Informatics and Control Problems 42 Issue 2 (2022)
journal homepage: www.icp.az

# Domain-oriented information search on the Internet

**A.M. Abbasov[1*], V.A. Gasimov[2]**

[1]*Institute of Control Systems of Azerbaijan National Academy of Sciences, Baku, Azerbaijan*
[2]*Azerbaijan Technical University, Baku, Azerbaijan*

**ARTICLE INFO**

*Article history:*
Received 16.09.2022
Received in revised form 28.09.2022
Accepted 07.10.2022
Available online 18.11.2022

*Keywords:*
Information retrieval
Information search
Information model
Search system
Domain-oriented search
Domain-oriented search system
Virtual web environment
User query
Fuzzy model

**ABSTRACT**

*The article is devoted to the study of the problem of information search on the Internet by using knowledge about the domains of the Web space. The domain information model (IM), the virtual Web environment model (VWE) and the user query model (UQ) are proposed, as well as the relationships and mapping principles between them are investigated. The possibility of providing a thematic focus, improving the quality and efficiency of the search process, improving results using an information model of domain representation is considered. Based on this model, a domain-oriented information search system for the Internet Web space has been developed. To improve the quality of the search engine and search results, a fuzzy logic approach is used, thanks to which a fuzzy domain-oriented information model of the Web space is proposed. Within the framework of the proposed models, methods have been developed for determining the content connectivity of Web documents, dividing the Web space into related Web areas, and virtual unification of Web areas.*

## 1. Introduction

An analysis of the problems of the Internet information space, as well as studies in the field of information search on the Internet, shows that in the current context of astonishing growth of the amount of information on the Internet, the efficiency of information search tools is clearly lagging behind the desired level. The main reason is primarily the poor organization of information resources, including Web-documents in the Internet environment [1-6].

This is due to the following factors:

– information resources of the Internet, including Web documents, are created independently of one another in an arbitrary form, at the discretion of the authors or owners;
– each information resource is characterized only by a certain set of features, at best, and their frequency characteristics;
– links between information resources are generally placed without regard to their subject matter, and a logical and meaningful link between information resources therefore exists only between neighboring ones, at least through two or three links;

[*]Corresponding author.
*E-mail addresses*: pr.dr.abbasov@gmail.com (A.M. Abbasov), gasumov@yahoo.com (V.A. Gasimov).

www.icp.az/2022/2-01.pdf https://doi.org/10.54381/icp.2022.2.01

– existing search engines mainly allow quasi-random (poorly targeted) search on the Web, the effectiveness of which depends on the level of knowledge of the subject area of the search.

The above shows that for search engines to be effective, the Internet information resources need to be systematized (indexed, classified, etc.). This is to some extent carried out in the most widely known Internet search engines, such as Google, Bing, Yandex, etc. The main disadvantage of this approach is that each Web document is characterized by a set of keywords and its location on the Internet, i.e., URL, or each keyword or word combination is matched by a set of URLs in which Web documents with these information attributes are located [7-9, 24].

Normally, semantic relatedness between information resources and their subject matter proximity are practically not taken into account in search organization [10,11]. Consequently, with such an approach, information search can be non-oriented, perhaps even moving in a "false" direction. Studies show that modern information search systems do not allow making full use of the capabilities of Internet resources, efficiency, accuracy, and completeness of information search generally does not reach the required level and leaves much to be desired [4,5,16].

Despite the great complexity of the Internet information space, it can be systematized in such a way that it is rationally covered by search engines. In other words, it is proposed to partition the information space into search zones (domains) such that they are the coverage areas of individual search engines. This implies that the semantic relatedness between such search zones should be quite weak, but the information resources included in the same search zone should have a greater content relatedness [12-16].

Through this distribution of information resources across domains using the proposed approach, information search becomes partly subject-oriented and follows the principle of navigating through content-relevant links, rather than "random" hyperlinks or keywords [16-22].

The purpose of this research is to develop methods and tools for semantically supported information search on the Internet and the use of fuzzy logic capabilities to improve the efficiency of the search process.

– Given the above, this article is devoted to the study of methods and tools to develop a domain-oriented search system (DOSS) to organize information search using the information model of domains, ontology, hypertext structure of information resources of the Internet. The basic principles of DOSS are the use of the information model of domains for the description of information resources of the Internet, the user query and the execution of the data mapping mechanisms between them. Further in the paper we propose a domain information model (IM – Information Model) to describe the resources of information space of the Internet, a Virtual Web Environment model (VWE – Virtual Web Environment) and a user query model (UQ – User Query). Methods of mapping one model to another, determining the relevance of mapping and ranking of results are also discussed in the context of the proposed models.

