

Запропоновано теоретико-множинну модель, що описує склад і структуру системи підтримки прийняття рішень, яка функціонує на основі багатоаспектного фактографічного пошуку. Розроблено інформаційну технологію багатоаспектного фактографічного пошуку. Технологія дозволяє генерувати групу запитів відповідно до аспектів розв'язуваної задачі. В процесі пошуку виділяються підмножини документів, що релевантні до аспектів. У кожному документі виділяються прецеденти, відповідні аспектам. Усувається надмірність результатів пошуку

Ключові слова: система підтримки прийняття рішень, фактографічний пошук, множина прецедентів щодо аспектів

Предложена теоретико-множественная модель, описывающая состав и структуру системы поддержки принятия решений, функционирующую на основе многоаспектного фактографического поиска. Разработана информационная технология многоаспектного фактографического поиска. Технология позволяет генерировать группу запросов в соответствии с аспектами решаемой задачи. В процессе поиска выделяются подмножества релевантных аспектам документов. В каждом документе выделяются прецеденты, соответствующие аспектам. Устраняется избыточность результатов поиска

Ключевые слова: система поддержки принятия решений, фактографический поиск, множество прецедентов по аспектам

UDC 004.912:519.7

DOI: 10.15587/1729-4061.2017.108569

DESIGNING A MODEL OF A DECISION SUPPORT SYSTEM BASED ON A MULTI-ASPECT FACTOGRAPHIC SEARCH

I. Shevchenko

Doctor of Technical Science, Professor*

E-mail: ius.shevchenko@gmail.com

V. Tertyshnyi*

E-mail: vladislafus@gmail.com

S. Koval

PhD, Senior Lecturer*

E-mail: kovalsvitlanakremenchuk@gmail.com

*Department of Information and

Control Systems

Kremenchuk Mykhailo Ostrohradskyi

National University

Pershotravneva str., 20,

Kremenchuk, Ukraine, 39600

1. Introduction

Among a wide variety of decision support systems (DSS), there is a class of the so-called passive systems that prepare data for decision-making for a user, but do not generate solutions themselves [1]. Among passive DSS, it is possible to identify the systems based on the information search. They are the so-called Document-Driven DSS. These systems carry out the search and manipulate unstructured information, assigned in different formats. For management decision support, for example, in marketing, passive DSS should conduct a factographic search, particularly a search over unstructured data that are relevant to problem that is solved.

Factographic description is a totality of data (descriptions of precedents, features, and characteristics), related to a certain object (process, phenomenon). Existing system of factographic search process users' queries, perform content analysis of the global network, generate data in accordance with the query. However, under conditions of solving a multi-aspect problem, information resources are not used sufficiently because of the imperfection of methods and tools for searching for information in non-homogeneous information environment. That is why development of technology and decision support system based on multi-aspect information search is a relevant task.

2. Literature review and problem statement

In contrast to the classic search, the outcome of a factographic search is not a document, but a brief and concise fragment of the text – a response to a question posed by a user in a natural language.

To meet the challenge of a factographic search, the methods, based on deep syntactic and semantic analysis of a text, are traditionally used. Research is conducted in three main directions [2]:

– syntactic methods – by studying syntactic characteristics and lexical composition of texts;

– statistical methods – establishing semantic relations by identification of distributive-statistical relationships;

– semantic methods – through exploring in-depth semantic links in a sentence, a text document and beyond a text document, for example, in a knowledge base, thesaurus, ontology, glossary-combination dictionary.

There is a known solution to the problem of automatic retrieval of key phrases by using GROBID (Generation Of Bibliographical Data) [3]. A feature set determines characteristics of the content, based on measures of informativeness, phraseness and keywordness. To create a set of lexical and semantic features, two knowledge bases are used: GRISP and Wikipedia. Technology allows users to retrieve data from structured texts, but its

effectiveness, when dealing with unstructured texts, is not evaluated.

The technologies are proposed for retrieval of concepts, formed by relevant single and multi-word units by statistical processing of the total array of the found documents with calculation of frequency distances between words [4]. Frequency method of generation of patterns for a search in the data stream is proposed in [5]. The proposed algorithm enumerates all possible frequent patterns and can be used for searching for the most popular elements. In the described technologies, there is no mechanism of aspect separation and there is a high probability of missing relevant data.

