An Advanced Quick-Answering System Intended for the e-Government Service in the Republic of Serbia

Slobodan Nedeljković

Ministry of Interior Republic of Serbia, KnezaMiloša 101, 11000 Belgrade, Serbia, vojkan.nikolic@mup.gov.rs

Vojkan Nikolić

Ministry of Interior Republic of Serbia, KnezaMiloša 101, 11000 Belgrade, Serbia, vojkan.nikolic@mup.gov.rs &
Academy of Criminalistic and Police Studies, Cara Dušana 196, 11080 Zemun, Serbia, vojkan.nikolic@kpa.edu.rs

Milan Čabarkapa

School of Electrical Engineering, University of Belgrade, Bulevarkralja Aleksandra 73,11000 Belgrade, Serbia, ca.milan@etf.bg.ac.rs

Jelena Mišić

Faculty of Electronic Engineering, University of Nis, Aleksandra Medvedeva 14, 18000 Nis, Serbia, jelena.misic@kpa.edu.rs

Dragan Ranđelović

Academy of Criminalistic and Police Studies, Cara Dušana 196, 11080 Zemun, Serbia, dragan.randjelovic@kpa.edu.rs

Abstract: Many of the services incorporated in the e-Government of the Republic of Serbia need a quick-answer system to meet the continually increasing demands of the citizens for easy, fast and effective obtaining of the requested information. However, the public administration of the Republic of Serbia contains a significant amount of unstructured data

arranged in the documents. Thus, it is necessary to provide an automatic classification system based on the principle query-document. The question-answering (Q&A) system related to the Crime domain of the e-Government service of the Republic of Serbia represents a system for achieving the quick replies on citizens' questions. The Q&A system is based on the data mining, text mining, natural language processing, question answer, Bag of Words and N-gram analysis. A similarity measure (distance) is a significant parameter of the Q&A system due to its direct impact on searching speed and distance from wanted documents. Here, three most commonly used similarity measures are used: Cos, Jaccard and Euclid. The primary goal is to determine the similarity measure which provides the most precise results in the crime domain, and that similarity measure is used as a referent one. Due to the high importance of a similarity measure, we use the above three similarity measures, in the process of selecting the most appropriate similarity measure. The selection of the similarity measure is performed using the principles of redundancy and fault tolerance. Specifically, the principle of triple modular redundancy (TMR) with one voter is used. The proposed system is verified by the experiments with real citizen queries. The results show that the proposed system achieves good performance.

Keywords: e-Government; text mining; redundancy; TMR; unstructured documents

1 Introduction

To meet the increasing demand of citizens, for the easy, fast and effective obtaining of needed documents, it is necessary to equip the e-Government services with a quick-answer system. Public administration of the Republic of Serbia disposes of a large number of unstructured documents. These documents are mostly text type documents, so required answers are usually within these documents. To obtain an adequate reply it is necessary to provide an automatic mapping of relevant documents, i.e., an automatic classification strategy, a query – a relevant document. The mentioned strategy is incorporated by the Q&A system (Crime Domain) for e-Government services of the Republic of Serbia [3]. This system provides a quick reply on the asked question (query) achieving the faster and more effective searching and obtaining of wanted answers. The Q&A system is based on the principles of data mining, text mining, natural language processing, question answer, Bag of Words and N-gram analysis.

A similarity measure (a distance) is a crucial parameter in the Q&A system because it is directly correlated to the speed of answering and distance from wanted documents. Here, we analyze three most commonly used similarity measures: Cos, Jaccard and Euclid, with the aim to select the similarity measure which gives the most precise results in the Criminal domain, and that similarity measure is used as a reference one. Since the selection of the similarity measure is very important, our goal is to improve its selection. Namely, to increase the precision in getting a correct answer on the asked question, it is necessary to increase the similarity measure, which can be achieved if all three similarity measures are used in the selection of the most appropriate similarity measure.

Therefore, to achieve the mentioned goal, the principle of the redundancy and fault tolerance are employed. Since there are three similarity measure, i.e., three algorithms which provide the values of similarity measures as information which is further used in processing, such a redundancy represents the information redundancy. Thus, three values are obtained, but the Q&A system needs only one value, for the further operation. To determine which of the obtained similarity measures id the most appropriate one, the principle of the triple modular redundancy (TMR) with one voter is used, where the best result is determined by a voting logic.

The method which authors have proposed in this paper, is based on the example of one method given from another author [3]. It could be applied in the same way or in other methods.

The paper is organized as follows. In Section 2, the related works are presented. In Section 3, the basic theoretical principles of the redundancy and fault-tolerant systems are introduced, and a triple modular redundancy is presented. Section 4 offers the proposed Q&A system related to Crime domain and intended for the e-Government services of the Republic of Serbia is explained in detail, and its possibilities are listed. The proposed TMR system which determines the best similarity measure for a specific query (question) for the Q&A system is introduced in Section 5. The experimental results are given in Section 6. Lastly, the paper is concluded in Section 7.

2 Related Works

Considering the existing solutions for present e-Government problems, and analyzing and identifying the problems in the e-Government solutions in the Republic of Serbia, Šimić et al. [1] proposed a hybrid solution which represents a multi-layer e-document clustering based on a fuzzy concept and a usage of different measures of text similarity. The main aim was to reduce the response time of public administrations with a minimum civil clerks' involvement. To solve this issues, the authors introduced a new approach to facilitate the optimization and automation of advanced methods and techniques for information retrieving. This paper presents the ADVanced ANSwering Engine solution (ADVANSE) for wide-range e-Government services. The most important contributions of the ADVANSE project related to the e-Government services quality, quick response to the citizens' requests, and innovative use of the available content and restructured relationship between civil clerks and citizens. In particular, the accent was on response efficiency and flexibility. Namely, the authors focused on testing under different conditions and improving the ability of adaptation in the next research phases. One of the objectives of mentioned work was to find solutions for the functioning of such a system in multilingual environments and to increase the content complexity regarding the grammar and dictionaries of different languages regardless of the area of use. Consequently, different strategies were proposed.

A particular challenge was the functioning of e-Government services in different domains. Namely, different domains can use the special dictionaries, so it is necessary to use the specialized techniques to find a similarity. Besides, qualitative improvement of a given document processing is required; thus a possible solution can be the tagging of certain document parts instead of the entire document labelling.

