International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 14 (2015) pp 34537-34539

Knowledge engineering method implementationfor digital archives

Anton Ivaschenko
(Samara State Aerospace University, 443086 Russian Federation, Samara, Moskovskoeshosse, 34)
Pavel Sitnikov
(SEC Open Code, 443001 Russian Federation, Samara, Yarmarochnaya, 55)
DmitriyMartyshkin
(Samara State Aerospace University, 443086 Russian Federation, Samara, Moskovskoe shosse, 34)
Sergey Fedotov
(Samara State Aerospace University, 443086 Russian Federation, Samara, Moskovskoe shosse, 34)
PavelVasin
(SEC Open Code, 443001 Russian Federation, Samara, Yarmarochnaya, 55)

Abstract- The paper introduces the ways to implement knowledge engineering method with regard to the creation of an intellectual digital archive system.Using the artificial intelligence system principles based on the use of knowledge, ontologies, allows to decrease the dependence on the qualification of expert regarding knowledge. The search based on the semantics of the documents allows to increase the relevance of the results.

Keywords- BigData, knowledgeengineering, commoninformationspace, humanresourcesmanagement,researchanddevelopmententerprise, digital archives, ontologies, semantics.

Introduction

Nowadays nearly all state and commercial organizations face problems regarding the storage of large amounts of paper documents. The way paper documents are stored is not up-to-date: paper media is short-lived, is exposed toageing, andcanbedistorted, lostordestroyed. Allthiscanleadtoirretrievablelossofinformation.

The problem can be solved by developing an information system for working with digital content that permits updating and management of digital archive fund, support of development, intelligent management and use of digital archives, including digitization of content in different formats, ability of intelligent semantic search against a digital archive of all forms of data (including audio, video, graphic and text information), search by metadata, full-text search and other features. The content, when digitized properly, can become a powerful tool for decision-making by executives of all levels.

1 Methods And Technologies Overview

Theessentialfeaturesthatdistinguishknowledgebasesfromtheconventionalonesaretheuseofmultipledata; openmodel; theconstructionandtheuseofsemanticweb (ontology).

The featuresof the intellectual digital archive system are interconnected documents and their optional attributes.

To ensure documents network connectivity, the network model fits best (semantic networks). This model implies description of the graph, where the nodes are objects and edges are relationships (links) between the objects. On the other hand, frame model fits the second requirement best. Attributes of the documents are placed to corresponding slots.

According to the CAP theorem (Brewer's theorem, see Fig. 1), heuristic assertion stating that in any implementation of distributed computing it is possible to provide maximum two out of three of the following properties [1, 2]:consistency, availability, and partitiontolerance.

Amongdifferentclassesofdatabases,thetwocategories canbehighlighted whichmeetthe earlier mentioned requirements: graph databases and NoSQL. Thefirstoneis perfectlysuitableforstoringthesemanticwebs**;**thesecondonepro videsfavorableopportunitiesforscaling, availability and consistency [3].

Among graph databases, popular “neo4j” can be mentioned, however this database is separation-intolerant, and also has some problems with sharding, which are crucial for scaling [4]. This is partly related to working with a complex structure like a graph, as well as to the need to support the previously stated functionality of ACID-transactions. As it is necessary to potentially work with Big Data, a solution was made upon choosing a database which is created specifically to work with Big Data, while working with the model was decided to implement independently,taking into consideration the peculiarities of each specific task.[5]

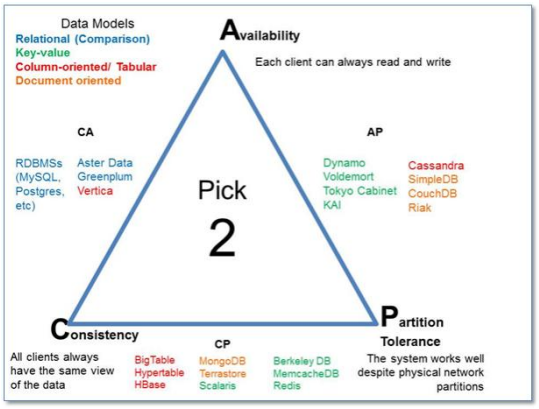

Fig. 1. The CAP Theorem

ThetechnologyofBigData impliesaseriesofapproaches, instrumentsandmethodsofprocessingstructuredandnon-

structureddataofhugevolumeand of considerable variety to get the results able to be perceived by the human, which are effective in conditions of continuing growth.Applyingthistechnologywillallowtheuserstoquicklyan dby appropriate timeprocesstheinformationand packages of documents including, the opportunity of package document tagging (taking into consideration the semantics)or defining the key words in the document.

2 Intelligent Archive System Challenges

High-technologyautomaticintelligentsystemofdigitalarchivesand program platform for building the systems of decision-making support based on the data of digital archives with the use of knowledge bases technologies will make for achieving the followingchallenges:

- to provide the storage of the archive information electronically;
- toprovideindependence of placesfor storing the documents from work places;
- tomakethe provision of information continualregardlessofthe location ofwork placeand the schedule of archive work;
- to make the processing of inquiries and the provision of archive data quicker;
- theintegrationintoa unified information infrastructure;
- to make the provisionof the response and the making ofdecisionscomplex situationsquicker;
- toimprovethequalityofthe management decisionswhich are made.