Paper [6] proposed the system of natural language processing using the Bayesian statistics and logics of rules for detection and characteristic of events of medical nature. The system uses a semantic model that allows the retrieval of facts that prove the risk of disease initiation. The problem is the need to support current knowledge base system when the system is scaled.

Solution to the problem of “side-information” is proposed in article [7]. The metadata, associated with a text document, are used. It is proposed to use clustering methods and probabilistic models. The results in the form of frequencies of found terms occurrence are cited. The existence of “side information” indicates existence of unrevealed aspects of a problem.

In paper [8], analysis of sequences in a text is carried out at the chapter level, paragraph level and sentence level. The impact of stop-words on quality and quantity of retrieved patterns is explored. However, thematic analysis and splitting into domains are not performed, which objectively worsens the system’s efficiency.

Paper [9] considers the problem of retrieving facts and their causes from a Web- text. Connecting markers, like ‘because’ and local trees of syntactic analysis of the described events are used. At the last stage, the statistical model for measuring potential causal relationships is used. Construction of a totality of local trees of syntactical analysis greatly complicates the technology.

The problem of processing large data arrays is considered in [10]. The authors try to parallelize the FP-Growth algorithm to solve this problem. The generated FP tree and its rules for analysis of retrieved trends are used. The parallelism of this algorithm does not involve thematic decomposition.

An analysis of these studies reveals that the problem of retrieval of useful information from texts can be solved in different ways, but, in most cases, it is impossible to do without using statistical and frequency methods.

Another side of the problem in the retrieval of factographic information from actual arrays of documents, located on the Internet, is connected with attempts of using the Semantic Web concept. Practices of the W3 consortium, concerning the use of special dictionaries and standards, are of recommendation character only. The actual development of most Internet resources happens without regard to such recommendations. This applies even more to the electronic documents, not located on the Internet.

Thus, the problem of retrieval of factographic information from the documents of arbitrary structure remains incompletely solved. The main problem is the difficulty of separating information, not related to a sought-for topic, and the selection of significant data with maximal completeness. An essential part of the required data can

relate to different aspects of the problem. This means that it is possible to perform a search by aspects, by the original problem undergoing decomposition. In this case, it is possible to simplify the methods for solving the problem of fact retrieval without losing important data. To implement the technology of factographic search, it is necessary to devise appropriate methods and models. The present work is a continuation of the research, the outcomes of which are published in papers [11, 12].

3. Research goal and objectives

The goal of present research is to enhance efficiency of factographic search at simultaneous simplification of the data processing technique through the use of a multi-aspect logical-linguistic model and a method of factographic search with the possibility of analysis of unstructured textual information and retrieval of data, required for decision-making support.

To accomplish the goal, the following tasks were set:

- creation of a formal basis, describing the composition and structure of DSS;
- development of information technology of a multi-aspect factographic search;
- test of efficiency of the information technology on the example of a particular task.

4. Devising a model of a decision support system based on information search

To develop a formal basis, describing the composition and structure of DSS, we will use a theoretical-multiple representation, which should reflect both the general task of preparation for decision-making and a complex of functional tasks of decision support, based on a multi-aspect factographic search.

The complex of functional problems includes:

- task of processing a user’s query and separation of key words and lexemes with subsequent separation of aspects and formation of secondary search queries;
- task of searching for relevant documents;
- task of a factographic analysis of the selected documents;
- task of forming a data set for the user.

In the process of implementation of the present set of tasks, DSS is the recipient of the information flow from the global Internet and a distributed database of an enterprise. The input information flow contains an array of documents, relevant to the DSS queries.

The output information flow represents a complex of factographic data for the user who resolves a task within a specific subject area.

The principles of DSS functioning must provide:

- conjugation of DSS with the database of the information-management system (IMS) of an enterprise;
- use of the subject area ontology, including specifics of tasks that are solved, in order to create the multi-aspect thesaurus;
- DSS versatility within the framework of the considered business processes.

