Received: 12 December 2015 | Revised: 24 September 2016 | Accepted: 7 January 2017

DOI: 10.1111/exsy.12200

ARTICLE

A case study of duplications detection for educational domain thorough ad hoc search and identification NLP-based method

S.N. Mikhaylov | V.V. Chuikova | Marina V. Sokolova | A.M. Potapenko

Southwest State University, 305040, Kursk, Russian Federation

Correspondence
Marina V. Sokolova, Southwest State University, 305040, Kursk, Russian Federation.
Email: marina.v.sokolova@gmail.com

Abstract

During the organization and planning of lecture courses for a discipline, its content may be overlapped and partially delivered in more than one course. Sometimes this action causes time loss through unnecessary repeating. This paper introduces an automated tool for duplications detections adapting methods of natural language processing used for Web search. The experiment for unstructured electronic document repositories clustering for thematic duplicate identification in different documents in the case of educational domain is presented. A prototype of this Web service-based software search engine is being designed and discussed. The experiment aimed to identify thematic duplicates of various courses within one of the teaching disciplines is also presented.

KEYWORDS

evaluation, information resource, information retrieval, natural language processing, visual interface for knowledge representation

1 | INTRODUCTION

Evaluation of the semantic content of a university discipline is a time-consuming process, which should be optimized in order to relocate and save human and time resources. It is obvious that the study of the same subject in different disciplines leads to a reduction in the volume of knowledge for students as it can slow down the development of basic educational programs in accordance with the required competencies.

As a rule, assessment of the content of the disciplines within a course is carried out annually with the purpose to update, on the one hand, the bibliography studied, and, on the other hand, to reduce the number of thematic overlaps in the different courses. Similar assessment procedures require the collaborative work for nearly all the teaching staff, and it lasts a long time. Given that the number of courses within a discipline can reach several dozens, the time cost of working time for this purpose can be counted in weeks. In this regard, automation of the process of semantic content evaluation for disciplines shows and rapid detection of thematic overlaps a considerable practical interest.

One of the ways to implement the automated evaluation of the semantic content of disciplines is the use of software tools that adapt automatic procedures of knowledge mining based on natural language processing (NLP) used to extract data from the Web. Similar systems have evolved significantly in recent years, moving into various directions, which include educational data mining (Romero & Ventura, 2007; Ha, Bae, & Park, 2000; Minaei-Bidgoli, Tan, & Punch, 2004; Baker & Yacef, 2009), and research mining (Burns, Feng, & Hovy, 2008; Davcev, Cakmakov, & Cabukovski, 1992).

Existing systems of scientific research are relatively new and the quality of information retrieval (completeness and relevance) of these systems in some cases fails to meet the requirements of students, teachers, administrators, and caregivers (Kormalev, Kurshev, Suleymanova, & Trofimov, 2004; (Zakharov & Khoroshilov, 2013). In this regard, the task of organizing the retrieval of scientific information from unstructured information resources is still relevant.

To address these issues and to tackle these problems, we introduced a method of purposeful electronic documents search and information retrieval based on the analysis of similar semantic content in unstructured electronic sources.

The paper is organized as follows: Section 2 presents a brief review on current state of art in the field of research and educational information retrieval; In Section 3, more details about the educational and scientific information retrieval are given; Section 4 introduces the architecture of the Web-based search system; a case study for duplications detection in educational domain is presented in Section 5; and Section 6 presents some conclusions and suggest new directions for future research.

2 | RELATED WORKS

Naturally, Internet has become an essential part of a daily routine, in such a way that great amounts of information have migrated into the Web. Moreover, the information is defined in a natural language, and thus, it is multiply heterogeneous and diverse, as knowledge is presented in many documents in informal, semistructrured, or unstructured forms (Aleksandrov, Andreyeva, & Kuleshov, 2006; Santos & Boticario, 2015). Traditionally, one the most appropriate ways to treat this kind of information consists in using techniques of NLP (Burns et al., 2008; Papadakis, Kefalas, & Stilianakakis, 2011; Stanojevič, Tomaševič, & Vraneš, 2010; Sokolova & Fernández-Caballero, 2007; Schwitter & Tilbrook, 2008). An example of such an application could be an automated Web mining system, which uses linguistic analysis of Web content.

Today, there are many approaches for NLP Web mining techniques. In the work of Galárraga, Teflioudi, Hose, and Suchanek (2013), different variations of rule mining and learning rules are mentioned. In another studies by Khan, An, and Huang (2006) and Wang, Cheung, Lee, and Kwok (2008), some applications of Fuzzy-based approaches are discussed. Various methods which make use of techniques for semantic analysis such as semantic relation concept triplets (Paik, Liddy, Liddy, Niles, & Allen, 2000) and Noun-Verb-Noun searching are introduced in the work of Rajaraman and Tan (2002). Another solution is discussed in the work of Sauer and Roth-Berghofer (2014), and it is a case-based reasoning system, which extracts vocabularies and generate taxonomies from Linked Data sources and from web community data in the form of semistructured texts.

However, for the implementation of research and educational information retrieval, many of these systems demonstrate certain disadvantages. To be exact, many of them are limited in the number of maximum number of queries; the results of queries, especially in case of general purpose searching engines, contain much spam, including that of commercial nature (purchasing and selling of goods, and the names of which are relevant to the request).

The proposed approach can be compared with known search systems (Google, Yandex, etc.), as well as with words-searching Microsoft Word. These search engines can search for terms. With this purpose, they can be represented as documents that have defined the query terms regardless of the themes found in documents. The proposed approach makes it possible to identify the semantic similarity of the content of the document. It enables evaluation of the semantic proximity for text documents, which have been retrieved. Therefore, as a result of the request, the system displays only that found documents, which have semantic similarity with the subject of the request.

In Section 3, a method for directed Web service-based search and its application to the educational domain is introduced.

3 | EDUCATIONAL AND SCIENTIFIC INFORMATION RETRIEVAL

The aims of the scientific information retrieval are typically related to (a) initial familiarization with domain of the scientific subjects, (b) terminology learning, (c) study of the fundamentals and the state of the art of the issue, (d) purposeful search to know about the latest achievements in the area and the actual directions of research. To address all these issues, we find that information retrieval methods are the first to be applied.

Here, in this research, under the term "scientific information," we refer to the part of the information concerned with the study of a certain field of knowledge. Historically, scientific information is acquired through the scientific method. There are four essential characteristics for information to be admitted as "scientific:" independent and rigorous testing, peer review and publication, evaluation of its verisimilitude, and the degree of its acceptance within the scientific community (Benos et al., 2007). Thereafter, the set of information sources, which is used in this research, has been previously classified by the university lecturers (who participate in the experiment) as scientific information. The term "retrieval of scientific information" refers to the directed purposeful selection of information from diverse repositories, which correspond to given research topics.

The authors in the previous studies (Aleksandrov et al., 2006; Mikhaylov, 2012; Mikhaylov & Tezik, 2015) pointed out that there is a necessity of more than one linguistic analysis in order to evaluate the semantic affinity of texts as the same word in different contexts may encompass different meanings. Thus, a meaning of a word depends on the other words with which it is associated, namely, on semantic structures which include it. In this connection, it is advisable, once the linguistic analysis of the text has been completed, to analyze verbal relations presented in the text. With this purpose, word combinations, which consist of two or three words, have been formed.

