COMPREHENSIVE SOCIOGRAMS OF THE SCIENTIFIC BULLETIN COMMUNITY

Remus Florentin IONITA[1], Dragos Georgian CORLATESCU[2], Stefan RUSETI[3], Mihai DASCALU[4], Stefan TRAUSAN-MATU[5], Nicolae TAPUS[6], Cosmin Karl BANICA[7]

The process of identifying relevant scientific papers and authors has become more and more problematic to unfamiliar individuals, as well as experts searching for state-or-the-art approaches, due to the high-speed at which emerging domains evolve and the rapid increase of available publications. This paper presents our overarching architecture grounded in Cohesion Network Analysis, while focusing on the papers published within the Scientific Bulletin from University Politehnica of Bucharest. Our method provides valuable insights about authors' semantic relatedness based on research topics, different types of associations identified as links within the 2-mode graph, as well as statistics about top ranked authors and articles.

Keywords: Sociograms, Cohesion Network Analysis, 2-mode graph, Sociometric study

1. Introduction

Communities of different types are gaining a broader adoption, and sociograms are frequently used to represent relations between different members [1]. For example, analyses of the interactions between members are useful to detect the most influential individuals in a community or to potentially identify members with similar interests. In addition, sociograms can be also used to reflect relations between authors based on the content and documents' meta-information. This paper introduces a comprehensive analysis of the articles and of authors from

[1] MSc student, Dept. of Computer Science, University POLITEHNICA of Bucharest, Romania, e-mail: ionitaremusflorentin@gmail.com

[2] PhD student and Teaching Assistant, Dept. of Computer Science, University POLITEHNICA of Bucharest, Romania, e-mail: dragos.corlatescu@cs.pub.ro

[3] PhD student and Teaching Assistant, Dept. of Computer Science, University POLITEHNICA of Bucharest, Romania, e-mail: stefan.ruseti@cs.pub.ro

[4] Assoc. Prof., Dept. of Computer Science, University POLITEHNICA of Bucharest, Romania, Project director Research Technology S.R.L., e-mail: mihai.dascalu@cs.pub.ro

[5] Prof, Dept. of Computer Science, University POLITEHNICA of Bucharest, Romania, Researcher Research Technology S.R.L., e-mail: stefan.trausan@cs.pub.ro

[6] Prof., Dept. of Computer Science, University POLITEHNICA of Bucharest, Romania, e-mail: nicolae.tapus@cs.pub.ro

[7] Assoc. Prof., Dept. of Electrical Engineering, University POLITEHNICA of Bucharest, Romania, Director Research Technology S.R.L., e-mail: cosmin.banica@upb.ro

the Scientific Bulletin of University Politehnica of Bucharest community, using a general framework that includes crawling, indexing, semantic similarity measures, and graph visualizations, all grounded in Cohesion Network Analysis graph [2].

Communities are usually evaluated based on existing links between individuals, which can be considered to generate an underlying graph. Social Network Analysis (SNA) [3, 4] provides different metrics to measure the importance of nodes in such graphs, like betweenness and centrality. In case of author networks, links can be used to model similarities between their articles. Thus a 2-mode graph can be generated, containing two types of nodes: authors and articles. In such graphs, the similarity between authors can be computed based on the similarity between articles. Usually, three different types of links between articles are used in order to evaluate links between articles and corresponding authors: co-authorship, co-citations, and semantic similarity of the content [5]. In the current experiments we used the *ReaderBench* framework [6, 7] to compute the semantic distances between article abstracts and to build the 2-mode CNA graph [5].

The paper continues with a brief description of the Scientific Bulletin corpus of scientific articles, our system’s architecture together with the method for building the 2-mode CNA graph. Afterwards, the network graphs generated from the corpus are discussed, highlighting interesting insights extracted from the data.

2. Method

Corpus of articles from the Scientific Bulletin

The Scientific Bulletin is an online journal available at https://www.scientificbulletin.upb.ro containing research articles mostly published by research teams and professors from the University Politehnica of Bucharest. The journal has been published for more than 80 years and includes research papers from four major areas of study, namely: Series A – Applied Mathematics and Physics; Series B – Chemistry and Material Science; Series C – Electrical Engineering and Computer Science; and Series D – Mechanical Engineering. We have gathered data from all four series, counting a total of 3792 articles and 7220 unique authors. Fig. 1 illustrates the distribution of articles and authors, per series.