Preparation of decision making takes place within a framework of the subject-matter ontology [12]:

$$O = \langle E(AT), ER, F, AS, AR \rangle, \tag{1}$$

where E is the set of entities, AT is the set of attributes of entities; ER is set of relationships of entities, $F: E \times ER$ are the functions of interpretation of relationships and entities; AS is the set of aspects that define a subset of entities and relationships, $AR: AS \times AS$ are the intersection of aspects. A feature of the proposed ontology model is that it takes into account the fact that there are several associations between entities by several aspects

The task on preparing making a decision is stated in the following way: we assign a subset of original states of an object, a subset of final states and a set of unspecified beforehand information resources, required for transformation of states. It is required to find and specify a set of information resources that will allow a decision maker to transfer the managed object from the current state to the desired or permissible state. Then the task of preparation for decision-making can be defined by expression:

$$Z = \langle DM, S, S_0, S_k, IR(AS), QR \rangle, \tag{2}$$

where DM is the model of a problem situation in a subject area; S is the set of current states (situations); $S_0 \in S$ is the subset of original states; $S_k \in S$ is the subset of final, or target states; IR is the finite set of information resources. Every element of this set carries information on several AS aspects of the solved tasks; QR is the set of quality criteria of found resources.

Model of DSS structure. Formally, the structure of the decision support system SAD will be represented as a set:

$$SAD = \langle MFR(O), SED, FSE, KB, DB, AB, SM, ISP, DS \rangle, \tag{3}$$

where MFR is the model of search for information resources, based on ontology of information resources O ; SED is the documents search engine; FSE is the factographic search engine; KB is the knowledge base; DB is the data base; AB is the base of algorithms, implementing query processing, document search and factographic search; SM is the unit of synchronization of information processes; ISP is the unit of integration with information system of a enterprise; DS is the subsystem of dialogue with DM .

A problem situation in the context of business processes can be described by selection of a certain separated feature set or through a certain structure that allows us to represent different relationships (relations) between the elements of a problem area. Formation of the model of problem situation is based on the analysis of components of the subject area ontology.

A model of search for the information resources MFR , intended for solving the task of a multi-aspect data search, will be formally determined by a set:

$$MFR = \langle O, Q, D, F, R(d, q) \rangle, \tag{4}$$

where O is the ontology of a subject area; Q is the set of representation of information need (query); D is the set of representations of a document; F is the means of modeling of representations of a document, queries and their relationships, based on the model of a specialized document search based on related data [11]; $R(d, q)$ is the ranging function, which matches real numbers against d from D and q from Q and determines the order on the set of documents relative to requests q .

SED is the document search engine, based on the method of corporative entity search on the basis of related data.

FSE is the factographic search engine, based on a logical-linguistic model and the method of a multi-aspect factographic search [12]. A logical-linguistic model of the text for a multi-aspect search takes the following form:

$$T = \langle CT, AS, DB, A, RA \rangle, \tag{5}$$

where CT is the marker of text belonging to a certain class, the marker is assigned to a text at the preliminary stage of search for relevant texts; AS is the set of aspects in a text; DB is the database that represents a set of thesauri that store text keywords, lexemes, and sentences; A is the set of paragraphs of a text; RA is the intersection of a set of paragraphs with a set of aspects.

The model of knowledge base KB , in addition to relations of a problem area, contains metaknowledge, necessary for fast switching (selection) of a required fragment from the basic knowledge base KB and generation of secondary queries for a document search:

$$KB = \langle ND, SR \rangle, \tag{6}$$

where ND are the matrices of fuzzy productions, allowing us to assess the degree of proximity of the document's content by the relative frequency of the occurrence of keywords and lexemes by the aspects of the solved problem; SR is the set of rules for the generation of comments and recommendations for the user.

Database DB contains data of the following categories:

- form of a user's original query and forms of secondary search queries;
- thesauruses of the subject area, divided by tasks and aspects;
- matrices of frequencies of the occurrence of terms and lexemes in reference documents;
- matrices of frequencies of the occurrence of terms and lexemes in the analyzed documents;
- arrays, containing lists of selected documents;
- forms of reports on factographic search.

The unit of synchronization of information processes SM is the main software module of DSS that performs the dispatching of analysis process, search for information resources, message generation, launch and termination of all the functional modules of DSS.

The algorithms base is represented by the set:

$$AB = \langle AC, AQG, ADS, AFA, ARG \rangle, \tag{7}$$

where AC are the algorithms of document space clustering; AQG are the algorithms of query generation and processing; ADS are the algorithms of document search; AFA are the algorithms of factographic analysis; ARG are the algorithms of report generation. These algorithms are implemented in DSS software in the form of functional modules, operated by a synchronization module.

