Work with Knowledge for Support of e-Government

Jan Paralic, Tomas Sabol

Department of Cybernetics and Artificial Intelligence, Technical University of Košice,
Letná 9, 042 00 Košice, Slovak Republic
Jan.Paralic@tuke.sk

Abstract. This paper describes some aspects of the e-Government, mainly from the knowledge management perspective. Most oft he work related with this talk is being done within the IST Project Webocracy[1], which is coordinated by Technical University of Kosice. The designed Webocrat system applies a knowledge-based approach [3]. Information of all kinds produced by various Webocrat modules is linked to a shared ontology modeling an application domain. Such ontology serves as a means for structuring and organizing available information resulting in improved search capability and contents presentation.

1 Introduction

There is a growing number of *e-Government portals* and solutions available today. But what the users lack in particular is a customized assistance – help that meets the individual situation and competence [11]. One way how to achieve progress in this direction is to focus on the way how knowledge is handled.

Knowledge management [10] generally deals with several activities relevant in knowledge life cycle [1]: identification, acquisition, development, dissemination (sharing), use and preservation of organization's knowledge. Our approach to knowledge management in the e-Government context supports most of the activities mentioned above. Based on this approach, a Web-based system Webocrat [7] has been designed and being implemented. Firstly, it provides tools for capturing and updating of tacit knowledge connected with particular explicit knowledge inside documents. This is possible due to ontology model, which is used for representation of organization's domain knowledge. Ontology with syntax and semantic rules provides the 'language' by which Webocrat-like systems can interact at the *knowledge level* [6].

Use of ontology enables to define concepts and relations representing knowledge about a particular document in domain specific terms. In order to express the contents of a document explicitly, it is necessary to create links between the document and relevant parts of a domain model, i.e. links to those elements of the domain model, which are relevant to the contents of the document.

[1] IST-1999-20364 project WEBOCRACY "Web Technologies Supporting Direct Participation in Democratic Processes", supported by the European Commission.

Model elements can be also used for search and retrieval of relevant documents. In case all documents are linked to the same domain model, it is possible to calculate a similarity between documents using the structure of this domain model. Such approach supports also 'soft' techniques, where a search engine can utilize the domain model to find concepts related to those specified by user. The search engine can thus return every document linked to the concepts, which are close enough to the concepts mentioned in the user's query. In order to provide better overview about the Webocrat system, its architecture from the functional point of view will be presented in the following section.

The other aspect of knowledge handling in Webocrat system is the use of knowledge discovery in text techniques to support some functions of the system, results of the analysis carried our are presented in section 3.

2 Webocrat from the functional point of view

From the point of view of functionality of the *Webocrat* system it is possible to break down the system into several parts and/or modules [7]. They can be represented in a layered, sandwich-like structure see Fig. 1.

A Knowledge Model module occupies the central part of this structure. This system component contains a conceptual model of a domain. The purpose of this component is to index all information stored in the system in order to describe the context of this information (in terms of domain specific concepts). The central position symbolizes that the knowledge model is the core (heart) of the system – all parts of the system use this module in order to deal with information stored in the system (both for organizing this information and accessing it).

Information stored within the system has the form of documents of different types. Since three main document types will be processed by the system, the document space can be divided into three subspaces – publishing space, discussion space, and opinion polling space. These areas contain published documents, users' contributions to discussions on different topics of interest, and records of users' opinions about different issues, respectively.

Since each document subspace expects different way of manipulating with documents, three system's modules are dedicated to them. Web Content Management module (WCM) offers means to manage the publishing space. It enables to prepare documents in order to be published (e.g. to link them to elements of a domain model), to publish them, and to access them after they are published. Discussion space is managed by Discussion Forum module (DF). The module enables users to contribute to discussions they are interested in and/or to read contributions submitted by other users. Electronic Submissions module (ES) enables users to access special part of the document space comprising their formal or informal submissions to local authority. Opinion Polling Room module (OPR) represents a tool for performing opinion polling on different topics. Users can express their opinions in the form of polling.