## 2. DOSS architecture and operating principle

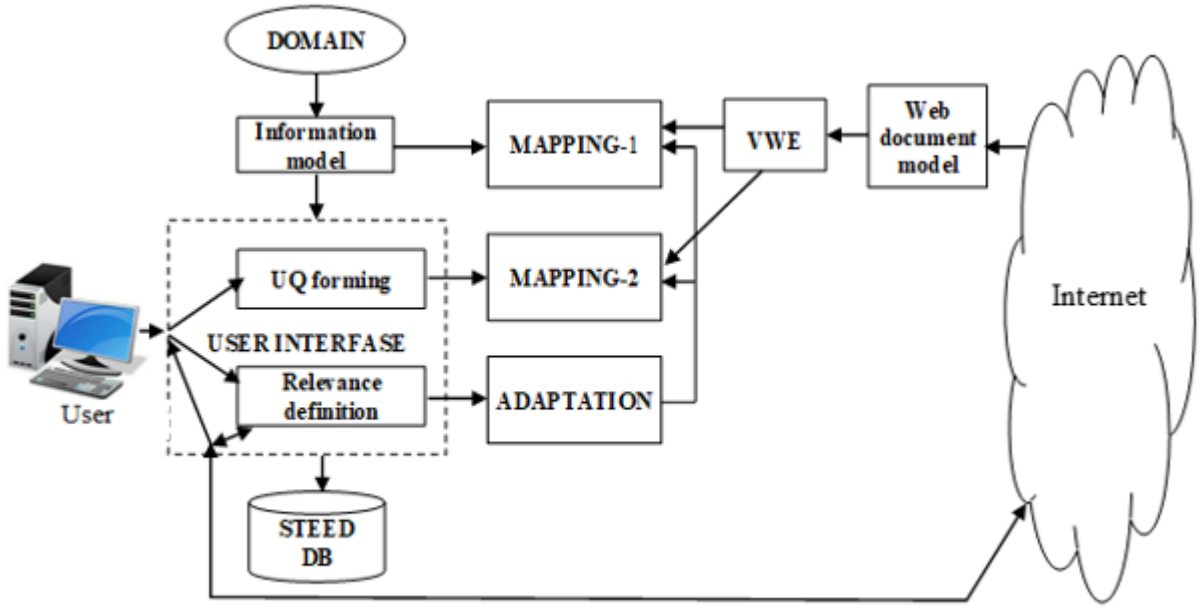
Based on the goal set, in this paper, the domain-oriented search is achieved within the proposed models by completing the following tasks:

- Building a Web document model;
- Building a model of the virtual Web environment of the Internet using the information model of domains:

$$\mathbf{VWE} \longleftarrow \text{context closest} \longleftarrow \text{IM};$$

- Developing the knowledge base ($KB_1$) for IM;
- Developing the knowledge base ($KB_2$) for VWE;
- Building the user interface – forming UQ, using IM;
- Mapping between the information model of domains IM and the virtual Web environment VWE using $KB_1$ and $KB_2$;
- Mapping between UQ and VWE: creating mapping of $KB_м$;
- Adaptation of the knowledge bases of the system;
- Optimizing the view.

In the following paragraphs, we propose the development of a DOSS that includes all the necessary blocks and modules to solve the above tasks. The generalized structure of a DOSS is shown in Fig.1.




Fiq.1. The generalized structure of a domain-oriented search system

As we see in the figure, the DOSS user interface consists of three modules [22]:

- the UQ forming module;
- the browser module (Google Chrome, Mozilla Firefox, Safari, Opera, Internet Explorer, etc.);
- module for assessing the relevance of search results to the user query.

UQ formation allows the user to express their requirements to the information model hierarchy and, using the available browser, sends UQ through the mapping block Mapping-2 to the VWE model.

VWE contains information about Internet domains in some form similar to IM and uses a Web document model. For creating a VWE with an IM structure system, there is a second mapping block Mapping-1 between the IM and the VWE. To create a virtual environment VWE, available Web browsers are used, making it possible to obtain information about domains and the Web document model.

The Mapping-2 block, which has knowledge of UQ and VWE, implements the mechanism of mapping between UQ and VWE. In addition, Mapping-2 implements browsing (search) optimization, which means determining which direction and which search route would be the optimal way to create the UQ.

When the user receives a list of Web documents from the VWE, the block for assessing the relevance of search results to the user query uses the relevance criterion to implement a ranking of

the Web documents found.

If we are dealing with an IM with fuzzy relations, then the adaptation block can implement the mechanism of adaptation by Mapping-1 and Mapping-2, depending on the results of the description of relevance.

The formal representation of DOSS models and the operating principles of the mapping blocks are discussed in the following sections.

## 3. Formal representation of the information model of Web space domains

As a result of this research, an information model of domains IM is proposed. In this model, it is assumed that the information space is organized as a hierarchical structure, at the levels of which Web areas containing objects with their attributes are defined [21,22,24-28]. In other words, the information space (Web space) of the Internet consists of domain areas containing sets of websites servers and other information services, which we will call domain objects.

It is assumed that an IM includes a set of objects and attributes, a set of relations between objects and their attributes, as well as between the objects themselves, a hierarchy of objects, which are represented by appropriate categorization, rules, functions, etc. Each object is described by a set of attributes, descriptors, keywords and terms. Suppose that $N$ is the maximum number of levels in the Web space domain hierarchy, $E = \{e_i\}_n$ is the set of objects (entities) in the domain, $A = \{a_j\}_m$ is the set of attributes used to describe objects in the domain. The principles of placement of objects on the levels, relations between objects, between objects and their attributes, and between attributes of objects are also defined in this model.