The unit of integration with information system ISP is described by the model:

$$ISP = \langle XM, XS \rangle, \tag{8}$$

where XM is the set of inputs that represent user's queries; XS is the set of subsystem inputs that represent the data, obtained from the existing IMS.

Set XM is determined by the totality:

$$XM = \langle T, T_s \rangle, \quad (9)$$

where T is the query test; T_s are the texts of user's replies in a dialogue mode.

Set XS is determined by vector:

$$XS = \{E_{SA}, A_E\}, \quad (10)$$

where E_{SA} are the entities of a subject area; A_E are the attributes of entities.

It is possible to represent the subsystem of dialogue with $DM - DS$ with a set of:

$$DS = \langle L(RD, SR(D_s, D_c)), P_M, MS(TM, SM, R, PM, A) \rangle, \quad (11)$$

where L is the language of dialogue; RD is the message generation rules; SR is the set of standard messages; D_s is the set of standard messages by queries; D_c is the set of standard messages of the process of a dialogue by search results; P_A is the procedure of producing results. MS is the message, where TM is the time of a message; SM is the key subject of a message; R is the reference to a previous message PM ; A is the user's response.

The model, represented by expressions (1)–(11), reflects, on the conceptual level, the composition and structure of a decision support system based on a multi-aspect factographic search, suitable for the development of DSS in any subject area. For the development of information technology of a multi-aspect factographic search, the models and the method, described in [12], are used.

5. Development of information technology of a multi-aspect factographic search

Any information technology (IT) is a totality of information processes (IP). Let us consider IP, included in this technology. For convenience of presentation, IP are numbered, and the natural sequence of the general course of data processing is observed.

At the preparatory stage, ontology is generated (1). Tables of relationships between aspects are created. Each entry of a table corresponds to a term from the thesaurus, and each field corresponds to an aspect. The table elements are binary, indicating the presence or absence of a given term in this aspect.

Reference matrices of frequency relations of an aspect are generated based on the texts, which are 100 % relevant to the given aspect of the texts. These texts are generated or selected by experts.

IP1. Query processing. This process consists of the following stages:

1. 1. Keywords and lexemes are selected from a query with the use of thesaurus of a subject area.

1. 2. Key aspects of a query are determined by analyzing the table of relations between aspects. This table is a matrix, in which lines correspond to terms, and columns – to aspects. Matrix cells contain relative frequencies of term occurrence in aspects.

1. 3. Clarification of aspects by a user and their ranking by importance through assigning weight coefficients:

$$\alpha_j = \frac{1}{n} \sum_{i=1}^n a_{ij}, \quad (12)$$

where a_{ij} is the marker of participation of the i -th term in the j -th aspect, n is the number of terms in a thesaurus. These coefficients are used for the generation of factographic search results.

1. 4. Grouping of keywords and lexemes by aspects. Generation of queries by aspects (extended query).

IP2. Search for documents and their grouping by aspects.

2. 1. From the original set, by the vector method, documents, for which the degree of relevance to the solved problem exceeds the specified threshold, are selected. This is a classic procedure, which is used in almost all search engines. All terms of all secondary requests are used for a search. Vectors X for existing terms for each document are generated.

Subsequently, the vector of keywords of query q_k and the vector of keywords of document x_k are compared for primary selection of documents to operating subsets of aspects A_j

$$r_j = \sum_{k=1}^{L_q} x_k q_k, \quad (13)$$

where L_q is the capacity of a set of terms in a query, taking into account query expansion by aspect; k is the number of a term.

2. 2. The vectors of terms of every aspect query q_k and the original vector of existing terms of document x_k are compared for primary selection of documents to operating subsets of aspects A_j . Normalized frequency of entrance of terms from the j -th aspect to a document is calculated for each document from the operating subset.

$$r_j = \frac{\sum_{k=1}^t x_k^j q_k}{\sum_{k=1}^t (x_k^j)^2 + \sum_{k=1}^t (q_k)^2 - \sum_{k=1}^t x_k^j q_k}. \quad (14)$$

The result is the vector of relative frequencies, showing the degree of correspondence of a document to each j -th aspect.

2. 3. According to established threshold of normalized frequency, documents are selected to operating subsets by aspects. Depending on the priority of an aspect, threshold value can be adjusted.

