Topic Modelling for Education in Computing

Arbana Kadriu, Lejla Abazi

Abstract: *This paper is about scientific research analysis on papers with education in computing as main topic of interest. We use Latent Semantic Indexing for topic modelling of scientific research papers collected from e-learning conference proceedings and DBLP library. After the LSI model is trained based on conference articles content, this model is then used to compare article titles obtained from the DBLP library.*

Key words: *topic modelling, latent semantic analysis, education in computing, scientific research analysis*

INTRODUCTION

There are many studies on education in computing, which show various approaches to enhance learning for computing and through computing. A researcher then would be able to be aware of related work in his subfield of interest in the framework of education and computers. But a researcher could be also attracted to the questions: What is going there and what are the trends? What are the most topics discussed when it comes to computing as a tool for better learning and how about education for new generations, so they can better compute?

When it comes to the analysis of scientific papers, most of the work is about analyzing scientific collaboration networks gained from co-authors in different conferences and journal papers. Using this kind of study, results are gained for mean and distribution of numbers of collaborators of authors, demonstrating the presence of clustering in the networks, and highlight a number of apparent differences in the patterns of collaboration between the fields studied [5]. Another approach in analyzing scientific papers is to catch writing style changes by taking into account stylometric features [1]. To develop the necessary tools for exploring and browsing digital articles, there is need for automated methods of organizing, processing, and delivering the results [3]. CiteTextRank is used for key phrase extraction from research articles [9]. Research paper recommender system can be developed generating user models, and calculating content-based recommendations [4]. Content analysis can show various relevant statistics and correlations within and across different research field, based on topic modelling [6].

The work of the research presented in this paper is about investigating how topic modelling can be used to better understand what are the movements and developments toward future education in computing. The extraction of topics from papers is done based on their content or only in their titles.

TOPIC MODELLING

Topic modelling is a text mining methodology to automatically find models of words that appear together in a small or sizeable document, comparing to the other documents in the collection. These obtained clusters of words are then treated as content topics. It tries to identify underlying semantic structure without syntactic parsing of sentences, using bag-of-words model where the order has no importance.

Latent Semantic Indexing is a technique for topic modelling which is based on classical indexing algorithms with deployment of term-document matrix, where rows represent terms and columns represent documents [8]. In this way to each term in a document is assigned a weight for that term that depends on the number of occurrences of the term in the document and in the collection, referred to as term frequency and denoted as $tf_{t,d}$. The document frequency df_t is defined to be the number of documents in the collection that contain a term t, and the inverse document frequency (id_f) of a term t is defined as follows:

$$idf_t = log(N/df_t)) \qquad (1)$$

These two metrics - term frequency and inverse document frequency, are combined to produce a composite weight for each term in each document. The *tf–idf* weighting scheme assigns to term t a weight in document d given by:

$$tf\text{-}idf_{t,d} = tf_{t,d} \times idf_t \qquad (2)$$

The problem is that even for a collection of small size, the term-document matrix C is likely to have several tens of thousands of rows and columns. That’s why an approximation of a term-document matrix of lower rank is needed, which can be achieved using symmetric diagonalization theorem [2]. The low-rank approximation of this matrix gives a new description for every document in the collection. The queries will be represented in this low-rank representation as well, making possible to figure out query/document similarity scores in this low-rank representation. This method is identified as *latent semantic indexing*.

METHODOLOGY AND RESULTS

For the purposes of this research we have used proceedings of the e-Learning’14 conference[20] and also the DBLP Computer Science Bibliography, to gain titles of papers which are about “education in computing”, starting from the year 2010. In this way we got 40 papers, together with their content and about 700 article titles that have to do with education and computing, for the last 5 years. All data are extracted using Python and its package Beautiful Soup, which serves as a tool for HTML and XML content parsing[21].

The papers content is automatically converted to text format (for easier further processing), removing stop words (most frequent words), words that appear only once, numbers and special characters and converting all words to lower case.

For our experimenting intentions we use Gensim, which is a pure Python library that fights on two fronts: 1) digital document indexing and similarity search; and 2) fast, memory-efficient, scalable algorithms for Singular Value Decomposition where the learning is unsupervised, semantic analysis of plain text in digital collections [7].

Firstly we transformed all papers from e-Learning’14 conference to vectors. These vectors are then utilized to automatically gain the corpus vector space and dictionary of the collection, which records the association between the documents and ids used in the corpus. This corpus serves as a basis for initializing the transformation model. We transformed the corpus to the tf-idf model. This model is then serialized into Latent Semantic Indexing model which produces into a latent x-D space, where x is the dimension that we define for our purposes, in this case the number of topics. For example, if we want the main two topics that are treated in the last conference, we write:

```
lsi = models.LsiModel(corpus_tfidf, id2word=dictionary, num_topics=2)
```

And the result from this model are two main topics, defined in this way:

```
'0.139*"social" + 0.126*"lms" + 0.120*"web" + 0.110*"online" + 0.106*"distance" + 0.104*"media" + 0.095*"collaborative" + 0.086*"e-learning" + 0.080*"students" + 0.079*"professional"'

'-0.215*"game" + -0.133*"beer" + -0.125*"mobile" + 0.125*"social" + 0.120*"collaborative" + -0.111*"application" + 0.107*"professional" + 0.101*"lms" + -0.101*"map" + -0.094*"laser"'
```

So the first topic is about “social, media, lms, web, online, distance, media, students, collaborative, e-learning, professional”. We then can find out from our collection of papers, which are the most similar to a topic. In this case, we found out that the highest similarity

[20] http://elearning-conf.eu/index.php?cmd=dPage&pid=proc-14

[21] http://www.crummy.com/software/BeautifulSoup/

with the first topic have papers "New Didactical Models in Open and Online Learning based on Social Media" (similarity of 0.49) and "Engaging Students in Online Courses through Interactive Badges" (similarity of 0.47).

