THE MATHEMATICAL MODELS OF SEARCHING DATA IN THE CENTERS OF INFORMATION RESOURCES

Bahodir Boltayevich Muminov[1]

[1]*Uzbekistan, TUIT, Senior Researcher*

Abstract

Modification mathematical models of searching data, which we can use to find the data of the centers of information resource, is given in this article. Also, this scientifically work is devoted to be observed the peculiarities of basic mathematical models which belong to the system of searching. The main component parts of these mathematical models are founded and they are called the technology of FSV. The model of searching data is divided three main directions and the model of based selections: the vector model, possible model and indication model have been observed. The integrated version for these direction models' CIR are given. The model of PageRank which it is used widely among web branch is also given and its main essence is exposed. New modification model is invited for graduating the document of the centers of information resource.

Keywords: *Data, Searching Data, The Models of Searching, The System of Information Resource, Document, Logical Searching, Boolean Operators, The Model of Vector Searching, Mathematic Statistic, Term, Query, Semantic Model, Fuzzy Model, Relations.*

----------***----------

1. INTRODUCTION

High roomy of the centers of information resources (IRC) which has much peculiarity electronically information resource (EIR) demands to increase the high qualities of searching necessary data. The solution of this problem dependent on getting complicated by changeable of resources and increasing its quantities.

Nowadays, main resources of much peculiarity EIR are electron and we select them electron resource. Some basic steps must be done for searching modules of these systems work fruitfully. We count important ones the followings:

- Treating searching query;
- Coordination to query of the result of searching;
- The possibility of correctness and cleverly of founded documents.

High power and famous searching modules, such as Google, Yahoo, Bing, Яndex, Rambler encompass billion web-sites. These systems differentiate each other with its special algorithms help working quality and quickly. But all of these algorithms, main approaches - are modification of searching models [1].

Model searching data - it is ordinary of real, mathematic formula and the rules of applying it for documents are created on based it. These formulas and rules are correspond to the system and the query of searching documents and it gives an opportunities to settle how graduate the founded selection of documents. There are main three theories based on traditional and modern methods.

The First Theory

This theory is based on the theory of selection. Its kinds forms divided into following variants:

- Boolean models (logical model);
- Enlarged Boolean model;

The Second Theory

This theory is based on vector algebra. These theories divided into following kinds:

- Vector model;
- General vector model;
- Semantic and neural network models;

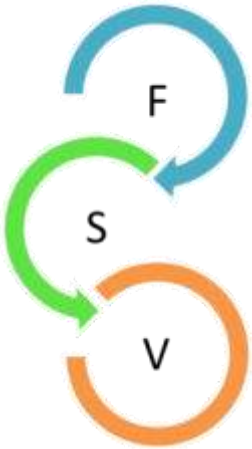
The Third Theory

This theory is based on such as the group of experts, machine learning, intellectual analysis, the theory of fuzzy sets and fuzzy logic, k-valued logic and they divided in following kinds:

- Probable models;
- Fuzzy models;
- calculating semantically relation models.

Classical models of searching information use the selection of presenting documents keywords in different environment and these words are called "terms". Term means ordinary word, its semantic helps to describing its means.

2. MAIN PART



Picture. FSV technology

Kinds of mathematic model of searching information module must have following parts:

1. (F)- the method of presenting requests - the method of forming of system user's necessities. (F-formation)
2. (S)- the conformity function of documents to need - the conformity level of query to founded documents (S-searching).
3. (V)- the method of presenting of documents (V-viewing).

We added these three parts and called FSV technology (Picture-1).

Let's suppose, from *T* dictionary to index *i (i=1,…., M)* of t_i term, *j* –document belong to $d^{(j)}$ – *D* selection of document, and $w^{(i,j)} \geq 0$, the weight dimensions dependent on $(t_i, d^{(j)})$ couple and for every t_i term which not entering $d^{(j)}$document, its weight equal to zero and $w^{(i,j)} = 0$.

