Automated Paper Annotation with *ReaderBench*

Ionut Cristian Paraschiv, Mihai Dascalu, Stefan Trausan-Matu
University Politehnica of Bucharest
313 Splaiul Independentei, Bucharest, Romania
ionut.paraschiv@cti.pub.ro,
mihai.dascalu@cs.pub.ro,
stefan.trausan@cs.pub.ro

Philippe Dessus
LSE, Univ. Grenoble Alpes
Grenoble CEDEX 9 France
philippe.dessus@upmf-grenoble.fr

ABSTRACT
The annotation of articles from a given domain and the generation of semantic metadata can be considered a reliable foundation for creating a paper recommender system. Within this paper, the models from other previous researches are extended with the capability of visualizing articles and the most important concepts from a domain within imposed timeframes. This can be very useful for researchers to check out the most important publications from a given period, to view which are the trends and how a domain has evolved. Our previous analyses used the articles to build a paper graph and to suggest the most relevant articles, given a user defined query in natural language. This research contains a use case and creates visual graph representations to enhance the overall perception of the evolution of a domain.

Author Keywords
Scientometrics; paper recommendation system; time analysis; discourse analysis; semantic similarity.

ACM Classification Keywords
I.2.7 [**Natural Language Processing**]: Discourse, Language parsing and understanding, Text analysis.

INTRODUCTION
A researcher's daily activities usually involve the study of new papers, as to use the information in building solutions and observing how the domain evolves. Since the retrieval of documents from the Internet can lead to large data flows, it is important to consider other approaches for a more comprehensive analysis of the domain. In this context, a paper annotation system that automatically retrieves papers on a given topic and tags them can be critical and can make the exploration phase of the research literature easier.

We propose a model that takes a large set of paper abstracts and tags them, later on annotating the results within a semantic database. The database can be queried for user defined texts, and can enable the researcher to explore the resulting graphs using different timeframes, as to retrieve the most important articles and concepts within a period. Moreover, a list of similar topics with the user's query is shown with the intention of stimulating the user in his/her research tasks.

The initial part from this paper will concentrate more on similar studies that discuss on how to build network graphs for scientific papers. We continue with the methods used behind the current model that demonstrate its potential and extensibility, as well as how they are used, and we finish with possible future improvements of our system.

RELATED WORK
A paper annotation system can be built with database software that uses keyword matching such as *Mendelev* or *DevonThink*, or other complex methods [11]. However, the current model does not rely on information retrieval [13] as we describe an alternative of annotating a dataset of documents. A different, older approach of indexing papers relies on co-citation analysis [6; 7], but that method is out of scope in terms of the current research.

Information Retrieval
Information Retrieval [13] techniques aim at finding materials of an unstructured nature, usually text, that satisfy information needs from within large textual collections. This process is usually concerned on how to store and structure data in such a manner that it will facilitate the retrieval of information based on a given query and in a relatively small amount of time. Being a text recognition tool, the query, consisting of a Boolean combination of keywords, is usually mapped with the collection; therefore, no semantic meaning is associated to the query or to the set of documents. In this way, the more complicated or complex the query is, the smaller are the chances of finding relevant results.

Semantic Similarity Analysis of Paper Abstracts
Our paper annotation model relies on two widely used methods. The first one, Latent Semantic Analysis [6; 7], is a natural language processing method that is used for analyzing relationships between documents and their terms, in our particular case – abstracts [9]. The method builds a document-term matrix that basically assigns for every word its corresponding number of occurrences within the document. After applying a Singular Value Decomposition, the dimensionality of the matrix is reduced, while keeping its similarity structure with a marginal error. At the end, the documents are compared by computing the cosine similarity between their associated vectors within the semantic space.

The second method, Latent Dirichlet Allocation [2], uses topic distributions among documents, and in combination with LSA can give an aggregated cohesion scores [4] that can be more accurate for computing semantic distances. In the end, in order to compute the semantic distance between two words, our model also relies on semantic distances extracted from the lexicalized ontology WordNet, together with LSA and LDA semantic models [3].

