ISSN 2278-3091

Volume 9, No.3, May - June 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse129932020.pdf

https://doi.org/10.30534/ijatcse/2020/129932020

Toward a Semantic Graph of Scientific Publications: A Bibliometric Study

Sara Mifrah[1], El Habib Benlahmer[2], Youssef Mifrah[3], Mohamed Ezeouati[4]

[1]Laboratory of Information Processing and Modeling, Hassan II University of Casablanca, Faculty of Sciences Ben M'sik, Casablanca, Morocco, mifrah.sara@gmail.com

[2]Laboratory of Information Processing and Modeling, Hassan II University of Casablanca, Faculty of Sciences Ben Msik Casablanca, Morocco, h.benlahmer@gmail.com

[3]Department of Software Engineering, National Institute of Posts and Telecommunications, INPT, Rabat, Morocco, mifrah.youssef@gmail.com

[4]Laboratory of Information Processing and Modeling, Hassan II University of Casablanca, Faculty of Sciences Ben Msik Casablanca, Morocco, mohamed.ezeouati@gmail.com

ABSTRACT

The term "scientific publication" includes several types of scientific communications and digital broadcasts that scientific researchers make of their work towards their peers and an audience of specialists. These publications describe in detail the studies or experiments carried out and the conclusions drawn from them by the authors. They undergo an examination of the value of the results and the rigor of the scientific method used for the work carried out. In this paper we evaluated the quality of a scientific article on the subject (topic), based on its citations and where is it cited, we based on the Topic modeling theme with the choice of LDA algorithms applied to the corpus Nips (1987-2016) for detecting all subjects of each paper and there citations in a first step then on the citations of each article of the corpus and on the Sentiment Analysis using a lexical based approaches. Then we created a csv file containing the link of each paper with the other cited papers (relation cited-citing), and finally generated a semantic graph between these publications.

Key words: Scientific publication, Citation analysis, Bibliometric, Citation graph, Semantic graph, Natural language Processing.

1. INTRODUCTION

When we say that an article is relevant, are we talking about the quantity of citations or the quality of the content?

This topic is the subject of recurrent debates while the number of scientific articles and the number of articles "cited" have increased sharply for two centuries, and even more for two decades due in particular to the appearance of the Internet and its search engines. Despite the establishment of scientometrics, bibliometrics [1] and management of scientific publication, authors [2][3][4][5][6][7] warn of the existence of misleading, biased or poor quality studies which pollute the research process and discredit science.

The criteria and procedures for evaluation are very much identical to those applied in the evaluation of research in general: peer reviews, panels, bibliometric indices, etc. The practical application of these tools in the case of development research may sometimes be quite difficult. Yet there seems to exist a general consensus that quality must be expressed in the same manner as for any scientific research, with the same problems and the same constraints.

The measurement of the quality of scientific productivity is problematic, no matter how quality is defined. In the case of publications, the most common method is to ask others in the same field to express their professional opinions. Obviously, the reliability of such assessments depend on the knowledge, objectivity and integrity of the assessors. What one person describes as a great contribution may be assessed as only modest by another or even regarded as contributing nothing by a third person. The peer review system is only as good as the people chosen to represent the peer group. The peer review process can sometimes be cumbersome because of difficulties in identifying appropriate peers, getting them to respond in reasonable time, and reconciling contrasting views when these arise. These are some of the reasons for the search for alternative measures of quality.

Bibliometric is a method of analyzing research activity, based on articles published in specialized scientific journals ("publications") [8]. These articles represent, in many sectors of academic research, an essential part of laboratory production. Bibliometric aims to establish relevant indicators to describe this production, and contribute to the analysis of research activity carried out in these laboratories [9]. Born on an empirical basis, difficult to handle, bibliometric-type indicators are nevertheless important tools for describing one of the essential dimensions of fundamental research.

This paper is structured as follows: Section 2 presents our research methodology. Section 3 the obtained results and finally, the conclusions of this paper are provided in Section. 4.

2. MATERIALS AND METHODS

In order to assess the quality of a scientific publication (or report), we should distinguish between an intrinsic evaluation and an evaluation on the basis of external indicators. An intrinsic evaluation by an independent expert in the domain treated by the publication is generally considered to be the best way for assessing the scientific value of a paper.

Our evaluation approach based on the evaluation of the quality of the content of a scientific article, we added two other criteria: (1) the subject treated by the citing article and its relationship with the other cited (Topic Modeling) [10] and (2) the purpose of the citation (Sentiment Analysis).

2.1 Data Set

In this study we have used the Corpus NIPS (1987-2017) -Neural Information Processing Systems (NIPS) is one of the top machine learning conferences in the world. It covers topics ranging from deep learning and computer vision to cognitive science and reinforcement learning- This dataset includes the Id, year of publication, title, and extracted text for all NIPS papers to date (ranging from the first 1987 conference to the 2017 conference). We have extracted the paper text from the raw PDF files and are releasing that both in CSV files.

2.2 Pre-processing

In this phase we built a second csv file containing the citations of all papers in the corpus (Table 1), and a third one contains the references of each paper and the relationship “referenced by”.

For the first part, we want to extract the citations from each paper, our input is a “file.txt” is the text content of an “article.pdf”, we have segmented the text into sentences with delimiter (Uppercase-point), then we let's do a search on the sentences extracted previously, if they contain a number between brackets (exp. [x] VANCOUVER Style) if "yes", we consider this sentence as a “citation”.