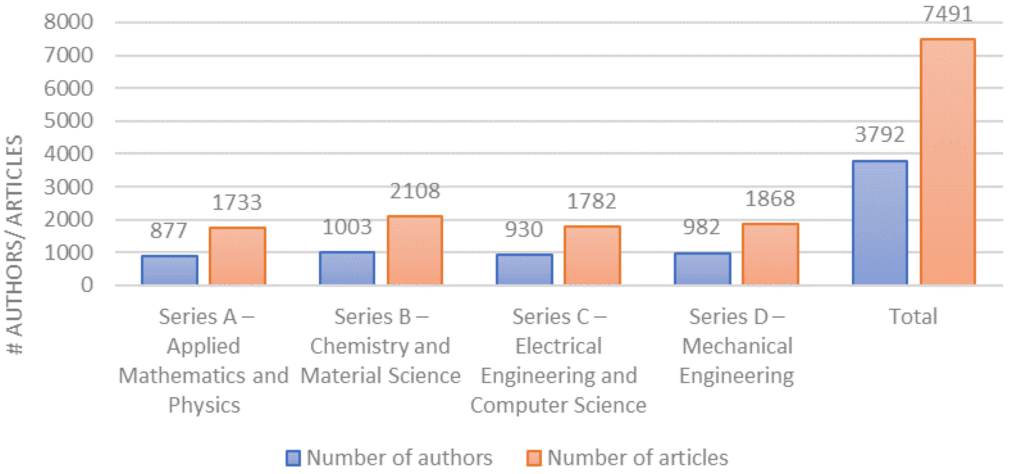

Fig. 1. The number of analyzed articles and authors per series.

Fig. 2 depicts the number of published articles per year for each series. Series C - Electrical Engineering and Computer Science has published, in average, the highest number of articles. All articles were downloaded in pdf format directly from the official site.

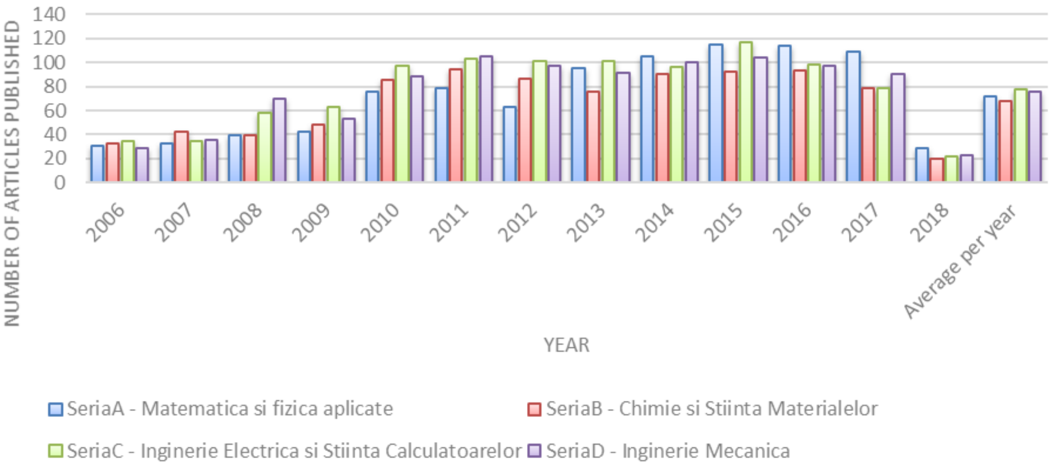

Fig. 2. The number of articles published per year, by series.

Architecture

The proposed processing architecture is depicted in Fig. 3 and it consists of three major stages. First, the Data Extractor component performs the following actions: a) crawling of .pdf articles, b) parsing of documents using Grobid (https://github.com/kermitt2/grobid), and c) saving the document's corresponding metainformation in the ElasticSearch full-text search engine and database (https://www.elastic.co/products/elasticsearch [9]. In order to automatically extract the relevant information from each article, we used Grobid, an open source machine-learning library specialized in parsing and extracting information from pdf format articles and packing them into TEI (Text Encoding Initiative) format [8]. The Grobid Server node works as a standalone server and a bash script calls the method '/api/processHeaderDocument' in order to get the necessary information from the raw articles which are now transposed into XML TEI format.

However, we could not achieve 100% accuracy on parsing all the Scientific Bulletin articles because Grobid is not always able to extract correctly information from the raw pdf format. For example, Grobid server usually fails on extracting the author affiliations. Moreover, articles in which the same author was inserted twice in the resulted XML TEI format were also identified. In order to overcome this issue, we only focused on the mandatory fields of an article: title, authors, published date and abstract. If any of those fields were missing from the Grobid resulting XML, the corresponding article was ignored.

