THE CURRICULUM CONTENT ANALYSIS FOR THE CONSTRUCTION OF THE TEACHING PROCESS

ANNA KEMPA, BOGNA ZACNY

Summary

Increasing discretion and an individual shape of the educational processes at modern universities require more efficiency in knowledge management about these teaching processes. One of the unwanted situations is an excessive duplication of the material taught and observed even on the same educational path. This problem became an inspiration for the presented work. Standard statistical methods have been applied for the analysis. The results allowed us to verify the curriculum contents of subjects contained in the syllabi and make the preliminary diagnosis of automation analysis of curricula content with text mining methods. Considerations concerning the extension of the existing system supporting the creation of syllabi with methods verifying their content and an initial concept of this system are contained in the final section of this article.

Keywords: curriculum content analysis, text mining, knowledge management

1. Introduction

The basic ‘product’ of universities is knowledge, which is made available in the form of teaching services, the results of scientific research, expert reports and implementations for economic practice and science. The conception of knowledge management in the case of universities applies to all the basic fields of its activity; this study focused mainly on one of them – the teaching process.

The reform of higher education which came into effect in October 2011 involves an increasing curricular autonomy of universities. Rigid central education standards are supposed to be eliminated, and universities will gain freedom in creating new, original fields of studies [12]. The development of the didactic process along with an increase in the flexibility of studies and the growing individualization of the studying profile requires an increased efficiency of knowledge management, both with regards to appropriate codification strategies for explicit knowledge and personalization strategies for tacit knowledge [14], [7], [3].

Creating and analysing the curriculum for the given fields of study is supported by two basic documents: program grids and syllabi. A syllabus contains a description of the subject, including its curriculum contents. An appropriate computer support for creating this document can improve knowledge management with respect to the shaping of the teaching process. Moreover, supporting the process of knowledge management through data exploration and describing the content with the help of semantic Web standards will allow for the creation of systems of recommendation for the most essential participants of this process – students [11, pp. 34–61], [15], [13].
The aim of this paper is to present the results of the statistical analysis of curriculum content included in the syllabi of the University of Economics in Katowice, carried out by the authors for a selected group of related subjects. The main reason why the authors took on this analysis was to identify curriculum content duplicates. The authors would like to take advantage of their experiences to extend the existing system supporting the creation of syllabi (S4) to the functionality of automatic curriculum content analysis. The initial concept of this extended system is included in the final part of this article.

2. Objectives of curriculum content analysis

Competition in almost every field forces producers and contractors to enrich their offers. This phenomenon, which motivates individuals to streamline their processes and accommodate themselves to the market, is obviously economically desirable. Nevertheless, in many areas it can be observed that the alleged richness and diversity means only a variety of “packages”, whereas there is actually the same content inside, or differences exist but are difficult to determine because of the applied terminology.

In the case of universities, increasing freedom during the construction of educational paths may result in some difficulty in diversifying “products” offered by universities in the form of specialties and individual subjects. The lack of clarity in this respect may be awkward both for students choosing the educational paths and for teachers realizing the teaching process. The curriculum content analysis the authors have undertaken was aimed at determining differences and similarities in curriculum contents. In the first step, this analysis referred to repeating curriculum contents on individual subjects within the given educational path.

It is notable that repeating curriculum content within the educational path itself is not a negative phenomenon. Positive and desirable roles can be pointed out:

- The role of a reminder of base issues seems to be natural, particularly if subjects with a common thematic base share a longer time on the educational path,
- Repeating content from different points of view and levels of detail allow students to better understand the learned issues,
- The system of curriculum paths can be presented with the use of a linear structure, adapted to the time axis. However, the teaching process itself is not linear. It seems to have the structure similar to the network. Some nodes learned later according to the time axis allow for the better understanding to the earlier nodes. And since both such nodes cannot be presented at the same time, the lecturer, discussing the initial node, in some contents leaves only “hooks”, where supplementing contents will be attached while learning another node. Integration of such nodes requires certain repetitions.

One can talk about a negative role of curriculum content repetition in the case when this phenomenon is not controlled, which can lead to the situation where curriculum contents available for students on the same educational path overlap too much.