The number of words in word combinations is limited because users usually create short queries when searching (usually up to four words). Usage of NLP-based methods suppose that, first, the domain of interest is described in natural language. This initial analysis facilitates defining concretely the principal and related concepts of the area with the help of expert information. Second, a vocabulary is created. Then, by adding associative relationships, which include synonyms, antonyms, etc., for each term which is inside the scientific literature description domain, a thesaurus is built. Thus, the resulting set includes semantic descriptions of each of the terms, as well as links to the rest of the terms of the vocabulary.

Information retrieval functions as it is given on Figure 1 below.

Initially, an initial user request is formulated as a natural language query. The query is then preprocessed, passing through (a) word counting, (b) thesaurus creation, (c) text segmentation, and (d) word combinations ranking. In more detail, information retrieval is performed as it is explained later on. After the natural language query is confirmed, words of interest are identified in a document. Then, a thesaurus, which uses associative relationships organized into a hierarchical structure, is created. It is represented with a conceptual graph. The latter contain words (concepts) from a text connected with arcs, which represent links between the concepts (vertices). Each link has its weight, which shows an "importance" or a rating of the corresponding link.

Figure 1 exemplifies these stages, as the NLP query is shown at the first block, and the results of the preprocessing are given in form of single words, and two- and three-word connections. Next, a conceptual graph is formed.

Initially each link between vertices is set up to one. After that, the text is revised, and counter (for each link) is incremented in case the

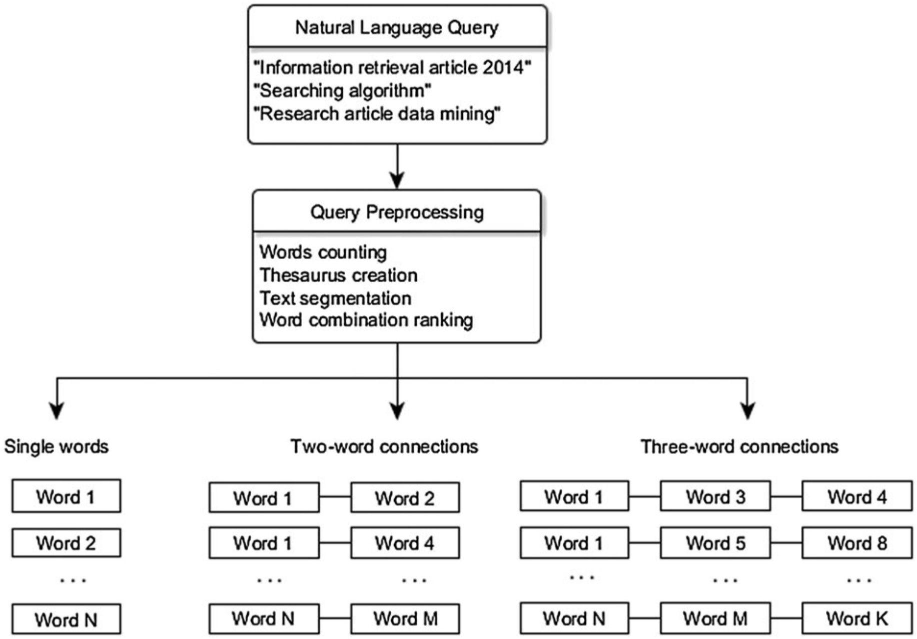


FIGURE 1 Information retrieval scheme

correspondent word pair is met. In such a way, rank for every link is calculated.

Following that, the links with a rating, which is lesser than two as well as separate words, are eliminated from the resulting set. Thus, a resulting list of words consists from connected word pairs and triples of words. The list is then sorted into descending ranking. The documents found in the Internet are stored in a set together with their URLs and short descriptions. These documents are being constantly ranked.

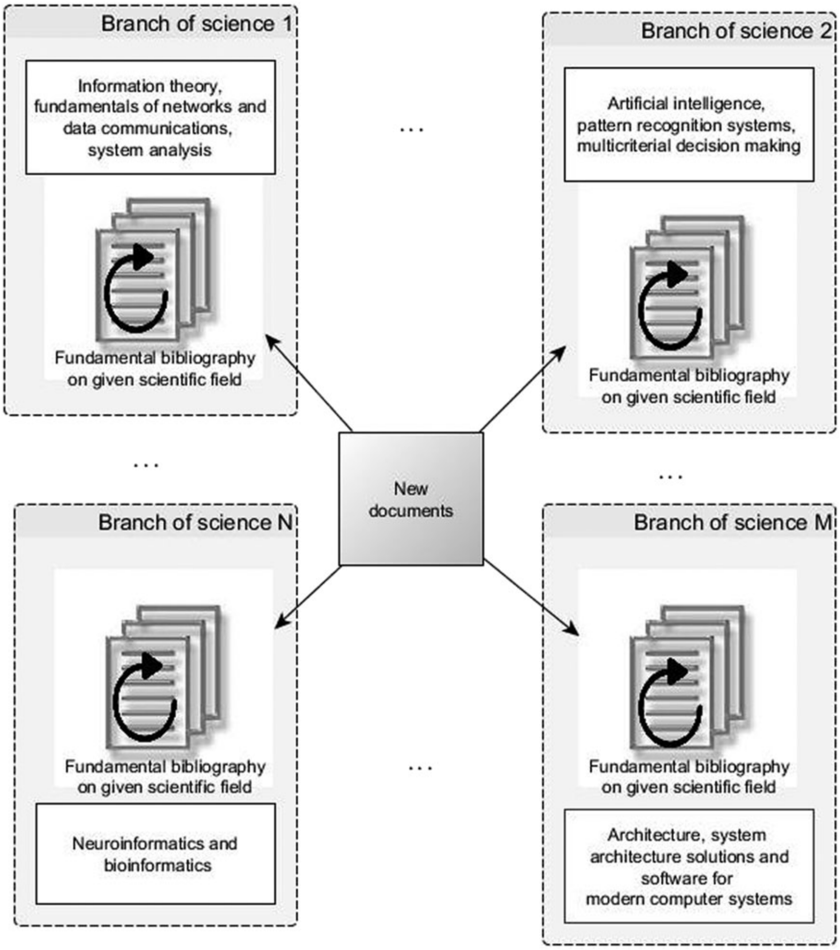

Branch of science 1
Information theory, fundamentals of networks and data communications, system analysis
Fundamental bibliography on given scientific field
Branch of science 2
Artificial intelligence, pattern recognition systems, multicriterial decision making
Fundamental bibliography on given scientific field
New documents
Branch of science N
Fundamental bibliography on given scientific field
Neuroinformatics and bioinformatics
Branch of science M
Fundamental bibliography on given scientific field
Architecture, system architecture solutions and software for modern computer systems

