DDTA - Digitalisation of Districts in the Textile and Clothing Sector

Floriana Esposito[1], Stefano Ferilli[1], Nicola Di Mauro[1], Teresa M.A. Basile[1], and Marenglen Biba[2]

[1] Department of Computer Science, University of Bari "Aldo Moro"
{esposito,ferilli,ndm,basile}@di.uniba.it
[2] Computer Science Department, University of New York Tirana
marenglenbiba@unyt.edu.al

Abstract. The main goal of the project was the development of a District Service Center for the SMEs of the Textile and Clothing sector. In particular, it investigates the introduction of innovative technologies to improve the process/product innovation of the sector. In this direction, the research unit proposal consisted in introducing document processing and indexing techniques on a variety (both for structure and content) of document formats whit the aim of improving the exchange of data among companies and the semantic content-based retrieval for the real companies'needs.

1 Introduction and Motivation

Since the seventies, the district represented one of the most common industrial model adopted from italian companies to face the crisis. It consists of an action of hiving-off (production decentralization) that relies on partitioning the work among companies in the same productive sector with specific professional abilities. This kind of model, however, rises the question about the interoperability among companies involved in it. Hence the need of improving communication standards and information sharing and exchanging.

On the other hand, one of the key sector in the Made in Italy districts is the Textile and Clothing (TC) one as it represents about the 28,8% of the total industrial production. Thus it seems to be the perfect test-bed to start the investigation of a policy aimed at introducing innovative technologies in order to improve the process/product innovation of the sector.

To this aim, the DDTA project (Apulia Region Project (2007-2010) - www.tessilpuglia.com) focuses on activities for the definition and diffusion of standards for interoperability in order to facilitate the cooperation and the collaboration among companies and for the development of ICT solutions specifically addressing the TC sector along with the development of a portal for delivering services to model and to support the District Service Center.

In response to these challenges, the policy stated the following objectives: to facilitate SMEs access to systems of digital integration; to create a network among districts of Apulia Region to support the diffusion of management and

M. Agosti et al. (Eds.): IRCDL 2011, CCIS 249, pp. 119–122, 2011.

technological best practices in the usage of ICT and to implement already existing service centres in the district areas or to create new structures to supply companies with supporting services in the areas of process and product innovation, market intelligence, ICT usage.

The achievements from this activity could made it possible to reach a clear and up-to-date idea of the state of the art and on the development trends relative to interoperability standards, IT solutions and initiatives in progress, bring the regions to adopt already existing and shared standards, maintain relations with other organizations for the definition of a national/international standard, promote the extension and improvement of the sector coding.

2 Scientific Challenges

The companies involved in a district must declare their skills in order to be selected for a specific work. Furthermore, in a more general business-oriented point of view, the companies have to be provided the right and accessible information about public announcements, import/export regulations, trade acts and so on. Exchanging data can be a challenge due to, among others, different specifications of formats and varieties of categorisations. If data is interpreted differently, collaboration is limited, takes longer and is not efficient.

The primary criterion for interpreting documents is by content. Hence, the documents in the repository should be grouped and organized accordingly. However, doing this manually is very expensive, and doing it automatically is very difficult due to the need of capturing document meaning (i.e., semantics). A more tractable starting point is exploiting layout analysis (i.e., syntax). Indeed, probably significant content is often placed in particular layout components, so that being able to identify the typical components of each group allows to selectively read only those components for identifying the document content. As a consequence, the ability to handle and manage documents according to layout structure can be a key factor towards the ability to reach content-based management as well.

However, the indexing process remains a key issue. A problem of most existing word-based retrieval systems consists of their ineffectiveness in finding interesting documents when the users do not use the same words by which the information they seek has been indexed. This is due to a number of tricky features that are typical of natural language. One of the most common concerns the fact that there are many ways to express a given concept (synonymy), and hence the terms in a users query might not match those of a document even if it could be very interesting for him. Another one is that many words have multiple meanings (polysemy), so that terms in a user's query will literally match terms in documents that are not semantically interesting to the user. Moreover, in case of documents such as public announcements, import/export regulations and trade acts, the retrieval of relevant information becomes more difficult as the notification writer uses a technical terminology while the final user (companies'employee) doesn't.

