Paperista: Visual Exploration of Semantically Annotated Research Papers

Nikola Milikic
Faculty of Organizational Sciences, University of Belgrade
Jove Ilića 154
Belgrade 11000, Serbia
+381-11-3950853
nikola.milikic@gmail.com

Uros Krcadinac
Faculty of Organizational Sciences, University of Belgrade
Jove Ilića 154
Belgrade 11000, Serbia
+381-11-3950853
uros@krcadinac.com

Jelena Jovanovic
Faculty of Organizational Sciences, University of Belgrade
Jove Ilića 154
Belgrade 11000, Serbia
+381-11-3950853
jeljov@gmail.com

Bojan Brankov
UZROK Labs
107 Nehruova
Belgrade 10070, Serbia
+381-63-581879
bb@uzrok.com

Srdjan Keca
UZROK Labs
107 Nehruova
Belgrade 10070, Serbia
+381-61-3115661
sk@uzrok.com

ABSTRACT
We consider the problem of visualizing and exploring a dataset about research publications from the fields of Learning Analytics (LA) and Educational Data Mining (EDM). Our approach is based on semantic annotation that associates publications from the dataset with Wikipedia topics. We present a visualization and exploration tool, called Paperista (www.uzrok.com/paperista), which presents these topics in the form of bubble and line charts. The tool provides multiple views, thus allowing users to observe and interact with topics, understand their evolution and relationships over time, and compare data originating from different research fields (i.e., LA and EDM). Moreover, user can explore papers to which the presented topics are related to, and make related Web searches to access the papers themselves.

Categories and Subject Descriptors
D.2.2 [**Software Engineering**]: Design Tools and Techniques - *user interfaces*

General Terms
Algorithms, Design

Keywords
Learning Analytics, Visualization, Research Papers

1. MOTIVATION
The field of Learning Analytics is emerging in the past few years and attracting more and more researchers from other areas of Technology Enhanced Learning (TEL). It aims to address the current needs in the broad area of education by making use of the latest trends in information technologies where everything is moving towards Big Data and real-time analytics.

Learning Analytics (LA) is defined as "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs" [16]. It is often equated with other similar fields in the TEL area, such as Academic Analytics or Educational Data Mining (EDM) [14]. EDM is a research field that focuses on using computational approaches, namely data mining and machine learning, to analyze educational data in order to facilitate and enhance educational process, and contribute to the overall improvement of students' learning experience [17]. Even though both LA and EDM are self-contained research fields, they are intertwined and overlap in topics they cover. They share many similarities, but also have some distinct differences as discussed by Siemens and Baker [18]. One of the similarities emphasized by these authors is that both fields reflect the emergence of data-intensive approaches to education, where both communities have the goal of analyzing large-scale educational data in order to support research and practice in education. They differ in the level of automation they aim to achieve. In particular, EDM has a greater focus on automating support for educational processes, such as adaptation and personalization of learning environments and learning processes. On the other hand, LA has a considerably greater focus on leveraging human judgment, on informing and empowering instructors and learners to reflect over and improve learning processes.

The Society for Learning Analytics Research (SoLAR) has published LAK dataset[1] containing structured data about research publications from Learning Analytics and Knowledge (LAK) Conference, Educational Data Mining Conference, and Journal of Educational Technology & Society (JETS) Special Issue on LAK. The data are represented in the RDF form, which makes them easy to integrate and process by applications.

In this paper, we propose an approach to visualizing and exploring the LAK dataset. It is centered around the topics covered by the papers from the dataset, and is intended to give an overall view of the topics that LA and the EDM fields cover. As the focus of researchers and the degree of relevance of particular topics have been changing over years, our approach tries to show a trend of those changes through the whole period the dataset covers, namely from 2008 to 2012. It also allows for topic-based exploration of research papers and easy navigation to them.

2. RELATED WORK
In [11], authors present an interesting work aimed at automating the creation of relations between research areas by using semantically annotated data about research papers in a particular

[1] www.solaresearch.org/resources/lak-dataset

area. As a continuation of this work, the same authors have created a tool, called Rexplore[2], which, among other things, visualizes authors migration patterns across research areas [15].

