Trakia Journal of Sciences, Vol. 17, Suppl. 1, pp 160-164, 2019

Available online at:
http://www.uni-sz.bg

ISSN 1313-7069 (print)
ISSN 1313-3551 (online) **doi:10.15547/tjs.2019.s.01.026**

APPLICATION OF REAL-TIME DICTIONARY INSTEAD OF FULL-TEXT INDEX FOR SEARCHING IN WEB PUBLICATIONS

P. Milev*

University of National and World Economy, Sofia, Bulgaria

ABSTRACT
The article examines topics of searching in web publications. In this sense, the paper presents features of database models, that are suitable for storing such information. In these cases, full-text indexes are mostly used for the implementation of algorithms for searching in the content of web publications. The main purpose of the article is to explore the possibilities of realization of such a database and to present a conceptual data model for searching in web publications using a real-time dictionary. To achieve its goal, the article uses various scientific methods, including study, analysis, research, modeling and experimentation. The results of this paper support the main hypothesis of the study, namely defining the advantages of using a real-time dictionary for searching in web publications. The conclusion highlights the possibilities for improvement of searching in databases, that store web publications.

Key words: Internet, data model, software solution, database, web content.

INTRODUCTION

The increase in global data volume leads to the need of using modern information technologies for storing and analyze this data. It can even be said that the growth of data is one of the main reasons for the need for the digitization of business processes. Nowadays, issues related to the so-called digital transformation in the context of a radical rethinking of how an enterprise uses technology to radically change performance [1]. According to some authors, the change in the business environment in recent years leads to a new reading of the classical problems of project management [2]. The largest part of new data can be found in the global network. In this context, the need for searching in the content of web publications is an extremely actual topic. Because of the progess of information technologies, there are many opportunities for implementing this process. Web publications differ in appearance, structure, and purpose. According to some authors, publications on the internet are online resources, that can be divided into two types – classical (websites and blogs) and new (social media and networks) [3]. The need for searching in the content of web publications is very often associated with digital marketing, in which internet publications are described as online means of communication and dissemination of information [4]. In this context, ways of searching for valuable information in content publishing systems are crucial to organizations and their marketing strategies. According to some authors, organizations nowadays have one major challenge – informing customers and getting feedback on their sustainability [5]. This challenge leads to the need for constant processing of large amounts of information. At the same time, the online content created by internet users is constantly increasing. Some authors explore the development of the web concept and define two types of internet users who create content – professional authors and end-users [6]. In any case, the content generated in web publishing information systems is of interest to organizations' analysis. According to some authors, cloud resources are used to store and analyze these data, which may be in very large dimensions [7]. According to other researchers in the same field, web publications are the main source of

Correspondence to: *Plamen Milev, University of National and World Economy, Sofia 1700, tel. +359 2 8195 312, e-mail: pmilev@unwe.bg*

external information about organizations [8]. Many organizations pay to receive this information in summary form or in the form of different data analysis. Platforms that offer such services are commonly referred as media monitoring systems, media clipping systems, social media monitoring systems, media analysis systems, internet marketing systems, etc. From an architectural point of view, these platforms are multilayered software solutions. Most of the researchers place the data in these architectures in a separate layer and the other layers take care of the business logic and user interface of the system [9]. In any case, these systems use a data model to find the desired publications through a specific search algorithm in the content of these publications. In the present study, it will be first examined the possibility of using traditional full-text search in the content of web publications and then it will be explained the need for application of real-time dictionary.

