I T H E A

International Journal

INFORMATION THEORIES & APPLICATIONS

2004 Volume 11 Number 2

International Journal
INFORMATION THEORIES & APPLICATIONS

ISSN 1310-0513

Volume 11 / 2004, Number 2

IJ ITA is official publisher of the scientific papers of the members of the Association of Developers and Users of Intellectualized Systems (ADUIS).

IJ ITA welcomes scientific papers connected with any information theory or its application.
Original and non-standard ideas will be published with preferences.

IJ ITA rules for preparing the manuscripts are compulsory.
The **rules for the papers** for IJ ITA as well as the **subscription fees** are given on *www.foibg.com/ijita*.
The camera-ready copy of the paper should be received by e-mail: foi@nlcv.net.
Responsibility for papers published in IJ ITA belongs to authors.

General Sponsor of IJ ITA is the **Consortium FOI Bulgaria** (www.foibg.com).

International Journal "INFORMATION THEORIES & APPLICATIONS" Vol.11, Number 2, 2004
Printed in Bulgaria
Edited by the **Institute of Information Theories and Applications FOI ITHEA**, Bulgaria,
in collaboration with the V.M.Glushkov Institute of Cybernetics of NAS, Ukraine, and
the Institute of Mathematics and Informatics, BAS, Bulgaria.
Publisher: FOI-COMMERCE - Sofia, 1000, P.O.B. 775, Bulgaria. www.foibg.com, e-mail: foi@nlcv.net
® "Information Theories and Applications" is a trademark of Krassimir Markov

ISSN 1310-0513

ASPICO: ADVANCED SCIENTIFIC PORTAL FOR INTERNATIONAL COOPERATION ON DIGITAL CULTURAL CONTENT

Jérôme Godard*, **Frédéric Andrès, and Kinji Ono**

Abstract: In this paper, we present the development of an advanced open source multi-lingual cooperative portal system (ASPICO) dedicated to semantic management, and to cooperative exchange for research and education purpose on digital cultural projects. Advantages of using ASPICO include greater flexibility for digital resource management, generic and systematic ontology-based metadata management, and better semantic access and delivery based on an innovative Information Modeling for Adaptive Management (IMAM).

Keywords: Digital Silk Roads, Semantic Management, Metadata Annotation, Image-Learning Ontology.

1. Introduction

Following the evolution of cultural heritage archives, new requirements for semantic understanding in a multi-lingual and multi-disciplinary cultural fields such as the historical silk roads have been pointed out in major symposiums [Ono 2001, Ono 2003] related to this field. The *Advanced Scientific Portal for International COoperation* (ASPICO) aims at providing a web portal service in order to enable international and multi-disciplinary researchers and fellows to cooperate on research about cultural projects (e.g. the historical Silk Roads project, the visual cultural topic maps online project). The platform is a java-based open source and available for everyone for research and education purposes. The platform provides multilingual semantic extraction services in order to process digital cultural artefacts from a cross disciplinary point of view, based on cooperative annotation support, metadata extraction and classification. It is based on powerful and industry-leading opensource components such as Linux OS, Enhydra, PostgreSQL, and Dspace. It is an independent platform, allowing internal usage (e.g. intranet), external usage (e.g. extranet) and access for general public via the internet. It provides flexible features, allowing easy customization for individual or community needs. Also it provides in a transparent way a fully multilingual support. So searches on the data can be performed on all the information in any language. The platform is standards compliant based on the usage of XML which allows complex data interaction and analysis to take place both within the server and on the client side. The platform has a distributed architecture so autonomous systems can be located in different institutions in order to be consulted simultaneously and to aggregate final results. In Section 2, we introduce the platform architecture. Then we present the different layers from data collection to semantic management and delivery in Section 3. Section 4 describes a function of ASPICO as a case study, the image learning ontology system including its image content recognition features. We review also the state of the art in the field of active contour as a promising solution for shaping understanding implementation inside the image learning ontology. Finally, Section 4 concludes and gives the direction of the future work.

2. The Platform Architecture

The platform architecture (see Fig. 1) is based on open source components including a storage layer (Dspace[1]), an ontology based metadata management, the query interface and resource entry service, and multi-resolution resource viewing. The system limits the access to data according to users rights to indoor users (Intranet), outdoor users (Extranet) and to the Web users.

* This research is partially supported by a grant (*bourse Lavoisier*) from the French Ministry of Foreign Affairs (Ministère des *Affaires Etrangères*).

[1] MIT's Dspace: http://www.dspace.org

2.1 Ontology-based Metadata Management

A key feature of our system is the multilingual ontology-based metadata support. Our platform follows a promising research approach based on the usage of metadata and ontologies. Metadata is any information which characterizes instance data, and which describes its relationship. Metadata is used to provide an effective use of data, in order to facilitate any data management, any data access, and data analysis [Duval 2002].

An ontology is an explicit specification of the conceptualisation of a domain [Gruber1993]. Ontologies enable domain experts to create an agreed-upon vocabulary and semantic structure for exchanging information about that domain. Ontologies facilitate cataloguing and sharing knowledge, as domain expert are able to contribute to a shared, worldwide, but well-organized knowledge base of technical information. We considered a metadata management architecture and designed multi-layer ontologies to classify and describe resources. It is based on protégé 2000[1]. Each ontology is related to one field such as history, geography, architecture, and art… However, possibility of overlapping exists as different ontologies may have equivalent concepts, or may contain subsets of separate ontologies within themselves.

This problem of ontology integration has been solved by classifying and reorganising ontologies in a logical and semantic sense according to metadata. This points to a need for a formal model for ontology-based metadata management. Ontology is the formal and explicit conceptualization of a particular domain. It includes a set of concepts and their relationships. Based on Protégé 2000, we defined our ontology structure as a 6-tuple:

$$O := \{C,P,A,H^C,prop,att\}$$

where C represents a domain-based set of concepts, P a set of relation identifiers, and A a *set of attribute-value relations. H^C is* a *Hierarchy of Concepts* that are linked together through relations (e.g. specialization, generalization). H^c is a directed transitive relation $H^c \subseteq CxC$ called *concept taxonomy*; function prop: P->CxC relates two concepts non-taxonomically; function att: A->C introduces the relationship between concepts and literal values. As an exemple, let us consider a subset of our ontology structure related to spirituality:

C := {SPIRITUALITY,RELIGION,LANGUAGE,OBJECT,LANGUE,BOUDDHISME}, P := {EXPRESS,CREATE} , and A defines the relations EXPRESS(RELIGION,LANGUAGE) and CREATE(RELIGION,OBJECT).

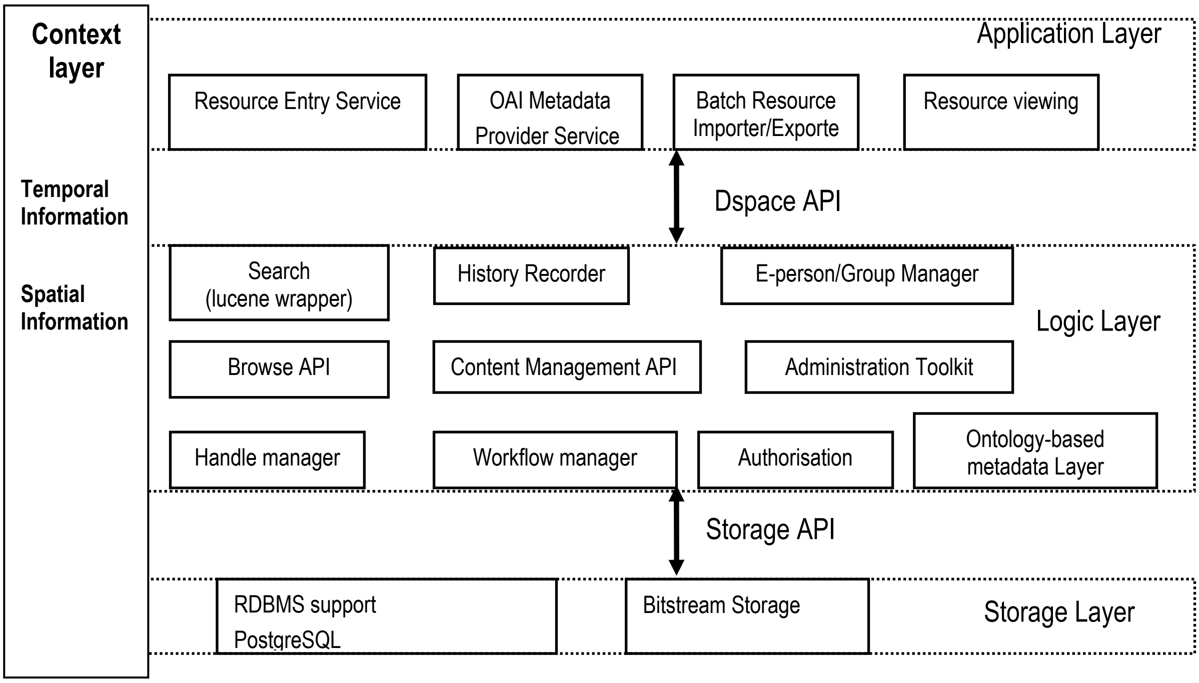


Fig. 1: Architecture of the ASPICO platform

[1] Stanford's Protégé ontology editor and knowledge-base editor: http://protege.stanford.edu/

2.2 Resource Entry

Resource entry is also performed via a web browser interface similar to that used for querying. Users who enter Resource need to log on. Write, modification and suppression rights can be assigned and controlled by the system administrator for each user; some predefined types of user provide community and group management abilities. Information such as the identification profile of the user and date of the entry are automatically filled in by the system. To maintain the integrity of the resource being entered into the system, the controlled lists of relevant vocabulary within the thesaurus are used for each translatable field. When uploading resource via the web interface, users are required to enter some preliminary metadata related to the resource. When a resource is saved into the main database, the metadata is translated into a language independent code representation. The creation of metadata profile is done according to the metadata category such as structural metadata, content metadata and contextual metadata avoiding overlapping between attribute sets.

2.3 Query Interface

Queries are performed via a web browser based interface. Screens for simple or advanced queries can be easily created and the fields to be viewed customized by the system administrator. In addition, date or numeric size fields can be searched by specifying a range of dates or sizes between which searches are performed. Users are able to select the working language and the domain of interest as well as the number of results returned and whether resource results are shown.

The interface is divided in three parts:

1- Historical and material resources related to artifacts;

2- Technical or management information related to photographic resources;

3- Technical or management information related to document resources;

Where applicable, the user can choose technical terms from a list of relevant terms classified alphabetically, or can type something directly in. Ontologies in 21 languages will be able to be consulted on line. Full text searches can be made within each field.

The display or the output format of the results (e.g. HTML, XML, plain text, formatted tabular, list of images, graphical, statistical analyses etc) is independent of the storage structure in order to optimize the delivery process. It typically follows a methodology based on context-dependent cultural resource accesses [Godard 2003].

2.4 Multi-Resolution Resource Viewing

Another key component of the resource management system is the capability to remotely view multi-resolution resources including high resolution images of both 2D paintings and 3D objects. Each image resource is stored as both a JPEG thumbnail for rapid previewing and in tiled pyramidal TIFF format for high-resolution viewing. A java applet permits multi-resolution viewing in conjunction with the storage layer. This viewing system is based on the Internet Imaging Protocol. The viewer works by requesting only the tiles at the appropriate resolution required for viewing a particular part of the image. The requested tiles are then dynamically JPEG encoded by the server and sent to the applet. In this way, images of any size can be viewed quickly across the internet.