Marovac et al. [2] analyzed texts written in different languages and according to different linguistic rules. The texts written in the Serbian language demand complex analysis because it can be written using has two alphabets, Cyrillic and Latin. Moreover, the Serbian language has very rich morphology. Therefore, the use of linguistic resources (corpus of contemporary Serbian language, morphological dictionaries, stop-words, a dictionary of abbreviations, etc.) intended for a qualitative analysis of a natural language, has become a considerable challenge.

The use of N-gram analysis achieved significant results without using the extensive lexical content or analyzing the texts written in Serbian without using the morphological vocabulary. Recently, special attention has been paid to the algorithm for keywords extraction (the N-gram) which is explained in detail in the following sections.

The Authors are of the opinion that an algorithm should be developed, such that, to cluster keywords according to their frequencies, in the text, text parts or clustering keywords (the N-gram), by separating the keywords that are frequent from the less frequent ones.

In V. Nikolić et al. [3], an approach for e-Government services intended for an automatic finding of the required answers or documents on online citizen's requests is presented. The authors described a method to overcome the problems caused by natural language processing tasks in the Serbian language and introduced the sentiment analysis as a special tool for text classification. The document pre-processing presented in this article included changing of a document format, removing the redundant and informal character, and structuring of the documents by the corresponding rules of the next step where the normalization is conducted. The used text normalization is based on the transformation of text into another form suitable for the computer processing. The results achieved by the proposed approach are very satisfactory.

The web-based framework for searching the Web content written in Serbian language, named the SEFRA, was proposed in M. Jovanović et al. [4]. The proposed SEFRA represents a hybrid solution that serves as a platform for a new search application or is used as a service for already existing applications. The SEFRA solves the indexing, searching, and displaying of search results which are adjusted to Serbian. Besides, these framework merges several web-based technologies and services for improving the e-Government citizen's services and other public-sector services. Moreover, SEFRA can be used in the administration of private companies solving the specific searching problems. Although the

SEFRA was developed for the Serbian language primarily, it can also be used for any other language containing the language morphology service. Furthermore, the SEFRA was optimized from both backend and front-end web perspective. The application of the SEFRA was validated by searching the crime law documents of Serbia, and good results were achieved.

In V. Nikolić et al. [5], the problems with large amounts of data, due to numerous implemented e-Government services, in the Serbian government was explained and a suitable solution was presented. The main problem is the specific data and information extraction from the variety of existing text documents which are usually in a format prepared for print (HTML, PDF and Microsoft Word formats). As a solution for that problem authors proposed an application that includes Lucene library, which is a specialized library for implementation of the indexing and searching over a significant amount of data. The proposed application provides a quick search within the unstructured text documents written in the Serbian language which further leads to an efficient detection and processing of criminal offenses and increases the security level of the Republic of Serbia. The proposed application is verified by searching the data and documents within the unstructured crime documents written in the Serbian language that aims to find the elements of a crime in the cyberspace. The obtained results showed that the proposed application accelerated the searching procedure significantly.

When it comes to the Serbian language in literature, except the papers of co-authors of this paper [1] [3] [4], very rare attempts are made to construct and describe the System for information retrieval i.e. Q&A system. One of these is the building of the Information Retrieval System for Serbian - Challenges and Solutions by M. Martinović et al. [6], using Natural language processing (NLP) technology as an area of computer science and artificial intelligence. In this article techniques, achieving state-of-the-art results in many natural language tasks, for example in language modelling, parsing, and many others implemented in the Serbian Information Retrieval System (SPRETS).

In article N. Milošević [7], one realized application was presented, which is, in fact, Stemmer for Serbian language. In this article is presented suffix-stripping stemmer for Serbian language, one of the highly inflectional languages. Stemming application is designed as a web application. It uses PHP script for backend and AJAX interaction with backend side.

The general effects of data redundancy have been noticed by many researchers in the field of Q&A systems, for example Lin in [8]. The redundancy-based approach has been developed as an alternative to the traditional ontology-driven knowledge-based techniques. This approach have basis in the philosophy of "data is all that matters" i.e., just give enough data and system could simply count instances and derive answers from these observations. This approach which solved problems in language processing was demonstrated in a paper written by Banko and Brill [9].

In Light et al. [10] is discussed a correlation between the number of times an answer appeared in the Text Retrieval Conference (TREC) corpus and the average performance of TREC systems on that particular question. That is, systems preferred to perform better on questions that had multiple answer instances within the corpus. Also, Clarke et al. [11] pointed an upward trend in Q&A system accuracy as a system was given even more text from which to extract answers and thereby holding everything else constant.

From standpoint of significance of using basic principles of TMR to constrict the proposed algorithm for optimization, one Q&A system and especially it's part named, voting system, it is important to notice different approaches of its implementation, like for example, using genetic algorithm [12], Bayesian techniques [13], Markov modeling [14], neural networks [15] etc.

In [16] Gruzenkin et al. Considered one compensation model of multi-attribute decision making and its application to N-version software choice.

3 Theoretical Background

With the development of IT technology, complex and sensitive processes have become automated which has a high demand for proper operation of the implemented system. This is provided by using the VLSI technology that enables practical adoption of many methods for mitigation and reduction of system faults. The fault-tolerant systems represent the high-confidential systems which even in unfavorable conditions operates in a proper way providing all functionalities due to the ability to tolerate single faults. These kinds of systems have the advantages of high reliability, availability, security, stability, manageability and serviceability [17, 18, 19, 20, 21].

In the fault-tolerant systems, one of the common methods to enhance system reliability is to use a redundancy, which represents the addition of resources, (amount of information and time) to normalize system operation. The redundancy can be related to the hardware, software, information and time.

As a basis for the considered Q & A system in [3], Text retrieved. Text retrieval is a branch of information retrieval where the information is stored primarily in the form of text. In the concrete case, it is about textual information relating to the members of the Criminal Code of the Republic of Serbia and they are written in Cyrillic and Latin scripts.