2.1 The Modification Variant of Mathematic Model of Logical Searching

Boolean models of searching based on selection theory and mathematical logic. Documents, queries and terms can be offered as selections of keywords. Every term of this model is expressed with value of Boolean, and 0 (the term in query isn't in document) or 1 (the term in query is in document). Weight value in document of term have only two values:

$$w^{(i,j)} \in \{0,1\} \quad (1)$$

Users can express query as Boolean expression in Boolean model of searching and they can use operators[AND](∧), [OR] (∨), [NOT] (¬). . It is known, certain expression dependent on any connection operation (disjunctive normal form) of logical expression can be expressed as the disjunctive form. That is why:

$$q \equiv d_{dnf} = V_{i=1\ldots N} q_{cc}^{(i)}$$

In here q-query, $q_{cc}^{(i)}$- q_{dnf} the *i*-connection element of query. Inthis case, the nearness measure of document $d^{(j)}$ and query q in Boolean model are marked with $q - sim(d^{(j)}, q)$ and the following expression is given [2]:

$$sim(d^{(j)}, q) = \begin{cases} 1, agar \exists q_{cc}^{(i)} : (q_{cc}^{(i)} \in q_{dnf}) \wedge (\forall k, g_k(q_{cc}^{(i)}) = g_k(d^{(j)}) \\ 0\ residues\ situation \end{cases} \quad (2)$$

In here g_k-term is inversion function which corresponds to index of t_k, it is defines in following way:

If this conjunctive element's every term's inversion function corresponds to the same inversion function for $d^{(j)}$ document, entering disjunctive normal form q_{dnf} has $q_{cc}^{(i)}$ conjunctive component, $g_k(d^{(j)}) = w_k^j$,and$sim(d^{(j)}, q) = 1$, other wise, $sim(d^{(j)}, q) = 0$. If $sim(d^{(j)}, q) = 1$, the$d^{(j)}$ document corresponds to$q$ query. Otherwise, the document doesn't correspond.

One of advantages of Boolean models is the ordinary of applying. The main defects of this model are the followings:

- Searching hasn't high effective properties, there isn't context operators, there isn't the opportunity of graduates based on correspond to founded documents, because there isn't any balances of marking;
- The complication of using - every users can't use Boolean operator freely for expressing their query.

The main defect of classical Boolean model is visitor of term's weight values in searching query and the some terms depended on the same condition. Inconsequence, the result of searching can't be graduated based on conformity degree of queries. Several types of enlarged Boolean models are recommended to loss these defects and use counting advantages of Boolean model. The special generalization of Boolean operators is introduced based on fuzzy selection in these kinds of models, they enable counting the measure of conformity for expressing query of documents.

2.2 The Modification Variant of Mathematical Model Based on Vector Algebra of Searching

The vector model of searching is traditional members of models' class. The terms of documents and queries in this model are described as vectors on much measure space. Every term in documents are corresponded its weight values. This value is accurate based on statistic information about the number of appearing term in seeing documents and all documents massive. Using logical operations in query in vector model isn't considered. Scalar multiplication of query and documents is used for marking nearness of query and document.

The corresponding of $d^{(j)}$ document to q query is looked as the scalar multiplication of information vectors which expressed the weight values of terms:

$$\overline{d^j} = (w_1^{(j)}, w_2^{(j)}, \dots., w_n^{(j)}) \text{and} q = (w_1^q, w_2^q, \dots. , w_n^q)$$

In this case, the weight of some terms may be counted very different ways. In this, the best ordinary theory of using is based on using criterion$freq_i^{(j)}$ as the weight of term$w_i^{(j)}$, and also the speed of finding other documents is counted. This method is called marking the strength of discrimination speed of term (3):

$$w_i^{(j)} = freq_i^{(j)} \cdot \log\left(\frac{N}{n_i}\right), \quad (3)$$

In here, n_i term is numbers of used documentst_i, N- is the general number of documents in massive. For example, if any word meets in every document of in word massive, using it isn't profitable. Corresponding to it, in this situation$n_i = N$,

$$w_i^{(j)} = freq_i^{(j)} \cdot \log\left(\frac{N}{N}\right) = 0$$

This method of measuring weight of terms has standard marking $TF * IDF$, in here TF (Term Frequency- meeting of term) the number of repetition of terms in documents, IDF (Inverse Document Frequency- meeting of documents) shows that values of reverse proportional meetings of documents which have this term in massive.