FIGURE 2 Identification of topics in educational domain

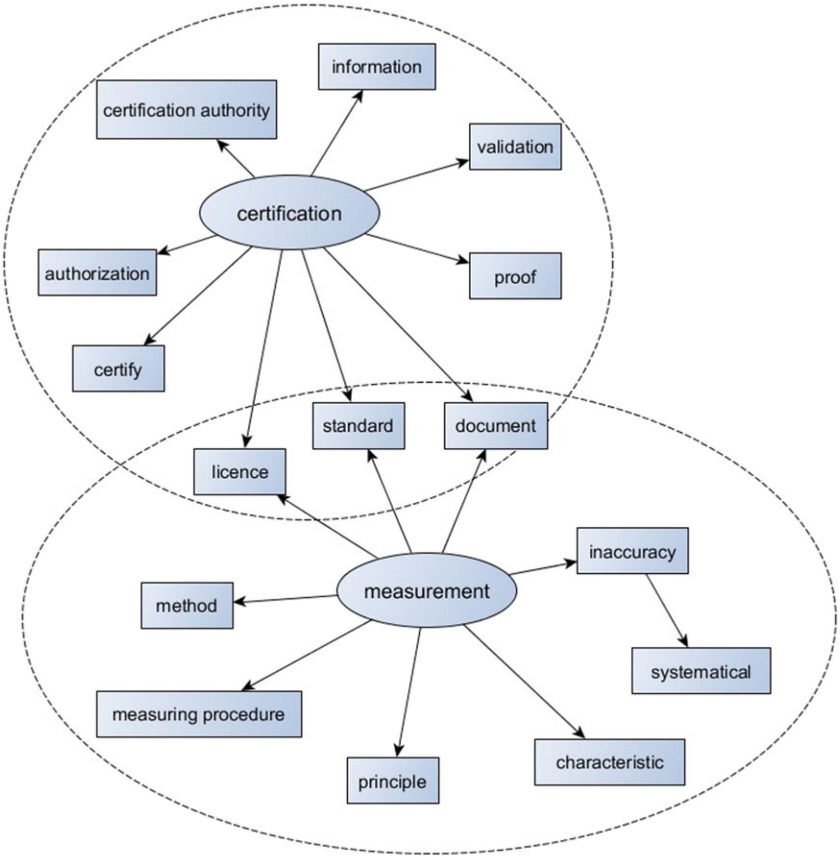


FIGURE 3 Intersection of two graphs, which represent topics

Let create a list formed from references to the documents which have been extracted from the Internet and denote the list as D={$D1$, $D2$, $D3$, …, DN}. Each document is led to a normalized form (NF), where semantically meaningful words are put into infinitive form (for verbs) and singular (for nouns). All stop words (prepositions, conjunctions, etc.) are removed. Indexed electronic documents D, constitute a set or topic S={$S1$, $S2$, $S3$, ..., SN}.

A topic is a subset of the fundamental bibliography related to the given domain. Figure 2 shows topics created for various knowledge areas (branches of science). The arrows indicate that the topics are updated, when new documents are retrieved.

Information on S, available in an electronic document, can be stored in special XML files. The final list of documents, which is generated as an answer to the user's request, is an intersection of the sets $S1$ and $S2$.

$$X = S1 \cap S2. \tag{1}$$

Figure 3 shows the intersection of the two topics S1 ("Certification") and S2 ("Measurement"). Featured spheres intersect because the same terms in a variety of contexts can have different meanings (e.g., for different study courses). Thus, the contextual meaning of the term depends on the meaning of the other words and semantic structures, which contain this given term. Here, on Figure 3, words "license", "standard", and "document" belong to both the topics, though fractionally differ in their semantic meaning for each of the topics.

In general, there are four cases, which are possible for the intersection of $S1$ and $S2$. Suppose, the correspondent documents $D1$ and $D2$ are candidate documents.

1. The set X is empty. $X = \varnothing$. In this case, the XML file-containing information about the set X contains no linking words. In other words, the documents $D1$ and $D2$ do not have common words, and the document $D2$ is not included into the X.
2. The set X coincides with the sets $S1$ and $S2$. $X = S1 = S2$. Because the XML files for both $D1$ and $D2$, as well as X, are completely equal, from which it can be concluded that $D1$ and $D2$ is the same document. Thus, the subset D2 is not included in the final set.
3. The set $S1$ is a subset of $S2$. $S1 \subset S2$. This condition is fulfilled if all the word combinations contained in the XML files for the $D1$ are also presented in the XML file for the document $D2$. This fact indicates that $D1$ is a part of $D2$ (e.g., a book chapter, a part of the collection, and so on). In this case, a user may decide if $D1$ and $D2$ should be referred as different documents. Alternatively, a search engine can include the document $D1$ in an X with a warning addressed to the user.
4. The set X takes an intermediate value between the empty set and the measure of the smaller of the sets $S1$ and $S2$. $\varnothing \le X \le (MIN\,(S1, S2))$. In this case, the XML file for X is not empty, because $D1$ and $D2$ have common words, but each of the XML document files for $D1$ and $D2$ (and hence the documents themselves) comprise words combinations, which are not included in $D2$. Thus, different documents $D1$ and $D2$ do have common words combinations, which mean that the documents contain information of closely related themes and the document $D1$ should be included into the final set for a search request.

Figure 4 shows a graphical representation of these four cases.

