A Framework for Delivering e-Government Support

Goran Šimić, Zoran Jeremić

Military Academy, University of Defense
Pavla Jurišića Šturma 33, 11000 Belgrade, Serbia
e-mail: goran.simic@va.mod.gov.rs; zoran.jeremic@va.mod.gov.rs

Ejub Kajan

State University of Novi Pazar
Vuka Karadžića bb, 36300 Novi Pazar, Serbia
e-mail: kajane@acm.org

Dragan Randjelović

Academy of Criminology and Police Studies, Dušanova 196 Street, 11080
Belgrade, Serbia, e-mail: dragan.randjelovic@kpa.edu.rs

Aaron Presnall

Jefferson Institute, Kneginje Ljubice 28, 11000 Begrade, Serbia
e-mail: apresnall@jeffersoninst.org

Abstract: This paper gives a solution for improving e-Government services based on a hybrid approach: multilayered clustering of e-government documents based on fuzzy concepts and application of different text similarity measures. The goal is to reduce time between citizen's questions and government feedback, either completely eliminating or at least minimizing the deployment of subject matter experts. After the problem description, the paper describes step by step the functionality of the proposed system. At the end, concluding remarks emphasize some important features of the given approach and potential for future research.

Keywords: e-government; clustering; text similarity; fuzzy clustering

1 Introduction

Citizens' access and right to information at the level of local government is one of the essential ingredients for a successful government. Access to information empowers citizens to make decisions on the issues of government that affect them, decisions which provide critical feedback toGovernment as it seeks to meet the needs of citizens and improve their quality of life. Government should actively seek to capture the positive feedback loop inherent in providing greater access to information as a critical component of its strategy to deliverhigh qualityjust in time services for citizens and businesses. The use of advanced information technology to provide easier access to public information and government services is therefore a necessary condition for good governance. However, a number of challenges must be addressed to fully utilize the benefits of available technology.

A growing volume of information related to government rules, regulations, amended provisions, legal precedence and interpretive guidelines are distributed on a multitude of government portals, so that citizens can browse, search and take action. Some of these portals are equipped with search engines that provide text based search of documents. However, government documents are often very long andwithmany cross references to other related documents. Moreover, these documents are semi-structured with similar and often ambiguous content and terminology when taken as isolated texts out of context. As such, the characteristics of government document records make simple text search a serious impediment to understanding and use by common citizens.

Moreover, most of these portals are based on a one-way relation, in which the government produces and delivers information for use by citizens. This information is categorized or could be searched through a simple search engine providing keyword based search. The results of such a search could be a large number of documents the citizen has to go through to find desired information. If his knowledge in law and policy is limited, it could take hours to find the appropriate information.

Despite considerable attention to the introduction of ICT in government, most developed and developing countries have so far focused on the relatively easy phase of e-Government: developing websites, piloting a few applications, and putting these services online. Developed countries have been better able to invest in ICT infrastructure and service improvement, while developing countries must carefully evaluate the marginal utility of such investment. While the global trend remains one of steady improvement of e-government services in all regions, there is a growing gap of e-government development between developed and developing countries [2].

The 2012 world leader in e-government development was the Republic of Korea, which uses a single government portal as a gateway to services from multiple channels, organized by theme and subjects [2]. Many departments are integrated together through a powerful search engine offering an advanced categorizing function, which can list results by websites, services, and news. Mexico takes a different approach to integrate services provided to citizens. It provides a search engine that respond to users' specific search criteria, which has ability to filter information based on the information type, theme or user's location. Serbia significantly increased the performance of its e-government recently. The Digital Agenda Authority is responsible for introducing online services to improve the quality of services provided to citizens based on the "all services from one place" principle. The Authority created a portal, eUprava (http://www.euprava.gov.rs), which aggregates services and information from more than 27 governmental authorities, including municipal authorities.

Most countries from the European Union follow the approach of separate portals for their information, service and participation offerings. However, a recent trend in many countries is to set up portals that aggregate large amounts of information and services into a single website. A common approach includes organizing content around themes and/or specific audiences. These portals include search features that may index content from other government websites.

This paper proposes a novel approach to facilitate and foster e-government optimization and automation through the use of advanced information retrieval methods and techniques. In the next section, we describe the problem of the existing e-government solutions in Serbia. Section 3 describes a proposed alternative solution and a use case. Afterward, we provide a description of the design and implementation of the proposed solution. We conclude this work with a description of the potential benefits of the proposed alternative approach.

2 Problem Description

Serbia has achieved great improvement in the development of e-government over the last few years. At the beginning of 2007 a central portal of e-government services in Serbia was created (www.euprava.gov.rs). The main goal of the portal development was to provide a common access point for all e-government services provided to citizens, companies and public administration. Through this portal, citizens can download all relevant service documents, find links to a full range of public institution Web sites, and learn more about how to use e-government services in general.

