A R C H I V E S O F M E T A L L U R G Y A N D M A T E R I A L S
Volume 58 2013 Issue 3
DOI: 10.2478/amm-2013-0103

S. KLUSKA-NAWARECKA*, K. REGULSKI**, M. KRZYŻAK**, G. LEŚNIAK**, M. GURDA**

SYSTEM OF SEMANTIC INTEGRATION OF NON-STRUCTURALIZED DOCUMENTS IN NATURAL LANGUAGE IN THE DOMAIN OF METALLURGY

SYSTEM INTEGRACJI SEMANTYCZNEJ NIEUSTRUKTURYZOWANYCH DOKUMENTÓW W JĘZYKU NATURALNYM Z ZAKRESU METALURGII

This paper presents assumptions for a system of automatic cataloging and semantic text documents searching. As an example, a document repository for metals processing technology was used. The system by using ontological model provides the user with a new approach to the exploration of database resources – easier and more intuitive information search. In the current document storage systems, searching is often based only on keywords and descriptions created manually by the system administrator. The use of text mining methods, especially latent semantic indexing, allows automatic clustering of documents with respect to their content. The result of this clustering is integrated with the ontological model, making navigation through documents resources intuitive and does not require the manual creation of directories. Such an approach seems to be particularly useful in a situation where we are dealing with large repositories of unstructured documents from such sources as the Internet. This situation is very typical for cases of searching information and knowledge in the area of metallurgy, for example with regard to innovation and non-traditional suppliers of materials and equipment.

Keywords: knowledge engineering, documents processing, ontologies, semantic integration, technological knowledge, metallurgy

Artykuł prezentuje założenia systemu umożliwiającego automatyczne katalogowanie i przeszukiwanie merytoryczne dokumentów tekstowych na przykładzie repozytorium dokumentów dotyczących technologii przetwórstwa metali. System dzięki zastosowaniu modelu ontologicznego ma umożliwić użytkownikowi nowe podejście do eksploracji zasobów bazodanowych – prostsze i bardziej intuicyjne wyszukiwanie informacji. W obecnych systemach przechowywania dokumentów często jedyną formą wyszukiwania jest wyszukiwanie na podstawie katalogu słów kluczowych i deskrypcji tworzonych ręcznie przez administratora systemu. Zastosowanie metod eksploracji tekstu, w szczególności ukrytego indeksowania semantycznego umożliwia automatyczne grupowanie dokumentów pod względem ich zawartości. Wynik takiego grupowania zostaje zintegrowany z modelem ontologicznym, przez co nawigacja poprzez zasoby dokumentów staje się intuicyjna i nie wymaga tworzenia ręcznie katalogów. Takie podejście wydaje się szczególnie przydatne w sytuacji, gdy mamy do czynienia z dużymi repozytoriami nieuporządkowanych dokumentów pochodzących m.in. z sieci Internet.

1. Introduction

Programs to improve the process of organizing knowledge stored in documents and text mining tools are very popular nowadays because in the continuously enlarging document repositories, manual search to find the most appropriate document is extremely tedious and time-consuming. Moreover, full-text search does not allow finding documents related to the preset word. Also, rapidly developing intelligent systems with knowledge bases create the need for tools for automatic recognition of the substance of the documents [1-4].

In this article, an attempt has been made to describe the process of creating an application that allows text document search based on keyword given, to display next to the user all the documents that are related to both the entered word and the synonymous words.

The system is divided into three modules that can be used by different users. The first module allows us to create a matrix of the keyword frequency, generated from a set of text documents retrieved from database, which is a bibliographic casting database containing abstracts of articles published in various casting journals, proceedings of conferences, and R&D works, and then its transformation into a TF-IDF matrix of weights with regard to the keyword entered by the user. We obtain in this way naturally occurring theme groups related to synonymous words. Thus created frequency matrix is transferred to the next module, which is responsible for assigning each group of documents to the corresponding classes of do-

* FOUNDRY RESEARCH INSTITUTE, KRAKÓW, POLAND
** AGH UNIVERSITY OF SCIENCE AND TECHNOLOGY, FACULTY OF FOUNDRY ENGINEERING, REYMONTA ST. 23, 30-059 KRAKÓW, POLAND

main ontology in the field of metallurgy, to enable intuitive semantic search in the specified database.

