Use of Computer Search Algorithms in the Research of Statistical, Semantic and Contextual Rules of Language in Digital Information Space

Zoran Ž. Avramović, Dražen Marinković, Igor Lastrić

Pan-European University Apeiron, Banja Luka, the Republic of Srpska, Bosnia and Herzegovina

zoran.z.avramovic@apeiron-edu.eu

drazen.m.marinkovic@apeiron-edu.eu

igor.t.lastric@apeiron-edu.eu

A general survey

DOI: **10.7251/JIT1701005A** UDC: **004.7:81'33**

Abstract: This paper will discuss and practically explore the interdependence between information technology and linguistics in the modern information society.

The relationship between information technology and linguistics, which has opened new opportunities in linguistic research, will be practically seen in the application of linguistic engineering in researching rules of language.

The aim of this paper is to extend knowledge about the possibilities of application of information technologies in researching rules of language, as well as emphasizing the importance that language technologies have in the field of linguistic research, preservation of language and culture and national identity.

Keywords: information technology, search algorithm, rules of language, linguistic engineering, digital information space.

Introduction

This paper explores and shows rules of language through finding linguistic data in the digital information space, using different methods based on the originally developed, programmed search algorithms in procedural and object-oriented languages that use the system to manage databases and relational model for final realization.

The subject of this paper is to explore the methods of finding, processing and presentation of data in digital libraries and the possibilities of improving the quality of that finding, by finding a faster search query response.

Comparing different methods of information retrieval in digital libraries, the paper starts from the basic, traditional methods of finding information and determining their effectiveness.

The objective of this paper is to determine whether and how advanced software methods help in finding quality linguistic data and their analysis.

The novel "The Bridge on the Drina" by the Nobel Prize winner Ivo Andric (Serbian language, Cyrillic, Ekavian dialect) is taken as the basis for the digital object record on which the research and the analysis of search results and extraction of data, information and knowledge will be carried out. [3]

Algorithms, Languages and Search Programs

An algorithm is a list of steps to follow in order to solve a problem, without ambiguity and vagueness. It relates to the principles of determinacy and finality. In computing environment, algorithms are implemented in a programming language.

In computer science, a search algorithm is an algorithm that retrieves information stored within some data structure.

The basic search algorithms are simple to implement and are suitable for search of static data sets (regulated and unregulated table). This group includes: sequential search and binary search.

Sequential search algorithm is the simplest to implement, but its performance is not so good. It starts to check each item in order to determine whether that item is the one we are looking for. If so, the search is successfully completed, otherwise it goes on, until it finds the desired item or until it examines all items.

In preparing this paper, we have primarily used Visual Basic, a programming language intuitively designed, based on events as an integral part of the programming system from Microsoft, designed to enable rapid application development - RAD with a graphical user interface - GUI and to communicate with databases (such as the DAO, RDO or ADO) and creating ActiveX controls and objects.

In our advanced analysis, processing and research, general purpose programming language C++ has found its application. It is an object-oriented language which provides facilities for low-level memory manipulation. It is designed to support system programming, especially in the case of limited resources, but it is widely used with large systems designed for efficient and flexible operation.

C++ programming language was initially standardized in 1998 by the International Organization for Standardization (ISO). Before the initial standardization, C++ was developed by Bjarne Stroustrup, as an extension of the C language as he wanted an efficient and flexible language similar to C. In this paper, the latest standard version, ratified and published by ISO in December 2014 as ISO/IEC 14882:2014 (informally known as C++14), has found its application.

Ivo Andric - "The Bridge on the Drina"

"The Bridge on the Drina" is a historical novel by Ivo Andric, which, among other literary works, won the Nobel Prize for Literature in 1961.

The novel tells the story of the bridge, Visegrad and human destinies associated with it and spans about four centuries. The bridge stands as a silent witness to history from its construction by the Ottomans in the mid-16th century until its partial destruction during World War I. The bridge is an object around which the destinies of Visegrad inhabitants intertwine, of Muslims, Orthodox Christians, Jews and immigrant Catholics, who play dramatic roles in a large theater of history.

