Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 939–943

ISBN 978-83-60810-51-4

Problems of automatic processing and analysis of information from Polish legal texts

Wojciech Cyrul
Jagiellonian University
ul. Gołębia 24, 31-007 Kraków,
Poland, Email:
cyrulwojciech@hotmail.com

Tomasz Pełech-Pilichowski
AGH University of Science and
Technology, Al. Mickiewicza 30
30-059 Krakow, Poland
Email: tomek@agh.edu.pl

Piotr Potiopa
AGH University of Science and
Technology, Al. Mickiewicza 30
30-059 Krakow, Poland
Email: ppotiopa@zarz.agh.edu.pl

***Abstract*—In the paper, problems of legal information digitalization are investigated. Conditions for extraction information from legal texts related to the common ones processing (non-legal terms) are outlined. Sample results of similarity analysis are presented. Further research aimed at semantic analysis of legal texts are outlined.**

I. Introduction

PROBLEMS concerning digitalization of legal information integrates contemporary legal theory and legal informatics. The Internet opened the possibility to promulgate and communicate legal texts in a machine-readable form. It also promised practical realization of the idea of more efficient and more transparent legislation, available on-line. Since that time, IT/ICT start to play an ever-growing role, both in legal discourse and in legislative processes. However, the use of IT in law raises numerous new problems: a question of standards used in the process of digitalization of normative acts or other legal documents; finding a method of adding computer-processed information to these texts; and finding tools which will help support both their preparation and management in a standard compliant way [15]. The specificity and complexity of legal language (i.a. multimeaning words, complex and temporary semantic dependencies, direct translation difficulties) entail that typical text processing tools and technologies have a limited application. An availability of legal text sources gives an opportunity to extract aggregated information, but due to both, the technical capabilities and the requirements for the expected high output data quality, legal information searching is performed in a syntactic way, whereas research on semantic searching is advisable.

The goal of the paper is to analyze a possibility of performing reliable similarity analysis of sample *polish legal acts*, based on approaches typically adopted to common (English) language processing, and to report the results in the context of long-term research. Problems of legal information digitalization are discussed; conditions for legal information extraction related to the non-legal terms are outlined; selected data processing steps, including reading legal texts available on-line, are discussed. Sample results of similarity analysis focused on distinguishing groups of legal acts are presented. Further research aimed at semantic analysis of legal texts are outlined.

II. Digitalization of Legal Information

From the beginning, the debate on the complexity of digitalization of law (started in the end of the 1960's) had a pragmatic character and focused mainly on possible applications of IT and ICT in law (see [14],[7],[9]). This instrumental approach became even more pervasive in the late 1990's when the discussion began to be dominated by the intricacies related to building and functioning of legal databases and complications connected with regulation of the computerization of law ([16]). As a result of this development, the Act of June 20, 2000, on promulgation of normative acts and some other legal acts (Dz. U. Nr 62, poz. 718) states that official legal journals and collections of local law should be issued and stored not only in the printed but also in the electronic form[1]. In consequence, since January 1, 2012, Polish law has been published exclusively in an electronic form and communicated on Internet, like in several other EU countries. These change were preceded both by development of several technical standards and by extensive regulation how they should be implemented in law.

The legal framework for the technical standard was created by the Act of February 17, 2005 on computerization of the activities of the entities implementing public tasks and the Regulation of the Prime Minister of December 27, 2011 (Dz.U. Nr 289, Poz. 1699) on the technical requirements for electronic documents, which contain normative acts and other legal acts, for official journals published in the electronic form and for means of electronic communication and electronic data carriers. Regretfully, existing regulations focus more on security and integrity of electronic legal documents than on increasing effectiveness of managing information in the legislative process and enhancing cooperation between different legal and social agents contributing to the process. Moreover, electronic legal documents containing normative acts and other legal acts announced on the basis of the Act of July 20, 2000 on publication of normative acts and certain other legal acts, shall be prepared as structural texts in a XML 1.0 format. The XDS schemes of such documents, although being developed and published by the government, have only the status of recommendations, and thus are legally not binding. According to the law, electronic documents containing legal texts should stay in line with the

[1]The statute is amended by the Act of March 4, 2011 r, on amendment of the statute on promulgation of normative acts and some other legal acts (Dz.U. Nr 177, poz. 676).

