BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **16**, No 1
Sofia • 2016 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.1515/cait-2016-0008

Models and Algorithms of Information Retrieval in a Multilingual Environment on the Basis of Thematic and Dynamic Text Corpora

Aleksey A. Mamchich

United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Surganov St. 6, 220012 Minsk, Belarus
Email: aleksey.mamchich@mail.ru

Abstract: *Models and algorithms of information retrieval in the global and local computer networks on the basis of thematic and dynamic text corpora are proposed in this article. The developed algorithms provide the effectiveness of documents information retrieval and notable for universality, i. e., for the independence from topics. Information system adjustment to a particular data domain can be fully automated, it adds up to the creation of a respective thematic text corpus and an actualization of dictionaries of the knowledge database.*

Keywords: *information retrieval, natural language processing, statistical analysis.*

1. Introduction

At present due to the intense increase of text information volume presented in electronic form, the development of systems that provide solution for a wide range of information retrieval problems becomes more essential [7, 11]. Moreover, requirements are being constantly raised both to the effectiveness of processes of finding documents and to the integration of software complexes with different information sources (e.g., Internet resources, files of LAN or corporate databases, hard drives of personal computers, etc.). All this complicates the discovery of high performance.

Traditional approaches to the effectiveness of search mechanism increase are mostly based on various linguistic and statistical methods [5, 9, 10] which do not assure an acceptable system. Considerable problems for modern retrieval services are

the necessity for adaptation to the information needs of specific users as well as the want of a profound search on topics of their interest [2, 14].

The vector space model is one of the most widely used retrieval model in the modern information retrieval systems [17]. In this model, user queries and document search profiles are represented as vectors in multidimensional space, whose components are the weights of appropriate index terms (keywords, phrases and etc.). The similarity between user query and documents is determined as cosine angle between appropriate vectors. There are many methods to weight calculating of index terms, but most of them are based on statistical information about the occurrences of terms in a separate document or documentary array [3].

By using the vector space model here can be implemented the interactive retrieval procedure called relevance feedback which may improve the performance of information systems [8, 16]. The main idea of this strategy is initial query expansion with the keywords of relevant document subcorpus which manually selected by the user in the search results. In spite of better search performance than classic information retrieval the relevance feedback method has a number of drawbacks which limit its practical application in modern information systems [13]:

- Low search efficiency in a multilingual environment, because documents in another language are not nearby in a multidimensional vector space.
- This approach requires relevant documents to be similar to each other, but users may specify as relevant dissimilar in content texts.
- Users are often reluctant explicit feedback, or in general do not wish to prolong the search interaction.

In order to improve the efficiency of indexing and retrieving documents from various information sources, we proposed an approach which, unlike existing methods, is based on using thematic corpora (collections of texts on specific topics) as a knowledge domain and specialized knowledge base dictionaries formed on their basis. This technique provides the system adaptation to the task and the independence of a software complex from the input language. Text corpora can be created under the projected pre-task or formed directly on-line when searching for information by combining sets of documents that are relevant to each particular text or a user query (we called them – dynamic corpora). This provides the adaptation to user information needs and gives a possibility to index and search not only full-text documents but short messages too, volumes of which are small and don't allow identifying their statistical characteristics. In addition this technique ensures the similarity of relevant texts, which is archived through the automatic dynamic corpora generation from respective thematic corpora. Besides, the procedure of initial query expansion does not require the participation of users and computing resources of information system in iterative retrieval.

The proposed algorithms are notable for universality, i. e. for the independence from topics. Information system adjustment to a particular data domain can be fully automated, it adds up to the creation of a respective thematic text corpus and an actualization of dictionaries of the knowledge database. At the same time, formation of the given dictionaries can completely be carried out in the hands off.

2. The data and knowledge model of object domain

2.1. Information languages

In automating text retrieval processes we will use three information languages – input, internal and target.

The input language L_{in} is a natural language for a user interaction with the information system. There may be several input languages in the information system, for example English, German, Russian and etc.

The internal language L_{int} is a retrieval language of document search profiles and search prescriptions. It is used to describe the basic substance of texts and their formal characteristics. The internal language is represented as a finite sequence of pairs of "index term – informativity of index term".

The target language L_{trg} is used to display retrieval results for the users of the information system. L_{trg} often coincides with the input language L_{in}.

2.2. Text corpora

Suppose, we have a certain nonempty set of input language L_{in} texts (a set of texts on specific topics). Let us call it as *the thematic text corpus*. As several of these corpora are presented in an information system, as a rule, we will denote them as Ct_j (j is the number of text corpus). Let us denote a union of all thematic text corpora sets as $\text{Cf}_i = \bigcup_{j=1}^{n_i} \text{Ct}_j$ and call them as *the full corpus of texts* (i is the number of a full corpus). In the specific implementation of an information system, a full corpus of texts is created for each input language.

Under *dynamic* we will mean a text subcorpus all documents of which are relevant to a certain text document or to a certain query for information retrieving.

Let us define the concept of relevance.

Definition 1. Let T_{in} and Z_{in} be a certain nonempty sets of input language L_{in} texts. Elements of Z_{in} are queries. Then we will consider that any text $t \in T_{\text{in}}$ is *relevant* to query $z \in Z_{\text{in}}$ if there exists such surjection $\mu: T_{\text{in}} \times Z_{\text{in}} \to \{0, 1\}$ that $\mu(t, z) = 1$. If $\mu(t, z) = 0$ the text t is *irrelevant* to query z then.

Let $z \in Z_{\text{in}}$ be a certain text (specifically, a query) of input language L_{in}. Let us denote a set of all texts from the full corpus Cf that are relevant to text z as Dz ($\text{Dz} \subseteq \text{Cf}$). We will call the set Dz as *the dynamic text corpus*.

2.3. Informativity of index terms

Informativity of index terms (for example, keywords, document's terms or semantic concepts) can be evaluated when using the results of statistical processing of thematic or dynamic text corpora and the full text corpus. Let us consider an index terms evaluation process on the basis of thematic text corpora taking into account that it is similar to dynamic corpora.

Let us examine the following population of events:

S_{Ct} – the index term α is randomly taken from the thematic text corpus Ct ($Ct \in Cf$);

S_{Cf} – the index term α is taken from the full corpus of texts Cf.