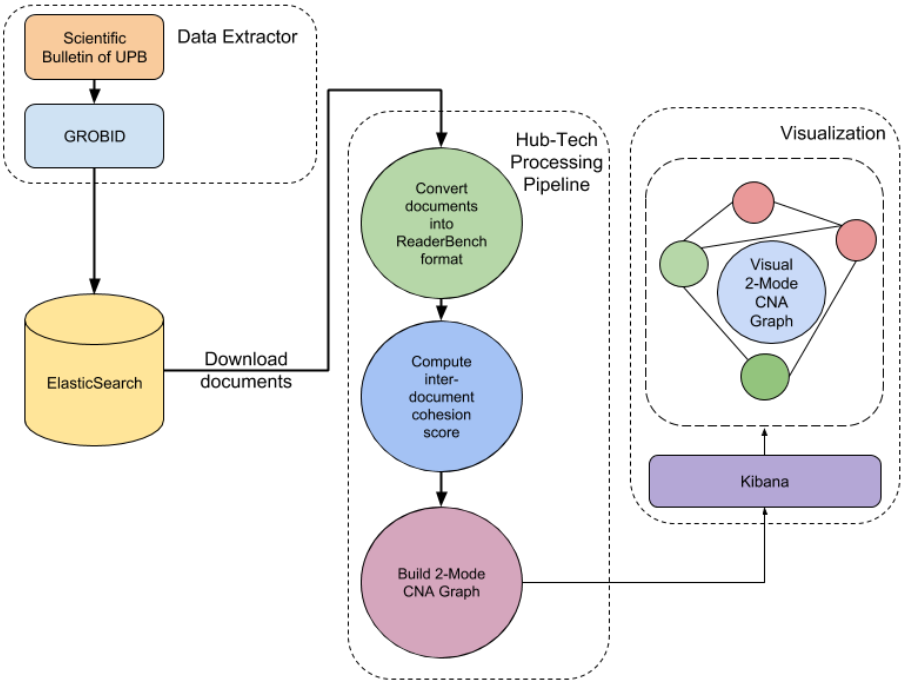
Scientific Bulletin of UPB
Data Extractor
GROBID
ElasticSearch
Download documents
Hub-Tech Processing Pipeline
Convert documents into ReaderBench format
Compute inter-document cohesion score
Build 2-Mode CNA Graph
Visualization
Visual 2-Mode CNA Graph
Kibana

Fig. 3. The detailed architecture of the system.

Second, the processing pipeline from the Hub-Tech project [5, 10] was adapted to perform specific analyses for the current project. The pipeline provides a fast environment for analyzing large datasets of documents and it uses AKKA (https://akka.io) for both parallel and distributed computing [11]. AKKA uses the master-worker architecture in which the master submits the documents for preprocessing, receives the converted documents using the *ReaderBench* framework [6], and builds batches of two documents that are sent to workers in a round-robin procedure for generating the 2-mode CNA graph which is passed to the next stage. The 2-mode CNA graph considers two types of nodes, namely author nodes and article, and the relationships between articles is reflected by three types of links: co-authorship, co-citation and semantic similarity [5]. Several rules have been introduced to refine the method:

- If the semantic similarity between two article abstracts is below an imposed threshold empirically set at 0.3, the link is disregarded;
- The value of the edge between two authors is computed as the mean of semantic similarities between the articles they both wrote (i.e., pairwise comparisons between the abstracts of all their papers). The previous rule still applies.
- Two specific cases need to be considered in terms of the relations between individuals and articles: a) if the author wrote an article, then the edge is assigned a value of 1 (perfect relatedness); b) otherwise, the edge is assigned the mean of the semantic distances between all the articles that the author wrote versus the article represented in the graph node. Similar to the previous cases, if the mean is below 0.3, the edge is disregarded.

Third, the visualization component is a web interface which renders the 2-mode CNA graph in order to provide a deeper understanding of the relationships between authors from the Scientific Bulletin community. Kibana (https://www.elastic.co/products/kibana), a framework seamlessly integrated with ElasticSearch, provides extensive visualizations. Kibana allows users to explore the dataset and discover authors and articles based on the associations derived from the 2-mode CNA graph. Article-article edges are colored in red, author-author edges are blue, whereas author-article edges are yellow. Edge values directly influence their visual size. A value closer to one (i.e., maximum relatedness) increases the edge width. The author’s number of publications also influences the size of author-article edges. In addition, edges drawn as dashed lines have the following interpretation:

- All parent-edges are drawn with grey dashed lines;
- Author-author edges drawn with blue dashed lines suggest that that two authors have published together a singular article, and none of them have published anything else. The author-author value between them is one;
- Author-article edges drawn with yellow dashed lines suggest that the author published only one article, that to which the author-article link is established.

