Decision Support Systems and Education
J. Mantas et al. (Eds.)
© 2018 The authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/978-1-61499-921-8-222

Similarity Detection Between Virtual Patients and Medical Curriculum Using R

Martin KOMENDA[a,1], Jakub ŠČAVNICKÝ[a], Petra RŮŽIČKOVÁ[a], Matěj KAROLYI[a], Petr ŠTOURAČ[b] and Daniel SCHWARZ[a]
[a] *Institute of Biostatistics and Analyses, Faculty of Medicine, Masaryk University, Czech Republic*
[b] *Department of Paediatric Anaesthesiology and Intensive Care Medicine, Faculty of Medicine, Masaryk University, Czech Republic*

Abstract. This paper presents the domain of information sciences, applied informatics and biomedical engineering, proposing to develop methods for an automated detection of similarities between two particular virtual learning environments – virtual patients at Akutne.cz and the OPTIMED curriculum management system – in order to provide support to clinically oriented stages of medical and healthcare studies. For this purpose, the authors used large amounts of text-based data collected by the system for mapping medical curricula and through the system for virtual patient authoring and delivery. The proposed text-mining algorithm for an automated detection of links between content entities of these systems has been successfully implemented by the means of a web-based toolbox.

Keywords. text similarity, virtual patient, akutne.cz, medical curriculum, OPTIMED, R programming language

1. Introduction

Economic pressures and social influences have steadily escalated training expectations as regards the education of medical and health professions, while training resources have been on the decline worldwide [1]. Moreover, due to the need for highly qualified professionals entering into clinical practice, students are increasingly required to acquire a sufficient set of competencies prior to providing patient care. Correctly compiled and well-balanced medical and healthcare curricula, which combine theoretically focused courses and a clinical teaching base, are essential prerequisites to acquire these minimum required skills and knowledge [2]. Today, there is a plenty of virtual learning environments based on specific didactic models [3]. They typically improve the individual study process by offering various functionalities such as curricula and courses management, repositories for learning objects, tools for learners' assessment, mass online communication and collaboration channels and many more [4]. We have adopted one of these concepts, namely serious games and virtual patients (VPs), which is known as a simulation-based technology using games or various scenarios for practising the application of previously acquired knowledge. In general, VPs represent interactive web-based clinical scenarios or simulations in the training of health professionals, which can

[1] Corresponding author, Martin KOMENDA, Institute of Biostatistics and Analyses, Faculty of Medicine, Masaryk University, Brno, 625 00, Czech Republic; E-mail: komenda@iba.muni.cz.

provide highly effective ways of addressing the limited student access to real patients, the need for standardised and well-structured educational patient encounters, and opportunities for students to practice in safe and responsive environments [5,6].

This paper describes the development and implementation of an algorithm providing an automated text analysis on a set of VPs from Akutne.cz[2] and a parametric description of the General Medicine curriculum created in OPTIMED[3]. This pilot project follows from the author's previously published work, where fundamental aspects of Akutne.cz [7] and OPTIMED [8] platforms as well the basis of a curriculum mapping framework [9-11] were emphasised. The following research questions were formulated in order to define and subsequently solve two particular research problems: (i) How to measure similarities (based on keywords extraction) between given VPs and particular building blocks of the curriculum? (ii) How to visualise the identified relations/overlaps/links with the use of an interactive web-based application? This paper suggests answers to both questions.

2. Methodological Background

2.1. Data Sets and Process Model

For the pilot purposes, sets of data in English language extracted from the OPTIMED curriculum management system used at the Faculty of Medicine of the Masaryk University (1,360 of learning units described by approximately 2,600 standard pages of text in total) and from the Akutne.cz platform for virtual patients were used to verify quantitatively the accuracy of developed algorithms (77 of virtual patients described by approximately 550 standard pages of text in total). In order to guarantee a systematic approach to solution of the above-mentioned research questions, a standardised data-mining methodology knows as CRISP-DM [12] (CRoss-Industry Standard Process for Data Mining) was adopted. It employs a six-phase process model of the life cycle of a data-mining project. Its individual phases can be briefly outlined as follows: (i) Business understanding focuses on the project goals and the definition of research questions from the particular perspective of medical curriculum mapping. (ii) Data understanding starts with initial data collection and data retrieval from databases. (iii) Data preparation covers all activities needed to construct the final dataset from the initial raw data: table, record and attribute selection, data aggregation and data pre-processing. (iv) Modelling represents the use of selected techniques for vector-space representation, computation of similarities and creation of interactive graphs. (v) Evaluation covers the review of achieved results by an expert board. (vi) Deployment summarises the strategy how to perform an automated detection of interconnections between VPs and curriculum components by means of a web-based toolbox.

2.2. Text Mining Techniques and Technological Mix

Measuring the similarity between textual records is a very challenging issue. Several different techniques are used to compare words, sentences, paragraphs and documents [13]; these techniques are mainly based on similarities between strings,

[2] http://www.akutne.cz/index-en.php

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

corpuses and even knowledge bases. In order to build a pilot prototype on real educational data, we opted for the string-based similarity of terms (the frequency-based approach) with the normalised Pearson correlation coefficient as the similarity measure. This coefficient is frequently used in text comparisons because it is one of the simplest to implement. Moreover, according to text clustering analyses [14,15], the normalised Pearson correlation coefficient yields comparable results to Jaccard, cosine or Dice coefficients.