In view of the above, the information model IM of domains in general form can be represented as follows:

$$IM \to \{N, E, H(E), H(E, E^*), A, R(E, A), R(A, A^*), R(E, E^*)\} \quad (1)$$

where

- $H(E)$ is the matrix (or vector) of placement of objects (categories) on the domain hierarchy levels, where $H(e_i) = 1$, if the object $e_i$ is in the domain, $H(e_i) = 0$, otherwise, $i = \overline{1,n}$;
- $H(E, E^*)$ is the matrix of categorization relations (subcategory and supracategory nesting) between objects, where $H(e_i, e_j^*) = 1$, if the object $e_i$ has a nesting relationship with the object $e_j^*$, $H(e_i, e_j^*) = 0$, otherwise, $i, j = \overline{1,n}$;
- $R(E, A)$ is the matrix of relations between objects and attributes of domains, where $R(e_i, a_j) = 1$, if the object $e_i$ has a relevance relationship with the attribute $a_j$, $R(e_i, a_j) = 0$, otherwise, $i = \overline{1,n}$, $j = \overline{1,m}$;
- $R(A, A^*)$ is the matrix of relations between attributes describing domain objects, where $R(a_i, a_j^*) = 1$, if the attribute $a_i$ has a relevance relationship (e.g., synonyms, associative words, etc.) with the attribute $a_j^*$, $R(a_i, a_j^*) = 0$, otherwise, $i, j = \overline{1,m}$;
- $R(E, E^*)$ is the matrix of relations between objects of domains $R(e_i, e_j^*) = 1$, if the object $e_i$ has a relationship with the object $e_j^*$, $R(e_i, e_j^*) = 0$, otherwise, $i, j = \overline{1,n}$.

## 4. Information model of the virtual Web-environment

Based on information model of domains (1), the VWE information model can be represented as:

$$VWE \to \{URL, K, R(URL, K), R(K, K), R(URL, URL)\}, \quad (2)$$

where

- $URL$ is the set of linked (having a relevance relation) Web documents of the Internet;
- $K$ is the set of descriptors (keywords) used to describe Web documents;
- $R(URL, K)$ is the matrix of relations between Web documents and keywords;
- $R(K, K)$ is the matrix of relations between keywords within Web documents;
- $R(URL, URL)$ is the matrix of links between Web documents.

VWE is part of the Web environment of the Internet and is related (relevant) to the domain in question. The optimal structure for VWE is an IM structure, because in it the user forms their quiery (UQ) according to the information model of domains. The VWE model formally has a graph structure. The vertices of this graph correspond to Web documents of the Internet, related by context (content) to a given domain, and the arcs correspond to the links (relations) between Web documents.

It follows from the above that the set of Web documents $URL$ is also a domain object, so we can write $URL \subset E$. Then, the set of descriptors (keywords) $K$ is also a subset of the set of attributes, i.e., $K \subset A$. Given this, the relational matrices $R(URL, K)$, $R(K, K)$ and $R(URL, URL)$ can be considered as $R(K^E, K^A)$, $R(K^A, K^A)$ and $R(K^E, K^{E^*})$. Here $K^E$ is the subset of descriptors (keywords) describing the object $E$, $K^A$ is the set of descriptors (keywords) describing Web documents, i.e., $K^E \in E$ and $K^A \in A$.

The relational matrices between objects and attributes within the VWE model are presented as follows:

- $R(K^E, K^A)$ is the relationship between objects and their attributes that define Web documents in the form of

$$V_1 = \left\| v_{ij}^1 \right\|_{n \times m},$$

where

$$v_{ij}^1 = \begin{cases} URL, & \text{if } R(K^E, K^A) = 1 \\ 0, & \text{otherwise} \end{cases}$$

- $R(K^E, K^{E^*})$ is the relationship between objects that define Web documents in the form of

$$V_2 = \left\| v_{ij}^2 \right\|_{n \times n},$$

where

$$v_{ij}^2 = \begin{cases} URL, & \text{if } R(K^E, K^E) = 1 \\ 0, & \text{otherwise} \end{cases}$$

- $R(K^A, K^A)$ is the relationship between the descriptors (keywords) of Web documents in the form of

$$V_3 = \left\| v_{ij}^3 \right\|_{m \times m},$$

where

$$v_{ij}^3 = \begin{cases} 1, & \text{if keywords } i \text{ and } j \text{ are correlated} \\ 0, & \text{otherwise} \end{cases}.$$

The latter matrix makes it possible to use relations between keywords to enhance the efficiency of the search result by taking into account the semantic relatedness of words, synonyms and associative words.

## 5. User query model

Since the user forms a query according to the IM model, the user query can also be represented

within this model. However, it is clear that this model will not contain all the information characteristics of the IM model, but only some of them. Taking into account that the user searches for Web documents from the information space of the Internet, which contain (usually) keywords and links from (2), then the user query can be generally represented as

$$UQ \to \{E, A, R(K^E, K^A), R(K^E, K^{E^*})\}. \quad (3)$$

where

$$R(K^E, K^A), R(K^E, K^{E^*}) \in \{0,1\}.$$

In this case, the weights of keywords and links from other Web documents can be used to filter results.