However services for citizens and the private-sector in Serbia have not yet gone beyond a nominal level. Citizens can download forms; get laws, reports and other related publications. The information flow is unidirectional, from government to citizen, and if there is a need for a bidirectional flow of information, most institutions encourage communication by e-mail, where response times vary widely. Even though e-government has come a long way, there is room for improvement especially in the quality of provided services, including better knowledge management and improved two-way interaction between government and citizens.

There are two objectives that should be addressed when assessing e-government system in Serbia:

- E-government services should be designed around users' needs and provide advanced search capabilities that will enable better and easier access to information;
- E-government should enhance government services by reducing the administrative burden, improving organizational processes and using ICT to improve efficiency in public administration.

Search features provided by an e-government portal are commendable, but do not meet either of these objectives. Citizens without expert knowledge in the domain of inquiry are often disappointed by the difficulties that must be overcome and efforts they have to make in order to access or gather the requested information, and ultimately by the lack of effectiveness in orchestration of the various procedures. Domain experts must be engaged to examine the various cases and select the appropriate service or information requested by citizen. Typically, such a scenario would consist of the following steps: a citizen makes a request for specific information elaborating her specific case through the use of email or phone; a government officer receives the request, examines the request, clustering the nature of the problem and sends it to a specific domain expert; the domain expert evaluates the citizen's case and prepares the requested information.

Some of the citizen's cases are obvious, related to previous cases or they are clearly and fully described in one document. However, collecting and consolidating all the information could be very difficult and time consuming. The domain expert may have to search across many documents, to search for relations between documents and the specific case or to compare the case with previous solved cases.

Moreover, the limited number of government employees limits the number of citizen requests that could be processed without significantly increasing costs. In addition, most of the services that the government portal offers declare a wait time of 2-6 weeks, which is not acceptable. The transformation to the improved, more intelligent system could deliver a drastic reduction on the average response time for a request from a citizen.

E-government should provide a solution to manage a citizen's requests by documenting and tracking it through to the final resolution. It should use previous cases and resolutions to provide a faster, smarter and more efficient response to a citizen's request. Only in specific and new cases or if the citizen is not satisfied with the proposed resolution, should the human government officer be asked to intervene.

The approach we suggest aims to enable interactive processes that are simple, effective, and based on the user's needs and capabilities, rather than the government's organizational structure or government business models. It should create the opportunity to evaluate and eliminate redundant or unnecessary steps and processes as well as to reduce costs and cycle times by transitioning from the processes mainly based on human-related work to automated and more intelligent user centered processes.

3 Advanced Answering Engine (ADVANSE) for Interactive E-Government Services

E-government systems under consideration here are designed to assist citizens in making decisions. Citizens ask questions and the system tries to response with an appropriate answer to inform the citizens' decision or next step. As described in Section 1, in the current e-government systems of Serbia, these questions are answered by Subject Matter Experts (SMEs). The response time varies from one to several days, depending on the availability of the SME. On the other hand, a number of questions and answers accumulate over time. They could be considered as a kind of the knowledge base (KB). This KB could be captured and applied as a good basis for development of the advanced answering engine (ADVANSE).

3.1 Scenario of Use

Interaction between the citizen and the system happens in three phases (Figure 1). At the beginning, the system offers groups of key terms (phrases and words) to the citizen. The citizen can select one (or not select any) according to her question. In the next step, the system delivers her the set of questions that are strongly related to the selected key terms. The citizen has two options in the second phase: to choose a question from the list, or to enter a new one. Regardless of which option is chosen, the system delivers an answer in the next phase.

If the citizen selected the 1st option, the system delivers an answer that is fully matched to the selected question. Otherwise, the system finds an existing question that is the most similar to the new one. The system sends this question's answer back to the citizen. The citizen can evaluate this answer. After this three steps interaction, she can continue, or finish the session with the system. If she is not satisfied with the answer she can try to find another question or enter the new one.

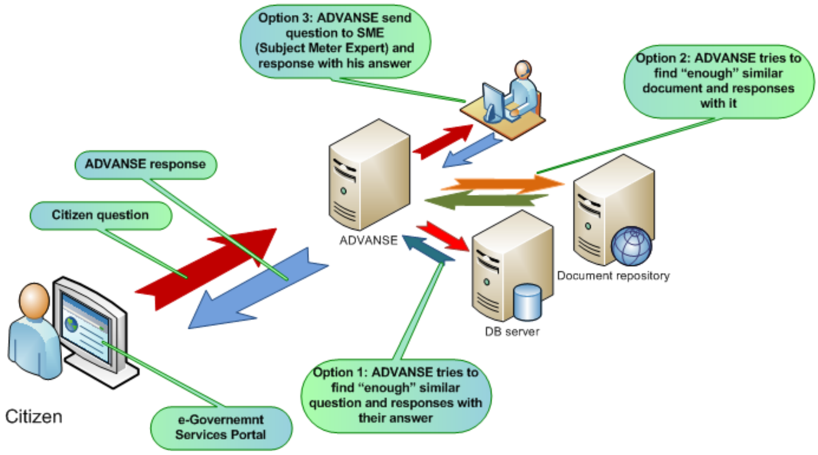

Option 3: ADVANSE send question to SME (Subject Meter Expert) and response with his answer
Option 2: ADVANSE tries to find "enough" similar document and responses with it
ADVANSE response
Citizen question
ADVANSE
Document repository
DB server
Citizen
e-Governemnt Services Portal
Option 1: ADVANSE tries to find "enough" similar question and responses with their answer

