The evaluation of convergence between content description of academic textbooks in the form of textual notes and UDC expressions

Paweł Lula[1], Urszula Cieraszewska[2], Monika Hamerska[3]

Abstract

In the era of dynamic development of e-commerce solutions product descriptions available online have crucial importance for customers. Therefore, product characterisation should reflect its core features. In the paper the problem of the convergence between the content of available online product descriptions and their significant features is discussed. The analysis presented here was concentrated on the evaluation of the adequacy between prepared in the Polish language information about academic textbooks published by online bookstores and main topics presented in them.

In the research the Universal Decimal Classification (UDC) system was used for representation of essential content of every monograph. Whereas vector space model was used for representation of online available book descriptions. Using these two methods of content description cluster analysis of monographs was performed. Assuming that a dendrogram generating by a cluster method reflects the content-based relationships between monographs, the similarity analysis of obtained dendrograms can be treated as a tool of convergence evaluation between information content of monograph descriptions having two different form: UDC expressions and textual descriptions. Generally, the research outputs allow to assess the convergence between the content of published online textual description of academic textbooks and their core elements represented by UDC codes.

Keywords: *text mining, universal decimal classification (UDC) system, similarity of textual documents, similarity of cluster analysis results*

JEL Classification: C630, C88

1 Introduction

The analysis of convergence between information content of textual description of academic monographs and their bibliographic characteristics having the form of UDC expressions is the main goal of the research. First section of the paper presents Universal Decimal Classification (UDC) scheme and the algorithm of similarity calculation between two UDC expressions. The problem of expressing the similarity of textual documents is presented in the third section. Next, in the fourth section, the Fowlkes-Mallows index as a tool for similarity analysis between dendrograms is briefly presented. The fifth section of the paper presents the

[1] Corresponding author: Cracow University of Economics, Department of Computational Systems, 27 Rakowicka St., 31-510 Cracow, Poland, e-mail: pawel.lula@uek.krakow.pl.

[2] Cracow University of Economics, Main Library of the Cracow University of Economics, 27 Rakowicka St., 31-510 Cracow, Poland, e-mail: cieraszu@uek.krakow.pl.

[3] Cracow University of Economics, Science and Knowledge Transfer Department, 27 Rakowicka St., 31-510 Cracow, Poland, e-mail: hamerskm@uek.krakow.pl.

results of empirical research concerning the analysis of convergence between textual descriptions of academic textbooks offered by Polish online bookstores and their content represented by UDC descriptions. Conclusion presents main findings of the research.

2 Similarity of UDC classes and UDC expressions

Universal Decimal Classification (UDC) is a scheme of classification which is used for description of library resources (McIlwaine, 1997). It is worth to underline that the UDC system can be treated as a model of the whole knowledge and its structure (Rafferty, 2001). The knowledge is represented by the hierarchical structure (the knowledge tree) composed of nodes corresponding to fields or subfields of the knowledge. Every node in the knowledge tree is described by a UDC class identifier. The UDC class can be treated as an essential element of the knowledge. Every class has an identifier which have form of a sequence of digits which reflects the position of a given class in the knowledge tree. The knowledge is divided into fields represented by main classes indicated by one-digit identifiers (from 0 to 9). Descendants of every main class (which represent subfields of fields described by the main classes) have two-digit identifiers (the first digit represents the ancestor class and the second is a consecutive number of a descendant). This schema is used also for lower levels of the knowledge tree. For clarity a dot symbol is put between every group of three digits.

Every publication is characterized by an UDC expression which is composed of one or more UDC class identifiers and auxiliary symbols. Auxiliary symbols describe form, language, place of issue of the publication or define particulars of UDC classes (e.g. related to time or localization).

The problem of similarity calculation between UDC expressions can be solved provided that the measure of similarity between individual classes is defined. It causes that an algorithm for similarity calculation between UDC expressions should be preceded by a short description of the problem of similarity between UDC classes.