The ordinary scalar multiplication $sim(d^j, q)$ is used in this model for defining topical nearness of documents and queries, it corresponds to cosine of corner which is in the middle of $\overline{d^j}$ and $\bar{q}$ vectors. The value of nearness measure of document $d^{(j)}$ and query q is counted by the following (4) formula:

$$\sin(d_j, q) = \frac{\overline{d^{(j)}}}{\left|\overline{d^{(j)}}\right|} \frac{\bar{q}}{\cdot|\bar{q}|} = \frac{\sum_{i=1}^n w^{(j)} w_i^q}{\sqrt{\sum_{i=1}^n (w_i^{(j)})^2} \sqrt{\sum_{i=1}^n (w_i^q)^2}} \quad (4)$$

Vector model is used more, because, it is very ordinary and it provide its effectives. Outside of it, vector-space model gives opportunity to express regime of searching document easily. Every document may be looked as query. But, vector-space model corresponds to counting high dimension massive, it isn't worthy for treating large information massive in traditional form.

2.3 The Modification Variant of Mathematical Model Based on Theory of Stochastic and Fuzzy in Searching

Theories of possible and mathematic statistic may be used as foundation of model on the basis of probability. The corresponding in this model is interesting for user and it is looked as the possibility of exiting. Primary selection which is selected by users or has corresponded documents in any ordinary theory is considered.

The exiting possibility which corresponded to next every document is counting on basis of corresponded and not corresponded parts of selection. The possibility of document's corresponding to query on bases of be divided differently in suitable and not suitable documents. The formulas of counting possibilities on bases of Bayes theory are used in this situation. We create final form by certain one possibility function on based on Bayes theory, it marks the suitable degree of possibility for every document in study selection which is called the status of searching (5):

$$SV = \sum_{t_i \in q \cap d} SV = \sum_{t_i \in q \cap d} \log \frac{rel_i (nrel - nrel_i)}{nrel_i (rel - rel_i)} \quad (5)$$

In here, rel_i - is the number of documents which has *i*indexterm, $nrel_i$ -is the number of not suitable documents, *d*- the study selection of documents, it is looked as word selection, *q*- the selection of word in query, it is known the selection of general terms in $q \cap d$ query and document.

The values of rel_i and $nrel_i$ and concerning component parts of searching status exponents is $\exp(SV_i)$for every term in query on bases of the result of expert marking of suitable query in documents of study selection. The value of searching status SV is counted by the above mentioned formula for extra documents.

The selection of extra documents may be graduated on bases of having the model of document and counting them.

The models of possibilities have some theoretical advantages, they recommends the natural method of formal describing the problem of searching information and gives the best prognosis on its suitable in present information. They doesn't spread widely, because, counting scale is short for possibility model and constant studying of the system is very necessary.

The observed models may be used in practical sphere in form of traditional. But they have some defines, the cause of it is the following, the main mean of words are marked with the selection words in it without counting connecting among terms. The mean of text isn't analyzed at all. Terms which has the high suitable of some query by theoretical, can create documents which have combination of selection without any mean.

That's why there are models which analyze mean of text, for example, the semantic model of searching. The effort of searching by meaning is express to analyze the grammar of text,use base of knowledge, thesaurus and ontologies

All of these models express to count semantic connection among some words and their groups. And also, nowadays the result of system on bases of theory is low. Hybridtheory is used much in practice and the opportunities of Boolean and vector models is added in them, often the special methods of semantic treating are added.

