Knowledge Discovery on Scopus

Paolo Fornacciari, Monica Mordonini, Michele Nonelli, Laura Sani, and Michele Tomaiuolo

University of Parma, Italy

Abstract. Scopus is a well known repository of metadata about scientific research articles. In this work, we gather data from this repository to create a social graph of scientific authors, starting from citations among their articles. Moreover, using data mining techniques, we infer some relevant research topics for each author, from the textual analysis of the abstracts of his articles. As a case study, we have limited our research to the authors who have published at least one article about Sentiment Analysis, in a decade. Starting from the more relevant terms extracted from abstracts, we then perform a clusterization of users. This shows the emergence of some subtopics of Sentiment Analysis, which are studied by distinct groups of authors.

Keywords: Data mining; Social Networks; Clustering.

1 Introduction

Since data is being produced at increasing volumes, the computational challenge to extract useful knowledge from it is also becoming more and more important. But, having to deal with large quantity of data also means that much latent information can be extracted and leveraged for further analysis.

In this work, we deal with information gathered from Scopus, a repository of metadata about scientific research papers. In fact, the scientific production is also accelerating, as emerging countries show great ambitions about innovation and both new and established institutions use citation-based metrics for hiring researchers and managing career advancements.

In this growing quantity of data, in particular, we analyze the social graph of researchers and their research topics. For highlighting communities of researchers, in particular, we use a directed and weighted social graph, based on citations. A citation from an article to another corresponds to the creation of arcs among the respective authors. On the other hand, we are also interested in obtaining knowledge about the research topics in which each author works. For this purpose, we apply data mining techniques to the textual abstracts of the articles he/she has written. In fact, mining topics from abstracts provides more cue than explicitly assigned keywords, which are often used inconsistently.

The authors included in our research are those who have published at least one article about Sentiment Analysis, in a decade. A practical API is provided by Scopus, for performing such queries. We apply data mining techniques to

gather the most significant terms for each author, and then we use these terms as features to cluster them. As a result of our analysis, we notice the existence of some subtopics, which were not obviously foreseeable in advance and which distinguish different groups of authors.

The rest of the paper is organized as follows. The next section briefly discusses some background topics and related work. Then, Section 3 presents the methodological and practical aspects of the research. Section 4 shows the result of this research and some example of latent knowledge that can be extracted from this kind of data. The final section provides some concluding remarks, about the perspectives of this kind of research.

2 Related Work

Automatic Knowledge Discovery and Data Mining are becoming more and more important, since data are produced and collected at significant rate. One of the first research works about these topics is presented in [12]. Before the increase of Internet usage and the era of Big Data, the authors discussed the need for a new generation of computational theories and tools to assist humans in the extraction of useful information (knowledge) from the rapidly growing volumes of digital data. These methods are central to the field of Knowledge Discovery in Databases (KDD), which tries to make sense of data.

Big Data concerns large-volume, complex, growing data sets with multiple, autonomous sources. Wu et al. presented the features of the Big Data revolution and proposed a model to process them, from the data mining perspective [25]. The world of social networking and Web is one of the main source of Big Data and the field of the research in the application of data mining techniques to the World Wide Web is called Web mining [6].

A Social Networking Service (also Social Networking Site, SNS) is an online platform that is used by people to build social networks or social relations with other people who share similar personal interests, or activities, or real-life connections. The main features of SNSs are illustrated in [11, 15], while an introduction to the Social Network Analysis is provided in [24]. In [13], Sentiment Analysis (SA), carried out by a framework of Bayesian classifiers, is applied to a Twitter channel with the aim of detecting online communities characterized by the same mood. SA is one of the techniques for the extraction of opinion-oriented information, a comprehensive survey of Opinion Mining is presented in [20].

One of the most important problems in the field of social network analysis is community detection, aimed at clustering the nodes on the basis of their social relationships. The description of a new algorithm, named PaNDEMON, able to exploit the parallelism of modern architectures while preserving the quality of results, is reported in [1]. In [23], parallel computation is exploited to find relevant structures in complex systems, while in [14] the complexity is managed in a totally different way, by using a combined approach of sentiment analysis and social network analysis in order to detect communities that are effectively homogeneous within a network.

Participation in social networks has long been studied as a social phenomenon according to different theories [8, 7], and, in particular, the notion of social capital refers to the person's benefit due to his relations with other persons, including family, colleagues, friends and generic contacts. The role of social capital in the participation in online social networking activities is shown in [16] in the various cases of virtual organizations or virtual teams. Social networking systems are bringing a growing number of acquaintances online, both in the private and working spheres and, in businesses, several traditional information systems (such as CRMs and ERPs) have also been modified in order to include social aspects [2].