TRADITIONAL USE OF DATABASE FULL-TEXT SEARCH

Web content search is a search for text in text. This type of search has always been a challenge for information systems and their databases. The storage of different types of web publications in a unified way is also relevant to the interoperability issues. Some authors define interoperability as an option to store and manage data from different systems in a unified fashion [10]. The concept of creating a single input and output for data to optimize system processes has been explored for a long time and nowadays it is well established [11]. The search in the content of the various web-based publications also implies the existence of a data model that allows these opportunities. According to other authors, the increase in the amount of data used by an information system reduces its productivity. In this context, it is necessary to implement special models for system development [12, 13]. Search in the content of web publications can be attributed to the problem of searching for information in unstructured data. According to many authors, the content that internet users create is precisely in the form of unstructured data [14]. Other authors explore the possibilities of automating the processes of structuring unstructured data using specialized search algorithms [15]. In all cases, a data model for storing data from web publications is required to allow searching on different criteria. Industry-leading database management systems offer tools that can test query performance under different conditions, because the times for executing database operations are critical to the willingness of users to use a concrete platform [16]. Operations at data level of information systems are at the basis of the development of both business information systems and public electronic services [17, 18]. One of the options for implementing text search in text in databases is the use of the so-called full-text index. This feature is a technique whereby the database management system (DBMS) allocates an additional storage resource to optimize search times in the appropriate columns and tables using built-in DBMS-specific functions. **Figure 1** illustrates a traditional data model for storing the content of web publications using a full-text index. This model consists of a single table with several columns for storing the different attributes of publications on the internet. These attributes can be a source, title, author, content, date, link, etc. In this case, the web publication content column should have a full-text index.

Figure 1. Data model for full-text search

The use of this model implies the availability and maintenance of a full-text index by the respective database management system. Nowadays, leading database management systems support this capability, but they do so with a variety of functionalities. In any case, creating a full-text index for any column of a table in a database will result in the use of a serious data storage resource for that index. At the same time, this index will greatly improve search times in the content of the column. In some cases, the use of a full-text index will not be possible for technical or other considerations and in other cases it will be possible, but at the cost of a serious resource expenditure.

APPLICATION OF REAL-TIME DICTIONARY

In cases where traditional approaches of using full-text indexes in a database are not appropriate, this functionality needs to be realized in some other way. In this study, this is the so-called real-time dictionary. This concept includes the use of a dictionary of keywords that is updated in real time with new data from web publications. **Figure 2** illustrates this concept. This approach extracts keywords and their positions in the publication from a web content of interest to the platform.

