Proceedings of the Federated Conference on Computer Science and Information Systems pp. 109–116
DOI: 10.15439/2018F163
ISSN 2300-5963 ACSIS, Vol. 15

Pitfalls in users' evaluation of algorithms for text-based similarity detection in medical education

Jakub Ščavnický, Matěj Karolyi, Petra Růžičková, Andrea Pokorná,
Hana Harazim, Petr Štourač, Martin Komenda
Faculty of Medicine, Masaryk University, Kamenice 5, 625 00, Czech Republic
Email: {scavnicky, karolyi, ruzickova, pokorna, komenda} @iba.muni.cz, hana.harazim@gmail.com, stourac@med.muni.cz

Abstract—**This paper introduces a user evaluation of several approaches for an automated similarity detection between study materials and curriculum description in the field of medical and healthcare education. Our objective is to present an effective methodology of getting relevant feedback from medical students and teachers. Two various data sets (electronic study materials represented by interactive educational algorithms on the AKUTNE.CZ platform and the curriculum of the General Medicine study programme) are processed. For the purposes of this work, text similarity between two data sets is expressed lexically, i.e. character-based (n-gram) similarity as well as term-based similarity methods are used. We present the comparison of five selected approaches to similarity calculation as well as an objective discussion covering our experience with and pitfalls of user evaluation.**

I. Introduction

Medical and healthcare studies cover a variety of useful information and sources used for learning and teaching leading to professional development. In general, any high-quality education requires that materials guaranteed by experts are available; these materials then constitute the curricula of individual study programmes. In the period between matriculation and graduation, students face a large amount of knowledge and skills to be acquired, which is repetitively emphasised in lectures, seminars and clinical practices. By way of illustration, the General Medicine master's degree programme at the Faculty of Medicine of the Masaryk University contains around 150 obligatory courses which are described by approximately 1,200 events (learning units) and 7,000 competency objects (learning outcomes); in total, this makes up more than 2,500 pages of text. Moreover, each of above-mentioned courses has a set of recommended study materials which are available either in the printed form (scripts/textbooks, atlases, monographs etc.) or in the electronic form (presentations, virtual patients/interactive educational algorithms, educational websites, etc.). With respect to human cognitive abilities, it is virtually impossible to carefully read and remember every single detail of all learning units and book chapters, including their linkages and co-dependencies [1]. This paper picks up the threads of the authors' previously published work, where the development and implementation of modern interactive tools [2], [3], as well as a complex analysis and mapping of medical and healthcare curricula [4]–[7], were introduced. There is also given a proposal of several approaches for an automated similarity detection between study materials and curriculum description in the field of medical education, including the evaluation of achieved results by users. The authors strived to get relevant feedback from medical students and teachers in terms of a systematic and objective evaluation of links between a given virtual patient and particular building blocks (learning units) of the curriculum. The following research questions were formulated in order to define and subsequently solve a particular research problem: What is the relation between the achieved results in a form of detected similarities done by computer and an evaluation by users (medical student and teachers)? Which approach of similarity detection can be effectively implemented in a particular domain of medical education?

II. Methods

A. Input data set

For the purposes of similarity detection between medical education data, we decided to process two various data sets: (i) electronic study materials represented by interactive educational algorithms on the AKUTNE.CZ platform[1] (77 virtual patients described by approximately 550 standard pages of text in total) and (ii) the curriculum of the General Medicine study programme taught at the Faculty of Medicine of the Masaryk University, represented by a full metadata description on the OPTIMED curriculum management system[2] [3] (1,232 learning units described by approximately 2,600 standard pages of text in total). Both input data sets were prepared in English language in order to eliminate problems related to a rather complicated morphology of the Czech language. As for the evaluation by users, we chose a subset of 16 learning units of a course entitled "Diagnostic imaging methods", which provides an introduction to the study of nuclear medicine, more specifically the study of radiology and imaging methods, including CT, MR, X-ray, angiography and ultrasound. All of these units are fully described by all mandatory and optional metadata, covering one complete topic and one complete course in the fourth year of study of the General Medicine. This choice of this particular course was consulted with senior experts in a field of medical education because some general overlapping

This work was not supported by any organization

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

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

topics and areas with interactive algorithms were expected here. One of the main motivation given by senior teachers to select this special area was the fact that imaging methods presuppose sufficient image documentation to be used in interactive educational algorithms. Moreover, the quality and length of metadata description of all above-mentioned learning units were sufficient in terms of text-based analysis.

