SYSTEM OF INTELLIGENT SEARCH, CLASSIFICATION AND DOCUMENT SUMMARISATION FOR INTERNET PORTAL

Vyacheslav Lanin, Dmitriy Tsydvintsev

Abstract: *The article presents a description of alleged approaches to the implementation of data processing subsystem on Internet portal. Main problems are connected with exponential growth in number of documents, lack of semantic indexing and unstructured nature of information. In proposed approach, user receives an effective intelligent means of finding electronic documents on the basis of semantic indexing, automatic classification and cataloging of documents with construction of semantic links between them and automatic summarization of documents with the use of knowledge. The proposal is to increase the effectiveness of working with electronic documents with the help of intelligent analysis, for which agent-based and ontological approaches are used. In accordance with the proposed approach, ontology is used to describe data semantics of the document and its structure. Ontology is a central concept in process of document analysis. Through the use of ontologies the required data can be obtained; we know where to find information and how it can be interpreted. Ontology Repository contains three levels of ontologies. At the first level there are ontologies describing objects which are used in a particular system and which take into account system features. At the second level there are objects described in terms of the first level that are invariant to the domain. Objects of the third level describe the most general concepts and axioms, by which the lower levels objects are described. The third and second levels can be divided into two parts: a description of structures and description of the documents themselves.*

Keywords: *ontology, agent, multi-agent systems, intelligent search, semantic indexing, document analysis, adaptive information systems, CASE-technology.*

ACM Classification Keywords: *H.2. Database Management: H.2.3. Languages – Report writers; H.3.3. Information Search and Retrieval – Query formulation.*

Introduction

An exponential growth in the number of electronic documents is currently underway, and it clearly shows that traditional mechanisms for processing of electronic documents cannot cope with needs of user. This trend is evident both on Internet and in corporate networks. Currently, so-called information portals (thematic and corporate) become more and more popular, and their main objective is a consolidation of information and knowledge.

One of these solutions is a research portal – information-analytical system for the collection and analysis of data about innovation activity of regions to support effective management decisions (Research portal "Innovative development of regions"). Data for analysis is extracted from heterogeneous unstructured or semistructured data sources, in particular, Internet resources, as well as operational databases. According to the plan, the system must provide integration, coordination, aggregation and maintenance of previously disconnected data. Also it should support the various forms of data visualization and analysis, customized to the needs of users. From this it follows that the search and processing of unstructured text data from different sources in different formats, is becoming one of the main functions of the system under development.

Thus, the relevance of the problem is caused by the following reasons:

- Exponential growth in the number of documents that make it impossible to process data by traditionnal methods without loss of quality;
- Lack of semantic indexing, which does not allow for intelligent document processing in full;
- Unstructured nature of the information; the traditional mechanisms of its processing and analysis can't be used.

Consider these problems in more detail.

The exponential growth of information contained in the Internet is the reason for continuing increase in difficulty of finding relevant documents (Fig.1) and organizing them into a structured within the meaning of storage [6]. It becomes more and more difficult for user to find the necessary information; traditional search engines become ineffective.

Most technologies of working with documents focus on the organization of effective work with information for the person. But often ways to work with electronic information simply copy methods of working with paper-based information. In a text editor, there are wide range of different types of text formatting (presentation in human readable form), but little or no ability to transfer the semantic content of the text, i.e. *no semantic indexing*. To effectively address the search problem we need to expand our notion of a traditional document: *the document should be linked with knowledge to interpret and process the data stored in the document*.

Unstructured information constitutes a significant part of modern electronic documents (Fig. 2). Data Mining systems work with structured data. Unstructured content requires using Text Mining systems. In fact, they solve the same problem for different types of data, so it is assumed that these systems will converge in a "single point".

