Intelligent Search via Ontology Driven Metadata Analysis

Thomas Mavroudakis and Haralampos Karanikas
National & Kapodistrian Univ. of Athens,
Knowledge Management Lab., Greece
{ tmavroudakis, bkaranikas, }@phs.uoa.gr

Abstract

In this paper we present an overview of MetaOn project. The core objective of MetaOn is to construct and integrate semantically rich metadata collections extracted from documents images and linguistic resources, to facilitate intelligent search and analysis. The proposed MetaOn framework involves, ontology-based information extraction and data mining, semi-automatic construction of domain specific ontologies, content-based image indexing and retrieval, and metadata management. The Hellenic history has been chosen as a challenging application case study and Protégé as the main tool for this research project.

1. Introduction

We are living in the "Information Age". The amazing growth of data that are being generated from heterogeneous sources makes it difficult for humans to manage it and mine useful information. In most companies and organisations employee time and effort is wasted in ineffective searches in the web or other conventional sources. The information overload is further exacerbated due to the unstructured format of the majority of the data. The vast amount of data found in an organisation, some estimates run as high as 80%, are textual such as reports, emails, etc. [1]. However, other types of data such as image, video and speech are becoming increasingly popular. These types of data usually lack metadata (data about data) and as a consequence there are no standard means to facilitate search, query and analysis. On the other hand, Web is the biggest document (text and image) collection and most of the current Web's content is designed for humans to read, rather than for data and information that can be processed automatically by computer programs [2].

In this paper, we present the vision underlying the *MetaOn* project within the Operational Programme "Information Society" of the Greek Ministry of Development, General Secretariat for Research and Technology, co-funded by the European Union. The *MetaOn* project aims to develop theory, techniques and a prototype system for efficient metadata management, flexible querying and knowledge discovery and delivery, based on image and text collections. Our approach is applied to a number of such multimedia data coming from The Foundation of the Hellenic World (FHW) that concerns Hellenic history. In the context of the project FHW provides its digital content that is available on the servers of the Cultural Center (www.e-history.gr). The digital content varies from high resolution 3D models of monuments, to architectural drawings of temples, GIS maps, video captures and audio recordings, still images and textual information. One important requirement of the case study is the discovery of the underlying semantic structure of the data independently of the usual terminology (text, image label, image features) by utilising a domain specific ontology. This process will provide a general model of query formulation for the basic classes and associations. The integration of the text and image features with the domain knowledge in order to facilitate intelligent search is an additional line of research. Finally there is a requirement to deal with modern Greek which is a language difficult to describe computationally (e.g. due to several different inflectional endings, prefixes, infixes and stress marks participating to the inflection or conjugation of Modern Greek words, resulting to 4-7 word-forms for a noun and up to 250 word-forms for a verb).

In summary, the main contributions of *MetaOn* are:

- Ontology-based (protégé) extraction of metadata from Greek and English documents.
- Incorporation of background knowledge via the use of domain specific ontologies in the discovery and integration (of text and image) process.
- The development of novel image feature extraction methods to facilitate efficient and effective image indexing and retrieval.
- Efficient metadata management that will facilitate search and analysis on the previous mentioned forms of data.
- The evolution of state of the art content-based indexing and retrieval methods and the development

of novel ones under an ontology-based integration framework facilitated by a unified metadata scheme.

The rest of the paper is organised as follows. In section 2 we discuss the state of the art of the involved domains upon the combination of which our approach is based. In section 3 we describe in detail our approach, while in section 4 we conclude and give the next steps in our development.

2. Related work

No single technology can cover the range of expertise that is required to build an intelligent multimedia querying and analysis system. The originality of the proposed approach mainly grounds on combining techniques coming from the following disciplines, namely, semi-automatic construction of domain specific ontologies and information extraction, content-based image indexing and retrieval and metadata/pattern management.

Organisations need methods that will embrace legacy unstructured data (external information, as well as current documentation), extract structured data, and use data mining techniques to learn from it. The technologies of information extraction (IE) [3] and Text Mining [4] provide the technical basis for this. Linking documents and extractable phrases from them to an ontology has been demonstrated in [2],[5] using a manually-constructed set of rules and a hand-crafted ontology.

Integrating various media stream features and domain knowledge is a significant task. Currently ontologies play that role. Ontologies are often constructed manually, although there are standards, such as ISO standard 13250 Topic Map format, by which information can be imported and exported between such structures. Other standards of interest include DAML/OIL and RDF.