2.5 Multilingual Ontology-based Metadata

Multilingual support is becoming very critical in the cultural domain. This can be accounted by: (1) the increasing share of cultural contents accessed over internet, (2) efforts to develop standards for cultural data from diverse fields for the purpose of digital archiving and research sharing, and (3) the increase in use of tools to extract semantic from cultural digital data. Furthermore, multi-lingual Ontology-based metadata approach enables searching by semantic and by contextual content as it relies on multi-lingual annotated documents and features extraction processes. Each set of ontologies is based on an object-identifier bridge and mono-lingual Unicode (UTF-8) encoding ontologies.

Controlled lists of technical terms (e.g. Art and Architecture Thesaurus AAT, Library of Congress Authorities) from each ontology as well as the free text information fields (such as the titles) have been translated with the support of domain experts.

3. From Data Collection to Semantic Management

We describe our solution of collecting information using ontology-based metadata management applied to cultural contents. Metadata are often categorized in three sets: (1) Production-related descriptors, (2) Physical descriptors, and (3) Administrative descriptors. Let us describe the whole acquisition process from the field data collection to the semantic management.

3.1 Data Collection and Contextual Metadata Acquisition

In this section, we describe the semantic model for ontology-based metadata management used to support metadata associated to a set of data collection. The semantic model consists of: (1) the resource layer, (2) the metadata layer, (3) the mapper layer, and (4) the context layer.

A resource r is the base physical representation of data in the metadata management framework. At the resource layer, resources are stored in the form {O,P,T} where O is the object representation of the resource r, P is the parent object of O such as archive or collection, and T is the type or schema of the resource r. data resources are represented as a tree-structure in terms of hierarchical parent-child relations between resources.

As metadata is often considered as synonym of instance of ontologies, our approach follows the formal definition of metadata structure [Guarino 1998]. The metadata structure is defined as a 6-tuple {*O,I,L,inst,instr,instl*} that consists of an ontology *O*, *I* a set of instance identifiers, *L* a set of literal values, a function *inst* called concept instantiation C->2^{I}, a function *instr* defined by P->2^{IxI} called relation instantiation, and a function *instl*, called attribute instantiation defined as P->2^{IxL}.

As examples from our ontology-based metadata, we provide the following subset of literal values:

{INDIA,FRENCH,MANDARA,CROWN}

inst can then be applied as follows:

inst(INDIA)=COUNTRY, *inst*(FRENCH)=LANGUAGE, *inst*(MANDARA)=OBJECT, *inst*(CROWN)=OBJECT

Furthermore, we can express relations between the instances and an attribute for the OBJECT instances. We give the following examples:

CREATE(BOUDDHISME,MANDARA), CREATE(BOUDDHISME,CROWN)

Mappings store how resources and metadata concepts are linked as well as how metadata layer concepts are linked between different vocabularies. There are two categories of mapping links namely: (1) resources-to-concepts and (2) concept-to-concept. For the resources-to-concepts mapping, we define MappingR2C={type, ER/AR,EC/AC} where type defines the typing of mapping, ER is the set of equivalent resources, AR is a set of resources to be aggregated. EC is the set of equivalent concepts and AR is the set of concepts to be aggregated.

The context layer stores the information related to the environment of the end-users and also the rules to determine if a mapping is applicable. A context rule is defined formally as CR={E,C,A} where E is the set of events, C is the set of conditions, and A is the set of actions to be performed if the conditions evaluate to true.

Examples of contextual values are the geographical location (e.g. *PARIS, GPS location*), time (e.g. *Oct 14th 1992* 9*:00pm*), and rules (e.g. *Tibet and 14th Century and Opaque watercolors).*

Each resource in our archive has a relationship with a *Resource Categorization Tree* (RCT) where each node label has been defined from a specific classification standard (e.g. ATT for the English culture). Each monolingual RCT is fully dependent on the cultural classification of the related language. In addition, the RCT provides a metadata form describing each resource. For each entity instance, the attributes lists and values (denoted *Resource Description* and *profile* for the other three entities) are then manipulated by database-like operators (creation, index, comparison...) [Godard 2004a], and provide the foundations of an Information Modeling for Adaptive Management (IMAM).

3.2 Resource Provider and Metadata Acquisition

Expert and researchers play a role of resource provider. Their role is to register their datasets using a metadata form. They only input from the research field mainly administrative descriptors (e.g. location address, administrative description) part of the metadata form. Also they can annotate resources according to their point of views and they can share their comments according to cross-disciplinary and multi cultural backgrounds. This registration process generates an update version of the metadata form. Each resource and its related metadata form are the knowledge base of the digital archive.

3.3 Semantic Management

Semantic Extraction

The semantic extraction (shown in Fig. 2) relies heavily on automatic process based on ontologies that provide shared conceptualizations of specific domains and on metadata defined according these ontologies enabling comprehensive and transportable machine understanding. The global process is based on four steps: (1) OCR batch recognition, (2) XML cleaning, (3) XML MPEG-7 metadata descriptors generation and (4) Topic Map Generation.

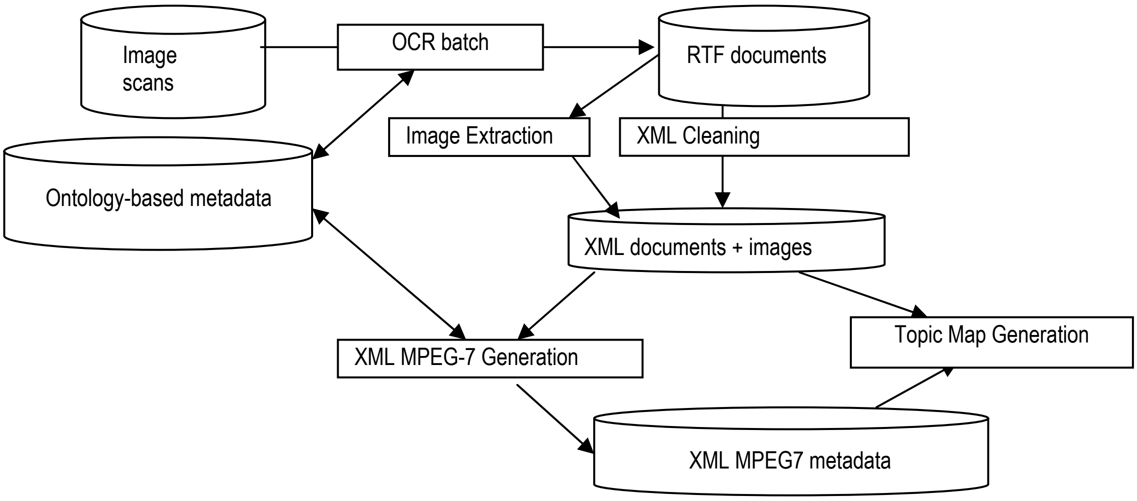


Fig 2: Semantic Extraction Example from Visual and Textual Resources

Semantic Storage, update and versioning

Several research projects such as the arXiv e-print archive[1], the Networked Computer Science Technical Reference Library (NCSTRL)[2] or the Kepler project[3] in the field of digital libraries or digital research archives, tried to solve issues of sharing and semantic storage of research information. They generally provide a common interface to the technical report collections based on the OAI (Open Archives Initiative) infrastructure[4]. This mechanism enables interoperability among large scale distributed digital archives. In many cases, the network environment services include automated registration service, tracking of connected clients, and harvesting service of clients' metadata. Query service enables accesses to resources and to its related metadata.

The Open Archives Initiative (OAI) has created a protocol (Open Archives Initiative Protocol for Metadata Harvesting, OAI-PMH) based on the standard technologies HTTP and XML as well as the Dublin Core metadata scheme[5]. OAI presently supports the multipurpose resource description standard Dublin Core which is simple to use and versatile.

Shortcomings of such research projects generally include a too general metadata attributes schema for fine-grained information (e.g. cultural domains) and the non-support of community building so the semantic management and storage are weak points. However, OAI-PMH itself has been created to provide an XML-wrapper. As part of our project, it has been extended in our project to support multi-disciplinary metadata schemas such as Object ID for historical buildings, CDWA[6] for historical artefacts, or VRA[7] for Visual Resources so the semantic storage has been enhanced. The platform provides service for data handling, registration for

[1] Cornell's arXiv.org e-Print archive: http://arxiv.org

[2] Networked Computer Science Technical Reference Library: http://www.ncstrl.org/

[3] Kepler – A Digital Library For Building Communities: http://kepler.cs.odu.edu/

[4] Open Archives Initiative: http://www.openarchives.org

[5] Dublin Core Metadata Initiative: http://dublincore.org/

[6] Categories for the Description of Works of Art: http://www.getty.edu/research/conducting_research/standards/cdwa/

[7] The Visual Resources Association: http://www.vraweb.org/

identification; and semantic handling based on cross-disciplinary metadata schemas to create OAI-compliant metadata and semantic management. All manipulated and exchanged information are in XML. As it has been pointed out in [Westermann 2003], XML database solution is the most appropriate approach for semantic storage and management according to the current state of the art but still issues still remain such as versioning and update propagation. A native semantic database engine will be part of the evolution of the platform.

Semantic Access and Delivery

Let us present how this architecture takes advantage of the semantics in order to provide adaptive processes for data access and delivery. In the case of collaborative environments, we are convinced that the available and well-structured knowledge, which is available through the semantic management, can allow us to provide relevant personalized services. Service-oriented architectures definitely require strong knowledge management frameworks in order to perform automated adaptive tasks. But fully automated service composition still remains illusive [McDermott 2003]. However, by operating within known and defined domains, it is nowadays possible to prepare mappings or wrappers in advance so the automated processes can be efficient. The semantic management layer described previously is the basis to enable ASPICO delivering powerful services dedicated to users and to improve automated data management for communities. Using ontology-based metadata management, ASPICO enhances the ability of personalization by dealing with multi-domain semantic control and a strong algebraic metadata model. We defined a set of four entities (resources, users, communities, devices), which allows us to structure all the available knowledge that can be used in order to process adaptive services. IMAM Services aim at offering personalized data access and delivery for users who are involved in communities related to multi-domain interests. We introduced two main services [Godard 2004b]:

- *Viewpoint*, which is a query optimizer. It selects relevant resources from the query answer depending on environment profile (combination of user and device profiles).
- *Placement*, which dispatches, in an authoritarian way, new resources on devices depending on the correlations between the resource and community or user, and on devices capacities.

In the next section, an image-learning ontology is described as a typical application for our semantic management framework.

4. Case Study: The Image-learning Ontology System

Content-based image retrieval is a challenging and active research area [Jain 1998] with the potential to provide powerful tools for image searching and semantic understanding. Although many techniques have been described in the research literature, the capabilities of current content matching systems are still basic general purpose approaches. Our research focuses on the combination of ontology-based metadata and image content understanding, calling this approach *Image-learning Ontology.* We have been applying this innovative approach to the architecture domain. As it is shown in Fig 3, the image-learning ontology system is based on two phases:

- a training phase where the system learns semantic links between technical terms and image objects.
- a semantic discovery phase where the system proposes to the experts new semantic links between technical terms and image objects.

Large set of digital image archives on architecture at NII allows us to sample and group together images with similar characteristics and categories thereby providing the reference material for testing the behavior of image content recognition. Furthermore cooperation with domain experts provides a multi-lingual support on these categories.

In the remainder of the paper, we detail the image content recognition processing according to semi-automatic and automatic properties. General techniques based on such features as color distribution, texture, outline shape, and spatial color distribution have been quite popular in the research literature and in content based retrieval systems.