Figure 1

Interaction between citizens and system

The case in which the citizen writes a new question represents the focus of the research because the system tries to find the most appropriate answer. The groups of key terms (mentioned in the first step of interaction) are generated automatically by a clustering process. The questions and formal documents belong to these clusters (according to similarity of the terms and words they contains). When the citizen enters a new question, the system calculates the similarity with questions in the determined cluster. If the similarity threshold is satisfied, the system delivers the best fitted answer to the citizen. Otherwise, it can deliver the related document and / or the safety answer (usually it is a message with an appropriate explanation, recommendation, or references to other resources). The citizen can follow the steps offered, or change the way of interaction.

3.2 Content Representation

There are three basic concepts (content) in the system: citizen questions, governmental documents, and answers. They are separated into different layers (Figure 2). The answers' layer is between questions and documents (QD) layers. QD are clustered by key terms. There are as many clusters as key terms in the domain dictionary. The answers are not clustered because of two reasons: they are excluded from searching and their concept has double purpose. An answer can be manually created by a subject matter expert or automatically generated by the system by making an association between the question and related documents. Formed associations can be of different types. Each question can be related to one or multiple answers and each answer can be relevant for one or more questions. The same approach is applied for associations between answers and documents.

Questions and documents can contain more than one key term. If the content is large or more general, there is more chance for it to score a hit. The question or document then belongs to more than one cluster. Therefore, the clusters can be represented as the sets that are intersected with each other and the questions and documents that contain more than one key term belong to these intersections.

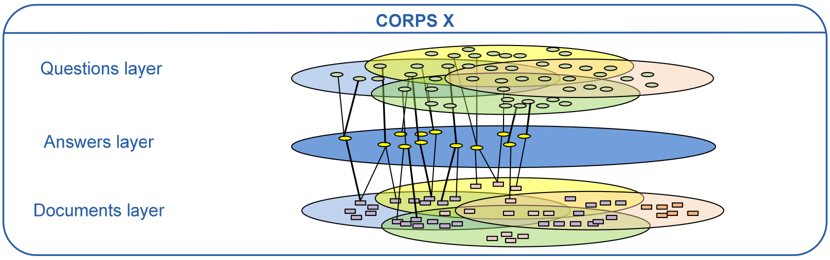
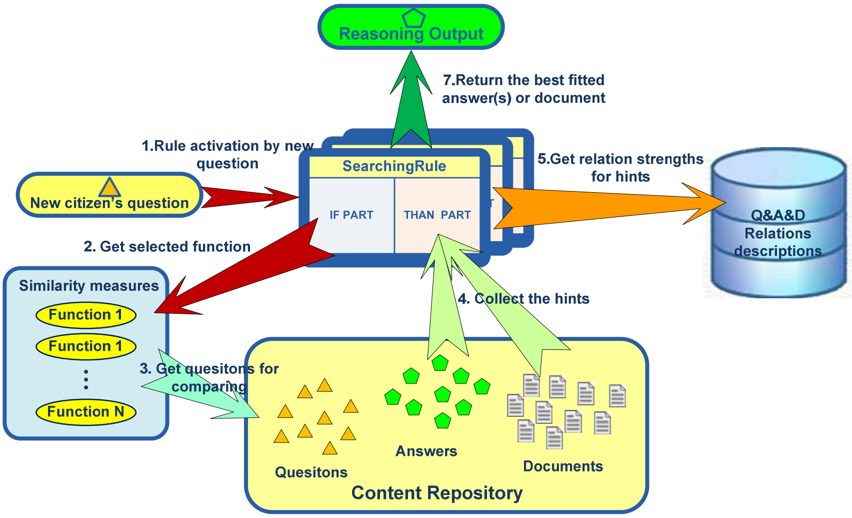


Figure 2
Layered content representation

The strength of relations between the content from different layers depends on citizens' satisfaction with the feedback. This value is calculated from two factors. The first one is degree of similarity between nodes and another is from the citizen evaluation of the system response. The degree of similarity between the answers and documents is calculated by the system and it represents an objective value (e.g. cosine similarity). The strength between questions and answers are continually changing and depending on the score given by citizens. This is a kind of content evaluation which depends on citizens' individual expectations and attitudes. Therefore, this measure has a subjective nature, but it becomes more objective with more citizen feedback.