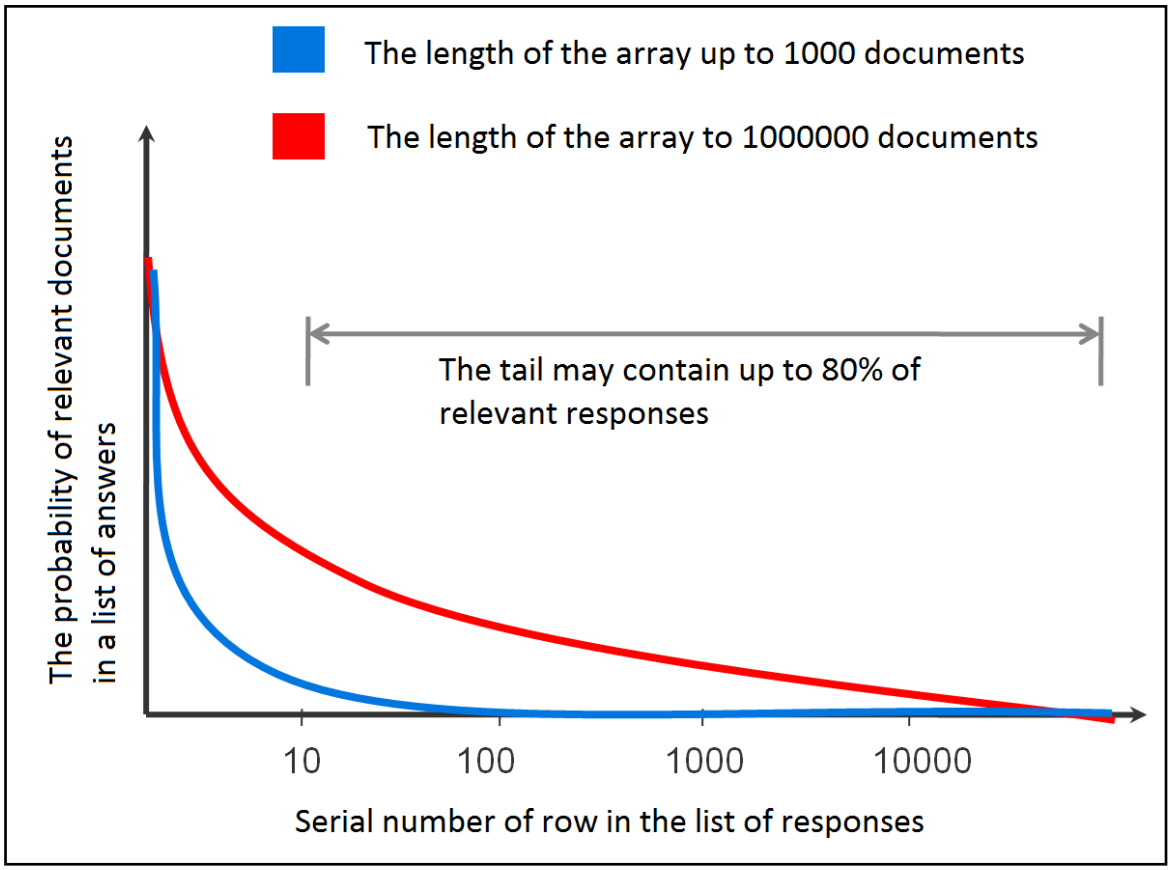

Fig.1. A problem of information retrieval with an increase in the number of documents