Related fields overlap in some areas in a natural manner. The common areas should be skilfully encompassed in the contents of subjects on individual educational paths, so as not to go beyond the positive roles of repetitions.
Analysis of curriculum contents in order to determine the degree of content repetition (similarity), focused particularly on detecting excessive repetitions, and was for the authors – as educators – the main reason for an interest in this subject. In the course of the work, however, further objectives appeared, connected with the teaching process development. The results of the analysis of curriculum contents with respect to similarity can be applied on a wider scale. Creating a system which would generate a “similarity map” between subjects would be useful for students choosing their educational paths, for teachers working out the details of the subject content and for those arranging syllabi.

3. Making the analysis of syllabi

Using contents of syllabi, the authors made an attempt to compare subjects concerning the widely understood subject of knowledge acquisition from data collected in various repositories. As this problem is interdisciplinary, making use of the achievements of other fields of science (e.g. statistics, artificial intelligence), the teachers proposing new subjects in this field face the problem of developing the contents so they complement each other and not coincide. This problem also concerns other subject areas, but for drawing up suggestions as to conduct in such cases, the analysis of curriculum contents was narrowed to the subject connected with the area of knowledge acquisition.

In the analysis of syllabi, being text documents, it seems natural to apply methods developed in text mining. According to the assumptions, the analysis of large text bases proceeds through the following steps [10]:

1. Text refining (text pre-processing) – aimed at transforming a text into the form where it will be possible to process data using definite methods and algorithms. This step can be divided into two stages:
   • Initial processing of the analysed collection of documents,
   • Determination of the way of representation of information occurring in a document.

2. Knowledge discovery – allowing the discovery of patterns, similarities etc. in the analysed documents.

Thus, the key task is to propose the document representation in a way which enables making statistic analyses, that is bringing unstructured collections to the structured form. There are two approaches that enable transformation of a collection of documents [4], [10], [9].

The first approach, most often associated with analyses of the text mining type involves making a list of words representing each document. The words, which are selected and brought to the basic grammatical form, are called terms. Connecting of a term list for all the documents makes a term-document matrix. The matrix can contain information in a binary form (whether the given term occurs or not) or in the form of the frequency of terms (informing about the number of repetitions of the given term in the document).

The other approach, discussed at length in the next chapter, assumes the representation of the document on the basis of the structured form which is ontology or storing information in the form of documents defined in the XML language.

Owing to a small number of available syllabi (22 subjects), the authors abandoned the automatic searching of terms in favour of manual construction of the list of notions, which can be
referred to as *concepts*. Concepts are features generated from a document with the use of manual, statistic, rule or hybrid methods of categorization. They allow the identification of single words, multiple-word phrases, whole sentences and larger syntax units, which are later attributed to particular identifiers of concepts [2]. Figure 1 presents a fragment of a concept-document matrix in a binary representation, being the basis for the further analysis of syllabi.

![Figure 1. Part of a concept-document matrix](image)

Source: own elaboration.

The first step of the main analysis was to examine relationships between pairs of documents (syllabi representing subjects) in the space of concepts, treated as the analysis variables. In order to do that, a measure based on statistics $\chi^2$, the mean square contingency coefficient phi, was applied, which allows measuring the power of a relationship between two variables measured on a nominal scale. Coefficients were presented in the correlation matrix, where each line and column represents one document (see Figure 2).

Even a superficial analysis of the matrix makes it possible to identify groups of subjects which, owing to the determined concepts, are interrelated. A strong correlation of documents C and N, equal to 1, is alarming. After the detailed analysis of the syllabi it appeared that the person running those subjects was one teacher who, probably by mistake, complemented the syllabi with the same contents.
Further analysis allows determination of pairs of subjects which remain in a strong relationship, which may (but not must) result in an alarming copying of the curriculum content. Such a relationship is observed between the subjects:

- **G** – Business Intelligence Engineering (year – I; sem. – V; obligatory; lecture – 15h, practical – 30h);
- **Q** – Business Intelligence System (year – I; sem. – VI; free choice; lecture – 30h).

As subject Q can be chosen by each student, including the one who obligatory had to receive a credit for subject G, either the choice should be blocked on the level of subject choice system or the teacher should be informed about the ensuing situation. For the teacher, being aware of the fact can diversify contents, case analyses, presented software, so that the students could familiarize themselves with new areas.