patterns of legal acts published in central repository of patterns of electronic documents and on the web page *www.dziennikiurzedowe.gov.pl*, the patterns as such do not need to be used when drafting a legal text[2]. The law requires only that the structure of a legal act, defined in XML, should be in accordance with the structure of legal acts as determined in the Regulation of the Prime Minister of June 20, 2002 on legislative technique (Dz. U. Nr 100, poz. 908) and promulgated in a PDF format or in a XML format enabling visualization of the content in the form of a PDF file. This situation creates a risk that electronic documents containing legal texts will not be fully compatible, and therefore hinders effective management and processing legal information in future. It poses also a significant problem, for the reason that the information included in the legal resource is complex and heteronymous. Effective and appropriated interconnection of its different layers requires not only a multi-layered architecture for modeling legal documents but also existence and implementation of shared and open standards for legal knowledge representation (see [11]).

This situation changes not only the way, in which law is being made and applied but also alter the ways in which legal information is searched, aggregated, analyzed stored, erased and communicated. Respectively, new standardized forms of legal documents are being formed, new forms of notification and delivering of legal decisions in legal proceedings are being developed, new tools are being used to encode and to manage legal information, and new electronic public archives or registers are being created.

Above mentioned developments open up an important field of activity for informatics. It is especially the case because to manage effectively legal information in the digital form, normative texts and various types of legal documents must be drafted and disseminated with the use of computers. Furthermore, the special status of legal information requires development and implementation of standards, which secure both its authenticity and integrity.

Machine-readable standards must take into account different aspects and particularities of legal texts, i.e. the appearance, the structure or the different normative status of particular types of legal texts. Although, these standards are being developed and gradually implemented, also in Poland, there is surprisingly little discussion about what standards should be applied, or how existing standards should be modified in order to fit best into the present and the future needs of both Polish state and the addressees of law. More discussion is also needed with respect to both the potential problems resulting from the specific semantic, syntactic and pragmatic nuances of legal texts and the impact of implemented technologies on legal practices. One must remember that status, types, structures and relations between different legal texts are determined by law. Furthermore, while some links or relations between different legal provisions are clearly defined by law others are identifiable only on the conceptual level. Thus, there is no doubt that the standards implemented today for digitalization of legal information will determine the possibility, scope, and costs of further developments. The awareness of these factors is important also because new standards are still being developed, aiming to meet new expectations of business people and lawyers. For example, there is a growing need for enabling an online access to legal information with the use of different electronic devices and for development of tools allowing computerization of aggregation of legal information and its analysis, taking into consideration the semantic dimension of law.

Both, the scope of digitalization and computerization of legal information depend on existence and quality of so called *knowledge representation languages*. They permit a description of legal information in a way that enables the machines its application. Information systems could not only support actions of individuals through versioning, finding or evaluating particular information, but also apply directly and automatically legal provisions[3]. With XML, a machine-readable representation of the formal structure of legal documents is also possible. There are still problems with making the machines "understand" and apply legal texts, since there are neither universal legal ontology nor the universally acceptable model of legal reasoning. The already existing tools have either a limited scope of application or are unable to computerize legal practice to the extent which would not create any doubts about the validity of conclusions.

One can see that efficiency and the scope of computerization of legal information depends, to large extent, on the existence of shared standards and conventions enabling the technology neutral representation. Standardization could increase the efficiency of generation, presentation, accessibility and description of legal documents.

III. Morpho-syntactic Text Analysis

A. Processing of polish legal texts:` steps and tools

Discussed in previous chapters, agreed by law, standards of legal texts formats are the great opportunity to create tools able to search legal information and to build semantic relations between the particular sets of texts. Therefore, considering efficient information analysis and extraction with automated tools, the following areas of obtaining information seem to be vital to investigate:

- •searching keyword phrases in legal texts and locating them in selected blocks of legal texts,
- •automatically examining relations between legal texts (such as similarity),
- •the law thematic area recognition on the basis of the legal test samples,
- •legal texts semantic analysis and (semi)automatic legal ontology construction.

As mentioned above (see Introduction), adopted in the Polish law regulations on the media publication of legal texts say that the format adopted in the generation and publication of these texts is XML and PDF (for publication of

[2]Currently the XSDs of polish legal acts is developer one the electronic platform of public administration services (ePUAP); *http://epuap.gov.pl*

[3]To some extent, it is already possible since any graphical sign can be represented in an alphanumerical way (ASCI, Unicode). There are also naming convention permitting univocal identifications of all available online legal documents or documents to which legal documents directly refer to (URL, URI).

texts on the Internet). Developed XML is compatible with the developed schema XSD file approved and published[4]. A PDF file may be generated based on XML thus further considerations mainly concern XML file processing. Therefore, it allows to assume that the processing of legal acts published on the Internet[5] will be sufficient.