Firstly, all documents from both data sources were pre-processed: HTML tags, punctuation, digits, special characters and words shorter than three characters were systematically removed; afterwards, specific words using both Google stop-word list and a customised stop-word list were removed. Secondly, word frequencies for each single document (virtual patient, learning unit) were counted. Thirdly, similarities between individual virtual patients and learning units were computed. And finally, the five most similar learning units were selected for each virtual patient based on the computed measure of similarities between documents. These data were further visualised using network graph, similarity graph and data table.

Our development toolkit is based on the R programming language and its software tools. Data pre-processing, analyses and visualisations were performed in R using special libraries and packages (READXL, DPLYR, TM, PROXY, VISNETWORK, DATA.TREE, DIAGRAMMER, SHINY, SHINYDASHBOARD). Afterwards, an interactive web-based visualisation toolbox was developed by means of the R Shiny package, which is available on CRAN and makes it easy to build a standalone online application, which benefits from the computational power of R and the interactivity of the modern web.

3. Results

The final web-based toolbox[4], which is freely accessible on the Internet, provides an algorithm for the computation of similarities between two input sets of data – virtual patients and learning units. Three different views (network graphs, similarity graphs and data table) were developed and implemented using R and aforementioned libraries for interactive visualisations. The network graph (see Fig. 1) shows a general overview of links for all existing relations between VPs and related learning units. The user can apply a node filter, view tooltips and zoom in/out in order to explore a particular cluster of nodes. The similarity data table (see Fig.1) contains a list of similar items including searching and sorting functions. The similarity graph (see Fig. 2) visualises the five most related learning units based on similarity measures for each virtual patient. The above-described pilot prototype has been placed to the context of the OPTIMED Reporting module[5]. This module contains reports with analyses of selected parts of the OPTIMED curriculum management system as well as user activities and behaviour monitoring. The Reporting module is divided into five main submodules: (i) Summary overviews, (ii) Content of teaching, (iii) Volume of teaching, (iv) Information about users and (v) Analytic reports. The output of this pilot study (i.e. a fully interactive set of visualisations based on R Shiny package) is available in the fifth submodule and is called the "MERGER virtual patients analysis".

[4] http://opti.med.muni.cz/en/reporting/web/analyticke-reporty/analyza-virtualni-pacient/

[5] http://opti.med.muni.cz/en/reporting/web/

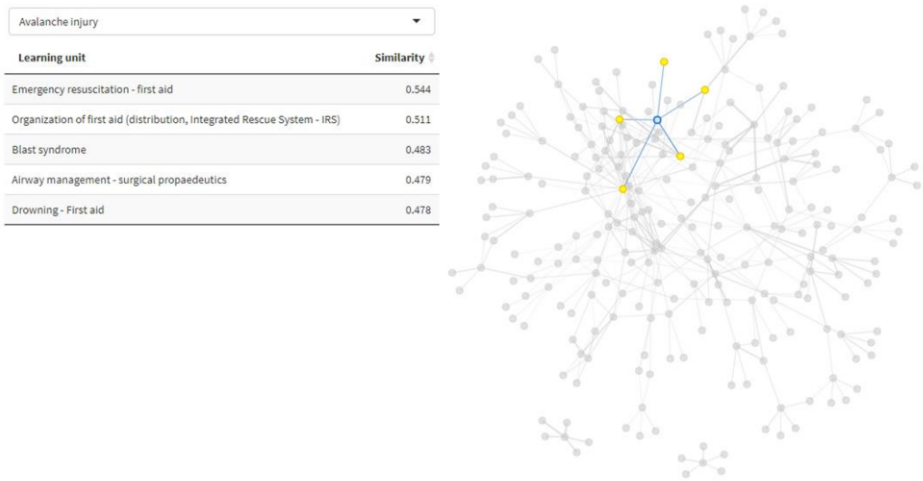

Figure 1. This interactive network graph shows 77 virtual patients from Akutne.cz and all related learning units from the OPTIMED portal (a zoom in/out function is available). A virtual patient called "Avalanche injury" is highlighted by a blue node, the most relevant learning units are highlighted by yellow nodes.

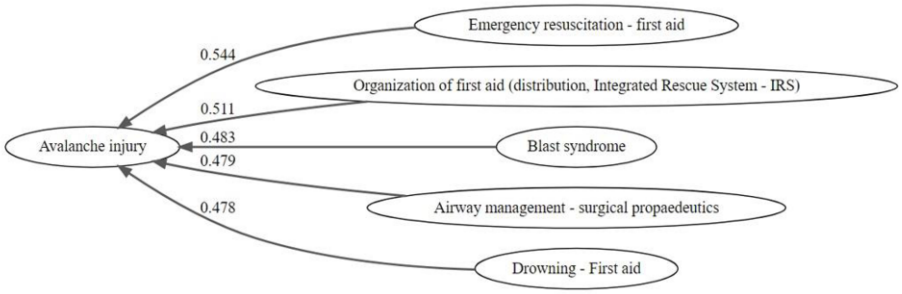

Figure 2. Similarity graphs shows the five most relevant learning units to a virtual patient called "Avalanche injury".

