Modelling the System of Receiving Quick Answers for e-Government Services: Study for the Crime Domain in the Republic of Serbia

Vojkan Nikolić

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

Branko Markoski

Technical Faculty "Mihajlo Pupin", University of Novi Sad, Đure Đakovića bb, 23000 Zrenjanin, Serbia, markoni@uns.ac.rs

Kristijan Kuk, Dragan Ranđelović, Petar Čisar

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

Abstract: Today, e-Governments services operate quick answer systems in many different domains in order to help citizens receive answers to their questions at any time and within a very short time. The automatic mapping of relevant documents stands out as an important application for automatic questions-documents classification strategies. This paper presents a contribution to the identification concepts in text comprehension in unstructured documents as an important step towards clarifying the role of explicit concepts in information retrieval in general. The most common representational scheme in text categorization is the Bag of Words approach when the dictionary, as incorporating background knowledge, is large. The authors introduce a new approach to create concept-based text representations, and apply it to a text categorization collection to create predefined classes in the case of short text document analysis problem. Further in this paper, a classification-based algorithm for questions matching topic modelling is presented. The paper also describes the weighting of concepts that present terms with high frequency of occurrence in questions is based on their similarity relevance in the predefined classes of documents. The results of the experiment within the criminal law domain in the present case study show that the performance of concept-based text representations has proven to be satisfactory in the case when there is no dictionary for this domain.

Keywords: e-Government services; criminal law domain; natural language processing (NLP); Bag of Concept (BOC) approach

1 Introduction

The reform and modernisation of the public sector, based on a wide application of information-communication technologies (ICT) is considered to be one of the key elements in furthering development of the information society in the Republic of Serbia. The trends of development of many e-Government services in numerous countries in the world indicate the necessity of quick answer systems and e-Government services in the Republic of Serbia, as well. When performing many of the text mining and natural language processing tasks, it is necessary for all forms of a given word with the same meaning to be of identical form. On the whole, many documents that are available to e-Government services tend to be not structured, thus it is very difficult to isolate some forms (knowledge) from them.

Question Answering (QA) systems are an integral part of a multitude of e-Government services. E-Government represents the usage of information retrieval and extracting technologies in order to improve the activities of public sector organizations so that their online services make public information available "anytime, anywhere" to citizens. In cooperation with the government organizations, these QA systems usually aim to compile an annual list of commonly asked questions. In that way, each organization can develop their own online database of related responses especially given that the majority of citizens need assistance from e-Government services for a specific domain providing specific answers to citizen queries. However, the following problem arises: these systems need to hold all the adequate or similar answers able to answer the citizen's questions within the system. It is often the case that the system does not have prepared answer sets. Therefore, the existing Frequently Asked Questions (FAQ) website pages may play a crucial role in the collection of all possible answers.

To extract keywords in Serbian, the authors used a e-Government support component, the so-called ADVANSE [1]. The process that enables the automatic recognition of key words within the text document is known as the "Bag of Words" or "Bag of Concepts", it is a process in which small collections of recognized words and phrases are extracted. In order for this process to function, it is necessary to use certain algorithms so as to reduce the text based on important characteristics. Rapid Automatic Keyword Extraction (RAKE) is one of the simplest algorithms used to extract key words [2]. There are also cases, when the word corpus in the text cannot be recognized making it thus not possible to extract key words. Usually, in text mining set of co-occurring words within a given window is computing by N-grams of texts. The basic point of N-grams is that they capture the language structure from the statistical point of view. In such an instance, the N-gram is used to create a "Bag of N-grams," which represent the specific terms of the given text or document. This method is often used when searching texts written in natural languages. Key words in documents that have similar content or where the same language text corpus is used, can be extracted

by using algorithms for indexing TF-IDF (term frequency - inverse document frequency) [3].

For their search service based on tokenization, counting and normalization, Apache Corporation and Microsoft have created several natural languages processing technology. Following experimental features of the search service API showed that Apache Lucene analysers are faster, but that the Microsoft analysers have advanced capabilities (lemmatization, word decompounding and entity recognition). In the present case, the authors developed the QA system using the open-source search server based on the Lucene Java search library.

The paper is organized as follows. Following this introduction, the second section of the paper elaborates on the related works. The third section presents the basic theoretical principles of text mining techniques with special emphasis on short texts and the lexical normalization (stemming) process based N-gram analysis. Also, in the same section the report in criminal law documents is given along with mapping paragraph members of the Criminal Code of the Republic of Serbia as documents for response answering question approach. The subsequent section describes a framework for the proposed quick answers, which is implemented on the web service related to the e-Government for the criminal law domain. Further details about the components of the architecture of this proposed system are given and the functionalities with the developed algorithm for question Classification are shown, too. Finally, in the experimental results section, the authors report on the successful implementation of the proposed system as well as the recommender for the best similarity measures for finding similarity between short documents in Serbian.

2 Related Work

Taking into account the existing solutions which depend on e-Government requirements, Šimić and others [1] firstly considered the problems identified in the existing solutions in the Republic of Serbia and then offered their own hybrid approach solution: multi-layer based document clustering based on the model of fuzzy concepts and using different measures of text similarity. The aim is to decrease the time of receiving a response from subjects (institutions) of public administration with a minimum (or completely without) any civil clerks' involvement, so the authors offered a new approach to facilitate the optimization and automation of using advanced methods and techniques of retrieving information. This paper presents the ADVanced ANSwering Engine solution (ADVANSE) for a wide range of e-Government services. The most important results of the ADVANSE project are reflected in the quality of e-Government services, the quick response time to the citizens' requests, the innovative uses of the available content and the restructured relationship between civil clerks and

citizens. In particular, the emphasis is on the efficiency and flexibility of the response.