3 Key Technologies

Organizing the documents on the grounds of the knowledge they contain is fundamental for being able to correctly access them. The key technologies to be exploited to reach this aim are: first-order incremental machine learning techniques for document layout processing, to classify documents and label their most significant components with the aim of representing the documents according to a standard format, and semantic indexing methodologies to guarantee an effective retrieval of relevant information according to a semantic level of the content.

The problem of document layout processing in Digital Libraries environments requires a first-order language representation as the variety of documents without a common standard does not allow to represent the document with a format made up of a set of fixed components. Moreover, first-order setting can model and efficiently handle the relationships coming from the topological structure of all components in a document that result very useful in document labelling. Finally, the continuous flow of new and different documents calls for incremental abilities of the system that must be able to update or revise a faulty knowledge previously acquired for identifying the logical structure of a document.

As for the indexing methodology, the weaknesses of term-matching based retrieval is overcome by Latent Semantic Indexing (LSI) technique whose basic idea is that there exists some underlying latent semantic structure in the data that is partially obscured by the randomness of word choice with respect to the retrieval phase and that can be estimated by means of statistical techniques. LSI relies on a mathematical technique called Singular-Value Decomposition (SVD). Starting from a matrix of term-document association data, the SVD allows to build and arrange a semantic space, where terms and documents that are closely associated are placed near each other, in such a way to reflect the major associative patterns in the data, and ignore the smaller, less important influences. The continuous flow of new documents that could be added to the initial database, requires an incremental methodology to update the initial LSI matrix. Two techniques have been developed in literature to update an existing LSI generated database: Folding In and SVD updating. The former uses the existing SVD to represent new information but yields poor-quality updated matrices, since the information contained in the new documents/terms is not exploited by the updated semantic space. The latter represents a trade-off between the former and the recomputation from scratch.

4 Contribution by the Research Group

The contribution of the research unit converges in the development of a document management system that intensively exploits intelligent techniques to support different tasks of document processing from acquisition to indexing, from categorization to storing and retrieval (1).

A central role is played by the Learning Server, which intervenes during different processing steps in order to continuously adapt the knowledge taking into consideration new experimental evidence and changes in the context.

The layout analysis process on documents in digital format starts with a preprocessing module that rewrites basic PostScript operators to turn their drawing instructions into objects. It takes as input a digital document and produces the initial document's XML basic representation, that describes it as a set of pages made up of basic blocks. Due to the large number of basic blocks discovered, an aggregation step is necessary. Since grouping techniques based on the mean distance between blocks proved unable to correctly handle the case of multi-column documents, such a task was solved by exploiting a kernel-based method, implemented in the Learning Server, that is able to generate rewriting rules that suggest how to set some parameters in order to group together blocks to obtain lines. After that, a module collects the semantically related blocks into groups by identifying the surrounding frames based on white spaces and the results of the background structure analysis. At the end of this step, some blocks might not be correctly recognized, hence a phase of layout correction is automatically performed by exploiting embedded rules stored in the theories knowledge base. Such rules were automatically learned by a first-order incremental learning system implemented in the Learning Server from previous manual corrections collected on some document during the first trials.

Once the layout structure has been correctly and definitely identified, a semantic role must be associated to each significant components in order to perform the extraction of the interesting text with the aim of improving document indexing. This step is performed by firstly associating the document to a class that expresses its type and then associating to every significant layout component a tag expressing its role. Both these steps use the theories previously learned and in case of failure the theories are properly updated by means of a first-order incremental learning system that runs on the new observations and tries to modify the old theories in the knowledge base. At the end of this step, both the original document and its XML representation, enriched with class information and components annotation, is stored in an internal document database. Finally, the text is extracted from the significant components and the Indexing Server, the module implementing the LSI techniques, is called to achieve a semantic indexing of document content useful for an effective content-based retrieval.

Acknowledgements. The work was partially supported by the Regional Project "DDTA - Distretto Digitale a supporto della filiera produttiva del Tessile-Abbigliamento (2007-2010) - www. tessilpuglia.com ".

Reference

1. Esposito, F., Ferilli, S., Basile, T.M.A., Di Mauro, N.: Machine learning for digital document processing: from layout analysis to metadata extraction. In: Marinai, S., Fujisawa, H. (eds.) Machine Learning in Document Analysis and Recognition. SCI, vol. 90, pp. 105–138. Springer, Heidelberg (2008) ISBN: 978-3-540-76279-9