2. 4. Working matrices, which will serve as “references”, are generated for each document. To do this, the lines and columns of the terms that were not found in a document are removed from the original reference matrix of relations of an aspect.

IP3. Detection of paragraphs and sentences (precedents), relevant to each aspect.

3. 1. Paragraphs in a document and sentences in paragraphs are indexed.

3. 2. For each j -indexed aspect, relative frequencies fp_{ijk} of occurrence of the i -th aspect terms T_{ij} in the k -th paragraphs of the document are calculated:

$$fp_{ijk} = \frac{n_{ijk}}{n_{jk}}. \quad (15)$$

In table “paragraph-sentence” (line – paragraph number, column – phrase number), the numbers of sentences, containing terms of an aspect, are recorded.

The result:

– table “paragraph – aspect”, containing frequencies $f_{p_{ijk}}$;

– table “paragraph – sentence”, containing markers of sentences.

3. 3. Grouping paragraphs that passed the threshold test by aspects. Using threshold transformation, where T is the assigned value threshold of relative frequency f ,

$$H = \begin{cases} 1 & \text{if } f \geq T, \\ 0 & \text{otherwise,} \end{cases} \quad (16)$$

we determine belonging of each paragraph to each aspect. The result is a binary matrix “paragraph – aspect”.

A paragraph or a precedent may relate to several aspects, that is why the factographic search result at this stage can appear redundant. To reduce information redundancy, it is necessary to consider importance factors of aspects, obtained in IP1 when generating an advanced query. If a paragraph is considered in several aspects, its markers are removed for all aspects, besides the one, which has the highest weight factor in this group.

3. 4. Selection of sentences (precedents), containing the terms of the j -th aspect. Aspect-relevant passages are analyzed by searching and marking the sentences, containing the terms, recorded in p. 3. 3. The result is the table, in which precedents are related to the considered aspects.

IP 4. Grouping of the found precedents by certain aspects and production of a subset of precedents on each aspect for a user.

6. Example of practical application of the multi-aspect search technology

As an example of practical use, we will show (in fragments), the solution of the task “Search for new buyers (distribution channels) for enterprises of agricultural engineering”. Table 1 shows keywords and lexemes for aspects of this task.

According to the proposed technology, after entering a query and retrieving keywords and lexemes from it, the system attempts to determine the range of aspects of the problem in dialogue with a user. After specifying the composition of aspects, there takes place generation of secondary queries, based on the revealed aspects, a search in distributed heterogeneous database of an enterprise and in the global network, and selection of relevant documents.

For each selected document, working reference matrices of frequency relations are generated by deleting not found terms from the reference matrices of aspects. Subsequently, working matrices of documents, in which calculated frequencies of co-occurrence of terms and lexemes in documents, paragraphs and sentences, are generated. A fragment of the matrix of frequency relations of the document and of aspect “Functional” is shown in Table 2. The text fragments, selected from the text, are grouped according to the calculated value of the Kemeny distance between the working “reference” matrix and the matrix of the given document [12].

Table 1

Keywords and lexemes for aspects of task “Search for distribution channels”

Search for new buyers (distribution channels)			
Aspects			
Buying/selling	Production	Buyer/supplier	Functional and options
Assortment	Assortment	Address	High-grade flour
Dealer	Reaper	Assortment	Construction works
Distributor	Grain reaper	Currency	Selection of shorts
Order	Mounted reaper	Exhibition	Cereal
Import	Windrower	State order	Flour
Client	Trailer wind-rower	Dealer	Handling
Volume	Mounted grain reaper	Distributor	Flour sieving
Partnership	Trailer wind-rower	E-mail	Production
Buying	Mill	Forum	Grain milling
Buyer	Aggregate mill	Volume	Corn sticks
Customer	Compact mill	Telephones	Precise seeding
Sale	Mounted equipment	Coordinates	Wholesale
Payment	Seeding machine	How to get in touch	Retail sale
Distribution	Mill	Payment	Store house of raw materials
Discounts	Heating boilers	Batch size	Milling quality
Tender	Cereal line	Batch of produce	Payback period
Price	Extruder	Service	Adjustment

Table 2

Matrix of frequency relations of document and aspect

Docu-ment/aspect	Equip-ment	Grain crops	Flour	Han-dling	Pro-duc-tion	Grain crops milling	Whole-sale
Equipment	–	0.34	0.76	0.96	0.85	0.73	0.32
Grain crops		–	0.54	0.43	0.32	0.62	0.13
Flour			–	0.21	0.57	0.78	0.11
Handling				–	0.23	0.15	0.12
Production					–	0.16	0.13
Grain crops milling						–	0.11
Wholesale							–

Below, a part of factographic search results by aspects is shown.