The present group of authors proposes focusing on testing in different conditions and improving the ability of adaptation in the next research phases. One of the objectives pursued in this work is to find solutions for the functioning of such a system in multilingual environments and increasing content complexity concerning grammar and dictionaries of different languages regardless of the area of use. In that sense, different strategies are proposed. A particular challenge is the functioning of e-Government services in different domains. Specific domains use special dictionaries, so it is necessary to use specialized techniques to find similarities. Further, a qualitative improvement of processing the given document is also required. A possible way to obtain the answer is the tagging of certain parts of the document instead of labelling the entire document.

Marovac et al. [4] conducted an analysis of the texts written in different languages and with different linguistic rules. It is especially complicated to analyse the texts written in Serbian given that the Serbian language has two alphabets in use, Cyrillic and Latin, and is also one of the languages with a rich morphology. The use of linguistic resources (text corpus of contemporary Serbian language, morphological dictionaries, stop-words, dictionary of abbreviations, etc.) aiming to obtain a qualitative analysis of a natural language becomes a considerable challenge.

The use of N-gram analysis contributed in significant results achieved without the use of extensive lexical content or analysis of texts in Serbian carried out without using the morphological vocabulary. Special attention is paid to the algorithm for extracting keywords (N-gram) shown in more detail below.

According to [4], the algorithm should be developed like clustering keywords according to their frequencies in the text or parts of the text or clustering keywords (N-gram) that are frequent in comparison to those less frequent ones, i.e. the ones having higher chi-square value. The size of the chi-square statistic is a misleading indicator of the extent of the relationship between two words. In this way, relationship of keywords bases can be extracted.

Stanković et al. [5] also deal with a model that leads to better quality of extensive information retrieval in text document and databases using a bag-of-words for pre-indexing and naming these objects. Each document contained in the database represents the summary report consisting of meta data (domains, keywords, summary and geographic location). The bag-of-words, resulting from these metadata using morphological dictionaries, are objects identified when using the rule-based system. The work of these authors has been valuable in the process of pre-indexing documents (information retrieval). The ranking of the documents downloaded is based on several measures TF-IDF. It can therefore be concluded that the development results obtained in such a way compared to the ones obtained

without the use of pre-indexing showed a great progress regarding an average of accurate measurement.

Kolomiyets et al. [6] represent the Question Answering method as a comprehensive approach that provides a more qualitative way for information retrieval. This approach is actually a system of queries and documents in relation to the possible functions of search in order to find an answer. This research discusses general questions contained in a complex architecture with increasing complexity and the level of frequency of questions and information objects. These authors represent here a method of how natural language roots are reduced on keyword for search while knowledge databases and resources, obtained from natural language questions and answers, are made intelligible.

Wenyin et al. [7] presented a user-interactive question-answering platform named BuyAns, a kind of online community which primarily uses a scheme for answering questions among users in this way creating a special pattern-based user interface (UI) for asking and answering. This pattern is based on the pooling and storage of pairs obtained in the question-answer relationship. Such a system promotes a C2C business model for exchanging and commercializing different types of knowledge. It can be used as a special kind of incentive approach to knowledge acquisition. In operation, this system uses templates and recognizes accurate answers to specific repeated questions.

3 Theoretical Background

3.1 Lexical Normalization

In general, optimization involves identifying the extreme values of objective functions given in a particular domain which combine different types of objective functions and different types of domains. In mathematics, statistics, empirical science and information technology, optimization is the selection of the best element (with regard to some criteria) from a given set of available alternatives [8]. When referring to the problem of optimization, one means the minimum and maximum chosen among real functions which are given, where systematic changes of input values (with regard to functions) are predetermined and defined.

An example of a text mining technique is given in [9], as the most commonly applied technique in the field of e-Government, citizen participation and e-democracy [10]. The QA application (in regard to text mining) enables choosing/finding the best answer for questioning and answering in the context of e-Government system with multiple search text options. These techniques imply text categorization/classification, text clustering, the concept of object extraction, sentiment analysis, document summarization, etc. which help in the determination

of relations within the collection of a large amount of text data using specific forms and methods. Given that text classification/document categorization is the process of assigning predefined categories of free text documents providing an overview of the concepts within the collection of documents, this solution has found a wide range of practical usage [11]. A great number of current approaches represent the text as a list (vector) of words, the so-called "Bag-of-Words" (BoW). In fact, based on the statistical analysis of certain factors (such as frequency and distribution within the collection of words), and without taking the word order and grammar into account, BoW creates a series of words and phrases recognized as important.

The lexical features are most commonly used in the QA and summarization communities. One of the robust lexical feature approaches in statistical models is the use of the N-gram model. The N-gram analysis determines the frequency of different N-grams in a text. N-grams of texts are extensively used in text mining and natural language processing tasks. The N-gram overlap measures the overlapping word sequences between the candidate document sentence and the query sentence. The recall-based N-gram scores for a sentence *P* using the following formula [12] are as follows:

$$NgramScore(P) = max_i \left(max_j \, Ngram(s_i, q_j) \right) \tag{1}$$

$$Ngram(S, Q) = \frac{\sum gram_n \in S^{Count_{match}(gram_n)}}{\sum gram_n \in S^{Count(gram_n)}}, \tag{2}$$

where *n* stands for the length of the N-gram ("n = 1, 2, 3, 4"), and $Count_{match}$ ($gram_n$) is the number of N-grams co-occurring in the query and the candidate sentence, q_j is the j-th sentence in the query pool, and s_i is the i-th sentence in the sentence pool of sentence *P*. In general, an N-gram is a subsequence (substring) of size *n* from a given sequence (string). An N-grams of the size 1, 2 or 3 are referred to as anuni-gram, bi-gram or tri-gram, respectively.