Let H_{Ct} be the hypothesis of the thematic text corpus Ct occurrence.

In our model we assume that each index term α_i is conditionally independent to every other term α_j for $j \neq i$.

Let $P(S_{Ct} / S_{Cf})$ be a conditional chance that the index term α is taken from the thematic text corpus Ct on conditions that it has already been taken from the full corpus Cf. This conditional chance equals

$$P(S_{Ct} / S_{Cf}) = \frac{P(S_{Ct}.S_{Cf})}{P(S_{Cf})} = \frac{P(S_{Ct}).P(S_{Cf} / S_{Ct})}{P(S_{Cf})}.$$

We will call the conditional chance as $P(S_{Ct} / S_{Cf})$ *informativity* of the index term α in the thematic text corpus Ct.

Conditional chance $P(S_{Cf} / S_{Ct}) = 1$, because Ct is a subset of the set Cf. Then we will have

$$P(S_{Ct} / S_{Cf}) = \frac{P(S_{Ct})}{P(S_{Cf})}.$$

After using formula of total probability evaluation for $P(S_{Ct})$ we will have

$$P(S_{Ct} / S_{Cf}) = \frac{P(S_{Ct} / H_{Ct})}{P(S_{Cf})}.P(H_{Ct}).$$

When volumes of the full corpus Cf and the thematic corpus (or text document) Ct are big enough it is possible to consider

$$P(S_{Ct} / H_{Ct}) \approx \frac{n_{Ct}}{N_{Ct}},\ P(S_{Cf}) \approx \frac{n_{Cf}}{N_{Cf}},\ P(H_{Ct}) \approx \frac{n_{Ct}}{N_{Cf}},$$

where n_{Ct}, n_{Cf} are absolute frequencies of the index term α occurrence in thematic and full corpora, and N_{Ct}, N_{Cf} are a quantity of all α occurrences in Ct and Cf, respectively. Then the formula for evaluation of informativity I_{Ct}^{α} of index term α in the thematic text corpus Ct will take on form of:

$$I_{Ct}^{\alpha} = \frac{n_{Ct}}{n_{Cf}}. \quad (1)$$

Unlike existing methods of informativity evaluation [12, 15, 18], the formula (1) uses statistical characteristics not only of text documents but of knowledge domain expressed by Ct and Cf corpora as well.

3. The model of indexing text information

Let T_{int} be a certain nonempty set of L_{int} internal language texts. Formally, the process of indexing any L_{in} input language $t_{in} \in T_{in}$ text is in generating a $t_{int} \in T_{int}$ text which is an image of the t_{in} text at certain injective mapping $\omega : T_{in} \to T_{int}$ .The $t_{int} = \omega(t_{in})$

text is called as *the search profile* of a t_{in} text. If the t_{in} text is a query, then the $t_{\text{int}} = \omega(t_{\text{in}})$ text is a *search prescription* appropriate to the $t_{\text{in}} \in Z_{\text{in}}$ query.

3.1. Indexing full-text documents and text corpora

Suppose, as before, α is an index term, and β_i, i=1, …, l, are index terms that are word forms and synonyms of the term α. Let us denote absolute occurrence frequencies of index term β_i in the full-text document $t \in T_{\text{in}}$ and the full corpus of texts Cf as n_t^i, n^i_{Cf} Then formula (1) for evaluation of index term informativity will take on form of:

$$I^{\alpha} = \frac{n_t + \sum_{i=1}^{l} n_t^i}{n_{\text{Cf}} + \sum_{i=1}^{l} n_{\text{Cf}}^i}. \tag{2}$$

Using that formula we can present the $O_t = \omega(t)$ document search profile as

$$O_t = \{(a, I_a) \mid a \in \text{Pr}_{\omega(t)},\ 0 \le I_a \le 1\}, \tag{3}$$

where I_a is the informativity of the index term a; $\text{Pr}_{\omega(t)}$ is the set of all index terms of document search profile $O_t = \omega(t)$.

Similarly, according to (2) one can index a thematic (or dynamic) text corpus, considering it a full-text document.

3.2. Indexing short messages and user queries

Suppose, $t \in T_{\text{in}}$ is a short message (or a user query), i.e., $t \in T_{\text{in}}$ is a text which volume is small and does not allow identifying statistical characteristics of its index terms.

Then while indexing the message t we will use the relevant to its thematic (or dynamic) text corpus Ct, i.e., $\mu(t, \text{Ct}) = 1$. Formula (2) for evaluation of index term α informativity will take on form of

$$I^{\alpha} = \frac{n_{\text{Ct}} + \sum_{i=1}^{l} n_{\text{Ct}}^i}{n_{\text{Cf}} + \sum_{i=1}^{l} n_{\text{Cf}}^i}. \tag{4}$$

The experiments showed that this indexing technique can be successfully applied to the short text (for example, Internet pages) or user query which length is one to three keywords (see Section 7 for details). Using (4) for indexing full-text documents and text corpora doesn't improve the performance of information retrieval system.

3.3. The vector space representation of document search profiles

Let us represent the document search profile O_t in a vector form in the following way.

We consider the n-dimensional Euclidean space. For this purpose, let us lexicographically order all the index terms of the set Pr, i.e., compose the tuple

Pr = $\langle a_1, a_2, \ldots, a_n \rangle$. Now we can construct the vector in the n-dimensional Euclidean space for an indexed text document as follows:

(5) $$\mathbf{O}_t = (I_{a_1}, I_{a_2}, \ldots, I_{a_n}).$$

The coordinates of vector $\mathbf{O}_t$ are appropriate values of index terms informativity. Similarly, we can represent a search prescription.

4. The model of text information retrieval

The goal of information retrieval is to find all text documents which are relevant to a user query [6]. Let us define the formal notions related to the implementation process of information retrieval.

4.1. Retrieval functions

The calculation of information retrieval efficiency is based on the criteria which essentially differ from each other. In some cases, the relevance estimations are used, and in others – the pertinence estimations which indicate the semantic closeness of a text to information user needs. The concept of relevance is formally defined in Section 2.2. Let us define the concept of pertinence.