For the construction of the third csv file we used the GROBID machine learning reference extraction library, this library has several functionalities, we used only one, it is the parsing of references and the extraction of meta-data from each (authors, paper title, journal or conference , ISSN, publication date…), and for the construction of the relation referenced by we find if a reference of an article exists in our corpus, if "yes" we trace this relation, if "no" we find elsewhere then we trace the relation with the “URL” where the article exists.

2.3 Processing and analysis

In the field of information analysis and information retrieval, we are interested to analyzed citations and extracted [18] their relationship with referenced papers. We are started with the detection of the context of citation; either in a sentence or for more precision within a paragraph. For example, we can take sentences from the first one containing the citation to the following sentence containing the other citation or to the end of the paragraph. This step call in the field of “Information Retrieval” by "segmentation”. It consists of identifying the sentences in a sequence of words, we are only interested in sentences containing references, in this way we extracts each sequence of words containing one or more references, we obtain afterwards a list of sentences as output.

For detecting the semantic links and Topic Modeling, a second preprocessing step - applied on the previous extracted sentences (citations) - is necessary before processing algorithms. This step named in Information Retrieval System by "Lemmatization": One of the most important treatments in text mining applications, is the processing of the natural

Table 1: Example of citations

ID paper	Year	Title	Citations
2	**1987**	The Capacity of the Kanerva Associative Memory is Exponential	This exponential grovth in capacity for the Kanerva associative memory contrasts sharply yith the sub-linear grovth in capacity for the Hopfield associative memory [1]
110	**1988**	Adaptive Neural Networks Using MOS Charge Storage	DAC's have already proven themselves in situations where 5 bits or less of resolution [3] [4] are sufficient, but higher resolution is prohibitively expensive in terms of area
6716	**2017**	Improving Regret Bounds for Combinatorial Semi-Bandits with Probabilistically Triggered Arms and Its Applications	Here we briefly explain the novel aspects of our analysis that allow us to achieve new regret bounds and differentiate us from previous analyses such as the ones in [7] and [16, 15]
7189	**2017**	Convolutional Phase Retrieval	The authors in [5, 2] show that the phase retrieval problem with random coded distraction and STFT measurements can be solved by minimizing nonconvex objectives, while [5] requires resampling for the initialization, and in [2] the contraction radius is not large enough for initialization

language. This task is highly dependent on the language being processed, it can be divided into two steps: first one is to build a list of empty words that do not share a particular meaning useful for finding information for each language (elimination of stop words), and the second is the lexical analysis of the content of a text grouping the words of the same family.

2.3.1 Text sentiment extraction

In this step of sentiment analysis [19] we are used the Lexical dictionary WordNet and we have worked with the three classical classes (neutral, positive, negative) (Figure1).

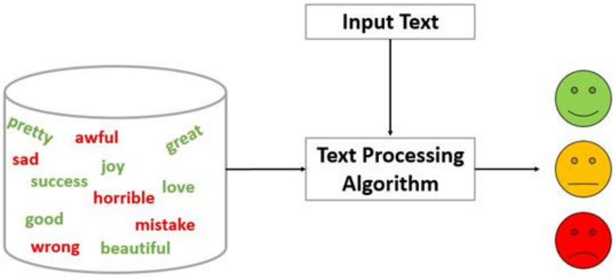


Figure 1: Polarity Extraction Processes

2.3.2 Identification of semantic links

Semantic Link analyzes the text and attempts to find all pairs of words which are semantically related. For that purpose it uses a statistical measure called Mutual Information, or MI for short. The higher the MI for a given pair of words, the higher the chance that they are related.

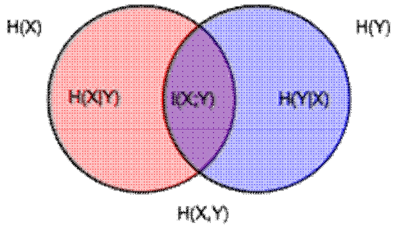


Figure 2: Venn diagram

Venn diagram [11] showing additive and subtractive relationships various information measures associated with correlated variables X and Y. The area contained by both circles is the joint entropy H(X,Y) (Figure 2). The circle on the left (red and violet) is the individual entropy H(X), with the red being the conditional entropy H(X|Y). The circle on the right (blue and violet) is H(Y), with the blue being H(Y|X). The violet is the Mutual Information I(X;Y)

Table 1: Polarity of Citations

ID paper	Year	Title	Citations	Polarity
2	**1987**	*The Capacity of the Kanerva Associative Memory is Exponential*	This exponential grovth in capacity for the Kanerva associative memory contrasts sharply yith the sub-linear grovth in capacity for the Hopfield associative memory [1]	Neutral
110	**1988**	*Adaptive Neural Networks Using MOS Charge Storage*	DAC's have already proven themselves in situations where 5 bits or less of resolution [3] [4] are sufficient, but higher resolution is prohibitively expensive in terms of area	Negative
6716	**2017**	*Improving Regret Bounds for Combinatorial Semi-Bandits with Probabilistically Triggered Arms and Its Applications*	Here we briefly explain the novel aspects of our analysis that allow us to achieve new regret bounds and differentiate us from previous analyses such as the ones in [7] and [16, 15]	Positive
7189	**2017**	*Convolutional Phase Retrieval*	The authors in [5, 2] show that the phase retrieval problem with random coded distraction and STFT measurements can be solved by minimizing nonconvex objectives, while [5] requires resampling for the initialization, and in [2] the contraction radius is not large enough for initialization	Negative

2.3.3 Topic Model

In this stage we have using the Latent Dirichlet Allocation (LDA) algorithm to detect the subjects of the scientific document and these citations, LDA [12] [13], is a probabilistic thematic modeling algorithm. It is based on the intuition that documents are composed of several themes (not words), where a theme is a multinomial distribution on a fixed vocabulary W [14].