Lexical normalization (stemming) is the process of reducing words to their roots. This is a necessary step in analysing textual content quickly. Stemming is the process of removing all of the prefixes and suffixes of the given words so as to produce the stems or the root [13]. The main objective of the stemming process is to remove all possible affixes (prefixes and suffixes) and thus reduce the word to its stem. N-gram analysis can be used as an alternative for finding words whose root is the same. This method identifies groups of words where the words in the group are small syntactic variants of one another and collects only the common word stem per group. For example, "development", "developed" and "developing" will be all treated as "develop". In order to obtain more effective results in the selection of the number *n*, the suffix are collected for important words in Serbian language [14]. Based on this, it was found that the highest percentage (92.70%) were extensions of under 4 letters. Also, in the research study of the paper by the authors Marovac et al. [4], the 4-gram analysis is shown

as it presents the best results for Serbian. For further research in this paper, therefore, 4-gram analysis will be performed as the reference algorithm for normalizing words in the documents in Serbian language.

3.2 Mapping Paragraph Members of the Criminal Code as Short Texts to Concepts

The BoW framework requires faster performance and adaptation of new words [16], especially during the search of short texts, as well as a simple online application. Sahlgren et al. [17] were pioneers in using the term "Bag-of-Concepts" (BoC) which means that when making the conditions for the term construction (concept), the taxonomy of knowledge is used. BoC then performs conceptualization within a short text in a manner that recognizes the content of the text and then connects the shorter part of the text with the relevant terms-concepts. The aim of conceptualization is an extraction of the most characteristic terms-concepts which best describe the given text [18, 19]. In fact, when in dealing with short texts there is often the risk of losing the context, which makes this procedure useful in the case of working with texts including various types of data, various texts with categorical or topical information, and thus facilitate the work of many applications. This process of mapping short texts within the whole collection has been successfully implemented. This paper aims to map the short text in the context of Law Article and represents it as a vector that best describes a particular part of criminal law when it is not predefined by a legal expert. To reach the representative vectors, one needs to overview the existing methods which lead to the keywords.

Keywords are the ones which are most frequently used in text documents for improving the information retrieval within the sets/collection of documents [20]. Automatic identification of the keywords is a method for analysing words and phrases within small sets/collections so that the content of a given document can be described. The process of identifying keywords aims at accelerating the search of documents in order to obtain the required information. In this regard, four key method extraction categories have been identified: the statistics, linguistics, machine learning and hybrid approach [21]. The statistical approach emphasizes the frequency of the file in a document using four types of frequency points within a set/collection of documents. These are: frequency of files in each subdocument with and without resulting words and frequency in each document with and without resulting words. The text-search techniques primarily establish a series of words and phrases which are then extracted on the basis of statistical analysis of the given factors, such as frequency and distribution. As part of the work process, modern tools create the Term Document Matrix (DTM) and use numerical statistics, so-called TF-IDF [22].

The conventional, TF-IDF weight scheme was invented by Berger et al. [23], which is as follows:

TF: $$TF(t, d_i) = \frac{n_{t,i}}{\sum_{k=1}^{|T|} n_{k,i}}, \quad (3)$$

where $TF(t, d_i)$ is term frequency of word t in document d_i, $n_{t,i}$ is number of occurrences of term t in d_i and $n_{k,i}$ is number of occurrences of all terms in d_i.

IDF: $$IDF_t = \log \frac{M}{m_t + 0.01}, \quad (4)$$

where M is total number of documents in the text corpus and m_t is total number of documents in the text corpus where term t appears.

TF-IDF: $$w(t, d_i) = TF(t, d_i) \times IDF_t, \quad (5)$$

where $w(t, d_i)$ is weight of term t in document d_i.

The frequently used algorithm for indexing TF-IDF requires the existence of a text corpus of documents to use key words of its age. In the absence of a text corpus, keywords can be extracted based on the number of impressions available at the given document. Table 1 shows the seven most often used terms in three separate documents which are found in the studied Articles of the Criminal Code of the Republic of Serbia. It is applied using a morphological normalization vocabulary. The results of the normalization carried out by cutting off the N-gram length of four are presented in Table 2.

Table 1

Normalization 1

No.	TERM	Freq
1.	delo	3
2.	učinilac	3
3.	povreda	2
4.	nastupila	2
5.	zdravlje	2
6.	telesna	1
7.	teška	1

Table 2

Normalization 2

No.	TERM	Freq
1.	delo	3
2.	učin*	3
3.	kazn*	3
4.	zatv*	3
5.	tele*	3
6.	tešk*	3
7.	povr*	3

Shortening the words to 4 characters increases the number of concepts covered key words since, for example, the word “kazniti” and “kaznom” were shortened to

the same base, hence the new key word N-gram “kazn”, as the beginning of the word “kazniće”.

It was established that the use of other algorithms offered in the Optimization Algorithm Toolkit (B-cell Algorithm (BCA), Cloning, Information Gain, Aging (CLIGA), CLONALG, simple and immune algorithm (SIA)), also lead to accurate results although they differ in the duration of operation time.

4 Framework for the Proposed Quick Answers Web Service

In order to create the novel mapping methods, the authors applied the BoC approach (TF-IDF weighting) within the three different documents which are part of the Criminal Code. The candidates are faced with the following combination of the terms: TF values and time IDF values. TF-IDF cannot measure the frequency of keywords in relation to the expected result of a larger set/collection of documents. Search Engine Optimization (SEO) [24], during TF-IDF term weighting in relation to term ranking, performs slightly better than the use of key words. TF-IDF represents each text data as BoW. In the entire text corpus, greater emphasis will be given to the high frequency terms, compared to those, whose frequency is lower. The tendency of TF-IDF is to refine the common terms.