In order to navigate among information stored in the system in an easy and effective way, this layer is focused on retrieving relevant information from the system in various ways. Two modules represent it, each enabling easy access to the stored information in a different way. Citizens' Information Helpdesk module (CIH) is dedicated to search. It represents a search engine based on the indexing and linking (to knowledge model) of stored documents. Its purpose is to find all those documents, which match user's requirements expressed in the form of a query defined by means of a free text, or a query composed from concepts from domain model, or by document attributes (like author, date of issue etc.).

The other module performing information retrieval is the Reporter module (REP). This module is dedicated to providing information of two types. The first type represents information in an aggregated form. It enables to define and generate different reports concerning information stored in the system. The other type is focused on providing particular documents – but unlike the CIH module it is oriented on off-line mode of operation. It monitors content of the document space on behalf of the user and if information the user may be interested in appears in the system, it sends an alert to him/her.

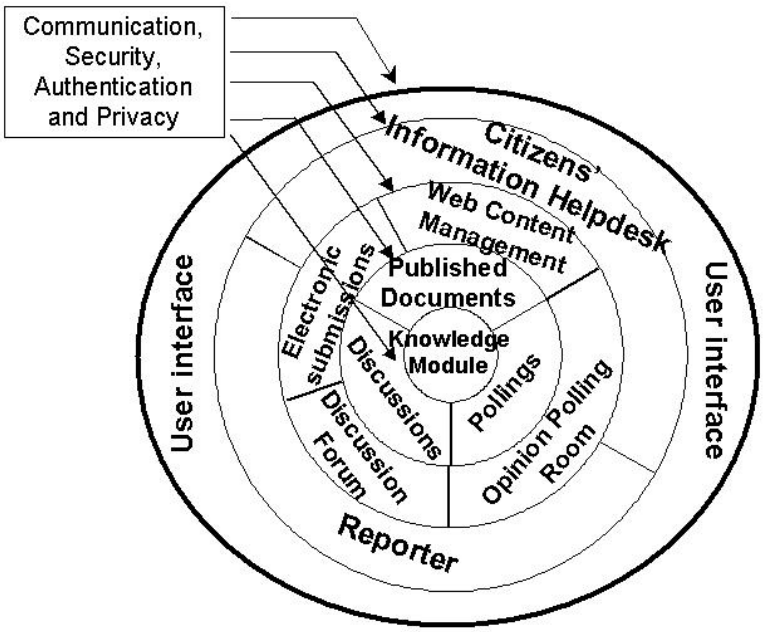

Fig. 1. *Webocrat* system structure from the system's functionality point of view.

The upper layer of the presented functional structure of the system is represented by a user interface. It integrates functionality of all the modules accessible to a particular user into one coherent portal and provides access to all functions of the system in a uniform way. In order for the system to be able to provide required func-

tionality in a real setting, several security issues must be solved. This is the aim of the Communication, Security, Authentication and Privacy module (CSAP) [2].

Technical achievements comprise also a system designed to provide automatic routing of messages from citizens to the appropriate person within the public administration (ES module); tools for easy access to public administration information and to competitive tendering (WCM module) and personalisation support (REP module).

3 Knowledge discovery in texts

Text data mining is much more complex task than data mining [9], because it involves text data that is inherently unstructured and fuzzy. In greater detail we can compare the KDT approach [5] and its particular steps against the KDD process steps [4].

1) *Understanding the application domain* and the goals of the KDT process: user must define which concepts are interesting.
2) *Acquiring or selecting a target data set*: texts must be gathered using information retrieval tools or in manual way.
3) *Data cleaning, pre-processing and transformation*: concepts must be described and texts need to be analyzed and stored in the internal representation form, usually after eliminating stop-words and possibly after stemming and exclusion of too frequent.
4) *Model development and hypothesis building*: identifying concepts in the collection.
5) *Choosing and execution of suitable data mining algorithms*: e.g. the application of the statistical techniques (text data mining task).
6) *Result interpretation and visualisation*: human must interpret the results.