SYSTEM'S IMPLEMENTATION

Using the approaches described in previous subsections, the papers from the initial dataset were tagged using the content from their abstracts, that usually contains the main ideas [8], as seen in our previous research [14]. In this manner, three different views on the dataset of articles have been built. The first one, the document similarity view [14], generates a graph with the papers which are connected if their semantic similarity exceeds a threshold. The second one, the concept map view [14], extracts the most important concepts from the subset of papers, and builds a graph where the words are connected using their relevancies [4]. In both these two views, the nodes are sized depending on their centrality, and the links between them are enforced based on the similarity. In this way, the user will see at the end which are the most important words and documents from the set of papers. The third view displays the document space for a particular paper [14], and can help researchers when they want to read semantically related articles.

After this first implementation phase which was previously presented [14], the improvements presented in this paper are focused on enabling the user to define his/her own queries. In the end, similar graphs to the document similarity view [14] are displayed, but with the papers that have a high semantic cohesion with the text introduced by the user. Moreover, the user can also check semantically related concepts with the query terms (concepts that do not specifically appear inside the query), thus stimulating his/her imagination with ideas for new queries. This is the first exploratory extension of our system.

The text given by the user enters inside the same pipeline as the abstracts: text preprocessing, lemmatization, part-of-speech tagging, syntactic dependency analysis and topic extraction [4; 12]. The next step is to represent the query using LSA and LDA vectors, and to compute its distance with every document from the dataset. The LSA query vector is obtained by summing up normalized occurrences of each constituent lemma vector representation, whereas for LDA the Gibbs inference tools is applied on the query in order to deduce the topic distributions based on pre-trained models. In order to increase the user's control, the final view shown in Figure 1 also contains a threshold used for displaying the links between the papers and the query, which can be manually adjusted. The documents are also shown inside a table, ordered by their similarity with the input query.

The second extension presented in this paper consists of enabling the user to select a timeframe for the displayed papers and concepts. In this manner, a researcher can check the evolution of a domain, the most important articles in a period and the most central concepts. As many domains evolve in a dynamic manner, this is definitely the way to check past and current trends, as well as concepts that become more important.

From a technical point of view, the paper annotation system uses the core components from *ReaderBench* [4; 5], a versatile tool for text and discourse processing. With a fully functional natural language processing pipeline incorporated, LDA and LSA semantic pre-trained models, WordNet and semantic distances, as well as Social Network Analysis [1], *ReaderBench* is an extensible tool for most of the undergone text processing steps.

These two extensions have the purpose of stimulating a researcher in his daily tasks by suggesting papers, similar concepts, as well as the modeling of the evolution of a domain within a period of time. Together, they can definitely support anyone interested in learning a domain. The next section is centered on specific use cases of how the system can be used with a real dataset of paper abstracts.

USE CASE

Dataset Description

The used dataset of abstracts consists of papers extracted from the citation index Web of Science, from the Education and Educational Research [10] domain, taken between the years 2000-2004. From this dataset, a subset of paper which contained within their abstracts the keywords "*IT"*, "*technology"* or "*computer science"* was extracted.

User Queries

Given the database of annotated abstracts, the user inputs an initial query "*electronic learning and information technology*", with the intention of finding papers related to informational systems that implement learning facilities. Using a threshold value of 70%, Figure 1 displays a sub-graph with the most related articles with the query. Table 1 displays what is the content of these articles.

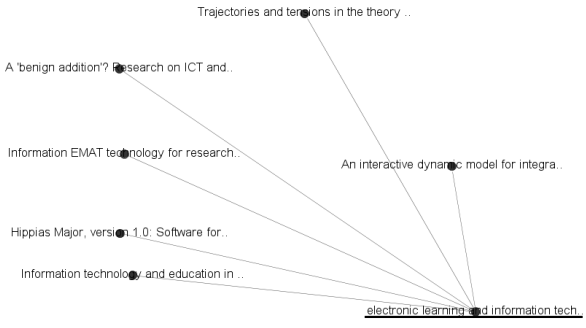

Figure 1. Semantically related articles to the input query

From Table 1, we can observe that the most related abstracts with the user's query are semantically related, and unlike standard information retrieval systems, the results don't necessarily have common words with the query. This can impact the retrieval of enhanced search results, and can definitely help anyone in finding documents about a subject.

For the example query in this subsection, the researcher could be interested to check some related concepts with his/her query. The system is capable of displaying the related keywords in a graph, which can be explored by

the user. In this context, the model suggests words such as "learner", "teacher", "science", "curriculum", "research", "process", and "development" (see Figure 2), which are quite relevant given the input query. Moreover, we can check the underscored words as being from the input query, while the others are being automatically introduced as semantically related to them. All the words with a threshold over .5 in terms of semantic similarity are shown, and they are grouped together using the same stem; in the end, the shortest lemma is being displayed.