Another debatable situation can be observed between the subjects:

- **I** – Knowledge Discovery Process Engineering (year – I; sem. – V; obligatory; lecture – 15h, practical – 15h);
- **T** – Data mining with SAS-Enterprise Miner (year – II; sem. – IV; obligatory; lecture – 30h, laboratory – 30h).

Here, however, the situation is different since the subjects are realized on different levels of university education. Subject I is realized at the 1st degree of studies and T at the 2nd degree. Nevertheless, it can be assumed that most students continue learning in the same field of study and at the same university. Therefore, care should be taken again with the appropriate differentiation of the contents (more advanced at the 2nd degree).

Also, the following subjects are characterized by a strong relationship:

---

![Figure 2. Matrix of phi coefficient](image-url)

Source: own elaboration.
• D – Data Warehouses (year – I; sem. – V; obligatory; lecture – 15h, practical – 15h);
• H – Data Warehouses Engineering (year – I; sem. – IV; obligatory; lecture – 30h, practical – 30h).

However, here the situation also differs from those presented above, as these subjects are realized in other fields of education. Subject D is realized in the Computer Science and Econometrics field of study and H in Engineering of Management Computer Systems. However, there is also a possibility here that some students will attend both subjects, as it is acceptable for students to study two fields at the same time. Thus, if it can be proved that those subjects realize the same contents, the student should not attend both, and the number of hours could be the recommendation for choosing the subject.

Further analysis of relationships based on the phi coefficient becomes difficult, owing to a low value of relationship, and the fact that the subjects remain in multi-dimensional relationships. Some groups of interrelated documents can be identified. For this reason, the authors used one of multivariate statistical analyses allowing the classification of documents to clusters in which syllabi are similar to one another. The hierarchical agglomerative clustering was used for classification. Since variables have a dichotomous character, the percent disagreement measure was used for the calculation of similarity of individual syllabi (measuring a distance between documents). The method of complete linkage (furthest neighbour) was applied to determine the distance between determined clusters. This is the method recommended in the situation when we suppose that objects will be grouped in separate clusters, and the clusters will not be tend to represent long “chains” [5]. As a result of the conducted grouping, a tree diagram with determined clusters was obtained (see Figure 3).

![Figure 3. Tree diagram for documents](image)

Source: own elaboration.
The distances between single documents/clusters (fragments from the horizontal axis to nodes and from the nodes lower in the hierarchy to the higher) are laid off on the vertical axis. In Figure 3 distances are presented as the percentage of the given distance towards the maximal distance, in which it is easier to compare them. The analysis of distances between clusters determined one-by-one confirms the results obtained during the analysis of relationships. Subjects N and C as well D and Q were connected in the two first steps, which indicate their large similarity.

Further analysis already requires identification of groups to which a sensible interpretation can be attributed. The most difficult element of the cluster analysis is choosing the number of such clusters. In spite of many developed weights helping in choosing the number of groups, this choice is often arbitrary and based on the visual analysis of the tree diagram. It is recommended to choose the optimal number of clusters cutting the branches of the tree where distances between clusters are becoming significantly larger. Analysing the obtained tree diagram in this respect, it is easy to observe that the distances between individual documents are relatively large and more clusters were determined at a less similarity of subjects. Nevertheless, for the initial analysis, an attempt can be made to determine a number of clusters. Eight basic groups of subjects were identified. The broken horizontal line in Figure 3 represents the point of cutting the tree branch, and the groups determined in this way were presented in Table 1. Similarity of subjects in groups not always must mean duplication of the contents; the subjects chosen for analysis concern a similar subject matter, but they are realized in different fields of study and at two degrees of studies. The teachers, however, should be aware of certain similarities to suitably shape the detailed contents of the subjects.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Year</th>
<th>Semester</th>
<th>Type</th>
<th>Subject title (Code)</th>
<th>Number of hours lecture/practical/lab</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>I</td>
<td>V</td>
<td>obligatory</td>
<td>Knowledge Discovery Process Engineering (I)</td>
<td>15/ 15/ 0</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>IV</td>
<td>obligatory</td>
<td>Data Mining with SAS-Enterprise Miner (T)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>II</td>
<td>III</td>
<td>obligatory</td>
<td>Business Intelligence (S)</td>
<td>30/ 0/ 30</td>
</tr>
<tr>
<td>3</td>
<td>I</td>
<td>IV</td>
<td>free choice</td>
<td>Data Analysis System (M)</td>
<td>30/ 0/ 0</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>VI</td>
<td>free choice</td>
<td>Automation Knowledge Acquisition from Data (O)</td>
<td>30/ 0/ 0</td>
</tr>
</tbody>
</table>
Analysis of Table 1 makes it possible to perceive similarities between subjects concerning similar problems, but it should be noted that the unprocessed data which were applied for the analysis have a substantial effect on the obtained results. Both the subjective selection of ideas (concepts) and the lack of attributing them appropriate weights which would introduce the evaluation of significances of individual problems, have a substantial effect on the accuracy and credibility of the analysis. The observed problems which appeared during the analysis of syllabus content result from the quality of those documents, which could be:

- out-of-date,
- inconsistent due to the use of different terms / abbreviations,
- inconsistent due to the author’s style (syllabi are described to a different degree of detail).

The results of the analysis may be of interest to people running the examined subjects and get them to discuss the detected relationships (it is possible that some excessive similarities are only due to the lack of updating the content of curricula), but it is necessary to make the results of the analysis available to teachers. Opportunities offered by the statistical analysis of the content of the program can support the educational process for both the lecturer and the student:

- lecturers responsible for preparing the syllabus for the course should receive feedback on the proposed content that significantly overlaps with the content of subjects already conducted – the subjects are too similar to each other (phi coefficient is too high);
- students could receive a list of subjects arranged by the degree of similarity – dendrogram presentation for the given thematic area. This would allow students to efficiently select subjects in order to build new knowledge and acquire more advanced skills.
Introduction of the proposed facilities is not possible without proper automation and linking them with existing IT tools supporting the creation of syllabi by lecturers and teaching path selection by students.

4. Conception of S4 system development towards analysis of curriculum content

In 2010 at the University of Economy in Katowice a new system called S4 was introduced, serving to create and record syllabi [6]. The direct motive for undertaking work over the system of e-syllabi was introducing a new rule of studies which radically increased the proportion of classes of the so-called free choice and thanks to this the students achieved large possibilities of individualization of their own profiles of studying.

With the use of S4 system it is possible to create a program grid and connect it with a field of study or specialization. The current syllabi are connected to the created program grid, or the task of creating new syllabi is initiated. S4 system also supports the creating of syllabi. In the course of writing the syllabi, the users apply the base of subject names and they have at their disposal dictionaries including: keywords, teaching methods, study methods, as well as the methods of student work evaluation. Using dictionaries, besides facilitating and speeding up the work over a document, is aimed at increasing the degree of data structuralization, which in turn allows for a more detailed analysis of contents [6].

Working out the analysis of curriculum contents presented in the previous chapter, the authors were based on concepts – keywords describing the issues involved in the teaching programme. These concepts, however, were not taken from the field “keywords” contained in syllabi, since the majority of syllabi contained in this field only general expressions, which can be useful for searching subjects but not for the analysis of curriculum contents. Concepts were determined on the basis of the analysis of curriculum contents and the knowledge of the subject matter for the group of the analysed subjects (connected with knowledge acquisition). This was a laborious stage of the analysis. To improve it, two solutions can be considered:

- automatic analysis of the text of syllabi to identification of terms
- creating the new field for concepts of reclassifying the current field for keywords.

Both approaches would require a certain change concerning the rules of creating syllabi. The use of text mining could not bring the desirable effects due to different approach of the authors of syllabi to describing curriculum contents (in the field “The subject content”), both with respect to the degree of description detail and the used terms. In this case, a change in the rules of creating syllabi, consisting in unification of the degree of detail and terminology, although theoretically possible, should be considered as the solution for the future. The second solution, based on words – concepts, written down by the authors of subjects, seems to be easier in realization, and additionally it may contribute to working out standards which allow changing for a completely automated system of curriculum content verification.

