Some problems of information retrieval in Web

ROMAN LUKATSKY, VLADIMIR RYBINKIN
Bureau of Internet Technologies BIT Ltd
Novoposelkovaja str., 6/7, Moscow, Russia, 123459
RUSSIA

Abstract: - This paper is devoted to using of DBMS for processing of Web documents to solve some problems of information retrieval. Algorithm of creation of info-portraits as computer imitation of sense and their application to arrange purposeful data replenishment from WWW is offered and tested.

Key-Words: - Data Mining, HTML, Content Extraction, DBMS, Metadata, Graph

1 Introduction

Engineering education at present-day reality is impossible to imagine without active use of Internet. However rapid growth of Web again sharply puts a problem of information retrieval. The end user has, as a matter of fact, only two basic ways of search of relevant documents: catalogues and search engines. The problem of automatic classification of Web data, despite of its high urgency [9], is very complicated. Therefore the share of manual skills at drawing up of catalogues remains high, and their role steadily decreases. On the other hand, search engines do not take into account initially graph nature of Web data. At both cases there is a problem of information noise, in particular, because of untidiness or unconscientiousness of authors [12].

Unfortunately, Internet cannot be considered as a database in the usual sense. A Web data in most cases are self-described, and do not have a fixed scheme, detached from resources. Hence there are a number of different and independent methods of information retrieval in Web. One of possible algorithms is described in paper [10]. Meanwhile DBMS could help to solve many of user problems, as data typification allows using the DB scheme for verification of data integrity, for the organization of effective search, etc.

Significant part of WWW is presented as HTML pages, and in this sense these Web data are homogeneous. However numerous attempts of automatically extracting of the content by various methods of analysis of HTML documents are far from success [2]. At present time the creation of DOM tree by parsing of HTML (and XML) content is considered as one of perspective directions [7].

Our positive experience of creation and optimization of graph DB with data, initially presented in markup languages [11], stimulate carrying out of researches of an opportunity of similar revealing of a structural and logic marking in HTML documents to put it into a database. Hereinafter in this paper we'll tell about processing of HTML data only.

2 Exploration objective

Originally intended for papers decoration, HTML now is widely used as means of their structural marking and even for the description of semantics [4]. By means of tags of this language the structure of Web resources can be described: text blocks (DIV, P, BR, HR, CODE), headings (H1…H6, CAPTION), lists, images and tables (IMG, TABLE, OL, MENU), various kinds of allocation of structural elements (FONT, STRIKE, SMALL). Some tags are created specially for description of semantic information (ADDRESS, CITE, EM, SUP, DL) and for description of communications with other elements (HREF). In most cases analysis of HTML tags allows to filter a significant share of the auxiliary or other information which are not having the direct attitude to the semantic content of a resource: navigating panels, pictures, banners, counters, etc. Certainly, for creation of some subject database of Web resources it is not enough to use only HTML tags.

Another approach consists in the analysis of text content itself. It can be some sensible standard terms (home, news, about) [2] or info-portrait - the set of words and phrases to characterize some document [3]. We consider as info-portrait the set of statistically significant words of the document or their group. As we do not plan in this paper the semantic analysis of text content, we'll try to use this term as computer imitation of sense. Info-portrait is accessible to processing for DBMS, in particular, for revelation of correlations between data elements and as search inquiry.

To check an applicability of our method we'll use human-making Internet catalogue. We consider using the statistically processed verbal descriptions of Web resources for establishment of relationships between elements of a DB. Already existing links of catalogue we assume as etalon.

The purpose of this paper is a researching an opportunity of creation of some subject database of Web resources, an organization of data replenishment from WWW and insertion of new data into structure of a DB. This database should be able to grant additional advantages for Web user due to integration of all standard abilities of DBMS at keeping of natural graph representation of Web data.