The generative process for each document thus declines [15]:

1. Choose N ~ Poisson (ξ).
2. Choose θ ~ Dir (α).
3. For each N words wn :
   a) Choose a topic zn ~ Multinomial (θ).
   b) Choose a word wn from p (wn | zn, β), A conditioned multinomial probability on the subject zn

Where:

- ξ: Random variable generating the number of words for each document.

- α: Vector of dimension K corresponding to the parameters of the Dirichlet law generating the words of a document.

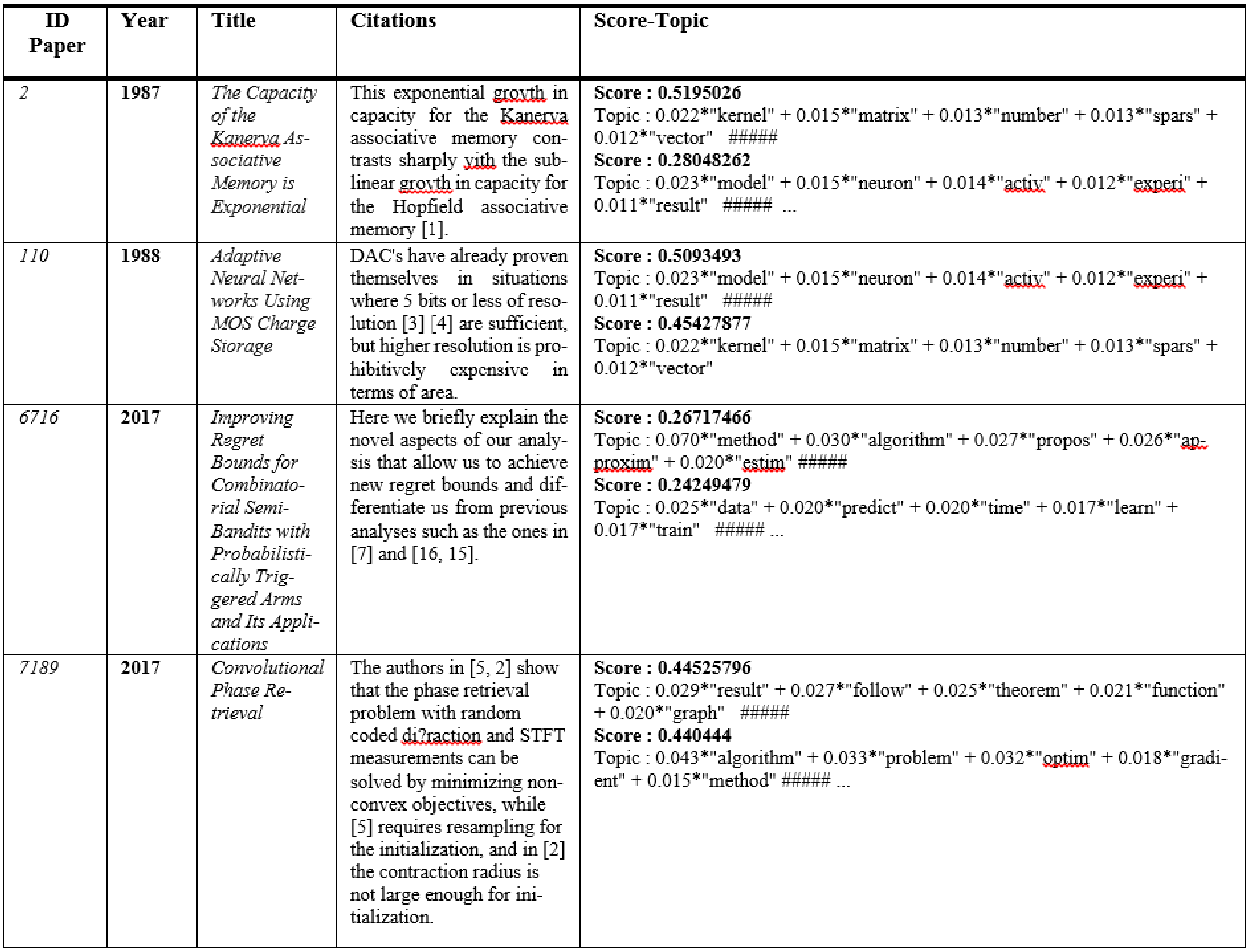
- β: The coefficients βij thus correspond to the probability of a word wj belonging to the topic zi, thus $\beta ij = P(wj | zi)$.

- θ: Latent variable represents the exact proportion of topics in each document.

- z: ziAre the topics (classes) associated with each word wj of a document.

LDA sees each document as a set of occurrences appearing in an arbitrary order, which brings it, closer to the "bag-of-words" model [16].