Scopus is an abstract and indexing database with full-text links, produced by the Elsevier Co. Scopus database provides access to articles and to the references included in those articles, allowing to search both forward and backward in time [5]. Scopus, together with Web of Science, represents one of the most authoritative bibliographic databases in order to perform bibliometric analyses and comparisons of countries or institutions [3].

Several works have been conducted in knowledge discovery on bibliographic databases. Newman constructed networks of scientists by using data from three bibliographic databases in biology, physics, and mathematics [19]. In these networks the nodes are scientists, and two scientists are connected if they have coauthored a paper. The author reports various analyses conducted over these networks, such as the number of papers written by each author, the number of coauthors, the typical distance between scientists through the network, and the variation of different patterns of collaboration between subjects and over time.

Rankings based on publications can supply useful data in a comprehensive assessment process of academic and industrial research. A framework for an automatic and versatile publications ranking for research institutions and scholars is shown in [21]. The authors demonstrate that the biggest difficulty in developing their framework was the insufficient availability of bibliographic data containing the institution with which an author was affiliated when the paper was published.

Using social network analysis to study scientific collaboration patterns has become an important research topic in information systems research. For example, research publications in the International Conference on Information Reuse and Integration provided the data in order to identify the most popular research topics, as well as the most productive researchers and organizations in that area [9]. The underlying idea is to treat scientific growth as a process of knowledge diffusion, so the most productive authors (or leaders) are often the ones who introduce new ideas and thus have a great impact on the research community.

An important type of social network is the co-authorship network, which has been studied both for social community extraction and social entity ranking. In general, to construct this kind of network we need to consider the co-authorship relation between two authors as a collaboration. Instead, in [17], the authors introduce a supportiveness measure on co-authorship networks, with quite interesting results. They model the collaboration between two authors as a supporting

activity between them, then they treat the supportiveness ranking problem as a reverse k nearest neighbor (k-RNN for short) searching problem on graphs.

In [10], the study of the scientific collaboration and endorsement is performed by using not only the co-authorship network but also the citation network. The results show that productive authors tend to directly coauthor with and closely cite colleagues sharing the same research interests; they do not generally collaborate directly with colleagues having different research topics, but they directly or indirectly cite them; highly cited authors do not generally coauthor with each other, but closely cite each other.

3 Methodology

Scopus, which is commonly accessed through its web interface, is a bibliographic database containing abstracts and citations for scientific articles[1]. It also contains useful pieces of information about the authors and the institutions to which these authors are affiliated (universities, companies, research centers, ...). The records stored inside the database can be categorized into three kinds of "information elements": abstracts, authors and affiliations. These categories can be seen as three different classes of objects, each one with its specific sub-attributes (i.e. an author has an ID, a name, a surname, one or more affiliations, and so on).

To programmatically access the data inside Scopus, Elsevier (the company running the database) also developed a set of HTTP RESTful APIs, whose up-to-date specifications can be found on the Elsevier developers portal[2]. At the time of writing this article, there are 11 Scopus APIs publicly available. Among these, for the sake of our research, just 6 have been used: ScopusSearch, AbstractRetrieval, AuthorSearch, AuthorRetrieval, AffiliationSearch and AffiliationRetrieval. Given the names of those APIs, it is easy to see that for each class of information elements one can find two kind of APIs: one that can be used to search among elements of a specific type, and another one that can be used to retrieve a specific object, given its ID. Each API has its specific query variables and conventions, which are described inside the already linked Elsevier developers website. The flagship API, named ScopusSearch, searches among the abstract objects.

To manage the HTTP requests and join the JSON responses coming from the server into bigger datasets, we have built a basic Python library wrapping the 6 main Scopus APIs and a set of Python scripts and Jupyter Notebooks (IPython Notebooks) which can be executed in series to automate the 5 steps of the Knowledge Discovery in Databases (KDD) pipeline. Outside of the experiment here presented, the library and the scripts can be used (and have been used during our tests) to query the Scopus APIs with any of the available parameters and search keys, and in particular to reproduce the whole knowledge discovery process described in the next steps. As said, the code needed to repeat this

[1] `http://www.scopus.com/`
[2] `http://dev.elsevier.com/`

experiment or try other queries is written in Python and can be downloaded from a public repository[3].

