ON INTELLECTUALIZATION OF MANAGEMENT OF ELECTRONIC DOCUMENTS

Makrufa Hajirahimova

Institute of Information Technology of ANAS, Baku, Azerbaijan
makrufa@iit.ab.az

Introduction. The main goal of the electronic government, which is one of the main elements of Information Society (IS), is to increase the level of services rendered to citizens by government institutions, simplify the access to government information resources, provide active participation of all social groups in government administration and consequently achieving the increasing of administration efficiency by using the capabilities of Internet and information-communication technologies (ICT) [1, 2].

In e-government environment, the most important of all applicable issues (access of citizens to government information, network possibility of majority of standard transaction between citizens and businesses) is the transfer of records management and circulation of all documentation into electronic form in government and administration, as well as local self-government institutions. As a result, the role of electronic documents becomes more significant and displaces paper documents from traditional carriers. Legal and regulatory framework of records management develops, electronic services with application of electronic signature providing the legal base of electronic documents is improved [3].

Application of electronic documents management allows reducing the time and effort spent on preparation and processing of documents, making administrative decisions, increasing of executive discipline and simplification of control mechanism, efficiency in service to population, accessibility of services and information about activity of government and self-government institutions, expediting the connection with its subordinate and other organizations, and receiving economical efficiency (efficiency=result/expenses).

Although Azerbaijan is not in the list of leader countries in creation of e-government and issues related to circulation of electronic documents [4], some works are being performed in this field. Each of government institutions is performing their administrative functions within the framework of their authority by subordination. Works such as development of web-sites of government institutions, placement of information and necessary documents about these organizations and their structure, FAQ sections, publication of vacancy announcements, payment of some public utilities online, submittal of documents for college entry examinations, creation of electronic libraries etc have been performed.

In documents management system, a document is unit of information. As always, paper or electronic document plays a key role as an information carrier in both government and private institutions. Documents accompany all business-processes in the organization and provide informational support for making correct decisions in institutions [5-8]. Efficiency of control of business processes in organizations immensely depends on organization of document control. Main principals of document control are reflected in ISO 15489 [9].

At all times, modern technologies were applied in documents control and automatization requirements of this process have changed. Lately, paper documents have been replaced with electronic documents. Rapid increase in number of electronic documents is observed on a yearly basis. Analytical companies have statistical information and forecasts regarding transfer to electronic documents from paper documents. IDC (International Data Corporation) analytical company forecasts following proportions of paper documents to electronic documents in world documentation: 30% electronic, 70% paper documents in 2004, 50% electronic, 50% paper documents in 2005, 30% paper, and 70% electronic in 2010 year.

Approximately 6 billion new documents are developed on yearly basis in the world, and an employee spends approximately 150 hours of his working time to develop, send and search

for necessary documents. According to an existing evaluation, manager spends 45%, reader spends 75%, and ordinary employee spends 30% of this working time to work with documentation [10].

Currently, there are electronic document circulation systems meeting modern requirements and defined as EDMS (Electronic Document Management Systems) in English language references. These systems differ in functional and technical capabilities. Systems with efficient storage and search capability are called electronic archives, and systems providing the flow of the documents are called – Workflow systems. Some systems hold both of these capabilities. Application of such systems for service to population and businesses is one of its services, and application for processing of requests from populations is one of its social aspects.

Problems of electronic documents management systems: As majority of EDMS have a corporate character there are restrictions in number of users and their access. Essentially, a government is a larger corporation administered based on the identical order and using the same tools. Multiplication of volume of circulated and stored corporate information on EDMS, is a "characteristic indication of modern day". According to calculations of IDC, the volume of processed and stored corporate information increases by 70% on a yearly basis, 3 thousand employees send 3 terra bites of information through their e-mails on a daily basis [10]. Document management in e-government has a larger scale. A very large number and volume of documents incoming from ministries, institutions and organizations, as well as corporate information systems of private sector are circulated and processed. Document flows differ by their source, as well as their presentation forms. Unlike the corporate environment, e-government is an open and transparent environment, the number of its users is unlimited, services and access to government documentation for users is carried out in interactive mode independent from location and time. Creation of a more intellectual circulation of electronic document in such environment is one of the most important issues. Several problems seem more important in solution of this problem:

Problem 1 – exponential increase of volume of electronic documents. This situation creates large obstacles in operative processing and systematization of documents. The search result of necessary document in a growing archive is unsatisfactory. A long list of documents is presented as the response to the request. Choosing the most relevant document from this list results in a great time loss and seriously affects the psychological state of the person.