This gained model is persisted and loaded to query the new texts that will come, depending on our needs. In our case, we used the trained model to see where our titles from DBLP library better fit in. In this way, we can now compute similarity between first 100 dblp titles and our conference collection, transforming all of them to vectors, indexing them according the loaded LSI space and then using the similarity interface of the gensim package:

```
for doc in dblp[:100]:
   vec_bow = dictionary.doc2bow(doc.lower().split())
   vec_lsi = lsi[vec_bow]
   index = similarities.MatrixSimilarity(lsi[corpus])
   sims = index[vec_lsi]
   sims = sorted(enumerate(sims), key=lambda item: -item[8])
   print(doc+"\n"+str(sims[0]+"\n")
```

In this way, it is given that in general DBLP titles have high similarity with conference articles, so one deduction is that this conference is treating similar topics in the field of education&computing to those indexed in the DBLP library for the last 5 years.

The similarity results differ from the number of topics that we define. For example, if we define 20 topics, then the title "*Toward an integrated model for designing assessment systems: An analysis of the current status of computer-based assessments in science*" is most similar to the paper "*A Model and an Index for e-Learning Quality Assessment*". With only 2 defined topics, it is the most similar to the paper "*About usability of mobile e-learning applications on an example of an UML-Quiz*".

Another interesting case is the paper "*Storytelling by a kindergarten social assistive robot: A tool for constructive learning in preschool education*" is gained as most similar to the paper "An *Approach to Teaching Introductory Programming Using Games*". We have to note here that similarity is between the content of the conference paper and the title of the DBLP paper. In this situation, the second paper has used the robots for teaching to program using games, which semantically is similar to using robots for another purpose of teaching, this time to preschool children.

CONCLUSIONS AND FUTURE WORK

We have investigated what are the benefits of topic modelling to better understand where researchers are focused on in the field of computing and education. It is shown that topic modelling can be used to compare research works in the area, even when the collection of these works is relatively modest. We plan to further extend this investigation with increased number of full research articles. Also we intend to compare topic modelling of this set with Latent Dirichlet allocation (LDA) model.

REFERENCES

[1] Rexha, A., Klampfl, S., Kröll, M. & Kern, R. Towards Authorship Attribution for Bibliometrics using Stylometric Features. Proceedings of the First Workshop on Mining Scientific Papers: Computational Linguistics and Bibliometrics co-located with 15th International Society of Scientometrics and Informetrics Conference (ISSI 2015).

[2] Manning, C. D., Raghavan, P. & Schütze, H. Introduction to information retrieval. Cambridge University Press, Cambridge, UK, 2008.

[3] D. Blei & J. Lafferty. Topic Models. In A. Srivastava and M. Sahami, editors, Text Mining: Classification, Clustering, and Applications. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, 2009.

[4] Beel, J., Langer, S., Gipp, B. & Nürnberger, A.. The Architecture and Datasets of Docear's Research Paper Recommender System. D-Lib Magazine, ISSN-e 1082-9873, Vol. 20, Nº. 11-12, 2014.

[5] Newman, M. The structure of scientific collaboration networks, PNAS, 98:404409,

the National Academy of Science, 2001.

[6] Paul, M. & Girju, R. Topic modeling of research fields: an interdisciplinary perspective. In Galia Angelova; Kalina Bontcheva; Ruslan Mitkov; Nicolas Nicolov & Nikolai Nikolov, ed., 'RANLP' , RANLP 2009 Organising Committee / ACL, , pp. 337-342.

[7] ŘEHŮŘEK, R. & SOJKA, P. Software Framework for Topic Modelling with Large Corpora. In Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks. Valletta, Malta: University of Malta, 2010. p. 46--50, 5 pp. ISBN 2-9517408-6-7.

[8] Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R. Latent Semantic Indexing. Journal of the American Society for Information Science (1986-1998); Sep 1990; 41, 6; ABI/INFORM Global, pg. 391.

[9] Gollapalli, S. & Caragea, C. Extracting Keyphrases from Research Papers Using Citation Networks. Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014, p. 1629-1635.

ABOUT THE AUTHOR

Assoc. Prof. Arbana Kadriu, PhD, Faculty of Contemporary Sciences and Technologies, South East European University, Phone: +389 44 356 162, E-mail: a.kadriu@seeu.edu.mk.

Assoc. Prof. Lejla Abazi, PhD, Faculty of Contemporary Sciences and Technologies, South East European University, Phone: ++389 44 356 178, E-mail: l.abazi@seeu.edu.mk.

The paper has been reviewed.