Proceedings of the Federated Conference on Computer Science and Information Systems pp. 287–290
DOI: 10.15439/2016F156
ACSIS, Vol. 8. ISSN 2300-5963

# Automatic Keyword Extraction from Medical and Healthcare Curriculum

Martin Komenda[a], Matěj Karolyi[a],
Andrea Pokorná[a,b]
[a] Institute of Biostatistics and Analyses,
[b] Department of Nursing,
Faculty of Medicine, Masaryk University,
Kamenice 126/3, 625 00
Brno, Czech Republic
Email: {komenda, karolyi, pokorná}@iba.muni.cz

Martin Víta
NLP Centre,
Faculty of Informatics,
Masaryk University,
Botanická 68a, 602 00
Brno, Czech Republic
Email: 333617@mail.muni.cz

Vincent Kríž
Faculty of Mathematics and Physics
Charles University
Malostranské nám. 25, 118 00
Prague, Czech Republic
Email: kriz@ufal.mff.cuni.cz

*Abstract*—**Medical and healthcare study programmes are quite complicated in terms of branched structure and heterogeneous content. In logical sequence a lot of requirements and demands placed on students appear there. This paper focuses on an innovative way how to discover and understand complex curricula using modern information and communication technologies. We introduce an algorithm for curriculum metadata automatic processing -- automatic keyword extraction based on unsupervised approaches, and we demonstrate a real application during a process of innovation and optimization of medical education. The outputs of our pilot analysis represent systematic description of medical curriculum by three different approaches (centrality measures) used for relevant keywords extraction. Further evaluation by senior curriculum designers and guarantors is required to obtain an objective benchmark.**

## I. Introduction

THE domain of medical and healthcare curriculum harmonization captures the systematic transmission of specialized required knowledge based on a suitable combination of theoretically focused courses and clinical teaching training [1]. It links traditional proven pedagogical approaches and medical expertise with computer sciences and technologies with a view to the discovery of an innovative way to understand the complex structure and content of medical and healthcare study programmes. The proper use of data extracted from curriculum management systems can significantly improve the global overview on entire curriculum including performing up-to-date information in real time. The attention of this paper is paid to the core technologies of automatic processing for documents classified as a supervised machine learning task called automatic keyword extraction, and its real application on curriculum metadata. Automatic Keyword Extraction (AKE) is a research topic focusing on the identification of a small set of words, key phrases, keywords, or key segments from a document that can describe the meaning of the document [2]. The aim of keyword assignment is to find a small set of terms that appropriately describes a specific document, a medical discipline in our particular case, independently of the domain it belongs to. Here are two fundamental issues, which are supposed to be answered by our research: (i) Are we able to automatically generate sets of keywords? Are the extracted keywords relevant? Do they express properly individual medical discipline? (ii) Which kind of centrality does identify the most accurate set of keywords and what is the proper value of determined threshold?

## II. Methods

The research methods stems from our previously published and reviewed papers [3]–[5], wherefrom the well-known and proven step-by-step data mining guide CRISP-DM (CRoss-Industry Standard Process for Data Mining) [6]. We have used the CRISP-DM reference model, which was primarily developed by means of the effort of a consortium from the business sphere, as a powerful scientific technique to excavate the knowledge and patterns concealed in the diversely complex mountain of medical and healthcare education data [7]. This standardized methodology provides the complete process model for exploring data. Below, all the CRISP-DM steps are described in accordance with determined research questions.

### A. Business understanding

In this stage, we focus on the understanding the research objectives from the perspective of curriculum mapping. In general, curriculum mapping is a procedure that creates visual representation of the curriculum based on real time information, as a way to increase collaboration and collegiality in higher education institutions. This phase also involves more detailed fact-finding about all of the resources, constraints, assumptions, published results and other factors that should be considered in determining the goals analysis [8].