2. Latent Semantic Indexing (LSI)

The frequency matrix is a set of vector representation of documents. Keywords are derived from a thesaurus and are related with the scope of metallurgy and metal processing. Thus created, the term frequency matrix is a sparse matrix, owing to which a Singular Value Decomposition (SVD) algorithm can be used to reduce the dimensions [5]. The method of Latent Semantic Indexing allows for replacement of individual subsets of the keywords with single terms, so-called, pseudowords, which are a weighted combination of the occurrences of the original keywords. The assumption, upon which the whole technique is based, stems from the fact that a single vector, which is a weighted combination of the occurrences of the original keywords, can better reflect the semantic content of the document [6]. As a result, the original matrix of occurrences of words of the NxM size can be replaced by a matrix of the NxK size, where k <<M (with a small loss of information) [7,8]. LSI creates relationships between keywords by creating new "pseudo-words", which are a more accurate way to express the semantic content of the documents.

The term frequency matrix (TF) is transformed into a TFIDF weight matrix (*TF – term frequency, IDF – inverse document frequency*), defined as a term frequency weighting (where the term is understood as an invariable core of keywords). This method allows the calculation of the weight of words based on the number of their occurrences. TFIDF is designed to provide information on the frequency of occurrence of terms, taking also into account the appropriate balance between the local significance of term and its meaning in the context of the full collection of documents. In other words, the terms are weighted relative to their discriminative power for a given document repository. The value of TFIDF is calculated from the following equation:

$$(tf - idf)_{i,j} = tf_{i,j} \times idf \tag{1}$$

where: $tf_{i,j}$ is the, so called, "term frequency", expressed with the equation:

$$tf_{i,j} = \frac{n_{i,j}}{\Sigma_k n_{k,j}} \tag{2}$$

where $n_{i,j}$ is the number of occurrences of term (t_i) in document d_j, and the denominator is the total number of occurrences of all the terms in the document d_j.

idf_i is the "inverse document frequency", expressed with the equation:

$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|} \tag{3}$$

where |D | is the number of documents in the database, $|\{d : t_i \in d\}|$ is the number of documents that contain at least one instance of the occurrence of a given term.

The next step is to transform the TFIDF matrix by decomposition to the main components of SVD. The reason for the use of SVD is that it allows for the best possible reproduction of the matrix with the minimum amount of information. The SVD technique relies on the detection and kicking off from the processed data of items that do not have a large impact on the matrix performance (so-called background, noise).

3. Application of ontology

Ontologies in computer science emerged as a response to the need for a strong, sufficiently expressive and unified formalism for modeling of knowledge that can not be algorithmized. To build the OWL language ontology, a description logic (DL) was used. The description logic is a subset of first-order logic (FOL), which can be used to represent a domain in a formalized and structured manner, computer-processable. It is based on the assumption that it is possible to get the semantics based on first-order logic, if the essential element of representation will be unary predicates corresponding to the set of objects, and binary predicates mapping the dependencies between objects. The description logic enables creating *descriptions* describing the domain with *concepts* (*unary predicates*) and *roles* (*binary predicates*).

The formalism of ontology allows us to unify the way by which the document repositories are searched. The ontology is on principle shared by many users, but mainly by other systems. The behavior of a single formalism over a long period of use is a key aspect of the utility of an ontological model. Ontology is *per se* the right tool for the creation of a model of knowledge which is necessary to describe the area the users are expected to share in a distributed network. An additional convenience for users should be some kind of dictionary, a thesaurus comprising concepts included in the areas for which definitions should be created in the system. Thesaurus as a base document already contains a structure that should be preserved in the process of creating ontologies. It is a kind of hint for an ontology engineer in what context, a new concept / class should be placed and found.

The integration of data resources and knowledge (as well as sources added during the use of the sources) consists in describing the resources with terms (concepts) of ontology, and then in mapping their structure onto the base ontology components. For each class in the ontology, a description in natural language can be added. The ontology editing tools such as Protégé permit the placement of descriptions in the form of plain text directly in the description of an OWL ontology. They also allow placing a reference in the form of an URL address. This gives the possibility to transfer the unstructured knowledge that are definitions in natural language and photographs to the knowledge base.

The aim of the system is to use the classes of domain ontology in the area of metallurgy to organize the documents that have previously been grouped with the use of "pseudowords" and the LSI technique.