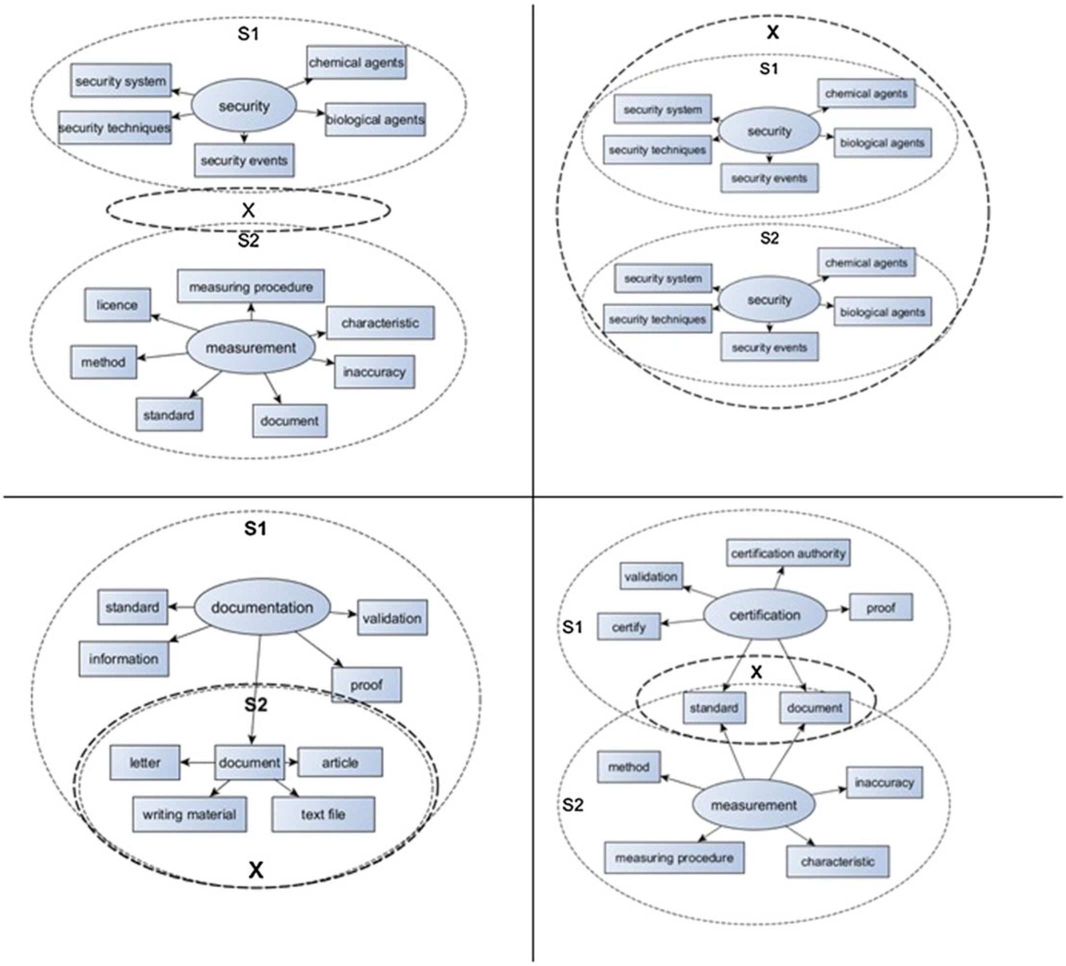


FIGURE 4 Possible intersections of topics

Moreover, if a minimum percentage of common word combinations are set and it is equal to *pmin*, documents can be considered similar in content. At the same time, it also seems reasonable to set a maximum percentage of common words *pmax* with aim to limit the possible income of documents, which could result to be are copies of each other. It is recommended to determine concrete numerical values for *pmin* and *pmax* experimentally. Although according to our preliminary findings, the value *pmin* should be about 30–40%, and *pmax* around approximately 85–90%. Last but not least, efficiency and flexibility of web search can be improved providing users with a tool for online change for values of *pmin* and *pmax*.

4 | THE ARCHITECTURE OF THE WEB-SERVICE BASED SEARCHING SYSTEM

For the implementation of Web service-based searching system (WSS), the prototype of the service-oriented architecture has been designed. Most of the modules are implemented in the programming environment Java, and some performance modules are in C++.

The composition of the architecture prototype including Web services and components is shown on Figure 5.

The entire searching process is automated; a user only determines an initial set of terms for search.

The main part of the program's components is a cross-platform and can be deployed on servers running Linux-based operating system and on Windows-based servers. The modules execute the following tasks:

- Morphology is a module for the morphological analysis, which converts documents in the form of plain text to the normalized form (NF) .
- Document indexer provides indexing for documents, which is aimed to add documents to the searchable database.
- Words dictionary provides storage and quick search of the database of words.
- Document Searcher is a tool that provides an associative search for the new documents.
- Document Classifier carries out thematic document clustering.
- Web portal is represented with a Web interface, which provides Web search and facilitates visualization of thematic clusters and terminological contents of documents.
- A crawler is a module that provides an automatic Web search for documents, their inclusion into the existing database, and their updating.

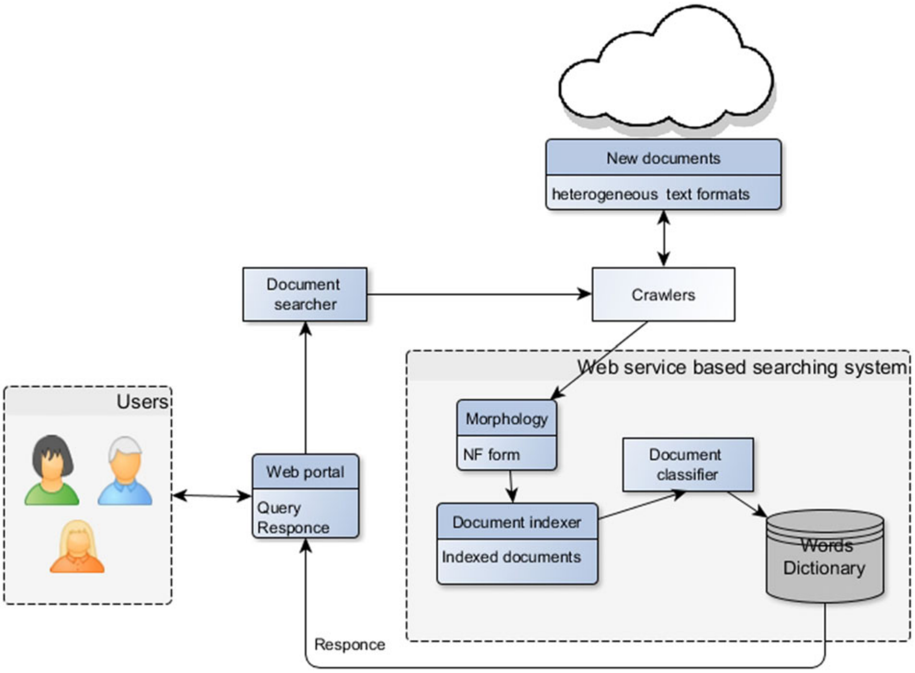


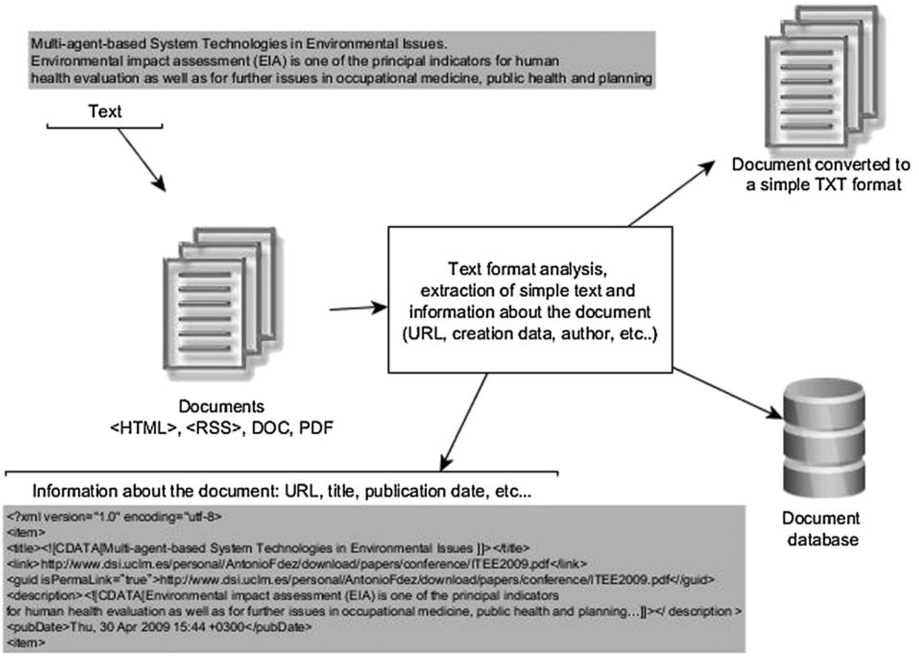
FIGURE 5 The main modules of the software prototype

The program platform can be executed under the following working profiles:

1. Administrator has all the privileges, which include upgrading, downgrading and cancelling user accounts, modifying access privileges for users and data base administrators, inviting and removing members, and changing roles.
2. Data base administrator (including interaction with final users) can modify access privileges for other members, inviting and removing members, and changing member roles.
3. User is a member who can use the software and who can invite other users.