In terms of visual representation, we find interesting an approach to visualization of tags (topics) and categories of tags over time. For example, Dubinko et al. [4] consider the problem of visualizing the evolution of Flickr tags. The authors present a new slider-based approach based on a characterization of the most interesting tags. A Flash-based animation in a web browser allows the user to observe and interact with the tags. Zhang et al. [5] present an approach to classification and visualization of temporal and geographic tag distributions. The authors argue that their approach can help humans recognize semantic relationships between tags. Lemma [6] presents the Ebony system, an application for browsing, navigation, and visualization of the DBLP database. Wattenberg [7] introduces arc diagrams for representing complex patterns of repetition in string data. Watteberg application, the Shape of Song, visualizes music files, creating a static representation of repetition throughout a time series. However, to our knowledge, there has been no (published) research work on the visualization of research topics and publications in the areas of LA and EDM.

3. THE PAPERISTA SYSTEM

Our approach is illustrated through a Web application called Paperista. The application visualizes topics associated with research publications from the LAK dataset, allowing users to browse through papers, compare LA and EDM research fields, and make related Web searches. Visualizations are created for each individual year in order to display relevant topics in the LA and EDM fields for a specific year, but also for all years combined in order to give an overall depiction of the topic distribution in these research areas.

3.1 Data Preparation and Analysis

LAK dataset consists of data about conferences and journal papers published in the LA and EDM research fields in the 2008-2012 period. For each paper, the following elements are available: title, author(s), abstract, keyword(s) and full text. Also, basic information about authors is available, such as name and affiliation.

3.1.1 Topic Extraction

Since one of the main features of Paperista is visualization of research topics relevant for the given corpus, the first step in the data preparation process was to extract main topics of the papers encompassed by the LAK Dataset. A straightforward approach was to use keywords associated with the papers. This is because the authors themselves have compiled those keywords, and it is them who know the best which topics describe their work in the most appropriate way. However, the downside of this approach is that those keywords are given as free form text and are not consistent with any existing formal vocabulary. This makes them inconsistent throughout the corpus. Furthermore, the dataset is incomplete in regard to keywords as for conferences EDM 2008, 2009 and 2010 no keywords are provided.

Thus, we decided to employ a service for semantic annotation in order to detect paper topics. We took into consideration two Wikipedia based semantic annotators: TagMe[3] and DBpedia Spotlight[4]. The decision to use Wikipedia based annotator was motivated by the fact that Wikipedia is the largest corpus of open encyclopedic knowledge and is often used as a well established large-scale taxonomy [8]. Both annotator services are designed to look for and retrieve recognized Wikipedia concepts from the given text. They can be configured to the specific needs of any particular usage scenario (i.e., corpus). TagMe is designed to identify Wikipedia concepts specifically in short texts. Its REST API[5] allows for configuration of two parameters: i) the *rho* parameter which refers to the "goodness" of an annotation with respect to the topics of the input text, and ii) the *epsilon* parameter which is used for fine-tuning the disambiguation process and indicates whether to favor the most-common topics or to take the context more into account [9]. DBpedia Spotlight annotates a given text with concepts from DBpedia, a structured representation of Wikipedia [12]. DBpedia Spotlight REST API[6] exposes two parameters: *confidence* of the annotation process that takes into account factors such as the topical pertinence and the contextual ambiguity; *support* parameter specifies the minimum number of inlinks[7] [10]. We used only paper title and abstract for topic extraction, based on an assumption that these two elements contain mentions of the most important and interesting topics a paper is related to. In order to decide which service for semantic annotation to use, the two services were tested with a random sample comprising 5% of all papers and with different parameter settings. The best results were achieved by the TagMe service (rho=0.15; epsilon=0.5). For this reason, TagMe service was employed to annotate all papers in the corpus.

3.1.2 Identifying Popular Topics

Once having all papers associated with topics, we calculated the significance of each topic. Numerical statistic called TF-IDF (Term Frequency – Inverse Document Frequency)[8] was used as it calculates how important a word is to a document in a corpus of documents. This metric was adapted to our case and used to calculate the importance of a topic in a paper. Instead of calculating the frequency of a word, we calculate the frequency of a topic.