Unfortunately, despite our attempts to find laws generated in XML format adopted in accordance with established XSD schema, we could not find such documents. Therefore, it became more difficult to extract information and document fields for research, however, such information simplifies the similarity analysis as: structured way of writing a legal text eliminates the problem of syntactic ambiguity in analysis and parsing of the document; structured schema makes possible to clearly separate systematization and editorials units of the legal text clearly defined in schema, respectively: "część" ("part"), "księga" ("book") etc.; facilitates the validation of the document with the adopted scheme XSD and use of tools for parsing XML documents[6].

A. Term-document Matrix

To achieve better description of dependencies between analyzed documents, the modification of the TDM matrix basic structure has been employed. TDM (term-document matrix) consists of words extracted from the input data. TF-IDF factor is based on determining the relative frequency of a term and then comparing to the inverted frequency of term calculated in the entire collection of documents.

For each term, TF (term frequency) is the relative frequency of the term w occurrences in the document d (the term rank) and the IDF (inverse document frequency) as inversely proportional to the occurrences of the term in relation to the corpus D ($d \in D$) of documents, which represents an importance of the word in the entire collection of documents. TF-ID factor is calculated as follows (eq.1) [13],[12]:

$$w_d = f_{w,d} \log(|D| / f_{w,D}) \tag{1}$$

where $f_{w,d}$ is a number of occurrences of w in d, $|D|$ is the size of document corpus $f_{w,D}$ is a number of documents containing term w. For large collections of documents, the normalization of the TF with dimensionality reduction procedures may be used.

To achieve this purpose Singular Value Decomposition (SVD) methodology is used, which is exploited in the area of dimensionality reduction, including natural language processing [3] and Latent Semantic Analysis method (LSA) [8], dedicated to acquiring semantic relations from large text datasets. The goal is to split the input m-by-n matrix A (see eq. (2)) into two orthogonal matrices U (m-by-m) and V (n-by-n), and the diagonal one S of the same size as input matrix A, with nonnegative decreasing elements (singular values) [1].

$$A = USV^T, U^T U = I, V^T V = I \tag{2}$$

Another approach is to employ the Principal Component Analysis method. Having a given n-by-n covariance matrix the principal component coefficients (loadings) are obtained. For further calculations the columns with the largest variances are taken [6].

In practice, the goal is to reduce the input matrix and to save as much information as possible (in practice, the LSA algorithm is utilized [4]).

The selection of a similarity method depends on the aim of analysis and the properties of the input dataset. Note, that inappropriate selection results in producing unreliable results.

An application of methods dedicated to assess the similarity between objects is based on the selection (definition) of the proper distance method or metric (function) among which the most popular is the Euclidean one. For the number of samples of each object p the Euclidean distance (and it's generalization – Minkowski metric) may be used.

In the field of text/semantic analysis, the Jaccard coefficient and the cosine distance are exploited. The Jaccard coefficient for non-binary data, i.a. two vectors of attributes, is computed as $A \cap B / A \cup B$ [5]. For the binary data, the Jaccard coefficient (see eq. 3) or the Dice one can be used.

$$S_J = \frac{n(1,1)}{n(1,1)+n(1,0)+n(0,1)} \tag{3}$$

Another approach is to use the cosine measure for text matching between two vectors (attributes) A and B (typically vectors: generated from documents (B) and a query one (A)), thus the angle between them (see eq. 4) [2].

$$S_C = \frac{A \cdot B}{\|A\| \cdot \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{(\sum_{i=1}^{n} A_i^2)} \times \sqrt{(\sum_{i=1}^{n} B_i^2)}} \tag{4}$$

The high level of the similarity between objects (in this case, to maximum possible value is equal to 1) is indicated for a small angle between vectors.

The similarity analysis of the text datasets is performed for a list of terms retrieved/acquired from original document(s), represented as points generated with an algorithm. To be able to process a number of related documents input datasets are weighted in relation to the maximum term frequency and/or the number of analyzed documents [2], [10].

