Performance analysis of the method for social search of information in university information systems

Georgi Petrov Dimitrov
University of Library Studies and Information Technologies
Sofia, Bulgaria
geo.p.dimitrov@gmail.com

Galina Panayotova
University of Library Studies and Information Technologies
Sofia, Bulgaria
panayotovag@gmail.com

Ivan Garvanov
University of Library Studies and Information Technologies
Sofia, Bulgaria
igarvanov@yahoo.com

Bychkov OS
Taras Shevchenko National University of Kyiv
Kiev, Ukraine
bos.knu@gmail.com

Pavel Petrov,
Faculty of Informatics
University of Economics - Varna
petrov@ue-varna.bg

Angel Angelov
University of Library Studies and Information Technologies
Sofia, Bulgaria
angel.s.angelov@gmail.com

***Abstract* — in this article the effectiveness of one of the main methods for assisting the information search in large data sets is analyzed, namely the method based on the social approach. This method analyzes the behavior of multiple users when searching for information, particularly the keywords they use. Based on the obtained results, the relevant algorithms in the searching engines of the organizations can be optimized. The studies are based on data extracted from the databases of two relatively large information systems based at the University in Library Studies and Information Technologies in Sofia, Bulgaria and University of Economics in Varna, Bulgaria. The numerical and experimental results are analyzed and discussed.**

Keywords - information systems, search methods, databases, data sets optimization.

I. INTRODUCTION

There are many different methods for analyzing data sets, the core of which are tools borrowed from statistics and informatics (especially machine learning). However, researchers continue to develop new methods and improve the existing ones. It is appropriate to note also that some of the methods are not designed exclusively for large data and quite successfully can be used for smaller arrays (eg A / B testing, regression analysis, etc.). But of course, the more voluminous and diversified array is analyzed, the more accurate and relevant the results are. [1,2,5]

Currently, information systems for searching of bibliographic and other information can use the following basic methods assisting the search [6,7,8,9]:

- Methods based on expert approach. In this method, a team of specialists determines keywords, the relationship between them and rank them in order of importance. The advantages of this method are high relevance between keywords and desired result. Some of the disadvantages are: slow performance, hard covering of all subject areas and high cost.

- Methods based on text mining. In these methods information from a large number of information sources is extracted programmatically by using the methods of artificial intelligence.

- Methods based on social approach. These methods examine the behavior of multiple users and particularly what keywords are used in the search of information sources.

We share/support the opinion of some authors that "... the convergence of modern science and technology to "technoscience" leads to the emergence/ formation of radically new social phenomena like the associated social intelligence that involves not only the natural intelligence, but also the models it created... ". The effectiveness of a method based on the work of social networks, whose essence is explained below, will be tested as an attempt to use the "social intelligence". It is possible to accelerate the searching process by using this method. When searching for information sources in the library system, records about each user (keeping his anonymity), the sources in which he is interested or the keywords he used are

ISBN: 978-1-4673-9186-3 ©2016 IEEE

automatically kept by the system. Subsequently, these data considered as Markov model through dynamic Bayesian network can be used for automatic giving of recommendations to other users regarding a particularly, chosen by them, source, providing informing about what appropriate sources other users have used or what keywords to use in relation to their specific study[13]

II. SOCIAL SEARCH METHOD – THEORETICAL RATIONALE

As stated above, the process of social search is based on the experience and the preferences of other users. Using their behavior, the information seeker can navigate faster and choose the right keywords to be used for information search in the bibliographic database, and to reach faster the necessary information sources [3,4]. The reason to seek new forms and options for bibliographic information search is confirmed by other authors: "Library organizations are put in very tough competition with other operators on the market in terms of gaining the consumes interest".