Since Paperista allows for visualizing topics in a specific year and overall (in all years, 2008-2012), the significance was calculated for corpora containing papers from each of these different time periods. Accordingly, we had six different corpora and calculated the significance of a topic for each corpus. In order to present only the most significant topics, we have filtered the topic set to only those whose significance for a particular period was over 0.01. This threshold was empirically chosen and presents the best balance between the relevance of topics and their presentation in the Paperista's visualizations (i.e., assuring easy comprehension by users).

3.1.3 Topic Cleaning

Even though the output of TagMe service consisted of topics that are relevant to the papers' content, some of them can hardly be considered as relevant research topics in the LA and EDM fields as they are too general. For instance, topics like *Methodology*,

[2] http://technologies.kmi.open.ac.uk/rexplore
[3] http://tagme.di.unipi.it
[4] http://spotlight.dbpedia.org
[5] http://tagme.di.unipi.it/tagme_help.html
[6] http://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki/Web-service
[7] Inlinik, or inline link, are incoming links from other DBpedia concepts to the observed DBpedia concept
[8] http://en.wikipedia.org/wiki/Tf-idf

Research, and *Experiment* can be associated with almost every paper in this corpus. Actually, these topics can be related to research papers from almost any other research area. Similarly, some of the retrieved topics were not relevant to research papers from the LAK dataset. Such topics resulted from imperfection of the TagMe tool (and semantic annotation tools, in general). Some examples of these alien topics include *The T.O. Show*, *Ade Easily*, *Henry Snapp*, etc. For instance, *Henry Snapp* topic was apparently mistaken with the SNAPP tool[9], a popular learning analytics and visualization tool. Hence, it was important to detect and exclude all these generic and alien topics from the final visualization in order to reduce the noise.

We applied topic cleaning approach similar to [11]. The idea is to identify topics that have little or no relationships with other topics in the corpus. This can be an indicator that a topic is too specific or alien to our set of identified topics and thus can be considered as an exclusion candidate. On the other hand, if a topic has relationships with too many other topics, this can be an indicator that a topic is too generic and again should be considered as an exclusion candidate. In order to detect these outlier topics, we needed a measure of relatedness between topics. To that end, we used the Wikipedia Miner[10] service that calculates semantic relatedness of two topics by finding the corresponding Wikipedia articles, and calculating similarity of those articles by comparing their incoming and outgoing links [13]. Wikipedia Miner has a REST API[11] that allows for retrieving this information programmatically.

Once having relatedness calculated for all the topics in our corpus, we compiled two lists to help us detect removal candidates. In the first list, each topic was associated with the number of other topics that topic is related to. This gave us an insight into which topics can be considered too general/specific (the higher the number of related topics, the more generic the topic is, and vice versa). In the second list, each topic was associated with a sum of its relatedness with all the other topics. This list was meant to complement the first one. The rationale here is that there might be a topic with fair number of relations to other topics, but those relatedness values are weak. This behavior also qualifies a topic to be considered as too specific or alien.

The initial idea with compiling these two lists was that topics to be removed will be at the beginning and the end of the lists (top and bottom 10%), and that they could be removed automatically. However, by examining the lists, among the obvious exclusion candidates, there were also several topics that should not have been excluded. For instance, topics like *Online tutoring*, *Process mining*, *Educational data mining* etc. were at the end of both lists making them removal candidates, even though these topics are obviously highly relevant for LA and EDM fields. The reason for this lays in the nature of Wikipedia itself and the fact that not many other articles in Wikipedia link to these topics. Thus, the topic removal process could not be done completely automatically and an expert in the area was consulted to mark the topics that should not be excluded.