Viewing of mathematic model which counting means of text is completely, because modern searching system and hypertext direction structure is analyzed. This information is expressed on bases of announced resources' important. The

pointer is very important. Searching modules recommends documents which suitable for query ten thousands and even hundred thousand. Graduating gives opportunity to find borders of founded documents in theory and practice (usually users can view only founded 10-30 documents at the bottom). Given results demands to count the important of document when using any methods of analyzing of the text of documents and query. Several types of factors can be counted in finding the necessary of document. PageRank model is popular model in practice. It is found by the following (6) formula:

$$PR_\alpha = \frac{(1-d)}{N} + d\sum_{i-1}^{n}\frac{PR_i}{C_i} \quad (6)$$

In here, PR_α-is the PageRank reyting of looking document, *d*- the coefficient of lessening, *N*- the general measure of documents, PR_i- the PageRank rating of the *i*rating which is given direction for looking document, C_i- is the number of directions of the *i*document.

The possibility model which moving users among branch documents is lying on bases of counting the value of PageRank. The possibility of user's entering into define document is the degree of the necessary of document. The PageRank users sort suitable documents by this pointer, in searching model or count it in the some way ii the time of sorting. The essential advantage of PageRank is the counting value of PageRank be done without the mean of text in document, but direction button of web graph is forced to work. All of document in branch are sorted by essential before taking PageRank searching query.

The other model of graduating is used less by direction outside of PageRank. It consist of Back Rank (the changed variant of PageRank), HITS, Hilltop, and SALSA. Counted models farce to work analyzing the part of web graph or completely.

The following mathematic model is recommended for creating document's graph and graduating it in ERM document (7).

$$D_r = \frac{1}{2}\left(\frac{\sum_{i=1}^{|I(i)_{DocID}|} D_r^i}{|I(i)_{DocID}|\sum_{count\ (I(i))} I(i)} + \frac{|O(i)_{DocID}|}{\sum_{count\ (O(i))} O(i)}\right) \quad (7)$$

In here, *DocID*- is the identification order of every document in database, *I(i)*- the directions which come in documents, *O(i)*- the directions which go out document, |*|- the number of elements of selection. The searching information withFSVtechnology may be easier by graduating documents using this mathematic model.

Several models of searching are used in the same time when modern searching system is be expressing. All models of searching information divided into two groups. The first one consist of the model of analyzing text, the second one consist of counting structure of direction.

Usually, counting direction gives opportunity for marked culmination of resources, analyzing document gives opportunity for marking suitable for query.

3. CONCLUSION

The best modern method of mathematic model is more used in internet branch for the problem of searching information. And also, PageRank model (6) and its modern variant, modification methods are be using in web resources, nowadays. The logical model of searching information (2) is more used in local and corporative branch. Vector model (4) is used mainly for paid services of global branch. But real view of these given models are found is very difficult. Because, these models is settled down in hidden model. As a result of cheering investigation, we give modification variant of them. And also, we recommend mathematic model (7) of graduating by direction of documents for the modern method of searching information with FSV technology in e-documents. This model counts the essential of document and rating of it in system. This counting easiness searching information in IRC and gives graduating opportunity for presenting it.

REFERENCES

[1]. SEO (SearchEngineOptimization) в Беларуси, СНГ и мире: эксперименты, новости, литература, терминология,http://seotool.by/seo [online].
[2]. Ландэ Д. В. Интернетика: Навигация в сложных сетях: модели и алгоритмы / Д. В. Ландэ, А. А. Снарский, И. В. Безсуднов. –Москва : Книжный дом «ЛИБРОКОМ», 2009. –264 с.
[3]. Сегалович И. В. Как работают поисковые системы / И. В. Сегалович. – Режим доступа: http://download.yandex.ru/company/iworld-3.pdf, [online].
[4]. Каталоги Rambler TOP100, Яндекс Каталог, каталог Aport, MSN, Google, Yahoo. https://htmlweb.ru/analiz/catalog.php [online].