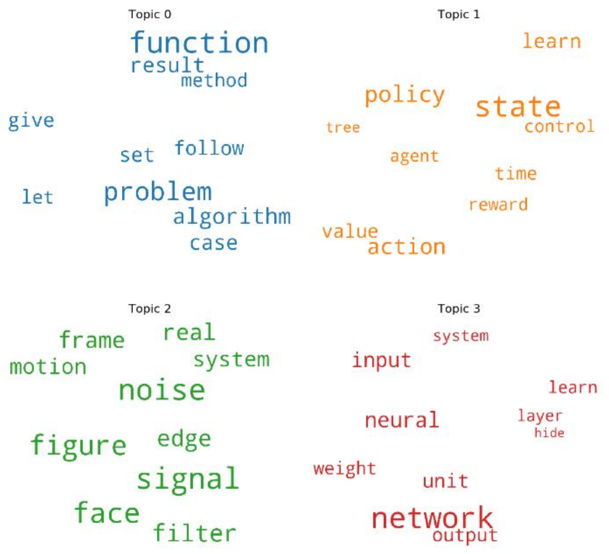
Table 2: Extracted topics from citations



ID Paper	Year	Title	Citations	Score-Topic
2	**1987**	*The Capacity of the Kanerva Associative Memory is Exponential*	This exponential grovth in capacity for the Kanerva associative memory contrasts sharply yith the sublinear grovth in capacity for the Hopfield associative memory [1].	**Score : 0.5195026** Topic : 0.022*"kernel" + 0.015*"matrix" + 0.013*"number" + 0.013*"spars" + 0.012*"vector" ##### **Score : 0.28048262** Topic : 0.023*"model" + 0.015*"neuron" + 0.014*"activ" + 0.012*"experi" + 0.011*"result" ##### ...
110	**1988**	*Adaptive Neural Networks Using MOS Charge Storage*	DAC's have already proven themselves in situations where 5 bits or less of resolution [3] [4] are sufficient, but higher resolution is prohibitively expensive in terms of area.	**Score : 0.5093493** Topic : 0.023*"model" + 0.015*"neuron" + 0.014*"activ" + 0.012*"experi" + 0.011*"result" ##### **Score : 0.45427877** Topic : 0.022*"kernel" + 0.015*"matrix" + 0.013*"number" + 0.013*"spars" + 0.012*"vector"
6716	**2017**	*Improving Regret Bounds for Combinatorial Semi-Bandits with Probabilistically Triggered Arms and Its Applications*	Here we briefly explain the novel aspects of our analysis that allow us to achieve new regret bounds and differentiate us from previous analyses such as the ones in [7] and [16, 15].	**Score : 0.26717466** Topic : 0.070*"method" + 0.030*"algorithm" + 0.027*"propos" + 0.026*"approxim" + 0.020*"estim" ##### **Score : 0.24249479** Topic : 0.025*"data" + 0.020*"predict" + 0.020*"time" + 0.017*"learn" + 0.017*"train" ##### ...
7189	**2017**	*Convolutional Phase Retrieval*	The authors in [5, 2] show that the phase retrieval problem with random coded di?raction and STFT measurements can be solved by minimizing nonconvex objectives, while [5] requires resampling for the initialization, and in [2] the contraction radius is not large enough for initialization.	**Score : 0.44525796** Topic : 0.029*"result" + 0.027*"follow" + 0.025*"theorem" + 0.021*"function" + 0.020*"graph" ##### **Score : 0.440444** Topic : 0.043*"algorithm" + 0.033*"problem" + 0.032*"optim" + 0.018*"gradient" + 0.015*"method" ##### ...

2.3.4 Topic visualization

Though we've already seen what the topic keywords in each topic are, a word cloud with the size of the words proportional to the weight is a pleasant sight. The coloring of the topics we've taken here is followed in the subsequent plots as well



When it comes to the keywords in the topics, the importance (weights) of the keywords matters. Along with that, how frequently the words have appeared in the documents is also interesting to look.

We want to keep an eye out on the words (Figure 3) that occur in multiple topics and the ones whose relative frequency is more than the weight. Often such words turn out to be less important. The chart we've drawn below is a result of adding several such words to the stop words list in the beginning and re-running the training process.

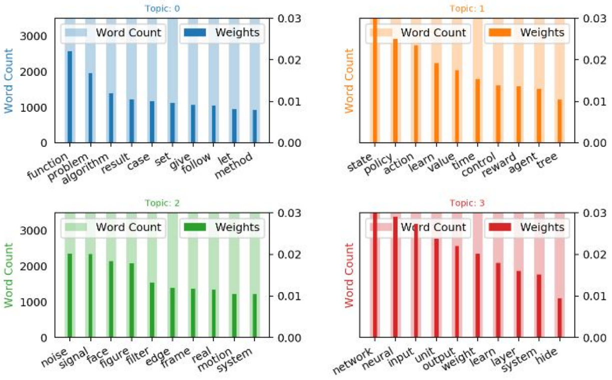


Figure 3: Word Count and Importance of Topic Keywords

2.3.5 Citing-Cited relationship

To create this tree structure (Figure 4), we created a reference relationship between all articles in the corpus (Table 3); as a starting point, we used Grobid1 (GeneRation Of BIbliographi--c Data) for extracting the references for each paper in the corpus in PDF format as input, and output as structured XML/TEI encoded documents. By the following we created a program takes as input an XML / TEI document to make parsing then extracts the titles of the references, find if the titles extracted exists in our corpus, as output we created a file csv with the form (id; pdf-name ; references; referenced by; title; pub-year).