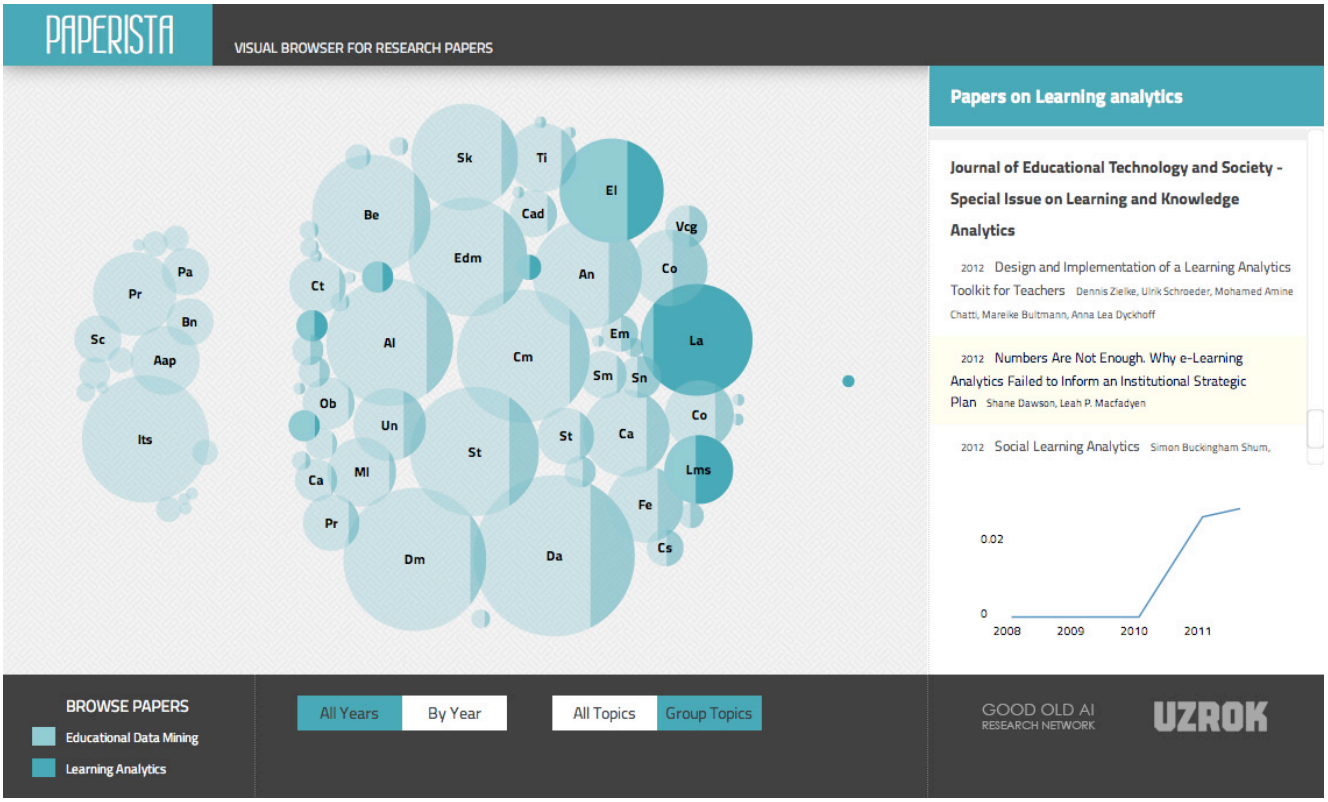

Figure 1 - Paperista Interface

[9] http://www.snappvis.org
[10] http://wikipedia-miner.cms.waikato.ac.nz
[11] http://wikipedia-miner.cms.waikato.ac.nz/services

3.2 Data Visualization and Exploration

The topic visualization applied in Paperista is inspired by the New York Times visualizations *Four Ways to Slice Obama's 2013 Budget Proposal* [1] and *At the National Conventions, the Words They Used* [2].

The Paperista visualization includes bubble and line charts, allowing users to gain insights into topic trends within the LA and EDM fields. Bubble charts show the importance of a certain topic for the entire dataset, each year, and/or each field. By changing different views, users can watch the changes within the dataset and compare the two fields. Animated transitions between charts help users understand these processes. In addition, since the animation does not show precise changes in topic's relevancy (calculated using TF-IDF metric, see Sect. 3.1), users are also presented with a relevancy line chart for each topic.

The user interface (Figure 1) consists of an animated bubble cloud, two button sliders, a sidebar, and an optional timeline. The first slider button (All Years / By Year) allows users to choose between the "All Years" and "By Year" views. "All Years" view presents relevant topics for the entire corpus of publications. "By Year" view activates a timeline, showing relevant topics for each year. By using the slider, users can follow the change in topic relevancy through the years the data is available for (2008-2012).