Suppose, as before, T_{in} is the certain nonempty set of input language L_{in} texts and Z_{in} is the set of queries. Let us define bijective mapping θ: $Z_{\text{in}} \rightarrow$ IP that assigns the biunique correspondence between the set of the users queries Z_{in} and the IP set of their information needs. Then we will introduce a formal notion of pertinent texts as follows.

Definition 2. We will call any text $t \in T_{\text{in}}$ as *pertinent* to information needs $\theta(z)$ if there exists the surjective mapping $\nu : T_{\text{in}} \times \theta(Z_{\text{in}}) \rightarrow \{0, 1\}$ for which the following relation holds $\nu(t, \theta(z)) = 1$ and *non-pertinent* if $\nu(t, \theta(z)) = 0$.

In theory of information retrieval, we distinguish quantitative and logical retrieval criteria [1]. The quantitative criterion provides a numerical calculation of the semantic proximity extent of a document and a query. Logical criteria are based on using logical operators of conjunction, disjunction and negation. The document is supposed to have been found if the logical formula evaluates as "true". If this formula evaluates as "false" the text is irrelevant to a user query. Let us formally define the retrieval criterion concept.

We will call the mapping $\eta : \omega(T_{\text{in}}) \times \omega(Z_{\text{in}}) \rightarrow R$ of the Cartesian product of the sets of search profiles and search prescription to the set R of real numbers as *the retrieval criterion*.

We can model one step of text retrieval as the partial multimapping $\pi : Z_{\text{in}} \rightarrow T_{\text{in}}$ of the set of queries in the set of texts.

We will call the partial multimapping π as *the retrieval function*, if for any query $z \in Z_{\text{in}}$, the set $\pi(z)$ includes the texts $t \in T_{\text{in}}$ for which the retrieval criterion is not less than a certain η_0, i.e., $\eta(\omega(t), \omega(z)) \geq \eta_0$.

We will call the tuple $\langle \pi_1(z_1), \pi_2(z_2), \ldots, \pi_l(z_l) \rangle$ of retrieval functions as *the search strategy*. Every subsequent element of the tuple differs from the previous query and/or the retrieval function.

4.2. The performance criteria of information retrieval

The main measures for evaluating the performance of information retrieval systems are recall and precision. If the values of recall and precision are determined on the basis of relevance then we will call them the recall and precision by relevance. If these values are evaluated on the basis of pertinence then we will denote them as the recall and precision by pertinence. Let us define the formal concepts of performance criteria using the above notations.

Definition 3. Let $z \in Z_{\text{in}}$ be any query. Then we call the expressions

$$R_{\text{rel}} = \frac{\sum_{t \in \pi(z)} \mu(t, z)}{\sum_{t \in T_{\text{in}}} \mu(t, z)}, \quad P_{\text{rel}} = \frac{\sum_{t \in \pi(z)} \mu(t, z)}{|\pi(z)|},$$

as *the recall* and *precision by relevance* with ($|\pi(z)|$ being a power of set $\pi(z)$).

The process of retrieval optimization in terms of criteria R_{rel} and P_{rel} turns into finding the maximum of these expressions. Since the value of recall R_{rel} for any $z \in Z_{\text{in}}$ depends on the random variables $\sum_{t \in \pi(z)} \mu(t, z)$ and $\sum_{t \in T_{\text{in}}} \mu(t, z)$, and the precision P_{rel} – on the random variable $\sum_{t \in \pi(z)} \mu(t, z)$ then as the performance criteria we will use the average recall and precision:

$$R_{\text{aver. rel}} = \frac{M[\sum_{t \in \pi(z)} \mu(t, z) = 1]}{M[\sum_{t \in T_{\text{in}}} \mu(t, z) = 1]}, \quad P_{\text{aver. rel}} = \frac{M[\sum_{t \in \pi(z)} \mu(t, z) = 1]}{|\pi(z)|},$$

where M is the mathematical expectation.

Since $M[\sum_{t} \mu(t, z) = 1] = \sum_{t} P(\mu(t, z) = 1)$ then the formulas for the average recall and precision by relevance will take on form of:

$$R_{\text{aver. rel}} = \frac{\sum_{t \in \pi(z)} P(\mu(t, z) = 1)}{\sum_{t \in T_{\text{in}}} P(\mu(t, z) = 1)}, \quad P_{\text{aver. rel}} = \frac{\sum_{t \in \pi(z)} P(\mu(t, z) = 1)}{|\pi(z)|}.$$

We denote the recall and precision by pertinence in the following way.

Suppose, as before, $z \in Z_{\text{in}}$ is any query, and $\theta(z)$ is an information need which is appropriate to a query. Then we will call the expressions

$$R_{\text{pert}} = \frac{\sum_{t \in \pi(z)} \nu(t, \theta(z))}{\sum_{t \in T_{\text{in}}} \nu(t, \theta(z))}, \quad P_{\text{pert}} = \frac{\sum_{t \in \pi(z)} \nu(t, \theta(z))}{|\pi(z)|},$$

recall and *precision by pertinence*.

The average recall and precision by pertinence will take on form of:

$$R_{\text{aver. pert}} = \frac{\sum_{t\in\pi(z)} P\big(\nu(t, \theta(z)) = 1\big)}{\sum_{t\in T_{\text{in}}} P\big(\nu(t, \theta(z)) = 1\big)}, \; P_{\text{aver. pert}} = \frac{\sum_{t\in\pi(z)} P\big(\nu(t, \theta(z)) = 1\big)}{|\,\pi(z)\,|}.$$

4.3. The average information retrieval loss by relevance and pertinence

Information retrieval in the set of texts T_{in} is equivalent to a set partition into two classes: category 1 of relevant and category 2 of irrelevant documents.

Let $P(t\,|1) = P(\mu(t, z) = 1\,|\,t \in T_{\text{in}})$ be the conditional probability distribution of relevance of texts of category 1, and $P(t\,|2) = P(\mu(t, z) = 0\,|\,t \in T_{\text{in}})$ be the conditional probability distribution of irrelevance of texts of category 2. Let us denote the a priori probability of occurrence of texts of category 1 as p_1, and as p_2 – the a priori probability of occurrence of texts of category 2.