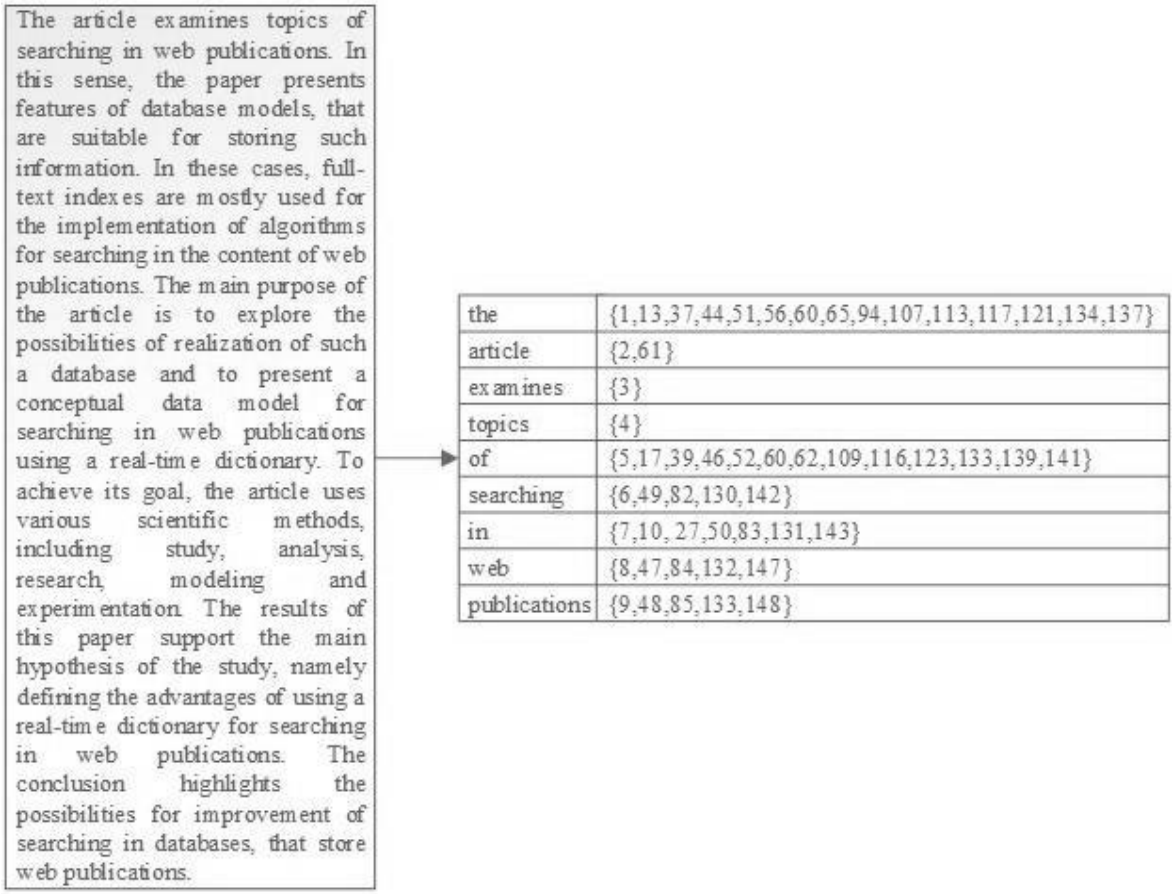

Figure 2. Extracting keywords from content

Retrieving the keywords from the content of the respective web publication is a one-time process. Its goal is to determine the positions of the keywords in the publication. Storing the positions is required, because this way it is possible to search for a specific phrase in the content of web publications. These data should be stored in a database that has a suitable model. **Figure 3** illustrates this model. The proposed data model contains three tables. The first table stores the keywords that are extracted from web publications. The information about the publications themselves is stored in a separate table with the attributes of the web publication without its contents, which is divided into words and is thus stored in a separate table. The third table in this model is intermediate between the keyword dictionary and the publications. It stores information about the corresponding position

of a keyword in the content of the publication. This table serves as a many-to-many relationship between the publications and the dictionary, because a concrete keyword can be found in many publications and a publication itself contains many keywords.

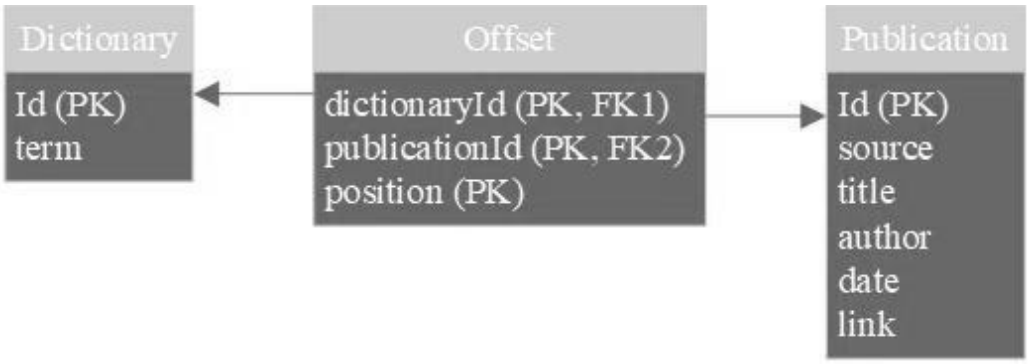

Figure 3. Data model for real-time dictionary

Using the proposed data model for storing web publications also has advantages in some other areas. The presence of full-text indexes in the database implies the storage of the content of the respective web publications in their full form. This, in turn, raises some case studies regarding the copyright of the authors of these publications. These authors may be well-established media, but nowadays the content on the world wide web is created mainly by its end-users. At the same time, the application of the real-time dictionary data model does not require storing web publications in their original form. This model stores data for relevant publications that are used for analytical purposes. Also, the proposed data model is suitable for performing various quantitative analysis as a result of the easier extraction of the count of the presence of a concrete keyword in the relevant publication compared to the use of full-text indexes.

