

Exploring Medical Curricula Using Social Network Analysis Methods

Martin Vítá
NLP Centre

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

Martin Komenda, Andrea Pokorná
Institute of Biostatistics and Analyses,
Faculty of Medicine, Masaryk University
Kamenice 126/3, 625 00 Brno, Czech Republic
Email: {komenda, pokorna}@iba.muni.cz

Abstract—This contribution demonstrates how to apply concepts of social network analysis on educational data. The main aim of this approach is to provide a deeper insight into the structure of courses and/or other learning units that belong to a given curriculum in order to improve the learning process.

The presented work can help us discover communities of similar study disciplines (based on the similarity measures of textual descriptions of their contents), as well as identify important courses strongly linked to others, and also find more independent and less important parts of the curriculum using centrality measures arising from the graph theory and social network analysis.

I. INTRODUCTION

NOWADAYS, the process of improving educational curricula requires more sophisticated analyses of relevant datasets than in previous decades. Using various powerful virtual learning environments (see [7], [9], [14], [10], [5]) authorial teams consisting of curriculum designers and guarantors are able to construct a detailed description of each lecture, seminar and practice. The current curricula usually represent a huge amount of data records (thousands of standard pages in total), which cover all necessary requirements on the graduates based on a predefined structure in an online environment including formal and semantic verification. The main objective of this paper is to present an innovative approach of exploring the (medical) curricula in a transparent way. This approach is based on a proven concept of social network analysis (SNA), since SNA provides a natural way of dealing with graph notions such as centrality, and with various models of importance. The primary motivation for these investigations is to create a suitable tool/methodology for anyone who is involved in the complicated process of curriculum design at institutions of higher education. The research can help us when answering – among others – the following questions:

- Are there any content overlaps and/or gaps in the curriculum? (The overlaps might be desirable in some cases and undesirable in others).
- Which disciplines/courses/learning units are similar in terms of contents? And which are “self-standing”?
- Are there some significant communities (clusters) of disciplines/courses/learning units (such that a change of one may influence others)?

- Which disciplines/courses/learning units are probably the “central” parts of the curriculum?

The results of our analysis are intended to be used by the (human) experts responsible for the development and further evaluation of the curricula and institutional management, nevertheless they could also be interesting for students who want to study efficiently and focus mainly on the important parts of the study plan. We are going to introduce a methodology/data mining process that is fully compatible with the process of standardised knowledge discovery in databases (KDD). The data mining process was inspired by Trigo and Brazdil’s works [16] and [2].

Our basic requirement on the selected approach is the simplicity and re-usability in practice. We use only a commonly available and free software (R), during the modelling stage, so the process can be repeated without any notable expenses. The possibility of straightforward visualisation is one of the advantages of this approach: we are able to show these content-based relations among textual descriptions of medical education (namely disciplines), not only the formal or organisational relations. Such visualisation can provide a quick and smart overview of a study plan and provide comprehensive information for the subsequent global in-depth curriculum inspection, which could be used for the future planning and changes in the curricula.

II. METHODOLOGY

Data exploration was done in accordance with a proven KDD background, namely the standardised methodology CRISP-DM (CRoss-Industry Standard Process for Data Mining) [1]. This process is entirely independent of selected modeling tools and consists of a cycle that involves six stages (Fig. 1). Each stage represents an independent issue which must be completed in order to move forward. This methodology is used in a wide range of applications including biosciences, industry, even finance [12]. All of the CRISP-DM steps are described below for our setting.

We have already proposed a complex curriculum planning model supporting the outcome-based paradigm [8], which promotes a clear communication between the involved stakeholders (teachers, guarantors, curriculum designers, supervisors and faculty management) [11]. Based on a robust web-oriented

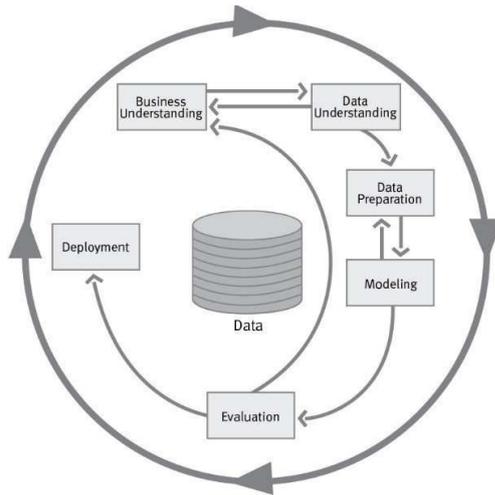


Fig. 1. Diagram of CRISP-DM.