In order to optimize the existing Q&A system: Q&A system for e-Government services in the Republic of Serbia, it can be used as a basic idea, the idea TMR which uses three functionally equivalent units to provide redundant backup, regardless of type of redundancy i.e. whether it is viewed as an active hardware or corresponding to its three programming software redundancy or three different

sources of information; especially possibly application of different approaches in voting system as its obligatory part. This idea through presentation of basic content of most important titles necessary for understanding this idea authors have done in this paper, as described in the following sub-sections.

3.1 Hardware Redundancy

Hardware redundancy represents the addition of additional hardware to normalize system operation. This type of redundancy is usually employed when it is necessary to detect system failures or tolerate them to make the system robust to the failures. Hardware redundancy can be active, passive and hybrid. [22]

Passive hardware redundancy denotes the technique of fault masking with the aim to prevent a fault to cause errors.

Active hardware redundancy denotes the adoption of techniques that enhance fault toleration, by detecting the fault and repairing of faulty (inoperative) hardware. It is based on fault detection and localization within the system and its repairing.

Hybrid hardware redundancy denotes the combination of the previous two redundancies by via their advantages. The technique of fault masking is used to prevent errors, and when this technique does not achieve good results, one of the active techniques is used to locate and eliminate the fault.

3.1.1 Passive Hardware Redundancy

In the implementation of passive hardware redundancy, a voting mechanism such as the principle of majority voting is used to mask a fault. In passive hardware redundancy, the techniques that provide fault tolerance without a need to detect and repair the fault are used.

One of the most commonly used passive hardware redundancies is a triple modular redundancy (TMR) whose basic principle is presented in Fig. 3.1, wherein, it can be seen that quantity in the used hardware is tripled, and the principle of majority voting is employed. Namely, if one module stops working properly, the rest two modules will mask its fault mitigating the errors. [23]

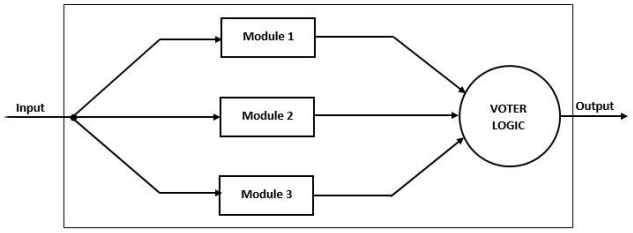

Figure 3.1
Passive hardware redundancy block diagram

The main problem in the TMR is the voter. In the case of its failure, the entire system work will be influenced. Therefore, the reliability of the simplest TMR is of the same level as the reliability of the voter. To overcome voter failure, the voter tripling technique can be used which will provide three independent outputs. [24]

In the TMR, special attention is paid to the voting techniques used by a voter. The voting process includes many problems. The first one is a decision whether to use the hardware voter or to conduct the voting process on a software level. The second problem relates to the practical realization of voting; for instance, three results of the TMR do not agree in total even when there is no any failure or fault. The processing of this disagreement can make it even more significant. Finally, the majority voter can decide that there are no two results in the TMR system that are in agreement, although the systems work correctly.

3.1.2 Active Hardware Redundancy

Active hardware redundancy tries to achieve fault-tolerance by fault detection, localization, and repair. Active hardware redundancy does not use fault masking technique, and it is used in the systems that can tolerate temporary, incorrect results under the condition that system is reconfigured and that its operation is stabilized in the satisfactory period.

3.1.2 Hybrid Hardware Redundancy

As already mentioned, the hybrid redundancy represents the combination of active and passive redundancy. The fault masking is used to prevent errors, and detection, localization, and repair od a fault are used for system reconfiguration in the case of system failure.

3.2 Q&A System for e-Government Services of the Republic of Serbia

The Q&A system (Crime domain) for e-Government services in the Republic of Serbia represents a quick-answer system which provides faster and more effective searching and obtaining of the answers on citizen's questions. [3] The Q&A system includes the following:

- Profound data analysis (Eng. Data mining - DM) which represents a tool for analysis of huge amount of data which constantly increases.
- Profound text analysis (Eng. Text mining – TM) which is used to reject the unneeded data to identify the data the citizen asked for.
- NLP which represents a set of techniques and methods for automatic generation of texts in natural language.

- N-gram analysis is used to overcome issues related to the lexical resources of Serbian language and which can provide satisfactory results. Namely, the Serbian language has a lot of language rules, and exceptions to these rules, so significant lexical resources are needed to analyze the documents properly.
- Apache Lucene represent a stable searching library which denotes the basis for the development of searching applications; also, it can provide a search within the created indexes and achieve good results for specific queries.

The working principle of the Q&A system for the e-Government services of the Republic of Serbia based on the Bag of Concept (BoC) model is presented in [3].

The Q&A system includes a framework for web services, a quick-answer model for eGovernment, a BoC based sub-system, an algorithm for classification of queries related to the criminal laws Criminal Code of the Republic of Serbia.

The main task of the Q&A system is a comparison of a short text represented as a vector with the queries made by citizens which are also represented as vectors. As a short text, we use the clauses of the Criminal Code of the Republic of Serbia in the Criminal domain (inflict of massive injuries).

The Q&A system contains the classification of questions related to the Criminal Code by using a special domain based on the BoC model as well as a comparison of performances of a classification of the based method [3].

In the Q&A system (Crime Domain) for the e-Government services of the Republic of Serbia, first, the mapping of questions from the BoC model defined by 31 terms is performed. Then, the filtering of a stop-word is conducted. Next, the stemming is performed, which denotes the removing of the common affix in words to conduct the morphological normalization and make more general characteristics of words. Therefore, a 4-gram steamer is used which is the most common stemming algorithm for the Serbian language. Finally, the similarity between a query and a BoC representation of three documents using the similarity functions. The Q&A system output is the corresponding document of the made query. The output denotes a message to the user having the following format "Please look at the clause No. n of the Criminal Code"; n will be determined by the Specific Annotator (SA) for stemming.

The pallet of words of the BoC model is defined by 31 terms in Serbian language: "teška, organ, teško, telesna, prouzrokovana, povredi, povreda, nesposobnost, naruši, telesno, povređenog, prinuda, povredi, kazniti, pretnja, trajno, zatvorom, ubistvom, meri, nehata, teškom, oštećen, napadom, telesnom, oslabljen, laka, povredom, tela, telo and posledice".