In this report we examine the impact of the social search method on displaying an additional list of suitable phrases based on the experience of other users; [10,11,12]

At the basis of the algorithm lies down the accumulation of data about the behavior of multiple users and build up a matrix of probabilities. The output staging is as follows: a system user is interested in literature in a specified subject area. Initially his system assigns a unique randomly generated session ID to track his behavior and keep his anonymity. The user enters certain search keywords and as a result a list of redundant data for each record is displayed. The user has to explicitly choose a record, thus he will be able to see all the information about it. When choosing a specific entry, information about the choice is recorded by the system. Thus the accumulation of information about logicall interconnected records in the bibliographic database is provided, which is based on subjective preferences. Of course, the user may mistakenly or deliberately choose to review records not related to the subject area of interest, but the accumulation of information about the behavior of a large number of consumers will eliminate the importance of these fluctuations, which actually represent a deviation from the normal rational behavior. Practically, as a result from the accumulation of data on consumer behavior we will have the following sample data presented in JSON format

```
{
    "SESS_ID1": [id1, id3, ..., idX],
    "SESS_ID2": [id1, id5, ..., idY],
    .....
    "SESS_IDN": [id3, id5, ..., idZ]
}
```

, where:

SESS_ID - a random session ID of user, for which we assume that once entered into the system he will search for information in a particular subject area;

id - unique identifier (system number in the database or signature, if it meets the condition for uniqueness of the literary source of bibliographic database). For brevity we'll call it signature.

Every session ID corresponds to a plurality (array) with a different number of elements, namely signatures, wherein the duplication is eliminated. This is done in order to prevent the option one user to influence the meaning and the importance of a signature, because it eventually would distort the list of proposals in favor of this literary source. This is one of the main differences from the famous model "Bag-of-words" [], which is a base for various popular implementations such as "Term Frequency - Inverse Document Frequency" (tf-idf).

From an algorithmic perspective duplications most easily can be removed if the list is sorted, so we can assume that the list of thee signatures per user is sorted and keeps unique values.

The matrix of the frequencies is a square matrix, symmetric to the main diagonal. Initially, all values in the matrix are zero. For each value from the signatures list of a user, the values of the cells, whose columns correspond to the whole list, are incremented with one on the relevant row in the matrix. In other words, if the list consists of five signatures, then five rows in five columns should be incremented. The process is repeated for the data of other users.

As a result, the values on the main diagonal NW-SE will show how many times this signature attended in the lists of all users. Above and below the main diagonal the values are symmetrical and show how many times the signature corresponding to the row or column has been in the same list with other signatures, wherein the values will range from zero to the number in the cell of the main diagonal of the corresponding row or column.

Quotient between the number in a cell and the number of the cell in the main diagonal in the same row or column gives us the probability of a signature to attend along with other signature in one list. Thus, if two signatures are always present together in lists, the probability is 100%, and if they have never attended together - 0%.

We will give a short simple example. When input is:

```
{
    "SESS_ID1": [1, 3, 5, 7],
    "SESS_ID2": [2, 5, 6],
    "SESS_ID3": [1, 7],
    "SESS_ID4": [1, 3, 6, 7],
    "SESS_ID5": [5, 7]
}
```

, the matrix of the frequencies will be as shown in Fig.1.

	1	**2**	**3**	**4**	**5**	**6**	**7**
1	3	0	2	0	1	1	3
2	0	1	0	0	1	1	0
3	2	0	2	0	1	1	2
4	0	0	0	0	0	0	0
5	1	1	1	0	1	1	2
6	1	1	1	0	1	2	1
7	3	0	2	0	2	1	3

Figure 1. Frequencies matrix

Practical programming realization of a real matrix in the form of two-dimensional array is very difficult for optimal implementation having in mind the large volume of memory it could use. For example, when using 32 bits data, if for 10,000 signatures the memory consumption is 400 MB, then for 100,000 signatures, it will be about 40 GB. This exponential growth creates significant obstacles for the practical use of the algorithm.

Using a structure of an array of arrays could lead to almost double reduce memory usage, since using the fact that the matrix is symmetrical, it can be represented in the memory as triangular. We propose another approach, starting with the assumption that a majority of signatures are not logically interconnected i.e. the matrix will have a large amount of zeros. If our assumption is true, the structure of associative array of associative arrays (hash of hashes) would be a more appropriate organization for implementation of the frequencies matrix. For example, the matrix of frequencies from the previous example can be represented as follows in JSON format:

```
{
    "1":{"1":3,"3":2, "5":1, "6":1, "7":3},
    "2":{"2":1, "5":1, "6":1},
    "3":{"1":2, "3":2, "5":1, "6":1, "7":2},
    "5":{"1":1,"2":1, "3":1, "5":3, "6":1, "7":2},
    "6":{"1":1,"2":1, "3":1, "5":1, "6":2, "7":1},
    "7":{"1":3, "3":2, "5":2, "6":1, "7":3}
}
```

In this operation mode all zero values are not stored at all, and in addition, directly can be used signatures, which are essentially strings, not numbers.