Distribution of the working profiles enables to optimize control and to ensure that the interaction of educators, students, and professors with the system is less stressful and more efficient.

When a user interacts with a WSS through the "Web portal," she or he creates a query. The following process of web search is carried out automatically. The format conversion and postretrieval preprocessing for the retrieved documents are presented on Figures 6 and 7 and can be described as follows:



Multi-agent-based System Technologies in Environmental Issues.
Environmental impact assessment (EIA) is one of the principal indicators for human
health evaluation as well as for further issues in occupational medicine, public health and planning

Text
Document converted to a simple TXT format
Text format analysis, extraction of simple text and information about the document (URL, creation data, author, etc..)
Documents
<HTML>, <RSS>, DOC, PDF
Document database
Information about the document: URL, title, publication date, etc...

```
<?xml version="1.0" encoding="utf-8">
<item>
<title><![CDATA[Multi-agent-based System Technologies in Environmental Issues ]]></title>
<link>http://www.dsi.uclm.es/personal/AntonioFdez/download/papers/conference/ITEE2009.pdf</link>
<guid isPermaLink="true">http://www.dsi.uclm.es/personal/AntonioFdez/download/papers/conference/ITEE2009.pdf</guid>
<description><![CDATA[Environmental impact assessment (EIA) is one of the principal indicators
for human health evaluation as well as for further issues in occupational medicine, public health and planning...]]></ description >
<pubDate>Thu, 30 Apr 2009 15:44 +0300</pubDate>
<item>
```

FIGURE 6 Format conversion for a document

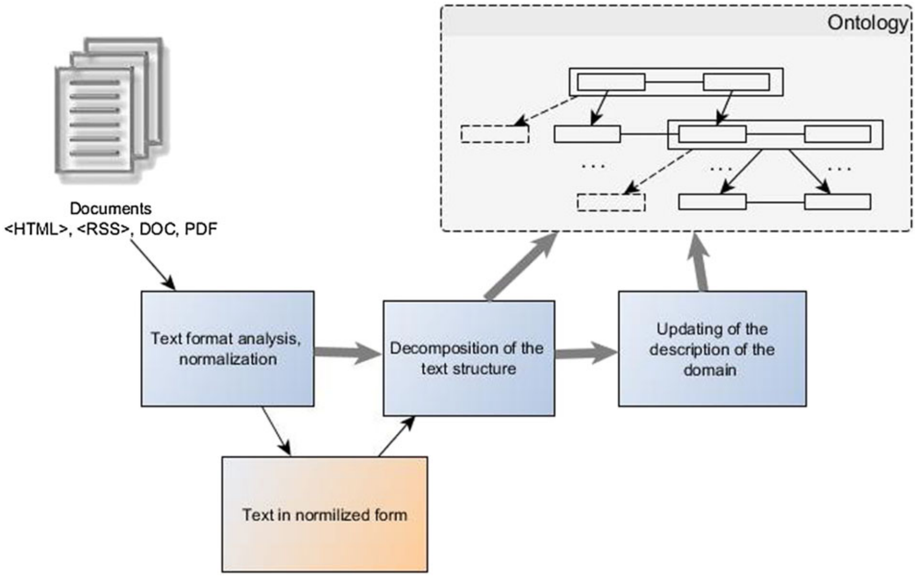

Ontology
. . .
. . .
. . .
Documents
<HTML>, <RSS>, DOC, PDF
Text format analysis, normalization
Decomposition of the text structure
Updating of the description of the domain
Text in normilized form

FIGURE 7 The process of the postretrieval preprocessing for the documents

1. The input system receives user queries with terms for "ad hoc retrieval;" the search is made in different document repositories: scientific articles, posters, monographies, conference proceedings, etc. in a variety of text formats (doc, pdf, html, etc.).
2. Format processing system converts input documents in plain text format (TXT), as it is shown on Figure 6, where a document is analyzed, and information about the file including URL, publication date title, authors, etc. is stored in the internal database. Figure 6 shows, in detail, all the steps of the format conversion for a document.
3. The resulting text is passed to block normalization and is converted to NF (see Figure 7).
4. Next, conceptual hierarchy of the given text and a thesaurus is formed.
5. Similar documents are grouped into clusters, making use of ad hoc retrieval results depending on the following attributes:
   - number of documents included in the cluster;
   - conceptual nucleus of a topic; and
   - summary of the topic, which is a brief abstract of one of the documents included in the cluster that best reflects the theme topic.
6. Semantic search for text documents is implemented at the user request. Here, semantically coherent units within a text are identified.
7. Visualize the content of the documents is realized through an interactive user interface of analytical monitoring of scientific topics. The main functions of the visualization interface are (a) to display topics scientific fields, (b) to form the conceptual core topics in a visual environment—a visual glossary, (c) to visualize thematic topics (to show the title of each topic, the conceptual core and a list of its constituent documents), and (d) to make the associative search of new publications.

5 | A CASE STUDY OF DUPLICATE THEMES DETECTION FOR EDUCATIONAL DOMAIN

The purpose of the case study is to apply WSS to an educational complex domain, where amounts of information should be processed in order to identify topics, and classify them.

It has been mentioned in Section 3 that the entire searching process is automated. Consequently, a user sets an initial set of documents, which constitute the official university program for a professional engineer degree study (which is delivered at the Department of the Information Security and Telecommunication Systems from the Southwest State University, Russian Federation). After the initial set of documents has been selected, the documents are automatically processed within the WSS without any intervention of a user.

For the case study, 12 courses, which are delivered at the department, have been selected. Each course contains several themes as it is shown in Table 1.

TABLE 1 The courses selected for the experiment and number of themes in each course

Course	Themes	Course	Themes
T1 "Antennas and wave propagation"	*T1.1–T1.20*	*T7* "Metrology, standardization, and certification"	*T7.1–T7.4*
T2 "Commutation systems"	*T2.1–T2.17*	*T8* "Modelling of telecommunication systems"	*T8.1–T8.13*
T3 "Digital signal processing"	*T3.1–T3.10*	*T9* "Networks and data transmission systems"	*T9.1–T9.17*
T4 "Electromagnetic fields and waves"	*T4.1–T4.5*	*T10* "Quantum and optical electronics"	*T10.1–T10.12*
T5 "Generating signals for mobile communication systems"	*T5.1–T5.12*	*T11* "Signal processing in mobile communication systems"	*T11.1–T11.5*
T6 "Introduction to geoinformatics"	*T6.1–T6.8*	*T12* "Theory of networks and mobile communication systems"	*T12.1–T12.10*

The principal goal of the following processing is to detect if there are similar topics within the given set of delivered courses. In other words, this study aims to find out if the contents of some themes are given in more than one course.

Let *N* be a total number of courses. Then, each theme of the each course receives an identification number *Ti*, *j*, where *i* = 1, 2, … , *n* is a number of a course, and *j* = 1, 2, … , *m* is a number of a theme, which is studied within the ith course. For example, the first course contains 20 themes (*T* 1.1–*T* 1.20); the second course includes 17 themes (*T* 2.1–*T* 2.17); the third one has 10 themes (*T* 3.1–*T* 3.10); the fourth counts with five themes (*T* 4.1–*T* 4.5), etc.