B. Similarity calculation

The similarity of text documents can be understood in two different ways – either semantical or lexical. The former refers to similarity in meaning and used context, whereas the latter represents similarity of character sequences. In this pilot study, we understand text similarity lexically. According to [8], we can classify lexical or string-based text similarity methods into character-based groups and term-based groups. The n-gram method is one of the character-based methods introduced in this pilot study. On the other hand, the term-based methods were implemented using several string measures – the normalised Pearson correlation, the cosine similarity, the extended Jaccard coefficient and the Dice coefficient. Each of used methods is briefly described in the following paragraphs.

C. Character based (n-gram) similarity

The comparison through a pg_trgm module (a PostgreSQL database engine that can be installed into an existing database using a simple SQL command, namely CREATE EXTENSION IF NOT EXISTS pg_trgm) is one of the approaches we used to calculate the measure of similarity between selected learning units at the OPTIMED portal and interactive educational algorithms at AKUTNE.CZ. This extension of the PostgreSQL standard database provides functions and operators to compare the similarities and distances of input text strings. Generally speaking, the pg_trgm is an n-gram (character-based) algorithm for similarity measurement [8]. In this case, N is equal to three and therefore, the measuring unit is called a trigram. In other words, a trigram is a group of three consecutive characters taken from an input string.

We are able to measure the similarity of two strings by counting the number of shared trigrams (there is a similarity to an ASCII alphanumeric text based on trigram matching). This simple idea turns out to be very effective for measuring the similarity of words in many natural languages (e.g. English) [9]. For example, the set of trigrams in the word "pet" is following: "p", " pe", "pet", "et " (the algorithm takes the input word with two spaces prefixed and one space suffixed). We expect that also in professional terminology, these similarities would be quite easily identifiable.

The pg_trgm module provides four functions and two operators. For our purposes, the function called similarity is the most interesting one, taking two strings to be compared. The function *similarity(text, text)* returns a number between 0 and 1 which indicates how similar the two inputs strings are: zero means that the two strings are completely different, whereas one indicates that the strings are completely identical. In the next step, the operator *text* <-> *text* returns the distance between two strings; it is defined as one minus the similarity of strings.

We computed all possible combinations of learning units and virtual patients/interactive algorithms using the similarity functions and stored the result in a database table (see Table 1) for further analysis and comparison with other algorithms. Similarity column represents computed trgm similarity between a learning unit and an algorithm. Correctness in interpretation of similarity results depends on our experts' expectations. If the two subjects are similar, we want to measure high similarity value.

D. Term-based similarity

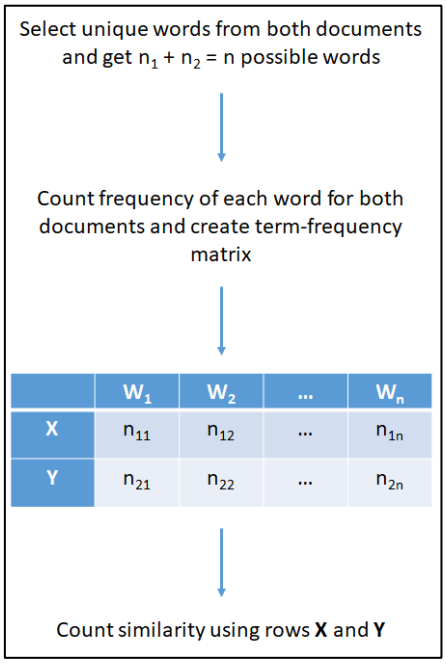
Term-based similarity approaches require that a similarity measure is chosen. A similarity measure quantifies the similarity between two numeric vectors of the same length. When using this approach, text documents are represented as bags of words, and a term-frequency matrix containing the counts of a word occurrence is computed. Figure 1 represents the process of computing text similarity between two documents.