Ivo Andric wrote this novel during World War II when he moved to Belgrade in 1944, after he had worked as an ambassador in Berlin.

Inspiration for the novel Andric found in his own life - he spent his childhood in Visegrad with his aunt who had raised him after his mother lost revenue when his father died. He finished primary school in Visegrad where he was looking at the impressive bridge on the Drina.

The Serbian language in Andric's novel is on one hand seemingly simple and easy to understand, but on the other hand we can see that every word in it is meticulously measured and harmoniously blended. Many think the language is so special, calling it Andric's – Ekavian dialect with syntax which is sometimes characteristic of English language as well as of Bosnian language. In fact, this is the language that Andric learned in his childhood, the true language of that region.

In this literary work we find many examples of folk wisdom in stories, poems and legends as well as in customs and beliefs. Significant attention is given to descriptions, both external and psychological.

Apparently, the novel does not have a theme and strict storyline, something that would link the story with other stories, following one after another, but there is the bridge on the Drina and it is the binding element. It symbolizes the strength and permanence, continuity and consistency despite all the disasters that threaten it. All inhabitants of Visegrad and the surrounding area are connected to this bridge. Compared to it, life expectancy is short and insignificant, so the writer points it out only in carefully chosen moments of human misery.

Search Results Display

The analysis of the text in the novel "The Bridge on the Drina" by Ivo Andric (without title and text on the cover and inner cover page) showed that the total number of characters in it is 532,109 and the following is determined (Table 1):

Table 1. *Evidence-based analysis of the novel*

No.	What	How much
1	Total number of characters in the novel	514,558
2	Total number of punctuation	17,448
3	Total number of numeric digits	103
4	Total number of words in the novel	115,395
5	Total number of sentences in the novel	6,011

Study of vowel representation is shown in Table 2 and Figure 1

Table 2. *Number and frequency of occurrence of vowels in the novel*

Vowel (uppercase and lowercase letter)	Number of occurrences	Percentage of the vowels	Percentage of the most present	Relative frequency of all letters
A	62,837	26.668	100.000	3.66
I	51,404	21.816	81.805	3.00
O	50,985	21.638	81.139	2.97
E	47,547	20.179	75.667	2.77
U	22,855	9.699	36.372	1.33
Total	**235,628**	**100.000**		

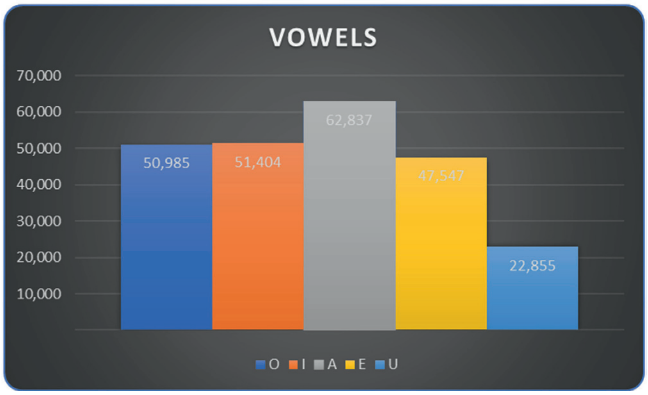

Figure 1. *Histogram of distribution of the total number of vowels in the novel*

The number and relative abundance of uppercase and lowercase letters among vowels are given in Table 3.

Table 3. *Number and relative abundance of uppercase and lowercase letters among vowels*

No.	Vowel	As uppercase letter	As lowercase letter	Letter case percentage ratio
1	A	827	62,010	1.333
2	I	554	50,850	1.089
3	O	525	50,460	1.040
4	E	51	47,496	0.107
5	U	340	22,515	1.510

It is noted that the vowel "e" rarely appears at the beginning of a sentence and/or at the beginning of proper nouns.

The research of the word length in the novel according to number of letters has given interesting results (Table 4 and Figure 2).

