

## **Classification Methodology Of Research Topics Based In Decision Trees: J48 And Randomtree**

**<sup>1</sup>Katherine Andrea Cuartas C<sup>2</sup>John Petearson Anzola A and <sup>3</sup>Giovanny  
Mauricio Tarazona B**

*1 Currently Student of Industrial Engineering.*

*Universidad Distrital Francisco José de Caldas, Bogotá Colombia*

*Email: kacuartasc@correo.udistrital.edu.co*

*2 Currently Student of Doctoral Engineering.*

*Universidad Distrital Francisco José de Caldas, Bogotá Colombia*

*Email: jpanzola@udistrital.edu.co*

*3 Professor Faculty of Engineering, Industrial Engineering*

*Universidad Distrital Francisco José de Caldas, Bogotá Colombia*

*Email: gtarazona@udistrital.edu.co*

### **Abstract**

In the scope of the investigation, one of the problematics is found the hole or gap in the border of the knowledge in which can contribute or grab knowledge of a thematic and give it the character of novelty. Is in that point, when the technique of analytical bibliometric is used to explore the condition of a topic in different academic data bases, in function of the reach of the results returned in those data bases. For a researcher exploring more than 10.000 thematic results, is complex. In this paper is implemented the J48 and RandomTree classifiers, that are techniques used in data mining, whose goal is find research thematics with low index of publication in a way that under the experience of the researcher could select thematics that guide him to publish in high impact journals give in the characteristic of novelty

**Key Words:** Data Mining, bibliometría, J48, Random Tree, WEKA.

### **INTRODUCTION**

The contained information, focused in the divulgation of research in data bases, currently is incredibly large. For august 2012 had 28.100 academic journals with an anual collective production of 1.7 to 1.8 millions of articles. The CrossRef data bases includes more than 56 millions of DOI registers, in which 46 millions are about

journal articles. An important subset is conformed by 10.675 journals included in Thomson Reuter Journal Citation that publish 1 million of articles per year approximately. Another important data base is Scopus, that covers 18.500 journals. The number of reviewed by peers and annually published journals has increasing around 3.5% per year and the number of articles also increased in an annual rate of 3%. One of the reasons of this growth is the increase of scientific researchers in the world.[1].

One of the problematics that the researchers confront is the selection of specific thematics of investigation, specially when the researcher confront an exorbitant quantity of information which is difficult to have full access. One of the tools that allows approach this problematic is Data Mining, allows extract information from a big dataset and with an expert judgment, give it value to the obtained results. Data Mining is used in this work with the goal of find a novelty clear field in investigation thematics, achieving the identification and classification of key words in a particular thematic. In this article is proposed the classification algorithms J48 and Random.Tree as identification and classification techniques of novelty investigation thematics and as support of the bibliometric systems. With the use of this techniques has the possibility of find related and unexplored investigation topics, whose study can enlarge the field of knowledge in a particular thematic.

Bibliometric is responsible on study using statistics the content of data bases, however, taking into account the increase in the publications of articles per year, is necessary contribute a classification of scientific production methodology, which not only give consultation dates and advanced filters per words, but also have the capability of classify this information in an intelligent way, give in the researcher the possibility of a vision spectrum of research thematics in a temporal window ( years in that the information has been extracted), this methodology also allows visualize the most relevant, worked and unexplored topics.

The propose of this paper is structured in the following way: in first place is made a background, in which synthesized the application approaches more extended in data mining and text mining, citing some models that in an indirect way has made some contribution to bibliometric, stand out the classification algorithms which are worked in this paper, next, expose j48 and Random-tree algorithms continuing with the methodology and the results. taking as thematic axis and study cases the Internet Of Things (IOT).

## **1. BACKGROUND**

Bibliometric is a set of methods for quantitative analyzing from scientist and technique literature. As such, when is used in the evaluation of investigation, the goal of the bibliometric indicators is quantify the vision and/or impact of a big investigation groups set, materialized in divulgation articles.