platform for complex curriculum management which provides an effective tools for creating, transparent browsing, and reviewing the curriculum, a correctly compiled and balanced description of the General Medicine study field was defined by the authorial team consisting of 384 guarantors and teachers of the Faculty of Medicine at Masaryk University. It has covered more than 1300 learning units and 7100 learning outcomes, i.e. approximately 2500 standard pages of text. With respect to human cognition abilities, it is not possible to carefully read and verify the content of all learning units with all their linkages and co-dependencies. We have decided to apply the CRISP-DM methodology as a general framework to obtain a deeper knowledge from the medical educational data.

A. Business understanding

Business understanding is an initial stage, which is focused on understanding the objectives and the problem definition – in our case, in terms of the in-depth exploration of medical curriculum in-depth exploration. The proposed goal is to detect outlying and overlapping areas in the General Medicine study field as well as investigate the mutual similarities of the courses involved. The whole study field is split into four individual modules (Diagnostic Sciences and Neurosciences, Internal Medicine, Surgical Sciences and Theoretical Sciences). These modules consist of 44 medical disciplines (such as Anatomy, Stomatology, Neurology, etc.). In total 144 courses are assigned to one or more disciplines are grouped together on a level of these disciplines for the purpose of this analysis. In this way, each discipline is generated from content-related courses across the entire study field (e.g. the discipline of Anatomy contains the following courses: Anatomy I – lecture, Anatomy I – seminar, Anatomy II – lecture, Anatomy II – practice).

B. Data understanding

The follow-up stage – called data understanding – begins with the initial collection of data. Each course is described by

a set of learning units using textual parameters and descriptors (see Table I). In our particular case, we have detected a set of descriptive attributes related to the learning outcomes, which were defined in accordance with the Bloom's taxonomy [13]. The Bloom's scheme provides a standardised classification of educational objectives that gives a commonly understood meaning to objectives classified in one of six main categories and many subcategories, thereby enhancing communication and achievement of more complex skills and abilities.

TABLE I
SELECTED ATTRIBUTES OF A LEARNING UNIT.

Attribute	Type
learning_unit	varchar (255)
total_range	int (10)
meaning	text
annotation	text
mesh_keyword	varchar (255)
significant_term	varchar (255)
learning_outcome	varchar (500)
grouped_outcome	varchar (500)
primary_index	varchar (500)
secondary_index	varchar (500)
assessment_form	varchar (200)

We have identified several attributes of learning outcomes mined from our curriculum management system, which can provide complete information about the coverage among various medical disciplines. All text data that we used was in English due to easier preprocessing steps.

Teaser of textual data to be processed:

Section:

Surgical Sciences

Medical discipline:

Surgery

Group outcome:

Abdominal aorta and its branches

Primary index: Arterial diseases

Secondary index: Closure, stenosis

Learning outcomes:

- Student masters anatomy of arterial system.
- Student describes injuries to the abdominal aorta and its branches.
- Student lists anatomy of abdominal aorta and its branches.

C. Data preparation

The stage of data preparation is a necessary prerequisite before applying any methods of data mining and/or SNA. It

consists of a sequence of procedures to create the final dataset from the initial raw data.

Each discipline was represented by a single plaintext file that contains merged contents of several fields from Table I, namely `learning_outcome`, `grouped_outcome`, `primary_index`, `secondary_index`.

The collection of these plaintext files was loaded as a corpus into the R system extended by a couple of packages `tm` and `lsa` – standard packages for text mining and latent semantic analysis. After tokenization at this stage, several standard preprocessing issues were performed – in the following sequence:

- transformation to lowercase,
- stemming (using Snowball),
- punctuation removal,
- numbers removal,
- stopwords removal,
- whitespace stripping.

Similarly to the work by [16] we have chosen a bag-of-words representation of the documents in the corpus and a document-term-matrix was generated. (*tf-idf* weighting [6] was used). Consequently, dissimilarity matrices were computed on the basis of cosine similarity. Values were rounded to two decimal digits and values lower than a certain threshold were replaced by zeros, since extremely low similarities were considered as irrelevant.

D. Modeling

Generally, various modeling techniques are selected and applied at this stage – SNA methods were applied in our case. Since some of the algorithms involved have several parameters, this stage also contains experiments with different values of these parameters.

