to show the external citations which are not stored in the existing LAK dataset.

3. INTERACTIVE VISUALIZATIONS

This section presents the functionality modules that serve the purpose of interactive visualization and animation. Visualization plays an important role in data analysis, since humans can often easily recognize patterns visually (in contrast to machines). When designing this Web portal, we paid special attentions to each aspect of the visualizations created. For instance, we tried to ensure that each visualization fits the data being presented, and that the user interface enables users to interact and understand the data. The subsections below describe some of these modules, and the role that visualization plays in their conceptualization.

Collaborative Institutes & Conference Participants

The *Collaborative Institutes* module is designed to help users explore the spatial distribution of co-authorship. In order to realize this, the number of co-authored conference papers between two institutes are displayed on a global map. As shown in Figure 1, a link between two institutes indicates research collaborations between authors from these institutes. Interactions provided by this module include moving the mouse over links and nodes to show additional details about each of the institutes and publication information. Using this geo-visualization, one can see spatial patterns of research collaborations. For example, some researchers prefer to collaborate domestically for certain conferences, as opposed to international collaborations. Previous studies [2] argued that the tendency of domestic co-publication may indicate that a large number of researchers in this domain reside within a single country. This is evident in the institute-collaboration pattern seen in the EDM 2013 dataset, which shows primarily US-based collaborations. Conversely, conferences, such as LAK 2013, consist of more internationally oriented authors.

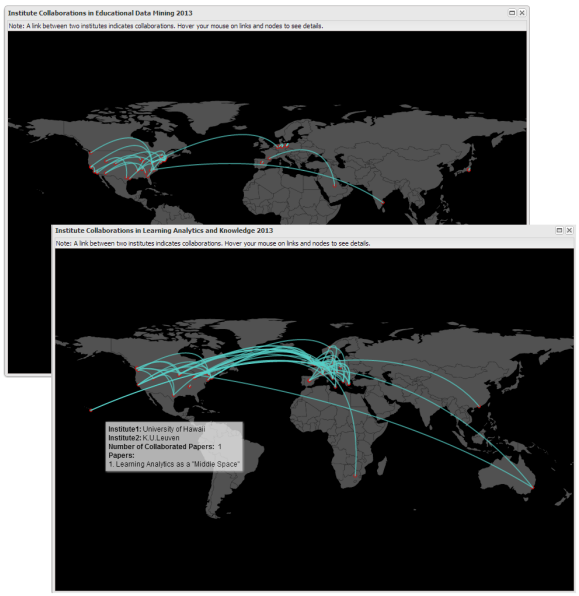

Figure 1: Geo-locations of institutes of collaborating authors.

The *Conference Participants* module was developed to provide an interactive visualization and animation of the geographic distributions of EDM & LAK conference authors; see Figure 2. The size of the yellow circles represents the number of participants that attended the conference; the larger the circle, the more attendees from the mapped institute. Hovering the mouse over a circle on the map displays the total number of authors who participated in the conference. From a purely visual perspective, it appears that the geographic distribution of participants is strongly influenced by the region where the event took place. For instance, LAK 2013 was held in Leuven, Belgium, and it attracted many European researchers, while many participants of EDM 2013, which was held in Memphis, USA, are from the United States.

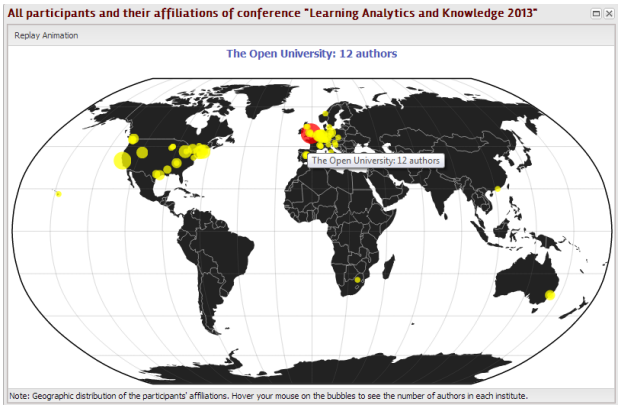

Figure 2: The spatial distribution of LAK 2013 authors grouped by their institute's location.