As mentioned above, four similarity measures (the normalised Pearson correlation, the cosine similarity, the extended Jaccard coefficient and the Dice coefficient) were used to compute similarities between virtual patients/interactive algorithms from the AKUTNE.CZ portal and learning units form the OPTIMED platform. Each similarity measure is computed in a slightly different way and might have a correspondingly different interpretation. The normalised Pearson correlation is a centred correlation similarity measure. The cosine similarity is a measure of similarity between two vectors of an inner product space that

TABLE I.
EXAMPLE OF DATABASE TABLE INCLUDING SIMILARITIES.

id_learning_unit	id_algorithm	title_learning_unit	title_algorithm	similarity
890	77	Protection against radiation, principle of skiagraphy and skiascopy	Car accident	0.3278
891	77	Principle of computer tomography (CT), magnetic resonance (MR) and ultrasound, new horizons	Car accident	0.1798
894	77	Abdominal radiology	Car accident	0.2433
895	77	Uroradiology	Car accident	0.1857

measures the cosine of angle between them. The extended Jaccard similarity is computed as the number of shared terms over the number of all unique terms in both vectors. And finally, the Dice coefficient is defined as twice the number of common terms in the compared vectors divided by the total number of terms in both vectors [8], [10]. These similarity measures lie between 0 and 1. Zero means that two vectors are completely different, whereas one indicates that they are completely identical.



	W_1	W_2	...	W_n
X	n_{11}	n_{12}	...	n_{1n}
Y	n_{21}	n_{22}	...	n_{2n}

Fig. 1 The process of similarity computing.

In our work, all term-based computations were conducted in the R software using the *dplyr*, *tm* and *proxy* packages. First of all, documents from both sources were preprocessed, i.e. HTML tags, punctuation, digits, other special characters and words shorter than three characters were systematically removed. Afterwards, specific words using both a Google stop-word list and a customised stop-word list were removed. In the second step, word occurrence frequencies for each single document were counted. And thirdly, similarities between individual frequency vectors of documents (virtual patients/interactive algorithms) from AKUTNE.CZ and OPTIMED educational texts/learning units were computed using all of the selected similarity measures.

E. Online evaluation tool

A web-based tool has been designed and developed in order to obtain an effective evaluation of achieved results (in a form of detected similarities) by users (medical teachers and students). Using this tool, objective opinions critically assessing the relevance between virtual patients/interactive algorithms and a particular learning unit can be systematically organised and processed. The evaluation module is integrated into the OPTIMED curriculum management system and offers the possibility to view the underlying data from both systems including an easy collection of the users' evaluation via an online form (see Fig. 2). A group of twelve evaluators (fifth and sixth year medical students, young and senior teachers) was involved for the purposes of our pilot evaluation. First of all, they needed to get acquainted with the name of the learning unit and with its brief description. Furthermore, they were invited to view the completed content of an evaluated learning unit, which was available via a direct link to the OPTIMED platform. Afterwards, the users started to evaluate the relevance of available virtual patients/interactive algorithms. Each individual Akutne.cz interactive algorithm was described by a title, a short description and keywords. The users' opinions were expressed using a marking system (grading scale) similar to that used in schools (i.e. the Likert scale from 1 to 5), where 1 meant that the interactive algorithm was very relevant to the learning unit and 5 meant that the interactive algorithm was not relevant to the learning unit at all.