## 6. Mapping between VWE and IM models

Note that the VWE is an information model of information sources, which can never be equivalent to $IM_D$ domains, but the VWE must have a structure that is very close to the IM structure.

Now we consider the problem of transforming model (2) into model (1) for a specified domain. As noted above, the main parameters in the information model of domains are $E$ – a set of objects and $A$ – sets of their attributes. Given that the sets of keywords are taken from the set of descriptors

$$K \in E \cup A,$$

then we can consider the transformation of model (2) into model (1) as

$$K \in K^E \cup K^A,$$

where $K^E \in E$ and $K^A \in A$.

Note that it is not difficult to distinguish (classify) $K^E$ from $E$ and $K^A$ from $A$ for the specified objects.

Thus, the relationship between objects and attributes $R(K^E, K^A)$ in the set of keywords for VWE will be

$$R(K^E, K^A) = \begin{cases} 1, \text{if } R(K^E, K^A) = 1 \\ 1, \text{if } R(URL(K^E), URL(K^A)) = 1 \\ 0, \text{otherwise} \end{cases}$$

Here the expression $R(K^E, K^A) = 1$ means that within Web documents there are links between $K^E$ and $K^A$, and the expression $R(URL(K^E), URL(K^A))$ shows the links between Web documents containing $K^E$ and $K^A$.

Similarly to the expression $R(K^E, K^A)$, the relations are determined between objects $R(K^E, K^E)$ with respect to their keywords for VWE:

$$R(K^E, K^A) = \begin{cases} 1, \text{if } R(K^E, K^E) = 1 \\ 1, \text{if } R(URL(K^E), URL(K^E)) = 1 \\ 0, \text{otherwise} \end{cases}$$

## 7. Mapping the UQ model to the VWE model: information search

Mapping the UQ model to the VWE model is the process of finding information in the VWE based on the UQ. It determines the correspondence between the UQ and VWE elements:

- the set of objects of the user query $E$ is mapped to the set of keywords (descriptors) $K^A$ describing these objects of WVE, i.e., $E \to K^E$;

- the set of attributes $A$ of domain objects is mapped to the set of keywords (descriptors) describing these attributes, i.e., $A \to K^A$;
- since the sets $E$ and $A$ are mapped, respectively, to the sets $K^E$ and $K^A$, then the relations between objects and attributes, as well as between objects themselves, will be mapped to the following relations:

$$R(E, A) \to R(K^E, K^A)$$

and

$$R(E, E^*) \to R(K^E, K^{E^*})$$

According to this mapping, the selection (search) of Web documents from the VWE upon query is implemented by one of the following lines:

| № | $E \to K$ | $A \to K^A$ | $R(E, A) \to R(K^E, K^A)$ | $R(E,E^*) \to R(K^E,K^{E^*})$ | Results (Web documents) |
|---|---|---|---|---|---|
| 1. | T | T | T | T | $V_1 \cap V_2$ |
| 2. | T | T | T | F | $V_1$ |
| 3. | T | T | F | T | $V_2$ |
| 4. | T | T | F | F | $V_1 \cup V_2$ |
| 5. | T | F | F | T | $V_2$ |
| 6. | F | T | F | F | $V_1$ |
| 7. | F | F | F | F | - |

## 8. Search for Web documents, determining relevance and ranking search results

Determining the relevance of Web documents given to the user as a search result satisfying the query in the most correct manner, i.e., the best search result of Web-documents based on a user query [7,19,22,24].

There are two methods of determining relevance:

- feedback, i.e., the user's response;
- analytical determination of relevance on the basis of the search result and user query.

In the first option, search results are evaluated on a certain rating scale by users as their query is satisfied. In most cases, this is not done by users, and sometimes the evaluations are subjective. Therefore, we do not consider this method.

Let us dwell on the second method. Here we have the following two sets for comparison:

$$UQ \to \{E, A, R(K^E, K^A), R(K^E, K^{E^*})\}$$

and

$$URL \to \{K, R(URL, K), R(K, K), R(URL, URL)\}.$$

The second expression follows from model (2). In the latter expression. $R(URL, K)$ is the relevance of the keywords to the Web documents where they occur, i.e., the weights of keywords $K$ for the corresponding Web documents, $R(K, K)$ is the relations (relevance relationship) between the keywords within Web documents (these relations are implemented within a single Web document), $R(URL, URL)$ is the relevance relationship between different Web documents.

It can be seen that maximum relevance can be achieved by satisfying the following conditions:

1. $E, A \in K$
2. $R(URL, K^E) \cdot R(URL, K^A) = \max\limits_k R(URL, K)$ (4)
3. $R(K^E, K^{E^*}) \& R(K^A, K^A) \& R(K^E, K^A) \& R(URL, URL) = 1$

4. $R\left(K^E, K^{E^*}\right)\&R\left(K^E, K^{E^*}\right)\&R(E, E^*)\&R(URL, URL) = 1$

Expression (4) can be used to rank the results, allowing the user to have an alternative to select the most suitable (optimal options of) Web documents. In this case, it is possible to perform an automatically adapting search process, because we are dealing with a specific VWE.