3. Results and Discussions

Specific analyses on a given community can be done either at global level or locally focusing on a given author or article. The latter approach is performed by applying a breadth-first search from the desired node and using only the highest weighted links. As a first example, Fig. 4 presents small sub-communities formed around professors from various departments of the university. Clusters are centered on some of the most active professors from the Computer Science department of UPB: *Nicolae Țăpuș*, *Ștefan Trăușan-Matu*, *Adina Magda Florea*, and *Florica Moldoveanu*. In addition, Fig. 4 illustrates a potential use case of the provided visualizations. Professors from different clusters have published articles on different subjects.

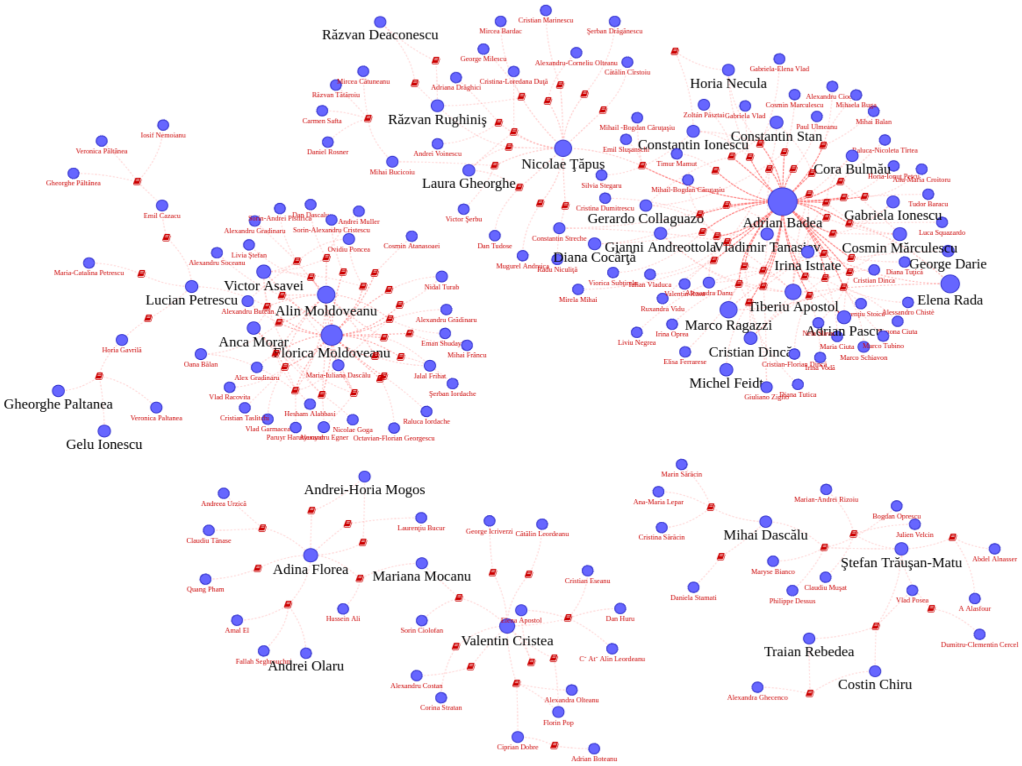

Fig. 4. Zoom on several departments of University Politehnica Bucharest, section Computer Science.

For example, the cluster center around *Ștefan Trăușan-Matu*, *Mihai Dascălu*, *Costin Chiru* and *Traian Rebedea* have published articles related to Natural Language Processing, whilst the cluster centered on *Nicolae Țăpuș*, *Răzvan Deaconescu* and *Răzvan Rughiniș* has published articles on Operating Systems and Computer Networks.

Fig. 5 introduces a network graph for two authors of this paper focused on the semantic similarities between article abstracts, a view which can be used to explore similarities between research topics. Our method is a variation of a breadth first search adapted for a graph with two types of nodes and four types of edges, together with imposed constraints of a maximum depth of three levels and a branching factor below 20.