<Aspect Production>

<Obtaining the batch of grain with assigned characteristics (milling batch), is achieved by mixing of two-three original batches (components) in required proportions> <link>.

<To compare different types of mills, it is possible to use the formula, which contain basic criteria for evaluation of

milling results (flour output and its quality indicator – ash content): $Kac=Ix(Z0-Z1)/Z0$ <link, address>.

<Milling equipment, produced by our enterprise, enables you to reduce costs of production of high quality flour, thus allows a flour manufacturer to enter the market, keeping high profitability and to dictate his conditions in the flour market> <link, address>.

<Aspects Buyers/suppliers>

<Enterprise KTL-LOGISTIC is a reliable supplier of agricultural produce and products of its handling in the global market. Head office of the enterprise is located > <link, address>.

<Aspect Buying-selling>

<Buy> <Flour mill with function of coarse and fine milling> <link, address>.

<Buy> <Plant for production of flour milling equipment> <link, address>.

<Buy> <Equipment for production of food corn sticks second hand extruder, cornfitting drums> <link, address>.

7. Assessment of efficiency of subsystem of precedents retrieval

The fundamental feature of the systems of factographic search is almost a 100 % completeness of results. In addition, general quality of two-stage (documents and precedents) search in each session varies greatly depending on a specific problem to be solved. That is why, when assessing efficiency of a subsystem of precedents retrieval, we created a special test collection of documents, containing the necessary volume of information on aspects of the problem to be solved. Experts accepted the degree of highlight completeness of every aspect in a test sample of documents as 1.0. Experts also generated original queries to the system. Results of performance assessment of the precedents retrieval subsystem for different ratios of importance weight factors of aspects are shown in Table 3. Redundancy of precedents representation was calculated as quotient of division of total number of precedents, represented in each aspect, by total number of detected precedents.

Table 3

Results of assessment of precedents retrieval completeness

Completeness/redundancy of precedents retrieval in collection of documents by 4 aspects							
1	2	3	4	1	2	3	4
Importance factors of aspects				Importance factors of aspects			
0.25	0.25	0.25	0.25	0.15	0.25	0.4	0.2
Completeness				Completeness			
0.95	0.94	0.92	0.91	0.94	0.92	0.97	0.90
Average assessment of completeness – 0.93				Average assessment of completeness – 0.95			
Redundancy – 1.36				Redundancy – 1.04			

Because importance factors mainly affect the process of selecting paragraphs of a text, average assessment of completeness of precedents retrieval changes little at different ratios of importance factors of aspects, but redundancy degree changes noticeably.

8. Discussion of results of development of a model and decision support technology based on a multi-aspect factographic search

The designed theoretical-multiple model of DSS allowed us to consider diversity of functions of the system, its composition and connections between individual elements at the system level. This allowed further development and testing of information technology of multi-aspect factographic search.

As a benefit of the proposed information technology, we can note its simplicity and effectiveness. The latter is achieved, first of all, by generation of a package of secondary queries on certain aspects. This provides sufficient completeness of the sample of documents for analysis. Second, filtering of the content of each document by certain aspects allows guaranteed and even redundant detection of precedents, containing the facts, necessary for a user. Redundancy is then eliminated by threshold processing of the found material and by using importance weight factors of aspects. The structure of a document makes no difference. It can be an arbitrary text or totalities of values of heterogeneous database regions. The system minimizes actions of a user who does not need to generate multiple queries and take care of solving a multi-aspect problem.

A shortcoming of the technology at this stage is the lack of a mechanism for automatic co-reference detection, i. e., multiplicity of options for designation of the same entity. This drawback is partially compensated by possibility of continuous learning of the system through simple replenishment of thesaurus of a subject area.

SDD, which implements the proposed technology of multidimensional factographic search, allowed the department of sales and marketing of an enterprise to identify potential customers and increase profits.