Adding a possibility of writing down concepts for a syllabus could perform an additional role (apart from curriculum content analysis) – namely the placed concepts could be treated as “germs”

---

1 Experimental text mining analysis for syllabi from S4 system was made at the University of Economy in Katowice – Master’s thesis by A. Pałka "Evaluation of possibilities for the use of text mining methods in analysis of text documents similarity on the example of syllabi" under supervision of Dr. K. Kania, UE Katowice, 2010.
of learning objects, which can be used at construction of e-learning courses (e.g. in SCORM standard).

A multi-level structure would allow the analysis of syllabi with respect to the hierarchy of mentioned contents in individual subjects. On one hand, the words of higher levels would enable filtration of documents and the analysis of only the selected ones. On the other hand, they would allow creating multi-level rules of association, which would help students during the choosing of subjects.

Analysing both presented propositions for changes – automatic analysis of the text and the solution going towards conceptualisation, and thus construction of ontologies, seeing the advantages and disadvantages of both of them – the authors would like to take into consideration both approaches in their further work. The conception of further research in this field involves a complementary solution – assuming a possibility of using the terms determined by text mining and the keywords currently existing in S4 system during creating concepts.

5. Conclusion

Universities employ experts in different fields who contribute to creating and storing knowledge with their competence and experience. However, the intelligence of an organization depends not only on the sum total of individual knowledge, but on organizational knowledge, the sum total of accumulated knowledge, experience and networking at the university.

The authors initiated work towards the automation of curriculum content analysis in order to facilitate access to the “common knowledge” about the teaching process – both for the teachers, people creating curricula, and students choosing their individual educational paths. The future direction in research fits into the trend referred to as data mining for education [1], [8].

Performed statistical analysis of the syllabus content allowed for the verification of knowledge accumulated in the S4 system about the curriculum content of subjects in terms of duplication of their content. Automation of the verification using text mining methods would give the opportunity to monitor the process of creating curriculum content.

The former work over the analysis of curriculum content within of the chosen area (of knowledge acquisition) helped the authors to formulate the initial assumptions for work over expanding S4 system towards the analysis of curriculum content:

- Comparison of curriculum content on the basis of analysis of syllabi using text mining,
- Creating concepts based on the detected relationships in the text mining analysis and on keywords existing in the system,
- Building ontology based on the determined concepts,
- Building the interface (using the expert system or Chabot) for three groups of users – people building / developing curricula, teachers creating syllabi and students choosing the educational path.

Building ontology will constitute the greatest challenge, since the use of initial concepts prepared on the basis of automatic methods of text analysis and statistical analyses will require the participation and cooperation of the university workers. While working out the way of this cooperation and in the course of codification of its results, the authors intend to use the experience contained in the literature from two fields: concerning the standards of construction of e-learning
trainings and concerning the methods of construction of ontologies used in different fields of business activity.

Bibliography


Streszczenie

Zwiększanie stopnia swobody w kształtowaniu procesu dydaktycznego na uczelniach wyższych wymaga coraz więcej sprawności w zarządzaniu wiedzą o procesie dydaktycznym. Głównym motywem prac autorów nad analizą treści programowych było wykrywanie nadmiernego powtarzania treści programowych na danej ścieżce kształcenia. Podczas tych prac autorki posługowały się metodami statystycznymi. Otrzymane wyniki pozwoliły na zweryfikowanie treści programowych przedmiotów zawartych w sylabusach oraz wstępne rozpoznanie zadania automatyzacji analizy zawartości sylabusów z użyciem metod text mining. Rozważania dotyczące rozszerzenia istniejącego systemu wspomagającego tworzenie sylabusów o metody weryfikacji ich treści oraz wstępna koncepcja takiego systemu zawarta została w końcowej części niniejszego artykułu.

Słowa kluczowe: analiza treści programowych, text mining, zarządzanie wiedzą

Anna Kempa
Bogna Zacny
Katedra Inżynierii Wiedzy
Wydział Informatyki i Komunikacji
Uniwersytet Ekonomiczny w Katowicach
ul. Bogucicka 3, Katowice
e-mail: kempa@ue.katowice.pl
zacny@ue.katowice.pl