After that this matrix of frequencies can be used to display additional recommended list depending on the behavior of consumers seeking bibliographic information. In order to optimize the occupancy memory and the high performance, it is more appropriate to transform the matrix into more suitable form, namely to use the structure of associative array of arrays in which numbers from the main diagonal are dropped out and values in a row are sorted descending, wherein there are only signatures. Additionally each list can be cut and data can be stored only for signatures, which are most relevant to a specific signature.

The final transformed sample data would look in the following way:

```
{
    "1":["7", "3", "5", "6"],
    "2":["2", "5", "6"],
    "3":["1", "3", "7", "5", "6"],
    "5":["5", "7", "1", "2", "3", "6"],
    "6":["6", "1", "2", "3", "5", "7"],
    "7":["1", "7", "3", "5", "6"]
}
```

This optimization makes sense, because the specific rates/ percentage are not so useful for the consumers, as the order in which the proposed sources are ranked - more closely related to the demand - higher on the list.

III. SOCIAL SEARCH METHOD – STUDY RESULTS

The studies are based on data derived from Agora - the information system of the "University of Library Studies and Information Technologies" and Automated Library Information System (AB) of the "University of Economics" – Varna.[3,4]

User ID	**Number of searches by maximum number of words, regardless of the number of users**	**1257**	**1380**	**……**
Signature / Word				….
Curriculum	10414	1025	1900	
Curriculum - Subjects	9543	1364	1336	
TestDetails	8500	535	1441	45
Products_List	8303	1441		
……..	……..			
SendMsg	1			

Table 1 **Number of searches by maximum number of words, regardless of the number of users**

Table 2 shows the number of searches for a signature by maximum number of users seeking the same signature.

USER_ID	Number of searches by maximum number of users	1257	1380	…
Сигнатури / Думи				
Crm SprMostri	18			
Learn_Program_D	18	1	1	1
Products_List	17			
Setrifikati	16			
AccessControl	14	1	1	1
……..				
……				
ZawerkaDisciplina	1			
Конфигуратор – форми	1			
Grand Total	497	26	16	16

Table 2 **Number of searches by maximum number of users**

On fig. 2 is presented the dispersion of searched words by authors.

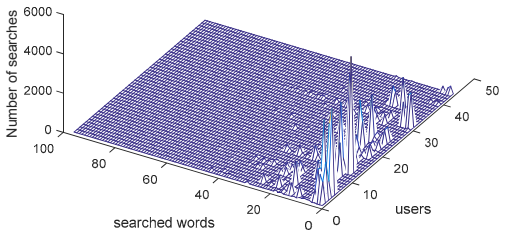

Fig. 2. The dispersion of searched words by authors.

On fig. 3 is presented the numerical dispersion of the experiment result.

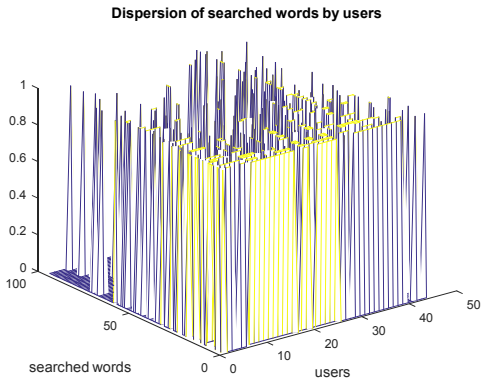

Fig. 3. The result of the experiment

It is obvious that there is a significant difference between a most frequently searched signature and the most frequently searched signature by many users.

CONCLUSIONS

Maintaining the high quality of the information included in the universities information systems is crucial for their effectiveness. Without relevant by its nature information, the process of training and scientific research would not be sufficiently effective in this university. As a source for completing the information in the information systems of universities serve/is used the incoming data streams of each information system. Depending on the key scientific fields specific to the relevant university, incoming data streams vary widely in terms of data sources. The more sources used in the scientific work of local researchers are available in the information system of the university, the approach of "filtering" of incoming data streams used by employees is more properly performed.

