Conference Paper

A Document Object Model for Solving the Problem of Identification and Structurization of Documentary Flows of Rosfinmonitoring

Kapochkin S. V., Zaripova E. V., and Maksimov N. V.

National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Kashirskoe shosse 31, Moscow, 115409, Russia

Abstract

The paper considers the issues of building an information retrieval system by using the algorithm of automated classification and recognition of the structure of full-text documents. It describes the selected approaches, as well as the algorithm for identifying the document type and the algorithm for recognizing its logical structure, developed on the basis of these approaches, with the aim of further semantic processing. It introduces a multi-stage method for automated recognition and formation of a model of the logical structure of a document. Experimental studies of this method have been conducted on the array of reporting documents "Rosfinmonitoring".

Corresponding Author:
Kapochkin S. V.
several23@mail.ru

Received: 11 December 2017
Accepted: 20 January 2018
Published: 13 February 2018

Publishing services provided by Knowledge E

Selection and Peer-review under the responsibility of the FinTech and RegTech: Possibilities, Threats and Risks of Financial Technologies Conference Committee.

1. Introduction

Identifying the fundamental properties and patterns of document design

Documentary turnover plays an important role in the area of financial investigations. The growing volume of such documents leads to the issue of their competent storage. Since the way the documents are stored depends on the information extracted from it, the problem of creating a database with the optimal classification and structuring of documents is relevant.

In the field of AML / CFT, there are types of documents such as a sentence, report, public statement, decree, order, FZ, etc. Using GOSTs of polygraphy and internal standards of Rosfinmonitoring, it is possible to define the rules for compiling all documents that will determine the final set of the structural elements of the document, their interrelation in the text, and their identifying features. Relying on the selected sets of structural elements and their interrelations, it is possible to divide the documents into classes in the database. The division of documents in the database into classes

allows going to the required class of documents and reducing the amount of data for analysis. As the amount of data for processing decreases, the time that the user spends on analysis decreases, and the amount of knowledge that the user can extract from the group of documents, selected according to the required group, increases.

The definition of a document class is based on the comparison of a set of its structural elements with a characteristic set of structural elements for each class. The characteristic set of the structural elements of the class is formed on the basis of GOST and GNI of Rosfinmonitoring for different types of documents. Thus, for each selected group of documents, there is a finite set of structural elements, which allows assigning the document uniquely to this group. A structural element can be a paragraph, a document number, a table of contents, and a title for the document. In other words, an independent data unit providing the user with complete information.

The layout of the document structure allows analyzing a separate part of the document instead of the entire document. Thus, for example, when the user needs to analyze only an annotation to a document, they can find the structural elements of "annotation" of all the documents in the group of documents, which are of interest, and reduce the time of analysis. The initialization of each structural element is based on the final set of its characteristics, compiled on the basis of the international rules of printing and the internal standards of the organization which publishes the document.

Developing the document object model and introducing the concept of a logical element

The given paper analyses Rosfinmonitoring's internal reporting documents. As a result, a document object model has been drawn up; this model defines the structural elements, their characteristics, internal links with structural elements, and external relations with the class of the document. The object model of the document is based on the approach of the general theory of systems. Each document is presented as a system (1):

$$D_i = \{M_i, A_i, R_i, Z_i\}, \tag{1}$$

where

$$\begin{cases} M_i = \{M_{i1}, M_{i2}, \ldots, M_{is}\} - \text{a set of logical elements included in the document;} \\ A_i = \{A_{i1}, A_{i2}, \ldots, A_{il}\} - \text{a set of features of logical elements;} \\ R_i = \{R_{i1}, R_{i2}, \ldots, R_{ie}\} - \text{a set of interrelations between logical elements;} \\ Z_i = \{Z_{i1}, Z_{i2}, \ldots, Z_{iq}\} - \text{a set of compositions (possible connection M\_i, A\_i,} R\_i); \end{cases} \tag{2}$$

The classes (groups) of documents are represented by the following set (2):

$$T = \{T_i\}, \tag{3}$$

Ti is an *i*-th class of the document; $P_i = \{M_i, R_i\}$ is a set of features of the i-th class of the document, where: M_i is a subset of the logical elements specific to the document i-th class, i.e., those whose presence indicates the class of this document; R_i is a subset of the links of the characteristic elements for the *i*-th class of the document.

$M_i = \{M_{i1}, M_{i2}, \ldots, M_{is}\}$ – a set of the *logical (structural) elements included in the document*;

A logical element is a quantity representing a logical (semantically significant) unit of information in a machine form. It must be represented by an integral physical object with an identifiable name. Elements can have different properties, which in general can be stored together with or separately from data. An example of a logical element is the Document Number or the Name of the Document, and so on.

Because the content of an element can store not only the description of the subject area of the document, but also the information identifying the document or metadata, classes of structural elements have been introduced. There are three classes: identification data, meaningful data and metadata. Each structural element belongs to only one class.

$A_i = \{A_{i1}, A_{i2}, \ldots, A_{is}\}$ is a set of features of logical elements;

Features of a logical element should be understood as a sufficient set of conditions or properties for the recognition of this element. Such features can include: the register, the font, the marking, the figures at the beginning of the heading.