CONCLUSIONS

In conclusion, the following results of the survey can be summarized:

- The traditional approaches for searching in the content of web publications are clarified, namely the use of full-text indexes in the database;
- The opportunities for performing keyword search in the content of publications on the internet are explored using a dictionary that is updated in real time;
- A real-time dictionary data model is defined that supports search by a keyword, search by a part of a word or search by a phrase in the content of web publications.

The benefits of application of the suggested in the article approach for searching in the content of web publications using a real-time dictionary can be presented in the following directions:

- The availability of additional tables and relationships in the data model for searching in the content of web publications, which leads to the use of less data storage resources than the application of full-text indexes;
- Ability to use a single dictionary of keywords for searching in the content of different types of web publications, which leads to improved search times;
- Application of a real-time dictionary data model that stores the content of web publications in a form for analytical purposes.

REFERENCES

1. Belev, I., Software Business Process Management Approaches for Digital Transformation, *Yearbook of UNWE*, 2018, ISSN 1312-5486.
2. Kirilov, R., Features of Application of Software Solutions in Public Projects Risk Management, *Economic and Social Alternatives*, Issue 2, pp. 127 – 140, 2017, ISSN 1314-6556.
3. Vankov, N., Online Resources and Tools for Digital PR, *Economic and Social Alternatives*, Issue 2, 2014, ISSN 1314-6556.
4. Slavova, M., Digital Marketing, *Economic and Social Alternatives*, Issue 3, 2016, ISSN 1314-6556.
5. Stefanova, K., Kabakchieva, D., Big Data Approach and Dimensions for Educational Industry, *Economic Alternatives*, Issue 1, 2019, ISSN 1312-7462.
6. Kisimov, V., Web 3.0 Approach to Corporate Information Systems Evolution,

Economic Alternatives, Issue 2, 2012, ISSN 1312-7462.
7. Boyanov, L., Opportunities and Threats from Implementing Internet of Things, *Economic and Social Alternatives*, Issue 2, 2018, ISSN 1314-6556.
8. Velev, D., Internet of Things – Analysis and Challenges, *Economic Alternatives*, Issue 2, 2011, ISSN 1312-7462.
9. Kirilov, R., Architectural Issues of Building Information Systems for Electronic Reporting of Public Projects, *Trakia Journal of Sciences*, Vol. 13, Suppl. 1, pp 1-3, 2015, ISSN 1313-7069 (print), 1313-3551 (online).
10. Kirilova, K., Interoperability Issues in Bulgaria, *Trakia Journal of Sciences*, Vol. 13, Suppl. 1, pp. 103-106, 2015, ISSN 1313-7069 (print), 1313-3551 (online).
11. Daskalova, T., The human resources in the administration of "one-stop-shop", The Human Resources in the Panorama of Labor: a magazine for labor and social relations, Issue 2, pp. 26-35, 2009, ISSN 1312-305X.
12. Mihova, V., Methods of Using Business Intelligence Technologies for Dynamic Database Performance Administration, *Economic Alternatives*, Issue 3, 2015, ISSN 1312-7462.
13. Denchev, E., Problems and Solutions for Information Provision of Clusters by Small and Medium Enterprises (SMES), *Research Papers of UNWE*, Issue 2, pp. 229-245, 2017, ISSN 0861-9344 (print), 2534-8957 (online).
14. Stefanova, K., Yordanova, S., Knowledge Discovery from Unstructured Data using Sentiment Analysis, *Economic and Social Alternatives*, Issue 1, 2017, ISSN 1314-6556.
15. Marzovanova, M., Intelligent Tagging and Search as a Fully Automated System, *Economic Alternatives*, Issue 2, 2015, ISSN 1312-7462.
16. Radoev, M., Comparison of Tools for Collecting Information About Query Performance in Microsoft SQL Server, *Economic Alternatives*, Issue 2, 2016, ISSN 1312-7462.
17. Shishmanov, K., An Analysis of the Possibilities for the Development of Information Systems in Companies and Organizations, *Business Management*, Issue 2, 2013, ISSN 0861-6604.
18. Kirilova, K., Methodological Issues in Development of Public Electronic Services, *Economic and Social Alternatives*, Issue 2, 2016, ISSN 1314-6556.