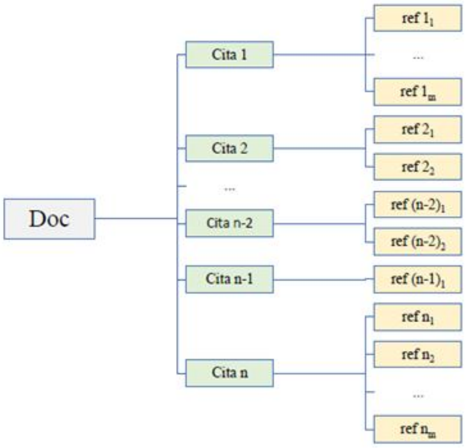


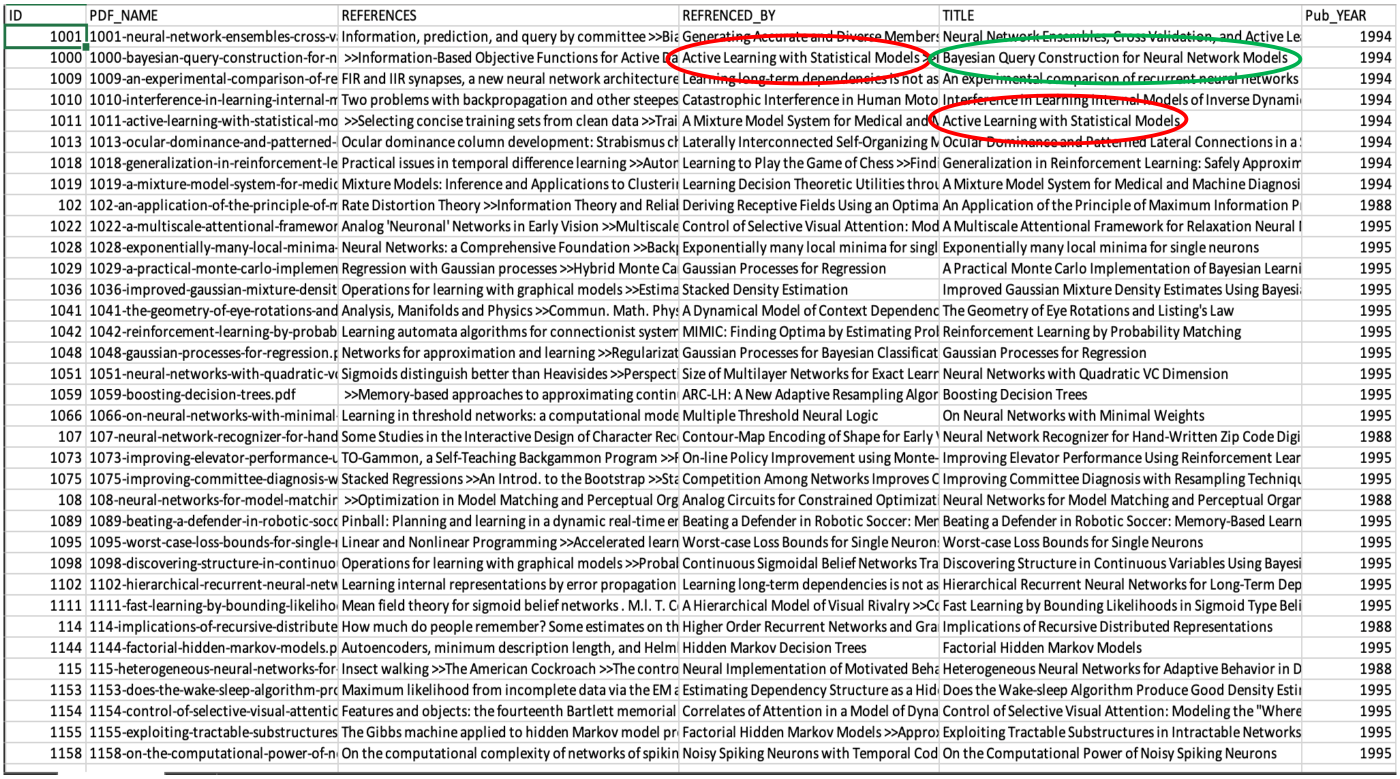
Figure 4: Relational schema of a scientific document with "ref" is a document

The referenced by or cited relationship allows us to establish a relationship between the articles in the same corpus.

After our experimentation on NIPS corpus we observed that only the articles in the period (1987-1997) which change this relation, the number of these articles is more than 200 articles among 3543 between the periods (1987-2016) (Table 4).

[1] A machine learning software for extracting information from scholarly documents https://grobid.readthedocs.io