Application of social network analysis – centrality concepts on the similarity graph: The dissimilarity matrix can be viewed as an adjacency matrix of a similarity graph – undirected graph with weighted edges. On this graph we are going to perform several calculations of centrality measures. Obviously, these values should be interpreted differently – with respect to the roles of disciplines/courses/learning units in the educational framework. We are going to deal with the following centrality measures:

Closeness – The closeness centrality of a node v in a graph G is defined by the inverse of the sum of the lengths of the shortest paths to/from all the other nodes in the graph G , i. e.:

$$c(v) = \frac{1}{\sum_{i \in V(G), i \neq v} d(i, v)},$$

where $d(i, v)$ is the length of the shortest path from node i to node v and $V(G)$ is the set of all vertices of the graph G .

If there is no path between a couple of nodes then the total number of nodes of the graph is used instead of the path length. By the mentioned calculation we obtain the so-called *raw closeness* of the node. To get normalised closeness of the node v , we multiply the raw closeness by $n - 1$, where $|V(G)| = n$: we are going to use this normalised version.

Nodes (disciplines) with low value of closeness are those disciplines with their content distant from other ones, thus, roughly speaking, they are independent on the others.

Betweenness centrality – In the simplest case (without edge weighting), the raw betweenness centrality of a node v corresponds with the number of shortest paths from all nodes to all others that go through a considered node, i. e.:

$$b(v) = \sum_{i, j, v \in V(G), i \neq j, i \neq v, j \neq v} \frac{g_{ivj}}{g_{ij}},$$

where g_{ij} is the total number of shortest paths going from node i to j and g_{ivj} is the total number of all shortest from node i to node j going through v . To get normalized betweenness $b_n(v)$ of the node v , we calculate $b_n(v) = \frac{2b(v)}{(n-1)(n-2)}$, where again, $|V(G)| = n$.

This definition can be extended for weighted networks. Nodes (disciplines) with a high betweenness centrality are those that are the best for joining the students' knowledge from different collections of disciplines.

Eigenvector centrality – One of the methods of computing of approximate importance of a given node. The idea behind this measure is that centrality of each node is the sum of the centrality values of its neighbour nodes. More precisely, the eigenvector centrality values correspond to the values of the first eigenvector of the adjacency matrix. The eigenvector centrality in our case models identification of important disciplines of the curriculum.

All these computations are done using the `igraph` package [3] that contains several built-in functions covering the area of SNA. The values of these measures were normalised in a relevant way. According to our goal, our intention of this stage is to find out nodes (disciplines) with high values of proposed measures to identify the core and important parts of the curriculum and, in contrary, also the nodes (disciplines) with lowest values of centrality measures to identify those that are relatively independent of others.

Extremely low values of these attributes can also indicate a non-proper description of the discipline (e.g. missing important parts of the description). Disciplines with the most interesting results (means highest values of centrality measures attributes) are presented below.

Ten disciplines with the highest betweenness centrality:

- 1) Pathological physiology (0.265),
- 2) Physiology (0.164),
- 3) Surgery III (0.140),
- 4) Pathology (0.111),
- 5) Clinical examination in neurology (0.104),
- 6) Neuroscience (0.083),
- 7) Immunology (0.070),
- 8) Intensive care medicine (0.065),
- 9) Biology (0.065),
- 10) Preventive medicine (0.063).

Ten disciplines with the highest closeness centrality:

- 1) Pathological physiology (0.106),
- 2) Pathology (0.105),
- 3) Clinical examination in neurology (0.105),
- 4) Immunology (0.105),
- 5) Biology (0.105),
- 6) Surgery III (0.105),
- 7) Internal medicine – part 4 – Gastroenterology and haematology (0.105),
- 8) Intensive care medicine (0.105),
- 9) Clinical oncology (0.105),
- 10) Histology and embryology (0.105).

Ten disciplines with the highest eigenvector centrality:

- 1) Pediatrics II (1.000),
- 2) Pediatrics III (0.904),
- 3) Clinical oncology (0.867),
- 4) Surgery I-II (0.842),
- 5) Surgery III (0.819),
- 6) Pathology (0.782),
- 7) Internal medicine – part 4 – Gastroenterology and haematology (0.522),
- 8) Clinical examination in surgery (0.406),
- 9) Dermatovenereology (0.375),
- 10) Pathological physiology (0.358).

On the opposite side there are disciplines with low values.