Let us consider the loss matrix $\|\varphi_{mn}\|$, where φ_{mn} ($m, n = 1, 2$) are the losses that arise at assigning m class texts to the class n. It is natural in information retrieval to consider that the loss matrix is

$$\|\varphi_{mn}\| = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Then the mathematical loss expectation that occurs when relevant texts that belong to class 1 appear in class 2 is found from the expression

$$\varphi_1 = p_1 \sum_{t\in T_{\text{in}}\setminus\pi(z)} P(t\,|1).$$

The mathematical loss expectation that occurs when irrelevant texts that belong to class 2 appear in class 1 is found from the expression

$$\varphi_2 = p_2 \sum_{t\in\pi(z)} P(t\,|2).$$

The average overall loss (the overall mathematical expectation of losses) will take on form of

$$\varphi = p_1 \sum_{t\in T_{\text{in}}\setminus\pi(z)} P(t\,|1) + p_2 \sum_{t\in\pi(z)} P(t\,|2). \tag{6}$$

Information retrieval will be optimal by relevance when the average overall loss is minimal. Thus, the criterion φ can be used as the optimality search criterion. We will call it as *the average retrieval loss by relevance*. In case of need to specify a retrieval function in retrieval criteria we will place its symbol in the upper position of the index, for example φ^{π}, $R^{\pi}_{\text{aver. rel}}$, $P^{\pi}_{\text{aver. rel}}$.

Let us prove that the criterion φ is equivalent to each of the criteria $R_{\text{aver. rel}}$ and $P_{\text{aver. rel}}$.

Statement 1. Let $z \in Z_{\text{in}}$ be any query, and π_1 and π_2 be some retrieval functions such that $|\pi_1(z)| = |\pi_2(z)|$. Then $\varphi^{\pi_1} > \varphi^{\pi_2}$ if and only if $R^{\pi_1}_{\text{aver. rel}} < R^{\pi_2}_{\text{aver. rel}}$ or $P^{\pi_1}_{\text{aver. rel}} < P^{\pi_2}_{\text{aver. rel}}$.

The proof of the Statement 1 is given in Appendix 1.

By analogy with the average retrieval loss by relevance (6) we will consider the average retrieval loss by pertinence

$$\psi = q_1 \sum_{t \in T_{\text{in}} \setminus \pi(z)} Q(t \mid 1) + q_2 \sum_{t \in \pi(z)} Q(t \mid 2),$$

where $Q(t|1) = P(\nu(t, \theta(z)) = 1 \mid t \in T_{\text{in}})$ is the conditional probability distribution of pertinence of texts of category 1; $Q(t|2) = P(\nu(t, \theta(z)) = 0 \mid t \in T_{\text{in}})$ is the conditional probability distribution of non-pertinence of texts of category 2; q_1 is the a priori probability of occurrence of texts of category 1; q_2 is the a priori probability of occurrence of texts of category 2.

For the criterion ψ, the statement similar to statement 1is true.

Statement 2. Let ($z \in Z_{\text{in}}$) be any information need, and π_1 and π_2 be some retrieval functions such that $|\pi_1(z)| = |\pi_2(z)|$. Then $\psi^{\pi_1} > \psi^{\pi_2}$ if and only if $R^{\pi_1}_{\text{aver.pert}} < R^{\pi_2}_{\text{aver.pert}}$ or $P^{\pi_1}_{\text{aver.pert}} < P^{\pi_2}_{\text{aver.pert}}$.

On the basis of Statements 1 and 2 let us define the concept of optimal retrieval functions in terms of average recall and precision by relevance and pertinence.

4.4. The optimal retrieval functions

Information retrieval in the set T_{in} will be effective if the average retrieval loss is minimal. From the theory of statistical decisions, the partition of the set T_{in} into two classes is known to provide a minimum of criterion φ (or ψ), if class 1 of relevant (or pertinent) texts includes the texts $t \in T_{\text{in}}$ for which $p_1P(t|1) \geq p_2P(t|2)$ (or $q_1Q(t|1) \geq q_2Q(t|2)$).

Taking into account that $p_1P(t|1) \geq p_2P(t|2)$, let us define the retrieval criterion η: $\omega(T_{\text{in}}) \times \omega(Z_{\text{in}}) \to R$ in the following way: for any text $t \in T_{\text{in}}$ and any query $z \in Z_{\text{in}}$

$$\eta(\omega(t), \omega(z)) = p_2P(\mu(t, z) = 0 \mid t \in T_{\text{in}}) - p_1P(\mu(t, z) = 1 \mid t \in T_{\text{in}}).$$

The corresponding retrieval function will take on form of

$$\pi(z) = \{t \mid p_2P(\mu(t, z) = 0 \mid t \in T_{\text{in}}) - p_1P(\mu(t, z) = 1 \mid t \in T_{\text{in}}) < 0,\ t \in T_{\text{in}}\}. \quad (7)$$

Let us prove the optimality by relevance of this retrieval function assuming as before that the loss matrix is $\|\varphi_{mn}\| = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$.

Statement 3. If the retrieval function π_1 is defined by equation (7), and π_2 is any other different from π_1 retrieval function and $|\pi_1(z)| = |\pi_2(z)|$, then $R^{\pi_1}_{\text{aver.rel}} > R^{\pi_2}_{\text{aver.rel}}$ and $P^{\pi_1}_{\text{aver.rel}} > P^{\pi_2}_{\text{aver.rel}}$.

The proof of the Statement 3 is given in Appendix 2.

According to Statement 3 on the basis of retrieval function $\pi_1(z)$, the optimal search strategies in terms of the average recall and precision by relevance and pertinence can be implemented.

Taking into account that class 1 includes the documents $t \in T_{in}$ for which $q_1 Q(t|1) \geq q_2 Q(t|2)$, we define the retrieval criterion $\eta : \omega(T_{in}) \times \omega(Z_{in}) \to R$ in the following way:

$$\eta(\omega(t), \omega(z)) = q_2 P(\nu(t, \theta(z)) = 0 \,|\, t \in T_{in}) - q_1 P(\nu(t, \theta(z)) = 1 \,|\, t \in T_{in}),$$

where $z \in Z_{in}$ is any query; $\theta(z)$ is an information need which is appropriate to a query.