9. Conclusions

1. We proposed a theoretical-multiple model that describes the composition and structure of the document-driven DSS with possibility to analyze unstructured textual information. The specific feature of the model is the existence of a factographic search engine based on logical-linguistic model and the method of identifying precedents, affecting different aspects of the problem area. The main objectives of the system are:

- processing of the user’s query and selection of keywords and lexemes with subsequent separation of aspects and generation of secondary search queries;
- search for relevant documents;
- versatile factographic analysis of selected documents by aspects, assigned a priori;
- generation of search results in the form of textual fragments, containing information, which is necessary for a user in the context of assigned aspects.

2. We proposed the information technology of a multi-aspect factographic search, allowing us, based on a primary user’s query, to generate a query group in accordance with the aspects of the problem to be solved. Subsets of aspect-relevant documents are separated. In each document, aspect-relevant precedents are detected. At the last stage, a redundancy of the search results is eliminated.

3. A test of the information technology was performed employing a marketing task as an example. Results of assessment of completeness of search by aspects and, on average, by a task were obtained. It was shown that the use of importance factors of aspects reduces redundancy in search results.

References

1. Kravchenko, T. K. Classification of decision support systems [Text] / T. K. Kravchenko, N. N. Seredenko // Open Education. – 2010. – Vol. 4. – P. 71–79.
2. Voxmintsev, A. V. Integration of professional knowledge bases with the help of factual analysis and ontological models. Vol. 1 [Text] / A. V. Voxmintsev // Process of the 13th International Workshop on Computer science and information technologies (CSIT'2011), 2011. – P. 61–66.
3. Lopez, P. GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications [Text] / P. Lopez // Lecture Notes in Computer Science, 2009. – P. 473–474. doi: 10.1007/978-3-642-04346-8_62
4. Ventura, J. Mining concepts from texts [Text] / J. Ventura, J. Silva // Procedia Computer Science. – 2012. – Vol. 9. – P. 27–36. doi: 10.1016/j.procs.2012.04.004
5. PhridviRaj, M. S. B. Data mining – past, present and future – a typical survey on data streams [Text] / M. S. B. PhridviRaj, C. V. GuruRao // Procedia Technology. – 2014. – Vol. 12. – P. 255–263. doi: 10.1016/j.protcy.2013.12.483
6. Urbain, J. Mining heart disease risk factors in clinical text with named entity recognition and distributional semantic models [Text] / J. Urbain // Journal of Biomedical Informatics. – 2015. – Vol. 58. – P. S143–S149. doi: 10.1016/j.jbi.2015.08.009
7. Bhanuse, S. S. Text Mining using Metadata for Generation of Side information [Text] / S. S. Bhanuse, S. D. Kamble, S. M. Kakde // Procedia Computer Science. – 2016. – Vol. 78. – P. 807–814. doi: 10.1016/j.procs.2016.02.061
8. Munkova, D. Data Pre-Processing Evaluation for Text Mining: Transaction/Sequence Model [Text] / D. Munkova, M. Munk, M. Vozar // Procedia Computer Science. – 2013. – Vol. 18. – P. 1198–1207. doi: 10.1016/j.procs.2013.05.286
9. Cao, Y. Mining Large-scale Event Knowledge from Web Text [Text] / Y. Cao, P. Zhang, J. Guo, L. Guo // Procedia Computer Science. – 2014. – Vol. 29. – P. 478–487. doi: 10.1016/j.procs.2014.05.043
10. Gadia, K. Parallel text mining in multicore systems using FP-Tree algorithm [Text] / K. Gadia, K. Bhowmick // Procedia Computer Science. – 2015. – Vol. 45. – P. 111–117. doi: 10.1016/j.procs.2015.03.100
11. Artamonov, V. V. Development of information search model with the use of linked data [Text] / V. V. Artamonov, V. O. Tertishnyi // Information Processing Systems. – 2015. – Vol. 10, Issue 35. – P. 69–75.
12. Tertishnyi, V. Model and method for multi-aspect search of factographic facts for support of decision-making [Text] / V. Tertishnyi, I. Shevchenko // Transactions of Kremenichuk Mykhailo Ostrohradskyi National University. – 2016. – Vol. 5 (100), part 1. – P. 19–25.