Disciplines with zero betweenness centrality:

- Anatomy II,
- Basic medical terminology I,
- Clinical examination in internal medicine,
- Communication and selfexperience,
- Community medicine,
- Diagnostic imaging methods,
- Family medicine and geriatrics,
- Gynecology and obstetrics,
- Internal medicine – part 2 – Cardiology and angiology,
- Internal medicine – part 6 – Occupational medicine,
- Medical ethics I,
- Medical ethics II,
- Medical chemistry,
- Medical microbiology II,
- Medical psychology,
- Nursing,
- Ophthalmology,
- Orthopaedics,
- Pediatrics II,
- Pharmacology I,
- Stomatology.

Ten disciplines with the lowest closeness centrality:

- 1) Anatomy II (0.098),
- 2) Medical ethics I (0.017),
- 3) Medical ethics II (0.017),
- 4) Basic medical terminology I (0.017),
- 5) Communication and selfexperience (0.017),

- 6) Diagnostic imaging methods (0.017),
- 7) Family medicine and geriatrics (0.017),
- 8) Nursing (0.017),
- 9) Ophthalmology (0.017)
- 10) Stomatology (0.017).

Ten disciplines with the lowest eigenvector centrality:

- 1) Medical microbiology II (0.013),
- 2) Clinical examination in internal medicine (0.009),
- 3) Pharmacology II (0.006),
- 4) Medical psychology (0.004),
- 5) Anatomy III (0.004),
- 6) Anatomy II (0.001),
- 7) Pharmacology I (0.001),
- 8) Anatomy I (0.001),
- 9) Medical ethics II (0.000),
- 10) Medical ethics I (0.000).

Interpretation: In general, we show medical disciplines with extreme (the highest/the lowest) values of selected centrality measures. The achieved results represent novel and hopefully useful information about the structure of the General Medicine study field created in the curricula. For instance, Pathological physiology and Pathology appear in all top-ten lists in all centrality measures, Surgery III, Clinical examination in neurology, Immunology, Biology, Internal medicine – part 4 – Gastroenterology and haematology, Clinical oncology in two of them. It may indicate that these disciplines belong purposely to an essential part of the curriculum or, on the contrary, it may reveal an undesirable preference of mentioned disciplines, which in fact cannot be identified by a visual human inspection. So, logically the final interpretation has to be done under the supervision of responsible curriculum designers and senior guarantors, who are familiar with optimal composition and intersections of individual disciplines, courses and learning units.

Application of social network analysis – community detection on the similarity graph: The community detection is a common task when dealing with social networks. Communities, i.e. densely connected subgraphs have also an importance for exploring the curricula: communities correspond with subsets of mutually close disciplines (with respect to content similarity). In contrast to hard clustering we do not insist on the rule that each item (discipline) is contained in just one cluster (community).

For detecting communities, the Walktrap algorithm [15] was used. The main idea of this algorithm is that short random walks tend to stay in the same community, see [3]. The implementation of this algorithm is also contained within the *igraph* package in R, and for our purposes it was executed with a parameter of length of the random walk set to $k = 4$.

Fig. 2 provides a basic overview of communities that were found. The thickness of the edges corresponds with the similarity between nodes (disciplines). Communities are bounded shapes with a grey background color. As we can see, we have several “singleton” communities along with some bigger ones that establish the core of the curriculum.

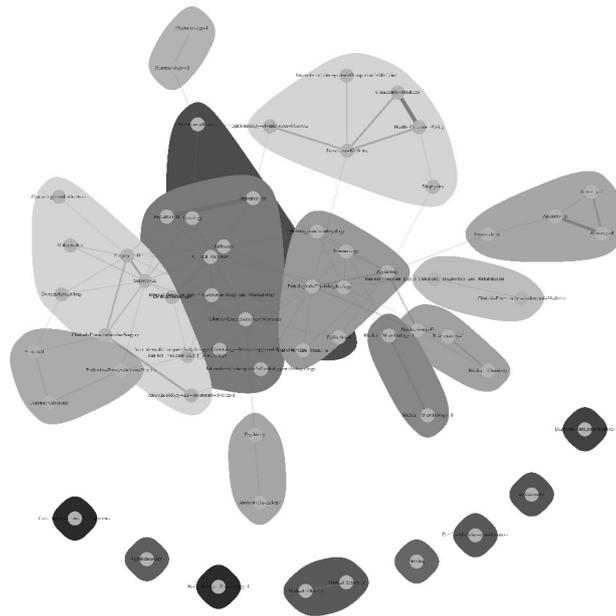


Fig. 2. Overview of the curriculum with marked communities.