3.3 The Role of the Rules

Besides the questions, answers and documents, rules represent another part of the system knowledge (Figure 3). Rule based reasoning is used for two purposes: separating business logic from heterogeneous data and separating the similarity measuring from decision making. There are two rule types in the system: searching and creating rules. Searching rules provide a flexible way for reasoning on similarity between questions and existing content. The final decision about responding to an answer or document is the result of rule based searching. Searching rules are designed for finding questions and answers existing in the system similar to the citizen question. There can be more than one similarity measure (algorithm) in the system. They are implemented as functions which are invoked by statements in the rule premise. This way the reasoning about similarity can be changed by using different algorithms. Further, the function compares the new question with the existing ones and returns the best fitted question. This question is forwarded to the rule action part as a parameter of another function there. This function finds and returns the appropriate content (answers or documents) that represents the final result of the reasoning.



Fi

Figure 3

Relations between rules and system concepts

The creating rules are designed for making relations between questions, answers and documents. Both types have the question sets on the left hand side and sets of triplets (questions, answers and their relations) on the right hand side. The rules are generated by the system by using those data. In contrast, creating rules are used when a new relation between question and answer has to be established. These rules do not change the content but the connections between questions and answers (see the last paragraph in Section 3.2).

The system complexity is reduced by using the rules. The complex functions designed for different purposes (measuring similarity, for calculating the relation strength and clustering) are embedded. The rules just contain the function calls in the premises, or in the action parts.

4 Design and Implementation

The main focus of the conceptual model is the citizen's question (Figure 4). This question can be in relation with one or more answers, or/and one or more documents. The question is presented in the model with the Citizen Question concept, whether it has been answered or not. This is a main concept in the system because processing of questions is the top objective of the system.

The questions and documents belong to the clusters. The answered questions are in associative relations with the answers. The answer concept has a dual nature. It can consist of the text written by an SME or automatically generated recommendations – links to the documents appropriate to the citizens' question. In this case it depends on document(s). The relations between answers and questions are weighted. The strength of the relation has a default (initial) value that is changeable by the citizens' feedback about her satisfaction with the answer. Therefore, relations are represented by the QASTriplet (question – answer – relation strength) concept.

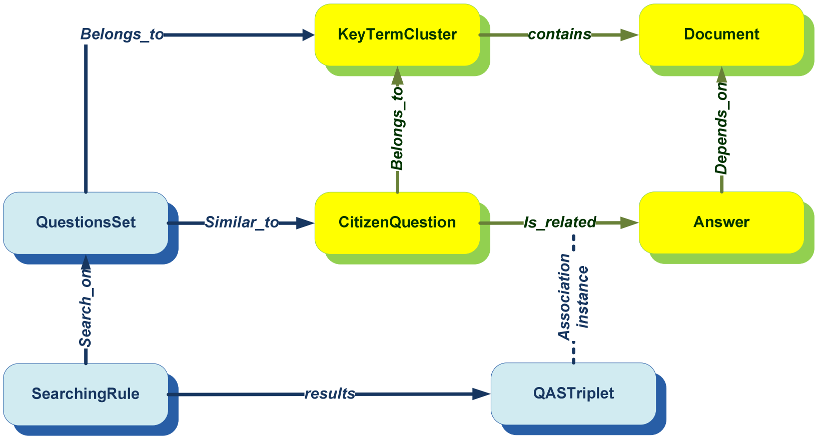


Figure 4

Basic conceptual model

Another part of conceptual model is designed for reasoning purposes (blue colored). The central concept in this section is the Searching Rule. As mentioned in the previous chapter, rules are used for providing flexibility to reasoning. The set of questions, which is searched by the function embedded in the rule's left hand side, is represented by the Questions Set concept. Since different similarity functions (measurements) can be used, the search results can be different and overall system behavior is flexible for that reason. The purpose of finding the most similar question to the new one is to get a related question – answer – relation strength triplet. This data structure is represented by the QASTriplet concept. If the rule is fired, the function on the rule's right hand side returns the triplets that are related to the resulted question. Detailed discussion about how the system uses the rules is in the following section (Section 4.3).

4.1 Clusters Generation

The clustering initialization could be performed in different ways: Random partition [3], *Forgy* partition [1] and *Kaufman* [6] are the most commonly mentioned. The *Kaufman* method does not need a predefined number of clusters, but the other two do. In ADVANCE the number of clusters is predefined: it is determined by the number of used key terms. Questions and documents are fuzzy clustered. The developed method (Equation 1) is based on the ideas of *fuzzy c –*

means (FCM) algorithm [5], [13]. Different of FCM, a concept of distances is avoided because key terms are used as constant values instead of the iterative calculation of some statistical value of central tendency every time the set of observations is changed (Equation 1).

$$f_{fcm} = \sum_{i=1}^{N} \sum_{j=1}^{K} m_{ij}(x_i) \qquad (1)$$

In this way, a membership function (m_{ij}) of document or question (x_i) represents the only measure of its belonging to the particular (j-th) cluster. It is calculated just one time and there is no need for recalculation every time a new question or document is added into the system. This approach is found useful because there is a need for permanent adding of new questions into the system.