Paper title and abstract	Relevance
Trajectories and tensions in the theory of information and communication technology in education "For largely historical reasons, information and communication technology in education has been heavily influenced by a form of constructivism based on the transmission and transformation of information. This approach has implications for both learning and teaching in the field. The assumptions underlying the approach are explored and a critique offered...."	.78
Information technology and education in the information age "This paper attempts to gain an understanding of current and potential impact of information technology (IT) on education in the information age. First, it attempts to highlight that integration of information technology in teaching is a central matter in ensuring quality in the educational system. ..."	.74
Hippias Major, version 1.0: Software for post-colonial, multicultural technology systems "The first half of Plato's Hippias Major exhibits the interfacing of the first teacher (Socrates) with the first version of a post-colonial, multicultural information technology system (Hippias). In this interface the purposes, results and values of two contradictory types of operating system for educational servicing units are exhibited to.."	.72

Table 1. Samples from the related papers

Unlike standard information retrieval systems, this model performs better when it comes to user queries, as in this case, more complex queries usually mean richer semantic content, and thus more accurate results.

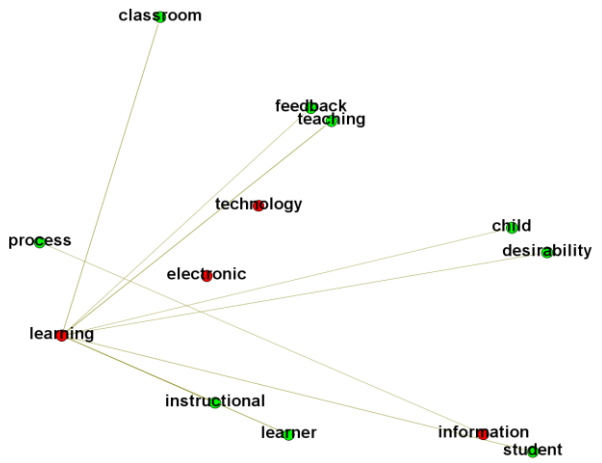

Figure 2 Semantically related concepts

Timeframe View

Although not a computationally demanding task, the ability to filter articles and important concepts using different timeframes can be very appealing to researchers, as they can check the trends and the evolution in time of a specific domain. In Figure 3, a subgraph from the article similarity view from the year 2002 is shown, displaying the central article having a relevance of 0.81.

Paper title and abstract
A collaborative, investigative recombinant DNA technology course with laboratory "A recombinant DNA technology course was designed to promote contextual, collaborative, inquiry-based learning of science where students learn from one another and have a sense of ownership of their education. The class stressed group presentations and critical reading and discussion of scientific articles. The laboratory consisted of two research projects: random cDNA …."
Technology in the first two years of collegiate mathematics "We present several roles of technology and suggest various ways that technology could have a lasting and significant impact upon the quality of mathematics courses being taught in the first two years of collegiate mathematics. Overcoming some mathematicians' anxiety and reluctance to address applied problem solving so as to take full advantage of the opportunities remains a challenge for the future…"
Maximising the educational affordances of a technology supported learning environment for introductory undergraduate phonetics "New technologies afford a range of opportunities that can transform teaching techniques and offer enhanced possibilities for learning. This potential is often not grasped by the technologist or the educationalist when introducing new technologies into the learning situation and a situation arises which can be described as 'New technology, no new pedagogy.'…"

Table 2. Central article summaries in 2002

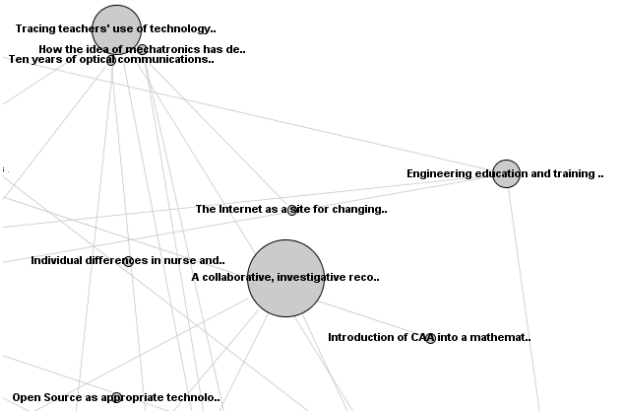

Figure 3. Subgraph of the article similarity view – Year 2002

The graph of the most important concepts from 2002 is displayed in Figure 4.