This knowledge discovery experiment is conducted applying the 5-steps KDD pipeline (Figure 1) on the Scopus database. As usual, in addition to the 5 KDD steps, this involves also a 0-step in which the domain (in this case, the Scopus systems) is analyzed in order to understand what kind of data is available and how to access it.

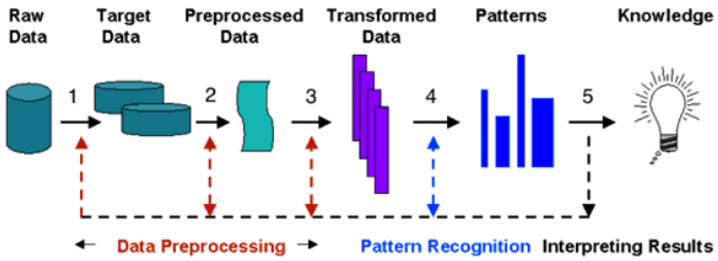

Fig. 1. KDD pipeline.

3.1 KDD Step 1: Selection

The first step of the KDD process, used to extract information, involves selecting a more specific subset of data from the whole database. Applying this on Scopus, we have decided to download responses from the ScopusSearch API for a specific query.

This API replies to HTTP GET requests with JSON (or XML) payloads containing a list of Abstracts matching the input search query, up to the limit of 5000 results per single query. Search results can be downloaded from the API in lists of only 25 abstracts per HTTP request, and must then be joined and saved in the file system for the next steps.

This process has been automated and the requests and responses are managed by the first step of the pipeline. In the experiment, this is used to download from the ScopusSearch API the data matching the query "TITLE-ABS-KEY({sentiment analysis})", which searches for all the abstracts having the exact whole sentence "sentiment analysis" in their "title", "abstract" or "keywords" attributes. The query produces 3648 results.

While doing a very first cleanup of the data, dropping invalid or corrupted abstracts, the JSON responses are joined creating a JSON array of abstracts. Each of these objects contains not only the text of the abstract and other data related to the article, but also some pieces of information about its authors

[3] `http://www.github.com/valleymanbs/scopus-json/`

and their affiliations. A very first cleaning of the results, dropping invalid or corrupted data, produces the Target Data on which the second step of the KDD process is run.

3.2 KDD Step 2: Preprocessing

Matching the second step of the classic KDD pipeline, the Target Data is preprocessed, doing a deeper cleanup of useless information and integrating the missing data, running the appropriate requests to the AuthorRetrieval and the AffiliationRetrieval APIs.

After this first cleanup and integration process, a search is performed on the already downloaded articles, in order to obtain the ones with citations. A set of queries is then run, again on the ScopusSearch API, to download all the articles citing an article which is already inside the dataset. The newly download data are joined into the original dataset, after being cleaned and integrated.

Using the "TITLE-ABS-KEY({sentiment analysis})" dataset, which is downloaded in the first step of the experiment, a final CSV dataset (10307 abstracts) is produced, which also contains useful data related to their authors. This last thing is particularly important, since the next steps will focus on the authors of articles.

3.3 KDD Step 3: Transformation

Going further in the process of extracting knowledge from data, two different kinds of transformation processes are applied to the Preprocessed Data obtained from step 2, both supported by the code included in the project repository.

The first transformation stage creates a graph from the previously preprocessed abstracts dataset: the entries of this dataset represent a social network, in which a node represents an author and the directed edges represent the citations between the authors of the articles which formed the dataset coming from the previous step. The graphs produced in the pipeline are formatted following Gephi standards, and will be used in future works in order to analyze the social networks between authors of the same research field. Gephi itself is a well known open source software for graph analysis[4].

The second kind of transformation consists in the extraction of all the authors data from the preprocessed dataset of abstracts. This approach is extended by grouping, under each author, all the text found in the abstracts of his/her articles. At an intermediate stage, this transformation creates a dataset (Table 1) in which each line relates to a single author with all the text contained in the titles, abstracts and keywords attributes found inside the abstracts information elements downloaded from the Scopus APIs in the first 2 steps of the KDD pipeline.

This authors-text dataset is rich in unstructured information and is then processed to obtain structured numerical data for the data mining step, going

[4] `http://gephi.org/`

	author_id	text
0	10041296000	Media-aware quantitative trading based on publ...
1	10042379000	A context-dependent sentiment analysis of onli...
2	10043514200	The networked cultural diffusion of Korean wav...

Table 1. CSV dataset.