It can be assumed that an identifier of a given UDC class is represented by a sequence of digits:

$$c_i = n_1 n_2 n_3 ... n_N . \qquad (1)$$

To calculate similarity between two classes: c_i and c_j first the identifier of the class which is the lowest common ancestor (LCA) of c_i and c_j is calculated. It is consisted of the leftmost digits which are common for the identifiers of both classes. Next the formula for similarity calculation proposed by Lin can be used (Lin, 1998):

$$sim(c_i, c_j) = \frac{2 \times \ln(p(c_{LCA}))}{\ln(p(c_i)) + \ln(p(c_j))} \tag{2}$$

where $p(c_i)$ is a probability of occurring of the symbol c_i.

In order to estimate the probability of occurrence of identifiers of UDC classes the analysis of the catalogue of the National Library of Poland was performed. The value of probability was calculated using the formula:

$$p(c_i) = \frac{n(c_i)}{N} \tag{3}$$

where $n(c_i)$ is a number of occurrence of the symbol c_i and N is a number of all UDC codes in the part of the catalogue which was taken into account during calculation.

The procedure of similarity estimation between UDC expressions is composed of two steps. First, identifiers of all UDC classes from a given expression are extracted. Next, a similarity coefficient is calculated.

Having two UDC expressions: $expr_1$ and $expr_2$ two sets can be defined: a set S_1 containing class identifiers extracted from $expr_1$:

$$S_1 = \{c_1, c_2, ..., c_N\} \tag{4}$$

and a set S_2 with identifiers extracted from $expr_2$:

$$S_2 = \{c_1, c_2, ..., c_M\}. \tag{5}$$

Then the similarity calculation between expressions $expr_1$ and $expr_2$ can be performed with the use of the formula (6) (Lula et al., 2014):

$$sim(expr_1, expr_2) = \frac{\sum_{i=1}^{N} \min_j (sim(c_i, c_j)) + \sum_{j=1}^{M} \min_i (sim(c_i, c_j))}{N + M}. \tag{6}$$

The formula (6) allows to build a similarity matrix for a given set of UDC expressions. Because all similarity coefficients are normalized to the range $[0;1]$ it possible to convert them into measures of distances subtracting them from value one. Matrix of distances calculated with the use of the algorithm presented above allows to perform cluster analysis of monographs based on their UDC codes.

3 Similarity of textual descriptions of monographs

Let’s assume, that a set of book descriptions prepared in the Polish language is available. To perform a comparison analysis of information content of descriptions, a similarity matrix can be calculated. The calculation of similarity matrix between descriptions consists of several stages (Manning and Schütze, 1999; Nasukawa and Nagano, 2001; Tuchowski et al., 2011):

1. Converting all texts into plain text format using UTF-8 coding system,
2. Performing a lemmatization process. During this step of analysis for every word its dictionary form (lemma) is identified. For lemmatization of documents in Polish several software packages are available (standalone applications or libraries which can be linked to other pieces of software). In the research describe here, Morfologik[4] package was used.
3. Creating a document-term matrix with rows representing documents and columns corresponding to words which appear at least in 2 and not more than in 6 documents with elements calculated as:

$$w_{ij} = f_{ij} \times \log\left(\frac{N}{n_j}\right) \tag{7}$$

   where f_{ij} indicates how many times a word j occurs in a document i, n_j is a number of documents containing a word j and N is a number of all documents.
4. Calculation the distance matrix for documents represented by rows of document-term matrix using Euclidean formula.

The matrix of distances constitutes the basis for cluster analysis of textual description of monographs.

4 Comparison of hierarchical clustering results with the Fowlkes-Mallows schema

The Fowlkes-Mallows schema for comparison of clustering results is presented in (Halkidi et al., 2001). Having results of a hierarchical clustering method for a set of N objects we can obtain the division of objects into k groups (where $k = 2,..., N-1$).