Academic Network

The *Academic Network* module is constructed to show an interactive visualization of an author's academic network based on co-author links. A graph-node approach (often found in social network analysis) has been implemented to connect authors with each other through their LAK/EDM publications. This technique views the academic network as a set of relationships composed of nodes and links. In this module, each author is presented as a node in the network, and a link connects two nodes if co-authorship exists between these two authors. The total number of co-authorships between two authors is recorded as an attribute of the link and is visually represented though the stroke width of each link (Figure 3). In addition to the visualization of the academic network, measurements of the network are also provided. For example, the *Centrality* of a node represents its relative importance within a network, and four types of centrality are visually presented in the *Academic Network* module when a node is selected.

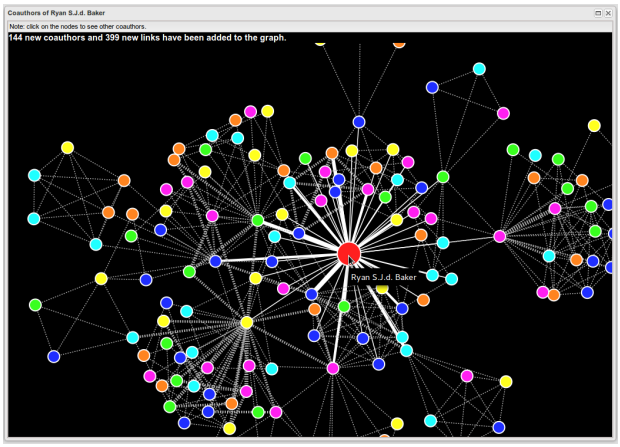

Figure 3: The Academic Network of a single researcher.

Reference Map

Based on the enriched and geocoded data, we have designed a module called *Reference Map*. This module presents an animated and interactive visualization demonstrating the geographic distribution of a paper's references. Using the reference data provided by the LAK dataset, we extracted the names of the first authors in the reference records, and use MAS to find the institutions of these authors. We then visualized these institutions as smaller bubbles on a global map, and created animated links from these bubbles to the location of the published paper, where a larger bubble is created (see Figure 4). The message that we are trying to deliver through this visualization is: scientific publications (i.e., the smaller bubbles) emerged at different locations of the world, and new papers (i.e., the bigger red bubble) was created and published when other researcher adopted these ideas (i.e., the animated linked). In this module, users can see additional information about authors, institutions, and each reference when hovering over the bubbles and links.

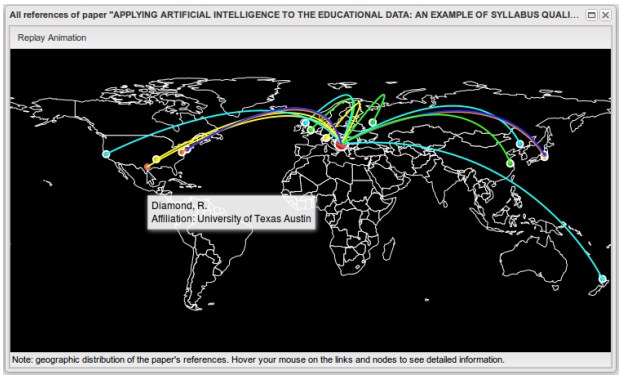

Figure 4: Geo-visualization of a paper's references.

4. LAK DATA MINING AND KNOWLEDGE DISCOVERY

The richness of the LAK/EDM dataset allows the development of a wide assortment of knowledge discovery tools. In this section, we present some of the modules which are developed based on data mining techniques. These modules utilize the enriched dataset, and offer informative metrics to facilitate the understanding of not only the dataset itself but also the LAK community. In addition, some of these modules provide useful services which, until now, may have not been fully realized.

Active Scholars

The module of *Active Scholar* is designed to discover the most productive researchers based on the provided dataset. We developed a SPARQL query which ranks authors in each conference according to the number of their publications. The top 30 authors were then selected, and displayed in the center of the visualization canvas (see Figure 5). The number after the name of each author denotes the number of papers that this specific author published in this conference. While authors who have two or more publications are displayed at the top, we have to randomly select authors who have a single publication in the conference due to the limited visualization space. In addition to the most active scholars, we also identified and visualized the most popular topics based on the number of publications in this topic. By hovering mouse on one of the authors, users can see the topics that are related to this author. Similarly, moving mouse on one topic will show all the authors who have publications

on this topic. Based on this module, we have discovered some very active scholars (e.g., Ryan S.J.d Baker) who often publish more than 2 papers in the LAK conferences.

