The Information Retrieval Process of the Scientific Production at Departmental-level of Universities: A New Approach.

César David Loaiza Quintana[1] and Víctor Andrés Bucheli Guerrero[2]

[1] *cesar.loaiza@correounivalle.edu.co*
Universidad del Valle, Escuela de Ingeniería de Sistemas y Computación, Facultad de Ingeniería. Universidad del Valle, Sede Meléndez. Calle 13 # 100-00. Cali, Valle Del Cauca (Colombia).

[2] *victor.bucheli@correounivalle.edu.co*
Universidad del Valle, Escuela de Ingeniería de Sistemas y Computación, Facultad de Ingeniería. Universidad del Valle, Sede Meléndez. Edificio 331 oficina 2113, Calle 13 # 100-00. Cali, Valle Del Cauca (Colombia)

Introduction

Our work was focused on document retrieval from Scopus databases of the Escuela de Ingeniería de Sistemas y Computación (EISC) of the Universidad del Valle (Cali - Colombia).

The databases systems as WoS (Web of Science) or Scopus contain the knowledge produced by engineer schools. However, this information is ambiguous and the retrieving of the specific documents of one school is identity uncertainly (Pasula et al., 2003). Thus, the design of machines (search engines) to retrieve the relevant documents of engineer schools is a complex process.

After the work of Bucheli et al. (2013); Cuxac, Lamirel, & Bonvallot (2013) proposed a semi-supervised approach, mixing soft-clustering and Bayesian learning. Additionally, Huang et al. (2014) proposed a rule-based algorithm. Both approaches were for affiliation disambiguation.

We reproduced the model proposed by Bucheli et al. (2013). The results show that the model can be used to information retrieval of department-level. In addition, we proposed a new approach addressing the problem of classification using network science. The future work will be related with building a model according to the network science approach.

Methodology

Model of Bucheli et al. (2013)

We followed the methodology specified by Bucheli et al. (2013) shown in Figure 1(a).

1) The configuration of the initial search strategy proposed by Bucheli et al. (2013) was applied using the Scopus search engine to get a set **I** composed by documents that contains all the documents that belong to EISC and others that not belong to it.

2) The initial search strategy was based on a review of the research activity of the School and it proposes recovering a set of documents $\mathbf{I} = \mathbf{A} \cup \mathbf{J} \cup \mathbf{S} \cup \mathbf{O}$. The staff **S** set is made up by papers which are related to a list of school professors names explicitly. The journal set **J** is the bunch of documents published in the journals where the school has previously published. The address set **A** is related to the documents that have in their affiliation the name of the school explicitly. Finally, socio-semantic set $\mathbf{O} = \mathbf{S} \cup \mathbf{C}$, where the concepts set **C** is made up by the documents related to a bunch of research areas from a school. Every set mentioned before has an additional restriction; his documents must belong to the university that hosts the internal-level unit, in our case to the Universidad del Valle.

3) An Expert from EISC classified all the documents from the initial search and we built a relevant set **R** with **I** elements that belong to EISC.

4) We built a dataset where one paper or instance is characterized by a vector (with five positions). Each position is a binary variable, related to sets **A**, **S**, **J**, **O** and **R**, that tell us if the paper belongs or not to the corresponding set. Thus, the instance class is determined by the variable **R**.

5) Afterwards, we made the classification using the Naïve Bayes model of information retrieval illustrated in (1). It was evaluated based on all instances of the dataset. We used standard measurements over cross validation test 10 fold (Witten, 2005; Baeza-Yates, 1999). On the other hand, the publication year was taken into account as parameter of evaluation. Thus, we train the model with paper published between two specific years, for instance 1989-2010 and testing the model with papers published in the following years. This procedure was evaluated by the following years of training 1989-2011, 1989-2012 and 1989-2013.

$$p(R|J,S,O,A) = \frac{p(R)p(J,S,O,A|R)}{p(J,S,O,A)} \quad (1)$$

Proposed model based on network science

The machine learning process follows five phases: Selecting data, expert validation, co-author network building, feature extraction from network and classification, as shows the Figure 1(b).

The data selection trough the initial search strategy and the expert validation have be taken into account similarly to the review model of Bucheli et al. (2013). Here, the document corpus used is the same of evaluation model applied to the EISC, however the feature extraction changes and the features are related with network measurements.