4.1 Semi-Automatic Content Recognition of Images

Regarding this aspect, requirements have been determined according to the resource type as part of the resource metadata. If the resource is a document, the recognition process is done for each image included inside the document. Iconography characterization is one the most complex issue in this field so semi-automatic content recognition process based a learning phase and sketching is the most suitable method. Some shape recognition methods work well such as portrait, landscape, buildings, or themes such as crucifixion, or virgin and child. The experts confirm through sketching [Sciascio 1999] the contents of the images.

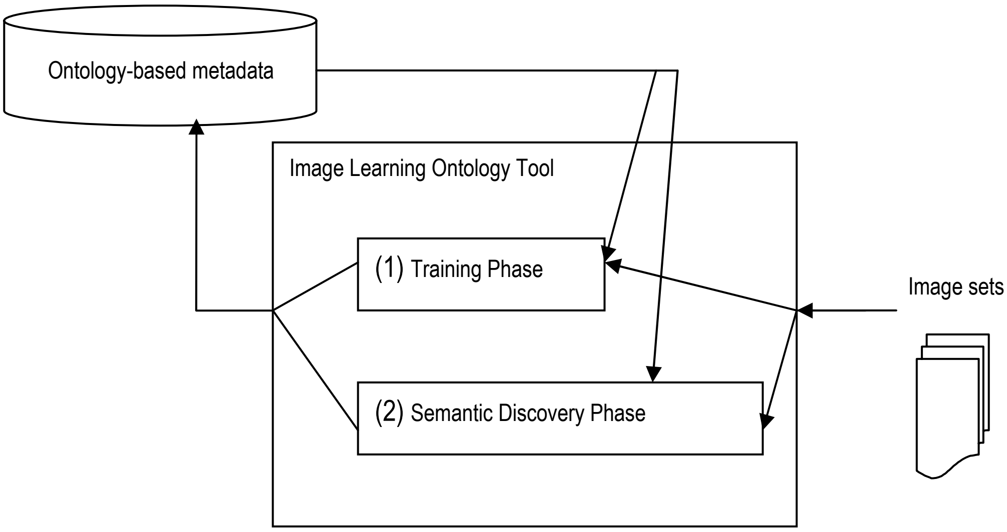


Fig 3: Architecture of the Image-learning Ontology System

4.2 Automatic Shape Identification

One user-requirement for the DSR project is the identification of shapes for painting and buildings in order to provide richer statistics for searches. This is useful for restricting areas of interest and avoiding backgrounds. This is carried out using recognition of deformable models. Research activities concerning deformable models can be partitioned in two types in [Jain 1998]:

- Free-form model, also called active contours, which allows representing any shape by using a minimizing energy algorithm.
- Parametric model allows defining and encoding specific geometric properties of the shape (moments, angles…)

Let us review the two classes of model.

Free-form model

An initial contour, or snake, C is defined by the coordinates $\{x(s), y(s)\}$, 0<s<1. The method was initially introduced by [Kass 1988] and involves the energy-minimization contour C by controlling the three forces:

- The internal forces E_{int}, which define the constraints concerning the shape of the model (more or less smooth).
- The images forces E_{image} which distort the contour according to the variations of pixels values (grey level or colour values).
- The external forces E_{con}.

The willed contour is thus obtained by minimizing the energy given by:

$$E_{snake} = \int_0^1 \left(E_{\text{int}}\left(x(s), y(s)\right) \quad + \quad E_{image}\left(x(s), y(s)\right) + E_{con}\left(x(s), y(s)\right)\right) ds$$

The external forces E_{con} will be not used in what follows.

The internal forces are mainly defined by the coordinates of the snake C:

$$E_{\text{int}}(s) = \tfrac{1}{2}\left(\alpha(s)\left(x_s(s)^2 + y_s(s)^2\right) \quad + \quad \beta(s)\left(x_{ss}(s)^2 + y_{ss}(s)^2\right)\right)$$

Where the subscripts on x and y define derivative form. The coefficients α and β indicate the strength of the elasticity and of the rigidity. In practice, for the digital images applications the problem must be discretized [Davison 2000]. Energies must be sampled at N equally spaced knots v_i around the edge C:

$$E_{\text{int}} = \frac{1}{2h}\sum_{i=0}^{N-1}\alpha_i |v_i - v_{i-1}|^2 + \frac{1}{2h^3}\sum_{i=0}^{N-1}\beta_i |v_{i-1} - 2v_i + v_{i+1}|^2$$

In general, the first curve is initialized by B-splines, widely described by [Blake 1998] and used, for instance, by [Stammberger 1999] for a magnetic resonance imaging application.

The image energy E_I depends on the variations of grey level $g(x_i, y_i)$.

$$E_I = k_I \sum_{i=1}^{N} g(x_i, y_i)$$

$$E_{image} = -\nabla E_I$$

The energy-minimization is usually realized iteratively by a gradient descent algorithm until a minimum. Intuitively, a major drawback appears. Indeed, the contour C which depends on the initial position can be attracted by a local minimum, far-off the shape desired. A control of the final contour must be thus checked. Many approaches have been proposed to erase these problems in [Jain 1998], and in [Tsechpenakis 2004]. Moreover the method fails sometimes for very complex shapes. Nevertheless, the method remains very powerful for image segmentation and its implementation is very fast.

Lastly the use of colour information allows improving the performance of active contours. [Ngoi 1999] proposed thus a new active contour model for shape extraction of images acquired in outdoor conditions.

Parametric models

The active contours are based on an energy minimizing calculated from the coordinates of the pixels belonging to the contour; the basis of the parametric models is the study of the shape deformation by using geometrical parameters. The model needs now more specific a-priori knowledge of the shape. We can differentiate two parametric models [Jain 1998]:

- Analytical deformable models which are defined by analytical curves.
- Prototype-based deformable templates defined an "average" shape of a class of objects.

Let us comment these two models:

- Analytical deformable templates

In those methods, templates are defined by parametric models such as ellipses, or circle parametric function. The model, which possesses only few degrees of freedom, fit the desired shape by energy minimizing applied to the model parameters. The most popular example of such a method is the eye template of [Yuille 1992]. In this model, parameters for which the variations are carried out are the centre and the radius of the circle and the coefficients of the parabola. In this model, we distinguish also two kinds of energy, the internal energy which is defined by a parametric function characterizing a shape and the external energy which represents the features of the image. Minimizing energy algorithm is then used. Because of the parametric function is chosen previously, the analytical deformable templates are required for segment objects with a known shape.

- Prototype-based deformable templates

For the prototype-based deformable templates, a particular model is previously built according to the shape we want to extract (for example a model of a sculpture or a model of a building). The performance of the prototype template depends, obviously, on the description of the shape. Recent research works have adopted learning method from a set of samples. So as to do it, [Cootes 1994] and [Cootes 1995] have thus used this kind of methods. From those samples templates, a mean shape is calculated and used as the generic model and the variations are determined by eigenvectors of the covariance matrix. Other deformable templates based on a prototype have been also described in [Jain 1998].

Digital Silk Roads archives contain high resolution colour images acquired in outdoor (high luminosity, reflections, shadows). If the natural conditions of imaging and the variability of the conditions complicate the segmentation, the images to be segmented present objects with particularly simple shapes (doors, roofs, mosaics, arch…), making the use of deformable models easier. The main advantages of active contours are the speed and the flexibility. Specific a-priori knowledge is not required. On the other hand, it seems that the deformable templates are very adapted for locating specific structures in the images but this method needs a more specific knowledge about the shape we want to extract. Another difficulty is to define accurately objects. The segmentation tool must be precise for segment small objects such as frieze; tiles, lock and writings without over segment the images by defining objects without any architecture signification. A quite “supervised” process is thus preferable.

5. Conclusion

This paper introduced the ASPICO platform as a semantic advanced archive management system. It described the knowledge structure and the interface based on multilingual ontology-based metadata management. Introducing the image-learning ontology system, we investigated the possible solutions for automatic shape identification and then motivated the appropriate strategy to be adopted by cultural digital resource archive in order to perform indexing and efficient retrieval on digital cultural content through ASPICO. In a next step, we are planning to add innovative functions related to real time semantic acquisition based on remote control data acquisition.

Bibliography

[Blake 1999] Blake A. and Isard M. (1999). *Active contours*. Springer. 352 p.

[Cootes 1994] Cootes T.F., Hill A., Taylor C.J., Haslam J. (1994). *Use of active shape models for locating structures in medical.* In: Image Vision Computing 12(6), p.355-366.

[Cootes 1995] Cootes T.F., Taylor C.J., Cooper D.H., Graham J. (1995). *Active shape model – their training and application.* In: Computers Vision Image understanding 61(1). p.38-59.

[Davison 2000] Davison N.E., Eviatar H., Somorjai R.L. (2000). *Snakes simplified.* In: Pattern Recognition 33. p.1651-1664.

[Duval 2002] Duval E., Hodgins W., Sutton S., Weibel S. (2002). *Metadata Principles and Practicalities.* D-Lib Magazine April 2002 Volume 8 Number 4 ISSN 1082-9873.

[Godard 2003] Godard J., Andres F., Grosky W., Ono K. (2003). *Management of Geomedia Content: Context-dependent Data Access*. NII Journal No. 7, p.9-17.

[Godard 2004a] Godard J., Andrès F., Grosky W. (2004). *Knowledge Management Framework for the Collaborative Distribution of Information*. DataX (EDBT Workshop), Heraklion, Crete, Greece, p.2-16.

[Godard 2004b] Godard J., Andrès F., Andaroodi E., Maruyama K. (2004). *Towards a Service-oriented Architecture for Collaborative Management of Heterogeneous Cultural Resources*, Digital Library Architectures, Sixth Thematic Workshop of the DELOS EU Network of Excellence, S. Margherita di Pula, Cagliari, Italy, p.183-194.

[Guarino 1998] Guarino N. (1998). *Formal ontology and information systems*. In N. Guarino, Ed., Proceedings of FOIS `98, Trento, Italy, June. IOS Press, Amsterdam, 1998, p.3-15.

[Gruber1993] Gruber T.R. (1993). *Towards Principles for the Design of Ontologies Used for Knowledge Sharing.* In Roberto Poli Nicola Guarino, editor, International Workshop on Formal Ontology, Padova, Italy.

[Kass 1988] Kass M., Witkin A., Terzopoulos D. (1988). *Snakes: active contours model*. In: International Journal of computers vision 1 (4), p.321-331.

[Jain 1988] Jain A.K., Zhong Y., Dubuisson-Jolly M-P. (1998). *Deformable template models: A review*. In: Signal Processing 71, p.109-129.

[McDermott 2003] McDermott D. V. (2003). *PDDL2.1 - The Art of the Possible? Commentary on Fox and Long*. J. Artif. Intell. Res. (JAIR) 20: p.145-148.

[Ngoi 1999] Ngoi K.P., Jia J.C. (1999). *An active contour model for color region extraction in natural scenes*. In: Image and vision computing 17, pp.955-966.

[Ono 2002] Ono K. (2002). Proceedings of the Tokyo Symposium for Digital Silk Roads, National Institute of Informatics, Tokyo, Japan, ISBN 4-86049-007-X, 321p.

[Ono 2004] Ono K. (2004). Proceedings of the Nara Symposium for Digital Silk Roads, Nara, Japan, ISBN 4-86049-024-X, 510p.

[Sciascio 1999] Sciascio E. D., Mongiello. M. (2003). *Query by sketch and relevance feedback for contentbased image retrieval over the web*. Journal of Visual Languages and Computing, 10(6) p.565-584.

[Stammberger 1999] Stammberger T., Eckstein F., Michaelis M., Englmeier K-H., Reiser M. (1999). *Interobserver reproducibility of quantitative cartilage measurements: comparison of B-spline snake and manual segmentation.* In: Magnetic Resonance Imaging, 17(7). p.1033-1042.