2.1 Benchmark data and tools

Original data were received from archive dated 5/22/2006 of subset "Kids and Teens" of the catalogue of Internet resources DMOZ [6], and were converted to format of graph DBMS Sindbad [11]. Original HTML pages of these resources have been received directly from Internet. Additional replenishment of database was fulfilled as a result of search inquiries to search engines Google and AltaVista. Computer Pentium-4, 2.4 GHz (512 Mb RAM), Windows-2000, the programming language C was used. Compilation of utilities of batch-mode processing was carried out by means of compiler BC ++ v3.1. A Web browser MS IE 5.00 as a client part of DBMS was used.

3 Creation of graph DB of Web data

Each resource of catalogue of Internet resources DMOZ, as a rule, is presented by means of a small verbal description of the composer (Title and Description). To create some statistically significant info-portrait for separate resource by its description is practically impossible, as the volume of the description is too small. However quite often there is an opportunity to make info-portrait of rubrics as result of statistical analysis of words, which belong to all its resources.

3.1 Creation of info-portraits

The initial DB is very non-uniform on its structure. Almost 10% of rubrics do not contain resources at all, others - some hundreds. We have established a threshold of the statistical importance in our experiment as not less than 16 resources - about 20% of rubrics. Text contents of resources belonging the analyzed rubric were broken into the separate words. For creation of info-portrait we used only words containing not less of 5 symbols and meeting in the common text of descriptions not less of 3 times. Besides, the words belonging to top 100 words of an info-portrait of whole database (imitation of the dictionary of "stop-words") were deleted from an info-portrait. The info-portrait was created if it contained not less than 4 words. Altogether info-portraits have been created for 345 (about 7%) of rubrics.

The created info-portraits, generally speaking, have for the person some semantic sense that allows in most cases at 16 and even at 8 words of a portrait precisely enough identifying a subject domain described by it, for example:

- ballet, dance, beginning, professional, instruction, modern, creative, academy (Ballet Schools).
- space, astronomy, universe, earth, solar, system, planets, exploration (Astronomy).
- greek, mythology, myths, heroes, ancient, temple, goddesses, creatures (Greece Mythology).
- health, fashion, advice, women, beauty, message, fitness, write (Girls Only).

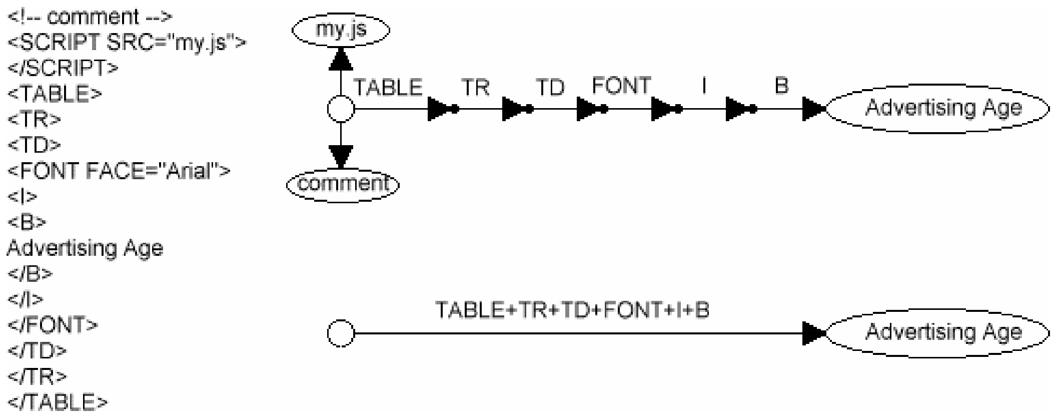
Certainly it concerns to rubrics, which classification is simple for human. In other cases (for example, classification of resources of children's homepages in catalogue DMOZ is made exclusively alphabetically), association of resources to rubrics was made practically senselessly. Meanwhile the simple statistical analysis of 1549 resources of this group allows revealing about 50 possible thematic rubrics on interests of owners of these pages. We present the first 8 words on frequency of a mention: poetry (93), gallery (78), movie (67), drawings (58), writings (54), sport (51), computer (42), artwork (39).