The multiplicative nature of the algorithm expresses the fact that every data portion belongs to every cluster in some degree. If there are K clusters and N questions or documents (they belong to separate layers), the$K \times N$matrix of values of the membership function is formed. Considered dynamically, the addition of a question or document produces the$K \times (N + 1)$matrix – one column is added into the existing matrix.

ADVANSE calculates the membership function $m_{t,q}$ (where t represents the clusters' key term and q represents the question or document) by using both the term frequency ($tf_{t,q}$) and the inverse document frequency ($idf_{t,q}$) [11], [8] (Equation 2). The first ($tf_{t,q}$) represents the internal characteristic – how many occurrences ($f_{t,q}$) of the specified term t there are in the particular content q (question or document). The other measure ($idf_{t,c}$) is on the global level – how many questions that contain the specified term ($N_{q,t,c}$) are there in the whole corps ($N_{q,c}$). The product of these two values is commonly called the TF-IDF function.

$$m_{t,q} = k \cdot tf_{t,q} \cdot idf_{t,c} = k \cdot \log(f_{t,q} + 1) \cdot \log \frac{N_{q,c}}{N_{q,t,c}} \qquad (2)$$

Logarithm functions are used for normalization purposes. This way the membership function value varies in range from zero to one. There is an additional coefficient k – correction factor that provides better dispersion of the values of $m_{t,q}$ in the range. This coefficient is calculated as the reciprocal of the TF-IDF function maximum.

The described method provide for flexible behavior of the system. Every new question or document added into the system can be processed particularly. There is no need for repeating the clustering initialization completely. The membership values are calculated and simply stored as metadata ready for filtering purposes (e.g. a question or document can be taken under consideration depending on the threshold – changeable minimum value of the membership function). Only changes in key terms will produce re-initialization of clusters.

4.2 Algorithm Description

The proposed algorithm is described with an activity diagram (Figure 5). As mentioned above (Section 3.1), the system activities are performed in three steps. The citizen's question is processed in the first phase. This activity starts with steaming and elimination of stop words (question filtering). If the citizen selects the term(s), the question is added to the existing cluster specified by the selected term(s). Otherwise, ADVANSE performs a measurement of similarity between the new question and cluster centroids (cluster determining). By using Cosine similarity, ADVANSE compares the vector of question terms with the vector of key terms that belongs to the cluster (Section 4.5).

After the cluster is determined, ADVANSE starts the last activity in the first step. The engine gets the questions from the cluster, one by one, measuring the similarity with the new one. When the question is found, ADVANSE starts the second phase. If none of the clusters satisfy the threshold criteria, ADVANSE tries to find a document that is closest to the search criteria.

If a similar question is found, ADVANSE starts searching for the most appropriate answer. Questions, answers and their mutual connections are forming triplets. One question can be connected to many answers, but connections between them could have different connection strength. Connection strength and threshold are used for selection of an answer that best fits the question (answer finding). The answer with the highest connection strength will be selected if the connection strength is above the threshold. The last activity in the second step is content delivering (responding with a document or an answer).

In the last step, the system checks if there is citizen feedback. Feedback includes evaluation of the system response. ADVANSE processes feedback in two ways. If the feedback is related to the document, ADVANSE updates the weights of the document's key terms (Section 4.7), i.e. positive feedback increases the weight, while negative feedback decreases the weight. In the answer case, if feedback is positive, ADVANSE increases the strength of the Q & A relation and vice versa. Due to the term weighting mechanism, documents are clustered just one time (during initialization).

