# Looking back at the EUNIS repository data using text mining techniques

Elsa Cardoso[1], Ricardo Ribeiro[2]

[1]ISCTE- Instituto Universitário de Lisboa and INESC-ID, Av. das Forças Armadas, 1649-026 Lisboa, Portugal, elsa.cardoso@iscte.pt
[2]ISCTE- Instituto Universitário de Lisboa and INESC-ID, Av. das Forças Armadas, 1649-026 Lisboa, Portugal, ricardo.ribeiro@iscte.pt

**Keywords**
Text mining, topic modelling, EUNIS repository

## 1. Summary

Topic modelling algorithms can be applied to a large collection of documents to uncover themes and their relationship, thus adding an automatic semantic layer to the archive of documents. In this work, topic modelling has been applied to the collection of papers published at the EUNIS (European University Information Systems) annual congresses from 1997 until 2015. Drawing on the initial work of (Wilson and Bergström, 2015), this paper extends the historical analysis of the conference proceedings with the application of text mining techniques to analyse the original texts. The performed analyses include the automatic identification of the most relevant topics and their evolution throughout the years for European University Information Systems, and the quantification of the degree of internationalization and cooperation of the EUNIS annual conference.

## 2. EUNIS repository analysis

The mission of the European University Information Systems organization (EUNIS) is to help member institutions to develop their information technology (IT) landscape by sharing experiences and working together. EUNIS was created in 1993, and currently has approximately 120 member institutions. Several activities and events are periodically organized. The most significant events are the annual conference or congress (hosted by a member institution) and the Rectors conference. Task forces (TF) have been created to foster a deeper collaboration and sharing of experiences between members with similar interests. Presently, there are four task forces established: BencHEIT (focused on benchmarking IT major costs), business intelligence, e-learning, and interoperability. Task forces also organize separate events, like the BencHEIT annual event (since 2012) and the Business Intelligence TF event in Hannover and Paris, in 2011 and 2013 respectively. More recently, the EUNIS Research and Analysis (ERAI) initiative has been launched with the purpose promoting the EUNIS research outcomes, being responsible for the editing of an e-journal, the EUNIS Journal of Higher Education IT (EJHEIT). The first issue of this e-journal was published in the fourth quarter of 2014.

In 2015, the ERAI team performed the first analysis of the existing set of publications (Wilson and Bergström, 2015) focusing on the papers published in the annual congresses. This exercise was very insightful, sharing light on some publishing trends. Since EUNIS does not currently have a data repository, the proceedings of the annual conferences have been stored in the original conference web sites, maintained by the member institutions that hosted the events. The ERAI team performed an extensive web search, to assemble, as much as possible, all the published papers. Key findings from (Wilson and Bergström, 2015) were: (1) a set of comprehensive records exist only from 1997 onwards; (2) the congress tracks are inconsistent across years; and (3) authors vary in their interpretation of keywords (i.e., keywords are also inconsistent across years).

Three major trends were identified by Wilson and Bergström's analysis. First, 2001 up to 2009 was identified as the most popular period for paper presentations, with 2007 achieving the highest value of paper publication in congresses. Second, United Kingdom (UK), Portugal, Germany, and Spain constitute the top-4 countries of paper authors. Finally, the most popular themes identified by the authors were e-issues, leadership and management, infrastructure and security, and information management.

The goal of the present work is to expand the data analysis of exiting EUNIS congress publications, applying business intelligence (BI) and text mining techniques. The motivation for this work emerged from the need to have a comprehensive view of the evolution of information systems development in Higher Education Institutions in Europe. BI and text mining enable the automatic extraction of reliable sematic information from unrestricted text, such as papers' full text and congress programmes. The end result complements the knowledge gained from a standard statistical analysis.

In order to achieve this goal, a thorough data cleaning process has been developed. This data preparation stage constituted a very intensive and time-consuming process, since contextual information about each congress is scattered (across different web sites and books) and sometimes insufficient. A multidimensional model has been designed to facilitate the analysis of paper submissions to EUNIS congresses. The multidimensional analysis was implemented using Microsoft Excel power pivots, applying standard Data Warehouse techniques. The output of this stage enabled the computation of several generic data repository indicators for the dataset of publications from 1997 until 2015, such as:

- total number of papers
- papers with keyword definition (number and %)
- papers with abstract definition (number and %)
- papers without author definition (number and %)
- papers with authors from different countries (number and %)
- papers without track definition (number and %)
- number of countries represented
- average number of authors per paper
- number of papers per track (average, min, max)
- number of congress tracks

Text mining techniques, and topic modelling algorithms in particular, can be applied to a large collection of documents to uncover themes and their relationship. The main goal of this work is to use different topic modelling algorithms to define an automatic semantic layer to the archive of EUNIS congress papers published in the time period of 1997 until 2015. The initial set of research questions outlined for this project is:

RQ1. Which are the most relevant topics in the area of European University Information Systems?
RQ2. How did this set of topics evolve throughout the years?
RQ3. How did the number of represented countries evolve throughout the years?
RQ4. How did the number of papers with authors from different countries evolve throughout the years?
RQ5. How did the ratio number of published papers versus number of conference tracks evolve throughout the years?

In order to answer the first two research questions, topic modelling algorithms were applied to automatically identify the most relevant topics and their evolution throughout the years for European University Information Systems. RQ3 has an impact on the analysis of the degree of internationalization of annual conferences. As previously mentioned, one of the main objectives of EUNIS is to establish an international Higher Education community to collaborate and share experiences among members. To this end, RQ4 will enable an analysis of the degree of cooperation between countries.