After the normalization by the 4-gram method from eight most common words, the set of seven words ("delo, učin, kazn, zatv, tele, teškipovr") is obtained by using the tf*idf criteria. Since this set is not enough for further analysis we use the

citizens' questions from the portal "PRO BONO" and daily newspapers "Blic" which contain the tag "physical injury/injuries". Consequently, the set of seven words is enlarged to the set containing 31 words.

The N-gram normalization is performed using the tf*idf criteria (term frequency and inverse document frequency) where tf denotes the number of word appearance in a specific text document, and idf denotes the importance of a given word. If a sum of all documents is divided by the sum of the documents wherein a given word is found, and the logarithmic function is applied, the idf value is obtained. By multiplying tf with idf, the TF-IDF measure is gotten, and it denotes word importance in a document or a set of documents.

In the Q&A system, the parts of the Criminal Code are presented as three different documents grouped in a whole, which shows the currently available knowledge. To complement this centralized repository, the answers by the related Experts are used, and they can be found on the website for free legal help named the PRO BONO [25]. The existing answers to citizens' questions correlated to the above-listed three law clauses of the Criminal Code which are a part of the section about the criminal law are analyzed. From a large number of questions, 45 questions are selected to help the best possible settings of the BoC model. Besides, to find the group of words correlated to the given problem, the Google search is used to find all related links to the term "serious bodily injury" which is the term - the basic lemma, obtained on the basis of an electronic dictionary for the Serbian language, using the available online-language resources "bag of words" [26]. The best results are found on the website of Serbian daily newspapers called the "Blic". From the mentioned website we use 35 text articles which correspond to the given query.

The sop words are the words which do not have any meaning for a given subject. There are certain types of words for which this is completely true (e.g. for conjuncts). In some other cases, this does not hold, so the selection of the stop words depends on the context of documents consideration. If it is needed to group the documents which contain data about current and previous events, then in the stop-word list the adverbs are not included. If the same documents are grouped regarding the meaning, then, the adverbs should be included in the list of the stop words. On the other hand, the nouns and verbs are rarely the stop words even, but that is also a possibility if the considered documents demand that. The standard list of stop words in English counts 600 words, while the SAS has a little fewer words, 330 exactly. Our Q&A system has a stop list containing 700 most commonly used words.

In the Q&A system, the N-gram analysis can be successfully used to find the words with the same root very easily. To select the value of N which will provide the valid results, we collect the extensions of words' forms in the Serbian language, only the words that are meaningful for the analysis are [27]. It is found that the majority of these extensions, about 92.70%, have up to 4 letters [2].

In addition, in [2], a 4-gram analysis was used because it achieves the best results in tasks on Serbian. Hence, we also use a 4-gram analysis for normalization of words in the document. The SA is a software agent which uses a BoC model of a specific domain from the knowledge base as a source of a map for keywords extraction. The SA assign the absence or presence of each extracted term following the relevant keywords within the BoC. The translation of the query and document form a raw data to the form needed for comparison can be done by the computer processing which represents the first obstacle in text similarity calculation. To overcome this problem, the text information is converted to the space-vector form [28].

To achieve good results by using the Q&A system, it is necessary to establish a good correlation between query classification and response type extraction. The goal of the SA is that the system "learns" how to map the corresponding type of answer on the basis of the query [29].

Parameters for supervised learning are set based on the values that have yielded good results in similar text classification processes. The proposed SA-based algorithm is based on the BoC approach with the task of automating the classification. The BoC is a list of all words ranked according to their descriptive value for three clauses of the Criminal Code of the Republic of Serbia (cluster membership). Similarity measures, in this case, precisely use vector representations of documents and questions for calculating the distance between them.

The Q&A system (Crime domain) for the e-Government services of the Republic of Serbia system was implemented using the Apache Lucene. The Apache Lucene is a scalable search library that represents the basis on which the search application is developed and which analyzes and indexes the textual content and provides the search within the created indexes and displays the search results for a particular query. The main task is to make a comparison of a short text, given as a vector, with the queries set by the citizens, which are also presented as vectors. Here, as short texts, certain articles of the Criminal Code of the Republic of Serbia relating to inflicting physical injuries are used.

To determine the threshold of similarity between the question presented in the form of short text and the Articles of the law, which is also shown as a short text, the following formula is used [30]:

$$Similarity = \frac{W(Sa) \cap W(Sb)}{\min(W(Sa), W(SB))} \quad (1)$$

where $W(S_a) \cap W(S_b)$ is an intersection set number of words in questions q_i, and the number of words in rt_j, and $min\ (W(S_a), W(S_b))$ is a value lower than the number of words in both documents.

3.3 Voting System Solutions in TMR and Q&A System

As we mentioned in introduction of this chapter, redundancy is a common approach to improve the reliability and availability of a system and there are various methods, techniques, and terminologies for implementing redundancy in one system.

To use one type of redundancy in one Q&A system whose optimization work through increasing its efficiency is the subject of this paper we used basic principles of TMR which refers to the approach of having multiply modules running in parallel, receive the same input information at the same time and their output values are then compared and a voter decides which output values should be used further in Q&A system.

N-version programming is one type of software redundancy and the well-known software development approach which ensures high dependability and fault tolerance of software. One algorithm have to be considered when choosing an optimal variant of N-version software, which could be and N different types of algorithms which solved same problem i.e. they received same input at the same time with the task to calculate same output but voting system as an obligatory part of this type of redundancy system has the task of selecting the best according to a particular criterion.

Voting algorithms in one TMR, as a special case of a N Modular Redundancy-NMR, play a significant role in most fault-tolerant and control systems so these algorithms are continually in develop and progress and with regard to different systems, new types of voting algorithms which could be de developing like different types of iterative algorithms, Markov modelling, neural networks etc.

For one Q&A system, to make it more efficient, it is necessary to determine which similarity measures will be used before the obligatory clustering of documents having in mind that no similarity measure is universally best for clustering of all types of documents and that this job is obligatory for each Q&A system.

We can use N but it is enough three different similarity measures in one Q&A and applying basic principles of TMR and one from mentioned voting algorithms choose the best from this three different similarity measures and in this way construct one optimized Q&A system in terms of its efficiency.