The ideal process for do that evaluation will consist in realize a consultation of a topic, read the article and clasificate it, but in the case that have 10.000 results could apply filters and reduce the consultation to a 2000 results. Read and classify this information volume, even if is relatively low, compared with another systems, is a

labor that results very complex. For that has been used the Data Mining in this process and specially the classification trees J48 and Random Tree.

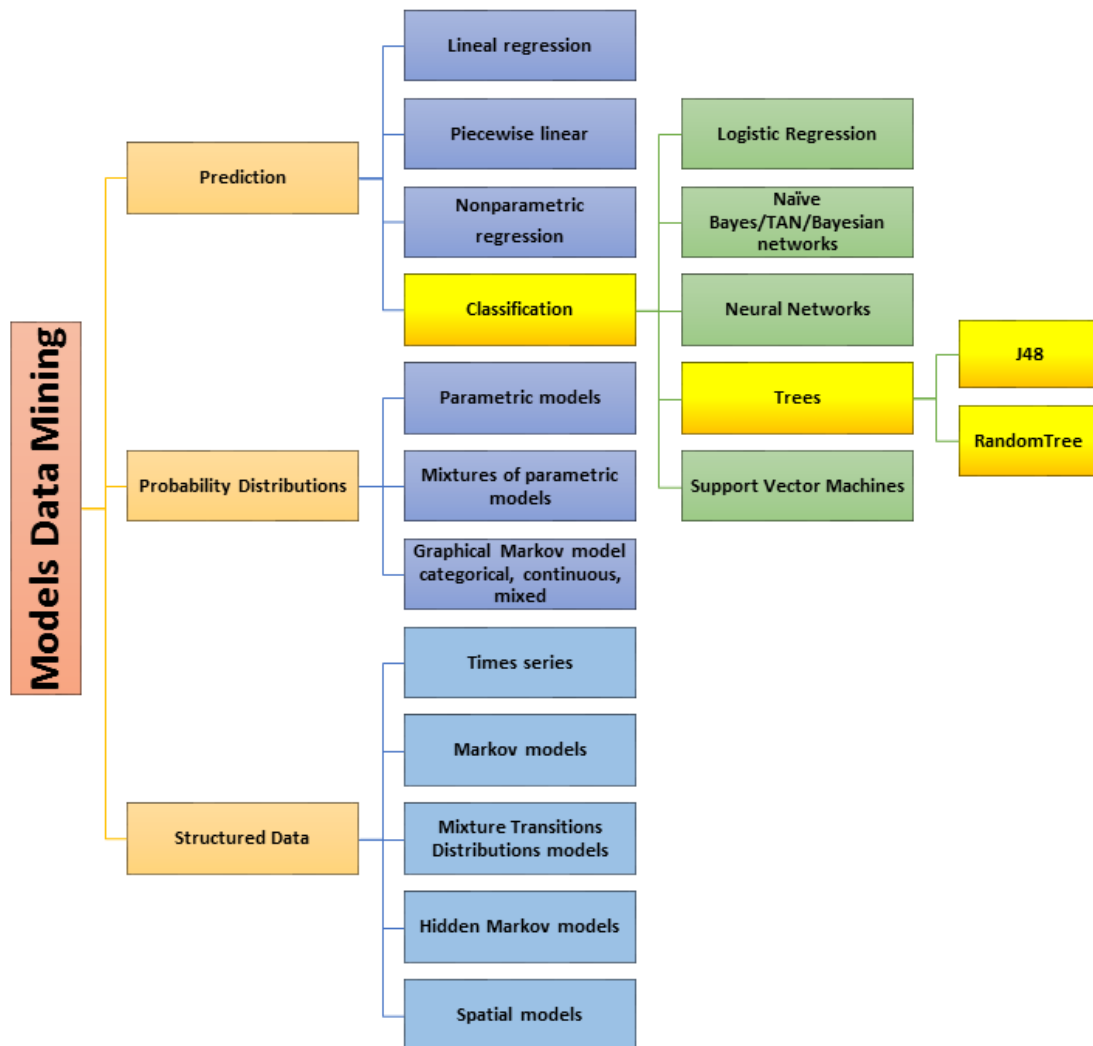


Figure 1. Models Applied to Data Mining

Exist several approaches raised for the knowledge extraction, one of them consist in consider text mining as a similar technique of data mining, but with the difference that the goal of text mining is extract information from non structured data texts [2], using techniques of natural lenguaje prosecution [3], combining visualization of information techniques [4].