Table 3 : "Referenced by" relationship or Cited



ID	PDF_NAME	REFERENCES	REFRENCED_BY	TITLE	Pub_YEAR
1001	1001-neural-network-ensembles-cross-v	Information, prediction, and query by committee >>Bia	Generating Accurate and Diverse Members	Neural Network Ensembles, Cross Validation, and Active Le	1994
1000	1000-bayesian-query-construction-for-n	>>Information-Based Objective Functions for Active Da	Active Learning with Statistical Models >>	Bayesian Query Construction for Neural Network Models	1994
1009	1009-an-experimental-comparison-of-re	FIR and IIR synapses, a new neural network architecture	Learning long-term dependencies is not as	An experimental comparison of recurrent neural networks	1994
1010	1010-interference-in-learning-internal-m	Two problems with backpropagation and other steepes	Catastrophic Interference in Human Moto	Interference in Learning Internal Models of Inverse Dynami	1994
1011	1011-active-learning-with-statistical-mo	>>Selecting concise training sets from clean data >>Trai	A Mixture Model System for Medical and M	Active Learning with Statistical Models	1994
1013	1013-ocular-dominance-and-patterned-	Ocular dominance column development: Strabismus ch	Laterally Interconnected Self-Organizing M	Ocular Dominance and Patterned Lateral Connections in a	1994
1018	1018-generalization-in-reinforcement-le	Practical issues in temporal difference learning >>Autor	Learning to Play the Game of Chess >>Findi	Generalization in Reinforcement Learning: Safely Approxim	1994
1019	1019-a-mixture-model-system-for-medic	Mixture Models: Inference and Applications to Clusterii	Learning Decision Theoretic Utilities throu	A Mixture Model System for Medical and Machine Diagnosi	1994
102	102-an-application-of-the-principle-of-m	Rate Distortion Theory >>Information Theory and Relial	Deriving Receptive Fields Using an Optima	An Application of the Principle of Maximum Information P	1988
1022	1022-a-multiscale-attentional-framewor	Analog 'Neuronal' Networks in Early Vision >>Multiscale	Control of Selective Visual Attention: Mod	A Multiscale Attentional Framework for Relaxation Neural I	1995
1028	1028-exponentially-many-local-minima-	Neural Networks: a Comprehensive Foundation >>Back	Exponentially many local minima for singl	Exponentially many local minima for single neurons	1995
1029	1029-a-practical-monte-carlo-implemen	Regression with Gaussian processes >>Hybrid Monte Ca	Gaussian Processes for Regression	A Practical Monte Carlo Implementation of Bayesian Learni	1995
1036	1036-improved-gaussian-mixture-densit	Operations for learning with graphical models >>Estima	Stacked Density Estimation	Improved Gaussian Mixture Density Estimates Using Bayesi	1995
1041	1041-the-geometry-of-eye-rotations-and	Analysis, Manifolds and Physics >>Commun. Math. Phys	A Dynamical Model of Context Dependenc	The Geometry of Eye Rotations and Listing's Law	1995
1042	1042-reinforcement-learning-by-probab	Learning automata algorithms for connectionist system	MIMIC: Finding Optima by Estimating Prol	Reinforcement Learning by Probability Matching	1995
1048	1048-gaussian-processes-for-regression.p	Networks for approximation and learning >>Regularizat	Gaussian Processes for Bayesian Classificat	Gaussian Processes for Regression	1995
1051	1051-neural-networks-with-quadratic-vc	Sigmoids distinguish better than Heavisides >>Perspecti	Size of Multilayer Networks for Exact Learr	Neural Networks with Quadratic VC Dimension	1995
1059	1059-boosting-decision-trees.pdf	>>Memory-based approaches to approximating contin	ARC-LH: A New Adaptive Resampling Algor	Boosting Decision Trees	1995
1066	1066-on-neural-networks-with-minimal-	Learning in threshold networks: a computational mode	Multiple Threshold Neural Logic	On Neural Networks with Minimal Weights	1995
107	107-neural-network-recognizer-for-hand	Some Studies in the Interactive Design of Character Rec	Contour-Map Encoding of Shape for Early V	Neural Network Recognizer for Hand-Written Zip Code Digi	1988
1073	1073-improving-elevator-performance-u	TO-Gammon, a Self-Teaching Backgammon Program >>F	On-line Policy Improvement using Monte-	Improving Elevator Performance Using Reinforcement Lear	1995
1075	1075-improving-committee-diagnosis-w	Stacked Regressions >>An Introd. to the Bootstrap >>Sta	Competition Among Networks Improves C	Improving Committee Diagnosis with Resampling Techniqu	1995
108	108-neural-networks-for-model-matchir	>>Optimization in Model Matching and Perceptual Org	Analog Circuits for Constrained Optimizati	Neural Networks for Model Matching and Perceptual Orgar	1988
1089	1089-beating-a-defender-in-robotic-socc	Pinball: Planning and learning in a dynamic real-time er	Beating a Defender in Robotic Soccer: Mer	Beating a Defender in Robotic Soccer: Memory-Based Learn	1995
1095	1095-worst-case-loss-bounds-for-single-r	Linear and Nonlinear Programming >>Accelerated learn	Worst-case Loss Bounds for Single Neuron:	Worst-case Loss Bounds for Single Neurons	1995
1098	1098-discovering-structure-in-continuo	Operations for learning with graphical models >>Probal	Continuous Sigmoidal Belief Networks Tra	Discovering Structure in Continuous Variables Using Bayesi	1995
1102	1102-hierarchical-recurrent-neural-netw	Learning internal representations by error propagation	Learning long-term dependencies is not as	Hierarchical Recurrent Neural Networks for Long-Term Dep	1995
1111	1111-fast-learning-by-bounding-likeliho	Mean field theory for sigmoid belief networks . M.I. T. C	A Hierarchical Model of Visual Rivalry >>Co	Fast Learning by Bounding Likelihoods in Sigmoid Type Beli	1995
114	114-implications-of-recursive-distribute	How much do people remember? Some estimates on th	Higher Order Recurrent Networks and Gra	Implications of Recursive Distributed Representations	1988
1144	1144-factorial-hidden-markov-models.p	Autoencoders, minimum description length, and Helml	Hidden Markov Decision Trees	Factorial Hidden Markov Models	1995
115	115-heterogeneous-neural-networks-for-	Insect walking >>The American Cockroach >>The contro	Neural Implementation of Motivated Beha	Heterogeneous Neural Networks for Adaptive Behavior in D	1988
1153	1153-does-the-wake-sleep-algorithm-pro	Maximum likelihood from incomplete data via the EM a	Estimating Dependency Structure as a Hid	Does the Wake-sleep Algorithm Produce Good Density Estir	1995
1154	1154-control-of-selective-visual-attentic	Features and objects: the fourteenth Bartlett memorial	Correlates of Attention in a Model of Dyna	Control of Selective Visual Attention: Modeling the "Where	1995
1155	1155-exploiting-tractable-substructures	The Gibbs machine applied to hidden Markov model pr	Factorial Hidden Markov Models >>Approx	Exploiting Tractable Substructures in Intractable Networks	1995
1158	1158-on-the-computational-power-of-n	On the computational complexity of networks of spikin	Noisy Spiking Neurons with Temporal Cod	On the Computational Power of Noisy Spiking Neurons	1995