A document belonging to one of the groups has structural elements (1..n) that take specific values of the characteristics (1..k) for this document. This model also implies that one structural element with different values of a set of characteristics can be contained in documents from different groups. The features have a finite number and are divided by functionality. The set of identification attributes specifies the "external appearance" of the structural element in the text of the document, the attributes of the element link determine the location of the element in the structural hierarchy of the document relative to other elements, and the semantic content of the element is specified by the content characteristics. And only the given set of values of all features allows to identify the structural element uniquely in the text.

The above-mentioned features of an element link are identified in the system Di in a separate set:

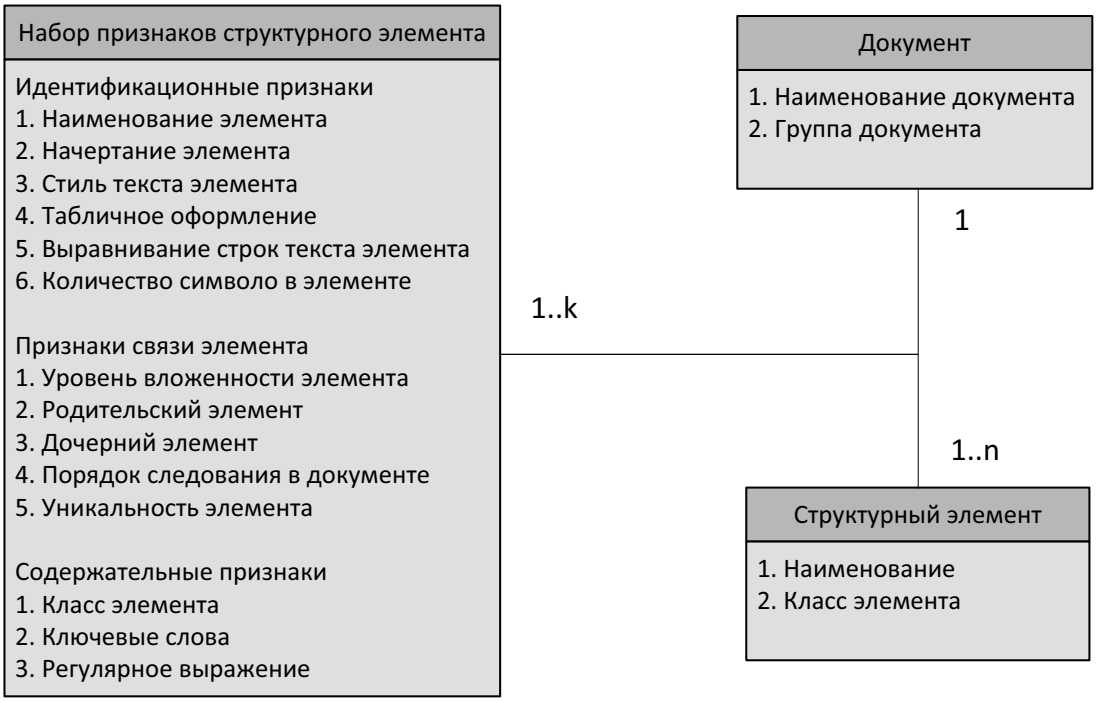

Figure 1: ER-model of the conceptual scheme of the relationship between the document, the structural element, and their features.

$R_i = \{R_{i1}, R_{i2}, \ldots, R_{is}\}$ is a set of logical element links (all possible element locations in the document); A document, like the Thesaurus, is a repository. But in the second case it is a repository of terms, and in the first one it is a repository of logical elements. In both of these vaults, we assume the interrelation of the stored objects. There are two types of links: hierarchical and non-hierarchical. Hierarchical links are understood as the location of objects relative to each other (ABOVE, BELOW, BEFORE, AFTER), as well as their nesting into each other (PART, WHOLE). Non-hierarchical relations are understood as the interrelation of objects on one level (ASSOCIATION) [2].

$Z_i = \{Z_{i1}, Z_{i2}, \ldots, Z_{is}\}$ is a set of compositions (i.e., what sets, features and locations of logical elements are possible);

Determining the relationship between a logical element and a physical object

At the same time, a logical element is built from physical document objects that have their own format (they can be a constant, a variable, or a more complex construction). So the logical element is also an object that we represent in the form of a system

$M_{ij} = \{F_{ij}, A_{ij}, R_{ij}, Z_{ij}\}$, where

F_{ij} is a set of physical objects;

A physical object that provides a representation of a logical element is, in particular for a text case, the minimum possible set of characters constituting a complete message, but having semantic significance only as part of a logical element. So, for example, for the logical element "Document number" presented in the following form: {Report #: XXXX}, where Report #,:, XXXXX-physical objects. "Report #" and ":" are physical objects of the constant type, "XXXXX" is a physical object of the type variable.

Thus, using the Extended form of the Bekus-Naur, you can define a logical element through a set of physical objects.

Logical element :: = Physical object, constant | Physical object, variable | {{Physical object, constant} {Physical object, variable}}.

Which means that a logical element cannot be empty and can be a concatenation of any number of combinations of a constant and a variable, or just a constant, or just a variable.