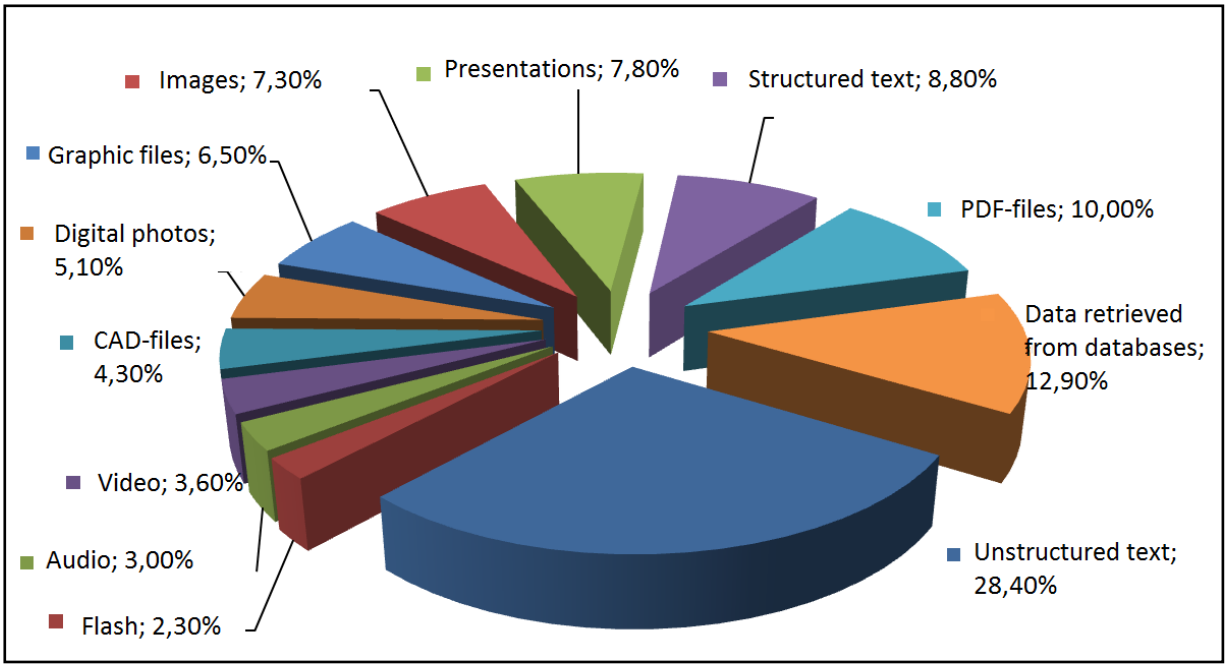

Fig. 2. Distribution of documents' categories

Text Mining allows to identify previously unknown relationships and correlations in the existing text data [5]. An important task of Text Mining technology is to derive from the text its characteristic elements or properties that can be used as document metadata, keywords or annotations. Another important task consists in assigning the document to certain categories of a given scheme of systematization. Text Mining also provides a new level of semantic search of documents. The possibilities of modern Text Mining systems can be applied in knowledge management to identify patterns in the text, to automatically "push" or post information on profiles of user interest, to create surveys of documents.

Tools and approaches of Text Mining will help to implement the intellectual capabilities of the portal when working with electronic documents.

The Approach to Semantic Indexing

Human factor has a great influence on the effectiveness of search process. User often is not ready for a long waiting of search results, for viewing and analyzing large amounts of resulting sample. In addition, most users ineffectively use search software, and usually, they ignore the advanced search capabilities, and make only short types of queries. Improvement of the electronic documents processing requires the availability of metadata describing the structure and semantics of documents. One of possible approaches to the description of information embodied in the document is an approach based on ontologies. Ontology is a knowledge base of a special type, which can be "read" and understood, alienated from the developer and / or physically separated by its users [4].

As an approach to semantic indexing has been chosen ontological approach [1], in which the ontology can describe both structure and content of the document, i.e. ontology is used to describe the semantics of the document data and its structure. Given the nature of solved targets in this paper we will concretize the notion of ontology. We assume that *ontology is a specification of certain domain*, which includes a glossary of terms (concepts) domain and a set of connections between them which describe how these terms relate to each other in a particular subject area. In fact, in this context *ontology is a hierarchical conceptual framework of the subject area*.

Ontology of document is used to analyze the document; due to it the required information can be obtained from the document: we know where to search for data and how it can be interpreted. If you represent documents using ontologies, the problem of matching ontologies and existing document is reduced to problem of search ontology terms in the document. As a result, the system needs to answer the question: Does this ontology describe the document or not. The latter question can be answered in the affirmative, if in the comparison process all the concepts included in ontology are found in document. Thus, the initial problem reduces to problem of finding general concepts in the text on the basis of formal descriptions.