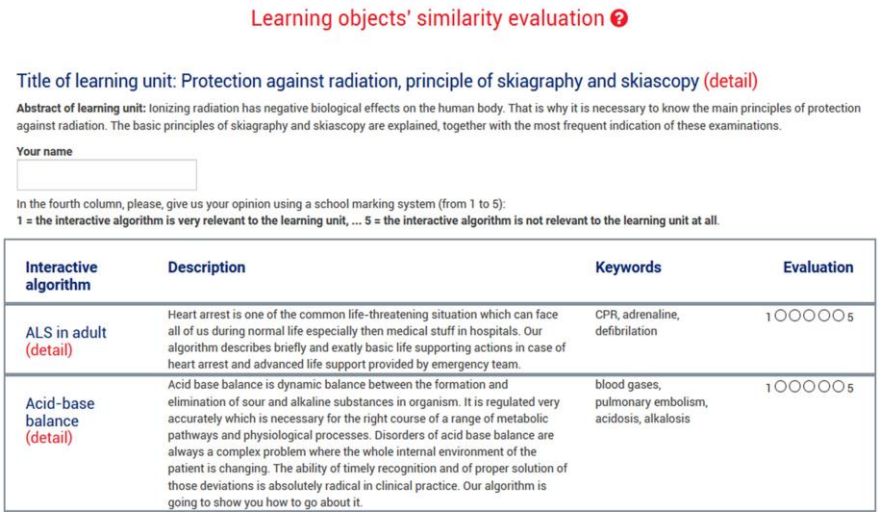


Fig. 2 Online form allowing evaluation of the teaching materials (relevancy of Akutne.cz interactive algorithms to a particular learning unit).

III. RESULTS

A. Overview of calculated similarities

The general overview (see Fig. 3 and Table 2) shows the comparison of five chosen approaches to similarity calculation in the form of a box plot chart, where the similarity measurements between all 77 interactive algorithms and 1232 learning units were taken into account. From our point of view, it is obvious that the pg_trgm module and its similarity function are very useful for cases where we expect to determine whether or not the original document and its copy are modified. It will very precisely and quickly find out whether there are differences in documents or whether the documents are identical.

On the other hand, this approach is not very appropriate for the comparison of two completely different documents, especially because it is dependent on the volume of the documents' content. Therefore, for the purposes of our pilot study, pg_trgm has been eliminated and the attention was only paid to four similarity measures (the normalised Pearson correlation, the cosine similarity, the extended Jaccard coefficient and the Dice coefficient), as described above.

After an in-depth analysis of the results, we discovered that in many cases, the cosine similarity, the extended Jaccard coefficient and the Dice coefficient indicated zero similarity because particular pairs of documents had no words in common. Nevertheless, the normalised Pearson correlation coefficient returned a non-zero value. That might be due to the fact that the correlation is a coefficient of linear dependency of two vectors. For example, let us compare a short text document (namely „What similarity measure value do we measure") with another short text document (namely „if correlation is used?)". The computed frequency vectors have the following form:

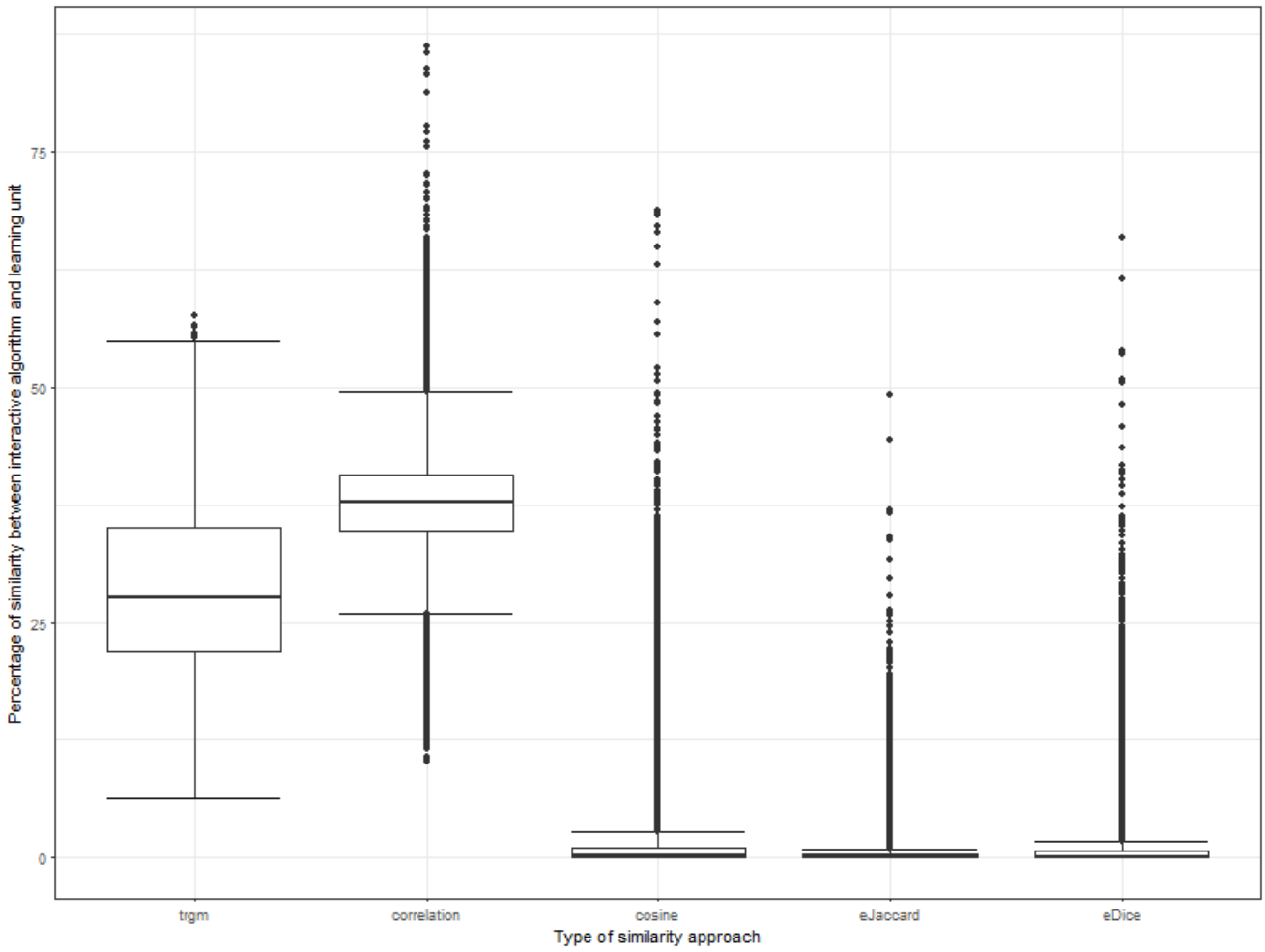


Fig. 3 Box plot representing five approaches for similarity calculation (minimum, maximum, median, average, upper and lower quartiles).

```
terms         freqs.x   freqs.y
correlation   0         1
measure       2         0
similarity    1         0
used          0         1
value         1         0
what          1         0
```

and the corresponding values of similarity measure are as follows:

```
correlation   0.07125354
cosine        0.00000000
eJaccard      0.00000000
Dice          0.00000000
```

This result shows that the correlation takes into account the whole vectors and its value depends on the vector structure rather than just intersection positions. Considering the above-described issue, we assume that the normalised Pearson correlation is not suitable for our text comparisons as out lexical understanding of term frequency in documents.
Therefore, all of our following analyses are conducted on cosine, extended Jaccard and Dice similarity measure outputs.

Figure 4 shows cosine, extended Jaccard and Dice similarity measures based on the normalised Pearson correlation coefficient, which calculates the linear correlation between two variables. The values between measures (0.923. 0.936, 0.996) imply that a linear equation describes the relationship between these measures perfectly, i.e. high positive correlation. Generally, cosine tends to return the highest values, whereas extended Jaccard tends to return the lowest ones. Nevertheless, all three measures provided very similar results.

B. Overview of calculated similarities

In terms of the pilot evaluation of achieved results (calculated similarity measures using various approaches), a set of learning units describing a complete course entitled "Diagnostic imaging methods" were used. Our users (medical students and teachers) used an online form to evaluate the relevance between learning units (OPTIMED) and interactive algorithms (AKUTNE.cz). Figure 5 represents the comparison of similarities (based on the normalised Pearson correlation coefficient) between three measures and the evaluation by users. It is immediately obvious that there is no linear correlation between any similarity measure and the evaluation by users. All algorithms used in a term-based process of similarity calculation provide very similar results, but the user evaluation of content similarity indicates no relationship or dependency between them.

TABLE II.
EXAMPLE OF SIMILARITY SUMMARY TABLE.