Two main objectives were identified: (i) to make the curriculum more transparent to all stakeholders; (ii) to demonstrate the links between the various components of the curriculum (such as modules, disciplines, courses and learning units) [9]. In terms of new trends and reforms in medical education, we have proposed a general model for curriculum management and harmonization supported by our

978-83-60810-90-3/$25.00©2016, IEEE

original web-oriented platform called OPTIMED[1] [10]. This innovative and dynamic platform provides a clear way how to describe medical curriculum with the use of given text attributes including links to relevant study materials. We set up the structure of curriculum, which covers study programmes (e.g. General Medicine), specialized modules (e.g. Theoretic sciences), medical disciplines (e.g. Anatomy), courses (e.g. Anatomy I – lecture), and learning units (e.g. Central nervous system). This phase involves more detailed fact-finding about a systematic keyword generation of individual medical disciplines from various metadata sources stored in the OPTIMED database.

For the pilot experiment two disciplines were chosen (Nursing and Psychiatry), which are both used for the acquisition of knowledge and skills and also allow the development of so-called "soft skills". Both disciplines are taught at different periods of study. While Nursing has been lectured almost on the beginning of the medical study (in the second year) and is focused on basic knowledge and skills in caring (helping with basic needs of patients and also caring for patients with different types of illness). Psychiatry is taught as one of the last courses of undergraduate study of general medicine. Both courses include theoretical instruction in the classroom and subsequently clinical placement where students should use knowledge and skills obtained in the previous tuition and both of them are not accepted as the core medical disciplines.

### B. Data understanding

Data understanding starts with initial data collection and proceeds with activities that enable to become familiar with the data, to evaluate the quality of data, to discover first insights into the data, and/or detect interesting subsets to form hypotheses regarding hidden information. The following metadata attributes describing medical curriculum was mined from PostgreSQL database and divided into the classes in accordance with fields of medicine. Names of the attributes, data types in squared brackets and samples are shown below in Table I.

TABLE I.
SELECTED ATTRIBUTES OF A CURRICULUM

| Attribute (data type) | Sample value |
|---|---|
| Name of learning unit (varchar) | Biologic therapy in psychiatry |
| Importance of learning unit (text) | The aim of the study unit is to introduce types of biological therapy in psychiatry, including psychopharmacotherapy, electroconvulsive therapy, Recently, modern neurostimulative methods such as repetitive transcranial magnetic stimulation … |
| Description of learning unit (text) | Psychopharmaceuticals can be classified in several ways. Lehman s diversification is often used and is based on effects on three mental functions: vigilance, effectivity and mental integration (thinking). Vigilance is positively influenced by nootropics, cognitives and psychostimulants, negatively by hypnotics… |
| Group learning outcome (varchar) | Main indications, differences from adults (indications, efficacy), adverse effects, ethical aspects. |
| Index (varchar) | Psychopharmacological drugs in relation to children, Antipsychotics, antidepressants, stimulants, thymoprophylactics. |
| Learning outcome (varchar) | Student knows key indicators relating to psychopharmacological drugs. |

### C. Data preparation

Data preparation covers all activities needed to construct the final dataset from the initial raw data. First of all, table, record and attribute selection as well as data pre-processing covering transformation and data cleaning procedure were automatically done. The input data set consists of information-rich attributes (name, importance, description of learning units and all related learning outcomes including indexes) mined from OPTIMED. The data preparation phase appears again at the end of the modeling process described in the next section. An output of an algorithm for keyword extraction is very compact list of keywords including just the self-sorting information. For the purposes of plotting the graph the simplified table combined with products is created during the calculation. This final data connection is executed every time when a user starts to explore new discipline keyword dataset.

### D. Modeling

In this section, we propose a novel algorithm for keyword extraction that forms the basis of the modeling and visualization stages, which are also introduced here.