IV. Sample Results

The goal of the part of research described in this section is to analyze a possibility of performing reliable similarity analysis of sample Polish legal texts based on approaches typically used for common language processing (in English).

For the prepared dataset of legal acts algebraic methods were applied to transform the original data into the set of terms suitable for further processing in order to similarity analysis. In this case, the Vector Space Model (VSM)[7] has

[4]The electronic form of legal acts – Legal Acts Editor – EDAP *http://bip.msw.gov.pl/portal/bip/185/18658/Edytor_Aktow_Prawnych_EDAP__wersja_z_dnia_31_marca_2010_r.html*

[5]For example, see *http://isap.sejm.gov.pl*

[6]The Java DOM Parser module was used

[7]Salton G., Wong A., Yang C. S. (1975): *A Vector Space Model for Automatic Indexing*, Communications of the ACM, vol.18, no.11, pp.613-620

been used to generate the incidence matrix of words in documents (term-document matrix, TDM).

Two sample similarity measures have been implemented: the Jaccard method and Cosine one.

The input data (body of documents) was extracted from PDF files officially published on the pages of the Polish government. As a body of documents, a set of eight Polish legal acts listed below have been chosen[8]: D1: Act on sell of consumers goods, 0.66MB; D2: Act against unfair market practices, 1.50MB; D3: Act on consumer credit, 3.40MB; D4: Act on competition and consumer protection, 4.80MB; D5: An act on Compulsory Insurance Insurance Guarantee Fund and Polish Motor Insurers, 9.60MB; D6: Act on Insurance Business, 14.1MB; D7: Act on tax consulting, 2.40MB; D8: Civil Code, 17.50MB. All the collected legal texts have undergone a process of analysis and extraction of words, which consisted of the following stages:

a) generation of morpho-syntactic text description from TXT format with Takipi[9];

b) exploration and extraction of words specific to the legal texts using models based on the original text in TXT format (references to the *"Dziennik Ustaw"*, referring to the date of enactment of legal documents, determining the names of offices and institutions etc.);

c) extracting tokens and phrases from the analysis (a) with the addition of specific terms of the legal texts and to create a final list of expressions (terms) used in the legal documents;

d) separating the expressions with stop-list for the Polish language and frequently occurring phrases specific to the legal texts (article, chapter etc.).

Results obtained for the similarity analysis of sample legal texts for 2 types of normalization (weighting TF/IDF and TDM normalization LSI) and 2 similarity measures are shown in tables I-IV.

An application of TD/IDF normalization and Jaccard method indicates that the most dissimilar documents to others is D1 and D3 (in Table II) while the most similar to others are difficult to indicate. The selection of these two methods produces different results factors (see tables I-II and tables III-IV). As shown below, the similarity matrix created off the LSI normalized vector shows more inter-document similarity than the one created off the TF/IDF normalized vector. Note, that most similar documents (D5 and D6) are accordingly obtained with LSI normalization and TD/IDF normalization for both similarity measure indexes: Jaccard and cosine method (see tab. I,II,III). Maximum values of similarity are indicated in the tables in bold.

TABLE I. RESULTS OBTAINED WITH TF/IDF NORMALIZATION AND JACCARD METHOD

	D1	D2	D3	D4	D5	D6	D7	D8
D1	---	0,0763	0,0805	0,0822	0,0798	0,0716	0,0606	0,0936
D2	0,0763	----	0,0860	0,1351	0,0841	0,0884	0,0715	0,0819
D3	0,0805	0,0860	---	0,0974	0,0946	0,0953	0,0634	0,0912
D4	0,0822	0,1351	0,0974	---	0,1336	**0,1641**	0,1442	0,1185
D5	0,0798	0,0841	0,0946	0,1336	---	**0,2106**	0,1124	0,1340
D6	0,0716	0,0884	0,0953	**0,1641**	**0,2106**	---	0,1504	0,1318
D7	0,0606	0,0715	0,0634	0,1442	0,1124	0,1504	---	0,0904
D8	0,0936	0,0819	0,0912	0,1185	0,1340	0,1318	0,0904	---
Suma:	0,5446	0,6233	0,6084	0,8751	0,8491	**0,9122**	0,6929	0,7414