Stemming example	words
media-awar	media-aware
quantit	quantitative
trade	trading

Table 2. Stemming example.

through a TF-IDF vectorization process. TF-IDF is a bag of words technique to vectorize text data in which a vocabulary is extracted from all the words contained in a Corpus of several text Documents (after a stemming process). Each term of the corpus vocabulary serves as an index of the vector that represents the unstructured text document as a structured indexed numerical vector. The weights inserted in each vector are calculated weighting each single word frequency inside the corresponding document against the term frequency inside all the documents in the corpus, following the formula

$$w_{i,j} = tf_{i,j} \times log\left(\frac{N}{df_i}\right) \tag{1}$$

where $tf_{i,j}$ is the number of occurrences of i in j, df is the number of documents containing i, and N is the total number of documents.

In our particular vectorization process the whole sum of text contained in the dataset makes up the corpus, while all the articles associated to an author are a single document inside this corpus; from this unstructured text data has been extracted a vocabulary vector made of the 500 most frequent terms inside the whole corpus, counting 27956 unique words, stemmed as in Table 2.

The TF-IDF vectorization is the last stage of the transformation process and finally produces the Transformed Data on which the KDD process will go on: a matrix of numerical data in which each line is a vector representing an author.

In other words, from the TF-IDF vectorization has been obtained a set of vectors in a 500-dimensional features space: each unique author found in the data originally downloaded from Scopus is represented in this 500D Vector Space Model by a vector which weights have been computed using the frequency of words found in the text written by the corresponding author, compared against the same words frequency inside all the text written by all the authors in the dataset.

Cluster	Authors	Terms
BLUE ECommerce	1082	reviews, product, features, aspects, users, mining, rating, modeling, customers, online
RED Algorithms	3376	classification, word, features, modeling, text, emotional, learning, language, polarized, lexicons
GREEN Social Media	3388	social, media, twitter, users, network, tweeting, mining, systems, predict, topic

Table 3. KMeans - Top terms per cluster.

3.4 KDD Step 4: Data Mining

In this step the Transformed Data obtained from step 3 has been analyzed with an unsupervised learning approach, in order to see if analyzing text data, there are emerging clusters of authors in this field of research.

To verify this, the vectors coming from the TF-IDF weighting have been clustered using K-Means and Expectation Maximization. This two methods have been used to check the results of a clusterer against the other.

Clustering with K-Means and E-M requires to set the number K of clusters as parameter: after testing several K values, in the experiment we set K=3. This value has been experimentally found to be the one giving stable and comparable results between the two clustering algorithms. Along with the experimental approach, this value has also been found to be the one giving the higher Silhouette Score [22] when running a Latent Semantic Analysis on the same dataset with a set of features reduced to 3 dimensions with PCA [18].

3.5 KDD Step 5: Knowledge

The unsupervised learning process, used to extract information from text, eventually leads latent information to emerge. This paves the way for the identification of the main topics treated by the authors in the chosen research field. Results obtained for the Sentiment Analysis field are presented in next section.

4 Experimental results

After the clustering process, the most frequent words, according to TF-IDF weights, used by all the authors of each one of the obtained clusters are analyzed. Two clustering algorithms are used and confronted. The aim is to understand which topics are treated by the clustered authors and to find relationships between clusters produced respectively by K-Means (Table 3) and E-M (Table 4).

The previous step of unsupervised learning process is used to extract information from text. The obtained data leads to identifying three main clusters and topics treated by the authors in the "Sentiment Analysis" research field (Figure 2, 3). According to the highlighted terms, we have labeled these

Cluster	Authors	Terms
BLUE ECommerce	1058	reviews, product, features, aspects, mining, users, modeling, customers, rating, online
RED Algorithms	4661	modeling, emotional, classification, text, features, word, learning, language, systems, polarized
GREEN Social Media	2127	social, media, twitter, network, users, tweeting, mining, events, predict, topic

Table 4. E-M - Top terms per cluster.

three subtopics (corresponding to the obtained clusters) respectively as: "ECommerce", "Algorithms", "Social media".

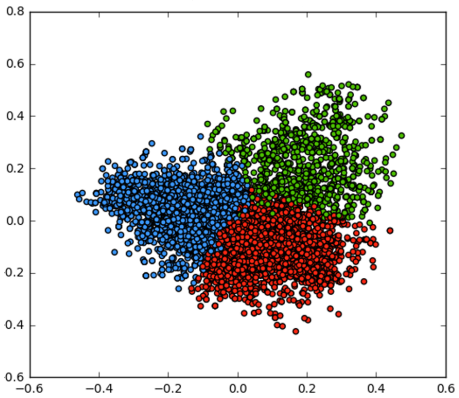

Fig. 2. Clustering results with KMeans (2D representation with PCA).