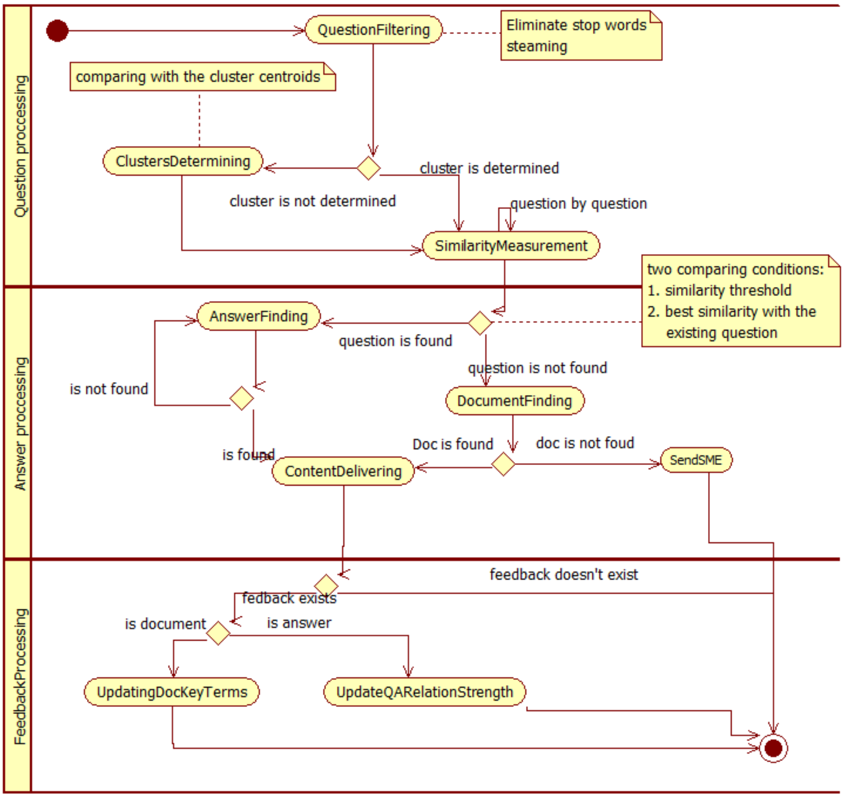


Figure 5
Processing algorithm

4.3 Concrete Knowledge Representation

In actual e – government systems, the SME suggests that citizens read answers to similar questions. This case is very common if there are a lot of questions and answers in the system. However, there may be one or more questions related to the same answer. If there is only one answer related to a question and the similarity between the cluster's questions and the new one is higher than the threshold value T, this answer is the only solution for a system response. However, this is not a common case. More often, the questions in the cluster can be related to different answers. Even more often, some questions could be related to more than one answer. The problem that arises here is how to select the appropriate answer to the given question.

The similarity can be functionally expressed as a maximum value of the relations between the questions in cluster C_p and the new one q_{new}(Equation 9).

$$sim(q_{new}, C_p) = \max_i \{ sim(q_{new}, q_i) | \; q_i \in C_p \} \tag{9}$$

where $\left| \; C_p = \{q_1, q_2, ..., q_m\}, i = 1,...,m, \; p = 1,...,s \right.$

Based on this representation and the conceptual model, the following rule design is used in the system (Equation 10).

$$(\exists C_p)\,(sim(q_{new}, C_p) \geq Treshold) \Rightarrow \{a_j \left| \begin{array}{l} (\exists t_{ij} \in T_p)(\exists q_i \in C_p) \wedge t_{ij} = (q_i, a_j, s_{ij}), \\ i = 1,...,m, \\ j = 1,...n\} \end{array} \right. \tag{10}$$

where $\left| \; T_p = \{(q_1, a_1, s_{11}), (q_2, a_1, s_{21}), ..., \{q_m, a_n, s_{mn}\}, i = 1,...,m, \; j = 1,...n\}, \; p = 1,...,s \right.$

The concept of triplets is introduced in the proposed solution. Every triplet is generally represented as a set {q_m, a_n, s_{mn}} where q_m represents the question that belongs to the corpus C_p, a_n represent the answer related to the question $q_{m,}$ and s_{mn} represent the strength of the relation between q_m and a_n. Searching for the most similar question can start if the cluster (C_p) that the new question belongs to is determined. There is an additional precondition for rule firing beyond similar questions in the cluster: the similarity has to be higher than the specified threshold. If both of these conditions are satisfied, the rule returns the answer related to question (q_m) that is the most similar to the new one (q_{new}). If multiple answers are connected with the selected question q_m, the one with the highest strength value will be chosen. However, other answers sorted by strength values will also be offered as alternatives to the recommended one in case the citizen is not satisfied with the offered answer.

The other consequence of searching is establishing new relation(s) between the new question and the answer(s) that the system replied with. The creating rules are designed for this purpose (Equation 11).

$$(\exists a_{res})(f(a_{res}) > 0) \Rightarrow (t_{new} = (q_{new}, a_{res}, s_{new}) \wedge T_{new} = T_p \cup \{t_{new}\}) \tag{11}$$

where a_{res} is responding answer, f is feedback function and t_{new} is a new triplet of a new question (q_{new}), the responding answer and new strength (s_{new}).

The new question is added into the cluster and the new question – answer – relation strength triplet is added to the knowledge base.

4.4 Measuring of Questions' Similarity

After the cluster is determined, ADVANSE uses the cluster's questions and compares them with the new one. Three similarity measures with different approaches are used for this purpose: *cosine* similarity, *Jaccard* correlation coefficient and averaged *Kullback-Leibler* divergence. They are implemented as functions that are used by searching rules (Section 3.3). Common for all of the measures is that the question's text is represented as a set or a vector of terms and their additional properties such as term frequency, hash function, distribution of probability. A short description of the applied measures is presented in this section.