Since the point of view of bibliometric has been raised studies to non structured dates through information extraction techniques [5], following the investigation topic [6],summary analysis[7], articles categorization [8], grouping [9],information visualization [10], and consulting methods[11].

In other side, has been employed a combination of text analysis techniques in function of his goals, for example, Nasukawa y Nagano proposed a text mining system know as como Text Analysis and Knowledge Mining (TAKMI)[12], through an interactive function that allows easily confirm the analysis results in an original document, being as the estadistic analysis tend to ignore the least important patterns, therefore, the use of this technique allows to the users find not just the principal patterns, but also the least important patterns, with the goal of found new knowledge, using the computing capability of a team for handle great amounts of data and the capability of human being for notice subtle differences in essential patterns.

Others utilization approaches of Data Mining, are defined as access and recovery of information, in function of computational linguistics, which are based in a corpus, that supports, the search of new information is a fundamental part of text mining, but establishes that the information, the recuperation and data mining don't contribute to the Discovery of new knowledge for the following reasons:

- (1) The documents returned for the recuperation systems, don't contains new information because the returned information is already know for the document authors
- (2) The applications of data mining help to find trends and patterns in an automatic way since large dataset

Are numerous the Works and contributions made since text mining, data mining, and the inter-disciplinary and isophormism used in different disciplines, find combinations of bibliometric techniques. The work raised used a classification methodology for decision trees.

## **2. DECISION TREES**

Are classification algorithms that has the propose of create knowledge structured that guide the process of taking decisions, this is based in the search of most relevant attributes inside the information and the classification of those divided then in homogeneous subsets that shows as result in a graphic way and in a group of rules that can be expressed inside the natural language if-then [13].

A decision trees shows the characteristic of be easily comprehensible, give it a class value to the more important data turning them in nodes, creating a hierarchical structure that allows made an interpretation of data having in count his relevance [14].

The decision trees to employed in the proposed of this article are:

- J-48: the decision trees compounds for J48 o C4.5 algorithms (developed by Ross Quinlan), are used principal for decision or classification proceedings. they are constructed of a tagged training data set, from which is creating a attributes list that divides in subsets based in entropy information, depends from the view point since the researcher wants work [15].

- RandomTree: Is a tree formed for a stochastic process, that includes an Uniforme Spanning Tree model, in which the Random part of the tree expands in to the minimum, creating a random binary and recursive tree. The branching process is quickly explored through a brownian process, characteristic that is considered as a vector of entry to the tree and generates the root class tag "weight". The mistakes of training are calculated through cross validation or boot.[16]. The exact estimation of the processing isn't necessary for the construction process of the tree, because in the moment of training the mistakes are calculated internally. The construction of the tree is made with all the data in a training phase, the validation phase of data or classification, is not made because of the construction tree and the way in that is organized is the characteristic that is used in the results that are exposed in this article.

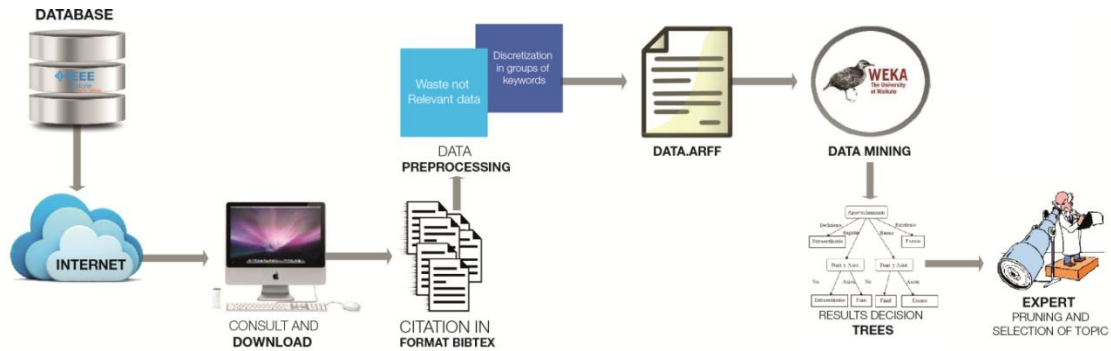
In the following chart are evident the characteristics of each tree:

**Tabla 1. J48 and RandomTree characteristics**

	<b>J48</b>	<b>Random Tree</b>
Create By	Ross Quinlan	Leo Breiman and Adele Cutler:
Starts from	Trainig data	Training data
Metodología	1. analyses an attribute list 2. Divide the information in subset 3. Identifies the attribute with most gain of information and recognized it like : decision parameter 4. classify the information according the decision parameter	1. obtain a prediction for each vector 2. branching through a brownian process, characteristic that considers as a entry vector from de tree and generates the root class tag "weight" 3. calculate the estimate from mistake of classification
Criterio	Entropy difference	Stochastic process

### 3. METHODOLOGY

The thematic axis selected is Internet of Things (IoT), the data mining is performed through the WEKA software and the raised methodology is shown in the following picture:



**Figure 2. Applied methodology**

The methodological model that is raised consists in a consulting process of data bases, for the treaty case of this article, the data base where the information was extracted is IEEE Xplore, in which was extracted 4.929 registers from articles of Journals and proceedings, downloading his respective citation format Bibtex.

Once download the citations is performed a pre –prosecution of data, whose consist in remove non relevant information, as : author names, journal Volume, pages, inther others, preserving only the following files: article tipe, abstract, and the principale topics that shape the article, in key words, with the goal of offer to the researcher a filtered and detailed consultation, in that case one article can contained more than 30 key words.

After of have the necessary information procedes to discretize the key words, this through the similar words grouping to avoid the data duplicity and is assigned a weighting to each Word depending of the order of this inside the text.

Once procesed tghe information is performed the conversion to arff format tipe, that is the read format by the free software to use Weka, in this the infomation is clasificate with two decision trees el J48 and RandomTree.

#### 4. RESULTS

The classification trees obtained through the data mining process in weka shows in the followings figures:

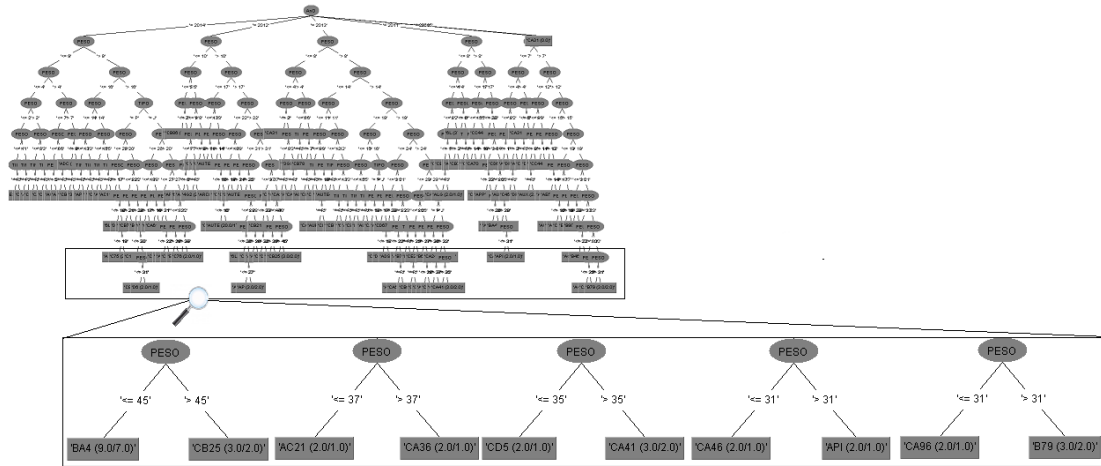


Figure3. J48 Tree Results

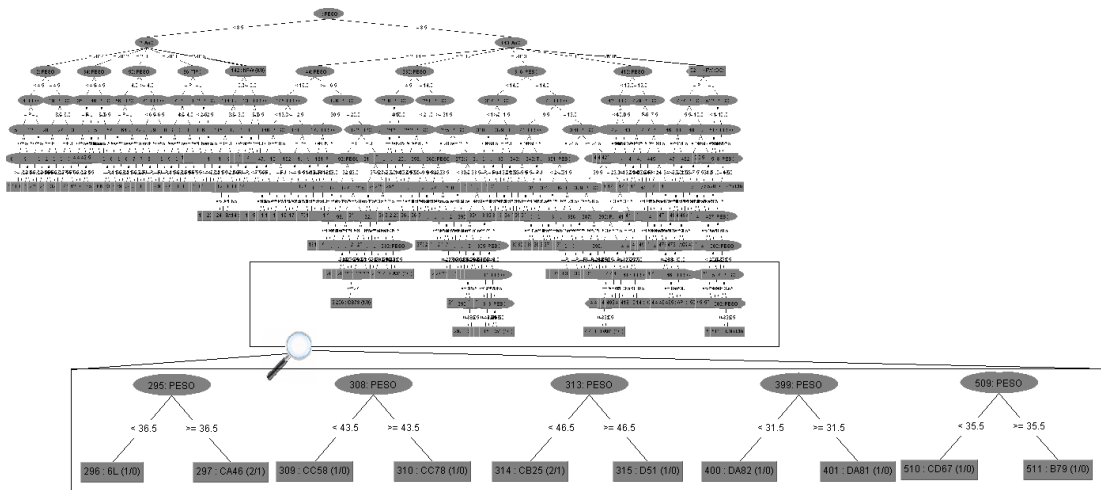


Figure 4. RandomTree Results

The classification tree J48 has an structure of 10 levels, the braches are divided first for year, continued by the weight and aricle tipe, while the classification tree Random trees shows and estrutura of 12 levels and shows brachings dived firstfor weight then for years and finally for the article tipe. For both previous cases the interesd focus corresponds to the last levels.becasuse those represents the lest frecuently group of words, and for that is and avoid of the lest explored topics in a temporal window

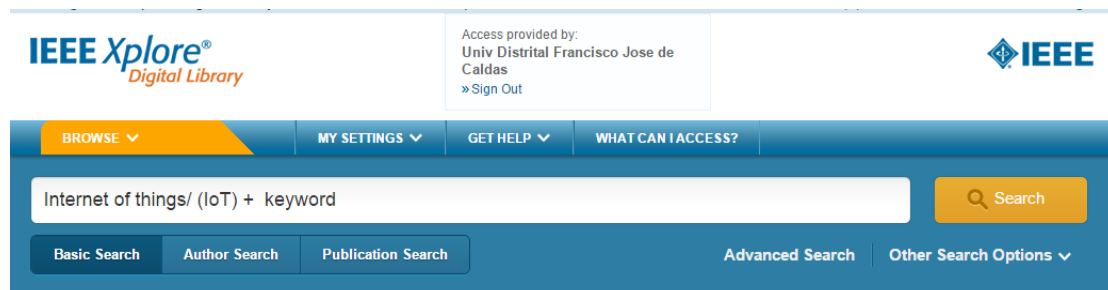
Making a pruning of this interest áreas represented in the decision trees applied in theInternet of things (IoT) case, obtain the following results:

**Table 2. Selected keywords**

	<b>Pruning Groups(keys words selected )</b>
Set 1: Algorithm J48	<b>Community-based Internet Access</b> , Actuators, cloud computing-empowered prototyping system, Delay, Cloud Governance, <b>Constrained Application Protocol (CoAP)</b> ,Application Programming Interface (API), Communication cables, <b>bridges (structures)</b>
Set2: AlgorithmRandomTree	6LoWPAN, <b>Constrained Application Protocol (CoAP)</b> ,Context middleware, Context-aware Web, <b>community-based Internet access</b> , digital service innovation, Dual-Level, dual-band PA, cryptography, <b>bridges (structures)</b>

A total of 5433 key words groups are obtain as result 9 words through J48 tree and10 words for RandomTree,this means a reduction of 99,84% and99,81%, respectively.between the resultants groups of the differents trees, find 3 common groups **Constrained Application Protocol (CoAP)**, **community-based Internet, bridges (structures)**, this performed aproximately the 30% of the total of words obtain by the tree, the other words differ, this because the classification criterion of each tree is defferent.

For corroborate the results obtained and has in count that the central topic wasInternet of Things (IoT), procedes to perferod a research in the following way:



**Figura5. Search IEEE Xplore.**



**Figura6. Search ScienceDirect**

The result of this search are show in the following chart :

**Table 3. Selected search words.**

	Key words	Number of Results IEEE Xplore	Number of results ScienceDirect
J48	<b>Community-based Internet Access</b>	1	0
	Actuators	4	0
	Cloud computing-empowered prototyping system	1	0
	Delay	5	1
	Cloud Governance	1	1
	<b>Constrained Application Protocol (CoAP)</b>	9	0
	Application Programming Interface (API)	4	0
	Communication cables	2	0
	<b>Bridges (structures)</b>	2	0
RandomTree	6LoWPAN	7	2
	<b>Constrained Application Protocol (CoAP)</b>	9	0
	Context middleware	2	0
	Contextaware Web	2	1
	<b>Community Based Internet access</b>	1	0
	Digital Service Innovation	1	0
	Dual-Level	1	0
	dual-band PA	1	0
	Cryptography	5	1
	<b>Bridges (structures)</b>	2	0

Another detected information through the construction of trees, is that the thematic most worked in IoT has been cloud computing, computer network security, wireless sensor network, entry other words. the temporal window of this results take the last 5 years, detecting that the year with most bibliographic production was 2012. the majority of bibliography found is in proceedings tipe.

Like topics of investigation that can be treated, in accordance of the obtain results are:

- Interoperability to internet of Things related with 6LoWPAN
- Interoperability to internet of Things related with Context middleware
- Interoperability to internet of Things related with Constrained Application Protocol (CoAP)

Are several the borderline topics that can be developed for Internet of Things, the observations depends of the expertise of the researcher.

## **5. CONCLUSIONS**

In this paper was presented the J48 and RandonTree algorithm analysis, employed in classification methods for data mining, focused in the relation between this algorithms and the entropy. In this way we can considerate an “order” in the information and the result obtained allows aviod significant topics, view sine de expertise of the researcher, providing a search structure of highimpact thematics an low bibliographic production. With this goal, pretend that the obtain results describe novelty and an unexplored topics.

The porpose methodology in this article no just shows data about the information that is in diferents médiums, but also gives the capability of filtered the information yin a relevant way, and then approaches the investigator with the posibbility of has a vision espectrum with a temporal thematic window. The results obtain for the J48 case, shows disordered topics, whose interpretation carry to the lest explored topics. In a similar way in en RandomTree obtain thematics with low investigative production.

The obtain thematics of major impact are performed for J48, because use the entropy information criterion y the prouding of the tree is performed in the extremes obtain the words more disordered, whose interpretation is transalate in topics lest explored, for that, can obtain knowledge innovation áreas.

The methodology proposed corroborate meaning information that the obtain results with filtered of advanced search in the academic data base, reducng bibliographic file that exist in a topic, serving as a support of decision in the time to choose a thematic by the researcher and that gives the posibility of novelty and boost the knowledge, wth the goal of movind the cience borderline.

