Trakia Journal of Sciences, Vol. 11, Suppl. 1, pp 296-300, 2013

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

ISSN 1313-7069 (print)
ISSN 1313-3551 (online)

TECHNOLOGICAL ISSUES OF IMPLEMENTING SOFTWARE FOR MEDIA MONITORING

P. Milev*

University of National and World Economy, Sofia, Bulgaria

ABSTRACT
The paper discusses issues related to the functionality of software for media monitoring. The main objective of this paper is to present methodological solutions of implementation of such software. Leading part in the paper takes issue with keyword search in the information flow and streamlining the search results. On this basis appropriate technological solutions are proposed. The main results of the work of the software solution show feasibility of the proposed methodological approach. The conclusion outlines some trends in development of media monitoring systems in the context of the critical role of modern technological capabilities.

Key words: web sites, indexing information, media monitoring.

INTRODUCTION

Media monitoring is the activity of monitoring the output of the print, online and broadcast media (3). It can be conducted for a variety of reasons, including political, commercial, scientific, and so on. In the commercial sphere, this activity is usually carried out in house or by a media monitoring service, a private company that provides such services to other companies, organisations and individuals on a subscription basis. The services that media monitoring companies provide typically include the systematic recording of radio and television broadcasts, the collection of press clippings from print media publications, the collection of data from online information sources. The material collected usually consists of any media output that makes reference to the client, its activities and/or its designated topics of interests. The monitoring of online consumer sources such as blogs, forums and social networks is more specifically known as buzz monitoring which informs the company of how its service or product is perceived by users. In academia media monitoring is deployed by social scientists in an attempt to discover e.g. biases in the way the same event is presented in different media, among the media of different countries etc. The use of large scale monitoring techniques by computer scientists enabled the exploration of different aspects of the media system such as the visualisation of the media-sphere (1), the sentimental and objectivity analysis of news content etc (2). Media monitoring is practically achieved by a combination of technologies including audio and video recording, high speed text scanners and text recognition software and human readers and analysts. The automation of the process is highly desirable and can be partially achieved by deploying data mining and machine learning techniques. The article discusses technological issues of implementing software for media monitoring of online information sources. Understandings in the article about the nature of this type of media monitoring are illustrated in **Figure 1**. According to this article a media monitoring software should implement the following options:

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

- Maintaining a large number of online information sources (all known if it is possible);
- Ability to automatically search in publications of these information sources according to our interests (interests of our clients);
- Creating a newsletter with wanted publications found according to established criteria.

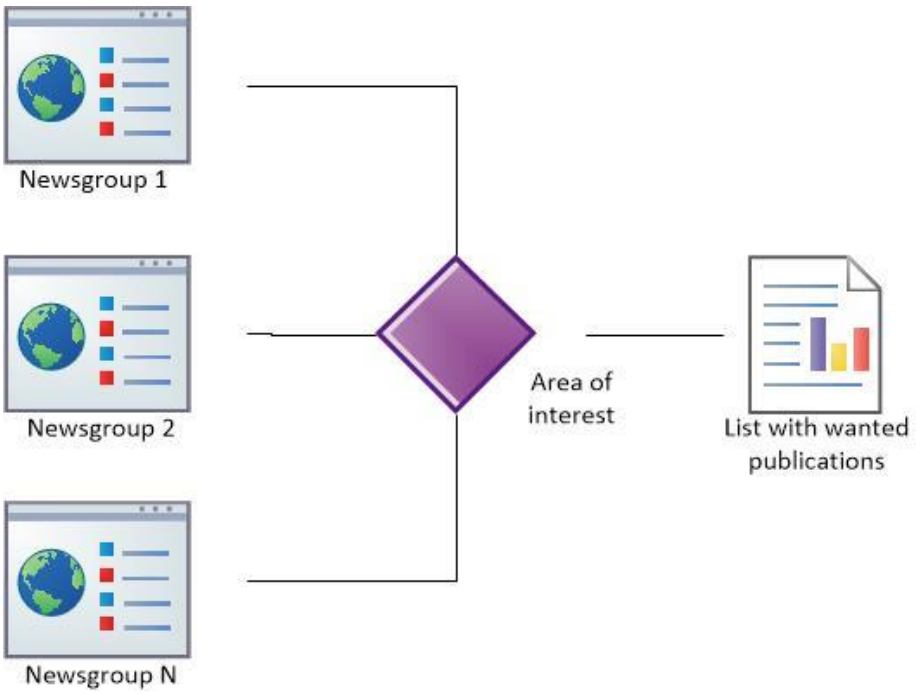

Figure 1. Basics of media monitoring

SOFTWARE ARCHITECTURE

According to this article a complete architecture of software for media monitoring of online information sources should contain the following components **(Figure 2):**