Introduction of the system allows to fulfil the following objectives:

- compilation of an archive with electronic images of documents;
- quickaccesstodocumentsofadigitalarchive, the requested document is presented within minimum amount of time(several seconds);
- automatedcontrolandmanagementofreportinformationons torageunits;
- development of complex multicriteria data samples from a digital archive, development of various document collections, quick search of necessary information and generation of necessary reporting forms;
- support of electronic intra- and inter-agency cooperation;
- thecontinuityoforganizational processes connected with the use of the documents of the archive;
- the ability of protected data exchange;
- differentiationofaccessrightstoelectronic documents;
- safedurablestorageof electronic copies of a document;
- the ability to currently replenish the archive electronic information resource.

3 Implementation

The intellectual analysis and the annotation of scanned documents allows to transform the information to knowledge by adding semantic descriptors.Thetransformationofdocumentstoknowledgeismad eautomaticallyusingtheartificialintelligencesystemprinciples based on the use of knowledge - ontologies,that considerably reduces the need to attract the expert regarding knowledge.Navigation and search are made based on the semantics of the documents, that allows to not just increase the relevance of the results, but also to introduce the user into the documents, of the existence of which they were never aware, but which would be relevant for them[6].

The semantic web permits to describe extremely complicated and diversified connectionsbetween the documents. An opportunity to set random attributes of a document allows to describe it more precisely that will certainly improve the quality of search. In the mode of interactive dialogue with a user, there is an opportunity to do not just keyword search, but also reach the documents that are somehow related to the found ones.

Within the automated system "Intelligent software system of digital archives", two separate subsystems or modules should highlighted:

- archive fund management subsystem "Archive";
- paperwork archive fund subsystem "Documents circulation".

Archivesubsystem is operating with the following objects:Fund, Inventory, Archive, and Archive document. This module enables workload distribution by functional roles and scenarios differentiation by stations.Document processing includes the following phases: administering, scanning, separation, indexation and verification.There is a possibility to configure and adjust attributes for different types of documents and to manage the fields containing the indexing data (which will be later used for document search).

There is flexible transaction management which includes an opportunity to correct user mistakes at the level of package management, checking of filled attribute text fields for the input and spelling correctness, checking of data entry completeness, as well as system of various prompts: underlining, pop-ups, etc.

The vital statistics on the processing document is produced, as well as the user statistics by operators working with the documents.

Intelligent search of documents by attributes is based on ontologyand advanced, simultaneous text search (multi-search) by multiple attributes entered in a single text field.Cataloging and structuring of data, data arrangement and the formation of hierarchies, user-friendly navigation through a hierarchical directory are provided. There is an opportunity to produce catalogues of 2-D and 3-D models, a possibility to view the model from different projections using 3D viewer.

Document circulation subsystem enables handling of incoming and outgoing documents related to the archive. Processing of incoming documents includes:

- primary processing of applications (requests);
- registration of incoming documents;
- submission of incoming documents;
- issue, execution and control of execution of orders;
- forwarding of executed documents to files.

Processing of outgoing documents (responses and additional requests) includes:

- preparation of the draft documents;
- the approval and the review of the draft documents;
- signing of outgoing documents;
- registration of incoming documents, and, if necessary,expedited processing and sending of outgoing documents.

Conclusion

Thepaperpresents thesolution, featuresandwaysofimplementationofIntelligent softwarecomplexofdigitalarchiveson the basis of knowledge bases.

Acknowledgements

This work was supported by the Ministry of education and science of the Russian Federation in the framework of the implementation of the Program of increasing the competitiveness of SSAU among the world's leading scientific and educational centers for 2013 – 2020 years.

References

[1] CAPTheorem. https://en.wikipedia.org/wiki/CAP_ theorem
[2] CAP Theorem: Its importance in distributed systems. http://blog.flux7.com/blogs/nosql/cap-theorem-why-does-it-matter
[3] NoSQL. https://ru.wikipedia.org/wiki/NoSQL
[4] Recap: Intro to Graph Databases / Webinar Series #1. http://neo4j.com/blog/recap-intro-to-graph-databases-webinar-series-1/
[5] Baesens B., 2014, "Analytics in a Big Data World: The Essential Guide to Data Science and its Applications" / B. Baesens. – Wiley, 232 p.
[6] Gupta R. et al.: Biperpedia, 2014, "An Ontology for Search Applications. PVLDB 7(7)
[7] Singh G., Pathak R.D., Naz R., 2010, "Service Delivery Through E-Governance: Perception and Expectations of Customers in Fiji and PNG", Public Organization Review, 1566-7170, pp 1-14, Springer Science+Business Media, LLC.

Author Biography

Anton Ivaschenko, Doctor of technical sciences, professor at Samara State Aerospace University (Samara, Russian Federation)
e-mail: anton.ivashenko@gmail.com
Pavel Sitnikov, PhD, Project Management Director at SEC Open Code(Samara, Russian Federation)
e-mail: sitnikov@o-code.ru
Dmitriy Martyshkin, lead scientist at Samara State Aerospace University (Samara, Russian Federation)
e-mail: martyshkin@o-code.ru
Sergey Fedotov, lead scientist at Samara State Aerospace University (Samara, Russian Federation)
e-mail:fedotov@o-code.ru
PavelVasin, Lead Software Engineer at SEC Open Code(Samara, Russian Federation)
e-mail:vasin@o-code.ru