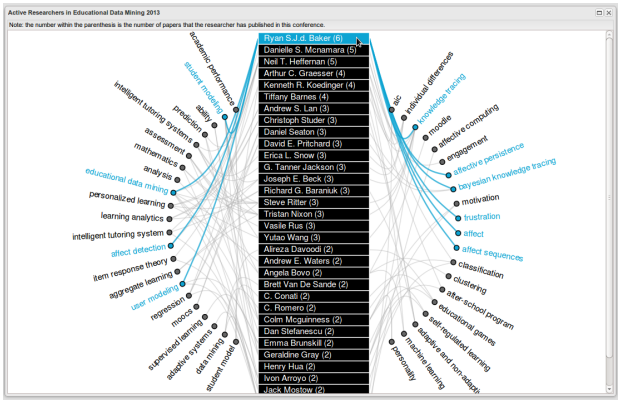

Figure 5: The Active Scholar Module.

Scholar Similarity

The *Scholar Similarity* module measures the similarity among authors in the LAK dataset using a Multidimensional Scaling (MDS) approach. The full text publications of each author have been selected as input for a *Latent Dirichlet allocation (LDA)* model consisting of 20 topics. LDA is an unsupervised, generative probabilistic model used to infer the latent topics in a textual corpus [1]. Given these topics, each author can then be described as a distribution across these topics which we term as *signatures*. Using the *Jensen-Shannon divergence* [6] method, each author in the dataset is compared to each other author producing a dissimilarity measure bounded between 0 and 1. This resulting matrix of dissimilarity values is used as input to MDS which produces a 2D representation of the authors in space. Authors that are closer together in space are more similar in their research topics compared with those further apart.

Reviewer Recommendation

The *Reviewer Recommendation* module (Figure 6) is built on the concept that researchers who have already published on a topic have the potential to become good reviewers for new papers on the same topic. Similar to the *Scholar Similarity* module, this tool uses LDA to generate a set of topics from a corpus of full text publications. The authors of these publications are then defined as a distribution across the topic space. Given a new submission (in the form of an abstract), the topic distribution of this material can be inferred from the existing LDA topics. Again, using the *Jensen-Shannon divergence* method, the topic distribution of the new material was compared with the topic distributions of all authors, and authors were then ranked based on the similarity value (defined as 1 – *JSD* dissimilarity value). To avoid interest conflicts, we compared the list of potential reviewers with the researcher's previous co-authors. Those co-authors were highlighted in the visualization to reflect the conflicts.

Research Topic Trends

This module demonstrates how research topics in the LAK or the EDM conferences trend over time (see Figure 7). In

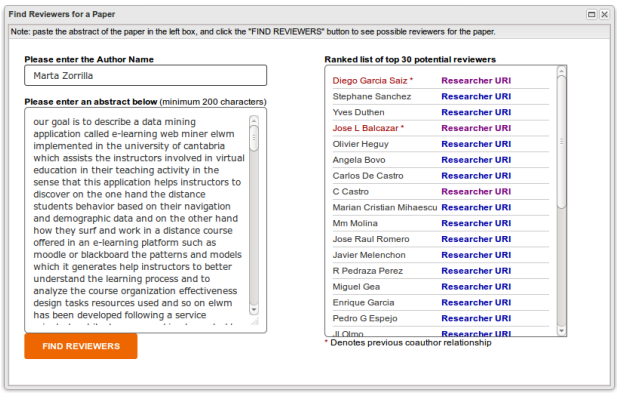

Figure 6: The Reviewer Recommendation module.

order to implement this module, data preprocessing has been done to extract the top 10 topics for each year ranked by total number of papers which contain the topic keywords. The EDM conferences, for instance, has 46 distinct topics extracted from 2008 to 2013. Frequencies for all topics were calculated across all years in a time-series format for further visualization and analysis. Using the interactive stream graph, it is clear to see the decline of certain topics, emergence of new topics, as well as trend expansions through time. The user interface enables interaction through mouse hover, reporting the chosen topic along with the number of papers associated with the topic in each year.