- Web server 1, web server 2, web server 3 – web servers, which check for new publications respectively in Web sites group 1, web sites group 2, web sites group 3 (different groups of web sites, arranged according to certain criteria). The use of several web server aims to provide a resource for media monitoring software when working with a large number of information sources;
- Application server 1 – application server, which manages analyze and retrieval of web content. Its functionality will be considered in this article. This application server also organizes the work of all available Web servers and synchronizes their actions;
- Database server 1 – database server, which stores the downloaded from information sources content that is going to be indexed by media monitoring software;
- Application server 2 – application server, which indexes the downloaded from information sources content. At this level are implemented algorithms for indexing that allows optimal information search from the client part of media monitoring software;
- Database server 2 – database server, which stores all the indexed publications. Client search operations are performed by that server;
- Web server 4 – web server of the client part of the system that manages the content that is displayed on the Web site;
- Application server 3 –application server, which is an intermediate level between the client website and the database with indexed publications. This application server provides search by key words in publications indexed. Its functionality will be considered in this article.

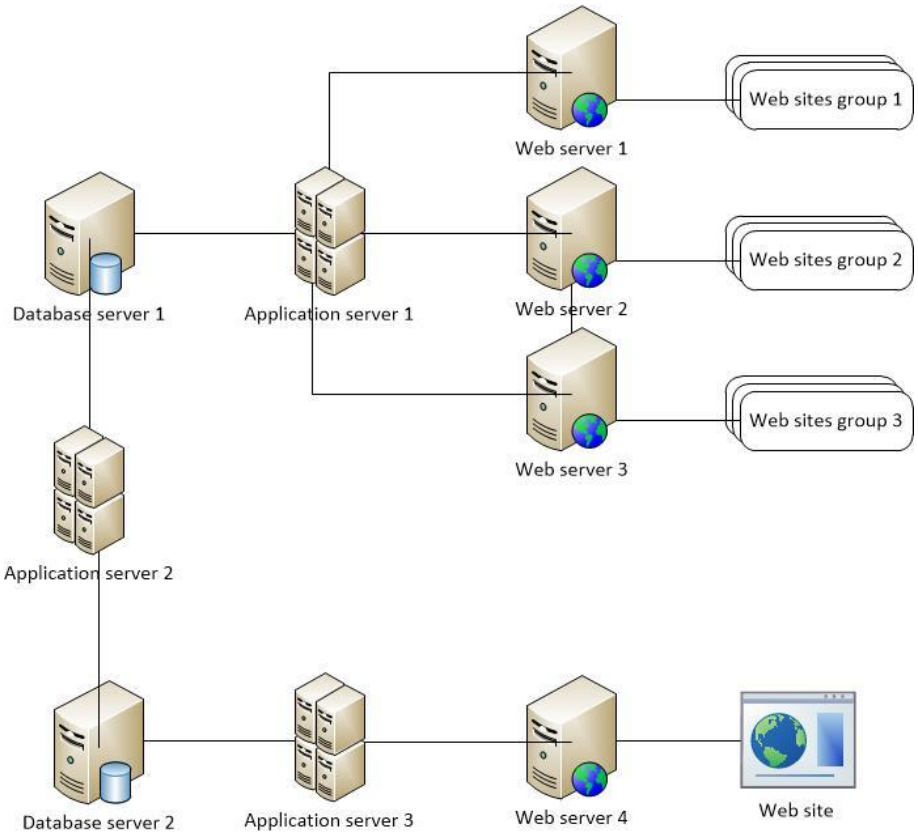

Figure 2. Architecture of software for media monitoring

ANALYZE AND RETRIEVAL OF WEB CONTENT

For the purposes of our study, the retrieval of the information must be appropriately combined with a suitable analyze, such as to enable the presence of the extracted contents in a structured way. There are many different information sources with their distinctive designs and ways of presenting information, but **Figure 3** illustrates the general scheme of a media website that consists of a home page and subpages with different elements. Information unit in this case is the publication with its details. These details are usually the author, title and text of the publication, the date it was published on the website, photo or galleries of photos, etc. The aim of the module for analysis and retrieval of web content consists mainly to recognize individual elements of publications, so that subsequently they may be indexed and processed separately. The algorithm for analyze and retrieval of web content can be described in the following steps:

1. Request to the home page of the information source with publications;
2. Determination of the menu with links to categories of publications;
3. Defining a menu with links to subcategories, if any;
4. Determination of the page area with a list of links to publications;
5. Request to detailed page of a publication in the list;
6. Determination of the base area of the publication within the page;
7. Determination of the area of each of the essential elements of the publication;
8. Retrieval of individual contents in the form of HTML code;
9. Redefinition of the rules for the contents of this information source.

This model does not know in advance the addresses (links) of the pages with publications. For this reason, the initial starting point is always the home page of the web site. The next item that interests us is the location of the menu with the

categories of publications. Categorization itself is not a focus of this paper. What is important for us is the proper distinction between information structures. Following the steps of the presented algorithm lead us to the extraction of web content that forms one publication. Ways in which these contents could be represented in HTML form are very indefinite. For this reason we actually need a specific approach. Our system should implement a basic functionality that is common to all the detailed pages of publications. This implementation, however, provides methods for setup of specific web content for each component separately extracted. The use of these methods is not required. In many cases, the base implementation should be sufficient for a proper interpretation of the HTML code to the desired structure. Additional functionality will be used when there is a need for correcting errors in incorrectly structured HTML, clearing up some of the code tags, removing unnecessary fragments extracted from HTML (as some banners), ignoring JavaScript. The presented approach can be defined as an object-oriented, because of the potential to redefine the way of extracting the resulting text from web content specific to object-oriented platforms that use the techniques of inheritance.