However, as can be seen from the clustering output and from the occurrence of the same top-words inside the different clusters, the clustering process conducted both with K-Means and E-M tends to be weak. This is probably due to the topic searched, which is very tight, with low variance between authors (new field, few articles).

Finally, we have confronted the results of the clusterization process, by terms, with the communities of authors, obtained by the social graph topology [4]. In particular, communities have been identified by the density of weighted and directed social arcs, obtained from citations among articles. It is apparent, quite at first sight, that clusters are spread all over the social graph (see Figure 4). In fact, community detection can distinguish three communities, with sizes comparable to those of content-based clusters. However, in each community, clusters are represented with almost the same incidence as in the whole graph. The same

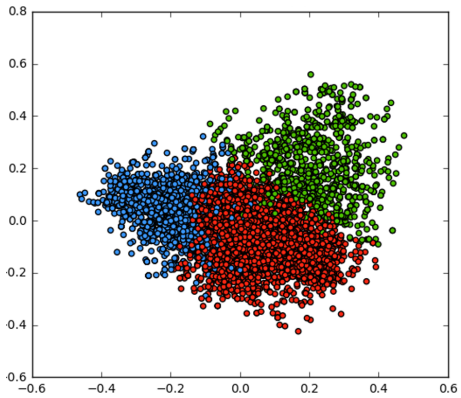

Fig. 3. Clustering results with E-M (2D representation with PCA)

kind of results are obtained for different runs of the community detection process, generating more and smaller communities, where clusters are nevertheless all well present.

This means that, in this case, the subtopics of research and the citation graph represent mostly orthogonal and independent aspects of analysis. For the topic of "Sentiment Analysis", in fact, the social graph of authors has a large and highly connected core, with low modularity. Thus, the results indicate that authors often cite other authors, working in different communities and different subtopics.

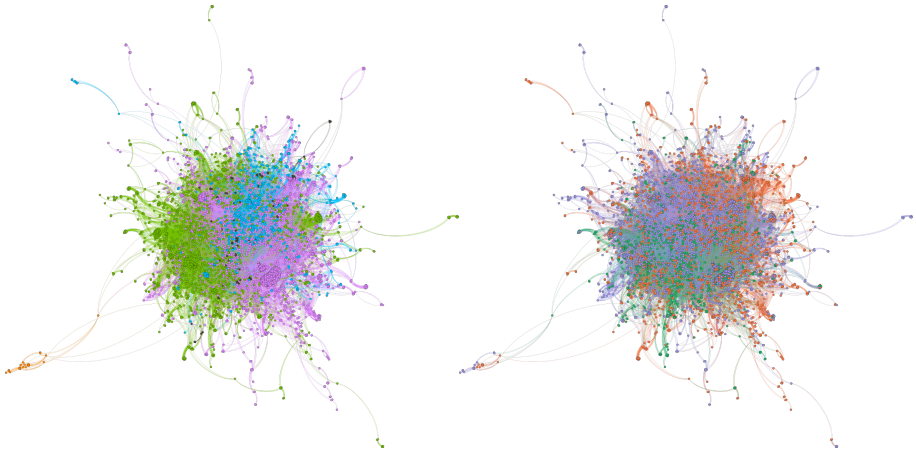

Fig. 4. Social graphs colored by citation-based communities (left) and term-based clusters (right). The same layout, ForceAtlas2, is applied in both cases.

5 Conclusion

The availability of large quantity of data often allows analysts to discover latent knowledge. We applied the techniques of social network analysis, data mining and clustering to data about scientific research articles. In particular, we used metadata available in the Scopus repository, to create a social graph of scientific authors, starting from citations among their articles. Moreover, using data mining techniques, we have inferred some relevant research topics for each author, from the textual analysis of the abstracts of his articles.

In particular, we have limited our analysis to the case study of authors who have published at least one article about "Sentiment Analysis", in a decade. After associating each author with the most relevant topics emerging from the text analysis of his abstracts, we have performed a clusterization of authors. This process has finally brought to light some groups of authors, who conduct their research about distinct application areas of Sentiment Analysis, namely: algorithms for sentiment analysis; marketing and e-commerce; social media analysis. Contrasting the results of the clustering process, by relevant terms, and the community detection process, based on the social graph of citations, indicates that authors often cite other authors, working in different communities and different subtopics.