To compare the results of two clusterings Fowlkes and Mallows propose to calculate similarity index between divisions into the same number of groups obtained in both clustering processes. Taking into account two divisions of objects into k groups the Fowlkes-Mallows similarity index can be defined as:

$$FM(k) = \sqrt{\frac{T}{T+F_1} \times \frac{T}{T+F_2}} \tag{8}$$

[4] http://morfologik.blogspot.com/.

where:

T - number of objects that fall into the same groups in two studied divisions,

F_1 - number of objects that fall into the same group in the first division and simultaneously fall into different groups in the second division,

F_2- number of objects that fall into different groups in the first division and simultaneously fall into the same group in the second division.

The value of $FM(k)$ is normalized to the range $[0;1]$ and its higher value indicates a greater similarity of two studied divisions. The similarity between two clusterings can be shown by analysing and plotting $FM(k)$ versus k for all $k = 2,..., N-1$.

5 Empirical analysis of convergence between UDC codes and textual description of academic textbooks

The set of arbitrarily chosen academic monographs considered in the empirical research was composed of 18 books related to the field of social sciences. The list presented below contains main information (author, title and UDC description) about every publication. Numbers assigned to consecutive positions on the list will serve as book identifiers.

1. Mruk H, Rutkowski I., *Strategia produktu*, 339.138:658.1/.5::66/69
2. Mazurek-Łopacińska K., *Badania marketingowe. teoria i praktyka*, 339.138(075.8)
3. Florek L., *Prawo pracy*, 349.2(438)(075.8)
4. Osińska M., *Ekonometria finansowa*, 330.43:336:51](075.8)
5. Waltoś S., *Proces karny. Zarys systemu*, 343.13(438)(075.8)
6. Dmowski A., et al., *Podstawy finansów i bankowości*, 336(075.8)
7. Mróz T., Stec M., *Prawo gospodarcze prywatne*, 346:347.44:347.7](438)(075.8)
8. Czarny E., *Mikroekonomia*, 330.101.542(075.8)
9. Flejterski S. et al., *Współczesna ekonomika usług*, 338.46(075.8)
10. Kończak G., Trzpiot G., *Metody statystyczne z wykorzystaniem programów komputerowych*, 311:004.42](075.8)
11. Budnikowski A., *Międzynarodowe stosunki gospodarcze*, 339.9(075.8)
12. Altkorn J., et al., *Podstawy marketingu*, 339.138(075.8)
13. Cziomer E., Zyblikiewicz L., *Zarys współczesnych stosunków międzynarodowych*, 327(4)"19"(075.8):[327.51:355.3:061.1A/Z](100-622)(075.8)
14. Witkowska D., *Podstawy ekonometrii i teorii prognozowania*, 330.43:338.27](075.8)
15. Osiatyński J., *Finanse publiczne. Ekonomia i polityka*, 336.1(075.8)

16. Błaszczuk D., *Wstęp do prognozowania i symulacji*, 338.27:330.43:519.876.5](075.8)

17. Ratajczak M., *Współczesne teorie ekonomiczne*, 330.83(075.8)

18. Florek M., *Podstawy marketingu terytorialnego*, 339.138:332.1](075.8)

In the first stage of analysis the similarity matrix between UDC descriptions was calculated (using the algorithm presented in the section 2). Probabilities of occurrence for UDC classes were estimated on the grounds of the part of the catalogues of the National Library of Poland which store the description of library resources in the field of social sciences (represented by the main table number 3 in the UDC system). First UDC descriptions were analysed with the uses of the UDC parser prepared in the Python programming language. Analysis of UDC class identifiers allowed to estimate similarities between them (formula (2)). Next UDC expressions of the monographs described in the Table 1 were analysed. Using the formula (6) the matrix of similarities between monographs was calculated (Fig. 1).