The retrieval function π that is based on the retrieval criterion $\eta(\omega(t), \omega(z))$ will take on form of

$$\pi(z) = \{t \,|\, q_2 P(\nu(t, \theta(z)) = 0 \,|\, t \in T_{in}) - q_1 P(\nu(t, \theta(z)) = 1 \,|\, t \in T_{in}) < 0, t \in T_{in}\}. \quad (8)$$

For the retrieval function $\pi(z)$, the statement which is similar to statement 3 is true.

Statement 4. If the retrieval function π_1 is given by the Equation (8), and π_2 is any other different from π_1 retrieval function, and $|\pi_1(z)| = |\pi_2(z)|$, then $R^{\pi_1}_{\text{aver. pert}} > R^{\pi_2}_{\text{aver. pert}}$ and $P^{\pi_1}_{\text{aver. pert}} > P^{\pi_2}_{\text{aver. pert}}$.

The proof of Statement 4 is similar to the proof of Statement 3.

The user query is considered to be correctly formulated if all relevant texts are simultaneously pertinent. Let us define the formal conception of a correctly formulated query.

We will call the query $z \in Z_{in}$ as *correctly formulated* if $\mu(t, z) = 1$ if and only if $\nu(t, \theta(z)) = 1$ for any text $t \in T_{in}$.

It can be shown that for a correctly formulated query the optimal retrieval function by relevance is optimal by pertinence too.

Statement 5. Let $z \in Z_{in}$ be any correctly formulated query. If the retrieval function π_1 is given by the Equation (10), and π_2 is any other different from π_1 retrieval function and $|\pi_1(z)| = |\pi_2(z)|$, then $R^{\pi_1}_{\text{aver. pert}} > R^{\pi_2}_{\text{aver. pert}}$ and $P^{\pi_1}_{\text{aver. pert}} > P^{\pi_2}_{\text{aver. pert}}$.

The proof of Statement 5 is similar to the proof of Statement 3 if we take into account the correctness of the query z.

According to Statement 5 on the basis of the retrieval function (8), the optimal search strategies in terms of the average recall and precision by pertinence can be implemented.

5. The algorithm of full-text documents indexing

When indexing text documents, the following knowledge base dictionaries are used: the frequency dictionary of word forms, the synonyms dictionary and the inflectional paradigms dictionary.

Let α be any word form, P_{Cf} and P_{Ct_i}, $i = 1, \ldots, n$, be its absolute frequencies in full and i-th thematic text corpora. Then we will call the set of tuples $\langle \alpha, P_{\text{Cf}}, P_{\text{Ct}_1}, P_{\text{Ct}_2}, \ldots, P_{\text{Ct}_n} \rangle$ as *the frequency dictionary of word forms*.

The synonyms dictionary consists of synonymous word form groups that can be used in determining their informativity.

The inflectional paradigms dictionary is used to find all word forms of paradigms after finding the word form and its code in the frequency dictionary of word forms.

Let us consider the algorithm of full-text documents indexing.

Algorithm 1

Input: The text document t and the number l of word forms with nonzero informativity, i. e. the count of $\mathbf{O}_t$ vector components.

Output: The document search profile $\mathbf{O}_t$.

Step 1. $\mathbf{O}_t := \varnothing$.

Step 2. Select the next word form a from the text t, and find it in the frequency dictionary of word forms.

Step 3. Find all synonyms of the word form a in the synonyms dictionary.

Step 4. Find all word forms of a in the inflectional paradigms dictionary.

Step 5. Calculate the value of word form informativity according to formula (2).

Step 6. Put the pair (a, I_a) into the set O_t.

Step 7. If all word forms of the text t are exhausted, then go to Step 8, otherwise to Step 2.

Step 8. Construct vector $\mathbf{O}_t = (I_{a_1}, I_{a_2}, \ldots, I_{a_n})$.

Step 9. END (the search profile of the text t is formed).

6. Strategies and algorithms of information retrieval

Let us represent an original user query as the set $z_1 = ((b_1, 1), (b_2, 1), \ldots)$, where $z_1 \in Z_{in}$, and b_i, $i = 1, \ldots, n$, are index terms of z_1. In order to achieve an adequate representation of user information needs in the original query z_1, its further correction based on the appropriate dynamic corpus Dz_1. may be needed. Let us construct the optimal by relevance retrieval function which will be used to generate a dynamic text corpus.

Let us examine the t_d document search profile as the set $O_{t_d} = \{(a, I_a) \mid a \in \Pr_{\omega(t_d)}, 0 \le I_a \le 1\}$, where $\Pr_{\omega(t_d)}$ is the set of all index terms of the document search profile $O_{t_d} = \omega(t_d)$. Let $\Pr_{\omega(z_1)}$ be the set of all index terms of the original user query z_1.

When creating the dynamic text corpus Dz_1, it is natural to assume that the relevant (formally relevant) documents are those of the full corpus Cf search profiles which contain all the index terms $b_1, b_2, \ldots$ of the query z_1, i.e., $\Pr_{\omega(z_1)} \subseteq \Pr_{\omega(t_d)}$. Then in the expression (7) for optimal retrieval function by relevance:

$$\pi_{\mathrm{Cf}}(z_1) = \{t_d \mid p_2 P(\mu(t_d, z_1) = 0 \mid t_d \in \mathrm{Cf}) - p_1 P(\mu(t_d, z_1) = 1 \mid t_d \in \mathrm{Cf}) < 0,\ t_d \in \mathrm{Cf}\},$$

$p_1 = 1$, $P(\mu(t_d, z_1) = 1 \mid t_d \in \mathrm{Cf}) = 1$ and $p_2 = 0$ because all the texts that are relevant to the query z_1 (and only relevant) are located in the set Dz_1.

In this connection, the corresponding retrieval function is represented as

$$\pi_{\mathrm{Cf}} = \{t \mid \eta(\omega(t_d), \omega(z_1)) < 0,\ t_d \in \mathrm{Cf}\}, \quad (9)$$

and the retrieval criterion is defined as

$$\eta(\omega(t_d), \omega(z_1)) = \begin{cases} -1 \text{ if } \Pr_{\omega(z_1)} \subseteq \Pr_{\omega(t_d)}, \\ 0 \text{ if } \Pr_{\omega(z_1)} \not\subset \Pr_{\omega(t_d)}. \end{cases} \quad (10)$$

The retrieval function (9) and the retrieval criterion (10) should be used at the first step of information retrieval while forming a dynamic text corpus.