Mining internal representation form of a document collection induces patterns and relationship across documents [5]. Some examples of unsupervised text mining tasks that we have analysed in the context of the Webocrat system are *Clustering/visualisation* of documents and *Association rules.* From supervised text mining task Predictive modelling (*classification models*) has been analysed.

3.1 Clustering/visualization

We think, that clustering/visualisation does not fit the functionality of the *WEBOCRAT* system as defined above, because documents in *WEBOCRAT* system are primarily organized by their links to knowledge model so that primarily knowledge model is used for document retrieval and topic-oriented browsing.

On the other hand, it could be useful to use techniques like GHSOM [8], because of its hierarchical structure that is tailored to the actual text data collection, *as a*

supporting tool within the initial phase, when the knowledge model of a local authority is being constructed.

This is true in such a case when local authority has a representative set of text documents in electronic form available for this purpose. It is assumed that these documents will be later on published using the *WEBOCRAT* system and linked to the knowledge model for intelligent retrieval purposes.

But users must be aware of the fact, that GHSOM does not produce any ontology. It is just a hierarchical structure, where documents are organized in such a way that documents about similar topics should be topologically close to each other, and documents with different topics should be topologically far away from each other. Particular node in this hierarchical structure is labelled by (stemmed) words – terms, which occur most often in cluster of documents presented by this node. This list of terms can provide some idea about concept(s), which can be (possibly) represented in the designed knowledge model.

Finally, particular documents represent leave nodes of this hierarchical structure. It is in our opinion necessary to look carefully through the whole structure, including particular documents in order to make reasonable conclusions about particular concepts proposed for the knowledge model and relations among them.

3.2 Association rules

Firstly, concepts need to be defined. To identify concept terms, natural language processing is commonly required. Instead of that, ontology provides this information in the *WEBOCRAT* system, and we can represent documents directly as a binary vector. This vector has one element for each concept, equal to 1 if document is linked to the particular concept.

Another advantage is, that within the *WEBOCRAT* system a hierarchy of concepts from local authority's area of work is available. We can make use of it for mining of associations, not only at the level of leaves concepts, but also at the higher levels of the hierarchy, getting so called multi-dimensional association rules [4].

Association rules can be exploited e.g. for automatic improvements of the knowledge model in the following way. When we use as input attributes for association rules mining algorithm only concepts to which documents are linked that means that we are looking for frequently occurring linking patterns. These patterns can be confronted with the actual ontology. When e.g. our algorithm finds association between concepts X, Y, and Z and in our ontology no relation between concepts X, Y and Z is presented, we can expect a missing relation between them. This approach is suitable mainly for documents that were not linked automatically, using a pre-defined template.

3.3 Classification models

In the *WEBOCRAT* system, ontology is used as a knowledge model of the domain, which is composed from concepts occurring in this domain and relationships be-

tween these concepts. Information is stored in the system in the form of text documents, which are annotated by set of concepts relevant to the document content.

One strategy for document retrieval is based on concepts. User selects interesting concepts and asks for information related to them, is used for information retrieval. The decision about document relevance to the user query is based on a similarity between set of query concepts and a set of concepts, which are annotated to the document. This task of document retrieval can be viewed as a classification task when the decision is made, whether the document is relevant for the user or not. With appropriate ontology which models domain well, use of this knowledge model can yield better results than e.g. retrieval based on vector representation of documents.

Retrieval accuracy depends on the quality of documents annotation. *Data mining methods can be very useful to guide user at annotating new document*. Annotation of the new document is the classification task *(text categorization task)* when we need to make decision which concept (concept represents category) is relevant to the content of the document.

The system must propose relevant concepts for new document in real time, so important requirement to used algorithm is execution time efficiency. User can add or delete some link between new document and concepts, and these changes can be immediately integrated into classifier. This requires ability of incremental learning. Relevance weighting of the concepts to the new document is better than simple binary decision. Concepts can be ordered by weight of the relevance to the new document and user can search for additional relevant concept according to this ordering.