## 9. Fuzzy domain-oriented information model of the Web-space

The fuzziness of a domain-oriented model is related to the assumed fuzziness in the description and presentation of Internet information sources, as well as in the building of search queries [7,9,16]. Thus, practice shows that when creating, designing and presenting Internet information resources, authors and owners use different keywords that are not clearly related to their content. In Web documents of the same area, different keywords (synonyms) defining the same concept and having fuzzy relations of proximity can be used. In other words, keywords have a fuzzy relevance relationship to the subject matter of the Web document content. In this case, the keywords describing the Web document can be defined by the functions of their membership in the set of terms, taking values in the interval [0,1].

Search queries are also fuzzy in two respects: firstly, the above fuzzy set of keywords (or synonyms) and their relations to Web documents are used in creating them; secondly, search queries are the result of subjective thinking of users who, due to lack of knowledge of the subject area or inexperience in creating queries even with good knowledge of the subject area, may not formulate a search query clearly.

Considering the above and models (1)-(4), we can conclude that a fuzzy model of the Internet information space should describe the fuzzy content relatedness between the available Web documents and the geographical transparency of the distributed information resources of the Web-space. Based on the above, the information space of the World Wide Web can be formally represented in the following form [6,8,13,27]:

$$IM_f = \{URL, K, D, R(URL, K), R(K, K), R(URL, URL)\}, \quad (5)$$

where

- $URL = \{u_1, u_2, \dots, u_m\}$ is the set of Web documents;

- $K = \{k_1, k_2, \dots, k_n\}$ is the set of information attributes (keywords, terms) describing Web-documents, in the future they will be called keywords;

- $D = \left\|d_{jj'}\right\|_{m\times m}$ is the matrix of distances between Web documents (or sites, or subspaces), where distances $d_{jj'} \in \{0,1,2, \dots\}, j, j' = \overline{1,m}$ are determined by the number of levels of hyperlinks between Web-documents;

- $R(URL, K)$ is the relationship between Web documents and the keywords *K* that describe them. The keywords extracted from or attached to Web documents during indexing may reflect their content unclearly. These relations will be described by the matrix $W = \left\|w_{ij}\right\|_{n\times m}$;

- $R(K, K)$ is the relationship between the keywords in Web documents. The relevance relations of keywords are usually fuzzy, that is, one keyword may be relevant to another keyword in some way and not clearly reflect the meaning of the first (for instance, synonyms have close but not exactly the same meanings);

- $R(URL, URL)$ is the relationship (links) between Web documents established through links. This relationship will be described by the matrix $R = \left\|r_{jj'}\right\|_{m\times m}$, where $r_{jj'} \in \{0,1,2, \dots\}, j, j' = \overline{1,m}$ determine the number of links between the Web documents.

It should be noted that the set of Web documents $URL = \{u_i\}_m$ included in model (5) is represented by the subset of the keywords $K_i$ of the set $K$, i.e., the subset of the keywords mapped to these Web documents. The keywords are assigned weight coefficients determining the degrees of importance of these terms for each Web document.

Suppose that $K_i = \{k_{i1}, k_{i2}, \dots, k_{in}\}, i = \overline{1,m}$ is the subset of keywords describing the Web document $u_i$. The weight coefficients of the keywords with respect to each Web-document are determined as a function of the membership of these keywords in the subset $K_i$. The membership function takes values in the interval [0,1] and its values indicate the degree of membership of a particular keyword in the subset $K_i$ and determine the degree of subject matter proximity (relevance) of it to the content of the Web document. In particular case, the value "1" indicates full membership (full relevance) of the keyword in this subset, and the value "0" indicates zero membership (lack of relevance).

It should be noted that all subsets $K_i$ are normalized, i.e., the keywords of the set $K$, which are absent in $K_i$, are added to the latter, but the values of their membership functions are taken as equal to zero. Thus, each document is described by a vector of size *n.* The elements of this vector define the values of the functions of membership of each keyword $k_j$ in the subset $K_i$.

In general, the set of Web documents $URL = \{u_1, u_2, \dots, u_m\}$ can be represented as a fuzzy $m \times n$ relational matrix. The element at the intersection of row $i$ and column $j$ determines the value of the membership function of the keyword $k_j$ for the Web-document $u_i$. The Web document $u_i$ is represented in the search engine index database using fuzzy relations as follows:

$$u_i = \left\{k_j / w_{k_j}(u_i)\right\}, i = \overline{1,m}, j = \overline{1,n}.$$

Here the relationship between the keyword $k_j$ and the Web document $u_i$ is determined as:

$$W = \left\{w_{k_j}(u_i) : K : K_i \to [0,1]\right\}, i = \overline{1,m}, j = \overline{1,n} \quad (6)$$

where $w_{k_j}(u_i)$ is the function of membership of the keyword $k_j$ in the subset $K_i$ of the Web document $u_i$, in other words, the weight coefficient of the keyword $k_j$ for the Web document $u_i$.