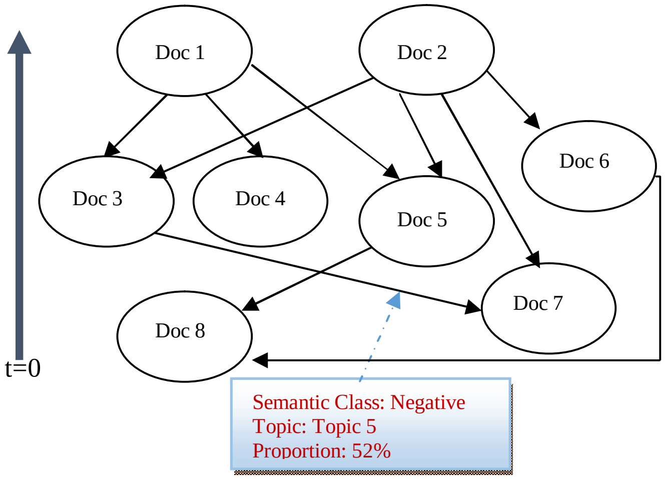
3. RESULTS AND DISCUSSION

Each citation in a scientific paper is linked at least to one reference's article (Figure 4).

After we got the results, we can incorporate the concept of the semantic class. We can say for example that:

« *The author of the basic document (Doc 3) criticizes the approach **x** cited of the article referenced (Doc 7)**,** in the subject **z** (Topic 5) with the proportion **52%** ».*

With : ***p*** ϵ [0,100[



In order to get our result (Figure 6) we used the Graphvis tool, it is an open source graph visualization software. Graph visualization is a way of representing structural information as diagrams of abstract graphs and networks (Figure 5). It has important applications in networking, bioinformatics, software engineering, database and web design, machine learning, and in visual interfaces for other technical domains.

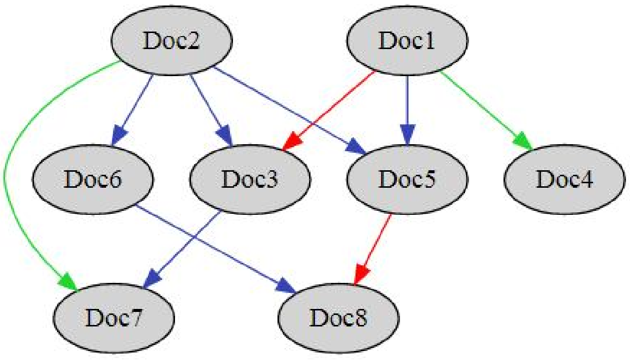


Figure 5: Generated Graph using Graphvis

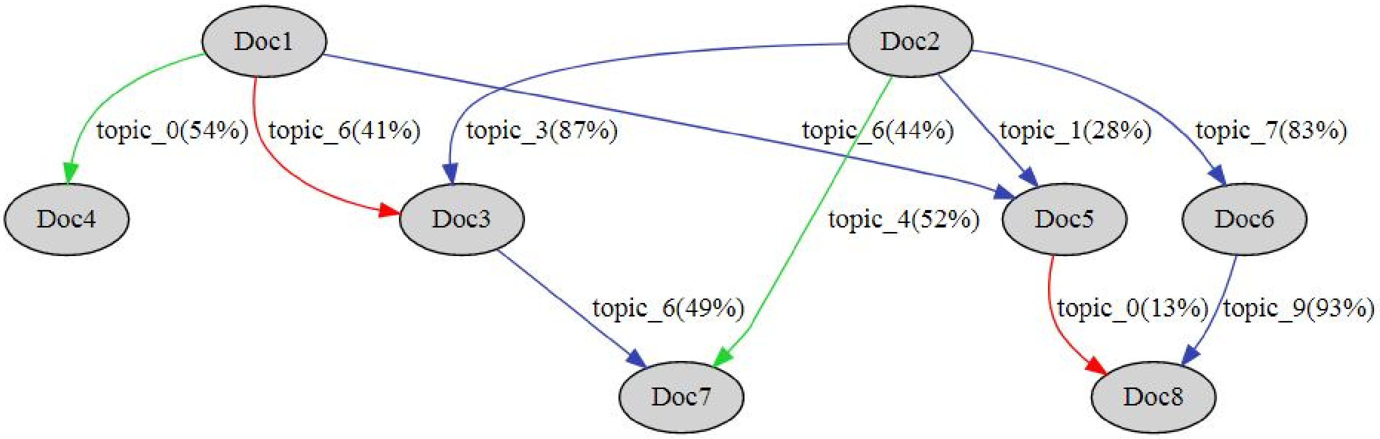


Figure 6: Example of relationship graph with topic & semantic class

Each node in the graph represents a scientific article and each arc represents the reference or citation relation, moreover the color of the arc represents the semantic class and the label on each arc is the dominant topic of the citation.