[Tsechpenakis 2004] Tsechpenakis G., Rapantzikos K., Tsapatsoulis N., Kollias S. (2004). *A snake model for object tracking in natural sequences.* In: Signal processing: Image communication 19, p.219-238.

[Westermann 2003] Westermann U., Klas W. (2003). *An analysis of XML database solutions for the management of MPEG-7 media descriptions*. ACM Computing Surveys, Volume 35 (4), p.331-373.

[Yuille 1992] Yuille A., Hallinan P., Cohen D. (1992). *Feature extraction from faces using deformable templates.* In: International Journal of Computer Vision, 8(2), p.99-111.

Authors' Information

Jérôme Godard – National Institute of Informatics (NII), Graduate University of Advanced Studies (Sokendai)
Hitotsubashi 2-1-2 Chiyoda-ku, Tokyo 101-8430, Japan; e-mail: jerome@grad.nii.ac.jp

Frédéric Andrès – National Institute of Informatics (NII), Graduate University of Advanced Studies (Sokendai)
Hitotsubashi 2-1-2 Chiyoda-ku, Tokyo 101-8430, Japan; e-mail: andres@nii.ac.jp

Kinji Ono – National Institute of Informatics (NII), Waseda University
Hitotsubashi 2-1-2 Chiyoda-ku, Tokyo 101-8430, Japan; e-mail: ono@nii.ac.jp

MPEG-7 BASED IMAGE RETRIEVAL ON THE WORLD WIDE WEB

Rajeev Agrawal, Farshad Fotouhi, Peter Stanchev, and Ming Dong

Abstract: Due to the rapid growth of the number of digital media elements like image, video, audio, graphics on Internet, there is an increasing demand for effective search and retrieval techniques. Recently, many search engines have made image search as an option like Google, AlltheWeb, AltaVista, Freenet. In addition to this, Ditto, Picsearch, can search only the images on Internet. There are also other domain specific search engines available for graphics and clip art, audio, video, educational images, artwork, stock photos, science and nature [www.faganfinder.com/img]. These entire search engines are directory based. They crawls the entire Internet and index all the images in certain categories. They do not display the images in any particular order with respect to the time and context. With the availability of MPEG-7, a standard for describing multimedia content, it is now possible to store the images with its metadata in a structured format. This helps in searching and retrieving the images. The MPEG-7 standard uses XML to describe the content of multimedia information objects. These objects will have metadata information in the form of MPEG-7 or any other similar format associated with them. It can be used in different ways to search the objects. In this paper we propose a system, which can do content based image retrieval on the World Wide Web. It displays the result in user-defined order.

Keywords: XML, MPEG-7, Metadata, Multimedia, Content Based Image Retrieval (CBIR)

1. Introduction

The CBIR has been a very active research area in the last decade. Conventional content-based image retrieval systems [1, 2, 3] use low-level features such as color, texture, shape, automatically extracted from the images. Another focus of this research is on improving the low level features. The modifying the similarity measures make the retrieval as better as possible. It is argued in [4] that unconstrained object recognition is still beyond of current technology. The content based systems can at best capture only pre-attentive similarity, not semantic similarity. So far there has not been a single system, which can perform this task automatically without human intervention due to the nature of this problem.

The expansion of the World Wide Web (WWW) is making the problem of effective retrieval of images very important for all its users. The complexity of Web documents is rapidly increasing with the wide use of multimedia components, such as images, audio and video, associated to the traditional textual content. This requires extended capabilities of the Web query search engines in order to access images according to their multimedia content. A large number of search engines (e.g. Altavista, Yahoo, HotBot, etc.) support indexing and content-based retrieval of Web documents. Only the textual information is taken into account. Initial experimental systems providing support to the retrieval of Web documents based on their multimedia content (Webseek [5] and Amore [6]) are limited to the use of pure physical features extracted from multimedia data, such as color, shape, texture. These systems do not go beyond the use of pure physical visual properties of the images. They suffer the same severe limitations of today as the general-purpose image retrieval systems [7], such as Virage [8] and QBIC [9]. These systems consider images as independent objects, without any semantic organization in the database or any semantic inter-relationships between database objects. Many image searches also use an approach that filters out less relevant results. They analyze and index the text on the page adjacent to the image, the image link text, text in the HTML alt tag, filename or file path name. Similarly, this approach can also be used with other media files such as audio and video. Even though these search engines do not “look inside” the media files, they can give quite relevant results.

Another approach can be to look into the media file contents itself and trying to mine for textual information in the file for better multimedia indexing. For example, a Portable Network Graphics (PNG) image file can contain textual information such as title, author, description, copyright, creation time, software used, disclaimer, warning, source and comments [10]. Not all file formats contain metadata, and even if they do, an indexing engine should know how to handle all the different file formats and where to find that information in a file. It would be better if we had a data model which could be used with different media formats and utilized a rich set of metadata. There have been many metadata models developed. Some of them are RLG Preservation Metadata Elements, NISO Draft Standard, DIG35 Specification, Data Dictionary for Audio/Video Metadata, Metadata for Long-Term Preservation, Metadata Encoding and Transmission Standard [11]. MPEG-7 is another multimedia metadata standard. The Moving Picture Experts Group (MPEG) was established in 1988 to develop audiovisual compression standards. MPEG-1, MPEG-2, and MPEG-4 all represent the content itself, while MPEG-7 represents information about the content [12]. While the first produces the contents, the latter describes the content. There are number of tools provided in MPEG-7 - descriptors (the elements), description schemes (the structures), a Description Definition Language (DDL) (for extending the predefined set of tools) and a number of system tools. MPEG-7 can support all natural languages. DDL provides the foundation for the standard. It provides the language for defining the structure and content of MPEG-7 documents. The DDL is not a modeling language such as Unified Modeling Language (UML) but a schema language to represent the results of modeling audiovisual data (i.e. descriptors and description schemes) as a set of syntactic, structural and value constraints to which valid MPEG-7 descriptors, description schemes, and descriptions must conform. The purpose of a schema is to define a class of XML documents. The purpose of and MPEG-7 schema is to define a class of MPEG-7 documents. MPEG-7 instances are XML documents that conform to a particular MPEG-7 schema (expressed in the DDL) and that describe audiovisual content. MPEG7 has been developed after many rounds of careful discussion. It is expected that this standard would be used in searching and retrieving for all types of media objects. If we have images stored with MPEG-7 metadata, it would be easier to do semantic retrieval. MPEG-7 files contain a reference to the location of the corresponding image file. It is also possible to exploit other tools and technologies developed for XML like Xquery, XPath, etc. There has been a lot of work on XML schema integration. This plays a central role in numerous applications, such as web-oriented data integration, electronic commerce, schema evolution and migration, application evolution, data warehousing etc. In schema integration, the main objective is to find a suitable technique to match the elements in different schemas. We propose to combine XML schema integration techniques and image retrieval techniques using low-level features with or without semantic annotations.

Rest of the paper is organized as follows: Section 2 describes the motivating examples. Section 3 relates a list of previous work and other literature survey. Section 4 describes our proposed system. Finally, we give concluding remarks in section 5.

2. Motivating Examples

The commercial image search engines available today basically search the images based on keywords. The keywords are extracted from the web page, where image appears. But the keyword based search has its own limitation, which will be clear from the following examples.

1. If we want to search and retrieve the pictures of a person in the different stages of his/her life with respect to the time, available on different websites, that is not possible through keyword search. The keyword search would definitely retrieve the images but not integrate in the order we want. Assumption here is that different websites has the pictures of the person at different stages of his/her life and also incorporate some semantic information, which can be in MPEG-7 or in any other metadata format. The reason is keyword search just looks for the name in the surrounding text, but no in other information. E.g. when we search the pictures of a great person like Mahatma Gandhi images are retrieved, but not in any order. The main reason is that no semantic information is incorporated with the images.

2. Some security agency is interested in getting more information about a person, who has perpetrated some crime and they have a photograph of this person. There is no technique available which can return the information about this person from the Internet, if the agency uses this photograph as input (query by example method). The basic idea of this kind of search is that low level feature of the query image should be compared with all the images available on the Internet and a set of images, which are closer up to certain threshold are returned.

3. We want to search images of two cities, which belong to same country. The keyword search can include some false results. E.g. when we search for the cities Detroit and Flint together, we see some graphs, which are not the images of the cities, but refer their names in the graphics.

4. There is also no method available, which can return the result of following types of query. E.g. search the pictures about American history between the year 1900 and 1950.

There is no method of defining the queries between certain time range and/or any other metric. One of the problems of not getting the desired results is that there in no or little metadata available with the images available on the Internet. Second reason is that the algorithms employed by the search engines, does not have the capability to do search based on a specific criteria like these. As we can see in the above examples, that there is still a long way to be able to apply complex queries to search the images from the World Wide Web. In addition to above examples, we may encounter large number of other kinds of queries, which are not possible through existing search engines.

3. Literature Survey

The CBIR on World Wide Web involves two research areas: images classification and, images search and retrieval techniques.

3.1. Image Classification

In the literature, a wide variety of content-based retrieval methods and systems may be found. In [13] authors have reviewed about 200 references in CBIR up to the year 2000. There are three broad classes of applications user aims when using the system: search by association, search at a specific image, and category search. [14] identifies other patterns of use: searches for one specific image, general browsing to make an interactive choice, searches for a picture to go with a broad story, searches to illustrate a document. An attempt to formulate a general categorization of user requests for still and moving images are found in [15]. This and similar studies reveal that the range of queries is wider than just retrieving images based on the presence or absence of objects of simple visual characteristics. To describe the image, we have to extract certain low level features from it. There are a number of image processing operations that translate the image data into some other spatial data array. These operations may use local color, local texture, or local geometry. The main purpose of image processing in image retrieval must be to enhance aspects in the image data relevant to the query and to reduce the remaining aspects. There are several color representations like RGB, HSV, YUV and their variations.

Local shape characteristics derived from directional color derivatives have been used in [16] to derive perceptually conspicuous details in highly textured patches. In [17] a series of Gabor filters of different directions and scale have been used to enhance image properties [18]. Combining shape and color both in invariant fashion is a powerful combination as described by [19]. The texture is defined as all what is left after color and local shape have been considered or it is defined by such terms as structure and randomness. Basic texture properties include the Markovian analysis and other generalized versions [20, 21]. Other texture analysis methods are MRSAR-models [22], Wavelets [23], fractals [24] etc. A comparative study on texture classification from mostly transform-based properties can be found in [25].

In CBIR, the image is often divided in parts before features are computed from each part. There are four types of partitioning identified in [13]: string segmentation, weak segmentation, sign detection, data independent image partitioning. In [26] knowledge-based type abstraction hierarchies are used to access image data based on context and a user profile, generated automatically from cluster analysis of the database. Also in [27] the aim is to create a very large concept-space inspired by the thesaurus-based search from the information retrieval community. In [28] a variety of techniques is discussed treating retrieval as a classification problem. One approach is principal component analysis over a stack of images taken from the same class of objects. This can be done in feature space [29] or at the level of the entire image [30]. In [31] binary Bayesian classifiers are used to capture high-level concepts from low-level image features under the constraint that the test image belongs to one of the classes. Specifically, the hierarchical classification of vacation images is considered. At the highest level, images are classified as indoor or outdoor; outdoor images are further classified as city or landscape. Finally, a subset of landscape images is classified into sunset, forest, and mountain classes. A large number of systems have ignored two distinct characteristics of CBIR systems: the gap between high level concepts and low level features, subjectivity of human perception of visual content. A relevance feedback based approach has been suggested in [32]. Other interactive approaches have been suggested in [33, 34, 35]. Example include interactive region segmentation [36]; interactive database annotation [34, 37]; usage of supervised learning before the retrieval [38, 39]; and interactive integration of keywords and high level concepts to enhance image retrieval performance [40, 41]. In [42] an image retrieval system called SIMPLIcity (Semantics-sensitive Integrated Matching for Picture Libraries), which uses semantics classification methods, a wavelet-based approach for feature extraction. An integrated region matching based upon image segmentation, has been proposed. There are several domain-dependent ontology based systems [43, 44]. In [45] system uses a neural network to identify objects present in the images.

