[Karvounarakis et al., 2003] G. Karvounarakis, A. Magkanaraki, S. Alexaki, V. Christophides, D. Plexousakis, M. Scholl, K. Tolle. Querying the Semantic Web with RQL. // In Computer Networks and ISDN Systems Journal, Vol. 42(5), August 2003, pp. 617-640.

[Miller, 1956] G.A. Miller. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. // In The Psychological Review, 1956, vol. 63, pp. 81-97

[Motik et al., 2002] Boris Motik, Alexander Maedche, Raphael Volz. A Conceptual Modeling Approach for Semantics-Driven Enterprise Applications. // In Proceedings of the First International Conference on Ontologies, Databases and Application of Semantics (ODBASE-2002). Springer, 2002.

[Noy et al., 2001] N.F. Noy, M. Sintek, S. Decker, M. Crubezy, R.W. Fergerson, M.A. Musen. Creating Semantic Web Contents with Protege-2000. // IEEE Intelligent Systems 16(2), pp. 60-71, 2001

[Wertheimer, 1944] M. Wertheimer. Gestalt theory. // In Social Research, 11, 78-99.

Authors' Information

Vladimir Gorovoy – PhD student, Saint-Petersburg State Polytechnic University, Intelligent Computer Technologies Dpt. 195251, Politechnicheskaya 29/9, St. Petersburg, Russia; e-mail: vgorovoy@mail.ru

Tatiana Gavrilova – Professor, Saint-Petersburg State Polytechnic University, Intelligent Computer Technologies Dpt. 195251, Politechnicheskaya 29/9, St. Petersburg, Russia; e-mail: gavr_csa@rambler.ru

INTELLIGENT SEARCH AND AUTOMATIC DOCUMENT CLASSIFICATION AND CATALOGING BASED ON ONTOLOGY APPROACH

Vyacheslav Lanin, Lyudmila Lyadova

Abstract: This paper presents an approach to development of intelligent search system and automatic document classification and cataloging tools for CASE-system based on metadata. The described method uses advantages of ontology approach and traditional approach based on keywords. The method has powerful intelligent means and it can be integrated with existing document search systems.

Keywords: electronic document, automatic document classification and cataloging, ontology approach, information system development.

ACM Classification Keywords: I.2.7 Artificial Intelligence: Natural Language Processing – Text analysis; D.2.2 Software Engineering: Design Tools and Techniques – Computer-aided software engineering (CASE).

Introduction

Development tools used for implementing large distributed information systems, which consist of separated subsystems and should be installed in territorially remote organizations, should meet the requirements, providing possibility of its customization on various maintenance conditions and user's requirement during installation and dynamically during maintenance. Organizations has various technical possibilities, organizational and business forms, it makes information system development difficult. Implementation of these requirements provides efficiency of expenses for system creation, a high degree of its adaptability and scalability, robustness of the system.

The CASE-system METAS bases on interpretation of the multilevel metadata. The metadata describes an information system from different points of view and with a various grain size. Opportunities of dynamic system customization are provided by re-structuring of a database, generation and customization of user interface, generation of queries and creation of reports [Lyadova, 2003].

The data domain analysis is the most labour-intensive and important stage in process of information system development by means of CASE-system. Any changes in business operations of organization, for which information system is created, demand the iterated analysis and modification in information system model. Often changes of system maintenance conditions or changes of user's requirements are connected with some normative documents. It can be the normative documents determining business processes of the data domain or the documents of the particular organization. Thus, the analysis of data domain in many respects bases on the analysis of documents, which constitute difficult system. Modifications in model should be grounded on the changes fixed in normative documents.

Complexity of analyst work can be lowered by automation of document analysis process. To solve this task it is necessary to have tools intended for search and keeping document set. Documents can be received from different sources, connected with considerated data domain. Except this for automation of analysis process tools of classification, cataloguing and data mining should be included to the system.

In this paper the problems connected with information processing in an inhomogeneous program and organizational environment, particularly with documents search and their electronic cataloguing, are considered. As an example of such documents we can mention various internal organization documents (orders, contracts, acts and so forth), normative - legal documents, etc. Documents come to a system in a random order from different sources. It is usually semistructured. It reasons the complexity of document processing. The extremely important tasks for implementation in this area are automation of processes of data exchange with various legal informational systems, and the possibility of import texts and documents from files and databases of various formats and documents management systems.

The main problems which prevent fast and high-quality document processing in electronic documents management systems are insufficient structuredness of the information, information redundancy, and presence of great deal of undesirable for user information. The human factor has a significant impact on the efficiency of document search. An average user is not aware of the advance option of a query language and uses simply typical queries.

Development of a specialized software toolkit intended for information systems and electronic document management systems can be effective solution of tasks listed above. Such toolkit should be based on the means and methods of AI.

Problem of document search

Situation when a user searches something in a book or printed document is considered below. The most obvious way is to read the whole book (document), but such process takes a long time. However if a user knows something connected with the data domain, he can use book (document) content and choose appropriating part containing necessary information. Also he can look throw subject index to define page numbers where searched terms are mentioned.

