МРНТИ 20.15.05
УДК 004.048

https://doi.org/10.51889/2021-4.1728-7901.22

APPLYING CLUSTER ANALYSIS OF EDUCATIONAL CONTENT FOR IDENTIFYING SIMILAR DOCUMENTS

Kaibassova D.Zh.[1], Sagatbekova D.E.[1], Sagatbekova M.K.[1]*

[1]Karaganda Technical University, Karaganda, Kazakhstan
**e-mail: dindgin@mail.ru*

Abstract

This article presents the results of a study using the cluster analysis of educational content to identify disciplines (syllabuses) that are similar in content to work curricula. Existing methods made it possible to quantify the similarity of documents and led to efficient processing of large text corpora and to present interesting and relevant information. The analysis of such methods of processing text documents showed that there are a number of approaches that are applicable to solving the problem of forming educational programs. During the experimental work, 350 educational work programs of disciplines were analyzed for compliance with 120 competencies in the areas of training IT specialists. The resulting cosine distance matrix allowed us to determine similar syllabuses. This led to the task of grouping by one clustering attribute. In addition, the metric used when combining features reduces them to one feature, respectively, splitting into clusters is equivalent to grouping by this feature. This study is dedicated to solving the important and urgent task of intellectual support for the educational programs formation, which has a high complexity when processing large volumes of poorly structured information in a short period of time under constant modifications.

Keywords: cluster analysis, educational content, document similarity, frequency matrix, educational program.

Аңдатпа

Д.Ж. Кайбасова[1], Д.Е. Сағатбекова[1], М.К. Сағатбекова[1]
[1]Қарағанды техникалық университеті, Қарағанды қ., Қазақстан

БІЛІМ МАЗМҰНЫ ҚҰЖАТТАРЫНЫҢ ҰҚСАСТЫҒЫН АНЫҚТАУ ҮШІН КЛАСТЕРЛІК ТАЛДАУДЫ ҚОЛДАНУ

Бұл мақалада кластерлік талдауды білім беру құжаттарының мазмұнымен ұқсас пәндерді анықтау үшін оқу жұмыс бағдарламалары (силлабустарды) мазмұнын қолдану бойынша зерттеу нәтижелері келтірілген. Қолданыстағы әдістер құжаттардың ұқсастығын анықтауға, сонымен қатар, үлкен мәтіндік корпустарды тиімді өңдеуге және қызықты әрі маңызды ақпаратты ұсынуға мүмкіндік берді. Мәтіндік құжаттарды өңдеудің осындай әдістерін талдау білім беру бағдарламаларын қалыптастыру мәселесін шешуге қолданылатын бірқатар тәсілдердің бар екендігін көрсетті. Эксперименттік жұмыс барысында 350 пән бойынша оқу жұмыс бағдарламалары IT мамандарын даярлау бағытындағы 120 құзыреттілікке сәйкестігін талданды. Алынған косинустық қашықтықтардың матрицасы ұқсас силлабустарды анықтауға мүмкіндік берді. Бұл бір кластерлік атрибут бойынша топтастырудың міндетіне әкелді. Сонымен қатар, функцияларды біріктіру кезінде қолданылатын метрика оларды бір ерекшелікке дейін азайтады, сәйкесінше кластерге бөліну осы функция бойынша топтастыруға тең келеді. Бұл зерттеу білім беру бағдарламаларын қалыптастыру үшін зияткерлік қолдаудың маңызды және өзекті мәселелерін шешуге арналған, ол қысқа уақыт ішінде үлкен көлемді құрылымдалмаған ақпараттың үнемі түрлендірумен өңдеу кезінде өте күрделі болып табылады.

Түйін сөздер: кластерлік талдау, білім мазмұны, құжаттың ұқсастығы, жиілік матрицасы, білім беру бағдарламасы.