3.2. Image Search and Retrieval Techniques

There are a large number of papers published in the area of image search and retrieval. We are restricting our discussion here related to image search on World Wide Web. A system is implemented in [46] by which visual information on the web is (1) collected by agents, (2) processed in both text and visual feature domains, (3) catalogued and (4) indexed for fast search and retrieval. A typical web image search engine will first traverse the Web by following the hyperlinks between documents using several autonomous Web agents or spiders. These agents detect images and download and process them and add the new information about the image to the catalog.

A *perception-based search component,* which can learn users' subjective query concepts quickly through an intelligent sampling process, is proposed in [47]. A multi-resolution feature extractor extracts perceptual features from images and a high-dimensional indexer performs non-supervised clustering using Tree-structured Vector Quantization (TSVQ) [48] to group similar objects together. iFind is a web-based image retrieval system developed at Microsoft Research, China [49]. It provides the functionalities of text based image search, query by example, and their combination. Images in the database are indexed by their low-level (visual) features, high level (semantic) features (collected from image's environment), and optionally, annotations if they are available. In [50] MISE (The MediaSys Image Search Engine) is described. This system enables the users to search, to browse, to process, and to store images according to the combination of visual and textual features with meta-data related to the images. The MediaSys servers store the meta-data, visual and textual features, and the images themselves over a large scale distributed and heterogeneous system. The article [51] investigates what MPEG-7 means to Multimedia Database systems (MMDBSs) and vice-versa. It is argued that MPEG-7 has to be considered complementary to, rather than competing with, data models employed in it. [52] describes the use of stylesheets in the search and retrieval process of multimedia information, especially for audiovisual information. MPEG-7 has been used to describe the contents of the information. The use of stylesheets over the MPEG-7 data gives flexibility during both query formulation and the presentation of search results, and it allows a personalized way of querying and presenting.

4. The Proposed System

We discuses some of the example queries in section 2, which can not be answered by any of the existing systems to the best of our knowledge. We propose a system, which will exploit the XML technology and new MPEG-7 media metadata standard. In this section, we briefly describe the Image Integration Architecture.

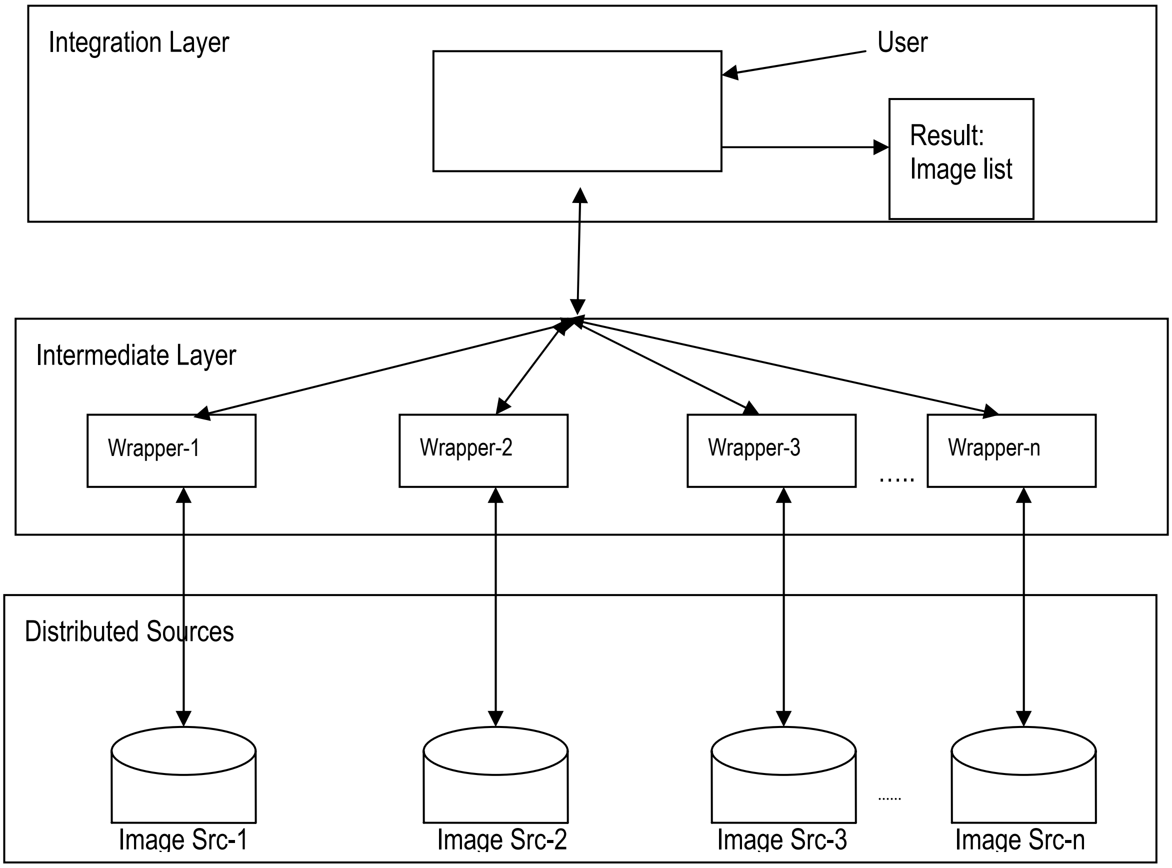


Figure 1. Image Integration Architecture

Figure 1 shows three-layered image integration architecture. At the lowest level, we have different Image sources. These sources have images and have not been designed on certain agreed schema. In other words, images in these sources may be in raw JPG, BMP, GIF or any other format without any semantic information. They may contain images clustered in certain groups. They may contain metadata in the form of MPEG-7 with partial annotation or they may contain MPEG-7 metadata with structured annotation. There may be other possibilities also.

Intermediate layer focuses on extracting image information by extracting low level features, metadata or any other semantic information available. If there is no semantic information available, we have to rely on low level features. We are considering images which are embedded in a webpage or stored in the image database. Each image source has to be treated in a different way.

Image source with raw image formats. At intermediate level, we extract low level features and store as MPEG - 7 metadata. Since this procedure has to be automatic, we can not do annotations at this level. There is no automatic annotation technique available so far.

Image source with raw image formats but clustered in groups. We extract low level features from the image and also store cluster information in MPEG - 7 metadata. Some intelligent technique has to be used to make cluster information useful in retrieval. We can also use traditional image search retrieval methods and look for important keywords stored in and around the image.

Image source with raw image format and with some metadata but not in MPEG – 7 standard. We extract low level features of the image and use the metadata while creating MPEG-7 metadata.

Image source with MPEG-7 partial or full annotation. We do not need to extract the low level features, since they are already available in MPEG-7 metadata.

Our emphasis here is to get information about all the images in MPEG-7 format, which is essentially XML data. Then we can use XML tools to query the images in Integration layer. There has been a lot of work on XML schema Integration [53, 54, 55, 56, 57, 58]. In this paper we are not discussing about XML schema integration.

The user will make a query at the top level using any of the methods using keyword, query by example, range queries discussed in section 2. This architecture may use agent based method or the popular directory based indexing method to search the image data sources. Integration process consists of querying the results returned by the intermediate layer, refine them according to user demand and return the results back to him/her. The relevance feedback and/or other long term learning technique can be used at the highest level to improve the results. The queries similar to the examples mentioned in section 2 can be successful if we combine low level features and semantic information together to produce the results. This architecture does not merely return the search results based on the keywords associated with the image, but also takes into account the low level features of the image.

5. Conclusion and Future Research

In this paper we suggest three - layered image integration architecture at a conceptual level. This approach takes care of images stored on the websites/image databases with or without semantic information. There are many challenges we have to face in this approach like selecting appropriate schema integration technique. MPEG-7, though already declared standard, will still take some time before images have their metadata stored in this format. Therefore it would be a grade mistake to rely on the assumption that metadata would be easily available in MPEG-7. Similarly, there are a large number of low level features suggested by different researchers, but MPEG-7 has included only some of them. There are possibilities that better features may be released in future and we have to consider these new features in any content-based image retrieval system. We are trying to set up an experimental environment based on the approach suggested in this paper, taking into account the methodology suggested in [59]. We are in the process of collecting the images with the properties described in section 4. We believe that the proposed system would enhance the quality of content based image retrieval.

Bibliography

[1] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by Image and video content: The QBIC system. IEEE Computer, vol. 28, no. 9, pp. 23-32, 1995.

[2] W. Y. Ma and B. S. Manjumath. Netra: A toolbox for navigating large image databases. ACM Multimedia System, vol. 7, pp. 184-198, 1999.

[3] Y. Rui, T. S. Huang. S. Mehrotra, and M. Ortega. A relevance feedback architecture for content-based multimedia information retrieval systems. IEEE Workshop on Content-based Access of Image and Video Libraries, pp. 82-89, 1997.

[4] S. Santini and R.Jain. Visual navigation in perceptual databases. International Conference on visual Information systems, San Diago, CA, Dec. 1997.

[5] J. R. Smith and S. Chang, Visually searching the Web for content, IEEE Multimedia, July-September 1997.

[6] S. Mukherjea, K. Hirata and Y. Hara. Towards a multimedia World Wide Web information retrieval engine. 6th WWW International Conference, S. Clara, CA, 6–11 May 1997.

[7] C. Meghini, f. Sebastiani and U. Straccia. Modelling the retrieval of structured documents contaning texts and images. 1st ECDL, Pis, Italy, Sep. 1997

[8] J.R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. Jain and C.F. Shu. The Virage image search engine: An open framework for image management. SPIE 96, 1996.

[9] M. Flickner et al., Query by image and video content: the QBIC system, IEEE Computer, 28(9), September 1995.

[10] PNG (portable Network Graphics) Specification, Version 1.2., http://www.libpng.org/pub/png/spec/

[11] Metadata Standards. http://www.chin.gc.ca/English/Standards/metadata_multimedia.html

[12] B S Manjunath et. El. Introduction to MPEG-7. John Wiley, 2002.

[13] Arnold W. M. Smeulders et. el. Content-Based Image Retrieval at the End of Early Years. IEEE transactions on Pattern Analysis and Machine Intelligence, vol. 22, No. 12, Dec 2000.

[14] S. Ornager. Image Retrieval: Theoretical and Empirical User Studies on Accessing Information in Images. 60th Am. Soc. Information Science Ann. Meeting, vol. 34, pp. 202-211, 1997.

[15] L. Armitage and P. Enser. Analysis of User Need in Image Archives. J. Information Science, vol. 23, no. 4, pp. 287-299, 1997.

[16] A. Mojsilovic, J. Kovacevic, J. Hu, R.J. Safranek, and S.K. Ganapathy. Matching and Retrieval Based on the Vocabulary and Grammar of Color Patterns. IEEE Trans. Image Processing, vol. 9, no. 1, pp. 38-54, 2000.

[17] B.S. Manjunath and W.Y. Ma. Texture Features for Browsing and Retrieval of Image Data. IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 18, no. 8, pp. 837-842, Aug. 1996.