The twelve documents corresponding to the set of selected courses have been stored in a common format with the extension .docx.

On the first turn, most frequently used and referred documents for the given discipline are selected and stored. Then, the documents and themes, which they contain, are ordered and numbered. After that, all the documents are placed in the common database and are converted to the common simple textual form. Next, an ordered set of thematic queries is formed, and the documents are sorted according to the order of queries. To detect similarity of the thematic content, we carried out a sequential analysis for the available historical data. Once completed the analysis, a decision on the thematic similarity of content of the processed documents is taken. Finally, courses and themes, which contain semantic similarity, are identified.

At the next stage of the experiment, an archive, which contains information about the twelve documents, is created. Once the archive has been formed, a user can display the conceptual visual graphs for each course. Figure 8 shows a graph of the conceptual hierarchy for the course *T7* “Metrology, standardization, and certification” (a) with a minimum level of detail and b) with more details (extended graph).

As can be seen from the graph (see Figure 8a), there is one basic semantic core with the centre in the term “measurement.” The extended graph (see Figure 8b) gives a more detailed thematic and semantic description of the document.

To create a query, we created 133 requests. A query includes a brief description all the themes from all the twelve courses. In addition, a nondirected graph showing conceptual content for each course is generated. Each of the formed queries has been consequentially put into the search box of the WSS interface.

In order to evaluate the thematic similarity of the content of each of the documents in the query, we performed semantic processing of all documents. Documents, in which the semantic similarity was detected, are displayed in form in the search box. Thus, the graphical interface of the WSS gives to a user a possibility to observe search results as a list of disciplines, which have thematic overlap with the discipline under test. For example, Figure 9 gives a view on duplicate themes for the course *T1*.

Figure 10 shows the results of the experiment, demonstrating the quantitative and qualitative characteristics for detection of duplicate themes. It was a good news that there are many themes, which appeared to have few duplications with a maximum number of nine for a few ones. The quantitative criterion percentage of common themes refers to the subjective boundary above which

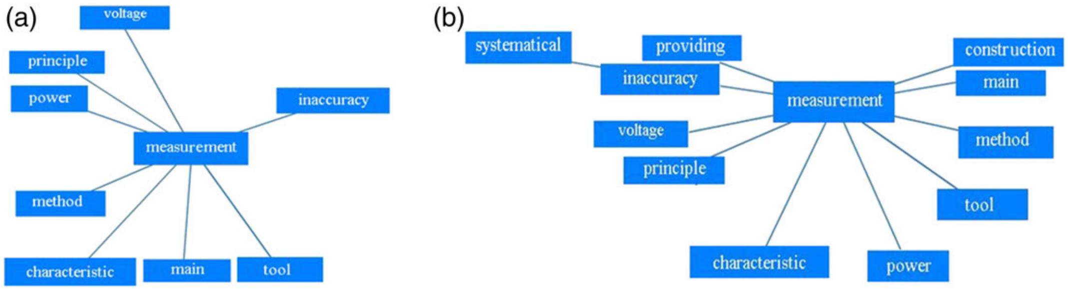


FIGURE 8 Visualization graph for the topic of the course T7 “Metrology, standardization, and certification.” (a) Minimal and (b) extended graph

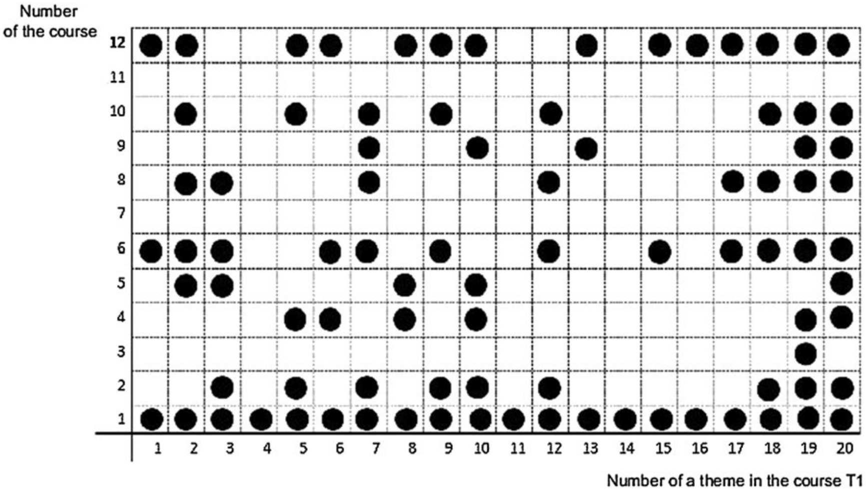


FIGURE 9 Duplications for themes within the course T1 “Antennas and wave propagation”

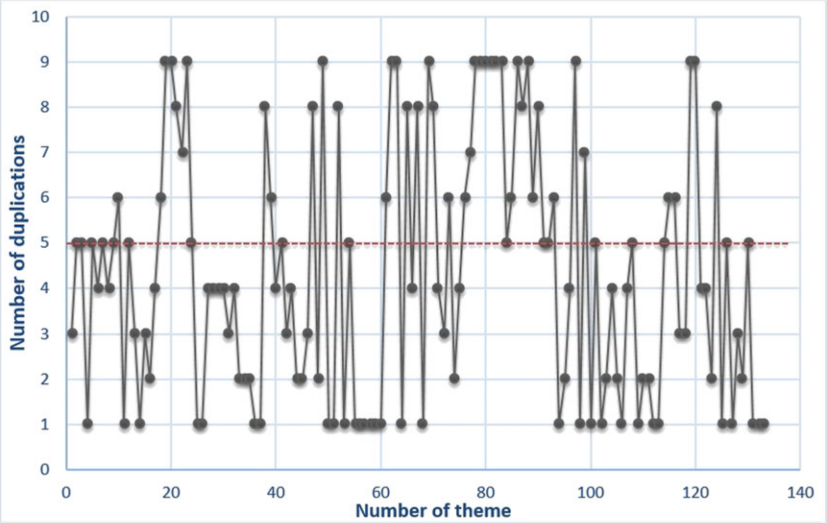


FIGURE 10 Number of duplications for each of 133 themes from 12 courses

TABLE 2 The false positive errors obtained during the experiment

Error	Theme	Error	Theme
1.18	*T6, T10*	4.2	*T5*
1.19	*T9, T3*	8.5	*T4*
2.3	*T3*	9.1	*T12, T4*
3.1	*T4*	1.11	*T1*

there are many thematic overlaps, which was set to 5. Table A1 shows possible duplications for themes and demonstrates an extract for 12 themes.

Eleven false positive errors, which are presented in the following Table 2, have been detected during the experiment. The first and the third columns of Table 2 show erroneously detected topics, and the second and the forth columns, correspondingly, the number of the theme. False negative errors were not detected. Detection of false positive errors was due to the fact that different courses within the same discipline use the standard established terminology and expressions.

Because the outcomes of this case study demonstrated that several disciplines had many overlapped themes, a group of experts (from a teaching staff of the university, where the given courses are delivered) came to a conclusion that detected duplications should be eliminated.