**Bibliographic**

- [1] P. M. Alfonzo, T. J. Sakraida, and M. Hastings-Tolsma, "Bibliometrics: Visualizing the impact of nursing research," in *Online Journal of Nursing Informatics (OJNI)*, 2014, vol. 18, no. 1.
- [2] F. S. Gharehchopogh and Z. A. Khalifelu, "Analysis and evaluation of unstructured data: text mining versus natural language processing," in *Application of Information and Communication Technologies (AICT), 2011 5th International Conference on*, 2011, pp. 1–4.
- [3] M. R. A. Iqbal, S. Rahman, S. I. Nabil, and I. U. A. Chowdhury, "Knowledge based decision tree construction with feature importance domain knowledge," in *Electrical & Computer Engineering (ICECE), 2012 7th International Conference on*, 2012, pp. 659–662.
- [4] N. Kang, B. Singh, Z. Afzal, E. M. van Mulligen, and J. A. Kors, "Using rule-based natural language processing to improve disease normalization in biomedical text," in *Journal of the American Medical Informatics Association*, 2013, vol. 20, no. 5, pp. 876–881.
- [5] A. Murakami and T. Nasukawa, "Tweeting about the tsunami?: mining twitter for information on the tohoku earthquake and tsunami," in *Proceedings of the 21st international conference companion on World Wide Web*, 2012, pp. 709–710.
- [6] A. J. Nederhof, "A bibliometric study of productivity and impact of modern language and literature research," in *Research Evaluation*, 2011, vol. 20, no. 2, pp. 117–129.
- [7] J. Nuansanong, S. Kiattisin, and A. Leelasantitham, "Diagnosis and interpretation of dental X-ray in case of deciduous tooth extraction decision in children using active contour model and J48 tree," in *Electrical Engineering Congress (iEECON), 2014 International*, 2014, pp. 1–4.
- [8] J. Rey-Rocha, M. J. Martín-Sempere, L. M. Plaza, J. A. Cebrián, and others, "Geographic Information Systems for Evaluative Bibliometrics and Science Policy Purposes," 2012.
- [9] M. Tsytsarau and T. Palpanas, "Survey on mining subjective data on the web," in *Data Mining and Knowledge Discovery*, 2012, vol. 24, no. 3, pp. 478–514.
- [10] D. Ugolini, M. Neri, A. Cesario, G. Marazzi, D. Milazzo, M. Volterrani, L. Bennati, S. Bonassi, and P. Pasqualetti, "Bibliometric analysis of literature in cerebrovascular and cardiovascular diseases rehabilitation: Growing numbers, reducing impact factor," in *Archives of physical medicine and rehabilitation*, 2013, vol. 94, no. 2, pp. 324–331.
- [11] M. Ware and M. Mabe, "The STM report," *An Overv. Sci. Sch. J. Publ.*, 2012.
- [12] P. Wouters and R. Costas, "Users, narcissism and control: tracking the impact of scholarly publications in the 21st century," 2012.

- [13] R. C. Barros, M. P. Basgalupp, A. A. Freitas, and A. de Carvalho, "Evolutionary design of decision-tree algorithms tailored to microarray gene expression data sets," *Evol. Comput. IEEE Trans.*, vol. 18, no. 6, pp. 873–892, 2014.
- [14] S. Chatterji, A. Dhar, B. Barik, S. Sarkar, and A. Basu, "Anaphora resolution for bengali, hindi, and tamil using random tree algorithm in weka," in *In Proceedings of the ICON-2011*, 2011.
- [15] B. B. Dalvi, W. W. Cohen, and J. Callan, "Websets: Extracting sets of entities from the web using unsupervised information extraction," in *Proceedings of the fifth ACM international conference on Web search and data mining*, 2012, pp. 243–252.
- [16] Y. Yao, B. Wang, Z. Huang, H. Ji, and H. Li, "A Combined Data Preprocessing Method Based on K-means Clustering and Singular Spectrum Analysis," in *Proceedings of the 2012 Second International Conference on Electric Technology and Civil Engineering*, 2012, pp. 26–29.