[18] R. Rodriguez-Sanchez, J.A. Garcia, J. Fdez-Valdivia, and X.R. Fdez-Vidal. The RGFF Representational Model: A System for the Automatically Learned Partitioning of `Visual Pattern' in Digital Images. IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 21, no. 10, pp. 1,044-1,073, Oct. 1999.

[19] T. Gevers and A.W.M. Smeulders. Content-Based Image Retrieval by Viewpoint-Invariant Image Indexing. Image and Vision Computing, vol. 17, no. 7, pp. 475-488, 1999.

[20] S. Krishnamachari and R. Chellappa. Multiresolution Gauss-Markov Random Field Models for Texture Segmentation. IEEE Trans. Image Processing, vol. 6, no. 2, 1997.

[21] G.L. Gimel'farb and A.K. Jain. On Retrieving Textured Images from an Image Database. Pattern Recognition, vol. 29, no. 9, pp. 1,461-1,483, 1996.

[22] J. Tatemura. Browsing Images Based on Social and Content Similarity. Proc. Int'l Conf. Multimedia and Expo, 2000.

[23] I. Daubechies. Ten Lectures on Wavelets. Philadelphia: SIAM, 1992.

[24] L.M. Kaplan et al. Fast Texture Database Retrieval Using Extended Fractal Features. Storage and Retrieval for Image and Video Databases, VI, vol. 3,312, pp. 162-173, SPIE Press, 1998.

[25] T. Randen and J.H. Husoy. Filtering for Texture Classification: A Comparative Study. IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 21, no. 4, pp. 291-310, Apr. 1999.

[26] C.C. Hsu, W.W. Chu, and R.K. Taira. A Knowledge-Based Approach for Retrieving Images by Content. IEEE Trans. Knowledge and Data Eng., vol. 8, no. 4, pp. 522-532, 1996.

[27] H. Chen, B. Schatz, T. Ng, J. Martinez, A. Kirchhoff, and C. Lim. A Parallel Computing Approach to Creating Engineering Concept Spaces for Semantic Retrieval: The Illinois Digital Library Initiative Project. IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 18, no. 8, pp. 771-782, Aug. 1996.

[28] N. Vasconcelos and A. Lippman. A Probabilistic Architecture for Content-Based Image Retrieval. Proc. Computer Vision and Pattern Recognition, pp. 216-221, 2000.

[29] H. Murase and S.K. Nayar. Visual Learning and Recognition of 3D Objects from Appearance. Int'l J. Computer Vision, vol. 14, no. 1, pp. 5-24, 1995.

[30] R.W. Picard and T.P. Minka. Vision Texture for Annotation. Multimedia Systems, vol. 3, pp. 3-14, 1995.

[31] Aditya Vailaya et. el. Image Classification for content-Based Indexing. IEEE reanscations on Image Processing, Vol. 10, No. 1, Jan. 2001.

[32] Yong rui et. el. Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval. IEEE transactions on circuits and Video Technology. Volume 8, Number 5, 1998, pp. 644-655.

[33] A. D. Narasimhalu, Multimedia Syst. (Special Section on Content-Based Retrieval), 1995.

[34] W. Niblack, R. Barber et al. The QBIC project: Querying images by content using color, texture and shape. In Proc. SPIE Storage and Retrieval for Image and Video Databases, Feb. 1994.

[35] P. P. Ohanian and R. C. Dubes. Performance evaluation for four classes of texture features. Pattern Recognition, vol. 25, no. 8, pp. 819–833, 1992.

[36] M. Ortega, Y. Rui, and K. Chakrabarti, S. Mehrotra, and T. S. Huang. Supporting similarity queries in MARS. In Proc. ACM Conf. Multi-media, 1997.

[37] A. Pentland and R. Picard. IEEE Trans. Pattern Anal. Machine Intell. (Special Issue on Digital Libraries), 1996.

[38] A. Pentland, R. W. Picard, and S. Sclaroff. Photobook: Content-based manipulation of image databases. Int. J. Comput. Vision, 1996.

[39] R. W. Picard and T. P. Minka. Vision texture for annotation. Multi-media Syst. (Special Issue on Content-Based Retrieval).

[40] J. Dowe. Content-based retrieval in multimedia imaging. In Proc. SPIE Storage and Retrieval for Image and Video Databases, 1993.

[41] Y. Rui, T. S. Huang, and S. Mehrotra. Content-based image retrieval with relevance feedback in MARS. In Proc. IEEE Int. Conf. ImageProcessing, 1997.

[42] J. Z. Wang, G. Li, and G. Wiederhold. SIMPLIcity: Semantics-sensitive Integrated Matching for Picture LIbraries. In IEEE Trans. on pattern Analysis and Machine Intelligence, volume 23, pages 947--963, 2001.

[43] C. Breen, L. Khan, and A. Kumar. Image Classification Using Neural Networks and Ontologies. IEEE DEXA, International Workshop on Web Semantics, France, Sept 2002.

[44] L. Khan and D. McLeod. Audio Structuring and Personalized Retrieval Using Ontologies. IEEE Advances in Digital Libraries, Library of Congress, Washington, DC, May 2000.

[45] Casey Breen, Latifur Khan, Arun Kumar, and Lei Wang. Ontology-based image classification using neural networks", SPIE 2002

[46] J. R. Smith, S.-F. Chang, “Searching for Images and Videos on the World-Wide Web,” Columbia University, No. CU/CTR/TR 459-96-25, Aug. 1996.

[47] Wei-Cheng Lai, Edward Chang, and Kwang-Ting (Tim) Cheng. An Anatomy of a Large-scale Image Search Engine. WWW 2002, 7-11 May 2002, Honolulu, Hawaii.

[48] A. Gersho and R. Gray. Vector Quantization and Signal Compression. Kluwer Academic, 1991

[49] Hong-Jiang Zhang, Zheng Chen, Wen-Yin Liu and Mingjing Li. Relevance Feedback in Content-Based Image Search. Invited Keynote, 12th Int. Conf. on New Information Technology (NIT), May 29-31, Beijing.

[50] Panrit Tosukhowong, Frederic Andres, Kinji Ono, Jose Martinez, Noureddine Mouaddib, Nicolas Dessaigne and Douglas C. Schmidt A Flexible Image Search Engine. ACM, MM99, Oct 30 - Nov 5, 1999, Orlando, Florida

[51] Harold Kosch. MPEG-7 and Multimedia Database Systems. SIGMOD Record, Vol. 31, No.2, June 2002.

[52] Mark van Setten, Erik Oltmas, Mettina Veenstra. Personalized Video Search and Retrieval using MPEG-7 and Stylesheets. https://doc.telin.nl/dscgi/ds.py/Get/File-8842/

[53] V. S. Subrahmanian, Sibel Adali, Anne Brink, Ross Emery, James J. Lu, Adil Rajput, Timothy J. Rogers, Robert Ross, and Charles Ward: HERMES: A heterogeneous Reasoning and Mediator System. Technical Report, University of Maryland, Maryland, 1995.

[54] S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom: The TSIMMIS Project: Integration of Heterogeneous Information Sources. IPSJ Conference, pp. 7-18, Tokyo, Japan, October 1994.

[55] Rajeev Agrawal, Mukesh Mohania, Yahiko Kambayashi, S S Bhowmick, S Madria: An Architecture for XML Schema Integration. ICDL: Research and Practices, Kyoto, Japan, 2000.

[56] Ralf Behrens: A Grammar Based Model for XML Schema Integration. BNCD, 2000

[57] Ronaldo dos Santos Mello, Carlos Alberto Heuer: A Bottom-Up Approach for Integration of XML Sources. WIIW 2001

[58] Dongwon Lee, Murali Mani, Wesley W. Chu: Effective Schema Conversions between XML and Relational Models. In European Conf. on Artificial Intelligence (ECAI), Knowledge Transformation Workshop (ECAI-OT), Lyon, France, July 2002.

[59] P. Stanchev, G. Amato, F. Falchi, C. Gennaro, F. Rabitti, P. Savino, Selection of MPEG-7 Image Features for Improving Image Similarity Search on Specific Data Sets, 7-th IASTED International Conference on Computer Graphics and Imaging, CGIM 2004, Kauai, Hawaii, 395-400, 2004.

Authors' Information

Rajeev Agrawal – Kettering University, Flint, MI 48504, USA; Email: ragrawal@kettering.edu

Farshad Fotouhi – Wayne State University, Detroit, MI 48202, USA; Email: fotouhi@cs.wayne.edu

Peter Stanchev – Kettering University, Flint, MI 48504, USA; Email: pstanche@kettering.edu

Ming Dong – Wayne State University, Detroit, MI 48202, USA; Email: mdong@cs.wayne.edu

MPEG-7: THE MULTIMEDIA CONTENT DESCRIPTION INTERFACE

Peter Stanchev, David Green Jr., and Boyan Dimitrov

Abstract: In this paper a review of the most used MPEG-7 descriptors are presented. Some considerations for choosing the most proper descriptor for a particular image or video data set are outlined.

Keywords: MPEG-7, Multimedia, Content based retrieval

1. Introduction

More and more digital images and video are being captured and stored. In order to use this information, an efficient retrieval technique is required. One major development in this area is the content based image and video retrieval techniques which use image features for indexing and retrieval [Rabitti, 1989]. The most used features are color, texture, and shape. Several semantic image and video models are suggested [Stanchev, 1999], [Grosky, 2001]. In MPEG-7 standard different descriptors for this purpose are proposed [Manjunath, 2002]. What descriptor is the best for a particular data set? Some preferable answers of this question are given.

2. MPEG-7 Descriptors

The MPEG-7 descriptors can be classified as general visual descriptors and domain specific descriptors. The former include color, texture, shape and motion features. The latter includes face recognition descriptor. Although distance functions are not part of the standard, we will present the most used distance functions. Only color, texture and shape descriptors are covered, since they are used mostly.

2.1. Color descriptors

Color is one of the most widely used image and video retrieval features [Schettini, 2001]. The MPEG-7 standard includes five color descriptors which represents different aspects of the color and includes color distribution, spatial layout, and spatial structure of the color. The histogram descriptors capture the global distribution of colors. The dominant color descriptor represents the dominant colors used. The color layout descriptor captures the spatial distribution or layout of the colors in a compact representation. While MPEG-7 standards accommodate different color spaces, most of the color descriptors are constrained to one or a limited number of color spaces for ensuring inter-operability.

2.1.1. Dominant color descriptor

This descriptor specifies a set of dominant colors in an image [Cieplinski, 2000]. It is good to represent color features where a small number of colors are enough to characterize the color information. The extraction algorithm quantizes the pixel color values into a set of dominant colors. The matching is done by calculating the distances between dominant color sets based on the difference between corresponding colors in any two sets of dominants.

The result of the method is a vector with integer numbers, presented as $F = \{(c_i, p_i, v_i), s\}$, (*i*=*1,2, ..., N*), where *N* is the number of dominant colors. The vector components are: the dominant color value c_i (RGB color space vector); p_i - normalized fraction of pixels corresponding to color c_i; optimal color variance v_i, (describes the variance of the color values of the pixels in a cluster around the corresponding color); and the coherency s representing the overall spatial homogeneity of the dominant colors.