Ontology repository contains *three levels of ontologies*. At the first level there are ontologies describing the objects used in a specific system and taking into account its peculiarities. The second level describes the objects that are invariant to the subject area. Objects of the second level are described in terms of objects of the first level. This is reflected in the relations of inheritance and metonymy. The objects of the third level describe the most general concepts and axioms, by which objects at the lower levels are described. The third and second levels can be divided into two parts: the description of structures and description of the documents themselves, with the documents described in terms of structures.

To address the problem of allocation of general concepts an agent-based approach is proposed on the basis of formal description [2]. This approach will satisfy the requirements of the search process, if all the advantages of multi-agent systems are realized in process of construction of the system.

When using this approach, for each node of the ontology, which contains a general concept, an agent is created which looks for this particular concept. In this approach, the agent is considered as a system aimed at achieving a particular purpose, capable of interaction with the environment and other agents. To be intelligent, the agent should have a knowledge base. Thus, to identify active agents in the system, you must choose a way to describe the knowledge base, the nature of interaction with the environment and cooperation.

Knowledge base of agent for finding find common concepts of the ontology can be also conveniently presented in the form of ontology. To enable the user to add new templates it is necessary to select the basic concepts for the formation of general ones.

One of the most important properties of agents is *sociality*, or the ability to interact [2]. As mentioned above, the agent is created for each node of the ontology, which contains a general concept. According to the accepted classification of agents it is *intentional agent*.

This agent is designed to address two problems:

1. It breaks the entire list of available templates concepts into separate components and runs simple search agents for searching of derived components.
2. Assembles results from all the lists submitted by agents of the lower level.

The agents at a lower level mentioned above are called *reflex agents*. They get a template, and their goal becomes finding phrases in the text covered by this template. Search results for agents of all levels shall be recorded on the "bulletin board".

At this point in other systems the instruments of ontological nature are used in the following areas:

- WordNet in conjunction with the vector and Boolean models of information retrieval;
- Traditional information retrieval thesauri in combination with various statistical models;
- Thesaurus for automatic indexing in Boolean models of documents searching, in problem of automatic headings and automatic annotation.

Ontologies will form a core of portal metadata when working with electronic documents. Clearly defined subject area allows creating sufficiently detailed ontologies, which can be used by all its subsystems.

Automatic Abstracting

Currently, there are two approaches used for automatic summarization. A traditional approach (quasi abstracting), which is used by such systems as Microsoft Office, IBM Intelligent Text Miner, Oracle Context, is based on allocation and selection of text fragments from the source document and connection them in a short text. On the other side, there is approach based on knowledge, which involves preparation of summaries and transfers the basic idea of the text, perhaps even in other words.

Quasi abstracting is based on the allocation of specific fragments (usually sentences). For this purpose, a method of comparison of phrasal templates chooses blocks with the greatest lexical and statistical relevance. There are a model of linear weights is used in most implementations of the method. The analytical phase of this model is a procedure of appointing the weighting coefficients for each block of text in accordance with such characteristics as location of this block in the original, frequency of appearance in the text, frequency of use in key proposals, as well as indicators of statistical significance. So, there are three main directions, often used in combination: statistical methods, positional methods and indicator methods.

The main advantage of this model lies in the simplicity of its implementation. However, the selection of sentences or paragraphs, not taking into account the relationship between them, leads to the formation of disconnected essays. Some proposals may be omitted, or there can be "hanging" words or phrases in them.

To implement the second method, some ontological reference is needed, reflecting the views of common sense and the concepts of targeted subject area, to make decisions during the analysis and to determine the most important information.

Method of forming a summary suggests two basic approaches.

The first approach relies on the traditional linguistic method of parsing sentences. This method is also uses semantic information to annotate parse trees. Comparing procedures directly manipulate the trees to remove and rearrange the parts, for example, by reducing the branches on the basis of certain structural criteria, such as brackets or embedded conditional or subordinate sentences. After this procedure, the parse tree is greatly simplified, becoming, in essence, a structural "squeeze" of the original text.