3.2 Import of HTML data

In connection with absence in DBMS Sindbad of utility for import of data directly from WWW we left in an initial DB three thematically close rubrics only and have received HTML documents for these rubrics manually (112 pages altogether). Import of these documents into a database, as well as in [11], consist in allocating of contents of HTML tags into separate nodes by recursive opening of enclosed tags. At present time the software of DBMS has been modified to allow edges to have their own arbitrary information, therefore the contents of HTML tags was not allocated into incidental nodes, but was brought into a DB as information of edges (fig. 1).

After this operation even the one HTML document can be represented as a subgraph composed of thousands of nodes. In this paper from all this variety only a small number of semantically significant elements were interesting for us: a part of the information of heading of Web page, Web environment of the page itself and textual data to create a info-portrait most adequately.

The created subgraph of a Web resource was cleaned from nodes, which do not contain its own content, and the information of their edges were united (fig. 1). The information of Web resource heading (META: Keywords, Title, Description) has been transformed to proper metadata of resource by the way, described in [11]. All hyperlinks (HREF) and the text of its vicinities have formed group metadata LINKS. At that we did not consider comments and the data, presented on script languages (<!-...-->), the links on non-HTML resources (gif, mp3, avi, pdf). In many cases it automatically filtered data (banners, counters), which not have the direct attitude to the semantic content of the resource. The content of other nodes of



```
<!-- comment -->
<SCRIPT SRC="my.js">
</SCRIPT>
<TABLE>
<TR>
<TD>
<FONT FACE="Arial">
<I>
<B>
Advertising Age
</B>
</I>
</FONT>
</TD>
</TR>
</TABLE>
```

Fig.1. Import of HTML data

the graph was used for creation of info-portrait. At that it was automatically carried out process of tokenizes a pages [7], and in this case there is an opportunity of much wider interpretation of concept "token" that can be any node or subgraph of HTML page or their group. Thus, the concepts "token" and "cluster" are joined, and this subgraph can be considered as self-dependent document for possible following more detailed structural or semantic analysis.

As all words of info-portraits have been born into separate nodes of the graph, search of conformity of info-portraits is made not by the analysis of its texts, but upon graph attributes only: by the presence of an edge on the same node of metadata. This approach allows using for cataloguing also nonverbal characteristics of resources, widely presented to a Web. One more advantage of this method consists in an opportunity of the registration of synonyms. For example, any of words "airplane", "aeroplane", and "aircraft" will be considered as the same, irrespective of words what really are present at info-portraits of a resource or rubric, if the corresponding synonymous link is registered in metadata.

The info-portrait was made only if the number of words of the text for its creation was not less than 256. Keywords and words of info-portrait for the organization of the subsequent navigation were brought in a DB as metadata [11]. Generally speaking, it is possible to reveal much more metadata: the size, color and type of a font, italics, underlining, the register of symbols, headings, etc. At that it seems possible to change tags of physical formatting to logical one by "losing" a part of the information of HTML tags. For example, the concrete color, size and font of the text are possible to replace with terms: "usual", "emphasis", etc. However we limited set of metadata as above, and did not identify even the information of headings of the text.

3.3 Checking of conformity of info-portraits

For checking correctness of application of the method "info-portraits" for an establishment of semantic communication the info-portrait of everyone Web resource was consistently compared with portraits of each of three rubrics of a DB. The statistical importance of conterminous words of info-portraits of rubric and resource determined quantity of the points granted to a resource on a three-point scale. For each coincidence of top 8 words (analogue of Title) 3 points, of following 16 (analogue of Description) - 2 points, of following 32 (analogue of Body) - 1 point were charged. The total quantity of points represents the numerical characteristic of link between a resource and a rubric irrespective of whether there is really such arc in the graph. Thus, this characteristic can be used for forming new edges of the graph (this threshold has been specified by us as 16 and more points) and for removing existing edges (less than 8 points). The intermediate score does not change current structure of the graph.

As a result, 33 resources (30%) have not exceeded a necessary statistical barrier in 256 words, and all of them without exception represent navigating pages (as a rule, homepage of sites). 73 resources that have their own info-portraits (96%), confirmed that they "have the right" to be presented in these rubrics, and three remained really have delicate attitude to rubric "Nutrition" (milk, sugar and school lunch).