Аннотация

Д.Ж. Кайбасова[1], Д.Е. Сағатбекова[1], М.К. Сағатбекова[1]
[1]Карагандинский технический университет, г. Караганда, Казахстан

ПРИМЕНЕНИЕ КЛАСТЕРНОГО АНАЛИЗА ДЛЯ ОПРЕДЕЛЕНИЯ СХОЖИХ ДОКУМЕНТОВ ОБРАЗОВАТЕЛЬНОГО КОНТЕНТА

В данной статье приводятся результаты исследования применение кластерного анализа образовательного контента для выявление схожих по содержанию рабочих учебных программ дисциплин (силлабусов). Существующие методы позволили количественно оценить сходство документов и привели к эффективной обработке больших текстовых корпусов и представлять интересную и актуальную информацию. Анализ таких методов обработки текстовых документов показал, что существует ряд подходов, которые применимы для решения задачи формирования образовательных программ. В ходе проведения экспериментальных работ были

проанализированы 350 учебных рабочих программ дисциплин на соответствие 120 компетенциям по направлениям подготовки IT-специалистов. Полученная матрица косинусных расстояний позволила определить схожие силлабусы. Это привело к задаче группировки по одному признаку кластеризации. Кроме того, используемая при объединении признаков метрика сводит их к одному признаку, соответственно, разбиение на кластеры равнозначно группировке по этому признаку. Данное исследование посвящено решению важной и актуальной задачи интеллектуального обеспечения формирования образовательных программ, которая имеет высокую сложность при обработке больших объемов плохо структурированной информации за короткий промежуток времени при постоянных модификациях

Ключевые слова: кластерный анализ, образовательный контент, сходство документов, частотная матрица, образовательная программа.

Introduction

Research was conducted on the intellectual support of the process for improving the quality of the educational content forming professional competencies of higher education programs. At the same time, the identification of relevant working curricula of disciplines, taking into account the context of entities in documents with automatic extraction of entities and relations between them, has allowed to implement it without laborious processing and adaptation of knowledge bases. The subject of the study is the content of working curricula (syllabus), which are defined as a set of data that characterize the results of training and the content of the discipline. As a result of the work [1], the author created a corpus of texts from the documents of working curricula in the disciplines of the specialty "Information Systems".

Existing approaches to the intellectual support of the formation of educational programs based on ontological models of systems [2-4] based on knowledge and rules, heuristic algorithms for the automated compilation of curricula, methods of expert assessments and cognitive maps do not allow to effectively take into account and quickly track changes in the market labor, and also in the space of educational content. In turn, the formation and actualization of ontological models, systems of rules and precedents by experts for all existing subject areas the preparation of educational programs is an extremely labor-intensive process that requires the involvement of a representative composition of experts in each of the subject areas to ensure the required accuracy.

Experimental

The definition of the set of variables by which the objects in the sample are evaluated was carried out as follows: for each term in the syllabus the tf-idf index was calculated [5], the totality of which was a document vector. Then a matrix was compiled from the vectors, in which each row was a separate document. At the same time, it is necessary to make sure that each vector contains the tf-idf indicators for each term found in the document corpus. In order to obtain a matrix by informative features, some filtering operations were carried out, such as removing uninformative columns, i.e. terms that were found in only one syllabus are not common, they have been deleted.

TF (Term Frequency) is the numerical value of a given word occurrence in the current document. It is calculated by the formula:

$$TF = \frac{n_i}{\sum n_k} \tag{1}$$

where n_i is the number of a given word's occurrences, n_k is the total number of words in the document.

IDF (Inverse Term Frequency) is a numerical value that shows how often this word appears in all source of documents. Calculation formula:

$$IDF = \log\left(\frac{D}{d_i}\right) \tag{2}$$

where D is the total number of documents, and d_i are the documents in which the given word occurs [6].

The final value of the TF-IDF coefficient is equal to the product of above factors

$$TFIDF = TF \cdot IDF \tag{3}$$

Words with a high frequency within a given document and with a low frequency within a whole set of documents gain more weight. To calculate TF, a vector of normalized words is used. Each word becomes a key in the map, and the number of occurrences becomes a value.

As a result, we obtain a matrix of weights of size (10, 129), which has the rows corresponding to the syllabuses from the sample and the columns corresponding to the general terms in the collection. This matrix for the solving problem is shown in Table 1.