The algorithm is based on two main components: word2vec model and network/graph centralities. Word2vec model proposed by Mikolov [11] is a word embedding model that belong to a wide class of distributed representations models. It arises from predicting the neighbors of words using a deep neural network – roughly said, the weights in the neural network between input and hidden layer constitute the vector representations of words. For our purposes, we have created a word2vec model over TC wikipedia[2] using the following main parameters: model = CBOW, dimension = 200. The concept of selected network/graph centrality measures was firstly developed in social network analysis. We successfully used these methods in our previous work [3], [4], namely we deal with the closeness, betweenness, and eigenvector centralities.

Input data for the algorithm are represented by trained word2vec model, given document (bag of words for individual medical disciplines), word similarity threshold `t`, number of keyword `n`. (1) Select all terms that are contained in the document at least two times. (2) For all pairs of terms, compute the cosine similarity of their word2vec representations. (3) Create the graph `G` in a following way: Vertices are the terms (obtained in the Step 1). Two vertices are connected with an edge if and only if their mutual cosine

[1] http://opti.med.muni.cz/en/

[2] http://nlp.cs.nyu.edu/wikipedia-data/

similarity is at least t. (4) Compute closeness, betweenness and eigenvector centrality of vertices in the graph G. Select n terms with the highest values of various centrality and set them up as keywords.

Finally, we have implemented a new online dynamic visualization as a feature of the OPTIMED reporting tools[3], which is based on outputs from described algorithm above. Innovative graphical interpretation allows users to create complex data overview, which cannot be achieved with basic data views. In our case, the simple network visualization was created with the D3.js JavaScript library. We decided to integrate a force-directed graph (see Fig. 1) displaying complex structures with expand-on-demand clusters and convex hulls around leaf nodes [12].

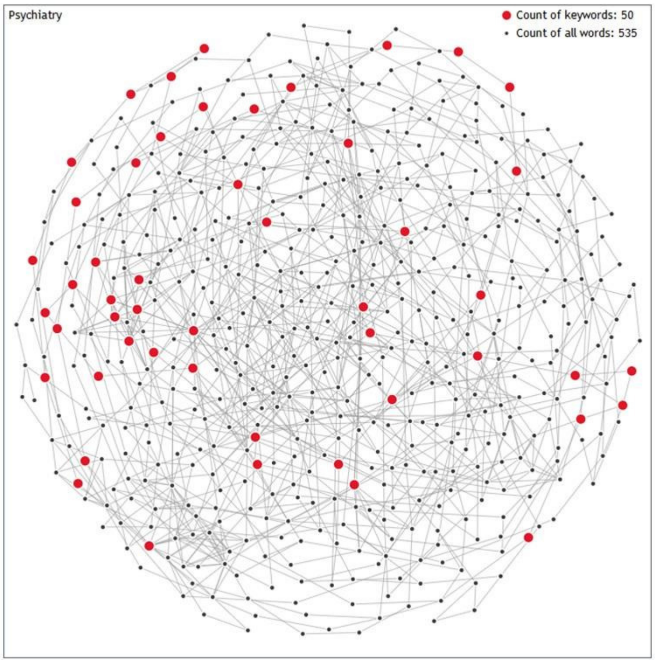
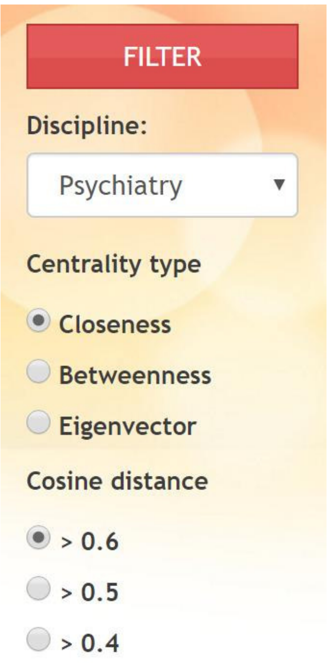


Fig. 1 The force-directed graph of Psychiatry (based on closeness centrality, cosine similarity is higher than 0.6)

The graphs include two types of keyword representation – red nodes (bigger) for more frequent keywords and black nodes (smaller) with less frequent keywords. Moreover, we wanted to determine what kind of centrality type (closeness, betweenness or eigenvector) and what value of the cosine similarity (> 0.6; > 0.5; or > 0.4) would be the most suitable and could the most accurately describe concrete medical discipline. Using advanced filters included in web reporting service, we are able to modify the final visual interpretation of available preprocessed data in accordance with selected parameters – medical discipline, centrality type, cosine similarity threshold (see Fig. 2).