The aim of the Question Answer application was so that citizens can receive quick and accurate answers to their questions. The most common types of content in such an environment are questions, documents and answers. The user may pose a question in various fields. Since the documents are clustered based on keywords, the whole collection is searched in order to find similarities within existing documents from related groups. In the present case study, the citizens' questions refer to criminal offenses as defined in the aforementioned three Articles of criminal law, Articles 121, 122 and 135. Parts of the new law in this QA system are represented as three different document groupings as a whole, comprising the entire currently available knowledge base. To complement and complete this centralized repository, the expert answers were used from the web portal for free legal aid – Pro Bono [25]. The analysis of the existing responses to the questions of citizens concern the criminal acts of the three Articles within the section of Questions and Answers in the field of criminal law. The authors selected 45 representative questions from the given set of questions. These consist of a possible short text documents group that will help towards a better presenting of the considered BoC.

In contrast, finding information from web pages for completing the knowledge base assumes the use of different methods suitable for certain areas that were carried out automatically, which are adaptable and can integrate the data found

[26]. Google's unique and improving algorithm has made it one of the most popular web search engines of all time. For this reason, the authors used Google's search engines to find the content of any tags with keywords from the given BoC. Search engines recognize relevant web pages as additional human resources for a better explanation of the terms that could be important in the mapping of relevant issues with a group of appropriate responses. During the setting up of the topic, the web administrators define and link the texts' dates with a series of key words via the option Related Articles or via the meta keywords tags. In order to discover the group that applies to a given problem, the authors set up a search query for the term "serious bodily injury" as a keyword obtained from electronic dictionaries for the Serbian language via web Language resources exploration site page [27]. The repository, which proved to be the best, was the site of the daily newspapers "Blic". From this site, a total of 35 texts were identified, which corresponded to the specified tag.

4.1 Definition of the Frequency of Term

The minimum term frequency is obtained by the second occurrence of a given term - while in the document this is used to represent the threshold. The term must exist at least twice in a document so as to be counted. Considering the length of the documents to be counted, the following can be stated: supposing that the length of the document is c_1, c_2, ..., c_n and $f(c_1, c_2, ..., c_n)$, representing its frequency, then $c_1, c_2, ..., c_n$ are extracted as terms from the text only if $f(c_1, c_2, ..., c_n)$ is equal to or greater than 2 [28], [29].

The mid interval is the most frequently selected: $(v_i+v_{i+1})/2$ as a representative start. However, C4.5 at the beginning chooses a lower value v_i for each interval $\{v_i+v_{i+1}\}$, not only for the high value [30]. This led the present authors to also implement this type of threshold calculated.

By using stop-words, both the relevant and irrelevant terms within the text were extracted since irrelevant words or concepts often appear. The relevant list of stop-words, for example, can be obtained by moving all the stop-words right from certain words such as the articles (e.g. the, a, an), the preposition (e.g. above, across, before), or some adjectives (e.g. good, nice) in an application [31]. In Serbian, currently there is no formal list of stop-words on the Web site https://sites.google.com/site/kevinbouge/stopwords-lists, thus the authors used the Serbian Lemmatized and PoS Annotated Corpus (SrpLem Kor). The stop-words were created using a list of 700 Serbian stop-words. Moreover, the list of stop-words is automatically constructed in each training process by retrieving the terms with the greatest frequency in the whole Corpus.

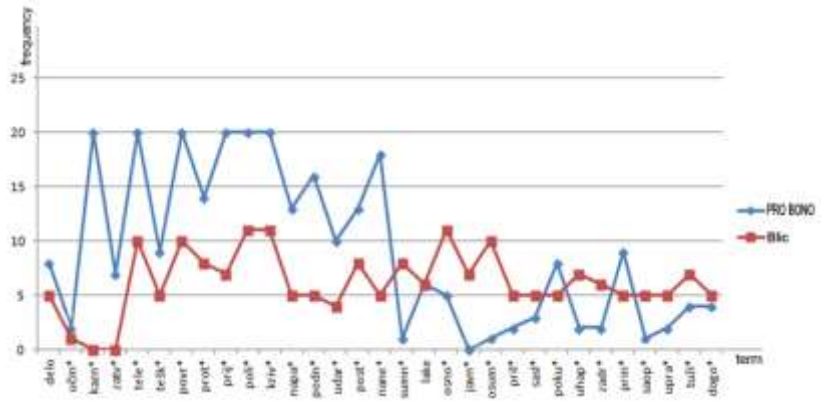

Figure 1
BoC representation in the crime domain for part of the Criminal Code

The TF-IDF weighting scheme is often used in the vector space model (VSM). This model, as well as some distance/similarity measures, are used to determine the similarity between documents or documents and queries represented in the vector space. A document is represented in the t-dimensional term vector space by $d_j = (w_{1,j}, w_{2,j} \ldots w_{t,j})$, where $w_{t,j}$ is defined by the weight key word or terms of the BoC.

In today's web search engines, when the user submits a query to a search system, first they must determine which pages in the index are related to the query and which are not. The proposed web system for quickly answering in the domain of crime, as presented in Figure 2, is based on the absence or presence of it in a document - representation of documents as the BoC. The concepts are mapped to a crime vocabulary for the three law Articles given as vectors below:

- Article 121 [kazn, zatv, tešk, post, javn, poku, tuži, pril, delo, tele, povr, nane, sumn, uhap];
- Article 122 [tele, poli, napa, udar, post, nane, učin, kazn, zatv, kriv, javn, povr, lake, prij, tešk, poku, tuži];
- Article 135 [prij, poli, kriv, post].