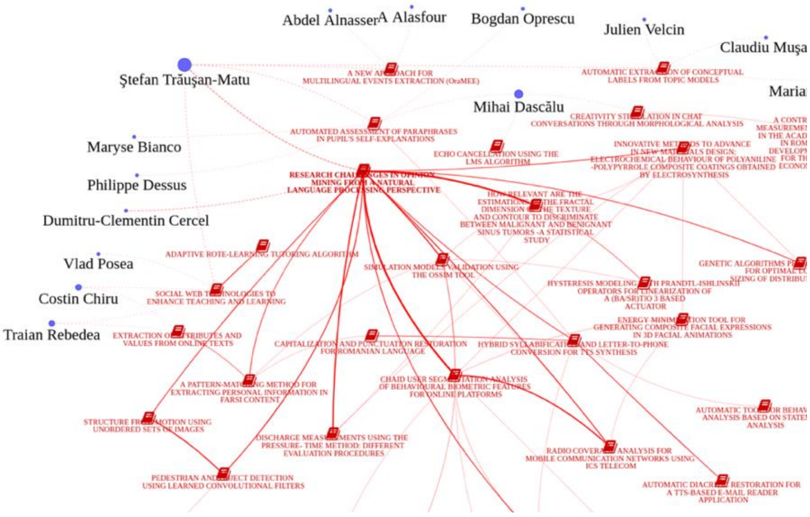
Abdel Alnasser
A Alasfour
Bogdan Oprescu
Julien Velcin
Claudiu Muşa
Ştefan Trăuşan-Matu
Maria
Mihai Dascălu
Maryse Bianco
Philippe Dessus
Dumitru-Clementin Cercel
Vlad Posea
Costin Chiru
Traian Rebedea
A NEW APPROACH FOR MULTILINGUAL EVENTS EXTRACTION (OraMEE)
AUTOMATIC EXTRACTION OF CONCEPTUAL LABELS FROM TOPIC MODELS
CREATIVITY STIMULATION IN CHAT CONVERSATIONS THROUGH MORPHOLOGICAL ANALYSIS
AUTOMATED ASSESSMENT OF PARAPHRASES IN PUPIL'S SELF-EXPLANATIONS
ECHO CANCELLATION USING THE LMS ALGORITHM
INNOVATIVE METHODS TO ADVANCE IN NEW MATERIALS DESIGN: ELECTROCHEMICAL BEHAVIOUR OF POLYANILINE -POLYPYRROLE COMPOSITE COATINGS OBTAINED BY ELECTROSYNTHESIS
A CONTR MEASUREMEI IN THE ACAD IN ROM DEVELOP FOR TH ECONO
RESEARCH CHALLENGES IN OPINION MINING FROM A NATURAL LANGUAGE PROCESSING PERSPECTIVE
HOW RELEVANT ARE THE ESTIMATIONS OF THE FRACTAL DIMENSION OF THE TEXTURE AND CONTOUR TO DISCRIMINATE BETWEEN MALIGNANT AND BENIGNANT SINUS TUMORS -A STATISTICAL STUDY
ADAPTIVE ROTE-LEARNING TUTORING ALGORITHM
SIMULATION MODEL VALIDATION USING THE OSSIM TOOL
GENETIC ALGORITHMS P FOR OPTIMAL SIZING OF DISTRIBU
HYSTERESIS MODELING WITH PRANDTL-ISHLINSKII OPERATORS FOR LINEARIZATION OF A (BA/SR)TIO 3 BASED ACTUATOR
SOCIAL WEB TECHNOLOGIES TO ENHANCE TEACHING AND LEARNING
ENERGY MINIMIZATION TOOL FOR GENERATING COMPOSITE FACIAL EXPRESSIONS IN 3D FACIAL ANIMATIONS
EXTRACTION OF ATTRIBUTES AND VALUES FROM ONLINE TEXTS
CAPITALIZATION AND PUNCTUATION RESTORATION FOR ROMANIAN LANGUAGE
HYBRID SYLLABIFICATION AND LETTER-TO-PHONE CONVERSION FOR TTS SYNTHESIS
A PATTERN-MATCHING METHOD FOR EXTRACTING PERSONAL INFORMATION IN FARSI CONTENT
CHAID USER SEGMENTATION ANALYSIS OF BEHAVIOURAL BIOMETRIC FEATURES FOR ONLINE PLATFORMS
AUTOMATIC TOOL FOR BEHAV ANALYSIS BASED ON STATE ANALYSIS
STRUCTURE FROM MOTION USING UNORDERED SETS OF IMAGES
DISCHARGE MEASUREMENTS USING THE PRESSURE- TIME METHOD: DIFFERENT EVALUATION PROCEDURES
RADIO COVERAGE ANALYSIS FOR MOBILE COMMUNICATION NETWORKS USING ICS TELECOM
PEDESTRIAN AND OBJECT DETECTION USING LEARNED CONVOLUTIONAL FILTERS
AUTOMATIC DIACRITIC RESTORATION FOR A TTS-BASED E-MAIL READER APPLICATION

Fig. 5. Article-article-type similarities (red edges).

Fig. 6 illustrates high semantic similarities between authors. The analysis starts from two authors and advances in depth for a maximum of three levels, with a branching factor of ten.