Cosine similarity is one of the most commonly used text similarity measures because of its simplicity and because it is not dependent on the text length. The compared texts are considered as resulted term vectors [9] and their similarity is expressed as a cosine of the angle between these vectors.

Jaccard Correlation Coefficient (JCC) is a similarity measure that depends on set theory [4]. Questions are considered a set of terms, and correlation between two of them is calculated as a ratio between the intersection and union of their sets. The zero value is calculated if there are not terms in the intersection. JCC has a maximum value (1) if both of the documents have the same term sets. The properties of terms are expressed by values of the appropriate hash function randomly selected from the hash function set (universal hash family for strings). If there is collision between hash values, the system performs rehashing. Different lengths of the questions are solved by padding the shorter document with zeros before comparison.

Averaged *Kullback-Leibler* (KL) divergence is a similarity measure that depends on probability theory [12]. Compared questions are represented by probability distributions of the terms they consist of. Because KL divergence is non – symmetric (the result depends on the order of comparison), averaging is used as a technique for compensation.

Using different similarity measures provides more scalability to the system. It is important because there are different domains (e.g. health, finance, low) in which these functions can be used. System performance can be evaluated during usage of different functions and the most appropriate function can be selected for searching purposes.

4.5 Q & A Relation Strength

Each time a new question-answer pair is created (see section 4.3) ADVANSE assigns an initial strength (V_{ini}) to their relation. This value is changeable and depends on a number of feedbacks and a feedback score (Equation 12). The relation strength is considered for ranking the answers related to a specified question.

$$S_{qa} = V_{ini} + I_f \cdot R_{pn} \tag{12}$$

The product of two factors: feedback importance (I_f) and positive & negative ratio (R_{pn}), represents the other part of the strength equation. It becomes important when the number of feedbacks exceed the threshold value. The feedback importance is the weighted factor included for this purpose. It is calculated on the normalized way (Equation 13):

$$I_f = 1 - \frac{1}{\sqrt{N_{pf} + N_{nf}}} \tag{13}$$

where N_{pf} is the number of positive feedbacks and N_{nf} is the number of negative feedbacks.

The importance value is greater than zero if the number of feedbacks is more than two. If the number of feedbacks grows, its value tends to be one. ADVANSE uses a threshold mechanism to define the number of feedbacks necessary for the strength calculation. In other words, if there are fewer feedbacks than defined by the threshold, the relation strength equals its initial value.

Another factor product of the relation strength is a ratio of positive & negative feedbacks (R_{pn}). It is calculated by dividing the sums of positive (M_p) and negative (M_n) grades with the whole number of feedbacks (Equation 14). The R_{pn} ratio is positive if the sum of positive marks is greater than sum of negative marks. If there is an equal number of positive and negative feedback, it has a zero value. Otherwise R_{pn} has a negative value.

$$R_{pn} = \frac{\sum_{i}^{N_{pf}} M_p - \sum_{i}^{N_{nf}} M_n}{N_{pf} + N_{nf}} \tag{14}$$

If the value of feedback importance (I_f) is below the threshold, the R_{pn} ratio is not included in the relation's strength calculation. Otherwise, the ratio value is included and decreased by the factor I_f. In practice, it is easier for citizens to get feedback by selecting one of two options than to evaluate the answer by selecting one of many grades. If there are only two feedback options (e.g. satisfied or not satisfied), the calculation of R_{pn} is simplified (Equation 15).

$$R_{pn} = \frac{N_{pf} - N_{nf}}{N_{pf} + N_{nf}} \tag{15}$$

In this case, the ratio's value is normalized in the range from -1 to 1. If there is not even one feedback, the relation strength is represented only with its initial value feedback (V_{ini}). Otherwise, the feedback importance (I_f) and ratio of positive and negative feedbacks (R_{pn}) are included in the calculation.

4.5.1 Responding with Documents

If the new citizen's question is not similar enough to any of the existing questions containing an answer, ADVANSE tries to find similar document(s) instead of answers. This measuring is performed in the same way as in the case of questions (Section 4.4). The documents are described by key terms that are statistically extracted during the clustering phase. They are clustered in the same way as the questions (Section 3.2) and they are presented as a set of key terms and their membership functions.

If the appropriate document is found (the similarity between the citizen's question and the document is greater than the threshold), ADVANSE returns a link to the document. Then, a relation between the question and the document is established. The initial strength is given and it is changeable over time depending on the citizens' feedback.