Discipline:

Psychiatry ▾

Centrality type

● Closeness

○ Betweenness

○ Eigenvector

Cosine distance

● > 0.6

○ > 0.5

○ > 0.4

Fig. 2 The force-directed graph filter

### E. Evaluation and Deployment

A checking procedure is performed in this stage in order to find the right meaning of analytical outputs. The obtained results were verified by representatives of the faculty management, in order to confirm the final interpretation [3]. Based on the assessment of the graphical presentation of the subject Psychiatry we decided that the best threshold for cosine similarity is 0.6 (see Fig. 1) and the most suitable centrality type is closeness. Table II shows keywords concerning different type of centrality. As you can see the first column (type of centrality – closeness) includes keywords especially in view of the clinical description of psychiatric illness, the second column (type of centrality – betweenness) cannot be held as summary of clear signs of typical keywords rather confusing mix of terms and third column (type of centrality – eigenvector) includes keywords focused on describing the pathophysiology and chemical substances important in Psychiatry, followed by keywords of clinically important symptoms of disorders as well as therapeutic modalities and drugs. So, different type of centrality enables to view the discipline from different perspectives through the different sample of keywords. What has to be highlighted that when we used higher cosine similarity we could identify less duplicate terms and terms which are not describing the key issues (as nouns or verbs) and we could identify them as „stop-words“ without any informative value.

[3] http://opti.med.muni.cz/en/reporting/web/analyticke-reporty/extrakce-klicovych-slov/dis-44

TABLE II.
TOP TWENTY KEYWORDS FOR PSYCHIATRY ACCORDING TO INDIVIDUAL CENTRALITIES.

| Rank | Closeness | Betweenness | Eigenvector |
|---|---|---|---|
| 1 | disorders | neurological | dopamine |
| 2 | neurological | dysfunction | neurotransmitter |
| 3 | schizophrenia | stimulation | acetylcholine |
| 4 | disorder | inhibition | serotonin |
| 5 | anxiety | serotonin | neurotransmitters |
| 6 | psychological | anxiety | receptors |
| 7 | symptoms | antidepressant | receptor |
| 8 | dysfunction | brain | neurons |
| 9 | psychopathology | cortex | noradrenaline |
| 10 | brain | disorders | Presynaptic |
| 11 | illnesses | symptoms | inhibition |
| 12 | epilepsy | disorder | reuptake |
| 13 | mental | neurons | neuron |
| 14 | cognitive | cognitive | cells |
| 15 | stimulation | psychological | hippocampus |
| 16 | behavioral | schizophrenia | antidepressant |
| 17 | pathological | perception | amygdala |
| 18 | clinical | clinical | cortex |
| 19 | psychiatric | particular | prefrontal |
| 20 | emotional | emotional | benzodiazepines |

## III. Discussion

The possibility to graphically visualize and interpret medical curriculum and content of the individual courses through keywords from different scientific disciplines allow teachers to identify the main issues discussed in the concrete tuition. The identification could help them to explore whether the main keywords correspond to any other keywords as descriptors of the main themes or major topics representing individual similarly or differently oriented teaching units. We have used the dynamic visualization using force-directed graph with clusters of key points (point graphs) and tables while we have identified 50 keywords as a border (maximum evaluated number of keywords for one visualization).

We could say that generally description of Nursing disciplines is not so straightforward, clear and simply as for Psychiatry. This can be explained mainly by the fact that the content of the discipline is very inhomogeneous. From the description of the summary of keywords for nursing evidently arise higher amount of duplicated terms and "empty" or so-called "stop words", which do not bring additional value for improvement of orientation in the curriculum content not only for teachers neither for students. This fact should not affect the view of usability of curriculum content visualization but rather to encourage the prudent use and analyses of the information gathered and from data mining and also in feedback when assessing the homogeneity of data.