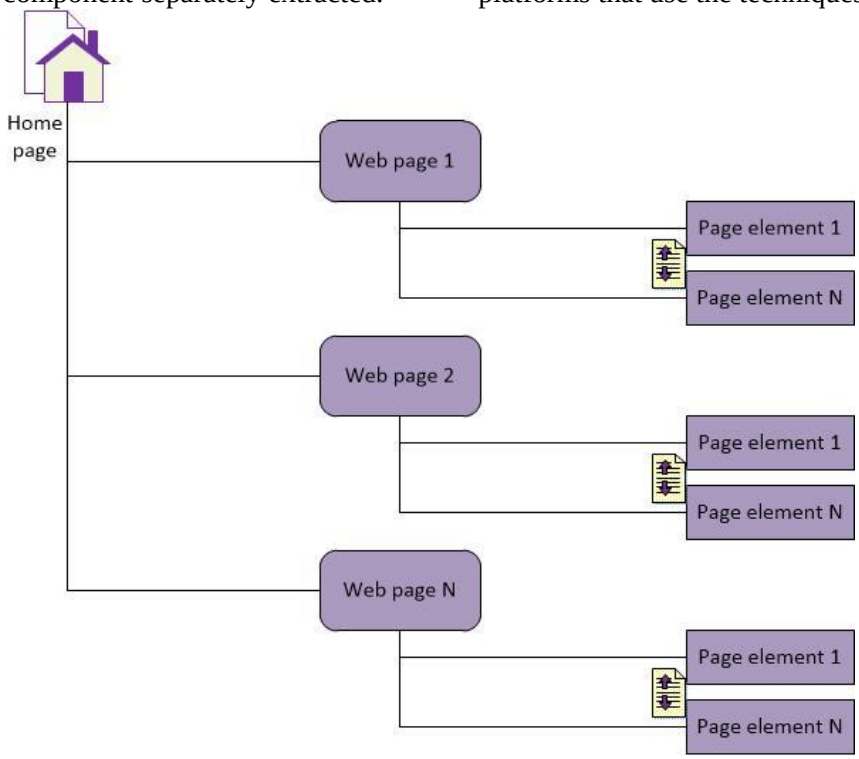

Figure 3. Structure of media website

IMPLEMENTATION OF APPLICATION SERVER FOR DATABASE SEARCH

This article argues that for the implementation of enough quickly and independently search within indexed publications is necessary to realize an application server that serves these processes demand. It would perform some specific tasks related to access to data in the database and the subsequent processing. This scheme is illustrated graphically in **Figure 4**. In the present study the application server is based on java technology and therefore will call it J2EE search service. This java based application server submits the necessary requests to the database management system. This paper examines several key benefits to which such conversion would result:

- Execution of a series of queries to the database in the background. The idea of the executioin of queries in the background is related to the scenario in which a user of the system performs the search in the database, resulting in a significant time to perform. It will be appropriate that the j2ee search service is engaged with management of

these applications, while the web part of media monitoring software provides opportunities for other actions. In turn, the java based application server could also contribute on the performance of search. The application server could perform queries to the database in several different threads. Presumably, this would lead to tangible benefits in the presence of sufficiently strong system resources;