Approach	Minimum (%)	Maximum (%)	Median (%)	Average (%)	Upper quartile (%)	Lower quartile (%)
trgm	6.28	57.61	27.7	28.74	35.15	21.93
correlation	10.24	86.15	37.8	37.44	40.62	34.72
cosine	0	68.78	0.19	1.11	1.11	0
extended Jaccard	0	49.22	0.05	0.38	0.34	0
Dice	0	65.97	0.1	0.74	0.68	0
trgm	6.28	57.61	27.7	28.74	35.15	21.93

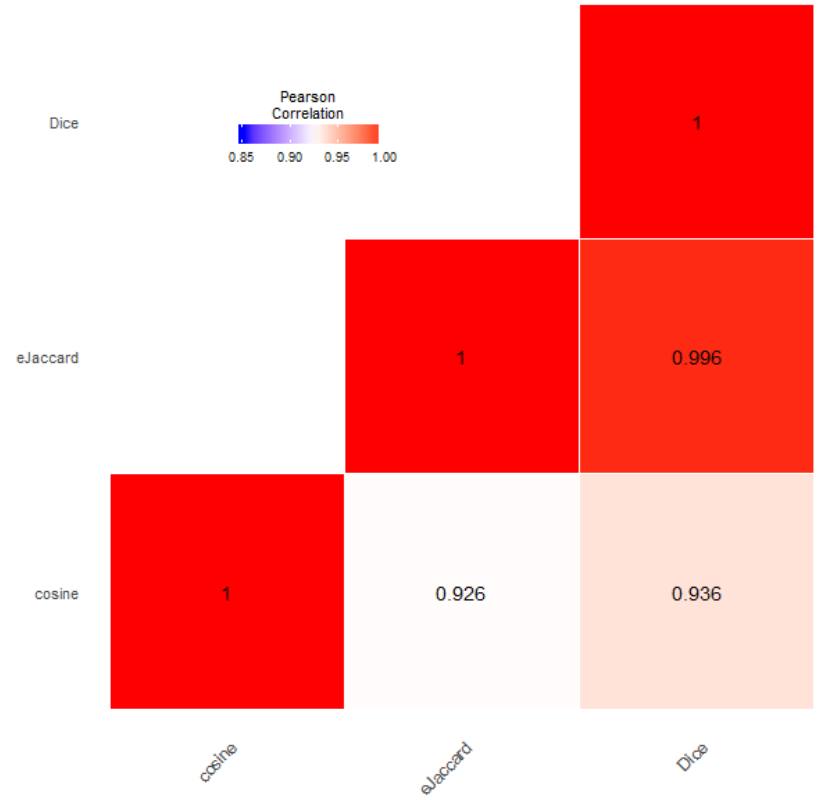


Fig. 4. Comparison of three chosen similarity measures.

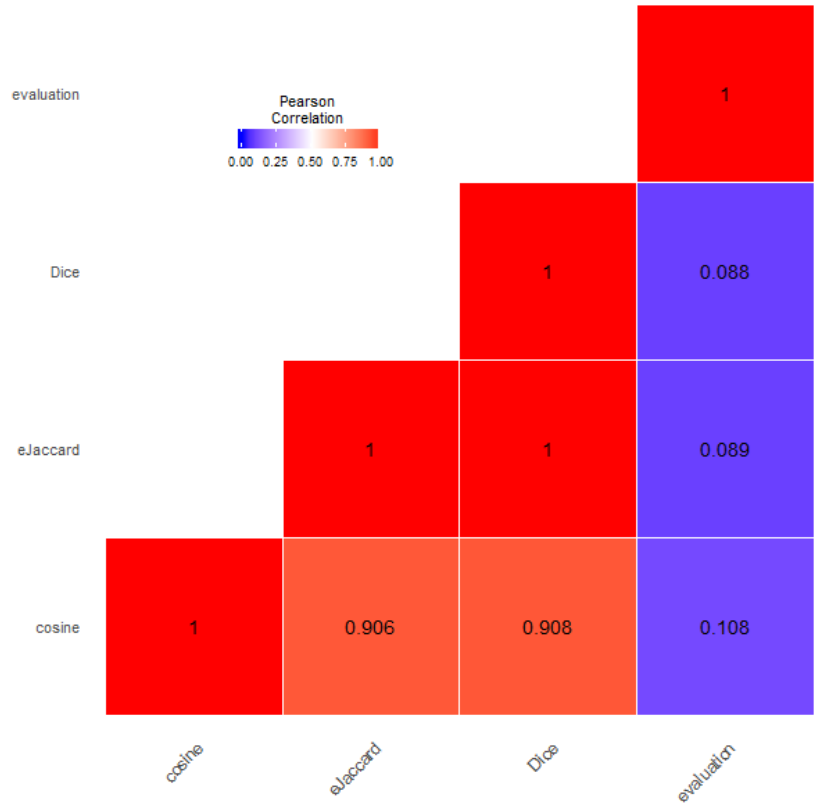


Fig. 5 Comparison of three chosen similarity measures together with users' evaluation.