## IV. Conclusion and future work

In this work we have proposed a novel method for identifying key terms from a free text covering medical curriculum. This method is based on computing node centralities in a similarity graph of terms contained in the given medical and healthcare discipline description, whereas the similarity is obtained by a popular word2vec model. First results seem to be promising in terms of face validity. Methodologically sound evaluation of this method and comparison with traditional methods of keyword extraction - even on different domains - is a current issue. Centrality measures in word2vec graphs can also serve as features in supervised machine learning algorithms for keyword extraction. In the near future this approach is also planned to be investigated.

## Acknowledgment

The authors were supported from the following grant projects: (i) CROESUS – clinical reasoning skills enhancements with the use of simulations and algorithms – reg. no.: 2014-1-CZ01-KA203-002002, which is funded by the European Commission ERASMUS+ program; (ii) MEDCIN – medical curriculum Innovations – reg. no.: 2015-1-CZ01-KA203-013935, which is funded by the European Commission ERASMUS+ program; and (iii) SVV – specific academic research – reg. no.: 260 333, which is funded by the Charles University in Prague.

## V. References

[1] M. Komenda, "Towards a Framework for Medical Curriculum Mapping," Doctoral thesis, Masaryk University, Faculty of Informatics, 2015.

[2] A. Hulth, "Improved automatic keyword extraction given more linguistic knowledge," in Proceedings of the 2003 conference on Empirical methods in natural language processing, 2003, pp. 216–223.

[3] M. Víta, M. Komenda, and A. Pokorná, "Exploring medical curricula using social network analysis methods," presented at the Federated conference on computer science and information systems, 5th International Workshop on Artificial Intelligence in Medical Applications, Lodz, 2015, doi:10.15439/2015F312.

[4] M. Komenda, M. Víta, C. Vaitsis, D. Schwarz, A. Pokorná, N. Zary, and L. Dušek, "Curriculum Mapping with Academic Analytics in Medical and Healthcare Education," PloS One, vol. 10, no. 12, 2015, doi:10.1371/journal.pone.0143748.

[5] M. Komenda, D. Schwarz, J. Švancara, C. Vaitsis, N. Zary, and L. Dušek, "Practical use of medical terminology in curriculum mapping," Comput. Biol. Med., 2015, doi: 10.1016/j.compbiomed.2015.05.006.

[6] A. I. R. L. Azevedo, "KDD, SEMMA and CRISP-DM: a parallel overview," 2008.

[7] S. C. Chen and M. Y. Huang, "Constructing credit auditing and control & management model with data mining technique," Expert Syst. Appl., vol. 38, no. 5, pp. 5359–5365, May 2011, doi: 10.1016/j.eswa.2010.10.020.

[8] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "CRISP-DM 1.0 Step-by-step data mining guide," 2000.

[9] R. M. Harden, et al, "Curriculum mapping: a tool for transparent and authentic teaching and learning". Association for Medical Education in Europe, 2001, doi: 10.1080/01421590120036547.

[10] M. Komenda, D. Schwarz, J. Hřebíček, J. Holčík, and L. Dušek, "A Framework for Curriculum Management - The Use of Outcome-based Approach in Practice," in Proceedings of the 6th International Conference on Computer Supported Education, Barcelona, 2014, doi:10.5220/0004948104730478.

[11] T. Mikolov, W. Yih, and G. Zweig, "Linguistic Regularities in Continuous Space Word Representations.," in HLT-NAACL, 2013, pp. 746–751.

[12] M. Bostock, V. Ogievetsky, and J. Heer, "D3 data-driven documents," Vis. Comput. Graph. IEEE Trans. On, vol. 17, no. 12, pp. 2301–2309, 2011, doi: 10.1109/TVCG.2011.18.