4. Discussion

Two research questions were formulated in the introduction to this paper. The answers help to understand how the presented approach can be used in real practice. (i) How to measure similarities (based on keywords extraction) between given VPs and particular building blocks of the curriculum? In accordance with the CRISP-DM model, we were able to analyse this problem, to understand the domain of textual similarity measurement, to design and to run an effective pilot algorithm for processing educational data in R, to use a proper similarity measure (the normalised Pearson correlation coefficient) and, finally, to get the most relevant content (i.e. learning units) for input data (i.e. virtual patients). (ii) How to visualise the identified relations/overlaps/links with the use of an interactive web-based application? Using the R Shiny package, we were able to develop and to implement an interactive web-based visualisation with CSS themes, htmlwidgets and JavaScript actions, which has been embedded to the Reporting module of the OPTIMED platform. The evaluation part of achieved results has already been initiated,

currently being carried out by senior curriculum designers and guarantors of the Faculty of Medicine of the Masaryk University. In the near future, we plan to implement and to evaluate different similarity measures and other textual document comparison approaches. Specifically, we plan to look into similarity measurement based on Jaccard, cosine or Dice coefficients, the character-based N-gram approach, machine learning classification methods and a PostgreSQL extension called pg_trgm. The comparison of all these pilot results will identify the most accurate method for an automated detection of similarities between virtual patients and a medical curriculum.

Acknowledgement

The authors were supported from the following grant projects: (i) MERGER project – Reg. No. MUNI/A/1339/2016 funded from the Grant Agency of the Masaryk University; (ii) Masaryk University Strategic Investments in Education SIMU+ (CZ.02.2.67/0.0/0.0/16_016/0002416) funded from the European Regional Development Fund; (iii) Masaryk University 4.0 (CZ.02.2.67/0.0/0.0/16_015/0002418) funded from the European Social Fund.

References

[1] A. Cook David, M. Triola Marc, Virtual patients: a critical literature review and proposed next steps, *Med. Educ* **43** (4) (2009), 303–311.
[2] M. Komenda, *Towards a Framework for Medical Curriculum Mapping*, Doctoral thesis, Masaryk University, Faculty of Informatics, 2015.
[3] T. A. Mikropoulos, A. Natsis, Educational virtual environments: A ten-year review of empirical research (1999–2009), *Comput Educ* **56** (3) (2011), 769–780.
[4] E. M. van Raaij, J. J. L. Schepers, The acceptance and use of a virtual learning environment in China, *Comput Educ* **50** (3) (2008), 838–852.
[5] D. A. Cook, P. J. Erwin, M. M. Triola, Computerized Virtual Patients in Health Professions Education: A Systematic Review and Meta-Analysis, *Acad. Med*. **85** (10) (2010), 1589–1602.
[6] R. Ellaway, T. Poulton, U. Fors, J. B. McGee, S. Albright, Building a virtual patient commons, *Med. Teach.* **30** (2) (2008), 170–174.
[7] D. Schwarz, P. Štourač, M. Komenda, H. Harazim, M. Kosinová, J. Gregor, R. Hůlek, O. Smékalová, I. Křikava, R. Štoudek, L. Dušek, Interactive Algorithms for Teaching and Learning Acute Medicine in the Network of Medical Faculties MEFANET, *J. Med. Internet Res.* **15** (7) (2013).
[8] M. Komenda, M. Víta, Ch. Vaitsis, D. Schwarz, A. Pokorná, N. Zary, L. Dušek, Curriculum Mapping with Academic Analytics in Medical and Healthcare Education, *PloS One* **10** (12), 2015.
[9] M. Karolyi, M. Komenda, R. Janoušová, M. Víta, D. Schwarz, Finding overlapping terms in medical and health care curriculum using text mining methods: reha, *MEFANET J.* **4** (2) (2017), 71–77.
[10] M. Komenda, M. Karolyi, R. Vyškovský, K. Ježová, J. Ščavnický, Towards a keyword extraction in medical and healthcare education, *2017 Federated Conference on Computer Science and Information Systems (FedCSIS)* (2017), 173–176.
[11] R. Randell, R. Cornet, C. McCowan, *Informatics for Health: Connected Citizen-Led Wellness and Population Health*, IOS Press, 2017.
[12] A. I. R. L. Azevedo, *KDD, SEMMA and CRISP-DM: a parallel overview*, 2008.
[13] W. H. Gomaa, A. A. Fahmy, A survey of text similarity approaches, *Int J Comput Appl* **68** (13) (2013).
[14] N. Sandhya, Y. S. Lalitha, A. Govardhan, K. Anuradha, Analysis of similarity measures for text clustering, *CSC J.* **2** (2008).
[15] A. Huang, Similarity measures for text document clustering, in *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008)*, Christchurch, New Zealand, 2008, 49–56.