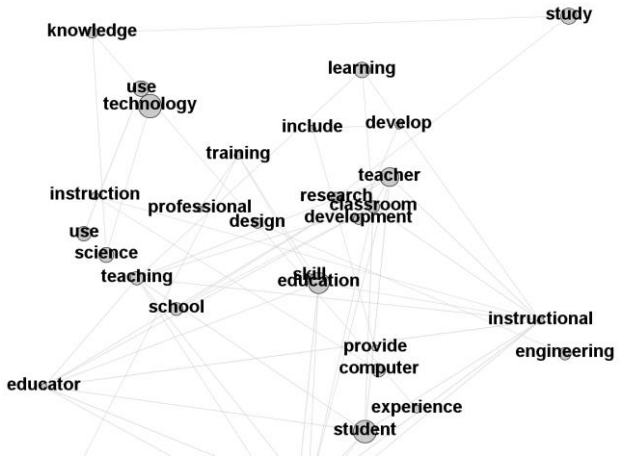

Figure 4. Central concepts in 2002

As displayed in Figure 4, the most important concepts from the dataset in 2002 are *educational, classroom, training and instruction*. It will be very interesting to check how the central article's topics and the corresponding concept graph have changed in 2004.

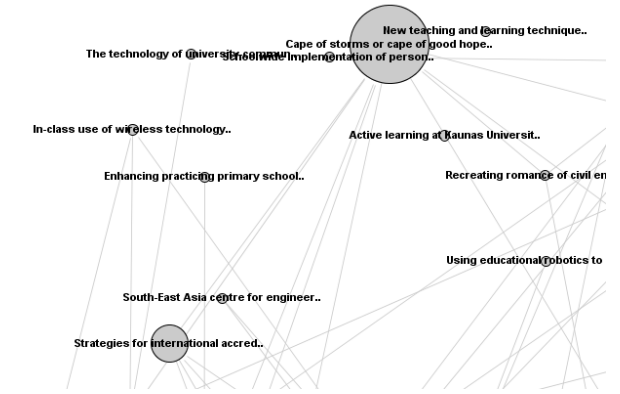

Figure 5. Subgraph of the article similarity view – Year 2004

Paper title and abstract
Cape of storms or cape of good hope? Educational technology in a changing environment "This article locates and describes the work of the Multimedia Education Group (MEG) at the University of Cape Town (UCT). This work is contextualised by three national and international challenges, these being (1) the need to increase access to new technologies and overcome the digital divide, (2) the need to respond to a new communication order, and (3) the urgency of transforming higher education…"
Technology and curricular reform in China: A case study "This article reports on a 5-year study of a technology-enhanced educational reform initiative at a university in eastern China. A facility team attempted pedagogical and curricular reform to better prepare English Majors to use new technologies for international communication, collaboration, and research. The team developed several project-based courses and incorporated technology into traditional lecture courses..."
Web-based curriculum development of a manufacturing engineering technology programme "The aim of this paper is to present the use of the Internet in developing the curriculum of a manufacturing engineering technology programme in Turkey. The programme was implemented in the curricula of 15 two-year colleges over six months to provide seamless progression from vocational high school to two-year colleges and meet the needs of Turkish and global industry…"

Table 3. Central articles summaries in 2004

Overall, it becomes clear that there is a difference in terms of semantic meaning between the central articles from 2002 and the ones in 2004. As technology has evolved in that period, the central articles are more related to information technology applied inside learning environments, which supports the idea that the user can check using our system how the domain has evolved in time.