Definition 1. A nonnegative real function $s(X_i, X_j) = s_{ij}$ is called a similarity measure if:

1) $0 \leq s(X_i, X_j) < 1$ для $X_i \neq X_j$;
2) $s(X_i, X_j) = 1$;
3) $s(X_i, X_j) = s(X_i, X_j)$.

where $X_1, X_2, \ldots, X_n$- are presented in the form of a data matrix with the size $p \times n$:

$$X = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1n} \\ x_{21} & x_{22} & \ldots & x_{2n} \\ & \ldots & \ldots & \\ x_{p1} & x_{p2} & \ldots & x_{pn} \end{pmatrix} = (X_1, X_2, \ldots, X_n).$$

In this case, the distances between pairs of vectors $d(X_i, X_j)$ can be represented as a symmetric matrix of distances:

$$D = \begin{pmatrix} 0 & d_{12} & \ldots & d_{1n} \\ d_{21} & 0 & \ldots & d_{2n} \\ & \ldots & \ldots & \\ d_{n1} & d_{n2} & \ldots & 0 \end{pmatrix}.$$

Pairs of similarity measure values can be combined into a similarity matrix:

$$S = \begin{pmatrix} 1 & s_{12} & \ldots & s_{1n} \\ s_{21} & 1 & \ldots & s_{2n} \\ & \ldots & \ldots & \\ s_{n1} & s_{n2} & \ldots & 1 \end{pmatrix}.$$

As noted previously [7], the measure of cosine similarity of vectors is used. The resulting matrix of weights was processed using the cosine_similarity function, which takes a matrix of vector weights as input and returns a matrix of cosine distances.

Cosine similarity is a measure of similarity between two vectors of the pre-Hilbert space, which is used to measure the cosine for an angle between them. If two feature vectors A and B are given, then the cosine similarity cos (θ) can be represented using the scalar product and the norm [8]:

$$\cos(\theta) = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (4)$$

The cosine similarity value of the two documents varies in the range from 0 to 1, since the frequency of the term (TFIDF weights) cannot be negative. The angle between two term frequency vectors cannot be greater than 90°.

Definition 2. Cluster analysis is a multivariate statistical procedure that collects data containing information about a sample of objects, and then arranges objects into relatively homogeneous groups [8]. In cluster analysis, grouping characteristics are combined using some "metric". Some of them were considered in the author's previous work [7]. At the same time, grouping according to one characteristic and clustering according to a number of characteristics are brought to each other. The number of features connected during clustering can be equal to one. The work [3] proposes an original approach to non-hierarchical clustering, the so-called "island clustering", which consists in the analysis of term correlations (normal forms of words and

stable word combinations). The clustering procedure is based on testing the statistical hypothesis about the independence of the appearance of individual parterms in the texts.

This leads the task of grouping by one attribute to clustering. In addition, the metric used in combining features reduces them to one feature, respectively, splitting into clusters is equivalent to grouping by this feature.

Results and Discussion

For this, it was necessary to implement such auxiliary procedures as removing stop words from documents, stemming, determining the importance of a term in the body of documents by tf-idf characteristics of the term [5].

Table 1. Fragment of the weight matrix

	algorithm	*analysis*	*safety*	*input*	*branching*	*inter-action*	*view*
IT-infrastructure 2019.txt	*0*	*0,06680*	*0,01028*	*0,02057*	*0*	*0,03111*	*0,00835*
ADSP.txt	*0,15297*	*0*	*0*	*0,04348*	*0,07453*	*0*	*0,01176*
Artificial_Intell__audit.txt	*0,03711*	*0*	*0*	*0,02052*	*0*	*0*	*0*
Methods of operations research.txt	*0,11517*	*0,04935*	*0*	*0,03639*	*0,06129*	*0*	*0*
Modeling Software Analysis.txt	*0,00986*	*0,12825*	*0,01215*	*0*	*0,01562*	*0*	*0,00986*
Designing IS_rus.txt	*0*	*0,09512*	*0,01673*	*0,06695*	*0*	*0,02151*	*0,01358*
Artificial_Intell__management.txt	*0,03793*	*0*	*0*	*0,02098*	*0*	*0*	*0*
DSS.txt	*0,03439*	*0,13757*	*0*	*0,01902*	*0*	*0*	*0*
Operating Systems.txt	*0*	*0,17435*	*0,01431*	*0,05726*	*0*	*0,03681*	*0,05811*
Database management system Oracle.txt	*0*	*0*	*0*	*0*	*0*	*0*	*0,02851*

When the task of choosing variables (features) and objects (syllabuses) is completed, the values of the measure of similarity between syllabuses can be produced.