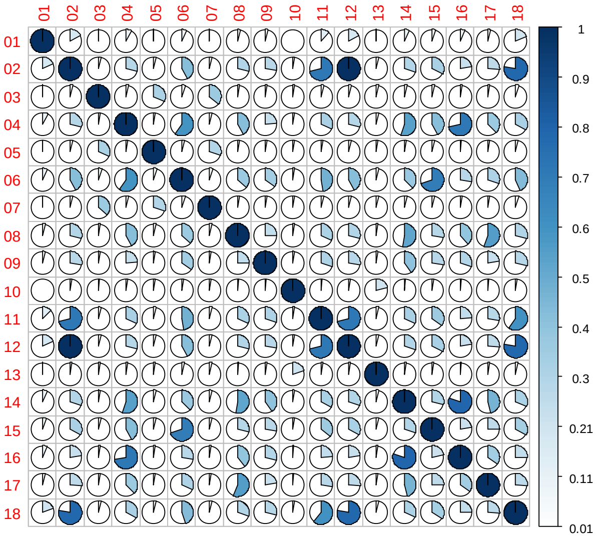

Fig. 1. Similarities between the monographs. Calculations based on UDC descriptions.

Next, the Ward method was used to perform cluster analysis of the monographs. The dendrogram is presented as Fig. 2.

After analysis of UDC codes textual descriptions of the monographs were retrieved from Polish online bookstores. Calculation of distances between textual descriptions of monographs was performed in the way presented in the section 3. The matrix of distances was normalized (by dividing all values by maximum distance) and transforming into the similarity matrix (by subtracting normalized values from 1) which is presented in the Fig. 3.

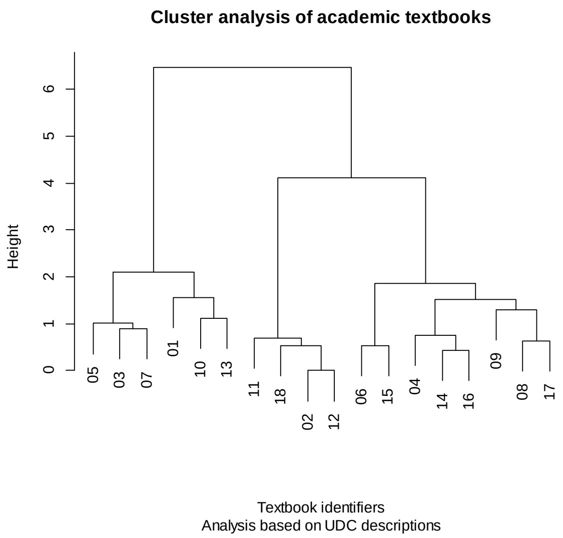

Fig. 2. Cluster analysis of the monographs. Calculations based on UDC descriptions.

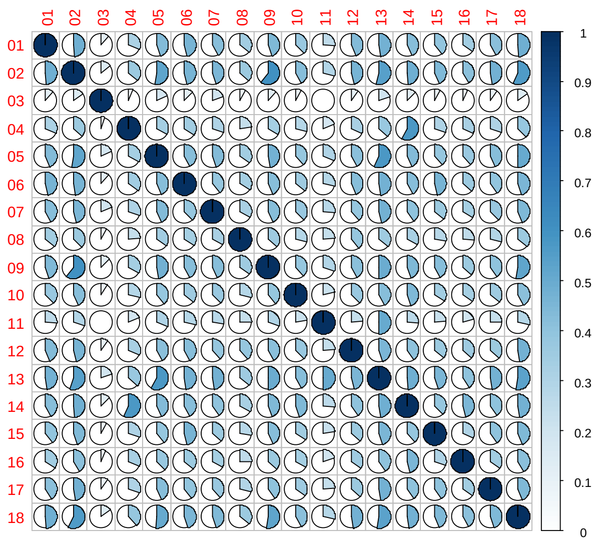

Fig. 3. Similarities between the monographs. Calculations based on textual descriptions.

Fig. 4. shows the results of cluster analysis of the set of monographs performed on the ground of their textual descriptions obtained from online stories.

The evaluation of the convergence between content description of academic textbooks in the form of textual notes and UDC expressions can be performed on the base of values of the Fowlkes-Mallows indexes presented in the Fig. 5.