For each rubric the info-portrait was formed also not of the text of descriptions, but of contents of Web resources itself (table 1). It is clear, that basically info-portraits are similar, though the images made on contents Web resources seem to us bolder. Both of an info-portraits were used for inquiries to search engines: Google and AltaVista.

3.4 Inquiries to search engines

The results of search by initial info-portraits have shown satisfactory quality of ones: each engine on relevance has raised rubric as the first resource. Then we consistently cleaned from inquiry the most popular word of an info-portrait. Results have appeared unexpected: only at cutting of a info-portrait down to three or even two (!) words, the corresponding rubrics have disappeared from the first page of search results. On bigger number of words of inquiry they

Table 1. Data exchange with WWW

Rubric	**Source**	**Search images**	**Type search**	**Search engine**	
				Google	**AltaVista**
Health Nutrition	DMOZ	nutrition recipes healthy explains quizzes fruit foods eating pyramid health citrus types puzzles fruits vitamins virtual	And	542	99
	Web pages	health nutrition healthy comic foods wrigley vegetarian recipes calcium footer fruit eating fitness vegetables products commission	And	83	0
			Or	336000000	303000000
Health Substance Abuse	DMOZ	drugs effects alcohol abuse national friend substances signs quizzes problem legal getting explains advice	And	138000	127
	Web pages	drugs samhsa alcohol abuse zurich univers issues dropmenu audiences homeimages drinking health problem marijuana addiction usercontrols	And	0	0
			Or	281000000	299000000
Health Substance Abuse Tobacco	DMOZ	tobacco smoking smoke industry health effects cigarettes nicotine message against smokeless secondhand quitting memos dangers affects	And	581	44
	Web pages	smoking tobacco smokers smoke health familydoctor cigarette newhome cigarettes market cancer advertising camel brand younger brands	And	73	0
			Or	232000000	301000000

persistently hold 1-3 places. That has made at us impression, what these search engines at any words of inquiry will continue to propagandize a balanced diet and to convince of harm of excesses in general and smoking in particular. However as soon as we expand initial inquiry with one more word (safety), these resources have disappeared not only from top, but also from search results – situation simply inconceivable for DBMS. On a rubric "nutrition", however, this word has not made impression - it has kept 1 place.

Even more surprising results were received by inquiries over info-portraits, generated on resources (table 1). This radical difference of results of search for very similar images has forced us to repeat these inquiries in a mode "any word". For the first 20 resources received from each search engine by each inquiry (excepting repetitions), their individual info-portraits also have been made (68 resources). These portraits were checked on an opportunity of a binding to all three chosen rubrics, and 41 resources (60%) have overcome the set threshold by quantity of points and have been brought into a DB. One of these resources has got a links to both "Substance_Abuse" and "Tobacco" rubrics.