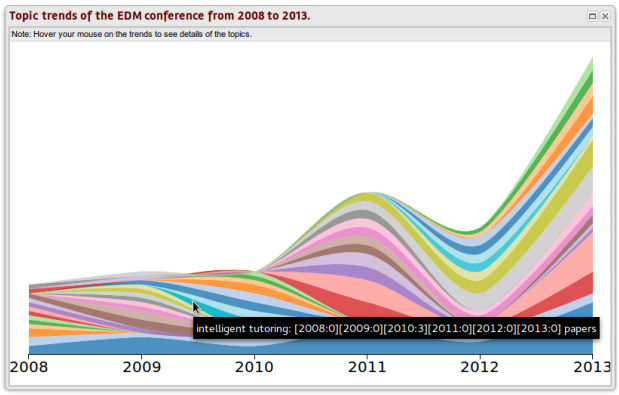

Figure 7: The stream graph visualization of the EDM hot topic trends.

Potential Collaborators

This analytical function is designed to find out the researchers who have similar research interests but may have never co-authored a paper before (i.e., the researchers that can potentially become collaborators). The similarity of researchers was calculated using cosine similarity measure based on expertise of the research imported from Arnetminer. We then calculated the shortest network distance based on the co-authorships in the module of *Academic Network*. After that, a metric (see equation (1)) was designed to find authors who have the potential to become collaborators.

$$p = sim(a_1, a_2) \times (1 - 1/d) \quad (1)$$

Where p is the *score* for collaboration potential, $sim(a_1, a_2)$ is the cosine similarity between authors a_1 and a_2, and d is the shortest network distance. Figure 8 shows a screenshot of the *Potential Collaborator* module, where the blue icon represents the current researcher and the surrounding grey icons display his/her potential collaborators. Researchers who have already co-authored papers before will have a link, while no link indicates there exist no co-authored paper in the LAK dataset.

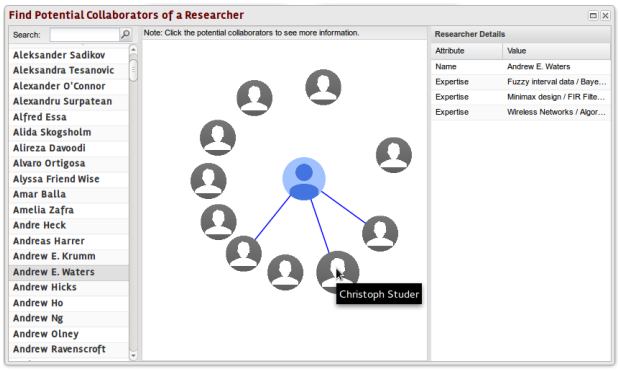

Figure 8: The potential collaborator module.

5. OTHER FUNCTIONAL MODULES

While we only highlight some selected modules here, there are a list of other DEKDIV modules. Interested readers are encourages to explore these modules. Examples include:

- *Conference Hot Topics*: The most popular topics in this conference.
- *Coauthor Treemap*: A treemap visualization of co-authors and their affiliations.
- *Citation Map*: The geospatial distribution of the citations of authors and their publications.
- *Key Concepts*: Important key phrases of a given paper and their relevance.

6. CONCLUSIONS

In this paper, we presented a Linked-Data-driven scientometrics Web portal, called DEKDIV, for the LAK challenge. We enriched the initial LAK data with paper topics, geospatial locations, and author expertise. We published the data in our local triple store, and designed various interactive visualization and analysis modules to facilitate the understanding of research in the LAK community. We also developed a set of more complex service modules on top of the enriched data. They can be used to recommend reviewers for newly submitted papers, help researchers find potential collaborators, detect trend changes in publication topics, and so forth. The developed Web portal is based on our previous work for the *Semantic Web journal*, and the entire system is highly modular, reusable, as well as free and open source software.

7. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[2] K. Frenken, S. Hardeman, and J. Hoekman. Spatial scientometrics: Towards a cumulative research program. *Journal of Informetrics*, 3(3):222–232, 2009.

[3] S. Gao, Y. Hu, K. Janowicz, and G. McKenzie. A spatiotemporal scientometrics framework for exploring the citation impact of publications and scientists. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 204–213. ACM, 2013.

[4] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America*, 102(46):16569, 2005.

[5] Y. Hu, K. Janowicz, G. McKenzie, K. Sengupta, and P. Hitzler. A linked-data-driven and semantically-enabled journal portal for scientometrics. In *The Semantic Web–ISWC 2013*, pages 114–129. Springer, 2013.

[6] J. Lin. Divergence measures based on the shannon entropy. *Information Theory, IEEE Transactions on*, 37(1):145–151, 1991.

[7] D. Taibi and S. Dietze. Fostering analytics on learning analytics research: the lak dataset. 2013.