In the next chapter we will consider one such optimization using one iterative algorithm and three types of commonly used similarity measures which are considered in the Q&A system (Criminal domain) for the e-Government services of the Republic of Serbia and that: Cos, Jaccard, and Euclid, to choose the similarity measure that gives the most accurate results for the crime field.

4 Using Basic Principles of 3-TMR System to Construct One Algorithm for Q&A Systems

For the Q&A system (Criminal domain) for the e-Government services of the Republic of Serbia, it is necessary to determine similarity measures (distances) before the clustering. The similarity measure is very important due to the direct impact on the documents ranking because of its direct impact on the proximity degree or a distance from the target documents. In addition, the measurement of the similarity of documents based on characteristics that depend on the type of data that are in the context of documents and processing leads to grouping and clustering of documents within the cluster. No similarity measure is universally best for clustering of all types of documents.

Selection of an appropriate similarity measure is crucial for cluster analysis, especially for a specific type of clustering algorithms. Three types of commonly used similarity measures are analyzed in the Q&A system (Criminal domain) for the e-Government services of the Republic of Serbia: Cos, Jaccard, and Euclid, to choose the similarity measure that gives the most accurate results for the crime field.

On a specific sample query (usually 10% of the total queries used in the analysis), all three similarity measures are applied, and the one which provides the best result is chosen as the most appropriate one. In order to select the best measure of similarity, the Expert determines the correct answer for each of the queries from the query prompt. Queries are in the form of textual documents in the area of the Criminal Code of the Republic of Serbia.

The results of this analysis determine the reference measure of similarity, one of the three applied similarity measures, and it is taken as a reference measure for further algorithm calculation (Fig. 4.1).

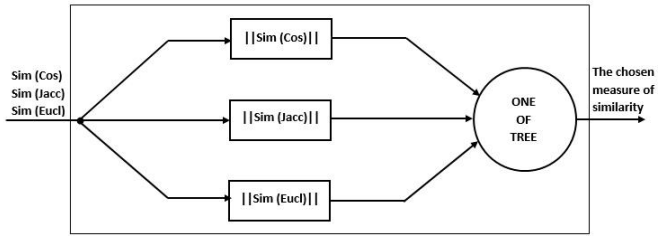

Figure 4.1

The TMR with three similarity measure (Cos, Jaccard and Euclid) and one voter (one of three)

The choice of a similarity measure for the Q&A system for the e-Government services of the Republic of Serbia is of crucial importance for determining the answers that the system delivers to citizens, businesses or employees in the public administration. Their satisfaction with the provided service depends on the quality of the answers they receive from the Q&A system.

For this reason, in this work, we put a focus on the possibilities of improving the selection of similarity measure in the existing Q&A system. The basic idea is that in the selection of a similarity measure all three similarity measures are used.

With the aim to increase the quality of the Q&A system response, we applied the principles of redundancy and fault tolerance of the system. In particular, the principles of triple modular redundancy (TMR) with one voter were applied.

The TMR is characteristic because of the hardware tripling (modules) and because a majority vote is used in determining the system output. The application of the TMR principle to obtain the best measure of similarity for the Q&A system is based on the fact that each module calculates a normalized measure of similarity, respectively: Cos, Jaccard, and Euclid. For all three normalized similarity measures, the same algorithm like module in one TMR system is used. From the point of view of the information redundancy, the TMR for obtaining the best measure of similarity for the Q&A system is an informational redundancy due to the obtaining of more than one information (three information) as needed and sufficient for the Q&A system.

For all three TMR modules, the inputs are the same: $Sim_i(Cos)$, $Sim_i(Jacc.)$ and $Sim_i\ (Eucl.)$, for the query for which the BoC Q&A system already has an answer and for which the Expert confirmed that it is correct.

In the first step, each of the algorithms calculates the normalized value of the input values according to the following formulas:

$$||Sim_i(Cos)|| = \frac{Sim_i(Cos)}{Sim_i(Cos)+Sim_i(Jacc.)+Sim_i(Eucl.)} \tag{2}$$

$$||Sim_i(Jacc.)|| = \frac{Sim_i(Jacc.)}{Sim_i(Cos)+Sim_i(Jacc.)+Sim_i(Eucl.)} \tag{3}$$

$$||Sim_i(Eucl.)|| = \frac{Sim_i(Eucl.)}{Sim_i(Cos)+Sim_i(Jacc.)+Sim_i(Eucl.)} \tag{4}$$

This process is repeated for the next query, i.e., for the following input values for algorithms until the following conditions are met:

$$\frac{\sum_1^i||Sim_i(Cos)||}{i} - \frac{\sum_1^{i-1}||Sim_{i-1}(Cos)||}{i-1} = 0.000 \tag{5}$$

$$\frac{\sum_1^i||Sim_i(Jacc.)||}{i} - \frac{\sum_1^{i-1}||Sim_{i-1}(Jacc.)||}{i-1} = 0.000 \tag{6}$$

$$\frac{\sum_1^i||Sim_i(Eucl.)||}{i} - \frac{\sum_1^{i-1}||Sim_{i-1}(Eucl.)||}{i-1} = 0.000 \tag{7}$$

Accuracy in the subtracting process is taken to three decimal places.

In order to ensure the exit from the algorithm, another criterion was introduced:

$$N < 100$$

This means that if the mean value on the third decimals is not found for up to 100 queries, then the last mean value is taken for further consideration.

This is an iterative algorithm that stops when the conditions are met. The algorithm calculates the mean values for each normalized similarity and their mean values. When the previous mean value is equal to the next one in the first three decimals, then, the algorithm stops because the condition is fulfilled.