The distance algorithm uses an estimate of the mean square error, based on the assumption that the sub-distributions described by dominant colors and variances are Gaussian. Consider 2 descriptors:

$$F_1 = \{(c_1^{'}, p_1^{'}, v_1^{'}), s_1^{'}\} \qquad (i = 1,2,...,N_1) \quad \text{and}$$

$$F_2 = \{(c_2^{'}, p_2^{'}, v_2^{'}), s_2^{'}\} \qquad (i = 1,2,...,N_2)$$

where $p \in [0,31]$, $c_i = rgb2luv(c_i^{'})$, $v_i = \begin{cases} 60.0 & v_i^{'} = 0 \\ 90.0 & v_i^{'} = 1 \end{cases}$, $p_i = \dfrac{(p_i^{'} + 0.5)/31.9999}{\sum_i p_i}$, and if

$$f_{x_i y_j} = \frac{1}{2\pi\sqrt{2\pi \times (v_{x_i}^{(l)} + v_{y_j}^{(l)}) \times (v_{x_i}^{(u)} + v_{y_j}^{(u)}) \times (v_{x_i}^{(v)} + v_{y_j}^{(v)})}} \times \exp\left\{ -\frac{1}{2} \left[\frac{(c_{x_i}^{(l)} - c_{y_j}^{(l)})^2}{v_{x_i}^{(l)} + v_{y_j}^{(l)}} + \frac{(c_{x_i}^{(u)} - c_{y_j}^{(u)})^2}{v_{x_i}^{(u)} + v_{y_j}^{(u)}} + \frac{(c_{x_i}^{(v)} - c_{y_j}^{(v)})^2}{v_{x_i}^{(v)} + v_{y_j}^{(v)}} \right] \right\}$$

and if $D_v = \sqrt{\sum_{i=1}^{N_1}\sum_{j=1}^{N_1} p_{1_i} p_{1_j} f_{1_i 1_j} + \sum_{i=1}^{N_2}\sum_{j=1}^{N_2} p_{2_i} p_{2_j} f_{2_i 2_j} - 2\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} p_{1_i} p_{2_j} f_{1_i 2_j}}$,

then the distance is calculated as: $D = [0.3 \times abs(s_1 - s_2) + 0.7] \times D_v$.

2.1.2. Scalable Color descriptor

This descriptor performs color histogram in HSV color space encoded by a Haar transform [MPEG, 2002]. The extraction is done by quantizing the image into a 256 bin HSV color space histogram and then using the Haar transform to reduce the number of bins.

The output of the method is *a* vector with integer components, presented by a histogram with 64, 32 or 16 bins.

The distance matching can be done either in the Haar coefficient domain or in the histogram domain. In the case where only the coefficient signs are retained, the matching can be done efficiently in the Haar coefficient domain by calculating the Hamming distance as the number of bit positions at which the binary bits are different using an XOR operation on the two descriptors to be compared. This induces only a marginal loss in similarity matching precision compared to reconstructing the color histogram and performing histogram matching, while the computational cost is considerably lower.

2.1.3. Color layout descriptor

This descriptor performs spatial distribution of colors [Kasutani, 2001]. The extraction is being done as follows: the image is divided into 8x8 blocks. For each block, a single dominant color is selected. The resulting 8x8 image is then transformed into a series of coefficients using dominant color descriptors transformation. These are finally quantized to fit an assigned number of bits.

The method output is a vector with integer components, describing *{DY, DCr, DCb}* coefficients, where *Y* is the coefficient value for luminance, *Cr, Cb* coefficient values for chrominance.

For matching two descriptions *{DY, DCr, DCb}* and *{DY', DCr', DCb'}* the following formula:

$D = \sqrt{\sum_i w_{yi}(DY_i - DY_i^{'})^2} + \sqrt{\sum_i w_{bi}(DCb_i - DCb_i^{'})^2} + \sqrt{\sum_i w_{ri}(DCr_i - DCr_i^{'})^2}$ is used, where *i* represents the zigzag- scanning order of the coefficients.

2.1.4. Color structure descriptor

This descriptor is a generalization of the color histogram that encodes information about the spatial structure of the colors in an image as well as their frequency of occurrence [Messing, 2001]. The histogram is extracted in HMMD color space and non-uniformly quantizing is performed over the histogram values. This descriptor specifies spatial distribution of colors. It is calculated by letting a structuring element with image samples to visit each position in the image and then summarize the frequency of color occurrences in each structuring element location in a histogram. The structuring element always has dimensions 8x8, but the distance between the samples in the original image differs with the resolution.

The output of the method is a vector with integer components, presented by a 256 bin histogram.

The matching is done by minimizing the distance calculated as the sum of the differences between the corresponding bins in any two color-structure histograms.

2.1.5. Group-of-frame or Group-of-picture descriptor

This descriptor is a compound descriptor that expresses the color features of a collection of images or video frames by means of the scalable color descriptor [Ferman, 2000]. During the extraction the average, median or intersection scalable color histogram of the frame/picture group is calculated from scalable color histograms of each group/picture. The intersection histogram is a histogram with the minimum value for each bin over all histograms in the group.

The output of the descriptor is a vector with integer components, as in the case of scalable color descriptor.

The matching is done in the same way as for the scalable color descriptor.

2.2. Texture descriptors

The image texture is one of the most important image characteristic in both human and computer image analysis and object recognition [Manjunath, 2001]. Visual texture is a property of a region in an image. There are two texture descriptors in MPEG-7: a homogeneous texture descriptor, and edge histogram descriptor. Both of these descriptors support search and retrieval based on content descriptions.

2.2.1. Homogeneous Texture

This descriptor is aimed at texture-based image-to-image matching [Ro, 2001]. During the extraction, the mean and standard deviation of the image pixel intensities is computed. Energy and energy deviation feature values are computed by applying 30 Gabor filters in the frequency domain. The polar form used in the frequency domain in this approach is more suited for rotation invariant analysis than the Cartesian form.

The output of the method is: the average value (an integer number in the interval [0,255]); standard deviation (an integer number in the interval [0,255]); energy (30 integer numbers in the interval [0,255]); energy deviation (30 integer numbers in the interval [0,255]).

The matching is done by summing the normalized weighted absolute difference between two sets of feature vectors not using rotation or scale invariant algorithms.

2.2.2. Edge histogram descriptor

This descriptor is a texture descriptor and describes the spatial distribution of four directional edges and one nondirectional edge in three different levels of localization in an image [Park, 2000]. The localization levels are the global, the semi-global and the local level. During the extraction, the image is partitioned into 16 non-overlapping sub-images with sizes depending on the original image size. It is also divided into a preferred number of image-blocks. For each image-block, a horizontal, a vertical, a 45 degree diagonal, a 135 degree diagonal and a nondirected edge value is calculated using edge extraction filters applied on the average brightness values in four sub-blocks. If the maximum edge value is greater than a threshold value, the image-block is considered to contain the corresponding edge. Otherwise, the image-block is considered to contain no edge. The image-block edge composition in the sub-images forms a local edge histogram with a total of 80 bins (5 types of edges, for each of the16 sub-images). The global edge histogram summarizes the distribution of the different edges in the whole image by adding the corresponding local edge histogram bins into five global histogram bins one for each type of edge. The semi-global edge histogram is generated by accumulating the edge compositions in the sub-image clusters.

The output is a vector of 80 integer numbers between [0, 7].

Distance is calculated as added weighted difference between the local, global, and semi-global edge histograms respectively. Significance is measure by is the sum of absolute difference of 150 coefficients extracted from the 80 bins.

2.3. Shape descriptors

MPEG-7 supports region and contour shape descriptors. Object shape features are very powerful when used in similarity retrieval.

2.3.1. Region Shape

In the region shape descriptor, the shape of an object can be a single or multiple regions with or without holes [Kim, 1999]. The feature extraction is based on a set of Angular Radial Transform (ART) coefficients. ART is a complex 2-D transform defined on a unit disc with polar coordinates. In practice, the needed values of the basic functions are pre-calculated and put into a lookup table during the first step of the extraction. The ART transformation is then done by summing up the multiplication for each image pixel with each corresponding pixel in the lookup table, calculating the magnitudes.

The output is a vector of 35 integer numbers in the interval [0, 15].

The matching is done by calculating the minimum distance between the feature vectors for any shapes of two images. The distance for two vectors is the sum of absolute difference of coefficients.

2.3.2. Contour Shape

The contour shape descriptor presents a closed 2-D object or region contour in an image or video sequence [Mokhtarian, 1992]. During the extraction, N equidistant points are selected on the contour, starting from an arbitrary point on the contour and following the contour clockwise. The contour is then smoothed by repetitive low-pass filtering of the *x* and *y* coordinates of the selected contour points. The smoothing flattens out the concave parts of the contour. Points separating concave and convex parts of the contour and peaks in between are then identified and the normalized values are saved in the descriptor.

2.4. An example of MPEG- 7 descriptors representation

An example of MPEG-7 XML form for some descriptors on the sample image taken from TREC2002-FeatureDevelopment-mpeg1VideoSet [Smeaton, 2002], shown in Figure 1. is given after the figure.