The second slider button (All Topics / Group Topics) allows for grouping and regrouping of topics. "All Topics" view shows one circle-shaped bubble chart. "Group Topics" view divides the chart into two groups of bubbles. The first group presents topics that appear only in the EDM field. The third one shows topics related only to the LA field (i.e., LAK and JETS publications). The group in the middle shows "mixed" topics, i.e., those that appear at least once in both EDM and LAK/JETS. Different views of the bubble chart are presented on Figure 2.

The size of a bubble represents the topic's relevancy (i.e., TF–IDF value). Two research fields, EDM and LA, are color-coded. Each bubble is divided into two slices the size of which corresponds to the frequency of that topic within publications of each of the two sources. For the years 2008-2010, the dataset contains data only for the EDM research field, so the bubbles are one-colored.

The order of topic bubbles is intended to help users compare the two fields. The leftmost bubbles represent mostly EDM-related topics, while the rightmost bubbles mostly belong to the LA field. Moreover, clicking on a bubble creates a line chart in a sidebar. The line chart shows the growth and decline of a certain topic.

In addition to the visualization, the Paperista application allows users to browse papers by topic. When a user clicks on a particular bubble (topic), a list of papers related to that topic appears in the right sidebar (represented by a title and a list of authors). Clicking on the particular paper opens a link to Google scholar with a name of the article as a search query. Thus, if a paper is available online, a user could easily obtain the paper using the Paperista system.