IV. DISCUSSION

Based on statistical calculations, similarities and common features / terms were found between interactive algorithms in the AKUTNE.cz platform and learning units in the OPTIMED platform. Contrary to our expectations, this study did not find significant similarity between our calculated results of cosine, extended Jaccard and Dice similarity and ratings of users (medical students and teachers). We concluded that a number of factors play a role in determining the results.

Perhaps the most serious limitation of this analysis is an inappropriate choice of the course "Diagnostic imaging methods" with its very specific content and the subsequent search for similarities with a set of interactive educational algorithms in the AKUTNE.CZ platform, which were designed as teaching aids for other courses of the Medical Faculty curriculum: First Aid, Intensive Medicine, Anesthesiology and Pain Management. The fact that X-ray and other imaging methods were used in almost every algorithm only demonstrates that these are frequently used imaging techniques in acute medicine, not that X-ray should be the learning outcome of individual algorithms.

Another important confounding factor is the keywords selection. We believe that mechanically chosen keywords lead to an overinterpretation of search results; the most frequent ones do not represent the most important words, i.e. the keywords. It became obvious that keywords chosen by the machine might have not agreed and in some cases really disagreed with the algorithm's keywords. Undoubtedly, if keywords of the algorithms as defined by the creators had been used, such similarity would have not occurred and it would have increased the accuracy. Moreover, there is definitely space for a systematic improvement of a customised stop-word list. We will need to eliminate terms that do not bring the required information value.

Human evaluators may have contributed to misleading results rather significantly in several ways. Some evaluators had been involved in the design of interactive educational algorithms, which inevitably led to a bias. Some evaluators might have provided an incorrect evaluation due to a misapprehension and/or an unclear assignment. It is important to point out that human evaluators tended to focus on similarities in the meaning of concepts and terms, unlike the machine-based and statistical evaluation of similarities. One improvement to be possibly considered in future might be an optimisation of the evaluation process itself, which should be focused and implemented as a two-stage analysis in the follow-up to this pilot study. First of all, appraisers/evaluators/users would identify similarities according to established keywords, followed by their own analysis of content (abstracts and then full texts). From the methodological point of view, this process would be adopted from the process of assessing professional resources in literature reviews [11], [12]. There is also the possibility to carry out the evaluation as a three-stage process (the third stage would be a peer discussion among evaluators), but it is clear that such a process would be very time-consuming.

Yet another challenge lies in the subjective rating of significance for evaluators and users, which stems from reasoning of both practical and scholarly significance of the teaching problem as well as the scope or the respective teaching topic and issue. In other words, students' and teachers' views could differ when evaluating the learning units and interactive algorithms. Furthermore, the specialised orientation of evaluators could be the explanation for results achieved from their qualitative evaluation: most of our evaluators/users in the pilot study were professionals most familiar with acute care. In their daily practice, they are much more focused on acute and rescue interventions with the goal of saving lives rather than focusing on examination methods, especially not on radiology and imaging methods. Another possible explanation of our findings from qualitative evaluations is that users / evaluators could not see a clear link between the two assessed contents of study materials and interactive algorithms, and their views were reflected in the evaluation. What should be highlighted is that even this finding could help us improve future development of interactive algorithms: there is more space for visual documentation of a clinical condition in intensive care because there is strong evidence that imaging documentation is helpful in the education of healthcare professionals [13]–[15].

We must also emphasise that an important role was played by the fact that the volume of evaluated study materials and interactive algorithms was relatively large (n = 77 virtual patients/interactive algorithms) and that all evaluators carried out their evaluations independently, without the opportunity to communicate with others.