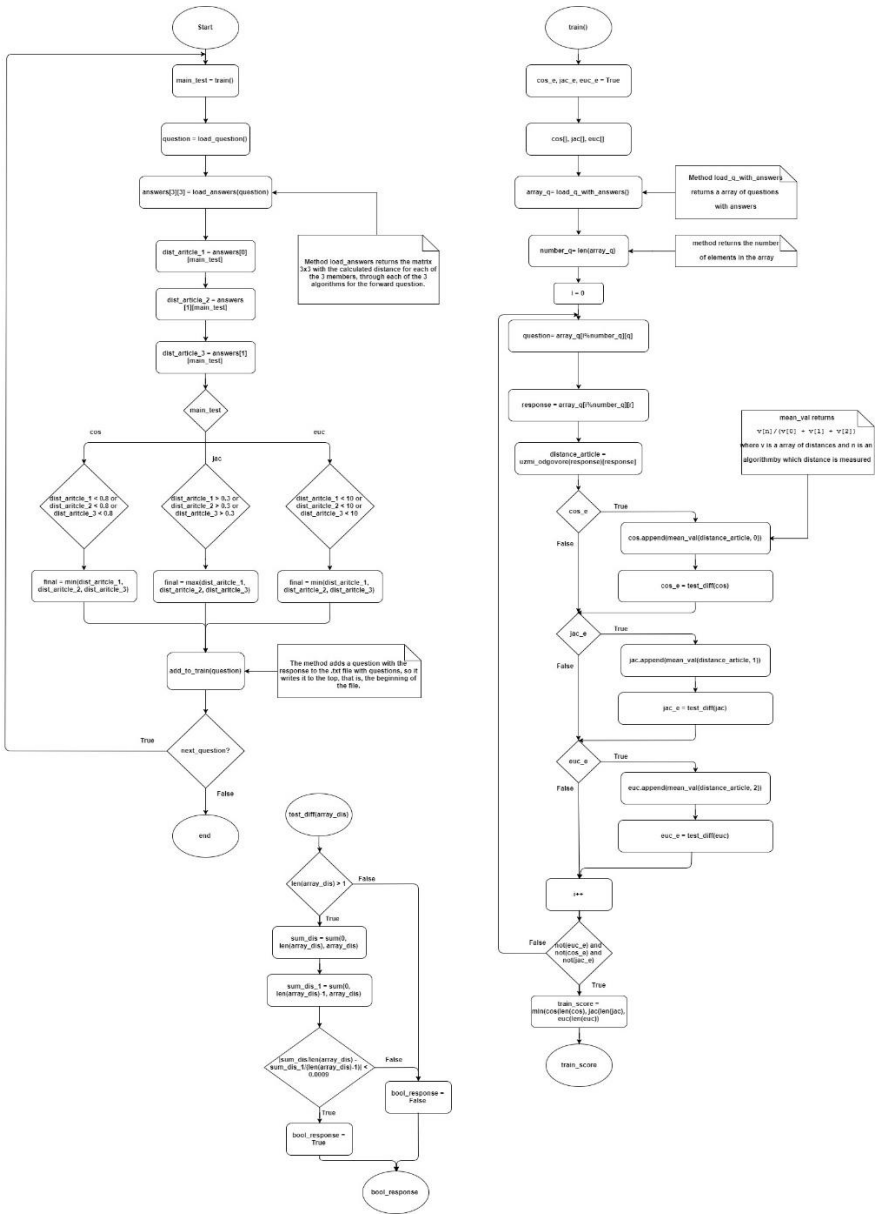

Figure 4.2
Block-diagram of proposed algorithm

To determine the mean distance in each algorithm (TMR module), the following formulas are used:

$$\frac{\sum_1^i ||Sim_i(Cos)||}{i} \quad (8)$$

$$\frac{\sum_1^i ||Sim_i(Jacc.)||}{i} \quad (9)$$

$$\frac{\sum_1^i ||Sim_i(Eucl.)||}{i} \quad (10)$$

The output of each of the algorithms (TMR modules) represents information.

These three distance values denote the three inputs of the TMR voter. The logic of the TMR voter is to determine the minimum value using the three obtained values. The smallest value represents the smallest distance, i.e. the greatest similarity between the two vectors. On the basis of the greatest similarity, the measure that will be used to searching is determined.

5 Experimental Evaluation – Case Study Q&A System for e-Government Services of the Republic of Serbia

In the Republic of Serbia, the laws related to the criminal represent a special kind of laws. One of such a law is Criminal Code of the Republic of Serbia. The current Criminal Code is presented in the "Sl. glasnik RS", br. 85/2005, 88/2005 - ispr., 107/2005 - ispr., 72/2009, 111/2009, 121/2012, 104/2013 i 108/2014) [31].

The Criminal Code of the Republic of Serbia examines the issue of guilt and the provisions of sanctions in relation to the omitted crimes in 36 segments, one of which is related to the physical injuries that are processed from many aspects. According to the possibility to inflict the physical injuries, three segments of the Criminal Code are selected:

1. Massive physical injury, Clause 121
2. Light physical injury, Clause 122
3. Coercion, Clause 135

These concepts are mapped on criminal vocabulary for three laws of Criminal Code, and they are given as vectors in the BoC:

- Clause 121 [kazn,zatv,tešk,post,javn,poku,tuži,pril,delo,tele,povr,nane, sumn,uhap]
- Clause 122 [tele,poli,napa,udar,post,nane,učin,kazn,zatv,kriv,javn,povr, lake,prij,poku,tuži]

- Clause 135 [prij,poli,kriv,post]

For specific relevant measures of similarity, 100 queries are prepared and used, wherein the Expert is determined the correct answer of each query, i.e., to which of the above three laws the query relates to.

The algorithm input consists of three measures of similarity: Sim (Cos), Sim (Jacc.) and Sim (Eucl.) which match the answer the Expert labelled as the correct one for a specific query. The measures of similarity for the first three queries are given in Table 1.

Table 1
The measures of similarities for the first three queries

R. br.	**Sim (Cos)**	**Sim (Jacc.)**	**Sim (Eucl.)**
1.	0.413528	1.000000	15.362291
2.	0.790875	0.900000	3.872983
3.	0.72175	1.000000	5.567764

The first step in any of three algorithms is to calculate the normalized values using the corresponding equations Eq. (2)-(4), as presented in Table 2.