The procedure for correcting an original user query can be implemented as follows: documents from the dynamic text corpus $Dz_1 = \pi_{Cf}(z_1)$ are presented to a user who excludes all non-pertinent texts from the set $\pi_{Cf}(z_1)$. The obtained set (which we denote as Dz_2) is considered to be the corrected dynamic text corpus. On its basis, the corrected user query z_{Dz} is formed through indexing Dz_2. Each index term in z_{Dz} is associated with its weight (informativity). The size of a corrected user query is chosen empirically and includes unique word forms from Dz_2, informativity of which is bigger than a certain threshold value J_0.

6.1. The algorithm of information retrieval in a full corpus of texts

Let us consider the algorithm of information retrieval in the full corpus of texts Cf which implements the above mentioned procedure for the correction of an original user query.

Algorithm 2

Input: the user query $z_1 \in Z_{in}$, the threshold informativity value J_0 and the full corpus of texts Cf.

Output: the corrected user query z_{Dz}.

Step 1. Represent an original user query in the form of the set $z_1 = ((b_1, 1), (b_2, 1), \ldots)$.

Step 2. Find the set of documents $Dz_1 = \pi_{Cf}(z_1)$ in the full corpus of texts Cf according to the retrieval function (9).

Step 3. Present documents from the dynamic text corpus Dz_1 to the user.

Step 4. Form the corrected dynamic text corpus Dz_2 after excluding non-pertinent documents.

Step 5. Index the dynamic text corpus Dz_2 (Algorithm 1).

Step 6. Taking into account the threshold informativity value J_0, form the corrected user query $z_{Dz} = \{(b_j, J_{b_j}) | j = 1, \ldots, k\}$, where $b_1, b_2, \ldots, b_k$ are the index terms of corpus Dz_2, and J_{b_j} is informativity of index terms which is calculated according to the formula (2).

Step 7. END (the corrected user query z_{Dz} is formed).

6.2. The algorithm of information retrieval in the database of indexed documents

It is assumed that the search of text documents using the optimal retrieval function by pertinence will be implemented in various information sources, i.e., under the $T_{in} \subseteq L_{in}$ we will understand the set of indexed texts on the Internet or LAN, or a hard disk.

Information retrieval in the database T_{in} is implemented in two stages.

At the first stage we conduct the search using the optimal retrieval function by relevance

$$\pi = \{t \mid \eta(\omega(t), \omega(z_{Dz})) < 0,\ t \in T_{in}\}, \tag{11}$$

and the retrieval criterion

$$\eta(\omega(t), \omega(z_{Dz_1})) = \begin{cases} -1 \text{ if } \Pr_{\omega(z_{Dz_1})} \subseteq \Pr_{\omega(t)}, \\ 0 \text{ if } \Pr_{\omega(z_{Dz_1})} \not\subset \Pr_{\omega(t)}. \end{cases} \tag{12}$$

The retrieval function (11) in compliance with statement 5 is also optimal by pertinence since the corrected query z_{Dz} is obtained by changing the original user query z_1.

At the second stage the search results are ranked according to the retrieval criterion which is used in the information system (for example, the cosine of the angle between the vectors of the corrected search prescription and document search profile can be used as a measure of proximity queries and documents).

Let us consider the algorithm of information retrieval which implements the above mentioned strategy.

Algorithm 3

Input: the corrected user query z_{Dz} obtained according to Algorithm 2.

Output: the tuple of texts $\langle t_1, t_2, \ldots \rangle$ ranked according to the retrieval criterion which is used in the information system.

Step 1. Find texts using search prescription z_{Dz} in accordance with the retrieval function (11) and the retrieval criterion (12) in the set T_{in}.

Step 2. Rank all retrieval texts according to the retrieval criterion which is used in the information system.

Step 3. END.

7. Experimental results

To evaluate the effectiveness of the proposed strategies and algorithms some experiments on comparing the developed approach with retrieval methods based on the vector space model were carried out. The term weights in vector space model were defined according to the known formulas TF-IDF and TF-IDF with normalizing a document length [15, 18]. The collection of ROMIP regulatory documents, 2007 was used for evaluating (Legal Documents Collection 2007 of Russian Information Retrieval Evaluation Seminar). It consists of more than 300 000 texts [4]. As the full text corpus Cf we used the collection of more than 10 000 documents on different subjects.

The 11 pt precision-recall curves obtained for the proposed method and methods based on the vector space model are shown in Fig. 1.

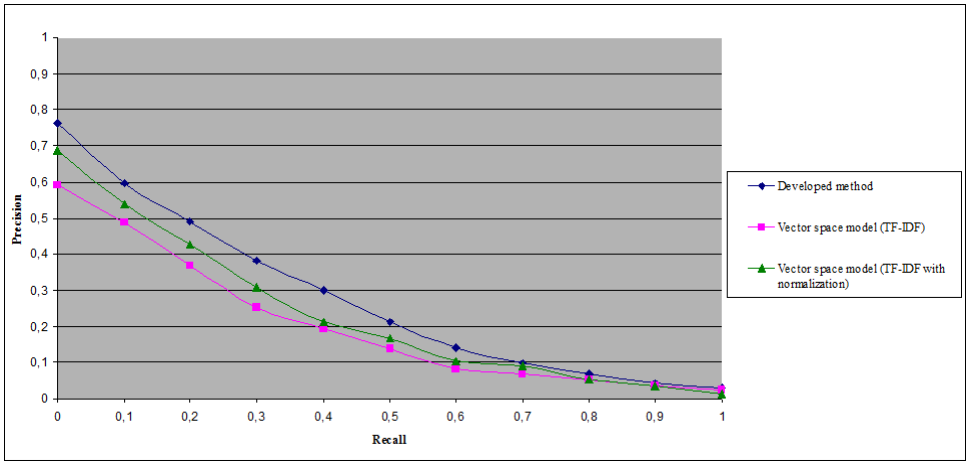

Fig. 1. The 11 pt precision-recall curves obtained for the proposed method and methods based on the vector space model

As follows from the curves in Fig. 1, the proposed method shows better results in the terms of precision than the vector space model. At the same time there is a significant difference in the starting levels of recall. This is important for users of the information system because it demonstrates that most of the relevant documents are at the beginning of the search results. Using the procedure for correcting an original user query on short queries (one to three keywords) increased the average recall by 18 percent. In experiments, the highest quality search results were found to be for threshold informativity values J_0 in the range from 0.4 up to 0.5.