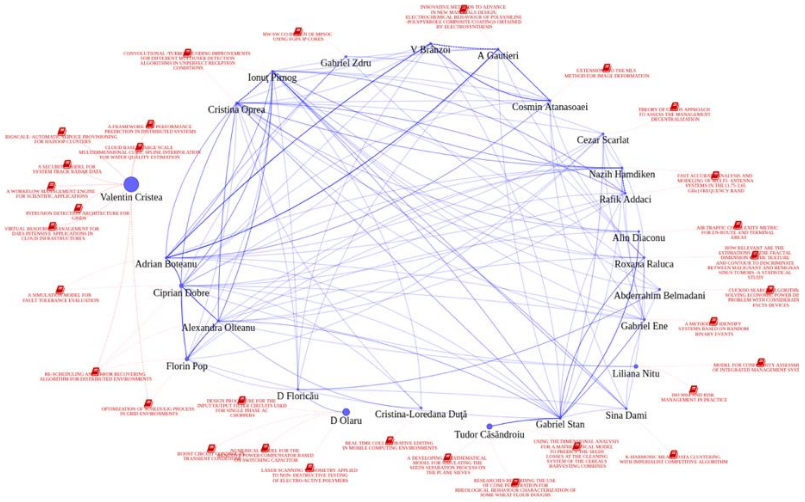

Fig. 6. Author-author-type similarities (blue edges).

The more articles the authors wrote together, the thicker the edge between them. This visualization is designed to support extensive follow-up analyses of author clusters.

Fig. 7 presents the relationship between articles and authors, using the top 20% of author-article-type edges and a breadth first search algorithm with a maximum depth of 3 and branching factor of 20. The visualization starts from one main author, in this particular case *Adina Magda Florea*. Larger author nodes denote an increasing number of publications. Similar to previous use cases, this visualization can help better understand the dynamic of research groups that published in the Scientific Bulletin of University Politehnica of Bucharest.

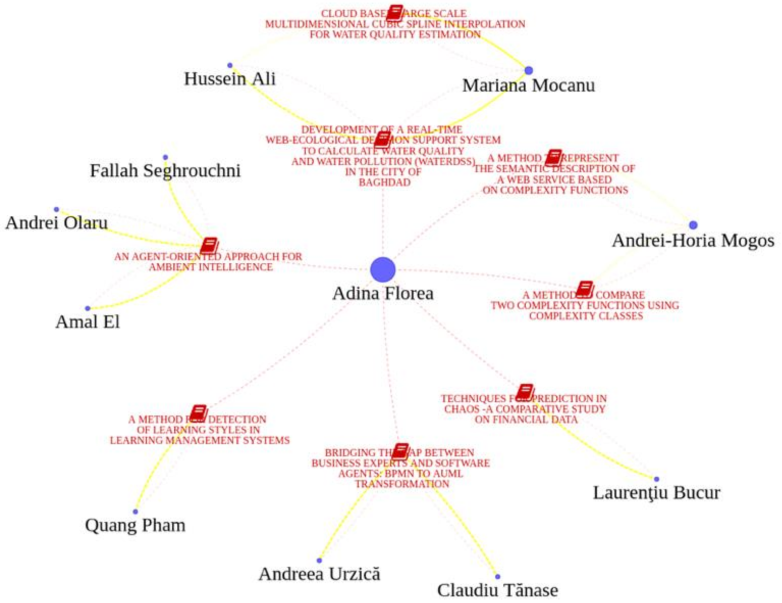

Fig. 7. Author-article-type similarities (yellow edges).

Fig. 8 illustrates the sub-graph generated by the top 25 authors and 25 articles according to their degree score in the 2-mode CNA graph.

These nodes have the highest semantic similarity score relative to the entire corpus. It is interesting to note that the articles are from different fields such as distributed programming, telecommunications and electronics, while the transition between the upper part of and the lower part of the graph is created by a multi-disciplinary article relevant to all the previously specified domains.