Table 2
The normalized similarities

R. br.	$\|\|Sim\ \ (Cos)\|\|$	$\|\|Sim\ \ (Jacc.)\|\|$	$\|\|Sim\ \ (Eucll.)\|\|$
1.	0,0246502421133657	0.0596096083297036	0.9157401495569310
2.	0.1421450727175280	0.1617582619829620	0.6960966652995100
3.	0.0108698287740294	0.1371833568054060	0.7638045554202930
25.	0.0411644913582571	0.0586313221421774	0.9588354463586860
26.	0.0695500826502071	0.1332351538288670	0.7972147635209250
…	…	END	…
32.	0.0446494071297836		0.9070069585827190
33.	0.0626163530416827		0.8075977536867500
…	…		END
50.	0.0108698287740294		
51.	0.0337292888988060		
	END		

According to the values in Table 2, it can be seen that first stops the loop of Algorithm 2 (module 2) during the processing of query 26. Then, the following condition is satisfied:

$$\frac{\sum_1^{26}||Sim_i(Jacc.)||}{26} - \frac{\sum_1^{25}||Sim_{25}(Jacc.)||}{25} = 0.000 \qquad (11)$$

0.1093293107839940 - 0.1083730770622000 = **0.000**9562337217949

The next one stops the loop of Algorithm 3 (module 3) during the processing of query 33. Then, the following condition is satisfied:

$$\frac{\sum_1^{33}||Sim_i(Eucl.)||}{33} - \frac{\sum_1^{32}||Sim_{32}(Eucl.)||}{32} = 0.000 \quad (12)$$

0.8354245664315350 - 0.8362941543298090 = **-0.000**8695878982745

In the end, stops the loop of Algorithm 1 (module 1) during the processing of query 51. Then, the following condition is satisfied:

$$\frac{\sum_1^{51}||Sim_{51}(Cos)||}{51} - \frac{\sum_1^{50}||Sim_{50}(Cos)||}{50} = 0.000 \quad (13)$$

0.0604216790226415 - 0.0609555268251182 = **-0.000**5338478024767

The last step in all three algorithms is the calculation of a distance using Eq. (9)-(11). The values of *i* for all three algorithms is 26, 33 and 51, respectively.

$$\frac{\sum_1^{26}||Sim_i(Jacc.)||}{26} = 0.1093293107839940 \quad (14)$$

$$\frac{\sum_1^{33}||Sim_i(Eucl.)||}{33} = 0.8075977536867500 \quad (15)$$

$$\frac{\sum_1^{51}||Sim_i(Cos)||}{51} = 0.0604216790226415 \quad (16)$$

The logic of TMR voter is based on the determination of the smallest value in one of three line of constructed TMR because the smallest value represents the smallest distance between two vectors, i.e., the highest similarity.

In this case, the smallest value presents the normalized cosine similarity, Eq. (16).

To verify the validity of the proposed system 1830 queries related to the Criminal Code of the Republic of Serbia were collected. After the processing of collected queries, the proposed system eliminated 270 queries which related to the clauses not included in this experiment. Thus, further processing is continued with the remaining 1560 queries.

The verification results are presented in Table 3. In Table 3, Doc represents the document corresponding to the particular query. The related verification parameters were calculated using Eq. (17)-(20).

Table 3

Evaluation Metrics: Classification View

Doc / Action	**Retrieved**	**Not Retrieved**
Relevant	Relevant Retrieved	Relevant Rejected
Not relevant	Irrelevant Retrieved	Irrelevant Rejected

$$Precision = \frac{Relevant\ Retrieved}{Retrieved} = 49.67\ \% \quad (17)$$

$$Recall = \frac{Relevant\ Retrieved}{Relevant} = 49.67\ \% \quad (18)$$

$$F_{i\ (i=1,n)} = \frac{2*precision_i*recall_i}{precision_i+recall_i} \quad (19)$$

$$F_{Average} = \frac{F_1+F_2+\cdots+F_n}{N} = 49.67\ \% \quad (20)$$

The precision, recall, and F_1 parameter focused on true positives, i.e., the positive examples of the gold standard. In a monolingual alignment, the positive examples denoted the tokens that were aligned, while the negative examples denoted the tokens that were not aligned. Usually, the focus is only on whether those which should have been aligned, are indeed correctly aligned; thus the measure of $F_1$is a good *t*.

Since the correct rejection value, related to the true negatives, is important, the accuracy was computed as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = 74.83\ \% \quad (22)$$

The accuracy weighs the true positives and the true negatives equally. The aim is to ensure that the Classifier recognizes both positive and negative examples. Specifically, in the alignment, where most tokens were not aligned, the accuracy value was likely to be very high in general, and it was difficult to determine the difference. In such a situation, only F1 on positive (aligned) examples was reported.

The result of questions based verification - With the aim to test the proposed algorithm, ten questions were used. The results obtained by the proposed algorithm are given in Table 4. Table 4 also displays the results provided by the Expert in the field of the Criminal Code of the Republic of Serbia.

Table 4
Comparison results of proposed and existing algorithm

	New algorithm	Old algorithm Cos	Old algorithm Euclid	Old algorithm Jaccard
TP	909	877	493	709
TN	921	953	1337	1121
FP	921	953	1337	1121
FN	4569	4537	4153	4369
Precision	0.496721	0.479234973	0.269398907	0.387431694
Recall	0.496721	0.479234973	0.269398907	0.387431694
$F_{Average}$	0.496721	0.479234973	0.269398907	0.387431694
Accuracy	0.748361	0.739617486	0.634699454	0.693715847

Conclusions

This work is focused on the improvement of the existing Q&A system (Crime Domain) within the e-Government services, of the Republic of Serbia, from the aspect of improving the similarity measure, which represents, a significant feature

for proper system operation. Similarity measure determines the similarity of direct influence on speed and distance from the necessary documents; the existing Q&A uses one of three similarity measures: Cos, Jaccard, and Euclid.

A new approach presented herein and validated by experiment, is based on the following principle, in the calculation of a new measure of similarity, all three similarity measures are used to increase the similarity level. Besides, the principles of the redundancy and the fault tolerant system are adopted by employing a triple modulation technique.

Using the described approach and application of the new algorithm, results are clearly better than the application of any measure of similarity individually, which proved to be true in the example of the Q&A system of the Government of the Republic of Serbia, i.e. eGovernment of the Republic of Serbia, where it is in experimental evaluation. In addition, the complete software is publicly available at the website:

https://drive.google.com/open?id=1Ny92N48JURDhhwy6tRCITwS8FH1eYh9E

For the correct operation of the Q&A system used in this experiment, it was necessary to use the Expert forgive opinion, concerning the accuracy of the results. Our future work will focus on an in-expert Q&A system, i.e. A Q&A system for which it will not be necessary to involve an Expert.