8. Conclusion

The models and algorithms of information retrieval presented in this paper can be used in various systems designed for processing and analyzing texts. In search engines, using the developed method allows increasing the recall and precision by correcting an original user query by a relevant text corpus and by consequent dynamically changing a search space. The proposed technique of calculating the informativity of index terms can be used in automatic summarization systems for detecting informative word forms in documents and for synthesizing connected summaries. With an appropriate selection of subjects and the hierarchical structure of a text corpus, it is possible to search regarding documents stylistic color (e.g., journalism, popular or scientific literature).

References

1. Berry, M., M. Browne. Understanding Search Engines: Mathematical Modeling and Text Retrieval. Society for Industrial and Applied Mathematics, 2005.
2. Brusilovsky, P., C. Tasso. Preface to Special Issue on User Modeling for Web Information Retrieval. – User Modeling and User-Adapted Interaction, Vol. **14**, 2004, No 2-3, pp. 147-157.

3. C u m m i n s, R., C. O'R i o r d a n. Evolving Local and Global Weighting Schemes in Information Retrieval. – Information Retrieval, Vol. **9**, 2006, No 3, pp. 311-330.
4. D o b r o v, B., I. K u r a l e n o k, N. L o u k a c h e v i t c h, I. N e k r e s t y a n o v, I. S e g a l o v i c h. Russian Information Retrieval Evaluation Seminar. – In: Proc. of 4th International Conference on Language Resources and Evaluation, 2004, pp. 1359-1362.
5. G r e e n b e r g, J. User Comprehension and Searching with Information Retrieval Thesauri. – Cataloging & Classification Quarterly, Vol. **37**, 2004, No 3, pp. 103-120.
6. H e n z i n g e r, M. Link Analysis in Web Information Retrieval. – IEEE Data Engineering Bulletin, Vol. **23**, 2000, No 3, pp. 3-8.
7. J a c k s o n, P., I. M o u l i n i e r. Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization. John Benjamins Publishing, 2002.
8. K a p t e i n, R., J. K a m p s. Improving Information Access by Relevance and Topical Feedback. – In: Proc. of 2nd International Workshop on Adaptive Information Retrieval, 2008, pp. 58-64.
9. K u m a r, C. A., M. R a d v a n s k y, J. A n n a p u r n a. Analysis of a Vector Space Model, Latent Semantic Indexing and Formal Concept Analysis for Information Retrieval. – Cybernetics and Information Technologies, Vol. **12**, 2012, No 1, pp. 34-48.
10. L a n g v i l l e, A. M., C. D. M e y e r. Information Retrieval and Web Search. Handbook of Linear Algebra. CRC Press, 2006.
11. L i u, T. Learning to Rank for Information Retrieval. Springer, 2011.
12. L v, Y., C. Z h a i. Adaptive Term Frequency Normalization for BM25. – In: Proc. of 20th ACM International Conference on Information and Knowledge Management (CIKM'2011), New York, USA, 2011, pp. 1985-1988.
13. M a n n i n g, C., P. R a g h a v a n, H. S c h ü t z e. Introduction to Information Retrieval. Cambridge University Press, 2008.
14. Q i u, F., J. C h o. Automatic Identification of User Interest for Personalized Search. – In: Proc. of 15th International Conference on World Wide Web, 2006, pp. 727-736.
15. R a m o s, J. Using TF-IDF to Determine Word Relevance in Document Queries. – In: Proc. of 1st International Conference on Machine Learning, New Brunswick: NJ, USA, 2003.
16. R u t h v e n, I., M. L a l m a s. A Survey on the Use of Relevance Feedback for Information Access Systems. – The Knowledge Engineering Review, Vol. **18**, 2003, No 2, pp. 95-145.
17. S i n g h, J., S. D w i v e d i. Analysis of Vector Space Model in Information Retrieval. – In: Proc. of IJCA National Conference on Communication Technologies & its Impact on Next Generation Computing 2012, Vol. **2**, 2012, pp. 14-18.
18. S o u c y, P., G. W. M i n e a u. Beyond TF-IDF Weighting for Text Categorization in the Vector Space Model. – In: Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI 2005), 2005, pp. 1130-1135.

Appendix 1

Statement 1. Let $z \in Z_{\text{in}}$ be any query, and π_1 and π_2 be some retrieval functions such that $|\pi_1(z)| = |\pi_2(z)|$. Then $\varphi^{\pi_1} > \varphi^{\pi_2}$ if and only if $R^{\pi_1}_{\text{aver. rel}} < R^{\pi_2}_{\text{aver. rel}}$ or $P^{\pi_1}_{\text{aver. rel}} < P^{\pi_2}_{\text{aver. rel}}$.

Proof. *Necessity*. Let us prove that from $\varphi^{\pi_1} > \varphi^{\pi_2}$ follows $R^{\pi_1}_{\text{aver.rel}} < R^{\pi_2}_{\text{aver.rel}}$ and $T^{\pi_1}_{\text{aver.rel}} < T^{\pi_2}_{\text{aver.rel}}$. It suffices to show that $M[\sum_{t\in\pi_2(z)}\mu(t,z)=1] - M[\sum_{t\in\pi_1(z)}\mu(t,z)=1] > 0$. Let us calculate $\varphi^{\pi_1} - \varphi^{\pi_2}$:

$$\varphi^{\pi_1} - \varphi^{\pi_2} = p_1 \sum_{t\in T_{\text{in}}\setminus\pi_1(z)} P(t\,|\,1) + p_2 \sum_{t\in\pi_1(z)} P(t\,|\,2) - p_1 \sum_{t\in T_{\text{in}}\setminus\pi_2(z)} P(t\,|\,1) - p_2 \sum_{t\in\pi_2(z)} P(t\,|\,2) =$$