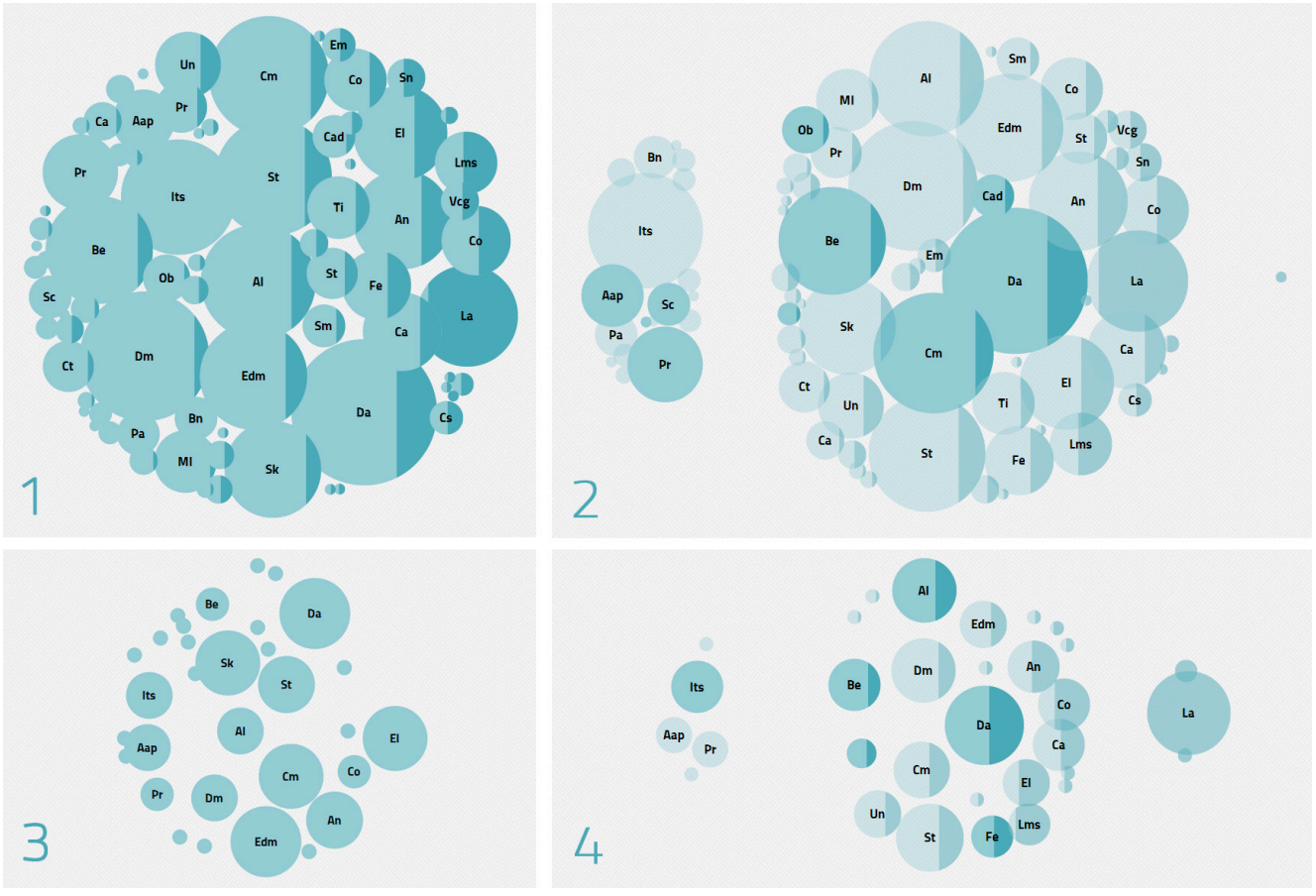

Figure 2 - Different views of the bubble chart: (1) All Year / All Topic (no highlights); (2) All Years / Group Topics (with highlights); (3) By Year (2009) / All Topics (no highlights); and (4) By Year (2012) / Group Topics (with highlights)

Furthermore, when a user hovers over the paper title, all topics related to that paper become highlighted. By hovering over papers, users can gain quick insight about topic connections between publications. Users can also distinguish papers annotated with highly relevant topics from those marked with insignificant ones. This can show which papers are more related to the fields of EDM and LA, and which can be viewed as "outliers".

3.3 Paperista Architecture and Dataset API

The Paperista system consists of a Web application and a server application that provides RESTful API for communicating with the dataset. The Web-based visualization is written in D3, a JavaScript library for manipulating documents based on data [3]. We have chosen D3 because of its good performance for animation and interaction within the Web environment. The visualization is available at the following address: www.uzrok.com/paperista.

All data about conference topics and their significance (explained in Section 3.1) is available as a part of Paperista Dataset API. This API supports a REST model for accessing the data and it is available at: http://147.91.128.71:9090/LAKChallenge2013. The Paperista's Web application calls these operations in order to access data from the dataset (for example, a click on a topic triggers a call to the API, which returns a list of papers).

4. DISCUSSION

When looking at the view displaying topic distribution in all years (Figure 2.1), one can observe that EDM conference dominates in almost all topics. This is due to the fact that EDM conference is being organized longer than the LAK conference (3 years longer), and thus the LAK dataset contains overall more papers coming from the EDM conference.

Filtering topics by years allows for observing the popularity of topics in a particular year and a particular field (LA or EDM). This further enables one to observe the shift in interest for a particular topic by researchers in the LA and EDM fields throughout the years. For instance, one can observe that before 2011, the topic of *Learning Analytics* was not much popular in the papers from the EDM field; thus this topic is not displayed at all in visualizations for years 2008-2010. In 2011, it boomed in popularity as indicated by the significant rise in the number of papers covering it. In fact, this was the first year the LAK conference was organized, and it immediately occupied the attention of researchers interested in the topic of Learning Analytics. Interestingly, this topic also gained some traction among the researchers publishing in the EDM field. In 2012, the topic's popularity grew even bigger and the researchers covering it directed their effort toward the LA field. This resulted in papers published within the LA field to almost exclusively cover the topic of *Learning Analytics*. Similarly, we can observe topics that have kept high popularity in both areas over years. For instance, this is the case with the *Data* topic, obviously as a consequence of research in both areas concentrating on the analysis of large amounts of data coming from various learning systems and other sources.

The application also allows us to observe that topics such as *Intelligent Tutoring System*, *Prediction* and *Accuracy and Precision* mostly kept their popularity throughout the years and stayed exclusively within the EDM field. On the other hand, one can observe that the large majority of topics have been covered by both fields. This suggests that the similarities between the two fields are significant as they share many research topics.