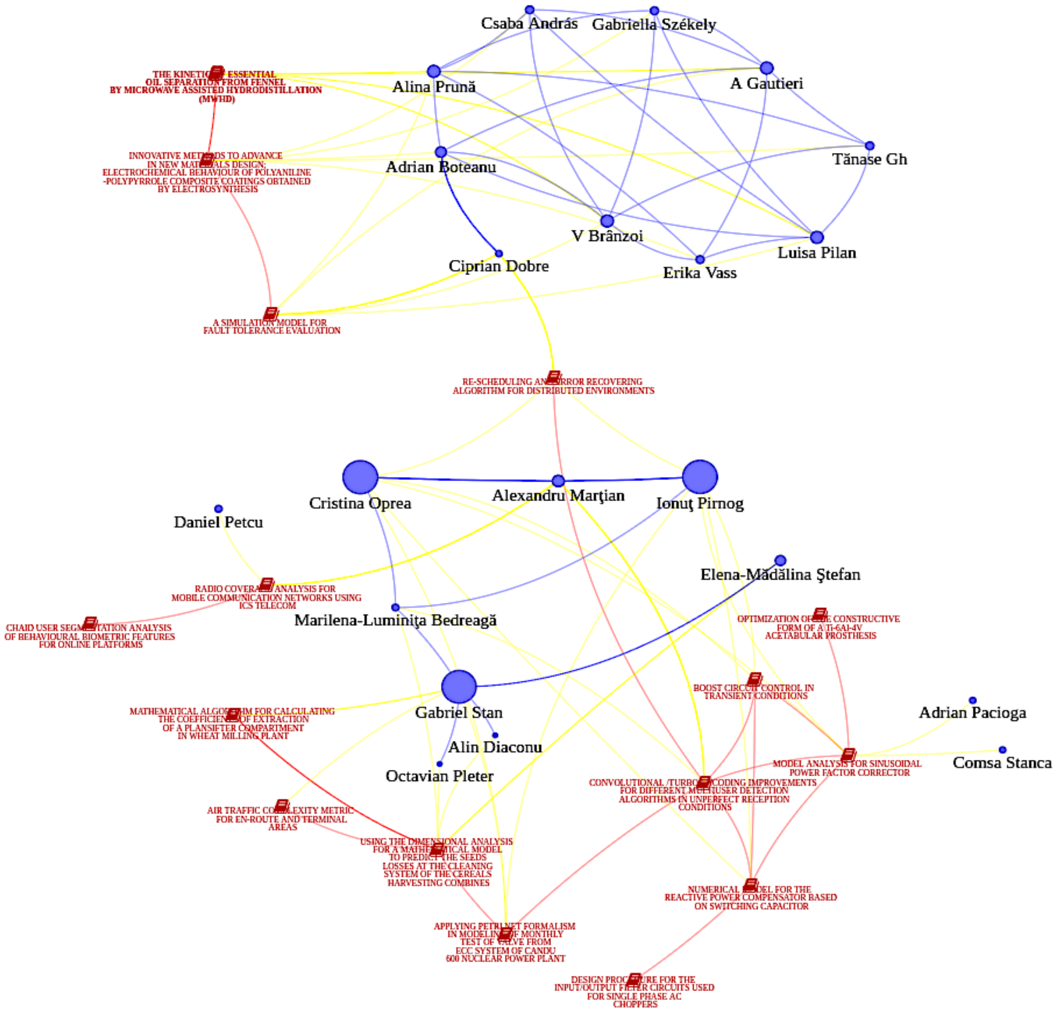
Csaba András
Gabriella Székely
THE KINETICS OF ESSENTIAL OIL SEPARATION FROM FENNEL BY MICROWAVE ASSISTED HYDRODISTILLATION (MWHD)
Alina Prună
A Gautieri
INNOVATIVE METHODS TO ADVANCE IN NEW MATERIALS DESIGN: ELECTROCHEMICAL BEHAVIOUR OF POLYANILINE -POLYPYRROLE COMPOSITE COATINGS OBTAINED BY ELECTROSYNTHESIS
Adrian Boteanu
Tănase Gh
V Brânzoi
Luisa Pilan
Ciprian Dobre
Erika Vass
A SIMULATION MODEL FOR FAULT TOLERANCE EVALUATION
RE-SCHEDULING AND ERROR RECOVERING ALGORITHM FOR DISTRIBUTED ENVIRONMENTS
Cristina Oprea
Alexandru Marţian
Ionuţ Pirnog
Daniel Petcu
Elena-Mădălina Ştefan
RADIO COVERAGE ANALYSIS FOR MOBILE COMMUNICATION NETWORKS USING ICS TELECOM
Marilena-Luminiţa Bedreagă
OPTIMIZATION OF THE CONSTRUCTIVE FORM OF A Ti-6Al-4V ACETABULAR PROSTHESIS
CHAID USER SEGMENTATION ANALYSIS OF BEHAVIOURAL BIOMETRIC FEATURES FOR ONLINE PLATFORMS
BOOST CIRCUIT CONTROL IN TRANSIENT CONDITIONS
MATHEMATICAL ALGORITHM FOR CALCULATING THE COEFFICIENT OF EXTRACTION OF A PLANSIFTER COMPARTMENT IN WHEAT MILLING PLANT
Gabriel Stan
Adrian Pacioga
Alin Diaconu
MODEL ANALYSIS FOR SINUSOIDAL POWER FACTOR CORRECTOR
Comsa Stanca
Octavian Pleter
CONVOLUTIONAL /TURBO CODING IMPROVEMENTS FOR DIFFERENT MULTIUSER DETECTION ALGORITHMS IN UNPERFECT RECEPTION CONDITIONS
AIR TRAFFIC COMPLEXITY METRIC FOR EN-ROUTE AND TERMINAL AREAS
USING THE DIMENSIONAL ANALYSIS FOR A MATHEMATICAL MODEL TO PREDICT THE SEEDS LOSSES AT THE CLEANING SYSTEM OF THE CEREALS HARVESTING COMBINES
NUMERICAL MODEL FOR THE REACTIVE POWER COMPENSATOR BASED ON SWITCHING CAPACITOR
APPLYING PETRI NET FORMALISM IN MODELING OF MONTHLY TEST OF VALVE FROM ECC SYSTEM OF CANDU 600 NUCLEAR POWER PLANT
DESIGN PROCEDURE FOR THE INPUT/OUTPUT FILTER CIRCUITS USED FOR SINGLE PHASE AC CHOPPERS