In the example content and index are tools which make search process easier. In case of information system including document management tools documents play a part of searched information and services called subject directory are used as content and subject index. Here under document management tools we mean advanced tool, which functions are not only creating and keeping documents, it also allows to search, import, analyze, categorize and catalogue documents.

For example user operates with a computer system and he needs to gather information about Perm city. Some search stages can be marked out. Necessary to find something appears, in other words information need appears. Then user has to formalize his information need somehow. In traditional systems formalization comes to choosing concepts and key words and their relations. Chosen set of key words with fixed relations between it is called query. Next user enters the query by means of search system interface. The system extracts documents which match user's query from document set called information search field according to customizable search criteria and then forms a result. The documents being found divide into two groups according to its content (Fig. 1): documents matching user's needs and documents, which do not match user's needs but match user's query (information noise). In the example documents where Perm is not a city name are information noise.

Measure of correspondence between system response and user information needs is called semantic relevance, and measure of correspondence between system response and user query is called formal relevance. Usually presence of query keywords in the text of a document is a criterion of document formal relevance. If we use search based on keywords usually some of the documents matching user's information needs do not get into result set. For example if keyword is "Perm" documents where it is used in phrase "Perm region capital " instead of "Perm city" may not be found.

The main problem of information search is the result of the fact, that the majority of information search systems are based on using keywords and "word" doesn't have meaning and semantics.

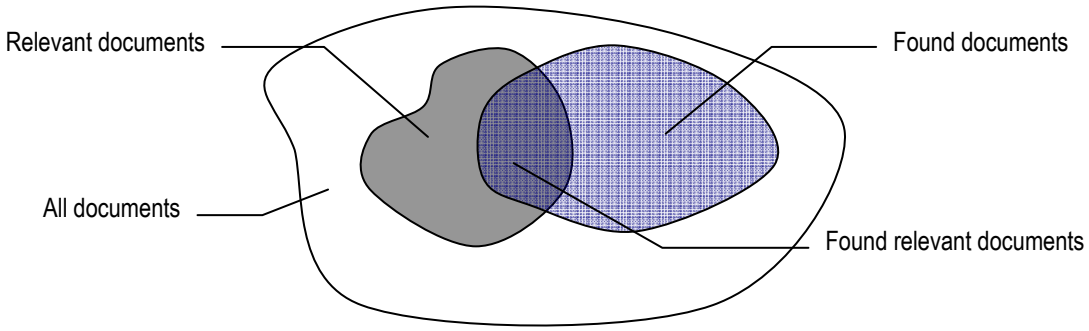

Fig. 1. Document space

The majority of technologies for document processing are oriented on handy work with information. Often principles of processing electronic information just copy principles of processing printed information. There are various means for text formatting, that help to represent information in a convenient form, in a text editor, but there are not means for semantics reflection in it. In most cases computer is used as typewriter or calculator, which goal is automatic examination of answer alternatives. For effective search it is necessary to extend traditional document concept: document should be related with knowledge, which make possible document content interpretation and processing.

Usually artificial intelligence methods are used for solving tasks, which are easy and obvious for people, but it is difficult to formalize them and implement their algorithms. One of these tasks is working with documents in information systems. It includes information search, cataloging, analysis and data mining.

There are different methods, models and languages oriented on integrated data and knowledge declaration. The most perspective and universal approach is using ontologies.

Ontology definition

The concept ontology is one of the most used concepts. The term ontology is used in different contexts, and different meanings are ascribed to it. According to implementing tasks we let that ontology is an exact specification of a data domain, which includes terms dictionary and a set of relations between them (like "instance of", "whole-part"). Relations between terms shows how this concepts are correlated with each other in particular data domain. In fact such ontology definition means that ontology is hierarchical entity base of current data domain for which information system is developed.

It is difficult to find appropriate ontology, this process takes much time. So sometimes it is impossible to find ontology among developed ones, that is why a new ontology creating may be defensible. Then ontology takes into account particular task specific. Except this using developed ontologies has some disadvantages more. In particular knowledge of different people can be represented by different ontologies. At the same time we can not state that one ontology is better than another. Some different ontologies representing various aspects of data domain and solving tasks can be developed for he same organization, in which information system is installed.

A lot of languages and systems for declarating ontologies and operating with it exist. The most perspective is visual method, which allows experts to draw ontologies evidently. This helps to formulate and explain appearance nature and structure. Visual (graph) models have especial cognitive force.

Search of documents based on ontologies

According to the approach [Chuprina, 2004] information search is carried out using an ontology either representing data domain of information system or specially developed by user. Generally document content interpretation is extremely difficult task, but document and ontology matching mechanism is necessary only for intelligent document search.

Document search process based on ontology is described below.

The process is started with search of basic ontology concepts in the document. If all concepts have been found in the document we make decision that the ontology describes the document. In case a concept has not been found the system begins to search its synonyms.

If synonyms of searched concept have not been found the system tries to gather concept by parts according to relation called “part of”. If we do not get a result even after this operation the system can use relation “class-subclass”. It allows to take into account stricter or more general concept.