A_{ij} is a set of features that allow recognizing the form of the physical representation of elements;

A physical object can have only four parameters defining its representation: the size, style, color, and font. The size parameter is responsible for the width, height and length of the physical object. The style parameter reflects the outline of the physical object (Italic, Bold, Underline, etc.). The color parameter stores information about the color of the physical object (RGB). The font parameter reflects which font has been used.

R_{ij} is a set of links of physical objects (the location of the physical object in the logical element);

Z_{ij} is a set of compositions (i.e., which sets, attributes and locations of physical objects are possible).

Applying the object model in the task of document identification and structuring

Thus, the structure of the document is represented by a hierarchy, the upper level of which is the class of the structural element, the second level is the abstract class of the structural element, the third level is the structural element, the subsequent levels are filled with nested elements and objects. The hierarchical structure reflects the sequence of elements according to the narrative sequence in the document. This is how the attributes of the element are formed (see Figure 1): the nesting level, the parent element, the child element, the sequence (the association is displayed), the uniqueness of the element.

The analysis of logical elements in classes allowed distinguishing polymorphisms, i.e., groups of logical elements that are similar in external features, but different in their semantic purpose. Based on the results of the selected polymorphisms, abstract classes of logical elements were defined. Each logical element is a derived special case from a parent abstract class. The relation between the document groups, abstract classes of logical elements and attributes of logical elements is summarized in the table "Document hierarchy".

Structuring and classifying a document is an automated process that runs shortly before the document is loaded into the database. The result of this process is a file with a marked document structure and an identified class. The layout of the document text is carried out in accordance with the developed model of the document. The process of classification and structuring of the document includes a multi-stage algorithm, which is developed on the basis of a neural-network approach.

The object model of the document was presented in the form of an xsd-scheme that displays all the previously allocated document types and their structural elements.

2. Conclusion

In this study, the object model of the document has been built, in order to automate the process of classification and structuring the document, which will increase the usefulness of searching the database of AML / CFT documents in the database.

The object model reflects the relationship between the document class and the logical element through a ternary connection through specific values of the characteristics of the logical element. The model defines the set of logical elements, classes of logical elements, their attributes and connections. The object model helps define the abstract classes of logical elements that allow to open polymorphisms in an array of the structural elements of a document.

For the sphere studied, a generalized table was constructed on the basis of the proposed model, which makes it possible to isolate the structural elements from the linear text.

In the study, an experimental approbation was carried out on an array of documents on financial investigations and technical documentation. Solutions are constructed that have shown experimentally the usefulness of the approach of structural analysis of a document and the construction of a hierarchical model of a document.

The review of the algorithm and the experimental setup are presented in an article published at the International Practical Conference of the Network Institute in the field of AML / CFT in 2016 in the section "Threats and Risks of the World Economy" [1].

Acknowledgements

This work was supported by Competitiveness Growth Program of the Federal Autonomous Educational Institution of Higher Education National Research Nuclear University MEPhI (Moscow Engineering Physics Institute).

References

[1] Sysoykina MA Modeling and development of tools and technologies for presenting information in distributed electronic libraries: author's abstract. dis. for the degree of Candidate of Technical Sciences: 05.25.05 / M.A. Sysoykina; Moscow, Russian State University for the Humanities - M., 2003. - 28 s.

[2] Debashish Niyogi D. and Srihari S. The use of document structure analysis to retrieve information from documents in digital libraries [Электронный ресурс].// URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.90.9296&rep=rep1&type=pdf (дата обращения: 05.02.2016)

[3] Kharlamov A. Automatic structural analysis of texts // Open systems. - 2002. - №10. - p.7-8

[4] Mao S., Rosenfeld A., Kanungo T. Document structure analysis algorithms: A literature survey [Электронный ресурс]. // ResearchGate [Сайт]. URL: http://www.researchgate.net/publication/221253919_Document_structure_analysis_algorithms_a_literature_survey (дата обращения: 07.03.2016)

[5] Hirokazu I., Shimazu A., Ochimru K. Document Structure Analysis with Syntactic Model and Parsers: Application to Legal Judgments [Электронный ресурс]. // Springer Link [Сайт]. URL: http://link.springer.com/chapter/10.1007%2F978-3-642-32090- 3_12 (дата обращения: 15.02.2016)

[6] Dengel A. Initial Learning of Document Structure [Электронный ресурс]. // Old Dominion University [Офиц.сайт]. URL: http://www.cs.odu.edu/~pflynn/survey/doc-struct- 00395776.pdf (дата обращения: 15.02.2016)

[7] Klampfl S., Granitzer M., Jack K., Kern R. Unsupervised document structure analysis of digital scientific articles [Электронный ресурс]. // Springer Link [Сайт]. URL:

http://link.springer.com/article/10.1007%2Fs00799-014- 0115- 1#page-1 (дата обращения: 15.02.2016)

[8] U. Yu. A., Beginning of the general theory of systems. // System analysis and scientific knowledge, Moscow, 1978.

[9] MN PI I. Golitsyna OL, Information Systems, Moscow: FORUM, 2009.