## 10. Determining the content relatedness between Web documents

If we take into account the weight coefficients of the keyword for the Web-documents $w_{ij} \in [0,1], i = \overline{1,n}, j = \overline{1,m}$, then taking into account the introduced parameters, the content relatedness of Web documents can be determined as [10,16]

$$\alpha_{jj'} = \sum_{i=1}^{n} w_{ij} \cdot w_{ij'}, j, j' = \overline{1,m} \quad (7)$$

Further we consider the distribution of Web space on three levels of hierarchization. At the first level a set of local Web areas is determined, which can correspond to individual Web servers or groups of them, united by thematic or geographical features. At the second level, regional Web areas are determined, which may consist of a set of groups united by local Web areas. Finally, the third level determines the global Web space, which unites both thematically and geographically distributed Web areas.

In the following paragraphs, we consider the tasks that allow formally obtaining the mapping of the Web space of the Internet to the proposed three-level structure depending on its information-topological structure. This structuring corresponds to the hierarchization of the Web space and creates prerequisites for the organization of a multilevel subject-oriented and meta-search system.

## 11. Partitioning the Web space into related Web areas

The connectedness of Web areas $v_l \in V, l = \overline{1,L}$ refers to the degree of connectivity of the Web documents $u_j \in URL, j = \overline{1,m}$ included in these areas. Let us introduce the variables $x_{jl}, j = \overline{1,m}, l = \overline{1,L}$ determined as: $x_{jl} = 1$, if the Web document $u_j$ is part of the Web area $v_l$, $x_{jl} = 0$ if not [10,16].

The optimal partitioning of the Web space into maximally connected Web-areas is analytically expressed through the total connectivity of areas $v_l \in V, l = \overline{1,L}$:

$$\sum_{l=1}^{L}\sum_{j'=1}^{m}\sum_{j=1}^{m} x_{jl}x_{j'l}\alpha_{jj'} \to \max_{x_{jl}} \tag{8}$$

The following conditions should be met here:

- each Web document should be included in at least one Web area of $v_l \in V, l = \overline{1,L}$:

$$1 < \sum_{l=1}^{L} x_{jl} < k, j = \overline{1,m}, \tag{9}$$

where *k* is the maximum allowed number of copies of Web documents.

- the number of Web documents in each area should have both upper and lower limits:

$$Q_{min} \le \sum_{j=1}^{m} x_{jl} \le Q_{max}, l = \overline{1,L}, \tag{10}$$

Thus, the solution of problem (8)-(10) will give the intended partitioning of the Web space into the maximal connected Web areas with an upper limit of their number. This allows narrowing the scope of the Internet search.

## 12. Virtual joining of Web areas

The second task in the posed problem is joining Web areas $v_l \in V, l = \overline{1,L}$ into some subsets $\omega_v \in \Omega, v = \overline{1,v^m}$ on the basis of their proximity. This task is solved when building metadata search engines, or subject-oriented search engines. The proximity between Web areas is determined as [14, 30]:

$$\beta_{ll'} = \sum_{j'=1}^{m}\sum_{j=1}^{m} x_{jl}x_{j'l}\alpha_{jj'}. \tag{11}$$

To determine the membership of Web areas $v_l \in V, l = \overline{1,L}$ in the subsets to be joined, we introduce the variables $y_{lv}, l = \overline{1,L}, v = \overline{1,v^m}$ where $y_{lv} = 1$, if the domain $v_l$ is a member of the subset $\omega_v$, otherwise, $y_{lv} = 0$. Here $v^m$ is the maximum number of subsets. Note that in this formulation $v^m$ is preset. Although this provision limits the efficiency of application of the solution of the problem in practice, however, in the first approximation it is necessary to simplify the solution of the problem.

Thus, the total proximity of the subsets $\omega_v$ in the optimal joining of the Web areas $v_l \in V, l = \overline{1,L}$ will be:

$$\sum_{v=1}^{v^m}\sum_{l=1}^{L}\sum_{l'=1}^{L}\beta_{jj'}y_{lv}y_{l'v} \to max \tag{12}$$

Next, we determine the conditions arising from the practical requirements of the problem:
- each Web area $v_l \in V, l = \overline{1,L}$ should be part of at least one subset $\omega_v$:

$$\sum_{v=1}^{v^m} y_{lv} \geq 1, l = \overline{1,L}, \tag{13}$$

- each subset $\omega_v$ can include only a limited number of areas $v_l \in V, l = \overline{1,L}$:

$$\sum_{l=1}^{l} y_{lv} \leq B, v = \overline{1,v^m}, \tag{14}$$

Thus, the problem of joining Web areas on the basis of their proximity comes down to the solution of a bilinear programming problem, i.e., finding $y_{lv}, l = \overline{1,L}, v = \overline{1,v^m}$ such that maximize functional (12) and satisfy constraints (13)-(14).