So, recursive ontology search mechanism has been posed. In contrast of traditional search system method mentioned above has more powerful semantics. It allows to find concepts implicitly contained in the document.

The main advantages of ontology search are

- systematic *viewpoint (ontology represents entire data domain);*
- *uniformity (knowledge is presented in standard form);*
- *completeness (*ontology allows to reconstruct not mentioned relations).

Documents found in outer sources can be imported to information system for classification and cataloging, analysis and extracting useful information.

Ontological document classification and cataloging

To organize document cataloging process user have to correlate each document category with an ontology.

When a new document gets to the system it is sequentially compared with ontology of each category. If comparison is successful the document falls within this category. Each document may match several ontologies, so it can be attributed to several categories.

ORDER # 1

Use since 01.12.2005 following list of members of staff:

Assistant	4 000
Associate professor	10 000
Head of	15 000

Head of a university J. Green

01.11.2005

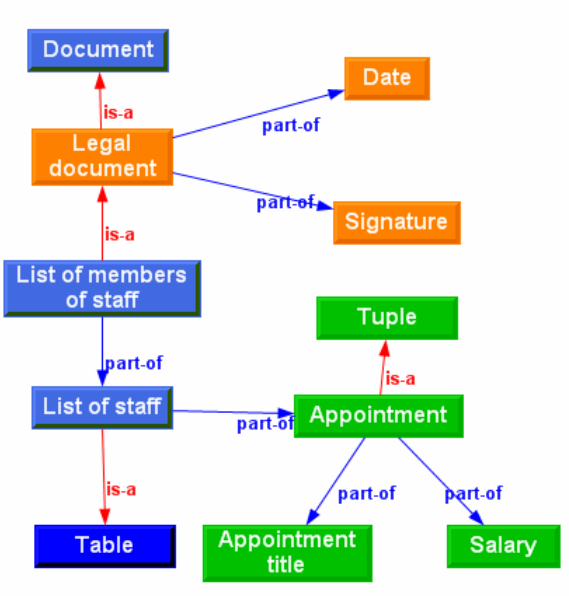

Fig.2. Documents and its ontology

It is convenient to represent a category system as a tree (Fig. 2). Therefore corresponding ontologies constitute an hierarchy. Within such approach to ontology representation child nodes qualify parent nodes ontologies.

For example top level nodes can match small ontologies which represent administrative documents, contracts etc., and nodes of other levels can match ontologies which specify these document types.

Possibility to operate with various document formats is a feature of the approach.

It is necessary to interpret terms describing format and structure of particular document to match ontologies and semistructured documents). Such terms are “table”, ”tuple”, ”date”, ”number” and so on. So special components which provide unificated access to documents with different format should be included to system architecture. Such functionality can be realized by using “installable format driver”. It is components implementing predetermined interface, which allows access to document of a specific type. Implementation of such driver can be based on using patterns, samples, which make possible document structure recognition.

Conclusion

The features of the developed approach are universality of its using, capability to integrate with existing document search systems, powerful intelligent capabilities. Mechanism of ontological clusterization together with functionality of particular information system makes possible effective document management both for documents which are generated in the system and documents which are imported from outer heterogeneous sources.

CASE-technology METAS developed by “Computing institute” makes possible effective using of described tools both during information system development process and when users operate with information system. Also these tools can be used in process of document interpretation to adapt a system dynamically.

The technology gives to user not only customization tools. It also provides components which are used to navigate through information objects, representing entities of data domain, and its relations. Object explorer displays object tree, which can be customized by users due to their information necessaries. Object tree is used not only for object preview, appropriate business operations can be run from the node too. Each document created in the system is represented in the database as an entity. So it is displayed in a tree node, which corresponds to the entity, and user can see it. A document can be reached in tree-walk by different ways according to user’s tasks.

User can include to the tree nodes intended for document clusterization. Such node can be associated with an ontology. So the same mechanism is used for working with documents imported from outer sources and clusterizated in the system.

Adding to CASE-system tools for analysis of semistructured documents, which are categorized on the base of ontologies created by developers and users, essentially reduce labor intensiveness of maintenance and customization.

Bibliography

[Лядова, 2003] Л.Н. Лядова, С.А. Рыжков. CASE-технология METAS. В кн.: Математика программных систем. Пермский государственный университет, Пермь, 2003. С. 4-18.

[Chuprina, 2004] S. Chuprina, V. Lanin, D. Borisova, S. Khaeva. Internet Intelligent Search System SmartFinder. In: Proc. of the European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology. Knowledge-Based Media Analysis for Self-Adaptive and Agile Multimedia Technology / The Royal Statistical Society, November 25-26, 2004, London, U.K. P. 151-156.

Authors' Information

Vyacheslav Lanin – Perm state university, student of computer since department; Russia, 614990, Perm city, 15, Bukireva st.; e-mail: lanin@perm.ru

Ludmila Lyadova – Institute of Computing, Deputy Director; 19/2-38, Podlesnaya St., Perm, Russia; e-mail: e-mail: lnlyadova@mail.ru