One the one hand, this approach is based on the classification of questions on the Criminal Code by using a special domain based on the BoC corpus, while on the other hand, it is based on the comparison of characteristics with the properties of the Codes in the case of the used expert representation of documents.

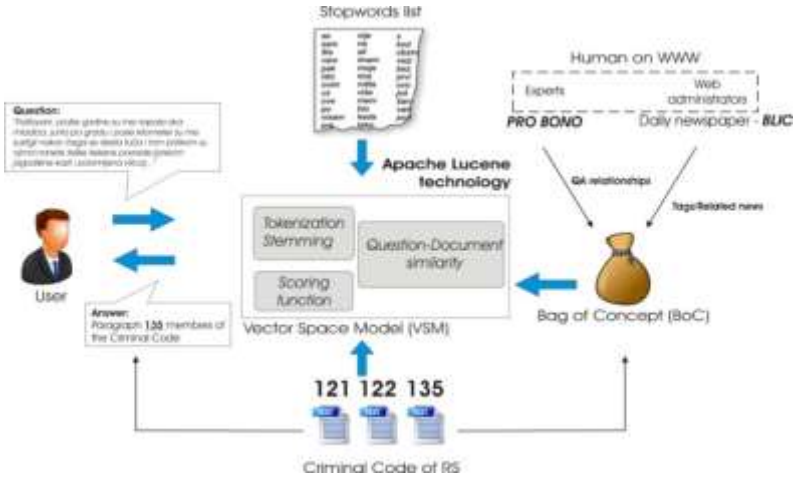

Figure 2

Question Answering system based on the BoC model

In the proposed system, first, it was necessary to map the question to the concepts from the BoC in the corpora defined according to 31 terms seen in Figure 2, presented as a bag. The second step for questions tokenization was to filter the stop-words. Following this, the next step was to remove variable suffixes. This led to a form of morphological normalization (or creating general characteristics). For this reason, when searching texts written in Serbian, the 4-gram analysis was applied. Finally, the authors reached a similar result, starting from posing the question and the BoC display view of three documents using scoring the tool.

The authors' QA system generates as output the relevant document for this question. As a final output, the user then receives the following message as an answer: "Look at Article *n* of the Criminal Code". The value of the number *n*, displayed to the user determines the part of the system software called domain-specific stemming annotator. A stemming annotator is a software agent that uses the BoC source of a particular domain as knowledge bases to map the extracted terms. It assigns the absence or presence of each extracted term in accordance with its keywords relevance within the BoC. Translating the question (query) and documents from raw strings into a computable form is the first hurdle to be overcome in computing a similarity score. To achieve this, the textual representation of information is converted into an algebraic vector space representation [32].

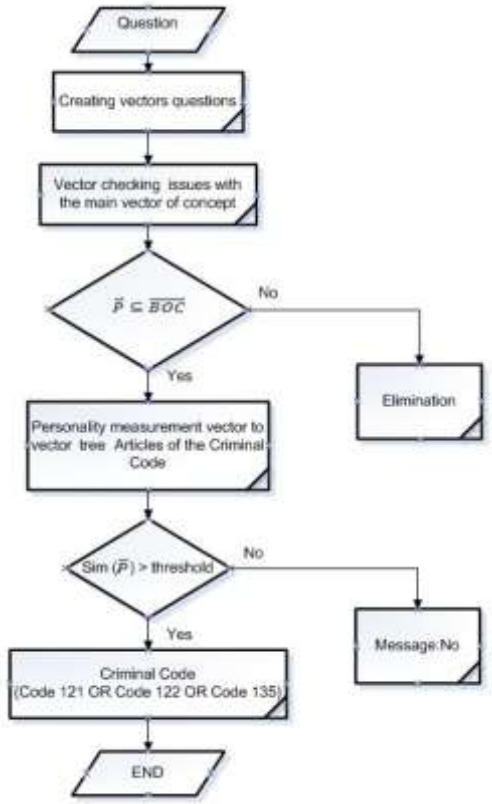

Figure 3
Classification questions algorithm

4.2 Question Classification

In order to ensure the successful operation of the QA system as a whole, it is necessary to perform the correct classification of questions and expected answers. The aim of the classification of the expected questions and answers is mapping - marking the QA types [33]. The targeted values were established on the basis of the values which provided the best results in the previous cases of the text classification. This paper proposes a new algorithm based on the BoC for automating question-document classification tasks such as recommendation on relevant documents. The BoC is a list of all words ranked according to their discriminative power of the three Articles of Criminal Law (clusters). Different measures of distance or similarity use the vectors from each document to calculate distances from the question.

Once a set of representative terms is selected for each cluster $c_j \in C$ (= $\{c_1, c_2, \ldots, c_n\}$), the set of terms is regarded as the representative terms (rt_j) of cluster c_j. This is followed by the comparison of the similarity of each question q_i mapping in the BoC (matching the vector of the question with each vector cluster) with each rt_j using the certain similarity metric to automatically compute a relevance score that measures the similarity between the question and Article number of the Crime Code - c_j. In order to determine the threshold of similarity between the question presented in the form of a short text and the Articles of the law, which is also shown as a short text, the following formula was used [34]:

$$Similarity = \frac{W(Sa) \cap W(Sb)}{\min(W(Sa), W(SB))}, \quad (6)$$

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, and in both documents.

5 Experimental Evaluation

Before the process of the clusterization, the similarity/distance measure must be determined. Similarity measure is vital due to its direct influence on the ranking of documents, in fact, direct influence on the degree of closeness or separation of the target documents. In addition, measuring the similarity of documents on the basis of characteristics that depend on the type of data in context to documents and observation leads to the clustering and grouping documents by clusters. There is no measure that is universally optimal for all kinds of clustering problems.