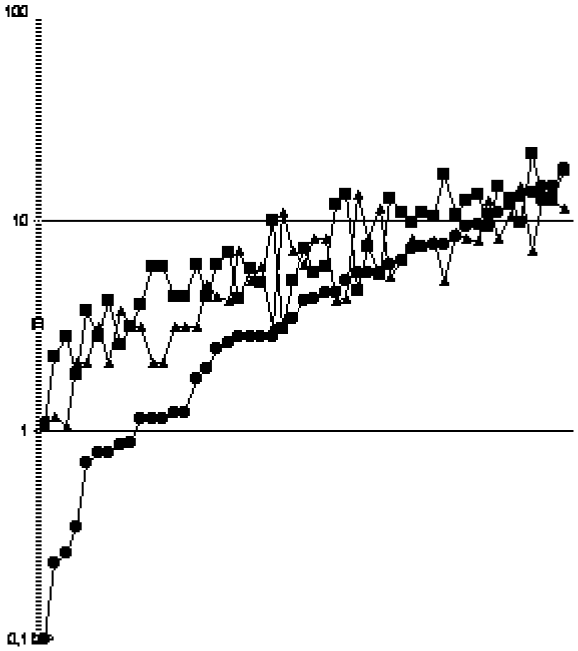
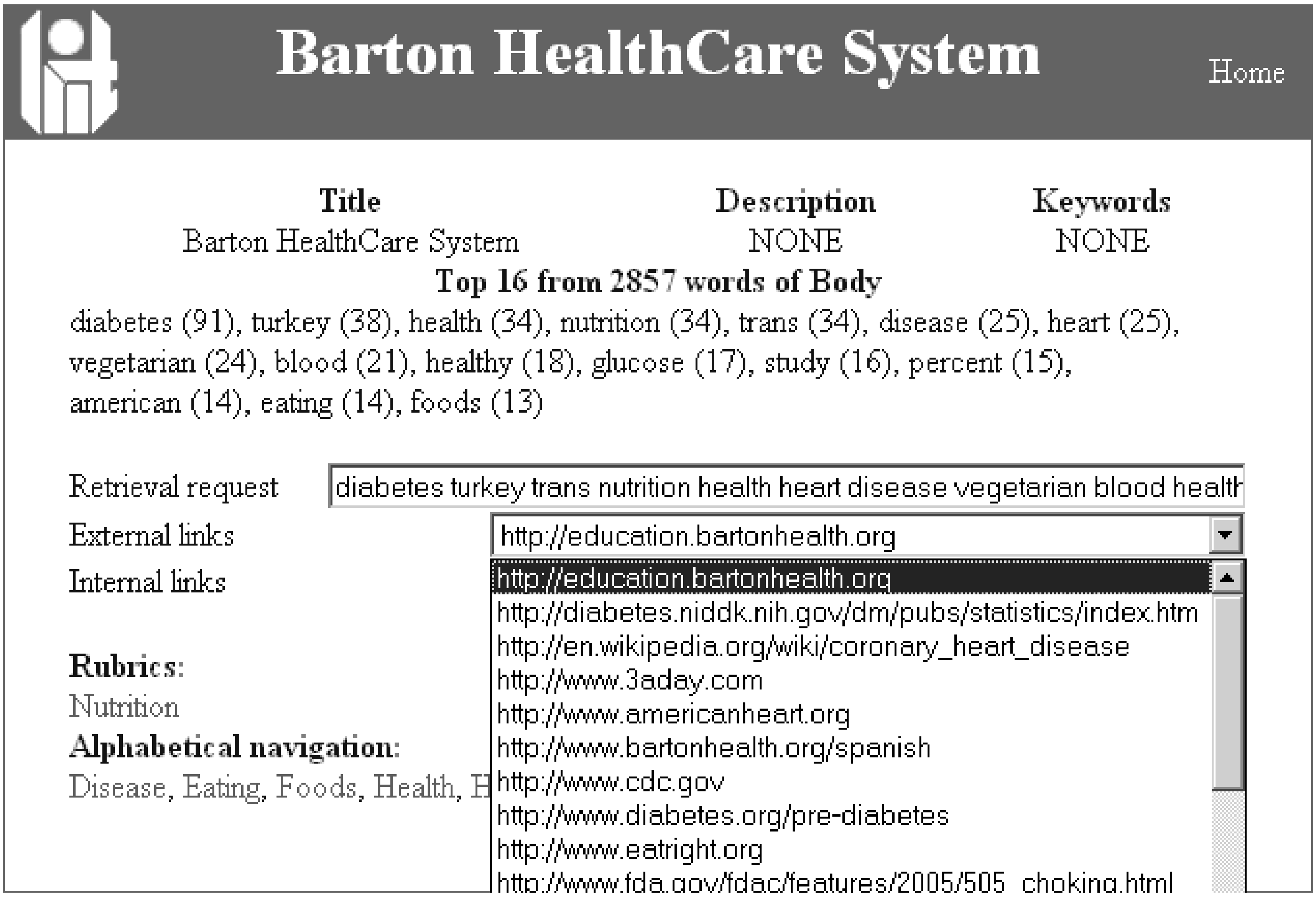


Fig.2. Search engine spam detection by keywords.