Often the existing set of links $\|r_{jj'}\|_{m\times m}$ is far from reality, i.e., does not reflect the actual relatedness of Web documents. The question arises: what should the web be like so that with a limited number of links the content relatedness is maximum? In this case this problem can be formally framed as follows: find $r_{jj'}, j, j' = \overline{1,m}$ such that allow synthesizing the optimal web, i.e., the web with the optimal links. The optimal links are the links connecting Web documents with the maximum content-relevant proximity at the minimum distance:

$$l_{jj'} = \frac{1}{d_{jj'}}\sum_{i=1}^{n} w_{ij}w_{ij'}, j, j' = \overline{1,m}, \tag{15}$$

where $l_{jj'}$ is the coefficient of content-relevant proximity of the Web documents $u_j$ and $u_{j'}$, $d_{jj'} \in \{0,1,2, ...\}$ is the distance between the Web documents $u_j$ and $u_{j'}$.

The set of optimal links can be determined using the following functional:

$$\sum_{j=1}^{m}\sum_{j'=1}^{m} l_{jj'}r_{jj'} \to max \tag{16}$$

The number of links should be limited to limit the confusion on the web:

$$r_{jj'} \leq r^m, j, j' = \overline{1,m} \tag{17}$$

Thus, the solution of the problem of construction of the optimal web is reduced to determining $r_{jj'}, j, j' = \overline{1,m}$ such that would give a maximum to functional (16) and satisfy constraints (17). However, in this formulation the efficiency of solution (16) strongly depends on the choice of the maximum value of the number of links $r^m$.

Based on the solution of the above two tasks, the following conclusions can be tentatively drawn:

- the Web areas that fall into the same subset can be covered by the same search engine;
- areas that fall into multiple subsets simultaneously should be covered by multiple search engines, which are global or meta-search engines.

## 13. Conclusion

The proposed domain-oriented information model of the Web space allows us to deliberately influence the process of information search. The methods developed for this purpose make it possible to partition the Web space into such Web areas (domain zones) that could act as separate search zones. When creating such zones in addition to the subject matter proximity of Web documents belonging to the same zone, factors such as their content relatedness via hyperlinks, the distance between the subjects of the web containing these resources, are taken into account. It should be noted that the solution of these problems can improve the efficiency of information search by systematizing the Web space, narrowing the scope of search and organizing a subject-oriented search, while leaving the software and hardware parameters of search engines unchanged. To improve the quality of the search engine and search results, as well as to describe Web documents, relations between them and the keywords, to organize a domain-oriented search, the fuzzy logic approach and a fuzzy model of the Web space are used.

## References


[1] A.M. Abbasov, Information boom: New trends and expectations, Word conference on soft conference, San Francisco, California, May 23-26, 2011, p.19.

[2] V.A. Gasimov, Methods for construction of information retrieval systems based on a hierarchical model of the information space of the Internet, Automatic control and computer sciences, Allerton Press, Inc., New York. 36 No.1 (2002) pp.33-43.

[3] P. Khramtsov, Modeling and analysis of information retrieval systems of Internet, Open systems. No.6 (1996). http://www.osp.ru/os/1996/06/46.htm. [In Russian: P. Khramtsov, Modelirovaniye i analiz raboty informatsionno-poiskovykh sistem Internet. Otkrytyye Sistemy. No.6 (1996)].

[4] W. Haslam, A browser driven by classification information model, STEED/T5/01/1, University of Manchester. (1997) pp.1-6.

[5] N. Ivanov, The path to effective search, Open systems, DBMS. No.01 (2010). https://www.osp.ru/os/2010/01/13000681. [In Russian: N. Ivanov, Put' k effektivnomu poisku, Otkrytyye sistemy, SUBD. (2010)].

[6] Ch.D. Manning, P. Raghavan, Schütze H. Introduction to information retrieval, Online edition, Cambridge, (2009). http://www.informationretrieval.org/

[7] V.A. Gasimov, Methods of information retrieval in computer networks with supersaturated information resources, Monograph, Baku, Elm. (2004) 208 p. [In Russian: V.A. Gasimov, Metody informatsionnogo poiska v komp'yuternykh setyakh s sverkhnasyshchennymi informatsionnymi resursami].

[8] S.V. Ratnikov, Domain-oriented data model taking into account the ordering properties, Abstract of the dissertation for the degree of candidate of technical sciences, Penza. (2006). http://tekhnosfera.com/domenno-orientirovannaya-model-dannyh-s-uchetom-svoystv-uporyadochennosti #ixzz74Zs5HGy0 [In Russian: Ratnikov S.V. Domenno-oriyentirovannaya model' dannykh s uchetom svoystv uporyadochennosti].

[9] V.A. Gasimov, Information search methods and systems, Textbook, Baku. (2015) 288 p. [In Azerbaijani: V.A. Gasimov, Informatsiya akhtarish usullary ve sistemleri].

[10] Z. Yan, Y. Ding, E. Cimpian, Towards a domain oriented and independent semantic search model, In: Apolloni B., Howlett R.J., Jain L. (eds), Knowledge-Based intelligent information and engineering systems, KES. Lecture Notes in computer science, Springer, Berlin, Heidelberg. Vol.4694 (2007). https://doi.org/10.1007/978-3-540-74829-8_90.