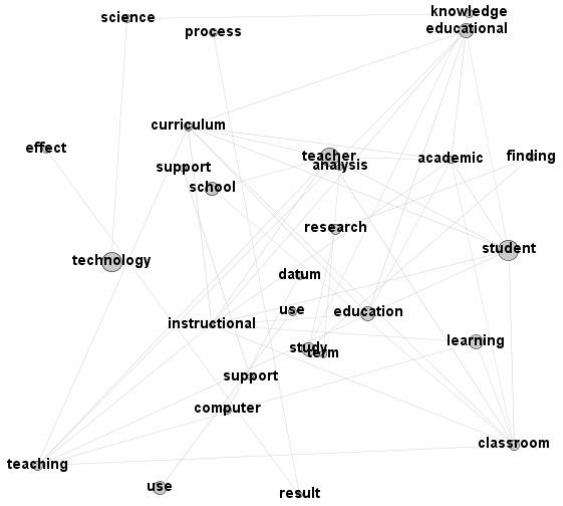

Figure 6. Central concepts in 2004

To sustain this idea, Figure 6 displays some of the most important concepts from 2004, where the most preeminent lemmas are "*technology", "education", "teacher", "study", "learning", "science"* and *"computer"*. They demonstrate the idea that the domain has evolved as new technologies appeared and were applied inside the educational research domain.

Another interesting experiment on the two datasets of papers was to find out which are the most similar concepts. In this manner, we have extracted the most important 100 concepts from the papers written in 2002, Dataset 1 (DS1), and from those written in 2004 (DS2). Every concept from DS1 was compared with every concept from DS2, and Table 4 displays the most similar pairs from the two subsets. In this manner, the researcher can get additional information regarding the domain and how it has evolved.

Concept from 2002	Concept from 2004	Sim.
Classroom	Student	.92
Educator	Student	.86
Design	Engineering	.75
Knowledge	Experience	.71
Educational	Student	.71
Professional	Practice	.66

Table 4. Concept similarities from the subsets

Further on, Table 5 depicts the most similar abstracts between the two subsets, being a good metric to see how the content has evolved in the two years. The idea of the table is that, by checking on the articles that are semantically related in different periods, a researcher can observe a domain's trends and what emerging solutions have appeared regarding certain problems.

Article pairs (marked as X.a and X.b)	Score
1.a *Introduction of CAA into a mathematics course for technology students to address a change in curriculum requirements* **(2002)** The mathematical requirements for engineering, science and technology students has been debated for many years and concern has been expressed about the mathematical preparedness of students entering higher education. This paper considers a mathematics course that has been specifically designed to address some of these issues for technology education students …	.955
1.b *Standardized test outcomes of urban students participating in standards and project based science curricula* **(2004)** Considerable effort has been made over the past decade to address the needs of learners in large urban districts through scaleable reform initiatives. We examine the effects of a multifaceted scaling reform which focuses on supporting standards based science teaching in urban middle schools. The effort was one component of systemic reform efforts …	
2.a *Integrating algorithm visualization technology into an undergraduate algorithms course: ethnographic studies of a social constructivist approach* **(2002)** Algorithm visualization (AV) software graphically illustrates how algorithms work. Traditionally, computer science instructors have used the software as a visual aid in lectures, or as the basis for interactive laboratories. An alternative approach, inspired by Social Constructivist learning theory …	
2.b. *Classroom use of multimedia-supported predict-observe-explain tasks in a social constructivist learning environment* **(2004)** This paper focuses on the use of multimedia-based predict-observe-explain (POE) tasks to facilitate small group learning conversations. Although the tasks were given to pairs of students as a diagnostic tool to elicit their pre-instructional physics conceptions, they also provided a peer learning opportunity for students. The study adopted a social constructivist perspective...	.913
3.a *Conditions for classroom technology innovations* **(2002)** This article reports on a study of the complex and messy process of classroom technology integration, The main purpose of the study was to empirically address the large question of "why don't teachers innovate when they are given computers?" rather than whether computers can improve student learning. Specifically, we were interested in understanding the conditions …	
3.b. *New teaching and learning techniques facilitated by information technology* **(2004)** A wide variety of classroom techniques are being advocated to increase learning: active learning, collaboration, integration of assessment and feedback, and the use of concrete physical manipulatives. These techniques must be transformed into practical tools and be infused with content from the subject area. At the same time, the information technology revolution has provided new tools …	.898

Table 5. Most similar articles from the subsets

This section presented in detail a generalizable use case that can be easily extrapolated on any dataset of papers and can enable researchers to better understand a domain. Moreover, the results clearly indicate that the evolution of a domain can be better understood by analyzing the semantic content of the articles within certain timeframes.