Efficient and effective indexing and retrieval of large scale image databases has been the focus of multimedia research in the last decade. Although various content-based indexing and retrieval systems have been devised [8],[9] the development of a system that will utilize heterogeneous metadata derived from various sources, such as texts, linguistic resources and images, still remains a challenge. Recent works of relevant context include the application of a semantically-relevant hierarchical clustering approach [10] and the Semantics-sensitive Integrated Matching for Picture Libraries (SIMPLIcity) [8],[11]. However, there is still much to be done for the improvement of the methods proposed, either in terms of image representation or image classification. Furthermore these methods have not been yet considered under a unified, ontology-based framework. Such an approach could possibly contribute to the enhancement of the image retrieval performance [12].

There are many efforts focusing on the efficient management of patterns in both industry and research. Industrial standards and specifications have been evolved aiming at the exchange of patterns between applications. Suggestively we refer the Predictive Model Markup Model (PMML) [13], and Common Warehouse Metamodel (CWM) [14]. Research works aim at defining a general framework for the support of the various management operations regarding to patterns. Such frameworks include FOCUS [15], PANDA [17], PSYCHO [6], and PAM [16].

3. The MetaOn framework

The intent of the MetaOn project is to carry out research aimed at prototyping and evaluating a novel framework for the management and analysis of text and image collections by integrating data mining, image processing, metadata management and information extraction techniques. Furthermore, the project adopts an ontology-based approach in order to facilitate the integration of information from different sources and formats into a multimedia warehouse.

The methodology we describe is a generic approach that can be applied to text and image databases of varying complexity. In our approach the process of detecting patterns within and across text documents depends mainly upon information extraction techniques. By identifying key entities in a text along with the image characteristics, one can draw relationships among them.

The overall architecture of our system is illustrated in Fig. 1. First the document and image collector performs basic processing including format transformation etc. Then in parallel the text and image processing modules elaborates the input text/image in order to extract the key features. The process of text processing is supported by the ontology.

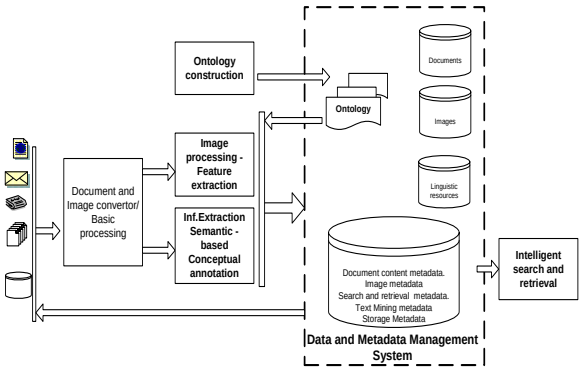

Figure 1. The *MetaOn* framework

A metadata repository is being utilized by our approach to store data that describe the information extracted from the documents and images. More specifically the role of the repository is to provide a

neutral medium and unified representation forming the basis for analysis. Additionally provides a more solid basis as it allows large scale applications in text and images.

3.1. Ontology based information extraction

This step deals with the identification and extraction of conceptual information from unstructured documents, through application of information extraction techniques to individual documents. An incoming document is firstly converted to XML. During this process, text zones (timestamps/dates, any pre-existing metadata, headings, text sections of various types, lists, tables, figures, etc.) are identified and annotated to aid the information extraction task. Conceptual annotation then takes place and results are stored in the multimedia warehouse for indexing and further analysis purposes.

Annotation covers basic semantic elements such as named entities and also relationships between such entities. The full representation of such relationships (i.e., the representation of multiword terms) will be realized in the form of ontologies. Types of entities and domain specific terms to be extracted will be specified in consultation with the domain expert user partner. The approach is semi-automatic: a human editor checks and if necessary amends the annotation results, in a user-friendly interface, thus ensuring high-quality annotation. One important issue is the exploitation of ontological and terminological information provided, thus improving recognition of concepts and leading to richer links between entities and domain specific terms. Additionally exploitation of text zone information will improve the accuracy of extraction. In contrast to other approaches, where entities of interest belong to a rather restricted set (persons, places etc.), *MetaOn* provides an integration of information extraction rules with a domain specific ontology (protégé based). The ontology provides extensibility with respect to the easy enhancement of the extracted conceptual information set.