The second approach for compiling a summary roots in artificial intelligence systems and relies on natural language understanding. Parsing is also part of such a method of analysis, but the parse tree in this case is not generated. On the contrary, there are conceptual representative structures of all the initial information are formed, which accumulate in the text knowledge base. As the structures, formulas of predicate logic or such representations as a semantic network or frame set can be used.

Automatic summarization is necessary for the developed portal. When user is searching, it is necessary to present him a document annotation, by which he can decide on the usefulness of this document.

Classification and Cataloging of Documents

The task of automatic classification and cataloging of documents is the task of partitioning the incoming stream of text into thematic substreams according to predetermined headings. Automatic cataloging of electronic documents, and documents posted on Internet in particular, is complicated because of the following reasons [8]:

- A large array of documents;
- An absence of special structures for tracking the emergence of new documents;
- Optionality of the author's classification of electronic documents (as opposed to print publications) through annotation, attribution of the qualifier codes, etc.;
- A problem of tracking changes in documents.

As for automatic abstracting, there are two opposite approaches to cataloging. *The methods based on knowledge* are the most effective, but it's difficult to implement them. When cataloging the texts on the basis of knowledge preformed knowledge bases are used. They describe language expressions, corresponding to a particular category, and rules for the selection of headings [5]. Another class of methods for automatic categorization of texts is *the methods of machine learning*, which can use manually pre-cataloged texts as training examples.

When implementing a system of automatic cataloging of the portal, it is necessary to solve two problems:

- *Establishment of a mechanism for introduction and description of categories*, as some expression on the basis of words and terms in documents. The problem can be solved on the basis of expert descriptions of categories or on the basis of machine learning methods with the help of pre-cataloged collections of documents.
- *Analysis of linguistic material and context of words' using*. It requires an extensive knowledge of the language and subject area.

Conclusion

The above approaches are used in the development of an electronic document management subsystem of research portal. Its distinctive feature is focus on the explicit knowledge representation by using ontologies. This approach will allow us to realize intelligent services for searching and processing of electronic documents related to the portal and gathered from different sources.

As a result, the following tasks will be solved by creating the research portal:

- Semantic indexing of documents and intelligent retrieval of data corresponding to users' queries and the specific subject area;
- Extracting information from unstructured documents;

– Intellectual classification and cataloging and automatic summarization of retrieved documents;
– Maintaining a history of electronic documents.

The implementation of the subsystem will significantly reduce complexity to find useful information, its analysis and possible use in research.

References

[1] Lanin V. Intelligent management of documents as the basis for the technology of adaptive information systems // Proceedings of the International Scientific-Technical Conference «Intelligent systems» (AIS'07). V. 2 / M.: Fizmatlit, 2007. P. 334-339.

[2] Tarasov V. From multi-agent systems to intelligent organizations: philosophy, psychology, computer science. M.: Editorial, URSS, 2002.

[4] Khoroshevskii V., Gavrilova T. Knowledge Base intelligent systems. Petersburg.: Peter, 2001.

[5] Lande D. Knowledge Search on the Internet. Professional work. M.: Publishing house "Williams", 2005.

[6] Efremov V. Search 2.0: fire on the "tail" // Open systems. DBMS № 08 (134), 2007.

[7] Chernyak L. Enterprise Search 2.0 // Database. - 2007. - № 07 (133).

[8] Fedotov A, Barakhnin V. Internet Resources as an object of scientific research [electronic resource. - 2007. - Mode of access: http://www.rfbr.ru/pics/28320ref/file.pdf.

[9] Weal M.J., Kim S., Lewis P.H., Millard D.E., Sinclair P.A.S., De Roure D.C., Nigel R. Ontologies as facilitators for repurposing web documents / Shadbolt. Southampton, 2007.

Authors Information

Vyacheslav Lanin *– Perm State University, Department of Computer Science; Russia, Perm, 614990, Bukirev St.,15; e-mail: lanin@psu.ru.*

Dmitriy Tsydvintsev *– Perm State University, Department of Computer Science; Russia, Perm, 614990, Bukirev St.,15; e-mail: akinokinos@yandex.ru.*