CONCLUSIONS

As more and more research communities appear and they are more dynamic than ever, it is becoming quite hard for a researcher to keep up-to-date with this fast growing information. In this context, a paper annotation model and viewer can be a good alternative to better visualize the papers from a certain dataset. Moreover, the support for user defined queries, graphical visual representations and timeframe filtering increase the overall understanding of a domain and, in the end, the productivity of researchers.

As future developments, the timeframe snapshot will be displayed in an interactive and animated manner, not just as a static graph within a period of time. Moreover, a current drawback must be addressed: the long preprocessing time due to the NLP processing pipeline, and a relatively small number of possible papers that can be loaded directly into the system's memory. Therefore, some improvements must be done for our model in terms of memory and CPU consumption. In order to address these issues, clusters of papers will be created and the search will be conducted within a multi-hierarchical structure of documents..

ACKNOWLEDGMENTS

The work presented in this paper was partially funded by the FP7 2008-212578 LTfLL project and by the Sectorial Operational Programme Human Resources Development 2007-2013 of the Ministry of European Funds through the Financial Agreements POSDRU/159/1.5/S/134398 and POSDRU/187/1.5/S/155420. We also thank Pablo Jensen and Sebastian Grauwin for providing the initial corpus of paper abstracts.

REFERENCES

1. Bastian, M., Heymann, S., and Jacomy, M., 2009. Gephi: An open source software for exploring and manipulating networks. In *International AAAI Conference on Weblogs and Social Media* AAAI Press, San Jose, CA, 361–362.
2. Blei, D.M., Ng, A.Y., and Jordan, M.I., 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*, 4-5, 993–1022.
3. Budanitsky, A. and Hirst, G., 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics 32*, 1, 13–47.
4. Dascalu, M., 2014. *Analyzing discourse and text complexity for learning and collaborating, Studies in Computational Intelligence*. Springer, Switzerland.
5. Dascalu, M., Dessus, P., Bianco, M., Trausan-Matu, S., and Nardy, A., 2014. Mining texts, learners productions and strategies with ReaderBench. In *Educational Data Mining: Applications and Trends*, A. Peña-Ayala Ed. Springer, Switzerland, 335–377.
6. Deerwester, S., Dumais, S.T., Furnas, G.W., Harshman, R., Landauer, T.K., Lochbaum, K., and Streeter, L., 1989. Computer information retrieval using latent semantic structure In *4,839,853*, Uspto Ed. Bell Communications Research, Inc., USA.
7. Deerwester, S., Dumais, S.T., Landauer, T.K., Furnas, G.W., and Beck, L., 1988. Improving Information Retrieval with Latent Semantic Indexing. In *51st Annual Meeting of the American Society for Information Science 25*, 36–40.
8. Ding, Y., Song, M., Wang, X., Zhang, G., C., Z., and Chambers, T., 2014. Content-based citation analysis: The next generation of citation analysis. *Journal of the American Society for Information Science & Technology 65*, 9. DOI= http://dx.doi.org/10.1002/asi.23256.
9. Gordon, M.D. and Dumais, S., 1998. Using Latent Semantic Indexing for literature based discovery. *Journal of the American Society for Information Science 49*, 8, 674–685.
10. Grauwin, S. and Jensen, P., 2011. Mapping scientific institutions. *Scientometrics*. DOI= RoCHI_v2.0.doc DOI 10.1007/s11192-011-0482-y.
11. Joeran, B., Langer, S., Genzmehr, M., Gipp, B., Breitinger, C., and NüRnberger, A., 2013. Research Paper Recommender System Evaluation: A Quantitative Literature Survey. In *Workshop on Reproducibility and Replication in Recommender Systems Evaluation (RepSys) at the ACM Recommender System Conference (RecSys'13)* ACM, Hong Kong, China.
12. Jurafsky, D. and Martin, J.H., 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. Pearson Prentice Hall, London.
13. Manning, C.D., Raghavan, P., and Schütze, H., 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.
14. Paraschiv, I.C., Dascalu, M., Trausan-Matu, S., and Dessus, P., 2015. Analyzing the Semantic Relatedness of Paper Abstracts - An Application to the Educational Research Field. In *2nd Int. Workshop on Design and Spontaneity in Computer-Supported Collaborative Learning (DS-CSCL-2015), in conjunction with the 20th Int. Conf. on Control Systems and Computer Science (CSCS20)* IEEE, Bucharest, Romania, 759–764.