4. Semantic integration

A module of semantic integration of documents is intended to effectively operate the resulting data obtained through text mining using LSI to combine them with the domain ontology (Fig. 1).

The system processes the classes of domain ontology to classes of object model, and then attaches documents to the

classes and writes the object model to the ontological model. The final step is transfer of a new OWL file to the System that handles queries to the ontological model.

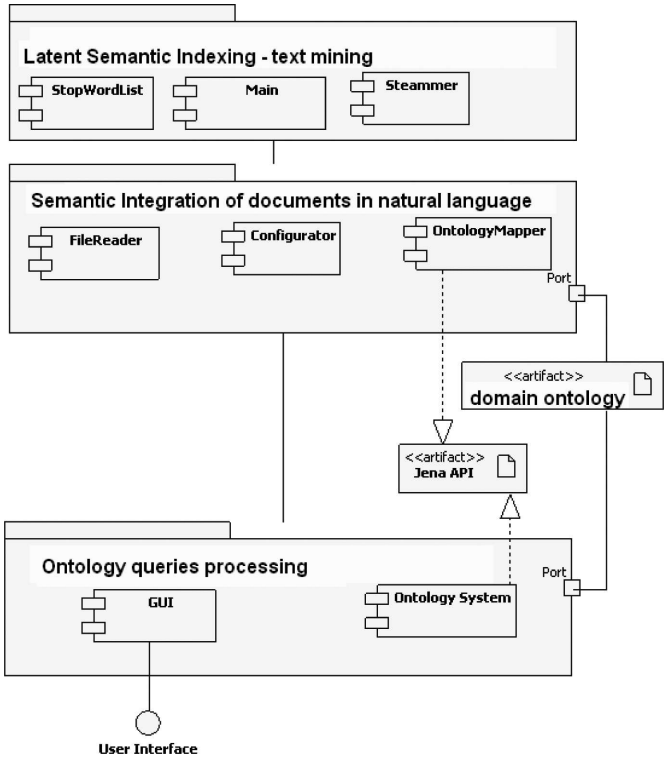

Fig. 1. Component and package diagram of the system

The class diagram (Fig. 2) is a scheme in which the elements of the system are presented in terms of an object-oriented programming. The diagram omits the external classes taken from libraries or frameworks, among others, also from JENA framework, since they are not a direct result of the work of the authors and have been used only for the purpose of efficient and optimal implementation of the project.

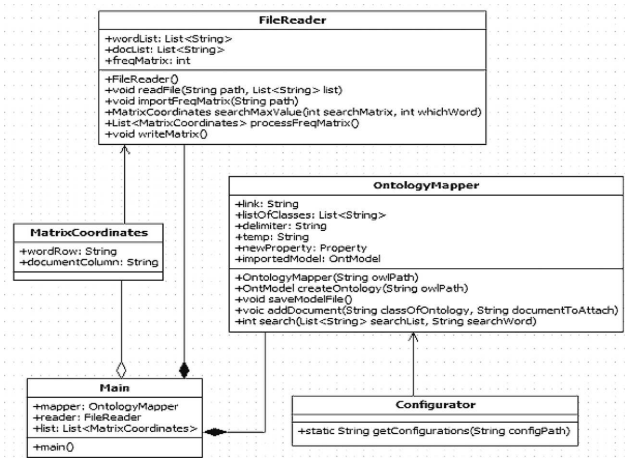

Fig. 2. Fragment of a class diagram

OntologyMapper is the most important class in the system. It is responsible for all activities related with reading, modification and writing of ontologies. To a large extent it uses the classes derived from the JENA Framework, such as, for example: OntModel and Property. This solution is compact and efficient by taking advantage of the Java language and JENA framework. The input data are supplied to the system in a manner consistent with its objectives, The documents are connected to the model of ontology classes owing to the newly created property ”hasDocuments” and the model is saved to the file ”NewOntology.owl”. The program presents a hierarchy of classes in the model and their properties, in a way such as to enable viewing the structure of the entire ontology. When the user selects from a list of the displayed classes, a class interesting to him, the system reads its name from the table and creates next a list of the properties containing also a list of the corresponding documents.

The *ClassTableModel* and *ClassPropertiesTableModel* provide a model to display data in tables. The former one has references to a list containing the names of all classes included in the model, while the latter one has references to a list containing detailed information about a particular class (Figure 3). As you can see from the figure, both considered class as well as the reference ontology correspond to the thematic focus of the documents related to metallurgy and casting in polish language.