TABLE II. RESULTS OBTAINED WITH TF/IDF NORMALIZATION AND COSINE METHOD

	D1	D2	D3	D4	D5	D6	D7	D8
D1	---	0,0142	0,0171	0,0264	0,0170	0,0131	0,0092	0,0857
D2	0,0142	---	0,0135	**0,1466**	0,0189	0,0177	0,0172	0,0225
D3	0,0171	0,0135	---	0,0255	0,0145	0,0166	0,0041	0,0193
D4	0,0264	**0,1466**	0,0255	---	0,0495	0,0623	0,0450	0,0409
D5	0,0170	0,0189	0,0145	0,0495	---	**0,1940**	0,0289	0,0707
D6	0,0131	0,0177	0,0166	0,0623	**0,1940**	---	0,0502	0,0413
D7	0,0092	0,0172	0,0041	0,0450	0,0289	0,0502	---	0,0189
D8	0,0857	0,0225	0,0193	0,0409	0,0707	0,0413	0,0189	---
Sum:	0,1827	0,2506	0,1106	**0,3962**	0,3935	0,3952	0,1735	0,2993

TABLE III. RESULTS OBTAINED WITH LSI NORMALIZATION AND JACCARD METHOD

	D1	D2	D3	D4	D5	D6	D7	D8
D1	---	0,8477	0,1734	0,8722	0,9510	0,9580	0,9480	0,8886
D2	0,8477	---	0,2317	0,9721	0,8916	0,8850	0,8211	0,9542
D3	0,1734	0,2317	---	0,2213	0,1906	0,1880	0,1693	0,2145
D4	0,8722	0,9721	0,2213	---	0,9172	0,9105	0,8445	**0,9816**
D5	0,9510	0,8916	0,1906	0,9172	---	**0,9927**	0,9197	0,9344
D6	0,9580	0,8850	0,1880	0,9105	**0,9927**	---	0,9264	0,9276
D7	0,9480	0,8211	0,1693	0,8445	0,9197	0,9264	---	0,8601
D8	0,8886	0,9542	0,2145	**0,9816**	0,9344	0,9276	0,8601	---
Sum:	5,6389	5,6034	1,3888	5,7194	**5,7972**	5,7882	5,4891	5,7610

The results also show cross-correlations of document similarity in the set {D4, D5, D6, D8} (the Insurance Law and the Civil Code) (see maximum values of similarity and "summary line"). The proposed algorithm has generated a group of legal acts from the similar semantic area (the Consumer Law; note, that the Civil Law has chapters related to the Insurance Law).

TABLE IV. RESULTS OBTAINED WITH LSI NORMALIZATION AND COSINE METHOD

	D1	D2	D3	D4	D5	D6	D7	D8
D1	---	0,8857	0,0213	0,8257	0,9100	0,8693	0,9944	0,9844
D2	0,8857	---	0,4831	0,9932	**0,9985**	0,9994	0,9297	0,9535
D3	0,0213	0,4831	---	0,5816	0,4340	0,5127	0,1266	0,1968
D4	0,8257	0,9932	0,5816	---	0,9853	0,9966	0,8805	0,9120
D5	0,9100	**0,9985**	0,4340	0,9853	---	0,9960	0,9486	0,9687
D6	0,8693	0,9994	0,5127	0,9966	0,9960	---	0,9166	0,9427
D7	0,9944	0,9297	0,1266	0,8805	0,9486	0,9166	---	**0,9975**
D8	0,9844	0,9535	0,1968	0,9120	0,9687	0,9427	**0,9975**	---
Sum:	5,4908	**6,2431**	2,3561	6,1749	6,2411	6,2333	5,7939	5,9556

[8] *http://isap.sejm.gov.pl/*

[9] Polish texts tagger, to its XML output, read in the designed program: Piasecki M. *Polish Tagger TaKIPI: Rule Based Construction and Optimisation*. Task Quarterly, 2007, 11, 151-167. (*http://nlp.pwr.wroc.pl/takipi/*)

Furthermore, an analysis aimed at examining the similarity of paragraphs from legal texts taken from the analyzed set of Polish legal acts (mainly from Civil Code) and changed in their structure (using synonyms), has been also performed. Sample results are presented below (see Table V).

TABLE V. RESULTS OF SIMILARITY ANALYSIS BETWEEN MODIFIED PARAGRAPHS AND LEGAL ACTS

Doc. no.	**TF/IDF**	Doc. nr	**LSI**
D8	**0,1072**	**D8**	**0,0806**
D5	0,0653	D4	0,0805
D4	0,0558	D1	0,0798
D2	0,0410	D2	0,0792
D7	0,0348	D5	0,0775
D1	0,0345	D6	0,0763
D3	0,0340	D7	0,0685
D6	0,0335	D3	0,0404

The results show that the modified paragraphs may be detectable with the described algorithm (see Table V; paragraphs are detected as similar to the document D8 – Civil Law) and they are promising for further work.

