A cognition-based approach for querying personal digital libraries.

A. Malizia
Dept. of Computer Science, University "La Sapienza" of Rome
Via Salaria 113, 00198 Rome, Italy
malizia @di.uniroma1.it

Abstract.

The decreasing cost and the increasing availability of new technologies is enabling people to create their own digital libraries. One of the main topics in personal digital libraries is allowing people to select interesting information among all the different digital formats available today (pdf, html, tif, etc.). Moreover the increasing availability of these on-line libraries, as well as the advent of the so called Semantic Web [1], is raising the demand for converting paper documents into digital, possibly semantically annotated, documents.

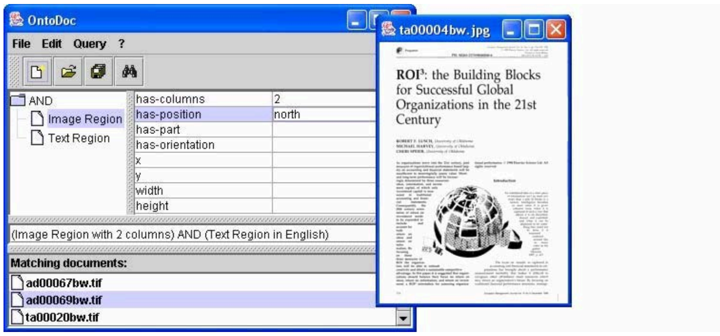
These motivations drove us to design a new system which could enable the user to interact and query documents independently from the digital formats in which they are represented. In order to achieve this independence from the format we consider all the digital documents contained in a digital library as images. Our system tries to automatically detect the layout of the digital documents and recognize the geometric regions of interest. All the extracted information is then encoded with respect to a reference ontology, so that the user can query his digital library by typing free text or browsing the ontology.

This approach could help users, because they don't need to know SQL since they could only recall visual hints about a document: like find me a document with a picture on the left side and a big title centered on top of the page.

Introduction.

The main goal of our system, OntoDoc, is to allow users to query their own personal digital libraries in an ontology-based fashion. An ontology [2] specifes a shared understanding of a domain of interest. It contains a set of concepts, together with its definitions and interrelationships, and possibly encodes a logical layer for inference and reasoning. Ontologies play a major role in the context of the so called Semantic Web [1], Tim Berners-Lee's vision of the next-generation Web, by enabling semantic awareness for online content.

OntoDoc uses a reference ontology to represent a conceptual model of the digital library domain, distinguishing between text, image and graph regions of a document, providing attribute relations for them, like size, orientation, color, etc. In order to classify a document, OntoDoc performs a first layout analysis phase, generating a structured, conceptual model from a generic document. Then the conceptual model goes through an indexing phase based on the features in the model itself. Finally, the user can query his digital library by typing free text or through composition of semantic expressions. The query system is particularly suited to express perceptual aspects of intermediate/high-level features of visual content, because the user does not have to bother thinking in terms of inches, RGB components, pixels, etc. Instead, the user can query the system with higher level, although accurate, concepts (e.g. medium size, black color, horizontal orientation etc.), moreover thanks to the use of the ontology the user domain could be taken into account in order to let the query expressions adapt to its knowledge.



AND
Image Region
Text Region

has-columns	2
has-position	north
has-part	
has-orientation	
x	
y	
width	
height	

(Image Region with 2 columns) AND (Text Region in English)

Matching documents:
ad00067bw.tif
ad00069bw.tif
ta00020bw.tif

ROI³: the Building Blocks for Successful Global Organizations in the 21st Century

Fig. 1. The System window, with a matched digital document on the right, the query was: find me a document with an image region, two columns text and text region in english.

The Reference Ontology.

The reference ontology encodes all the required information about the digital library domain. Documents and Regions are represented as resources with a number of attributes in common, e.g. Orientation, Size, etc. The ontology encodes specific relations between subconcepts of Region, i.e. Text, Image and Graph Regions, and several attribute concepts, like Color, Size, Orientation, Reading Direction and so on (a snapshot of the concept

Proceedings of the 2004 IEEE Symposium on Visual Languages and Human Centric Computing (VLHCC'04)
0-7803-8696-5/04 $ 20.00 IEEE

taxonomy is shown in Figure 2). Encoding ontological concepts instead of numerical attributes is a peculiar feature of our system. The user can think of concepts instead of low level measures (e.g. inches, pixels etc.) and submit queries like "all the documents with a text region in the south part and an image region with 2 columns having an inner table". User can query the database also using keywords extracted from the content using digital indexes or OCR for scanned doccuments.

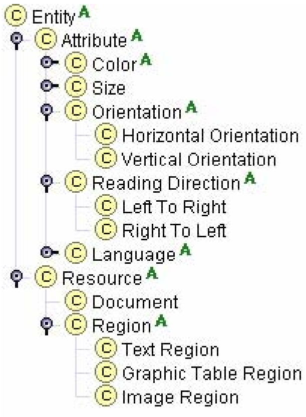


Fig. 2. A portion of the reference ontology for digital libraries.

Experimental results and Conclusions.

We have tested our system over the UW-II database that is the second in series of document image databases produced by the Intelligent Systems Laboratory, at the University of Washington, Seattle, Washington, USA. This database was particularly useful because it contains 624 English journal document pages (43 complete articles) and 63 MEMO pages. All pages are scanned pages. Each document in the database has been taken from scientific journals and contains text, graphs and images. All the images were already annotated with labels for the region type (image, text, ...) and sizes. The experiments have been carried out in 2 phases: the first phase was to the test our layout analysis module over the UW database to verify the percentage of the automatic classification of the digital documents regions; while the second phase was performed on 10 users in order to measure the ability of the system in helping them to retrieve documents.

For the first experiment, concerning the classification abilities of our system, we have tested it over the entire database (600 images) obtaining an 84% of correctly recognized regions, 14% of incorrectly recognized and 2% to be defined. Of course, errors at this stage affect the querying precision. A method to overcome this inconvenience is an ongoing work. The 84% of correctly recognized regions could be subdivided into a 59% of entirely recognized and a 25% of partially recognized, which means that some regions were assigned to the right class and some others not, for example a single text region was interpreted as two text regions (this usually happens in titles with many spaces). For the second phase, all tests have been carried out using the relevant feedback process by which the user analyzes the responses of the system and indicates, for each item retrieved, a degree of relevant/non-relevant or the exactness of the ranking [3]. Annotated results are then fed back into the system, to refine the query so that new results are more fitting. The experiment was implemented showing to the users 10 different documents and then asking them to retrieve the documents from the entire UW database using our query module. On the qualitative side, our system proved to be highly effective because the users concentrated on the conceptual content of documents rather than on numerical information about them, allowing faster and more accurate retrieval of the desired documents with respect to keyword-based non-ontological retrieval. A major improvement of OntoDoc may be in the classification phase. In fact, the system could classify shapes like subject images or specific geometry on the basis of their ontological descriptions (for instance, finding a document with an image of an apple, or with a pie-chart).

References.

[1] Berners-Lee, *Weaving the Web*, Harper, San Francisco, 1999.

[2] B. Smith, and C. Welty, Ontology: towards a new synthesis, *Proc. of Formal Ontology in Information Systems FOIS-2001*, ACM Press, October 2001.

[3] G. Nagy, Twenty years of document image analysis, *PAMI, IEEE Trans. Pattern Analysis and Machine Intelligence*, 1/22 pp: 38-62, 2000