Conclusion

The represented hybrid solution depends on the nature of the e – government services. It is mainly focused on providing conditions for advanced responses to citizen requests. Most of the information that can be used are held in repositories as formal documents. Otherwise, citizens' questions and SME answers are recorded in the system DB. Therefore, the ADVANSE content model is layered. The content is fuzzy clustered based on fuzzy sets theory and fuzzy c-means algorithm. The boundaries between clusters do not exist and pieces of information can belong to more than one cluster. On the other hand, the questions, answers and documents are semantically connected in the system. The mentioned features have influence on the overall system design and they make the system flexible in responding to citizen queries. Using different text similarity measures provides adaptive behavior to the system. Processing of the citizens' questions in different and flexible ways provides the conditions for early high – quality responses. The expectation is that the citizens will be much more satisfied than before, able to make better decisions with better information, while the public administration will have captured more systematic information on the problems citizens face.

Improved e-government services in response time, quality of response (citizen's satisfaction) and the usage of existing content in a new manner as well as relaxing the SME responsibility for answering citizen questions represent the main results of the ADVANSE project. Adaptive response represents one of the most important features of ADVANSE. The questions on one side and the answers and documents on another are related. These relations are changeable and they depend on citizens' evaluation of the response (their satisfaction with the delivered content). Thus, the response to the same question can be different over the time as conditions and citizen needs change. Documents are delivered in response to a situation where there is not any suitable answer to the citizen question (if the threshold of similarity between the new and existing questions is not satisfied). ADVANSE respondswith documentsthat containa key term set most similar to that of the question.

Future objectives will focus on testing in different environments and if necessary, to improve the ability of adaptation. The functioning of the system in multilingual environments will be the one of the targeted solutions. The dictionary and grammar of different languages enlarge the complexity of the system independent of the domain of usage. In this case, several processing strategies should be used. Functioning of the system in different e-government domains represents the other challenge. Different domains are covered with different thesauruses. More specialized similarity techniques will be required. Document processing will also need to be improved.Annotation (tagging) can provide a response with extracted part(s) of a document instead of the whole document body.

The presented solution is a part of a wider project aiming to provide an intelligent decision support system able to collect, cluster and analyze data from various data sources (social, biological, and economical systems) in order to make government decisions easier. Future research will take place in several directions, such as the improvement and evaluation of information retrieval and text mining algorithms, allowing personalized services by applying user profiles, implementation of morphology data of different languages for better text preprocessing and establishing semantically based relations between pieces of information.

Acknowledgement

This work was supported by the Ministry of Science and Technological Development of the Republic of Serbia under the grant III-44007.

References

[1] Garijo, F., Riquelme, J., Toro, M.: A GRASP Algorithm for Clustering, Proceedings IBERAMIA 2002, LNAI 2527, pp. 214-223, 2002

[2] Global E-Government Survey 2012, E-Government for the People, United Nations, New York, pp. 9-69, 2012 retrieved from http:// unpan3.un.org/egovkb/global_reports/12report.htm

[3] Hamerly, G. and Elkan, C.: Alternatives to the k-Means Algorithm that Find Betterclusterings. Proceedings of the Eleventh International Conference on Information and Knowledge Management (CIKM), pp. 600-607, 2002

[4] Hamers, L., Hemeryck, Y., Herweyers, G., Janssen, M., Keters, H., Rousseau, R., Vanhoutte, A.: Similarity Measures in Scientometric Research: the Jaccard Index Versus Salton's Cosine Formula Information Processing and Management: an International Journal, Volume 25 Issue 3, pp. 315-318, 1989

[5] Bezdek, J., 1981, Pattern Recognition with Fuzzy Objective Function Algorithms, Kluwer Academic Publishers Norwell, MA, USA, 1981

[6] Laan, M., Pollard, K., Bryan, J.: A New Partitioning Around Medoids Algorithm, Journal of Statistical Computation and Simulation 73, No. 8, pp. 575-584, 2003

[7] MacQueen, J.: Some Methods for Classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, pp. 281-297, 1967

[8] Robertson, S.: Understanding Inverse Document Frequency: On Theoretical Arguments for IDF, Journal of Documentation, Vol. 60, No. 5, pp. 503-520, 2004

[9] Salton, G. and Wang, A.: Generation and Search of Clustered Files. *ACM Transactions on Database Systems*, Vol. 3, No. 4, pp. 321-346, 1978

[10] Vattani, A.: K-means Requires Exponentially Many Iterations Even in the Plane, Discrete and Computational Geometry Journal, pp. 596-616, 2011

[11] Wu, H.C., Luk, R. W. P, Wong, K. F., and Kwok, K. L .: Interpreting TF-IDF Term Weights as Making Relevance Decisions, ACM Transactions on Information Systems, Vol. 26, No. 3, Article 13, 2008

[12] Zhang, W., Ma, J., Zhong, Y.: UsingKullback-Leibler Divergence Language Models to Find Experts in Enterprise Corpora, IITAW '09: Proceedings of the 2009 Third International Symposium on Intelligent Information Technology Application Workshops, pp. 402-405, 2009

[13] Bezdek, J., Enrlich, R., Full, W., FCM: The Fuzzy c-means Clustering Algorithm, Computers & Geoscience, Vol. 10, No. 2-3, pp. 191-203, 1984