- Implementation of multithreaded search in the database using synchronized requests. The idea of the implementation of multithreaded synchronized search is determined by the ability of application server to perform queries to the database in a synchronized manner. This particular situation would be one in which each user has a resource to perform searches and the j2ee search service is one that maintains separate quees for user query execution in separate threads. Then the time for a consumer demand will be almost independent of other consumer searches at this time.

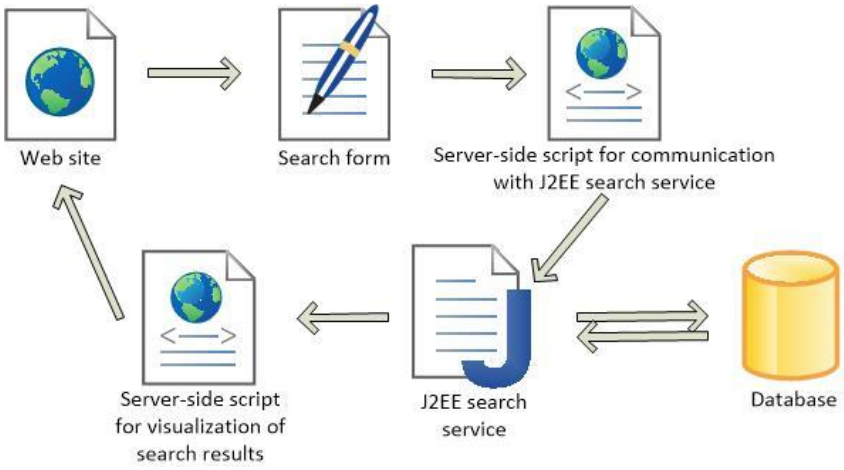

Figure 4. Algorithm for web search

CONCLUSIONS

The model for analyze and retrieval of web content presented in the article can be characterized by the following potential advantages:

- Ability to extract structured information from different sources in a defined general way;
- Programming of rules for behavior in analyze of web content that is used for various web design structures;
- Methods for redefining certain areas of the web pages for specific treatment of HTML fragments.

The J2EE search service presented in the paper has the following advantages:

- Options for removing a portion of the specific business logic of web based system at the application server that is self-encapsulated java application;
- Execution of complex user queries in the background portion of the application, which enables the users to work with other parts of the web system at the same time;
- Increase in the security of the work with the database because of the presence of the application server in the role of a middle layer.

REFERENCES

1. Flaounas I., Turchi M., De Bie T. and Cristianini, N., Inference and Validation of Networks, ECML/PKDD, Bled, Slovenia, Springer, pp. 344-358, 2009.
2. Godbole N., Srinivasaiah, M. and Skiena, S., Large-Scale Sentiment Analysis for News and Blogs, Int. Conf. on Weblogs and Social Media (ICWSM 2007), Denver CO, March 26–28, 2007.
3.http://en.wikipedia.org/wiki/Media_monitoring