References

[1] G. Šimić, Z. Jeremić, E. Kajan, D. Randjelović, A. Presnall: A Framework for Delivering e-Government Support, Acta Polytechnica Hungarica, Vol. 11, No. 1, 2014

[2] U. Marovac, A. Pljasković, A. Crnišanin, E. Kajan: N-gram analysis of text documents in Serbian, TELFOR 2012

[3] V. Nikolić, B. Markoski, K Kuk, D. Randjelović, P. Čisar, Modelling the System of Receiving Quick Answers for e-Government Services: Study for the Crime Domain in the Republic of Serbia, Acta Polytechnica Hungarica

[4] M. Jovanović, G. Šimić, M. Čabarkapa, V. Nikolić, D. Ranđelović, SEFRA - Web-based framework customizable for Serbian language search applications, Paper is accepted in 2017. for publications in Acta Polytechnica Hungarica (aceppted)

[5] V. Nikolić, M. Ivković, S. Nedeljković, P. Djikanović, Information Retrieval for Unstructured Text Documents: Lucene Searching, AIIT 2015

[6] M. Martinović, S. Vesić, G. Rakić, Building an Information Retrieval System for Serbian - Challenges and Solutions, INTERSPEECH 2007

[7] N. Milošević, Stemmer for Serbian language, http://www.inspiratron.org (2018)

[8] Lin, J. 2007. An exploration of the principles underlying redundancy-based factoid question answering. ACM Trans. Inform. Syst. 25, 2, Article 6 (April 2007), 55 pages. http://doi.acm.org/10.1145/1229179.1229180

[9] Banko, M. Andrebill, E. 2001. Scaling to very very large corpora for natural language disambiguation. In Proc. of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001) 26-33

[10] Light, M., Mann, G. S., Riloff, E., Breck, E. 2001. Analyses for elucidating current question answering technology. Nat. Lang. Eng. 7, 4, 325-342

[11] Clarke, C., Cormack, G., Lynam, T. 2001. Exploiting redundancy in question answering. In Proc. of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001) 375-383

[12] Zahra Latifi, Abbas Karimi,"A TMR Genetic Voting Algorithm for Fault-tolerant Medical Robot," Medical and Rehabilitation Robotics and Instrumentation, V.42, pp. 301-307, 2014

[13] Eric P. Kim; Naresh R. Shanbhag, "Soft N-Modular Redundancy, "IEEE Transactions on Computers, V.61, pp. 323-336, 2012

[14] Omid NOROUZIFAR, Manouchehr KAZEMI, Mahdieh Nadi SENEJANI, Developing a New Weighted Voting Algorithm based on Markov Model, Bulletin de la Société Royale des Sciences de Liège, Vol. 86, special edition, 2017, pp. 528-534

[15] Zarafshan F, G Latif-Shabgahi, and Karimi A, "A novel weighted voting algorithm based on neural networks for fault-tolerant systems" Computer Science and Information Technology (ICCSIT) 3rd IEEE International Conference, Vol. 9, pp. 135-139, 2010

[16] Gruzenkin D. V., Grishina G. V., Durmuş M. S., Üstoğlu I., Tsarev R. Y. (2017) Compensation Model of Multi-attribute Decision Making and Its Application to N-Version Software Choice. In: Silhavy R., Silhavy P., Prokopova Z., Senkerik R., Kominkova Oplatkova Z. (eds) Software Engineering Trends and Techniques in Intelligent Systems. CSOC 2017. Advances in Intelligent Systems and Computing, Vol. 575, Springer, Cham

[17] Atkins, E. M., Abdelzaher, T. F., Shin, K. G. et al., Planning and Resource Allocation for Hard Real-time, Fault-Tolerant Plan Execution, Autonomous Agents and Multi-Agent Systems (2001)

[18] Berlizev A., Guelfi N. (2009) Fault Tolerance Requirements Analysis Using Deviations in the CORRECT Development Process. In: Butler M., Jones C., Romanovsky A., Troubitsyna E. (eds) Methods, Models and Tools for Fault Tolerance. Lecture Notes in Computer Science, Vol. 5454, Springer, Berlin, Heidelberg

[19] Cao, Z., Tian, Y., Le, TD. B. et al., Rule-based specification mining leveraging learning to rank, Automated Software Engineering, Springer (2018)

[20] Shekhar, C., Jain, M., Raina, A. A. et al., Reliability prediction of fault tolerant machining system with reboot and recovery delay, Int J Syst Assur Eng Manag (2018)

[21] Ferdinando C., Fault-Tolerant Search Algorithms, Springer-Verlag Berlin Heidelberg (2013)

[22] Nadia N., Luiza de M. M., Hardware for Soft Computing and Soft Computing for Hardware, Springer International Publishing (2014)

[23] Pan Z., Qi Z., Zhankui Z., Liman Y., The signal integrity design and simulation of triple modular redundant (TMR) computer, 2017 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM)

[24] Hossein S., Mehdi D., Mostafa S., Comparing the reliability in systems with triple and five modular redundancy, 2016 5th International Conference on Computer Science and Network Technology (ICCSNT)(2016)

[25] http://www.besplatnapravnapomoc.rs/

[26] http://hlt.rgf.bg.ac.rs/Page/Services

[27] D. Subotić, N. Forbes, "Serbo-Croatian language – Grammar", Oxford Clarendon press, str.25-31, 61-64, 101-113

[28] Magerman, Tom, et al. "Assessment of Latent Semantic Analysis (LSA) text mining algorithms for large scale mapping of patent and scientific publication documents." 2011

[29] Metzler, D. and Croft, W.B., Analysis of Statistical Question Classification for Fact-based Questions, in Information Retrieval, 8(3), 481-504, 2005

[30] P. Djikanović, V. Nikolić, D. Sivčević, National Framework of Interoperability of the Republic of Serbia and Service-Oriented Architecture (SOA), YU INFO 2014

[31] www.paragraf.rs