Despite the fact that inconsistencies were identified in our "quantitative/mathematical" and "qualitative – user view" evaluation, we are still convinced that the chosen procedure was appropriate to the above-mentioned set of objectives. We have repeatedly verified that an automated statistical evaluation must always be accompanied by an expert judgment and by an evaluation provided by target users of teaching materials [1]. At least we have verified that our methodology can reveal potential gaps as well as new possibilities of linking study materials to improve the learning process and to increase the students' preparedness for clinical practice. In the follow-up work, we would also like to approach other specialists who might provide their feedback as evaluators; as we have already mentioned, the feedback in this case was mostly provided by intensive care specialists and anaesthetists.

ACKNOWLEDGMENT

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. We are also thankful to the team of medical students and teachers, who evaluated achieved results, namely Tereza Prokopová, Daniel Barvík, Václav Vafek, Tereza Ondráčková, Jiří Libra, Matěj Anton, Lucia Macková, Klára Vataha and Martina Žižlavská.

References

[1] M. Komenda, M. Karolyi, R. Vyškovský, K. Ježová, and J. Ščavnický, 'Towards a keyword extraction in medical and healthcare education', in 2017 Federated Conference on Computer Science and Information Systems (FedCSIS), 2017, pp. 173–176.

[2] D. Schwarz et al., 'Interactive Algorithms for Teaching and Learning Acute Medicine in the Network of Medical Faculties MEFANET', J. Med. Internet Res., vol. 15, no. 7, Jul. 2013.

[3] M. Komenda, D. Schwarz, C. Vaitsis, N. Zary, J. Štěrba, and L. Dušek, 'OPTIMED Platform: Curriculum Harmonisation System for Medical and Healthcare Education', Stud. Health Technol. Inform., vol. 210, pp. 511–515, 2015.

[4] M. Komenda et al., 'Curriculum Mapping with Academic Analytics in Medical and Healthcare Education', PloS One, vol. 10, no. 12, 2015.

[5] M. Víta, M. Komenda, and A. Pokorná, 'Exploring Medical Curricula Using Social Network Analysis Methods', Jul. 2015.

[6] M. Karolyi, M. Komenda, R. Janoušová, M. Víta, and D. Schwarz, 'Finding overlapping terms in medical and health care curriculum using text mining methods: reha', MEFANET J., vol. 4, no. 2, pp. 71–77, Jan. 2017.

[7] R. Randell, R. Cornet, and C. McCowan, Informatics for Health: Connected Citizen-Led Wellness and Population Health. IOS Press, 2017.

[8] W. H. Gomaa and A. A. Fahmy, 'A Survey of Text Similarity Approaches', Int. J. Comput. Appl., vol. 68, no. 13, pp. 13–18, Apr. 2013.

[9] 'PostgreSQL: Documentation: 9.1: pg_trgm'. [Online]. Available: https://www.postgresql.org/docs/9.1/static/pgtrgm.html. [Accessed: 15-May-2018].

[10] H. Liu, J. He, D. Zhu, C. X. Ling, and X. Du, 'Measuring Similarity Based on Link Information: A Comparative Study', IEEE Trans. Knowl. Data Eng., vol. 25, no. 12, pp. 2823–2840, Dec. 2013.

[11] M. Newman and D. Elbourne, 'Improving the Usability of Educational Research: Guidelines for the REPOrting of Primary Empirical Research Studies in Education (The REPOSE Guidelines)', Eval. Res. Educ., vol. 18, no. 4, pp. 201–212, Nov. 2004.

[12] J. P. Gall, M. D. Gall, and W. R. Borg, Applying educational research: A practical guide. Longman Publishing Group, 1999.

[13] B. F. Branstetter, L. E. Faix, A. L. Humphrey, and J. B. Schumann, 'Preclinical Medical Student Training in Radiology: The Effect of Early Exposure', Am. J. Roentgenol., vol. 188, no. 1, pp. W9–W14, Jan. 2007.

[14] R. B. Gunderman, A. R. Siddiqui, D. E. Heitkamp, and H. D. Kipfer, 'The Vital Role of Radiology in the Medical School Curriculum', Am. J. Roentgenol., vol. 180, no. 5, pp. 1239–1242, May 2003.

[15] K. Soyebi, 'Changing students' performance in and perception of radiology', Med. Educ., vol. 42, no. 5, pp. 522–522, May 2008.