Table 4. *Number of words in the novel according to number of letters*

Word length	1	2	3	4	5	6	7	
Number of occurrences	12,473	22,298	10,804	16,905	14,729	13,719	9,506	
Word length	8	9	10	11	12	13	14	15
Number of occurrences	6,664	4,162	2,201	1,212	475	201	68	41
Word length	16	17	18	19	20	21	22	23
Number of occurrences	16	6	4	3	1	0	0	1

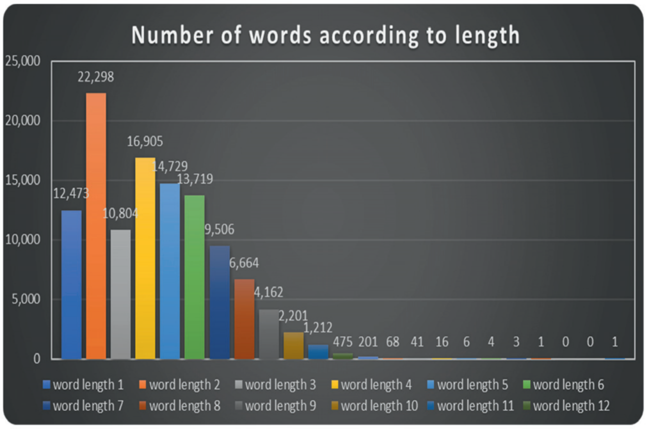

Figure 2. *Histogram of number of words in the novel according to number of letters*

Computer analysis of the text in the novel "The Bridge on the Drina" by Ivo Andric has found that the average length of sentences is 19.20 words.

In this literary work, the average number of letters in the sentence is 85.61.

The longest word in observed and analyzed text has 23 letters.

In this paper we have also researched the presence of the consonants in the novel. For example, in alphabetical order, the first two of them appears: B - 7,881 and V - 19,380 times.

Conclusion

Based on the conducted research and analysis, we have concluded that the advanced search methods help in finding better data quality and thus the information. This was evident in comparing the efficiency and accuracy of finding the data with the use of advanced methods compared to conventional methods.

We can say that finding information using computational methods gives immeasurably better results compared to data retrieval that relies only on the classical methods. We want to emphasize how important is the presence and use of digital resources of analyzed works, especially morphological dictionaries (which are of great importance for morphologically rich languages like Serbian language).

The main focus of our paper was on computer search and finding data because digital libraries offer better and broader search options – through the full text of documents contained in them and also through metadata that describes documents. We believe that a digital library should be much more than a collection of documents available in digital form. We partly focused the subject of our research towards desire to determine whether the present digital libraries can respond to such requests.

We emphasize that the field of separation of concerns, processing and presentation of data, as a subfield of field of natural languages, largely depends on description of observed natural language. Though this field is quite advanced in some languages, such as English, we note that in Slavic languages, especially Serbian language, it is still relatively new.

References:

[1] Aleksandra S. Trtovac, Metadata descriptors and content descriptors for finding information in digital libraries, doctoral dissertation, University of Belgrade, Faculty of Philology, Belgrade, 2016.

[2] Stasa I. Vujicic Stankovic, Ontology-guided Information Extraction (model for Serbian language), doctoral dissertation, University of Belgrade, Faculty of Mathematics, Belgrade, 2016.

[3] Ivo Andric, The Bridge on the Drina, published by the Institute for Textbooks and Teaching Aids, Belgrade, 2012. (leter: Cyrillic, dialect: ekavian, number of pages: 376)

[4] Mladenovic, M. i Mitrovic, J. Semantic Networks for Serbian: New Functionalities of Developing and Maintaining a WordNet Tool. Natural Language Processing for Serbian: Resources and Applications (pages 1-11), Belgrade, University of Belgrade, 2014.

[5] Mitar Pesikan, Jovan Jerkovic, Mato Pizurica, Pravopis srpskoga jezika, Third Edition, Matica Srpska, Novi Sad, 2016.

[6] Donald Erwin Knuth, *The Art of Computer Programming Volume 3: Sorting and Searching third edition*, 1998, Addison-Wesley Professional. ISBN978-0-201-48541-7.

[7] Branimir Covic, Milos Crnjanski's Migrations in the context of modern Russian historical novel, Belgrade, 2001 (S-20287)

Submitted: May 25, 2017.
Accepted: June 7, 2017.