To achieve our result we have generated a text file of the following form using Graphvis library:

```
digraph prof {
    ratio = fill;
    node [style=filled];
    Doc1 -> Doc3 [color="0.002 0.999 0.999",label = "topic_6(41%)"];
    Doc2 -> Doc3 [color="0.649 0.701 0.701",label = "topic_3(87%)"];
    Doc1 -> Doc4 [color="0.348 0.839 0.839",label = "topic_0(54%)"];
    Doc1 -> Doc5 [color="0.649 0.701 0.701",label = "topic_6(44%)"];
    Doc2 -> Doc6 [color="0.647 0.702 0.702",label = "topic_7(83%)"];
    Doc2 -> Doc7 [color="0.348 0.839 0.839",label = "topic_4(52%)"];
    Doc3 -> Doc7 [color="0.650 0.700 0.700",label = "topic_6(49%)"];
    Doc5 -> Doc8 [color="0.002 0.999 0.999",label = "topic_0(13%)"];
    Doc6 -> Doc8 [color="0.650 0.700 0.700",label = "topic_9(93%)"];
    Doc2 -> Doc5 [color="0.650 0.700 0.700",label = "topic_1(28%)"];
}
```

Each line represents the link of the two nodes; the color of the link arc and the label on the arc.

4. CONCLUSION

In conclusion, we can say that the analysis of scientific documents [17] is a topic that interests many people from different disciplines. We find, on the one hand, scientists (researchers) that produce and exploit these documents; on the other hand, scientific publishers, librarians, and producers of databases... who handle and process these scientific documents.

The better scientific articles you use in your literature review, the more robust your own research will be. If you use articles that are not authoritative, you run the risk of drawing unfounded conclusions harmful to your dissertation or thesis.

Your supervisor always checks the sources of the conclusions you draw.

REFERENCES

1. Weingart P. **Impact of bibliometrics upon the science system: Inadvertent consequences?** Scientometrics .62(1), 117-131. 2005 https://doi.org/10.1007/s11192-005-0007-7
2. Maddox J. **Why the pressure to publish?.** Nature 1988, 333, 493. https://doi.org/10.1038/333493a0
3. Angell, M. **Publish or perish: a proposal**. Annals of Internal Medicine, 104(2), 261-262. 1986 https://doi.org/10.7326/0003-4819-104-2-261
4. De Rond M & Miller A.N. **Publish or perish bane or boon of academic life?** Journal of Management Inquiry, (2005) 14(4), 321-329. https://doi.org/10.1177/1056492605276850
5. Darnill A. **Publish or perish**. Accountancy, 118 (1,237), 1997.
6. Colquhoun D**. Publish-or-perish: Peer review and the corruption of science** .The Guardian, 5(09).
7. Lehrer J. **The truth wears off: is there something wrong with the scientific method?** The New Yorker.
8. Pritchard, A. **Statistical Bibliography or Bibliometrics** Journal of Documentation. 348-349. 1969
9. Vachon, Éric, Yves Petit, et Pierre Etève . **Les indicateurs Bibliométriques en recherche**. Éducation & formations, n°59, 123-128 , avril-juin 2001
10. Sara Mifrah and Ben Lahmar El Habib. **Semantic Relationship Study between Citing and Cited Scientific Articles Using Topic Modeling**. In Proceedings of the 4th International Conference on Big Data and Internet of Things (BDIoT'19), Rabat Morocco (2019). DOI: https://doi.org/10.1145/3372938.3372943
11. Venn Diagram [Availabale] https://en.wikipedia.org/wiki/Venn_diagram

12. Blei, D., Ng, A., & Jordan, M.. **Latent dirichlet allocation.** The Journal of Machine Learning Research 3, 993–1022.
13. 12 Griffiths, T. L., & Steyvers, M. **Finding scientific topics**. doi: 10.1073/pnas.0307752101 PNAS April 6, 2004 vol. 101 no. suppl 1 , 5228-5235. 2004 https://doi.org/10.1073/pnas.0307752101
14. 13 Deveaud, R., Bonnefoy, L., & Bellot, P. **Quantification et identification des concepts implicites d'une requête**. [Availabale] http://coria.unine.ch/coria/coria2013_27.pdf .2013
15. Francesiaz, T., Graille, R., & Metahri, B. **Introduction aux modèles probabilistes utilisés en fouille de données**. Retrieved juillet 20, 2016 [Availabale] http://www-ljk.imag.fr/membres/Marianne.Clausel/Fichiers/Rapport_Metahri_Graille_Francesiaz.pdf
16. Rigouste, L., Cappe, O., & Yvon, F. **Quelques observations sur le modèle LDA**. 2006. [Availabale] https://perso.limsi.fr/yvon/publications/sources/Rigouste06lda.pdf
17. Mifrah, S., Benlahmer, El H.: **Semantico-automatic Evaluation of Scientific Papers: State of the Art**. BDCA'17 Proceedings of the 2nd international Conference on Big Data, Cloud and Applications Tetouan, Morocco — March 29 - 30, 2017 ISBN: 978-1-4503-4852-2 DOI: 10.1145/3090354.3090380 (2017)
18. Kadu, Payal, and Ashwini V. Zadgaonkar. "**Knowledge Extraction from Text Document Using Open Information Extraction Technique**." International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE) Volume 9, No.2 ISSN 2278-3091, DOI: 10.30534/ijatcse/2020/208922020 (2020).
19. Ali Muttaleb Hasan et al., "**Knowledge-Based Semantic Relatedness measure using Semantic features**" International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE) Volume 9, No.2 ISSN 2278-3091 Volume 9 No.2, March - April 2020 DOI/ 10.30534/ijatcse/2020/02922020