Problem 2 – Unstructured condition of majority of electronic documents.

Information can be in structured (VB tables) and unstructured (text, audio, video etc) form. Unstructured information forms the majority of documents circulated in e-government environment (laws, decrees, letters, requests, contracts, reports etc). Text documents forms the majority, approximately 80%-90%, of data, which creates obstacles in operative analysis of their content. Reading, studying and mastering of large text documents by managers is impossible both time-wise and physically. In turn, this results in delays in delivery of documents to their destination, and sometimes failure in delivery.

Problem 3 –Lifecycle management of electronic documents. This problem is related to the rapid increase of volume of documents, and obstacles created by the changes in requirements on their business value, accessibility and protection over the period of time.

Undoubtedly, given problems cannot be solved by traditionally approaching the automatization of documents management in e-government environment. Intellectualization of the system through the methods of intellectual analysis, as well as performance of researches in development of intellectual processing and circulation system of electronic documents is required.

Intellectual processing of electronic document. As noted above, types of documents circulated and stored in EDMS are different. Considering that over 80% of circulated documents are text type documents, solution of following problems is considered: automatic performance of organization of automatic classification, indexing and circulation of text-type electronic

documents by their content by the system without a human factor; management of intellectual search and life cycle of electronic documents.

Important issues related to working with text-type electronic documents are their classification based on their content and search of the documents based on their content.

Unlike the Data Mining technology [11, 12] used for detection of data in structured information, Text Mining [13] is the most absolute technology for performing intellectual operations on text-type documents. Text Mining allows to detect characteristic elements or features in documents using certain algorithms in text type documents, determine the belonging of the documents to one group or another, as well as performance of a higher intellectual search (semantic search) of documents. Using the capabilities of Text Mining technologies are essential for performance for intellectual capabilities of EDMS. The main issues solved by Text Mining are following: classification of text documents (Classification); Clustering of text documents (Clustering); Information Extraction.

In classification method, documents are firstly grouped in the existing classification scheme based on certain indications. Probability models are more often used in classification of text documents. *Bayes, KNN (k Nearest Neighbor)* etc are the simplest classification methods. There are also classification methods based on fuzzy model and ontology.

Unlike the classification method, clustering method classifies the documents without certain pre-determined classification schemes' environment. Even the number of clusters is not known in advance. Certain classification is achieved as a result of training. There are a number of clustering algorithms. For example, k-means, LSA (Latent Semantic Analysis), Suffix Trees etc [14, 15].

As in search systems, detection of the most relevant document from the multitude of documents is the most important condition in EDMS. Relevance is the similarity of the document content to search. The classification systems significantly simplify the efficiency of the search, because the search is carried out in certain groups of documents, instead of the multitude of all documents.

Multiagent technologies are also vastly used in issues such as automatic passage and search of information in distributed information systems [11,16].

It is also possible to apply the content classification method in solution of automatic address delivery problem of the documents in EDMS.

Conversion of texts into summary form is considered as a prospective direction in classification issues and is broadly used. Conversion of texts into abstract is a process of creation of summaries by keeping their content. While studying the abstracts of the documents instead of their originals, one can gain sufficient information in a short period of time. This creates a suitable working environment with large volume of information [16-18]. TRM (Text Relationship Map), LSA and other methods are used during conversion of texts into abstract.

It is clear that, information is valuable for any organization and the value of the information decreases over time. Considering the business value of information and its change over time, it is possible to make the document storage control more efficient. This approach is called ILM (Information Lifecycle Management) concept [19-21]. Based on this concept, necessary information must be stored in more high-speed, reliable and protected storage system, and information of less importance must be placed in cheaper and lower-speed storage system. Unnecessary information is automatically deleted from the system. This condition is carried out as a cyclic process

It is possible to apply several technological solutions for performance of this operation: automation migration of data from one class storage system to other, mirror reflection, reserve copies and archiving devices.