6 | CONCLUSIONS AND FUTURE WORK

Experimental evaluation of time required to assess the content of twelve courses and analysis of possible duplications in themes, in case it is made manually, and with the WSS shows that in the second case, the result of achieving efficiency can be increased up to five times (that fact is discussed in the study of Mikhaylov & Tezik, 2015). In practice, as the WSS has been introduced to the teaching staff during the experiment, a significant reduction of time efforts and supporting stress has been noticed.

This case study demonstrated the possibilities of the further applications in managerial decision support domains and aimed at optimizing the use of time for teaching not only in universities but in others educational centers. In addition, the WSS can be the basis for design of the more generalized architecture. The evaluation indicators of time and cost effect obtained by using WSS demonstrate significant decrease of initial material and time resources.

ACKNOWLEDGEMENT

This work was supported by the Ministry of Education and Science of Russian Federation research Grant under project 2.2491.2014/K.

REFERENCES

Aleksandrov, V. V., Andreyeva, N. A., & Kuleshov, S. V. (2006). Sistemnoye modelirovaniye. In *Metody postroyeniya informatsionno-logisticheskikh sistem* (p. 109). Saint Peterburg, Russia: Izd-vo Politekhn. un-ta SPb.

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM-Journal of Educational Data Mining*, *1*(1), 3–17.

Benos, D. J., Bashari, E., Chaves, J. M., Gaggar, A., Kapoor, N., Lafrance, M., … Qadri, Y. (2007). The ups and downs of peer review. *Advances in Physiology Education*, *31*(2), 145–152.

Burns, A. G., Feng, D., & Hovy, E. (2008). Intelligent approaches to mining the primary research literature: Techniques, systems, and examples. In *Computational intelligence in medical informatics* (pp. 17–50). Berlin Heidelberg: Springer.

Davcev, D., Cakmakov, D., & Cabukovski, V. (1992). Distributed multimedia information retrieval system. *Computer Communications*, *15*(3), 177–184.

Galárraga L. A., Teflioudi C., Hose K., & Suchanek F. (2013). AMIE: Association rule mining under incomplete evidence in ontological knowledge bases. In Proceedings of the 22nd international conference on World Wide Web (pp. 413-422). International World Wide Web Conferences Steering Committee.

Ha H., Bae S. M., & Park S. C. (2000). Web mining for distance education. In Management of Innovation and Technology, 2000. ICMIT 2000. Proceed- ings of the 2000 IEEE International Conference on (Vol. 2, pp. 715-719). IEEE.

Khan, S., An, A., & Huang, X. (2006). Hierarchical grouping of association rules and its application to a real-world domain. *International Journal of Systems Science*, *37*(13), 867–878.

Kormalev, D. A., Kurshev, Y. P., Suleymanova, Y. A., & Trofimov, I. V. (2004). Arkhitektura instrumental'nykh sredstv sistem izvlecheniya informatsii iz tekstov. In *Trudy mezhdunarodnoy konferentsii "Programmnyye sistemy: Teoriya i prilozheniya"* (pp. 49–69). Pereslavl'-Zalesskiy, M: Fizmatlit.

Mikhaylov, S. N. (2012). *Sposob tematicheskoy klasterizatsii tekstovykh dokumentov na osnove ikh infologicheskoy obrabotki, naukoyomkiye tekhnologii (science intensive technologies)* (Vol. 9) (pp. 48–51)Izdatel'stvo radiotekhnika.

Mikhaylov S.N., & Tezik K.A. (2015). Algorithm of realization of process of comparison thematic orientation of data in inform ation resources. Proceedings of the South-West State University. Management, computer facilities, Computer science. Medical instrument making, Vol.1 (14), Kursk, pp. 34–41.

Minaei-Bidgoli B., Tan P. N., & Punch W. F. (2004). Mining interesting contrast rules for a web-based educational system. In machine learning and applications, 2004. Proceedings. 2004 International Conference on (pp. 320-327). IEEE.

Paik W., Liddy E. D., Liddy J. H., Niles I. H., & Allen E. E. (2000). U.S. patent no. 6,076,088. Washington, DC: U.S. patent and trademark office.

Papadakis, N., Kefalas, P., & Stilianakakis, M. (2011). A tool for access to relational databases in natural language. *Expert Systems with Applications*, *38*(6), 7894–7900.

Rajaraman K., & Tan A. H. (2002). Knowledge discovery from texts: A concept frame graph approach. In Proceedings of the eleventh international conference on Information and knowledge management (pp. 669-671). ACM

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, *33*(1), 135–146.

Santos, O. C., & Boticario, J. G. (2015). User-centred design and educational data mining support during the recommendations elicitation process in social online learning environments. *Expert Systems*, *32*(2), 293–311.

Sauer, C. S., & Roth-Berghofer, T. (2014). Extracting knowledge from web communities and linked data for case-based reasoning systems. *Expert Systems*, *31*(5), 448–456.

Schwitter, R., & Tilbrook, M. (2008). Meaningful web annotations for humans and machines using controlled natural language. *Expert Systems*, *25*(3), 253–267.

Sokolova, M. V., & Fernández-Caballero, A. (2007). A meta-ontological framework for multi-agent systems design. In *Nature Inspired Problem-Solving Methods in Knowledge Engineering* (pp. 521–530). Berlin Heidelberg: Springer.

Stanojevič, M., Tomaševič, N., & Vraneš, S. (2010). NIMFA: Natural language implicit meaning formalization and abstraction. *Expert Systems with Applications*, *37*(12), 8172–8187.

Wang, M., Cheung, C. F., Lee, W. B., & Kwok, S. K. (2008). Mining knowledge from natural language texts using fuzzy associated concept mapping. *Information Processing Management*, *44*(5), 1707–1719.

Zakharov, S. N., & Khoroshilov, A. A. (2013). Avtomaticheskaya otsenka podobiya tematicheskogo soderzhaniya tekstov na osnove sravneniya ikh formalizovannykh smyslovykh opisaniy. *Sistemy i sredstva informatiki*, *23*(1) Moskva, 143–158.

AUTHOR BIOGRAPHIES

Dr S. N. Mikhaylov, Associate Professor, is a leader of information search group (in the context of a grant) of the department of the information security and telecommunication systems at Southwest State University (SWSU). His current research interest focuses on applying data mining-based methods to determine semantic meanings of unstructured texts. Since 2008, at SWSU, he has published over 180 papers in various international conferences and journals related to topics from his research, and he has been a member of committee of the annual department conference.

V. V Chuikova, lecturer, is a member of information search group of the department of the information security and telecommunication systems at Southwest State University (SWSU). Her current research interest focuses on applying data mining methods, information retrieval, and Web search. She has published four papers in various international conferences, and she is a member of committee of the annual department conference.

Dr. Marina. V. Sokolova is an Associate Professor at the Southwest State University (Russia). Her research interests are Web search, data mining, user-oriented techniques for affective computing, and their applications to learning environments and ambient living. She is a researcher at the LOUiSE laboratory at the Informatics Institute of Albacete (UCLM, Spain). She has participated in 12 international and national research projects, published over 35 papers in various international conferences and journals, and cochaired several workshop. She is a member of the international program committees of five international conferences.