In the individual information systems, the sorting of lists for optimal display of the searched information should be performed based on the signatures most frequently searched by a maximum number of users.

To adjust the incoming data streams it is possible to apply the approach of the feedback in order to evaluate the quality of the selected by the incoming data streams information included in the information system of the university.

The results can be applied in two ways:

- Social search based on user interests from literary sources (signatures) in a particular subject area.
- Social search based on user requests (keywords and phrases) in a particular subject area.

Using the potential of external NoSQL system of type REDIS can significantly facilitate the practical implementation of the methods, because the built-in abstract data structures, such as sets and sorted sets, greatly simplifies the implementation of such projects, while the final storage of the data as structure hash of sorted sets is optimized for fast response with terms of displaying of recommendations to the consumers.

ACKNOWLEDGMENT

This work is partly supported by the project SIP-2016-09 /02/03/2016 - "Research of methods for complex optimization of WEB-based business information systems"

REFERENCES

[1] Georgi Petrov Dimitrov, Galina Panayotova, Queuing systems in insurance companies – analyzing incoming requests, Proceedings in Electronic International Interdisciplinary Conference,The 2nd Electronic International Interdisciplinary Conference, EIIC 2013,2. – 6. September 2013,Slovak Republic, ISBN 978-80-554-0762-3,ISSN 1338-7871,p. 139-142

[2] Galina Panayotova, Georgi Petrov Dimitrov, Researching and Modeling of Queuing Systems of the Insurance Company HASSACC - HASSACC - HUMAN AND SOCIAL SCIENCES AT THE COMMON CONFERENCE,HASSACC 2013 - Virtual Conference Human And Social Sciences at the Common Conference,18-22 November, 2013, ISBN: 978-80-554-0808-8 ISSN: 1339-522X, p. 93-95

[3] Georgi Dimitrov, Ilian Iliev, Study of methods for front-end webpage optimisation The 3rd International Virtual Conference 2014 (ICTIC 2014) Slovakia, March 24 - 28, 2014,

[4] Georgi P.Dimitrov, Ilian Iliev, Front-end optimization methods and their effect MIPRO 2014 - 37th International Convention, 26-30.06.2014

[5] Georgi Petrov Dimitrov, Galina Panayotova, Stefkka Petrowa, Analysis of the Probabilities for Processing Incoming Requests in Public Libraries The 2 nd Global Virtual Conference 2014 (GV-CONF 2014) Goce Delchev University Macedonia & THOMSON Ltd. Slovakia, April 7 - 11, 2014,ISSN: 1339-2778

[6] Georgi Dimitrov, Galina Panayotova, ANALYSIS OF THE QUERING OF DATABASES IN SYSTEMS FOR QUALITY MANAGEMENT OF EDUCATION, Macedonia, 12th International Conference on Informatics and Information Technologies, 04.2015, Bitola, Macedonia

[7] Galina Panayotova, Georgi Dimitrov ,Balancing Automated and Manual Testing with Opportunity Cost, the 4th International Virtual Conference 2015 , (ICTIC 2015) Slovakia, March 23 - 27, 2015

[8] Georgi Dimitrov, Galina Panayotova, "ASPECTS OF WEBSITE OPTIMIZATION", PROCEEDINGS OF THE UNION OF SCIENTISTS – RUSE VOL . 12 / 2015, 106-113, ISSN 1314-3077

[9] Galina Panayotova, Georgi Petrov Dimitrov, Comparison of methods for optimization of websites, International simposium - Operation reserch , Bled, Slovenia, 22-25. 09. 2015

[10] Huston S., J.S. Culpepper, W.B. Croft, Indexing Word Sequences for Ranked Retrieval. //ACM Transaction on Information Systems, 2014, vol.32, №1, p.3:2-3:4

[11] The JSON Data Interchange Format: Standart ECMA-404, 2013. <http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf>, достъпено 20.06.2014

[12] Bag-of-words model, <http://en.wikipedia.org/wiki/Bag_of_words_model>, достъпено 20.06.2014 tf–idf, <http://en.wikipedia.org/wiki/Tf-idf>, достъпено 20.06.2014T

[13] Silverman, B. W., Density Estimation for Statistics and Data Analysis, Chapman and Hall, 1986