References

1. Amoretti, M., Ferrari, A., Fornacciari, P., Mordonini, M., Rosi, F., Tomaiuolo, M.: Local-first algorithms for community detection. In: 2nd International Workshop on Knowledge Discovery on the WEB, KDWeb (2016)
2. Angiani, G., Fornacciari, P., Mordonini, M., Tomaiuolo, M.: Models of participation in social networks. In: Social Media Performance Evaluation and Success Measurements, p. 196. IGI Global (2016)
3. Archambault, É., Campbell, D., Gingras, Y., Larivière, V.: Comparing bibliometric statistics obtained from the web of science and scopus. Journal of the Association for Information Science and Technology 60(7), 1320–1326 (2009)
4. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment 2008(10), P10008 (2008)
5. Burnham, J.F.: Scopus database: a review. Biomedical digital libraries 3(1), 1 (2006)
6. Cooley, R., Mobasher, B., Srivastava, J.: Web mining: Information and pattern discovery on the world wide web. In: Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on. pp. 558–567. IEEE (1997)
7. Cristani, M., Fogoroasi, D., Tomazzoli, C.: Measuring homophily. In: CEUR Workshop Proceedings. vol. 1748 (2016), `https://www.scopus.com/inward/record.uri?eid=2-s2.0-85012298603&partnerID=40&md5=81df100456c2118853ca823496097c79`
8. Cristani, M., Tomazzoli, C., Olivieri, F.: Semantic social network analysis foresees message flows. In: ICAART 2016 - Proceedings of the 8th International Conference on Agents and Artificial Intelligence. vol. 1,

pp. 296–303 (2016), https://www.scopus.com/inward/record.uri?eid=2-s2.0-84969287486&partnerID=40&md5=6d7a0bb42fd4f45cdb48b8dc1193907a
9. Day, M.Y., Ong, C.S., Hsu, W.L.: An analysis of research on information reuse and integration. In: Information Reuse & Integration, 2009. IRI'09. IEEE International Conference on. pp. 188–193. IEEE (2009)
10. Ding, Y.: Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. Journal of informetrics 5(1), 187–203 (2011)
11. Ellison, N.B., et al.: Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication 13(1), 210–230 (2007)
12. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI magazine 17(3), 37 (1996)
13. Fornacciari, P., Mordonini, M., Tomaiuolo, M.: A case-study for sentiment analysis on twitter. In: WOA. pp. 53–58 (2015)
14. Fornacciari, P., Mordonini, M., Tomauiolo, M.: Social network and sentiment analysis on twitter: towards a combined approach. In: 1st International Workshop on Knowledge Discovery on the WEB, KDWeb (2015)
15. Franchi, E., Poggi, A., Tomaiuolo, M.: Blogracy: A peer-to-peer social network. International Journal of Distributed Systems and Technologies (IJDST) 7(2), 37–56 (2016)
16. Franchi, E., Poggi, A., Tomaiuolo, M.: Social media for online collaboration in firms and organizations. International Journal of Information System Modeling and Design (IJISMD) 7(1), 18–31 (2016)
17. Han, Y., Zhou, B., Pei, J., Jia, Y.: Understanding importance of collaborations in co-authorship networks: A supportiveness analysis approach. In: Proceedings of the 2009 SIAM International Conference on Data Mining. pp. 1112–1123. SIAM (2009)
18. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. pp. 50–57. ACM (1999)
19. Newman, M.E.: Coauthorship networks and patterns of scientific collaboration. Proceedings of the national academy of sciences 101(suppl 1), 5200–5205 (2004)
20. Pang, B., Lee, L., et al.: Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval 2(1–2), 1–135 (2008)
21. Ren, J., Taylor, R.N.: Automatic and versatile publications ranking for research institutions and scholars. Communications of the ACM 50(6), 81–85 (2007)
22. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics 20, 53–65 (1987)
23. Sani, L., Amoretti, M., Vicari, E., Mordonini, M., Pecori, R., Roli, A., Villani, M., Cagnoni, S., Serra, R.: Efficient search of relevant structures in complex systems. In: AI*IA 2016 Advances in Artificial Intelligence. pp. 35–48. Springer (2016)
24. Scott, J.: Social network analysis. Sage (2012)
25. Wu, X., Zhu, X., Wu, G.Q., Ding, W.: Data mining with big data. ieee transactions on knowledge and data engineering 26(1), 97–107 (2014)