Dr A. M. Potapenko, Associate Professor, is the head of the Department of the Information Security and Telecommunication Systems at Southwest State University (SWSU). He works in the field of data mining, Web search, ontologies, and information retrieval. He is the author of 190 papers in various international conferences and journals. He is the chair of the annual department conference.

How to cite this article: Mikhaylov SN, Chuikova VV, Sokolova MV, Potapenko AM. A case study of duplications detection for educational domain thorough ad hoc search and identification NLP-based method. *Expert Systems*. 2017;34: e12200. https://doi.org/10.1002/exsy.12200

APPENDIX A

TABLE A1 The results of all the queries for theme duplications identification

Course number	Theme number	Duplications number	Number of courses with similar themes
1	1.1	3	1, 6 ,12
	1.2	5	12, 6 ,1 ,5 ,8
	1.3	5	6, 8, 1, 5, 2
	1.4	1	1
	1.5	5	12, 1, 2, 4, 10
	1.6	4	1, 6, 4, 12
	1.7	5	8, 6, 2, 10, 1, 9
	1.8	4	12, 1, 5, 4
	1.9	5	2, 1, 10, 6, 12
	1.10	6	2, 4, 5, 9, 1, 12
	1.11	1	1
	1.12	5	6, 8, 1, 2, 10
	1.13	3	9, 12, 1
	1.14	1	1
	1.15	3	1, 12, 6
	1.16	2	1, 12
	1.17	4	1, 6, 8, 12
	1.18	6	6, 8, 12, 2, 10, 1
	1.19	9	8, 6, 12, 2, 10, 1, 9, 4, 3

(*Continues*)

TABLE A1 (Continued)

Course number	Theme number	Duplications number	Number of courses with similar themes
	1.20	9	6, 8, 12, 1, 2, 10, 9, 4, 5
2	2.1	8	4, 12, 1, 6, 8, 9, 5, 2
	2.2	7	2, 6, 8, 10, 1, 4, 9
	2.3	9	8, 6, 9, 1, 12, 2, 10, 3
	2.4	5	8, 10, 1, 2, 6
	2.5	1	2
	2.6	1	2
	2.7	4	1, 4, 9, 2
	2.8	4	8, 6, 2, 10
	2.9	4	12, 4, 1, 2
	2.10	4	10, 1, 6, 2
	2.11	3	2, 10, 1
	2.12	4	9, 8, 2, 4
	2.13	2	1, 2
	2.14	2	12, 2
	2.15	2	10, 2
	2.16	1	2
	2.17	1	2
3	3.1	8	1, 2, 3, 4, 5, 6, 8, 9
	3.2	6	2, 8, 6, 3, 10, 9
	3.3	4	4, 3, 8, 6
	3.4	5	3, 6, 8, 9, 10
	3.5	3	3, 2, 9
	3.6	4	3, 6, 9, 1
	3.7	2	3, 4
	3.8	2	3, 4
	3.9	3	3, 6, 12
	3.10	8	3, 1, 2, 10, 12, 6, 8, 4
4	4.1	2	4, 9
	4.2	9	4, 2, 6, 1, 9, 8, 12, 10, 5
	4.3	1	4
	4.4	1	4
	4.5	8	4, 1, 2, 8, 9, 10, 6, 12
5	5.1	1	5
	5.2	5	4, 4, 8, 9, 1
	5.3	1	5
	5.4	1	5
	5.5	1	5
	5.6	1	5
	5.7	1	5
	5.8	1	5
	5.9	6	1, 3, 4, 5, 6, 8
	5.10	9	5, 8, 6, 2, 1, 12, 9, 4, 3
	5.11	9	5, 6, 8, 12, 1, 9, 2, 10, 4
	5.12	1	5
6	6.1	8	6, 4, 8, 5, 2, 1, 10, 9
	6.2	4	6, 2, 1, 9
	6.3	8	6, 4, 1, 2, 8, 5, 10, 9
	6.4	1	6
	6.5	9	8, 1, 6, 2, 4, 9, 10, 12, 3
	6.6	8	6, 1, 8, 2, 10, 4, 9, 12
	6.7	4	6, 1, 4, 2
	6.8	3	1, 2, 6
7	7.1	6	4, 7, 8, 2, 9, 12
	7.2	2	4, 7
	7.3	4	7, 4, 9, 6
	7.4	6	7, 4, 2, 1, 6, 12
8	8.1	7	12, 6, 8, 2, 4, 1, 10
	8.2	9	8, 4, 6, 9, 2, 1, 12, 10, 3
	8.3	9	8, 6, 1, 2, 10, 9, 12, 4, 3
	8.4	9	8, 4, 6, 1, 2, 10, 12, 9, 3
	8.5	9	8, 9, 6, 4, 1, 2, 3, 10, 12
	8.6	9	8, 9, 6, 4, 1, 3, 2, 10, 12
	8.7	9	8, 6, 9, 4, 10, 12, 1, 2, 3
	8.8	5	9, 6, 8, 4, 12
	8.9	6	8, 9, 2, 6, 4, 3
	8.10	9	8, 6, 9, 12, 1, 4, 10, 2, 3
	8.11	8	6, 8, 1, 4, 9, 10, 2, 12
	8.12	9	9, 6, 8, 2, 5, 4, 12, 1, 10
	8.13	6	9, 8, 6, 4, 7 ,2
9	9.1	8	9, 4, 8, 6, 12, 2, 10, 1
	9.2	5	9, 6, 8, 4, 12
	9.3	5	9, 8, 6, 4, 1
	9.4	6	9, 8, 6, 4, 12, 2
	9.5	1	9
	9.6	2	9, 4
	9.7	4	9, 8, 6, 3
	9.8	9	9, 8, 1, 6, 2, 12, 4, 10, 3
	9.9	1	9
	9.10	7	9, 6, 12, 8, 4, 2, 10
	9.11	1	9
	9.12	5	9, 6, 8, 4, 2
	9.13	1	9
	9.14	2	9, 4
	9.15	4	9, 4, 5, 6
	9.16	2	9, 6
	9.17	1	9
10	10.1	4	2, 8, 6, 10
	10.2	5	10, 6, 8, 2, 1
	10.3	1	10
	10.4	2	2, 10
	10.5	2	2, 10
	10.6	1	10
	10.7	1	10
	10.8	5	10, 2, 8, 9, 3
	10.9	6	10, 8, 3, 2, 9, 12
	10.10	6	2, 10, 6, 5, 8, 3
	10.11	3	12, 5, 10
	10.12	3	12, 2, 10
11	11.1	9	4, 1, 2, 6, 8, 10, 9, 12, 11
	11.2	9	8, 2, 11, 6, 1, 12, 10, 5, 3
	11.3	4	11, 6, 9, 8
	11.4	4	11, 4, 9, 5
	11.5	2	11, 5
12	12.1	8	12, 8, 1, 4, 6, 2, 9, 10
	12.2	1	12
	12.3	5	12, 2, 4, 5, 10
	12.4	1	12
	12.5	3	12, 9, 1
	12.6	2	12, 1
	12.7	5	12, 1, 9, 6, 8
	12.8	1	12
	12.9	1	12
	12.10	1	12

(Continues)