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	1.00000	0.369214	0.371486	0.382330	0.329737	0.428720	0.376852	0.326614	0.457715	0.084785
1	0.369214	1.00000	0.450268	0.539880	0.629847	0.136864	0.454336	0.399773	0.368217	0.286247
2	**0.371486**	**0.450268**	**1.00000**	**0.478797**	**0.520549**	**0.243711**	**0.997007**	**0.614417**	**0.570959**	**0.093727**
3	0.382330	0.539880	0.478797	1.00000	0.650349	0.121331	0.482240	0.605263	0.353657	0.096490
4	0.329737	0.629847	0.520549	0.650349	1.00000	0.148876	0.523596	0.512596	0.451146	0.125639
5	0.428720	0.136864	0.243711	0.121331	0.148876	1.00000	0.250872	0.132556	0.339926	0.150454
6	**0.376852**	**0.454336**	**0.997007**	**0.482240**	**0.523596**	**0.250872**	**1.00000**	**0.593572**	**0.575720**	**0.096691**
7	0.326614	0.399773	0.614417	0.605263	0.512596	0.132556	0.593572	1.00000	0.403258	0.078889
8	0.457715	0.368217	0.570959	0.353657	0.451146	0.339926	0.575720	0.403258	1.00000	0.024300
9	0.084785	0.286247	0.093727	0.096490	0.125639	0.150454	0.096691	0.078889	0.024300	1.00000

Figure 1. Cosine distance matrix

The matrix of cosine distances obtained during testing a sample of 10 syllabuses is showed in fig. 1. During the research of the data, it was revealed that the documents under indexes 2 and 6 are the most similar.

The sequential clustering process begins with n objects; then the two least distant (closest) objects are combined into one cluster and the number of clusters becomes n-1. The process is repeated until all n objects fall into one cluster containing all objects.

In a tree diagram, objects are vertically on the left and clustering results are on the right. Distance or similarity values corresponding to the construction of new clusters are plotted along a horizontal line over the dendogram. Having n objects, it is possible to build a large number of tree diagrams that corresponds to a given clustering procedure, however, there is only one tree diagram for a given distance or similarity matrix.

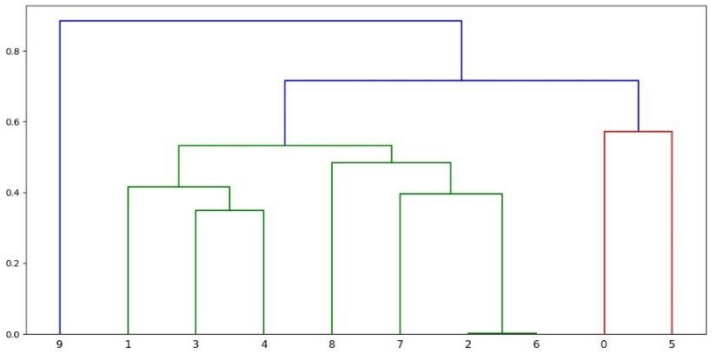

Figure 2. Dendogram of syllabus clustering

As shown in Fig. 2, the dendogram corresponds to the case of ten objects (n = 10) and p characteristics (features). Objects 2 and 6 are the closest (least distant from each other), and therefore are combined into one cluster at a proximity level of 0.009. Objects 3 and 4 merge at 0.3703. At this step, there are 8 clusters: (2, 6), (0), (5), (9), (1), (3, 4), (8), (7). At the third step of the process, clusters are formed (7, 2, 6), (1, 3, 4), (9), (8), (0), (5) At the fourth step of the process, clusters (2, 6, 7, 8), (1, 3, 4), (9), (0), (5), which correspond to the proximity level of 0, 5001. Finally, all objects are grouped into one cluster at a level of 0.923.