The presented system used the free open-source Lucene library search by Apache corporation. This search method can be embedded in the application code that runs in the background when the need arises or it can be run on a remote website. In the process of searching, the field values are searched. In Lucene, starting with the version 5.0, the *SerbianNormalizationFilter* is included, which normalizes Serbian Cyrillic and Latin characters into "bold" Latin. Cyrillic characters are first converted to Latin, then the Latin characters have their diacritics removed with the exception of *đ*, which is converted to *dj*. Note that this process expects lower-cased input.

public class **SerbianNormalizationFilterFactory**
extends **TokenFilterFactory**
implements **MultiTermAwareComponent**
Factory for **SerbianNormalizationFilter**

For the here-described research, the Criminal Code of the Republic of Serbia was used, which was written in Serbian, in Latin script. Unlike English, Serbian contains the following signs: č, ć, ž, đ and š. In order for these signs to be

identified and then the process of indexing and search executed, it is a pre-requisite that Lucene be supported for the Serbian language. This support is provided by the Lucene version 5.2.0, which was implemented in this study.

Choosing an appropriate similarity measure is crucial for cluster analysis, especially for a particular type of clustering algorithms. Four types of the most used similarities were included so that the one be chosen which provides the most precision results for the Criminal field.

The following text similarity functions were used:

1. Cosines Similarity

$$SIM_C(\overrightarrow{t_a}, \overrightarrow{t_b}) = \frac{\overrightarrow{t_a}.\overrightarrow{t_b}}{|\overrightarrow{t_a}|^2 \times |\overrightarrow{t_b}|^2}, \quad (7)$$

where t_a and t_b are m-dimensional vectors over the term set $T = t_1, ..., t_m$.

2. Jaccquard's Coefficient

$$SIM_J(\overrightarrow{t_a}, \overrightarrow{t_b}) = \frac{\overrightarrow{t_a}.\overrightarrow{t_b}}{|\overrightarrow{t_a}|^2 + |\overrightarrow{t_b}|^2 - \overrightarrow{t_a}.\overrightarrow{t_b}} \quad (8)$$

The Jaccard's coefficient is a similarity measure and ranges between 0 and 1. It is 1 when $t_a = t_b$ and 0 when t_a and t_b are disjoint where 1 means that the two objects are identical whereas 0 means they are completely different. The corresponding distance measure is $d_j = 1 - SIM_j$.

3. Euclidean Distance

$$D_E(\overrightarrow{t_a}, \overrightarrow{t_b}) = \left(\sum_{t=1}^{m} |w_{t,a} - w_{t,b}|^2\right)^{\frac{1}{2}}, \quad (9)$$

where the term set is $T = t_1, ..., t_m$. As mentioned previously, the TF –IDF value was used as term weights, that is $w_{t,a} = tf\text{-}idf(d_a, t)$.

In order to determine the appropriate measure of similarity in this proposed system, a study was conducted using the similarity of all three above-mentioned measures over 10 questions in the form of text documents in the field of the Criminal Code. The results of this analysis are presented in Table 3.

Table 3
A summary table of similarities

	Article 121	**Article 122**	**Article 135**	**Expert**
Question KP 1	sim(Cos)= 0.034003 sim(Jacc.)=0.909091 sim(Eucl.)= 17.320508	sim(Cos)= 0.031068 sim(Jacc.)= 0.909091 sim(Eucl.)= 7.937254	sim(Cos)= **0.188982** sim(Jacc.)= **1.000000** sim(Eucl.)= **5.291503**	Article 135
Question KP 2	sim(Cos)= 0.230796 sim(Jacc.)= **1.000000** sim(Eucl.)= 13.152946	sim(Cos)= 0.312195 sim(Jacc.)= 0.909091 sim(Eucl.)= 6.708204	sim(Cos)= **0.435801** sim(Jacc.)= **1.000000** sim(Eucl.)= **8.426150**	Article 121

Question KP 3	sim(Cos)= 0.154287 sim(Jacc.)= 0.916667 sim(Eucl.)= 16.792856	sim(Cos)= 0.197359 sim(Jacc.)= 0.916667 sim(Eucl.)= 7.549834	sim(Cos)= **0.268687** sim(Jacc.)= **1.000000** sim(Eucl.)= **7.000000**	Article 122, 135
Question KP 4	sim(Cos)= **0.522323** sim(Jacc.)= 0.900000 sim(Eucl.)= 15.427249	sim(Cos)= 0.434959 sim(Jacc.)= **1.000000** sim(Eucl.)= **5.477226**	sim(Cos)= 0.268687 sim(Jacc.)= **1.000000** sim(Eucl.)= 7.000000	Article 121, 122
Question KP 5	sim(Cos)= 0.402331 sim(Jacc.)= 0.888889 sim(Eucl.)= 15.905974	sim(Cos)= 0.441129 sim(Jacc.)= 0.888889 sim(Eucl.)= 5.477226	sim(Cos)= **0.759257** sim(Jacc.)= **1.000000** sim(Eucl.)= **2.828427**	Article 135
Question KP 6	sim(Cos)= **0.017314** sim(Jacc.)= 0.909091 sim(Eucl.)= 16.970563	sim(Cos)= 0.000000 sim(Jacc.)= 1.000000 sim(Eucl.)= 7.000000	sim(Cos)= 0.000000 sim(Jacc.)= **1.000000** sim(Eucl.)= **3.464102**	Article 135
Question KP 7	sim(Cos)= 0.660926 sim(Jacc.)= **1.000000** sim(Eucl.)= 15.394804	sim(Cos)= 0.648886 sim(Jacc.)= 0.750000 sim(Eucl.)= 5.099020	sim(Cos)= **0.92582** sim(Jacc.)= **1.000000** sim(Eucl.)= **4.690416**	Article 121

To select the measure, the results obtained with the 3 date rates were compared to the similarities given over the formula (as indicated in the previous section). As in the case of similarity, it is demonstrated that the question Q1 has the greatest similarity with Article 135 where the value was 1.000000. If certain similarities in the column of Article 135 are detected, it can be determined where they are equal or greater in value than is given by the Jaccard similarity. However, the question arises if it was taken as a reference similarity measure. The remaining six issues were analysed in the same manner. The results show that the reference rate of similarity is in this case the Jaccard's correlational coefficient, which is taken as a representative for the calculation of the algorithm in Section 4 in Figure 3.