5. CONCLUSION

In this paper we have presented our approach to visualizing topics and their trends in the LA and EDM fields. Our application allows for easy identification of the main topics researchers in these fields have been focusing on, and also exploration of papers related to those topics.

When compared to other similar tools that provide visualization of research topics, our tool is the most similar to the previously mentioned Rexplore tool. However, while Rexplore is more focused on relations between authors and topics in research areas, Paperista's focus is on research topics and their trends over time. Also, Paperista allows for exploring papers related to different topics.

Future work for Paperista will be primarily directed towards extending the system to support other datasets, related to other research areas. Since the LAK dataset is RDF-based, Paperista can easily be expanded to support other RDF-based datasets expressed using the same or related vocabulary, such as the Semantic Web Dog Food corpus[12]. Regarding the interface, we plan to introduce keyword-based search functionality for searching a topic by its name. This would allow for easy navigation to a desired topic and filtering papers related to it. The final goal for Paperista is to become a universal visualization tool for research papers.

6. REFERENCES

[1] Carter, S. *Four Ways to Slice Obama's 2013 Budget Proposal*. New York Times, 2012. Available online: http://www.nytimes.com/interactive/2012/02/13/us/politics/2013-budget-proposal-graphic.html

[2] Bostok, M., Carter, S., and Ericson, M. *At the National Conventions, the Words They Used*. New York Times, 2012. Available online: http://www.nytimes.com/interactive/2012/09/06/us/politics/convention-word-counts.html

[3] Bostok, M., Ogievetsky, V., and Heer J. *D3: Data-Driven Documents*. IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis), 2011. Available online: http://vis.stanford.edu/papers/d3

[4] Dubinko, M. et. al. *Visualizing Tags over Time.* WWW 2006, Edinbourgh. Available online: http://labs.rightnow.com/colloquium/papers/visualizing_tags.pdf

[5] Zhang H., Korayem M., You E., and Crandall D. J. *Beyond Co-occurrence: Discovering and Visualizing Tag Relationships from Geo-spatial and Temporal Similarities*. Available online: http://www.cs.indiana.edu/~zhanhaip/wsdm2012-clustering.pdf

[6] Lemma, R. *Visualizing the DBLP Database*. Bachelor Thesis, 2010. Available online: http://www.inf.usi.ch/faculty/lanza/Downloads/Lemm2010a.pdf

[12] http://data.semanticweb.org

[7] Wattenberg, M. Arc Diagrams: Visualizing Structure in Strings. InfoVis 2002. Available online: http://hint.fm/papers/arc-diagrams.pdf

[8] Ponzetto, S. P., & Strube, M. (2007, July). Deriving a large scale taxonomy from Wikipedia. In *Proceedings of the national conference on artificial intelligence*(Vol. 22, No. 2, p. 1440). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999. Available online: http://www.h-its.org/english/research/nlp/papers/ponzetto07b.pdf

[9] Ferragina, P., & Scaiella, U. (2010, October). TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 1625-1628). ACM.

[10] Mendes, P. N., Jakob, M., García-Silva, A., & Bizer, C. (2011, September). Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems* (pp. 1-8). ACM.

[11] Osborne, F., & Motta, E. (2012). Mining semantic relations between research areas. *The Semantic Web–ISWC 2012*, 410-426.

[12] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. *The Semantic Web*, 722-735.

[13] Milne, D., & Witten, I. H. (2008, October). Learning to link with wikipedia. In*Proceedings of the 17th ACM conference on Information and knowledge management* (pp. 509-518). ACM.

[14] Siemens, G., & Long, P. (2011). Penetrating the Fog: Analytics in Learning and Education. *Educause Review*, *46*(5), 30-32.

[15] Osborne, F., & Motta, E. (2012). Making Sense of Research with Rexplore. *The Semantic Web–ISWC 2012*

[16] 1st International Conference on Learning Analytics and Knowledge, Banff, Alberta, February 27–March 1, 2011, link https://tekri.athabascau.ca/analytics/

[17] Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, *40*(6), 601-618.

[18] Siemens, G., & Baker, R. S. D. (2012, April). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 252-254). ACM.