Fig. 8. Top authors and articles by degree.

4. Conclusions

This paper describes a comprehensive case study performed on the publications from the Scientific Bulletin of University Politehnica of Bucharest. Our system creates a 2-mode CNA graph containing both authors and articles, in which similarities between different authors are computed, leading to clusters of researchers working together on certain topics.

Our system can be used to detect important authors from a specific domain, or similar articles to a given subject, helping individuals from outside the community to find relevant authors and articles based on the semantic content of the articles. In addition, the system can be also extended towards becoming a benchmarking system to evaluate the impact and relations between different authors, publications, and affiliations.

Moreover, we aim to implement additional recommendation services targeted to support individuals in identifying key experts who fit more specific requirements and different filtering criteria (e.g., affiliation to a national research institute from a different city, or an European partner). We also strive to include analyses that depict the time evolution of a community in terms of global statistics, research group dynamics, as well as trending topics and their changes in time.

Acknowledgement

The work presented in this paper was funded by the European Funds of Regional Development with the Operation Productivity Program 2014-2020 Priority Axe 1, Action 1.2.1 D-2015, “Innovative Technology Hub based on Semantic Models and High Performance Computing” Contract no. 6/1 09/2016.

REFERENCES

[1]. *A. Rapoport, W. J. Horvath*, “A study of a large sociogram“, in Behavioral science, **vol. 6**, no. 4, 1961, pp. 279–291.

[2]. *M. Dascalu, D. S. McNamara, S. Trausan-Matu and L. K. Allen*, “Cohesion Network Analysis of CSCL Participation“, in Behavior Research Methods, **vol. 50**, no. 2, 2018, pp. 604–619.

[3]. *S. Wasserman, K. Faust*, Social Network Analysis: Methods and Applications, Cambridge University Press, Cambridge, UK, 1994.

[4]. *J. Scott*, Social network analysis, Sage, London, UK, 2017.

[5]. *I. C. Paraschiv, M. Dascalu, D. S. McNamara, S. Trausan-Matu and C. K. Banica*, “Exploring the LAK Dataset Using Cohesion Network Analysis“, in proceedings of the 3rd Workshop on Social Media and the Web of Linked Data (RUMOUR 2017), in conjunction with the Joint Conference on Digital Libraries (JCLD 2017), Toronto, Canada, "Alexandru Ioan Cuza" University Publishing House, pp. 17–21, 2017.

[6]. *M. Dascalu*, Analyzing discourse and text complexity for learning and collaborating, Studies in Computational Intelligence, Springer, Switzerland, 2014.

[7]. *M. Dascalu, P. Dessus, M. Bianco, S. Trausan-Matu and A. Nardy,* Mining texts, learner productions and strategies with ReaderBench, in Educational Data Mining: Applications and Trends, A. Peña-Ayala Ed. Springer, Cham, Switzerland, 345–377, 2014.

[8]. *N. Ide, J. Véronis,* Text encoding initiative: Background and contexts, Springer Science & Business Media, 1995.

[9]. *C. Gormley, Z. Tong*, Elasticsearch: The Definitive Guide, O'Reilly Media, Inc., 2015.

[10]. *I. C. Paraschiv, M. Dascalu, C. K. Banica and S. Trausan-Matu*, “Designing a Scalable Technology Hub for Researchers“, in proceedings of the 5th Int. Workshop on Semantic and Collaborative Technologies for the Web, in conjunction with the 13th Int. Conf. on eLearning and Software for Education (eLSE 2017), Bucharest, Romania, Advanced Distributed Learning Association, pp. 13–18, 2017.

[11]. *D. Corlatescu, I. C. Paraschiv, M. Dascalu, S. Trausan-Matu and C. K. Banica*, “Concurrent Processing of Scientific Articles using Cohesion Network Analysis“, in proceedings of the 17th Int. Conf. on Networking in Education and Research (RoEduNet), Cluj-Napoca, Romania, IEEE, 2018.