3.2. Construction of domain specific ontologies

MetaOn aims at deploying several techniques in order to develop domain specific ontologies (protégé based). Innovation arises from the individual techniques and also from the integration of ontology-related techniques and linguistic processing techniques. All techniques focus on the types of document and domains defined by a domain expert user. As the quality of knowledge extraction and integration depends on the quality of the domain ontologies, all additional new concepts, terms, and relations, will be manually validated by domain experts.

We focus on approaches to semi-automatically discover candidate concepts, using data-mining and linguistic processing techniques on domain text corpora. Innovative knowledge discovery techniques, including clustering, and association rules, will be developed to discover new concepts and relationships among derived concepts.

Initially we carry out semi-automatic term recognition, to identify high-quality candidate multi-word terms in domain texts, using a mixture of linguistic, contextual and statistical information. The main work here is in adapting the techniques for the chosen document types and discovering the most appropriate parameters for the chosen domains.

The validated concepts and terms are import and merge with existing ontologies using a Protégé Tab that is build for the purpose of this project. The ontology engineer, easily, can drag and drop concepts from this tab - that contains the results of the linguistic process - into standard protégé tabs.

3.3. Content-based image indexing and retrieval

A content-based image indexing and retrieval system will be supporting the *MetaOn* framework. Both indexing and retrieval involve partitioning of the whole or part of the input image into blocks. From each block, a set of low-level features that quantify local image color and texture will be extracted to form a representative signature for this block.

During indexing, the whole input image will be partitioned into blocks. Following the low-level feature extraction from each block, overlapping windows of multiple sizes, called indexing subimages, will be sliding over the whole image with a given step to capture more abstract information from groups of blocks at various levels. This information will be encoded in terms of simple statistical features, such as moments, estimated over the low-level features of the corresponding blocks. These features will comprise meta-signatures that encode the content as well as the ordering of the blocks of the subimages.

Image retrieval will be based on a user query that specifies which image regions are of interest, which ones are neutral and which ones are not of interest within an input image. These regions will be then partitioned into blocks, from which low-level signatures and meta-signatures will be extracted. The meta-signatures will be used for efficient retrieval of images that contain or do not contain the regions specified in the user query. In the sequel, the signatures from the retrieved set of images will be used for the refinement of the results. We consider the use of less complex measures for the quantification of the similarity of the meta-signatures, whereas the similarity of the block signatures can be measured by more complex but effective ones to expedite accurate retrieval, such as the SamMatch measure [7].

A semantics-sensitive approach to content-based indexing and retrieval is considered to be the interface joining the images and the semantics of the underlying ontology [8]. Feature extraction and classification are considered in a multi-level hierarchy of ascending complexity from level-0 to level-*L*. At each level, the different classes correspond to different semantics. Thus, a training phase will be required for the system to learn the correspondences between feature vector clusters and semantics. Various classification schemes will be researched so as to achieve optimal classification performance for each classifier in the hierarchy. It is worth noting that we focus on incremental classification schemes, as they are advantageous in that they do not require retraining from scratch each time a new image is inputted for training. The number of levels *L* of the hierarchy will be determined upon the design of the ontology.

3.4. Metadata management

In MetaOn the efficient management of patterns is essential due to the variety and complexity of the types of data involved (text, images). Taking advantage of the PANDA framework [17] we consider the idea of a PBMS as the base for modeling patterns extracted from TM, NLP, CBIR processes. The advantage of PANDA is that it directly supports a unified way for the representation of patterns and as a result we can easily include new types of patterns as those derived by the various modules of MetaOn. This can be achieved through a pattern-base keeping information about extracted patterns. Such a pattern-base introduced in [17] consists of three basic types: the pattern, the pattern type and the class. In particular:

A pattern type is a description of the pattern structure. It consists of the pattern type name, structure, source, measure and formula. The "structure" element is the structure schema that describes the structure of the pattern type (for example, in association rule, the structure would consist of head and body, in patterns produced by CBIR, the structure would be vector signatures of images, while in patterns extracted from NLP, the structure would be a list keywords); "source" is the source schema that describes the dataset where patterns of this pattern type are constructed from; "measure" is the measure schema that defines the quality of the source data representation achieved by patterns of this pattern type (e.g. number of occurrences of a keyword); and "formula" is the formula that describes the relationship between the source space and the pattern space. A pattern is an instance of the corresponding pattern type and class is a collection of semantically related patterns of the same pattern type.