- Examples of larger communities: {Preventive medicine, Health care and policy, Community medicine, Biophysics, Internal medicine – part 6 – Occupational medicine, Epidemiology of infectious diseases}, {Medical chemistry, Biochemistry I, Biochemistry II}, ...
- Examples of independent communities: {Nursing}, {Diagnostic-Imaging methods}, {Communication and self experience}, ...

A detail of a community within the entire graph is shown in Fig. 3. There are also communities connecting just disciplines divided into two parts, such as Medical ethics I and Medical ethics II, Pharmacology I and II, etc. These results surely are not coincidental; quite the opposite, they confirm the reasonability of the method.

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. It may indicate either balanced or unbalanced representation of compulsory and optional courses intended for graduation to obtain professional qualification for employment as a physician. The curriculum visualisation based on the location of individual

disciplines and their coloured marking is useful with regard to the possibility of a comprehensive evaluation of the curriculum structure. It makes it possible to identify weak points and shortcomings in terms of inconvenient interdisciplinary relations, as well as remote disciplines with hardly any relations to other disciplines. Additionally, the visualisation can very easily confirm the consistency of individual specialised disciplines, which leads to the idea of correctly specified requirements on the acquired knowledge and skills of future physicians. Currently, the final process of deployment is still ongoing. In the end, presented visualisation will be integrated directly into the curriculum management platform as an additional overview module.

III. CONCLUSION AND FUTURE WORK

In this work we have introduced a novel method for exploring general curricula that uses several concepts of social network analysis. The presented use-case provides an easy-to-understand visualisations of the entire medical curriculum and centrality models involved disciplines. The entire process of obtaining the knowledge is done according to an industrial standard in data mining (CRISP-DM). Our plans for further work are described below.

1) *Improving the similarity graph:* In this experimental stage, only a part of the accessible data was used

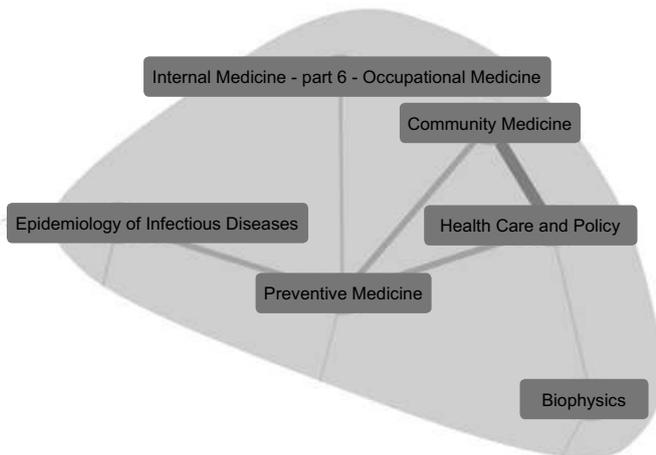


Fig. 3. Detail of a given community of disciplines.

(learning_outcome, grouped_outcome, primary_index, secondary_index). Further development of this approach will be based on incorporating data such as the MeSH [4] terms. There is also an opportunity to experiment with weights: terms contained in certain controlled vocabularies, thesauri, ontologies etc. can be taken with boosted weights. The “NLP” stage will be improved in several ways, for example dealing with synonyms, hypernyms etc.

2) *Enriching the visualisation:* At this stage, we have only focused on the structure of the similarity graph and visualisation of obtained communities – that is, in fact, the core of the original work. Hence we have only used a small amount of attributes of the produced graph, i. e. the thickness of the edges (and vertices grouping). The remaining parameters of the visualisation can therefore carry additional formal, organisational or empirical data:

- The size of a node should correspond with the total number of teaching hours per discipline.
- The color of a concrete node can be chosen accordingly to a given classification of the discipline.
- The opacity should correspond with the importance of the node – the eigenvector centrality (or some other attribute of the discipline (for example, the ratio of students that fail the exams).
- Textual labels of the nodes can be used for information about the number of students attending the discipline.
- Edges on another layer (represented in a different color) can link a discipline having the same lecturers.

Both of these activities might lead to more information-rich visualisations that would provide a better insight to the specific curriculum which will be evaluated manually by experts in the near future.

ACKNOWLEDGMENT

We would like to express our special thanks to the management of the Faculty of Medicine at Masaryk University, represented by Prof. Jaroslav Štěřba (Vice-Dean for Education in Clinical Branches) and Prof. Jiří Mayer (Dean of the Faculty of Medicine), and also to the Institute of Biostatistics and Analyses at Masaryk University represented by Assoc. Prof. Ladislav Dušek.