$$= p_1 \sum_{t\in T_{\text{in}}} P(t\,|\,1) - p_1 \sum_{t\in\pi_1(z)} P(t\,|\,1) + p_2 \sum_{t\in\pi_1(z)} P(t\,|\,2) - p_1 \sum_{t\in T_{\text{in}}} P(t\,|\,1) + p_1 \sum_{t\in\pi_2(z)} P(t\,|\,1) -$$

$$- p_2 \sum_{t\in\pi_2(z)} P(t\,|\,2) = -p_1 \sum_{t\in\pi_1(z)} P(t\,|\,1) + p_2 \sum_{t\in\pi_1(z)} P(t\,|\,2) + p_1 \sum_{t\in\pi_2(z)} P(t\,|\,1) - p_2 \sum_{t\in\pi_2(z)} P(t\,|\,2) =$$

$$= 2p_1 \sum_{t\in\pi_2(z)} P(t\,|\,1) - 2p_1 \sum_{t\in\pi_1(z)} P(t\,|\,1) - (\, p_2 \sum_{t\in\pi_2(z)} P(t\,|\,2) + p_1 \sum_{t\in\pi_2(z)} P(t\,|\,1)) +$$

$$+ (p_1 \sum_{t\in\pi_1(z)} P(t\,|\,1) + p_2 \sum_{t\in\pi_1(z)} P(t\,|\,2)).$$

After simple transformations taking into account the equalities

$$\sum_{t\in\pi_1(z)} (p_1 P(t\,|\,1) + p_2 P(t\,|\,2)) = |\,\pi_1(z)\,|, \quad \sum_{t\in\pi_2(z)} (p_1 P(t\,|\,1) + p_2 P(t\,|\,2)) = |\,\pi_2(z)\,|$$

we obtain the following expressions:

$$2p_1 \sum_{t\in\pi_2(z)} P(t\,|\,1) - 2p_1 \sum_{t\in\pi_1(z)} P(t\,|\,1) + |\,\pi_1(z)\,| - |\,\pi_2(z)\,| > 0, \quad \sum_{t\in\pi_2(z)} P(t\,|\,1) - \sum_{t\in\pi_1(z)} P(t\,|\,1) > 0,$$

whence it follows $R^{\pi_1}_{\text{aver.rel}} < R^{\pi_2}_{\text{aver.rel}}$ and $P^{\pi_1}_{\text{aver.rel}} < P^{\pi_2}_{\text{aver.rel}}$.

Sufficiency. Let us prove that the inequality $\varphi^{\pi_1} > \varphi^{\pi_2}$ results from $R^{\pi_1}_{\text{aver.rel}} < R^{\pi_2}_{\text{aver.rel}}$ and $P^{\pi_1}_{\text{aver.rel}} < P^{\pi_2}_{\text{aver.rel}}$. Let us represent the inequalities $R^{\pi_1}_{\text{aver.rel}} < R^{\pi_2}_{\text{aver.rel}}$ and $P^{\pi_1}_{\text{aver.rel}} < P^{\pi_2}_{\text{aver.rel}}$ in the form of

$$\frac{\sum_{t\in\pi_1(z)} P(\mu(t,z)=1)}{\sum_{t\in T_{\text{in}}} P(\mu(t,z)=1)} < \frac{\sum_{t\in\pi_2(z)} P(\mu(t,z)=1)}{\sum_{t\in T_{\text{in}}} P(\mu(t,z)=1)}, \quad \frac{\sum_{t\in\pi_1(z)} P(\mu(t,z)=1)}{|\,\pi_1(z)\,|} < \frac{\sum_{t\in\pi_2(z)} P(\mu(t,z)=1)}{|\,\pi_2(z)\,|}.$$

Let us multiply both sides of each inequality by $\sum_{t\in T_{\text{in}}} P(\mu(t,z)=1)$ and $|\pi_1(z)| = |\pi_2(z)|$, respectively. In both cases, we obtain

$$\sum_{t\in\pi_1(z)} P(t\,|\,1) < \sum_{t\in\pi_2(z)} P(t\,|\,1).$$

Repeating the necessity proof in the reverse order, we obtain the inequality $\varphi^{\pi_1} > \varphi^{\pi_2}$.

Statement 1 is proved.

Appendix 2

Statement 3. If the retrieval function π_1 is defined by equation (7), and π_2 is any other different from π_1 retrieval function and $|\pi_1(z)| = |\pi_2(z)|$, then $R^{\pi_1}_{\text{aver. rel}} > R^{\pi_2}_{\text{aver. rel}}$ and $P^{\pi_1}_{\text{aver. rel}} > P^{\pi_2}_{\text{aver. rel}}$.

Proof. At first let us prove that the retrieval function π_1 provides a minimum of the average retrieval loss by relevance (6). Let us calculate the difference $\varphi^{\pi_2} - \varphi^{\pi_1}$:

$$\varphi^{\pi_2} - \varphi^{\pi_1} = p_1 \sum_{t \in T_{\text{in}} \setminus \pi_2(z)} P(t \mid 1) + p_2 \sum_{t \in \pi_2(z)} P(t \mid 2) - p_1 \sum_{t \in T_{\text{in}} \setminus \pi_1(z)} P(t \mid 1) - p_2 \sum_{t \in \pi_1(z)} P(t \mid 2) =$$

$$= (p_2 \sum_{t \in \pi_2(z)} P(t \mid 2) - p_1 \sum_{t \in T_{\text{in}} \setminus \pi_1(z)} P(t \mid 1)) + (p_1 \sum_{t \in T_{\text{in}} \setminus \pi_2(z)} P(t \mid 1) - p_2 \sum_{t \in \pi_1(z)} P(t \mid 2)).$$

Under the choice of the retrieval function $\pi(z)$ for any $t \in \pi_2(z) \setminus \pi_1(z)$, it is true that $p_2P(t \mid 2) - p_1P(t \mid 1) > 0$, and for any $t \in \pi_1(z) \setminus \pi_2(z)$, it is true that $p_1P(t \mid 1) - p_2P(t \mid 2) > 0$. This implies that $\varphi^{\pi_2} - \varphi^{\pi_1} > 0$, i.e., the retrieval function $\pi_1(z)$ provides a minimum of the average retrieval loss by relevance. Taking an advantage of Statement 1, we see that $\pi_1(z)$ also provides a maximum of the average recall and precision by relevance.

Statement 3 is proved.