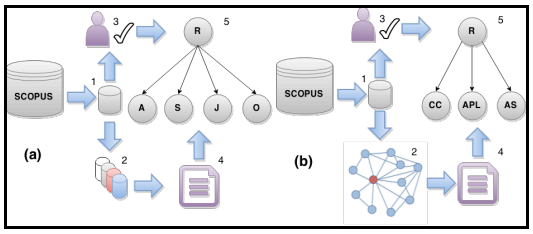

Figure 1. The five phases of the evaluated and proposed methodologies.

The document corpus contains information about co-authorship relations. Each author is identified by an ID that Scopus assigns. We build a co-authorship network, where, the network is traduced as a weighted and undirected graph in which the weight of the edges designates the number of documents where whichever two authors have participated. The new dataset is built as follows: one document or instance is a vector of values where each position is a variable related with one measurement of co-author network, in which, the specific paper was subtracted. Thus, for each instance, the authors that participated in the specific document are deleted and the measures are computed again. Additionally, the last variable **R** shows if the paper belongs or not belongs to the School. The measurements of networks are:

1. The Cluster Coefficient (**CC**): The local clustering coefficient captures the degree to which the neighbours of a given node link to each other. We use the average of all local clustering coefficients.
2. The average path length (**APL**) is the average distance between all pairs of nodes in the network.
3. The average strength (**AS**), is the average of the sum of the edge weights of each node. (Barabasi. 2012).

Finally, we develop a supervised learning environment through a Naïve Bayes Classifier and the proposed model is evaluated and compared with the model proposed by Bucheli et al (2013).

Results, discussion and future work

Table 1 shows standard evaluation measurements. Here, we introduce the cross validation fold 10 test, the measurements show in Bucheli. et al. (2013), and the evaluation for different publication years 1989-2011, 1989-2012 and 1989-2013. The results show that the model was applied to other School with similar performance measurements, in this sense the model is consistent and allows to build one search engine of department-level. Additionally, we evaluated the practical utility of the model, verifying that it is capable of doing an acceptable prediction of EISC's documents published after a specific date when it is trained with a set of documents published until that date.

In this work, we found the finger prints of department-level of universities that allow us to design search engines that retrieve relevant documents of department-level.

Table 1. Evaluation measurements of the model.

	Recall	Precision	ROC curve
EISC Univalle			
Cross Validation fold 10	0,932	1,000	0,989
Bucheli et al. (2013)			
Department of Industrial Engineering –University of Pittsburgh	0,494	0,997	0,984
Faculty of Engineering – Universidad de los Andes (Colombia)	0,954	0,992	0,965
EISC Univalle			
Training:1989-2011 Evaluation: 2012-2014	0.833	1.000	0.974
Training 1989-2012 Evaluation: 2013-2014	0.826	1,000	0.964
Training 1989-2013 Evaluation: 2014	0,786	1,000	0,939

The networks science approach is an opportunity to propose a mathematical model able to learn the structure of co-authorship network from a particular school. Then, we can design a classifier of relevant documents at department-level based on co-authorship relations. This allows making a classification with little a priori information about an organization, which turns into a more general model than Bucheli et al. (2013). We proposed a model, namely (2).

$$p(R|CC, APL, AS) = \frac{p(R)p(CC, APL, AS|R)}{p(CC, APL, AS)} \quad (2)$$

We suggest as future work to evaluate the model based on network measurements at the same school and other 3 schools of engineering from different universities.

Acknowledgments

Thanks to Convocatoria Interna, Universidad del valle 2014; Facultad de Ingeniería, Universidad del valle; and EISC.

References

Baeza-Yates, R. (1999). *Modern information retrieval*. New York: Addison-Wesley.

Barabási, A.L. (2012). *Network science book*. Center for Complex Network Research, Northeastern.

Bucheli, V., Calderón, J., Gonzales, F., Bidanda, B., Valdivia, J., & Zarama, R. (2013). Model to support the information retrieval process of the scientific production at departmental-level or faculty-level of universities. *Proc. ISSI*.

Cuxac, P., Lamirel, J. C., & Bonvallot, V. (2013). Efficient supervised and semi-supervised approaches for affiliations disambiguation. *Scientometrics*, *97*(1), 47-58.

Huang, S., Yang, B., Yan, S., & Rousseau, R. (2014). Institution name disambiguation for research assessment. *Scientometrics*, *99*(3), 823-838.

Pasula, et al. (2003). Identity Uncertainty and Citation Matching, *NIPS,* MIT Press.

Witten, I. (2005). *Data mining: practical machine learning tools and techniques*. 2nd ed., Amsterdam: Morgan Kaufman.