As a result, it was revealed that there are methods to quantify the similarity of documents that can quickly and efficiently process large corpuses and present interesting and relevant information. Analysis of existing methods of processing text documents showed that there are a number of approaches that are applicable to solving the problem of forming educational programs.

References:

1 Кайбасова Д.Ж. «Предварительная обработка коллекции рабочих учебных программ дисциплин для формирования корпуса текстов» - Вестник КазНИТУ, ISSN 1680-9211, № 6 (136) декабрь, 2019, стр. 541-546.

2 Bakanova A., Letov N.E., KaibassovaD., KuzminK.S., LoginovK.V., Shikov A.N. The use of Ontologies in the Development of a Mobile E-Learning Application in the Process of Staff Adaptation, International Journal of Recent Technology and Engineering (IJRTE), Vol 8 Issue-2S10, 2019, pp 780-789.

3 Chung H., Kim J. An Ontological Approach for Ssemantic Modelling of Curriculum and Syllabus in Higher Education // International Journal of Information and Education Technology. Vol 6, No 5, 2016, pp 365–369.

4 Oprea M. On the Use of Educational Ontologies as Support Tools for Didactical Activities // Proceedings of the International Conference on Virtual Learning (ICVL2012), Nov. 2012, pp 67–73.

5 Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. Автоматическая обработка текстов на естественном языке и анализ данных: уч.пособие. М.: НИУ ВШЭ, 2017. – 269 с.

6 Сокэл Р.Р. Кластер-анализ и классификация: предпосылки и основные направления. Классификация и кластер // Под ред. Дж.Вэн Райзина М: Мир, 1980, с. 57-79

7 Kaibassova D., L. La, A. Smagulova, L. Lisitsyna, A. Shikov, M. Nurtay/ Methods and algorithms of analyzing syllabuses for educational programs forming intellectual system/ Journal of theoretical and applied information technology, vol 98, No 05, 2020, pp 876-888.

8 https://ru.wikipedia.org

References:

1 Kaibassova D.Zh. (2019) Predvaritel'naja obrabotka kollekcii rabochih uchebnyh programm disciplin dlja formirovanija korpusa tekstov [Pre-processing of a collection of working curricula of disciplines for the formation of a text corpus] Vestnik KazNITU, № 6 (136), 541-546. (In Russian)

2 Bakanova A., Letov N.E., KaibassovaD., KuzminK.S., LoginovK.V., Shikov A.N. The use of Ontologies in the Development of a Mobile E-Learning Application in the Process of Staff Adaptation, International Journal of Recent Technology and Engineering (IJRTE), vol 8 Issue-2S10, 2019, pp 780-789.

3 Chung H., Kim J. An Ontological Approach for Ssemantic Modelling of Curriculum and Syllabus in Higher Education // International Journal of Information and Education Technology. Vol. 6, No 5, 2016, pp 365–369.

4 Oprea M. On the Use of Educational Ontologies as Support Tools for Didactical Activities // Proceedings of the International Conference on Virtual Learning (ICVL2012), Nov. 2012, pp 67–73.

5 Bol'shakova E.I., Voroncov K.V., Efremova N.Je., Klyshinskij Je.S., Lukashevich N.V., Sapin A.S. (2017) Avtomaticheskaja obrabotka tekstov na estestvennom jazyke i analiz dannyh: uch.posobie [Automated Natural Language Processing and Data Analysis: A Tutorial]. M.: NIU VShJe, 269. (In Russian)

6 Sokjel R.R. (1980) Klaster-analiz i klassifikacija: predposylki i osnovnye napravlenija. Klassifikacija i klaster [Cluster analysis and classification: preconditions and main directions. Classification and cluster] // Pod red. Dzh.Vjen Rajzina M: Mir, 57-79 (In Russian)

7 D. Kaibassova, L. La, A. Smagulova, L. Lisitsyna, A. Shikov, M. Nurtay/ Methods and algorithms of analyzing syllabuses for educational programs forming intellectual system/ Journal of theoretical and applied information technology, vol 98. No 05, 2020, pp 876-888.

8 https://ru.wikipedia.org