3.5 New abilities

Already the fact of import of Web resources into a DB allows granting new opportunities for end user. Except of control of data integrity it is possible to personify the user interface and to give an opportunity of program access to data by users utilities for specialized data processing. Besides, DBMS allows to create very flexible inquiries at search of non-uniform data, using transformations of types and the mechanism of navigating expressions close to language of inquiries Lorel [1], for example, to find out the most cited sites for selected subject (FIND.CURRENT.CITES). Even for our very small database (529 links altogether) this inquiry lead to quite reasonable answers: Centers for Disease Control (http://www.cdc.gov, 3 links) and National Institutes of Health (http://www.nih.gov, 4 links).

We'll shortly stop on a topical problem "search engine spam" [8]. For an estimation of reliability of the presented information let's compare info-portraits of resources with data of tag Keywords from Web page. Results of experiment are shown in a logarithmic scale on fig. 2. The share of conterminous words of an info-portrait (are represented as squares) was considered just as a share of conterminous keywords (are represented as triangles). The product of these parameters (are represented as circles) was applied as an integrated quantitative estimation of conformity of a real contained resource and the information given about it by its founders. Selective



Barton HealthCare System Home

Title	Description	Keywords
Barton HealthCare System	NONE	NONE

Top 16 from 2857 words of Body

diabetes (91), turkey (38), health (34), nutrition (34), trans (34), disease (25), heart (25), vegetarian (24), blood (21), healthy (18), glucose (17), study (16), percent (15), american (14), eating (14), foods (13)

Retrieval request: diabetes turkey trans nutrition health heart disease vegetarian blood health

External links: http://education.bartonhealth.org

Internal links

http://education.bartonhealth.org
http://diabetes.niddk.nih.gov/dm/pubs/statistics/index.htm
http://en.wikipedia.org/wiki/coronary_heart_disease
http://www.3aday.com
http://www.americanheart.org
http://www.bartonhealth.org/spanish
http://www.cdc.gov
http://www.diabetes.org/pre-diabetes
http://www.eatright.org
http://www.fda.gov/fdac/features/2005/505_choking.html

Rubrics:
Nutrition

Alphabetical navigation:
Disease, Eating, Foods, Health, H

Fig.3. Representation of Web data in DBMS.

expert estimation has shown basic applicability of the given criterion: for all checked resources from the left quarter of diagram the declared Keywords do not correspond with the real contents. On the contrary, really respective Web resources had high value of this criterion.

4 Discussion of results

The new structure of data allows to user to navigate on a DB without formation of search inquiries, including abilities don't stipulated by composers of the initial catalogue. One of 41 new received resources together with its environment is shown on fig. 3. All edges are presented as hyperlinks that allows carrying out navigation on any dimension of the graph uniformly, by means of a Web browser. The search results also are presented together with their links that allows making navigation on it and facilitate their analysis.

The information received from the Internet about a Web resource generally consists not only of an info-portrait, but also references to external Web resources. Each of these references is checked on its presence in a DB and if necessary brought in it as new node, and reference is replaced with usual edges of the graph. Thus, each Web resource is brought any more as separate unit, but as a subgraph. New nodes do not contain any information except of URL, therefore, "stimulate" DBMS to independent investigation of Internet for additional information without any assistance. In a like manner, DBMS often are able to detect a type of resource by comparison of number of external links and of volume of the text (sitemap, catalogue). Therefore, it is possible to detect the groups of related Web resources to count up info-portraits not only of Web pages but also the group info-portrait of site in automatic mode.

The offered method of imitation of semantic sense by info-portraits has shown its basic applicability for the organization of data exchange with World Wide Web. At least, in many cases this "sense" is clear for search engines. In particular, we used the info-portrait of resource shown on fig. 3 as search inquiry to Google and AltaVista, and it was one of the first in search results (as well as all three rubrics). Moreover, info-portrait is direct consequence of primary properties of entities [5] such as: attributes, relationships with other entities or entering into named entity set. So info-portrait is stable characterization of document or any verbally expressed concept and can be of interest directly for the end user for creation and refinement of search inquiries in Web.

For an illustration of this statement we'll lead an experiment. Let's assume that user needs something about coffee. On his inquiry with keyword "coffee" Google has found 52 900 000 resources that, certainly, in fact means absence of the answer. Let's assume also that user hardly knows, which words need to be added in search inquiry to narrow area of the answer down to comprehensible one at preservation in search results of all interesting for him of documents. Whether is possible to help him?