The results of checking the validity (accuracy) of this algorithm on the basis of the types of alignment by the gold standard vs. real prediction is summarized in the table below:

Table 4

Evaluation Metrics: Classification View

Artical / Expert	Retrieved	Not Retrieved
Relevant	Relevant Retrieved	Relevant Rejected
Not relevant	Irrelevant Retrieved	Irrelevant Rejected

$$Precision = \frac{Relevant\ Retrieved}{Retrieved} = 75{,}71\% \quad (10)$$

$$Recall = \frac{Relevant\ Retrieved}{Relevant} = 57{,}00\% \quad (11)$$

$$F_{i\ (i=1,n)} = \frac{2*precision_i*recall_i}{precision_i+recall_i} \quad (12)$$

$$F_{Average} = \frac{F_1+F_2+\cdots+F_n}{N} \quad (13)$$

Precision, recall and F_1 are only focus on true positives, i.e., those positive examples by the gold standard. In a monolingual alignment, positive examples are those tokens that are aligned, while negative examples are those that are not. Usually the focus is only on whether those that should have been aligned, are indeed correctly aligned, thus the measure of F_1 is a good *t*.

If correct rejection (true negative) is important, then accuracy is computed instead:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (14)$$

Accuracy weighs true positives and true negatives equally. The aim is to ensure that the Classifier recognizes both the positive and negative examples. Specifically, in alignment, where most tokens are not aligned, the accuracy value is likely to be very high in general and it is difficult to determine the difference. In such an instance, only F_1 on positive (aligned) examples is reported.

Number of questions extractions results- In order to test the algorithm, the authors used 10 questions. Table 5 shows the results obtained on the basis of the above-presented algorithm, further, the table also displays the results provided by the expert in the field of the Criminal Code.

Table 5
Table of results

Precision	75,71 %
Recall	57,00 %
$F_{Average}$	0.6936
TP	7
FP	5
FN	3
Accuracy	46,66 %

Conclusions

This paper deals with the impact of certain parameters of algorithms on the time period required for processing, trying to determine the optimal value of certain algorithms. The examples analysed in this paper present unique algorithms resulting in optimal solutions using different time values in relation to different parameter values. In this regard, multiple optimal values exist as extreme, unacceptable values.

In the authors' view, it is also important to check and research the behaviour of each of these algorithms in relation to the number of different parameters, which comprises the practical part of this paper. It has been confirmed that all available algorithms are not applicable in certain situations (function optimization), given that specific algorithms analysed during the study did not recognize certain extreme test search functions.

A new model, based on the BoC approach in text representations, was presented in the paper. The proposed system was focused on the terms and they failed to identify concepts when attempting to understand the text. Using the BoC is suggested in order to improve the performance of QA applications. This approach was achieved by the novel exploration of a set of predefined classes and has provided impressive results in the process dimensionality reduction by the mapping query to the document. The research shows that this current system can map queries to the relevant code Article with average precision values of 75,71 %. In general, citizen queries in e-Government services must be very short and they are difficult to classify using a traditional classifier technique, such as the BoW. The documents or parts of them need to be processed so as to obtain typical concepts since the queries can be similar in the main topics of the documents. Thus, much time is saved by not having to process large amounts of long text so as to enable the concept model of documents presented as predefined classes to give quick recommenders presented as answers.

Regarding the concepts as sub contents, these can diversify the recommendations directly from the subtopic level where a large knowledge base is needed to convert the terms to concepts. Helping to adequately construct a concept model for each member of the law requires the use of many sources. Therefore the authors recommend the use of existing news Articles by people already grouped and classified using the tag of that mark. This principle of addition to the relevant terms of the existing unstructured documents is necessary to describe what precise vector the document is given in the absence of a dictionary for the area. This framework is suitable for many future e-services, especially those that do not contain pre-set response answers offered in the form of relevant documents. Also, the authors found that using the Jaccard index provided the best results in comparison to other similarity measures for comparing short text, which might be suitable for short text such as the clustering process short-text corpora. This paper presents one solution of an e-government service application designed to find the existing answers or relevant document for citizens' questions asked in Serbian. The research presented in this paper indicates that a new approach must be introduced with the aim to solve similar problems in current and future Serbian public services.

Acknowledgement

This work was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia under the project no. III 47016.