V. CONCLUSION AND FUTURE WORK

Depending on the selection of algorithms employed for extracting terms from legal documents, dimensionality reduction and for preprocessing of input data, it can be assumed that there is a possibility of application numerical procedures (widely used in the area of common English language) to datasets containing Polish legal texts. The essential step towards obtaining reliable results is similarity method and the input matrix normalization numerical procedure selection. However, as far as more complex semantic analysis are concerned, presented statistical methods, in some cases, give ambiguous results. In this case, dedicated dictionaries (thesauruses) or ontologies providing structured relationships, need to be built. They ensure structuring of taxonomic and non-taxonomic relationships in processed documents. (the research is in progress).

In the paper, the preliminary work aimed at automatic Polish legal texts processing was outlined. Taking into account presented considerations and sample results, further research will be aimed at Polish legal acts clustering, in particular, through adjusting of numerical algorithms (briefly described above), and – additionally – performing analyses of semantic relations between terms, phrases, sentences and finally – Polish legal acts.

REFERENCES

[1] Anderson E., Bai Z., Bischof C., Blackford S., Demmel J., Dongarra J., Croz J.Du, Greenbaum A., Hammarling S., McKenney A., Sorensen D., LAPACK User's Guide (*http://www.netlib.org/lapack/lug/lapack_lug.html*), Third Edition, SIAM, Philadelphia, 1999.

[2] Charikar M.S.: Similarity Estimation Techniques from Rounding Algorithms. STOC '02 Proceedings of the thiry-fourth annual ACM symposium on Theory of computing, ed. John Reif, ACM, 2002.

[3] Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A., Indexing by latent semantic analysis. Journal of the American Society of Information Science, vol. 41(6), pp. 391–407, 1990.

[4] Deerwester S. et al, Improving Information Retrieval with Latent Semantic Indexing, Proc. of the 51st Annual Meeting of the American Society for Information Science 25, 1988, pp. 36–40.

[5] Hand D., Mannila H., Smyth P., Principles of Data Mining. MIT Press, 2001.

[6] Jolliffe I. T., Principal Component Analysis. 2nd ed., New York: Springer-Verlag, 2002.

[7] Kisz A., Model cybernetyczny powstawania i działania prawa, Wrocław 1970, Sobczak K., Prawo a informatyka, Wydawnictwo Prawnicze, Warszawa 1978.

[8] Landauer, T., Dumais, S., A solution to Plato's problem: The latent semantic analysis theory of acquisition. Psychological Review, vol. 104(2), pp. 211–240, 1997.

[9] Malinowski A., Wstęp do badań cybernetycznych w prawoznawstwie, Warszawa 1977.

[10] Markines B., Cattuto C., Menczer F., Benz D., Hotho A., Stumme G.: Evaluating Similarity Measures for Emergent Semantics of Social Tagging, WWW '09 Proceedings of the 18th international conference on World Wide Web, ACM 2009.

[11] Palmirani M.,Cervone L., Vitali F., A Legal Document Ontology: The Missing Layer in Legal Document Modeling, in: Sartor G., Approaches to Legal Ontologies. Thwoeiwa, Domains, Methodologies. Springer, Dordrecht etc. 2011, p. 168.

[12] Potiopa P, Methods and tools for the automatic processing of textual information and its use in the process of knowledge management, Automatyka, T. 15, z. 2, s. 409-419, AGH Publ. 2011.

[13] Salton G., McGill M., Introduction to Modern Information Retrieval. McGraw-Hill, 1983.

[14] Studnicki F., Cybernetyka a prawo, Warszawa 1969

[15] Vitali F., A Standard-Based Approach for the Management of LEgisaltive Documents, [In:] Sartor G., Palmirani M., Francesconi E., Biasiotti M.A. (eds.) Legislative XML for Semantic Web, Springer Dordrecht etc. 2012, p. 35

[16] Wiewiórowski W.R., Wierczyński G., Informatyka prawnicza. Technologia informacyjna dla prawników i administracji publicznej, Oficyna – Wolters Kluwer business, Warszawa 2008