Creation of info-portrait on HTML page with first 100 search results that user has received in reality (this analysis can be executed even on clients computer at the corresponding organization of page of results) allow to user to see that in offered to him documents is spoken not only about coffee, but also about other concepts, some of which also can be interesting for him at present. He can use them for detailed elaboration of search inquiry.

The new inquiry on chosen by the user words of info portrait has narrowed volume of documents down to 10200, sharply having changed content of top 100 and, accordingly, info portrait itself. All words of inquiry became a top of info portrait. We result their frequency in old/new queries: arabica 4/210, beans 12/264, espresso 12/247, gourmet 11/265, organic 11/204, recipes 4/151, retail 4/174, roaster 7/189. It is interesting that the word "coffee" in spite of the fact that in second inquiry it was absent, remained the most popular: 563/517.

4 Conclusion

Processing of HTML data with graph DBMS has shown applicability of the offered technique for creation of subject databases of Web resources that allows granting some benefits for end users. First of all there is an opportunity to arrange the purposeful thematic information retrieval by using of navigating expressions and also of info-portraits for automatic classification of Web data and for creation of inquiries to search engines. Besides, the possibility of presentation to user of metadata about data structure permits more efficient representation of Web documents, in particular, by navigation on their environment and on keywords. For example, it is easy to count up citation index of Web pages and, at some elaboration, of arbitrary textual documents.

The mechanism of quantitative estimation of relationships between resources, in our opinion, is interesting first of all due to the numeric characteristic which allows to detect still absent in DB the attitudes between the elements, right up to real removal or creation of edges of the graph and even to full merge of nodes.

The small volume of a DB has not demanded creation of the distributed database. However software of DBMS, by transforming of internal edges in external, allows dividing of a DB into any number of servers. Such division is transparent for end users, and at increasing of volume of a DB or of number of users can be expedient.

Unfortunately, in this exploration were not applied neither operations of indistinct comparison of portraits, nor grammar form analysis, nor even operation of Boolean logic and instructions of a priority of operations - it will be a subject of our further researches.

References:

[1] S.Abiteboul, D.Quass, J.McHugh, J.Widom and J.Weiner, The Lorel Query Language, *Journal of Digital Libraries*, Vol.1, No.1, 1997.

[2] M.Ageev, I.Vershinnikov, B.Dobrov. Automating Extraction of significant part of Web pages for information retreival. *RCDL*, 2005.

[3] A.Antonov, E.Kurziner. An automatic revelation of a subject domain of the big raw text file, *Computer linguistics and intellectual technologies, International seminar Dialog*, 2002

[4] T.Berners-Lee, D.Connolly. HTML 2.0, 1995, http://www.ietf.org/rfc/rfc1866.txt

[5] P.Chen. The entity-relationship model - towards a unified view of data, *ACM Transactions on Database Systems*, Vol.1, No.1, 1976.

[6] DMOZ Open Directory Project, http://www.dmoz.org

[7] S.Gupta, G.Kaiser, P.Grimm, M.Chiang, J.Starren. Automating Content Extraction of HTML Documents. *WWW Journal*, 2005

[8] Z.Gyongyi, H.Garcia-Molina. Web spam taxonomy. *AIRWeb*, Chiba, Japan, 2005.

[9] S. H. Hoi, R. Jin, M. R. Lyu, Large-Scale Text Categorization by Batch Mode Active Learning, *International World Wide Web Conference*, 2006.

[10] S.Madnick, Integrating Information from Global Systems: Knowledge Representation in the Context Interchange System, *Moscow ACM SIGMOD*, 2005.

[11] V.Rybinkin, R.Lukatsky. Elastic model for processing of heterogeneous data, *WSEAS TRANSACTIONS on SYSTEMS and CONTROL*, Issue 2, Vol.2, 2007

[12] Baoning Wu, Brian D. Davison, Detecting Semantic Cloaking on the Web, *International World Wide Web Conference*, 2006