## 3. Topic modelling

Topic models allow us to explore large collections of (textual) documents in an unsupervised manner. The intuition behind this type of models is that by collecting statistics about the distributional properties of the words occurring in the documents of a given collection it is possible

to uncover its latent topics. “Words with similar distributional properties have similar meanings” (Sahlgren, 2006).

In general, topic modelling approaches can be divided in probabilistic approaches and non-probabilistic approaches. The main difference between these approaches is that the relationships between words and topics and documents and topics in probabilistic models, like Latent Dirichlet Aloocation (Blei, Ng, and Jordan, 2003), are modelled using probability distributions, while in non-probabilistic approaches, like Latent Semantic Analysis (Landauer and Dumais, 1997), such restriction does not exists (Steyvers and Griffiths, 2007). However, as mentioned by Blei, Ng, and Jordan (2003), the main advantage of (generative) probabilistic models is their modularity and extensibility. In terms of performance, both types of approaches show similar performance when used for classification, with probabilistic models inducing better quality topics (Stevens et al., 2013). Moreover, as probabilistic modules, this kind of approaches can be easily integrated in a more complex model.

One of the most well-known probabilistic topic models is Latent Dirichlet Allocation (LDA). Given a fixed number of topics, LDA considers that each topic is characterized by a probability distribution over the vocabulary of the collection and that documents are mixtures of topics. LDA is a generative model, so it assumes that each document of the collection is generated from a topic distribution and each word of the document is generated from the distribution over the vocabulary corresponding to a topic randomly selected from the topic distribution.

One of the limitations of this model is that it assumes a fixed number of topics, required to train the model. This choice can have a strong impact on the results: a small number can lead generic topics, while choosing a large of number may lead to low quality topics. The hierarchical Dirichlet Process (HDP) (Teh et al., 2007) can be seen as an extension to LDA that allows an unrestricted number of topics. HDP can estimate the number of topics of a given collection by using a sampling method.

In this work we explored both approaches to determine the right amount of topics that better explains the collection of documents (i.e., EUNIS papers). To this end, several experiences were conducted with different algorithms, leading to different sets of topics. In the end, the final set of topics is always a choice of the analyst. The rationale is to perform a sufficient number of experiences to aid the analyst in the selection of the most relevant and recurrent topics. The inferred models are used to analyze the evolution of topics through time.

## 4. Conclusions

Preliminary results allowed us to better understand the main themes of the EUNIS conferences, as well as the evolution of such themes over the last 20 years.

An important aspect of this work is data visualization. The generic conference paper indicators are simple data visualization, characteristic of business intelligence projects. However, topic modelling visualization is still a research topic and hence constitutes a challenge to this work. The goal is to provide a sufficiently rich albeit simple text visualization of the evolution of themes in the context of the EUNIS conferences.

## 5. REFERENCES

Wilson, N., and Bergström, J. (2015) EUNIS Congress' 21st Birthday - A Historical Perspective on its Proceedings. In Proceedings of EUNIS 2015, Dundee, UK, 10-12 June, 2015

Landauer, T. K. and Dumais, S. T. (1997), A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge, in *Psychological Review*, *104* (2) , 211–240

Sahlgren, M. (2006), *The Word-Space Model*, PhD Thesis

Teh, Y., Jordan, M., Beal, M., Blei, D. (2007), Hierarchical Dirichlet processes, *Journal of the American Statistical Association,* 101(476), 1566–1581

Blei, D., Ng, A., Jordan, M. (2003), Latent Dirichlet allocation, *Journal of Machine Learning Research* 3, 993–1022

Steyvers, M., and Griffiths, T. (2007), Handbook of Latent Semantic Analysis, chap. Probabilistic Topic Models, 427–448. Lawrence Erlbaum Associates

Stevens, K., Kegelmeyer, P., Andrzejewski, D., and Buttler, D. (2012), Exploring Topic Coherence over many models and many topics. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Association for Computational Linguistics, 952–961

## 6. AUTHORS' BIOGRAPHIES

**Prof. Dr. Elsa Cardoso.** Assistant Professor, Director of the Master in Integrated Business Intelligence Systems at ISCTE – University Institute of Lisbon, Portugal, and leader of the Business Intelligence Task Force of EUNIS (European University Information Systems organization). She is also a researcher at the Information and Decision Support Systems Group of INESC-ID Lisboa, Portugal.

She has a PhD in Information Science and Technology, with a specialization in Business Intelligence. Her research interests include business intelligence and data warehousing, big data analytics, performance management, balanced scorecard, business process management, applied to Higher Education, Healthcare, and IT service management. She is a member of: Institute of Electrical and Electronics Engineers (IEEE), Association for Computing Machinery (ACM), Portuguese Institute of Business Process Management (IPBPM), Portuguese Quality Association and the vice-representative of ISCTE-IUL at CS03, the commission of IPQ (Instituto Português da Qualidade) for Quality in Information Technology.

https://pt.linkedin.com/in/elsa-cardoso-88092215

**Prof. Dr. Ricardo Ribeiro** has a PhD (2011) in Information Systems and Computer Engineering and an MSc (2003) in Electrical and Computer Engineering, both from Instituto Superior Técnico, and a graduation degree (1996) in Mathematics/Computer Science from Universidade da Beira Interior. From 1999 to 2011, he was a Lecturer at the Lisbon University Institute (ISCTE-IUL), where he is now an Assistant Professor and the Director of the Computer Engineering undergraduate programme. Since 1996, he has been a researcher at INESC-ID Lisboa, Portugal, working in speech and language processing. He is a member of the Spoken Language systems Laboratory of INESC-ID since its creation in 2001. His current research interests focus on the following areas: high-level information extraction from unrestricted text or speech, and improving machine-learning techniques using domain-related information. He has participated in several European and Nationally-funded projects, organized several scientific events and was a member of the program committee of several others, and edited a book on the computational processing of Portuguese.