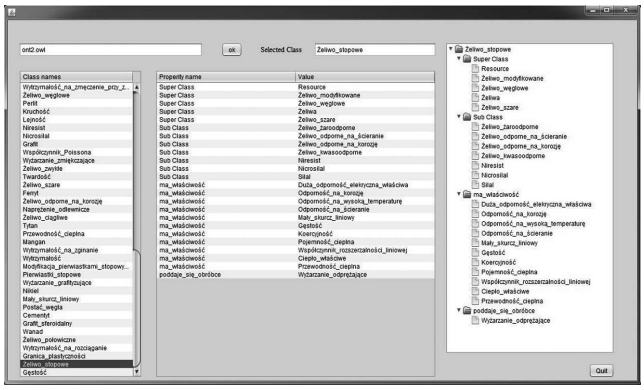

Fig. 3. Program supplying information about the selected class

5. Summary

The application presented here has met the preset assumptions and is able to operate not only as a system for integration of documents, but also as a standalone graphical user interface enabling the user to view the hierarchy of the ontology. It may also work with a variety of other applications, forming or editing ontologies, as well as serve ordinary home users.

The problem of ontology creates a picture of untapped opportunities. It enjoys a growing interest in recent years and the increasing number of implementations. Although semantic technologies already exist and function in the modern world, the average user usually lives unaware of issues such as semantic technology and ontology. Ontology and methods of its use often arouse much controversy, and opinions on ways and meaning of its implementation are divided. Programmers face difficulties such as, among others, different definitions of ontology and different applications, a multitude of available technologies and many tools that were created and existed for some time (or still exist), but having no technical support can not keep up with developments in this field and become use-

less. This applies particularly to many applications and utilities that run on an open source license.

Despite their flaws, ontologies offer a very interesting approach to the issues related with the processing of textual data and modeling of knowledge. In conjunction with text mining methods, they allow creating an efficient and ergonomic tool changing the way in which the text repositories are searched. Due to the limited volume, the work focuses on the description of the methodology for the construction and functionality of the implemented application. It should be clear that in the area on both knowledge sources used (text documents) as well as the applied ontology, and finally the expected use cases of considered system is dedicated to the needs of metallurgy and materials technology.

Acknowledgements

The work was financed in the framework of the international project No. 820/N-Czechy/2010/0 of 30 November 2010.

REFERENCES

[1] E. Nawarecki, S. Kluska-Nawarecka, K. Regulski, Multi-aspect Character of the Man-Computer Relationship in a Diagnostic-Advisory System, Human-Computer Systems Interaction: Backgrounds And Applications 2, Pt 1 Eds. Hippe, ZS; Kulikowski, JL; Mroczek, T. **98**, 85-102 (2012).

[2] Z. Gorny, S. Kluska-Nawarecka, D. Wilk-Kolodziejczyk, K. Regulski, Diagnosis of casting defects using uncertain and incomplete knowledge, Archives of Metallurgy and Materials **55**, 3, 827-836 (2010).

[3] Z. Jančíková, O. Zimný, P. Koštial, Prediction of Metal Corrosion by Neural Networks. Metalurgija **52**, 3, 379-381 (2013), ISSN 0543-5846.

[4] R. Frischer, J. David, M. Vrožina, Neural Networks Usage at Crystallizers Diagnostics. METAL 2011, s. 134, Tanger, spol. s r.o. Ostrava, Ostrava, 2011. ISBN 978-80-87294-22-2.

[5] D. Hand, H. Mannila, Padhraic Smyth, (2001) Principles of Data Mining, The MIT Press, United States of America, 277-280.

[6] G.H. Golub, Ch.F. Van Loan, Matrix Computations, Johns Hopkins University Press, Baltimore, Maryland, 48-86 (1996).

[7] M.W. Berry, S.T. Dumais, G.W. O'Brien, Using linear algebra for intelligent information retrieval. SIAM Review, 573-595, December 1994.

[8] J. Ramos, Data Using TF-IDF to Determine Word Relevance in Document Queries, Morgan Kaufmann Publishers, United States of America, 2006.

This article was first presented at the VI International Conference "DEVELOPMENT TRENDS IN MECHANIZATION OF FOUNDRY PROCESSES", Inwałd, 5-7.09.2013

Received: 20 January 2013.