References

[1] G. Šimić, Z. Jeremić, E. Kajan, D. Randjelović, A. Presnall, A Framework for Delivering e-Government Support, *Acta Polytechnica Hungarica*, 11(1), 2014, pp. 79-96

[2] A. Medelyan, NLP keyword extraction tutorial with RAKE and Maui, https://www.airpair.com/nlp/keyword-extraction-tutorial. [Accessed April 2016]

[3] D. Avdic, A. Avdic, Spalević, U. Marovac, M-Government Application Intended to Search Documents Written in Serbian Language, in *Sinteza 2014 - Impact of the Internet on Business Activities in Serbia and Worldwide*, Belgrade, Singidunum University, Serbia, 2014, pp. 902-906, doi: 10.15308/sinteza-2014-902-906

[4] U. Marovac, A. Pljasković, A. Crnišanin, E. Kajan, N-gram analysis of text documents in Serbian, Proceedings of 20th Telecommunications forum TELFOR 2012, pp. 1385-1388

[5] R. Stanković, C. Krstev, I. Obradović, O. Kitanović, Indexing of Textual Databases Based on Lexical Resources: A Case Study for Serbian, Semantic Keyword-Based Search on Structured Data Sources, *Springer*, 2016, pp. 167-181

[6] O. Kolomiyets, M.F. Moens, A survey on question answering technology from an information retrieval perspective, *Information Sciences* 181(24), 2011, pp. 5412-5434

[7] L. Wenyin, T. Hao, W. Chen, M. Feng, AWeb-Based Platform for User-Interactive Question-Answering, *World Wide Web* 12(2), 2009, doi 10.1007/s11280-008-0051-3, pp. 107-124

[8] "The Nature of Mathematical Programming", Mathematical Programming Glossary, *INFORMS Computing Society*, https://glossary.informs.org/ver2/mpgwiki/index.php?title=Extra:Mathematical_programming [Accessed 15 March 2017]

[9] T. W. Miller, Data and Text Mining: A Business Applications Approach, *Pearson / Prentice Hall*, New Jersey, 2005

[10] G. Koteswara Rao, D. Shubhamoy, Decision Support for E-Governance: A Text Mining Approach, *International Journal of Managing Information Technology (IJMIT)*, 3(3), August 2011, DOI: 10.5121/ijmit.2011.3307. pp. 73-91

[11] Y. Yang, T. Joachims, Text Categorization, Scholarpedia, 3(5): 4242. http://www.scholarpedia.org/article/Text_categorization. doi:10.4249/scholarpedia.4242. [Accessed 11 March 2017]

[12] W. Jiang, B. Samanthula, N-Gram based Secure Similar Document Detection, the 25th Annual WG 11.3 Conference on Data and Applications Security, Richmond, Virginia, July 11-13, 2011

[13] R. Kumar, V. Mansotra, Applications of Stemming Algorithms in Information Retrieval - A Review, *International Journal of Advanced*

Research in Computer Science and Software Engineering 6(2), February - 2016, pp. 418-423

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

[15] Paragraf - legal and economic editions, www.paragraf.rs. [Accessed 1 Jun 2017]

[16] F. Wang, Z. Wang, Z. Li, J-R Wen, Concept-based Short Text Classification and Ranking, *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, 2014. pp. 1069-1078

[17] M. Sahlgren, R. Cöster, Using bag-of-concepts to improve the performance of support vector machines in text categorization, In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING 2004, pp. 487– 493, Geneva, Switzerland, August 2004

[18] Y. Song, H. Wang, Z. Wang, H. Li, W. Chen, Short text conceptualization using a probabilistic knowledgebase, *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three*, AAAI Press, 2011, pp. 2330-2336

[19] Z. Wang, H. Wang, Z. Hu, Head, modifier, and constraint detection in short texts, In: *Data Engineering (ICDE), 2014 IEEE 30th International Conference on*. IEEE, 2014. pp. 280-291

[20] A. Hulth, Improved automatic keyword extraction given more linguistic knowledge, *Proceedings of the 2003 Conference on Emprical Methods in Natural Language Processing*, Sapporo, Japan, 2003, pp. 216-223

[21] C. Zhang, H. Wang, Y. Liu, D. Wu, Y. Liao, B. Wang, Automatic Keyword Extraction from Documents Using Conditional Random Fields, *Journal of Computational Information Systems*, 4(3), 2008. pp. 1169-1180

[22] G. Wunnava, Applying Machine Learning to Text Mining with Amazon S3 and RapidMiner (on Jun 25 2015), AWS Big Data Blog. https://blogs.aws.amazon.com/bigdata/post/Tx22THFQ9MI86F9/Applying-Machine-Learning-to-Text-Mining-with-Amazon-S3-and-RapidMiner. [Accessed March 2017]

[23] A. Berger et al, Bridging the Lexical Chasm: Statistical Approaches to Answer Finding, In: *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2000. pp. 192-199

[24] A. B. King, Website Optimization, *O'Reilly Media*, July 2008

[25] Pro Bono - portal for free legal aid, http://www.besplatnapravnapomoc.rs/

[26] I. Augenstein, D. Maynard, F. Ciravegna, Distantly Supervised Web Relation Extraction for Knowledge Base Population, *Semantic Web* 7(4), 2016. pp. 335-349

[27] Human Language Tecnologies, http://hlt.rgf.bg.ac.rs/VeBranka/BagOfWords.aspx . [Accessed 3 May 2017]

[28] S. Li, H. Wang, S. Yu, C. Xin, News-Oriented Keyword Indexing with Maximum Entropy Principle, *Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation*. 2003. pp. 277-281

[29] S. Li, H. Wang, S. Yu, C. Xin, News-Oriented Automatic Chinese Keyword Indexing, *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17*. Association for Computational Linguistics, 2003. pp. 92-97

[30] J. R. Quinlan, C4.5: Programs for Machine Learning, *Morgan Kaufmann Publishers Inc.*, San Francisco, CA, USA, 1993

[31] K. Ganesan, All About Stop Words for Text Mining and Information Retrieval (on Oct 6, 2014), http://www.text-analytics101.com/2014/10/all-about-stop-words-for-text-mining.html#sthash.V4f3jaL1.dpuf. [Accessed 17 Jun 2017]

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

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

[34] S. S. Singh, Statistical Measure to Compute the Similarity between Answers in Online Question Answering Portals, *International Journal of Computer Applications* (0975 – 8887) Volume 103 (15), October 2014