4 Conclusions

In this paper, a system called Webocrat has been presented as an attempt to shift e-Government portals toward a customized assistance and knowledge enhanced services. The Webocrat system applies a knowledge-based approach. The functional overview of the system has been presented focusing on the role of knowledge management support.

Moreover, some text mining methods have been analyzed with respect to their potential use in Webocrat. From the three approaches described above the clustering approach based on GHSOM algorithm has been implemented and tested, the new version is being implemented now. For association rules discovery an external system will be used. Finally, classification task is implemented by means of the Bayes approach with important text pre-processing phase.

Webocrat approach is user-centered, focusing on pilot applications that are divided into two trials. At the time of writing, the first trials (May – July 2002), where part of the Webocrat functionality (DF, WCM, ES and OPR modules) have been available for testing, have been already evaluated. But the full Webocrat potential will be elaborated in the second trial, which is scheduled for March – May 2003.

Acknowledgements

This work is done within the Webocracy project, which is supported by European Commission DG INFSO under the IST program, contract No. IST-1999-20364 and within the VEGA project 1/8131/01 "Knowledge Technologies for Information Acquisition and Retrieval" of Scientific Grant Agency of Ministry of Education of the Slovak Republic.

The content of this publication is the sole responsibility of the authors, and in no way represents the view of the European Commission or its services.

References

[1] Abecker A., Bernardi A. Hinkelmann K. Kühn, O. & Sintek M. (1998): Toward a Technology for Organizational Memories, IEEE Intelligent Systems, 13, May/June, p.40-48.

[2] Dridi, F. and Pernul, G. and Unger, V. Security for the electronic government. In Proceedings of the European Conference on E-Government, Trinity College, Dublin, Ireland, 2001, pp. 99-110.

[3] Dzbor M., Paralic J. and Paralic, M. (2000) Knowledge management in a distributed organization. In: Proc. of the 4th IEEE/IFIP International Conference BASYS'2000, Kluwer Academic Publishers, London, pp. 339-348.

[4] Han, J., Kamber, M. (2000) *Data Mining – Concepts and Techniques*. Morgan Kaufmann Publishers

[5] Loh, S., Wives, L. K., and Palazzo, J. (2000) Concept-Based Knowledge Discovery in Texts Extracted from the Web. SIGKDD Explorations, Vol. 2, Issue 1, pp. 29-39

[6] Newell A. (1982) The Knowledge Level. Artificial Intelligence, 18, pp. 87-127.

[7] Paralic, J., Sabol, T. & Mach, M. (2002) A System to support E-Democracy. Proc. of the First International Conference EGOV 2002, Aix-en-Provence, France, LNCS 2456, LNCS 2456, Electronic Government, R. Traunmuller, K. Lenk (Eds.), Springer Verlag

[8] Rauber, A., Dittenbach, M. and Merkl, D. (2000) Automatically Detecting and Organizing Documents into Topic Hierarchies: A Neural Network Based Approach to Bookshelf Creation and Arrangement. *Proc. of the 4th European Conference on Research and Advanced Technologies for Digital Libraries (ECDL2000)*, Springer LNCS 1923, Lisboa, Portugal

[9] Tan, A.H. (1999) Text Mining: The state of the art and the challenges. *Proc. of the PAKDD'99 workshop on Knowledge Disocovery from Advanced Databases*, Beijing, pp. 65-70

[10] Tiwana A. (2000) The Knowledge Management Toolkit. Prentice Hall.

[11] Traunmüller, R. & Wimmer, M. (2001) Directions in E-Government: Processes, Portals, Knowledge. Proc. of the Int. Workshop "On the Way to Electronic Government" in Conjunction with DEXA (Munich, Germany), IEEE Computer Society Press, Los Alamitos, CA, pp. 313-317