We further consider extensions of the previous model so as to meet the special requirements coming from text and image collections. We aim at an integrated representation of such disparate types of patterns. We pay particular attention to the efficient retrieval of documents and images as a result of combined information coming from both respective types of patterns. Finally, we amend the PANDA framework [17] for the comparison of simple and complex patterns so as to support patterns exported from images and documents. In PANDA the similarity between two simple patterns p_1, p_2 of the same type can be computed by combining, by means of an aggregation function faggr, the similarity between both the structure s and the measure m components:

$$sim(p_1, p_2) = f_{aggr}(sim_{struct}(p_1.s, p_2.s), sim_{meas}(p_1.m, p_2.m))$$

Across the comparison between patterns of the same type, we investigate the assessment of patterns of diverse type (e.g. documents with images).

4. Conclusions

In this paper the MetaOn, a multimedia querying and analysis system, was introduced. The system allows various data formats such as text and image to be integrated with the domain knowledge in order to facilitate search and analysis. We assist the process of the ontology construction with data mining and linguistic processing techniques. Additionally a semantics-sensitive approach to content-based image indexing and retrieval is considered. The whole process is supported by an effective metadata management scheme that integrates the various types of extracted patterns under a unified framework.

5. References

[1] Merrill Lynch: e-Business Analytics. *In-depth Report*, November, 2000.

[2] L. Gilardoni, P. Prunotto, and G. Rocca, "Hierarchical Pattern Matching for Knowledge Based News Categorization", in *Proc. RIAO 94, Intelligent Multimedia Information Retrieval Systems and Management*, New York, October, 1994, pp. 67-82.

[3] I. Muslea, "Extraction Patterns for Information Extraction Tasks: A Survey", in *Proc. Workshop on Machine Learning for Information Extraction*, Orlando, FL: July 19, 1999.

[4] H. Karanikas, and T. Mavroudakis, "Text Mining Software Survey", in *Proc. RANLP 2005-Text Mining Research Practice and Opportunities*, Borovets – Bulgaria, Sept 2005.

[5] F. Ciravegna, et al, "FACILE: Classifying Texts with pattern matching and Information Extraction", in *Proc. 16th International Joint Conference on Artificial Intelligence*, Stockholm, Sweden, 1999, 890-895.

[6] B. Catania, and A. Maddalena. *Pattern management: Practice and challenges*, Idea Group Publishing, 2005.

[7] K. Vu, K.A. Hua, W. Tavanapong, "Image Retrieval Based on Regions of Interest", *IEEE Trans. Knowledge and Data Enginering*, vol. 15, no. 4, 2003, pp. 1045-1049.

[8] J.Z. Wang, J. Li, and G. Wiederhold, "SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture Libraries," *IEEE Trans. Pattern Analysis and Machine Intelligence* vol. 23, no. 9, 2001, pp. 947-963.

[9] R. Krishnapuram, S. Medasani, S.-H. Jung, and Y.-S. Choi, "Content-Based Image Retrieval Based on a Fuzzy Approach", *IEEE Trans. Knowledge and Data Engineering*, vol. 16, no. 10, 2004, pp. 1185-1199.

[10] K. Barnard, P. Duygulu, and D. A. Forsyth, "Clustering Art", in Proc. *IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2, 2001, pp. 434-441.

[11] National Science Foundation: Global Memory Net, International Digital Library Project, http://www.memorynet.org/, accessed Feb. 2006.

[12] G. Harit, S. Chaudhury, and H. Ghosh, "Managing Document Images in a Digital Library: An Ontology Guided Approach", in Proc. *1st Int. Workshop on Document Image Analysis for Libraries*, 2004, pp. 64.

[13] Predictive Model Markup Language (PMML) http://www.dmg.org/pmml-v3-0.html

[14] Common Warehouse Metamodel (CWM) http://www.omg.org/cwm

[15] V. Ganti, J. Gehrke, and R. Ramakrishnan, "A framework for Measuring Changes in Data Characteristics", in *Proc. 18th ACM SIGMOD-SIGACT*, New York, USA, 1999, pp. 126-137.

[16] S. Baron, M. Spiliopoulou, and O. Gunther, "Efficient Monitoring of Patterns in Data Mining Environments", in *Proc. ADBIS*, 2003, pp. 253-265.

[17] I. Bartolini, P. Ciaccia, I. Ntoutsi, M. Patella, and Y. Theodoridis, "A Unified and Flexible Framework for Comparing Simple and Complex Patterns", in *Proc. 8th European Conference on Principles and Practice of Knowledge Discovery in Database*,Pisa, Italy, September, 2004.