REFERENCES

- [1] Azevedo, Ana Isabel Rojao Lourenço, 2008, KDD, SEMMA and CRISP-DM: a parallel overview. [online]. 2008. [Accessed 9 July 2013]. Available from: <http://recipp.ipp.pt/handle/10400.22/136>
- [2] Brazdil, Pavel, Trigo, Luís, Cordeiro, Joao, Sarmiento, Rui and Valizadeh, Mohammadraza, 2015, Affinity mining of documents sets via network analysis, keywords and summaries. Oslo Studies in Language, 7(1).
- [3] Csardi, Gabor and Nepusz, Tamas, 2006, The igraph software package for complex network research. InterJournal, Complex Systems. 2006. Vol. 1695, no. 5, p. 1–9.
- [4] Davis, Allan Peter, Wieggers, Thomas C., Rosenstein, Michael C. and Mattingly, Carolyn J., 2012, MEDIC: a practical disease vocabulary used at the Comparative Toxicogenomics Database. Database. 2012. Vol. 2012, p. bar065.
- [5] Erguzen, Atilla, Erel, Serafettin, Uzun, Ibrahim, Bilge, Hasan Sakir and Unver, Halil Murat, 2012, KUZEM LMS: A new learning management system for online education. Energy Education Science and Technology Part B-Social and Educational Studies. 2012. Vol. 4, no. 3, p. 1865-1878.
- [6] Feldman, Ronen and Sanger, James, 2007, The text mining handbook: advanced approaches in analyzing unstructured data [online]. Cambridge University Press. [Accessed 2 May 2015].
- [7] Frank, Jason R. and Danoff, Deborah, 2007, The CanMEDS initiative: implementing an outcomes-based framework of physician competencies. Medical teacher. 2007. Vol. 29, no. 7, p. 642-647.
- [8] Harden, R. M., Crosby, J. R. and Davis, M. H., 1999, AMEE Guide No. 14: Outcome-based education: Part 1—An introduction to outcome-based education. Medical teacher. 1999. Vol. 21, no. 1, p. 7-14 (doi:10.1080/01421599979969).
- [9] Kabicher, S. and Derntl, M., 2008, Visual Modelling for Design and Implementation of Modular Curricula. Zeitschrift für Hochschulentwicklung [online]. 2008. [Accessed 19 July 2012]. Available from: <http://www.zfhe.at/index.php/zfhe/article/view/64>
- [10] Kerkiri, Tania Al and Papadakis, Spyros, 2012, Learning Outcomes Design Authoring Tool: The Educator is Not Alone! International Journal of e-Collaboration (IJeC). 2012. Vol. 8, no. 4, p. 22-34.
- [11] Komenda, Martin, Schwarz, Daniel, Hřebíček, Jiří, Holčík, Jiří and Dušek, Ladislav, 2014, A Framework for Curriculum Management - The Use of Outcome-based Approach in Practice. In : Proceedings of the 6th International Conference on Computer Supported Education [online]. Barcelona: SCITEPRESS, p. 473-478. ISBN 978-989-758-020-8. Available from: <http://www.csedu.org>
- [12] Korczak, Jerzy, et al. A-Trader-consulting agent platform for stock exchange gamblers. In: Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on. IEEE, 2012. p. 963-968.
- [13] Krathwohl, David R., 2002, A revision of Bloom's taxonomy: An overview. Theory into practice. 2002. Vol. 41, no. 4, p. 212-218.
- [14] Mong, Yu, Chan, Mangtang and Chan, Francis Kar Ho, 2008, Web-based outcome-based teaching and learning - An experience report. In: Advances in Web Based Learning—ICWL 2007. Berlin: Springer-Verlag Berlin. p. 475-483. ISBN 978-3-540-78138-7.
- [15] Pons, Pascal and Latapy, Matthieu, 2005, Computing communities in large networks using random walks. In: Computer and Information Sciences-ISCIS 2005 [online]. Springer. p. 284–293.
- [16] Trigo, Luís and Brazdil Pavel, 2014, Affinity Analysis between Researchers using Text Mining and Differential Analysis of Graphs, ECM-L/PKDD 2014 PhD session Proceedings, Nancy, France, p. 169–176.
- [17] Uzunboyulu, Höseyin, Bicen, Hüseyin and Cavus, Nadire, 2011, The efficient virtual learning environment: A case study of web 2.0 tools and Windows live spaces. Computers & Education. 2011. Vol. 56, no. 3, p. 720–726.