Application of noted methods in information systems, as well as ESDS, improves and intellectualizes functional and analytical characteristics of these systems.

Conclusions. As the majority of circulated documents in e-government environment are text type documents, so, it is essential to develop methods and algorithms for the lifecycle

management of electronic documents and automatic performance of document circulation without human factor, automatic classification and intellectual search of text type documents based on the content of the text-typed documents connected with definite problems. These methods and algorithms can be applied to both newly projected and existing systems.

References

1. В. Дрожжинов, А. Штрик. Электронные правительства информационного общества. PC Week / REN 15, 2000.
2. R.M. Əliquliyev, Y.N. İmamverdiyev. Etibarlı və təhlükəsiz eletron hökumət yaradılmasının bəzi məsələləri, / Proceedings of the second International Conference on "Problems of Cybernetics and Informatics", October 24-26, 2006, Baku, Azerbaijan, vol.2, pp. 80-82.
3. R.M. Əliquliyev, Y.N. İmamverdiyev. Rəqəm İmzası Texnologiyası, Bakı, Elm, 2003, 130 səh.
4. UN E-government survey 2008.
5. R.M. Əliquliyev, T.X. Fataliев. Korporativ şəbəkə mühitində elektron sənədlərin dövriyyəsinin avtomatlaşdırılması sistemi, //AMEA Xəbərləri, 2001, №3, c. 47-49.
6. Р.М. Алгулиев, Т.Х. Фаталиев, М.Ш. Гаджирагимова. Интеграция корпоративных систем документооборота и архива, //Известия НАНА, 2005, №2, с. 12-17.
7. Е.А. Сидорова, Ю.Ф. Загорулько, И.С.Кононенко, Ю.В.Костов. Подход к интеллектуализации документооборота, //Информационные технологии, 2004, №11, с. 6-13.
8. А.В. Андрейчиков, О.Н. Андрейчикова. Интеллектуальные информационные системы, Финансы и статистика, 2004.
9. İSO 15489 - Information and documentation - Records management
10. IDC, Europe Document Management market review and Forecast, 1998-2003.
11. В. Дюк, А. Самойленко. Data Mining, Питер, 2001, 368 с.
12. R. Feldman, J. Sanger. The text mining handbook: Advanced Approaches to Analyzing Unstructured Data, England: Cambridge University Press, 2007, 410 p.
13. G. Salton, A. Wong, C.S. Yang. A vector model for automatic indexing // Communication of the ACM, November 1975. V18. N11. pp. 613-620.
14. F. Sebastiani. Machine Learning in Automated Text Categorization //ACM Computing Surveys, Vol. 34, No. 1, March 2002, pp. 1–47.
15. Р.М. Алгулиев, Р.М. Алыгулев. Новый метод резюмирования текстовых документов и оценка результата классификации в трех аспектах, // Телекоммуникации, 2006, №3, с. 7-17
16. Р.М. Алгулиев, М.Ш. Гаджирагимова. Некоторые аспекты организации и реализации мультиагентной системы поиска информации в распределенной информационной среде /Proceedings of the Second International Conference "Problems of Cybernetics and Informatics", vol. 2, October 24-26, 2006, Baku, Azerbaijan, pp. 31-34.
17. Р.М. Алыгулиев. Метод кластеризации коллекции документов и алгоритм для оценки оптимального числа классов // Искусственный интеллект, Донецк, 2006, №4, с. 651-659.
18. R.M. Aliguliyev. A new sentence similarity measure and sentence based extractive technique for automatic text summarization // Expert Systems with Applications, 2009, v.36, No.4, pp. 7764-7772.
19. В. Шаров. Управление жизненным циклом информации // Byte, 2004, №11, http://www.bytemag.ru
20. С. Орлов. Жизненный цикл ILM // LAN, 2007, №7, с. 30-40.
21. H. Jin, M. Xiong, S. Wu, Information value evaluation model for ILM, //Proceedings of the Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2008, pp. 543-548.