[11] J. Zhang, W. Qu, L. Du, Y. Sun, A framework for domain-specific search engine: design pattern perspective, SMC'03 Conference proceedings, IEEE International conference on systems, Man and cybernetics, Conference theme-system security and assurance (Cat. No.03CH37483). Vol.4 (2003) pp.3881-3886. doi: 10.1109/ICSMC.2003.1244494.

[12] Ye.V. Romanova, M.V. Romanov, I.S. Nekrestyanov, Using intelligent network robots to build thematic collections, Programming. No.3 (2000) pp.63-71. [In Russian: Ye.V. Romanova, M.V. Romanov, I.S. Nekrest'yanov, Ispol'zovaniye intellektual'nykh setevykh robotov dlya postroyeniya tematicheskikh kollektsiy, Programmirovaniye. No.3 (2000) pp.63-71].

[13] M.N. Asim et al., Use of ontology in retrieval: A study on textual, multilingual, and multimedia retrieval, IEEE Access. Vol.7 (2019) pp.21662-21686.

[14] G.H. Yang, Dynamic information retrieval modeling for domain specific search, Final technical report, May 2018.

https://apps.dtic.mil/sti/pdfs/ AD1051882.pdf
[15] A.Z. Broder, S.C. Glassman, M.S. Manasse, and G. Zweig, Syntactic clustering of the web, Computer networks and ISDN systems, 29 No.8-13 (1997) pp.1157-1166.
[16] A.M. Abbasov, V.A. Gasimov, Organization of thematic-oriented search in global information systems, International journal of computer research, Nova sciences publishers, New York. 17 No.1/2 (2009) pp.41-58.
[17] A. Theobald, Department of computer science university of the Saarland, Germany, (2021). http://www-dbs.cs.uni-sb.de
[18] A. Anikin, D. Litovkon, M. Kultsova, An ontology-based approach to collaborative development of domain information space, Economics and education, Proceedings of the 12th International conference on educational technologies (EDUTE'6), Proceedings of the 10th International conference on business administration (ICBA'16), Barcelona, Spain, February 13-15 (2016) pp.13-19.
[19] R. Okada, E. Lee, T. Kinoshita, N. Shiratori, A method for personalized web searching with hierarchical document clustering, Transaction of information processing society of Japan. 39 No.4 (1998) pp.868-877.
[20] O. Lassia, Web metadata: a matter of semantics, IEEE Internet computing, Jule-August 1998, pp.30-37.
[21] I.S. Nekrestyanov, Subject-oriented methods of information retrieval: dissertation of the candidate of physical and mathematical sciences, St. Petersburg, (2000) 88 p. [In Russian: I.S. Nekrest'yanov, Tematiko-oriyentirovannyye metody informatsionnogo poiska].
[22] A.Yu. Sergeyev, V.M. Tyutyunnik, Methodology to increase the effectiveness of subject-oriented Internet search, Fundamental researchs. No.8-2 (2013) pp.306-311. [In Russian: A.Yu. Sergeyev, V.M. Tyutyunnik, Metodika povysheniya effektivnosti tematiko-oriyentirovannogo internet-poiska, Fundamental'nyye issledovaniya]. URL: https://fundamental-research.ru/ru/article/view?id=31914.
[23] A.M. Abbasov, V.A. Gasimov, User interfaces in distributed information retrieval systems, Devices of systems and machines, Kyev. No.5 (2003) pp.67-74. [In Russian: A.M. Abbasov, V.A. Gasimov, Interfeysy pol'zovatelya v raspredelennykh informatsionno-poiskovykh sistemakh, Ustroystva sistem i mashin, Kiyev. 2003]
[24] S. Kohli, S. Arora, Domain oriented semantic Web based personalized search engine, Fifth international conference on intelligent systems, Modelling and simulation, (2014). https://stackoverflow.com/questions/34258772/search-query-and-search-result-in-domaindrivendesign/34274033
[25] G. Gaetan, V. Saldaño, A. Buccella, A. Cechich, A domain-oriented approach for GIS component selection, Fifth international conference on software engineering advances, 22-27 Aug. 2010. https://ieeexplore.ieee.org/ document/5615004
[26] G. Shaochen, The research on domain-oriented information resource management and retrieval, International journal of machine learning and computing. 4 No.1 (2014) pp.90-93.
[27] Y. Liu, Domain-oriented retrieval model research based on meta-search, IEEE International conference on service operations and logistics, and informatics. 12-15 Oct. 2008. https://ieeexplore.ieee.org/document/4686382
[28] S. Pohorec, M. Verlič, M. Zorman, Domain specific information retrieval system, Proceedings of the 13th WSEAS International conference on computers, Greece, July 22-24, (2009) pp.465-469.
[29] Zi Lingling, Du Junping, Wang Qian, Domain-oriented subject aware model for multimedia data retrieval, Mathematical problems in engineering, Article ID 429696, Hindawi. (2013) pages 1-13. https://doi.org/10.1155/ 2013/429696
[30] V.A. Gasimov, Methods of information retrieval in Internet on the basis of fuzzy preference relations, Automatic control and computer sciences, Allerton Press, Inc., New York. 37 No.4 (2003) pp.62-67.