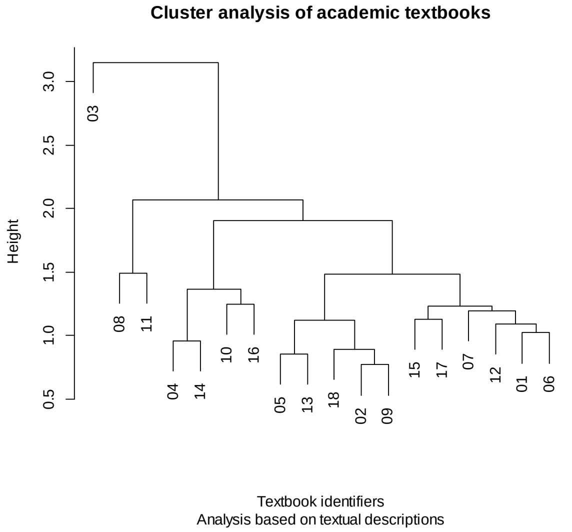

Fig. 4. Cluster analysis of monographs. Calculations based on textual descriptions.

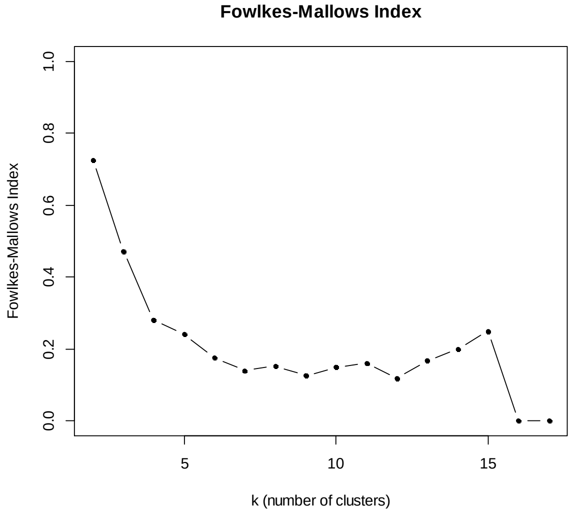

Fig. 5. Comparison of the results of monograph classification based on UDC codes and textual descriptions.

The values of Fowlkes-Mallows indexes show that the convergence between monograph content described with the help of UDC codes and monograph descriptions published by online bookshops is rather low. It is worth to notice that only comparisons of the monographs into two groups allows to obtain Fowlkes-Mallows index higher than 0.5.

Conclusions

The results obtained during analysis show that published online textual descriptions of academic textbooks do not reflect their content defined by UDC codes. Simultaneously it seems that presented in the paper the method of similarity calculation between UDC expressions was verified positively. Further investigations will be focused on theoretical and practical issues of automatic analysis of UDC description of library resources.

Acknowledgements

This research was financed by the statutory fund of the Faculty of Management, Cracow University of Economics.

References

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2-3), 107-145.

Lin, D. (1998). An Information-Theoretic Definition of Similarity. *ICML '98 Proceedings of the Fifteenth International Conference on Machine Learning,* 296-304

Lula, P., Tuchowski, J., & Wójcik, K. (2014). Similarity between Compound Objects and its Application in Recruitment Process. In: D'Amico, A., Moschella, G., (eds.) *Enterprise in Hardship Economics, Managerial and Juridical Perspectives* (9-25). Ariccia: Aracne editrice.

Manning, C. D., Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press.

McIlwaine, I., (1997). The Universal Decimal Classification: Some factors concerning its origins, development, and influence, *Journal of the American Society for Information Science,* 48(4), 331-339.

Nasukawa, T., Nagano, T., (2001). Text analysis and knowledge mining system, *IBM Systems Journal*, 40(4), 967-984.

Rafferty, P. (2001). The representation of knowledge in library classification schemes. *Knowledge Organization*, 28(4), 180-191.

Tuchowski, J., Wójcik, K., Lula, P., & Paliwoda-Pękosz, G. (2011). OBCAS - An Ontology-Based Cluster Analysis System. *Research in Systems Analysis and Design: Models and Methods*, 93, 106-112.