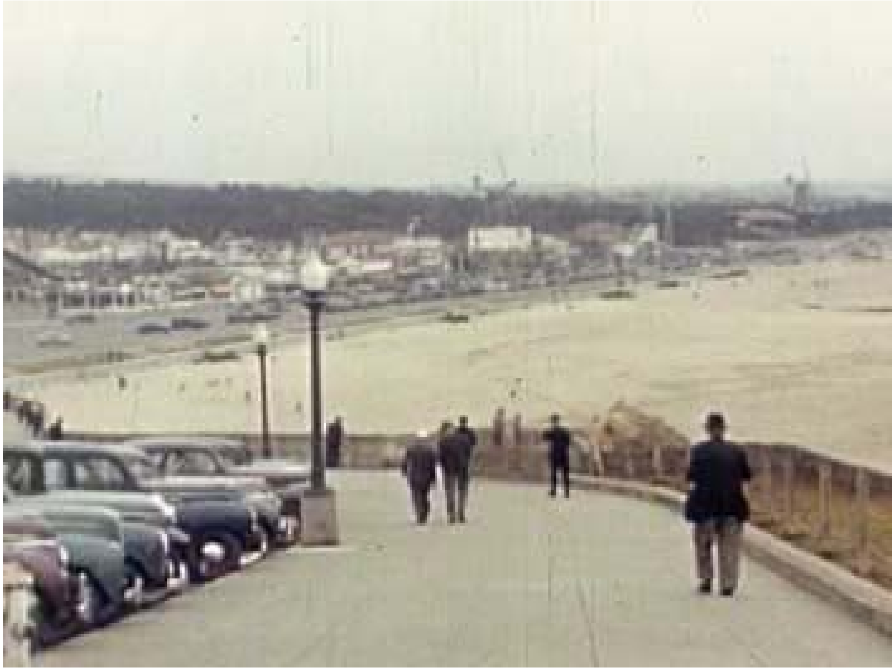


Figure 1. A sample image from the movie “San Francisco, 1944”, KeyFrame from 1130-1407.jpg

```
<?xml version="1.0" encoding="ISO-8859-1"?><Mpeg7
xmlns="urn:mpeg:mpeg7:schema:2001"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:mpeg7="urn:mpeg:mpeg7:schema:2001"
```

```
xmlns:xml="http://www.w3.org/XML/1998/namespace"
xsi:schemaLocation="urn:mpeg:mpeg7:schema:2001 Mpeg7-2001.xsd">
<Description xsi:type="ContentEntityType"><MultimediaContent
xsi:type="ImageType">
<Image>
<MediaLocator>
<MediaUri>urn:milos:image_jpeg:SanFranc1940_KeyFrame_1130_1407</MediaUri>
</MediaLocator>
<VisualDescriptor xsi:type="ScalableColorType" numOfBitplanesDiscarded="0"
numOfCoeff="64">
<Coeff>-40 1 -12 63 -14 11 3 5 31 26 -5 9 -54 -2 12 9 -7 2 3 1 -3 5 2 -1 9
-2 1 1 -3 5 1 -4 3 3 1 3 3 2 -2 2 1 1 1 3 1 0 3 5 -9 3 2 0 -2 0 1 -3 0 0 0
-2 1 0 -3 -3</Coeff>
</VisualDescriptor>
<VisualDescriptor xsi:type="ColorLayoutType">
<YDCCoeff>48</YDCCoeff>
<CbDCCoeff>25</CbDCCoeff>
<CrDCCoeff>34</CrDCCoeff>
<YACCoeff5>15 26 22 16 14 </YACCoeff5>
<CbACCoeff2>18 17 </CbACCoeff2>
<CrACCoeff2>15 14 </CrACCoeff2>
</VisualDescriptor>
<VisualDescriptor xsi:type="ColorStructureType" colorQuant="2"><Values>0 0
0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 34 85 170 103 0 0 0 0 7 5 0 0 0
3 2 0 43 110 146 150 0 7 29 28 32 52 32 4 54 64 45 18 8 28 53 139 137 93 44
11</Values>
</VisualDescriptor>
<VisualDescriptor
xsi:type="DominantColorType"><SpatialCoherency>23</SpatialCoherency><Value>
<Percentage>1</Percentage><Index>45 38 44</Index><ColorVariance>1 0
0</ColorVariance></Value><Value><Percentage>10</Percentage><Index>191 186
163</Index><ColorVariance>0 0
0</ColorVariance></Value><Value><Percentage>9</Percentage><Index>229 230
218</Index><ColorVariance>0 0
0</ColorVariance></Value><Value><Percentage>3</Percentage><Index>111 108
111</Index><ColorVariance>0 0
0</ColorVariance></Value><Value><Percentage>4</Percentage><Index>158 154
147</Index><ColorVariance>0 0
0</ColorVariance></Value><Value><Percentage>1</Percentage><Index>117 100
84</Index><ColorVariance>0 0 0</ColorVariance></Value></VisualDescriptor>
<VisualDescriptor xsi:type="EdgeHistogramType"><BinCounts>0 1 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 7 3 3 3 0 7 2 2 3 0 7 6 3 2 0 6 3 0 3 0 6 2 4 3 3 4
3 1 3 0 6 2 2 3 0 4 4 4 2 3 6 2 4 4 0 1 3 1 2 1 0 0 0 0 2 3 1 6
4</BinCounts>
</VisualDescriptor>
<VisualDescriptor xsi:type="HomogeneousTextureType">
<Average>188</Average><StandardDeviation>114</StandardDeviation><Energy>220
214 205 160 206 218 212 181 180 146 173 183 179 161 137 142 135 163 152 138
128 142 133 133 135 95 74 89 76 87</Energy><EnergyDeviation>223 215 205 156
207 218 202 173 180 140 164 176 169 150 122 127 122 158 143 114 126 123 129
132 129 76 58 77 53 76</EnergyDeviation></VisualDescriptor>
</Image>
</MultimediaContent>
</Description></Mpeg7>
```

3. The Use of MPEG-7 Descriptors

There are several problems, which have to be solved before evaluating the quality of different descriptors. The first problem is: how to choose the benchmark database? There is no common database used for content based benchmarking. Many researchers use the Corel image database (http://www.corel.com/). Another possibility is the collection used in MPEG-7 [MPEG, 1998], but it is also copyrighted as Corel database. Other possibilities are the databases on:

- http://elib.cs.berkeley.edu/photos/tarlist.txt,
- http://www.cs.washington.edu/research/imagedatabase/groundtruth,
- http://www.white.media.mit.edu/vismod/imagery/VisionTexture/vistex.html.

The second problem is how to measure the performance of the different descriptors. This mean to find a set of features which adequately encodes the characteristics that we intend to measure and a suitable metric. Which is the best similarity function? In 1977 Amos Tversky proposed his famous feature contrast model [Tversky, 1977]. He uses a set of binary features. In [Eidenberger, 2003] mean and standard deviation, distribution analysis and cluster analysis are used. Some of the results are: Color Layout performs badly on monochrome data. Like Color Layout, Color Structure performs inferior on monochrome data. The Dominant Color identifier performs equally well on any type of media. Scalable Color performs exactly like Color Layout and Color Structure. All color descriptors works excellent on photos but three of four perform badly on artificial media objects with few color gradations and very badly on monochrome content. An exception is the Dominant Color descriptor. This descriptor works well on each type of content. Edge Histogram performs excellent on any type of media. The Homogeneous Texture descriptor works acceptably on the Brodatz dataset. A combination of different descriptors is needed. The best descriptors for using combinations are Color Layout, Dominant Color, Edge Histogram and Texture Browsing. The others are highly dependent on these. The color histograms (Color Structure and Scalable Color) perform badly on monochrome input. Therefore, Dominant Color should be used for GoF/GoP color instead of Scalable Color. Generally, all descriptors are highly redundant and applying complexity reduction transformations could save up to 80% of storage and transmission capacity.

In [Stanchev, 2004] we generalized this result. We propose a technique for evaluating the effectiveness of MPEG-7 image features on specific image data sets, based on well defined statistical characteristics of the data set. The aim is to improve the effectiveness of the image retrieval process based on the computed similarity on these features. We also validate this method with extensive experiments with real users.

Finally, some aspects of images are captured by none of the descriptors and existing descriptors should be either refined or new visual descriptors should be added to the standard.

Conclusion

Several visual descriptors exist for representing the physical content of images, for instance color histograms, textures, shapes, regions, etc. Depending on the specific characteristics of a data set, some features can be more effective than others when performing similarity search. For instance, descriptors based on color representation might be effective with a data set containing mainly black and white images. Techniques based on statistical analysis of the data set and queries are useful.

It seems that the most intelligent descriptors are the one based on color layout. Not only does it compare the colors, but also where in the image they occur. This can be of great use if you are looking for a sunset, a face, a certain kind of landscape view etc, where similar colors usually occur in the same regions of the images. The texture and shape based search methods can also be very good, but the search results that are not among the used ground truth set can often be perceived as looking completely different compared to the query image so the use in general image databases can be questioned. On the other hand, the texture and shape based methods can recognize features such as contours and appearance that cannot be detected by the color based methods.

Even if it is not possible, in general, to overcome the semantic gap in image retrieval by feature similarity, it is still possible to increase the retrieval effectiveness by a proper choice of the image features, among those in the MPEG-7 standard, depending on the characteristics of the various image data sets (obviously, the more homogeneous the data set is, better results can be obtained).

Bibliography

[Cieplinski, 2000] Leszek Cieplinski, Results of Core Experiment CT4: Extension of Dominant Colour Descriptor, MPEG-7 TR #13-06, January 2000

[Eidenberger, 2003] H. Eidenberger, “How good are the visual MPEG-7 features?”, SPIE & IEEE Visual Communications and Image Processing Conference, Lugano, Switzerland, 2003

[Ferman, 2000] A. Ferman et al., “Group-of-frame/picture color histogram descriptors for multimedia applications”, Proceedings of the Storage and Retrieval of the IEEE International Conference on Image Processing”, Vol. 1, Vancouver, Canada, 2000, 65-68

[Grosky, 2001] Grosky W., Stanchev P., “Object-Oriented Image Database Model”, 16th International Conference on Computers and Their Applications (CATA-2001), March 28-30, 2001, Seattle, Washington (94-97).

[Kasutani, 2001] E. Kasutani, A. Yamada, “The MPEG-7 color layout descriptor: a compact image feature description for high-speed image/video retrieval”, Proceeding of International Conference on Image Processing 2001, Oct. 2001, Thessaloniki, Greece 2001

[Kim, 1999] W. Kim, Y. Kim, “A new Region-Based Shape Descriptor”, MPEG-7 TR#15-01, December 1999

[Manjunath, 2001] Manjunath B., Ohm J., Vasudevan V., Yamada A., Color and Texture Descriptors, IEEE Transactions on circuits and systems for video technology, V. 11, No. 6, June 2001, 703-715

[Manjunath, 2002] B.S. Manjunath, P. Salembier, T. Sikora, “Introduction to MPEG-7”, Wiley, 2002

[Messing, 2001] Dean S. Messing, Peter van Beek, James H. Errico, Using Colour and Local Spatial Information to Describe Images, MPEG-7 TR #13-07, January 2001

[Mokhtarian, 1992] F. Mokhtarian, A. Mackworth, “A theory of multiscale, curvature-based shape representation for planer curves”, IEEEE Transaction on Pattern analysis and machine intelligence, 14 (8), 1992, 789-805

[MPEG, 2002] “MPEG-7 Overview (version 9)”, ISO/IEC JTC1/SC29/WG11N5525

[MPEG, 1998] MPEG Requirements Group, “MPEG-7: Context and objectives (version 10 Atlantic City),” Doc. ISO/IECJTC1/SC29/WG11, International Organisation for Standardisation, 1998.

[Park, 2000] D. Park, Y. Jeon, C. Won, S. Park, “Efficient use of local edge histogram descriptor”, Processing of ACM International workshop on Standards, Interoperability and Practices, Marina del Rey, CA, USA, 2000, 52-54

[Rabitti, 1989] Rabitti F., Stanchev P., "GRIM_DBMS - a GRaphical IMage DataBase System", in "*Visual Database Systems*", T. Kunii (edt.) North-Holland 1989 (415-430).

[Ro, 2001] Y. Ro, M. Kim, H. Kang, B. Maniunath, J. Kim, “MPEG-7 Homogeneous Texture descriptor”, ETRI Journal 23 (2), 2001, 41-51.

[Schettini, 2001] Schettini R., Ciocca G., Zuffi S., A survey of methods for color image indexing and retrieval in image databases, in Luo R., MacDonal L., (editors) Corol Imaging Science: Exploiting Digital Media, J. Willey, 2001

[Smeaton, 2002] A. Smeaton, P. Over, The TREC-2002 Video Track report, http://www-nlpir.nist.gov/projects/t2002v/results/notebook.papers/VIDEO.OVERVIEW.pdf

[Stanchev, 1999] Stanchev P., “General Image Database Model”, in Visual Information and Information systems, Huijsmans, D. Smeulders A., (etd.) Lecture Notes in Computer Science 1614, 1999 (29-36).

[Stanchev, 2004] P.Stanchev, Giuseppe Amato, Fabrizio Falchi, Claudio Gennaro, Fausto Rabitti, Pasquale Savino, “Selection of MPEG-7 Image Features for Improving Image Similarity Search on Specific Data Sets”, The sevent IASTED International Conference “Computer graphics and imaging”, Kauai, Hawaii, 2004

[Tversky, 1977] A. Tversky, "Features of Similarity", Philosophical review, 84/4, 327-352, 1977

Authors' Information

Peter L. Stanchev – pstanche@kettering.edu

David Green Jr. – dgreen@kettering.edu

Boyan Dimitrov – bdimitro@kettering.edu

Kettering University, Flint, MI 48504, USA

IMAGESPACE: AN ENVIRONMENT FOR IMAGE ONTOLOGY MANAGEMENT

Shiyong Lu, Rong Huang, Artem Chebotko, Yu Deng, and Farshad Fotouhi

Abstract: More and more researchers have realized that ontologies will play a critical role in the development of the Semantic Web, the next generation Web in which content is not only consumable by humans, but also by software agents. The development of tools to support ontology management including creation, visualization, annotation, database storage, and retrieval is thus extremely important. We have developed ImageSpace, an image ontology creation and annotation tool that features (1) full support for the standard web ontology language DAML+OIL; (2) image ontology creation, visualization, image annotation and display in one integrated framework; (3) ontology consistency assurance; and (4) storing ontologies and annotations in relational databases. It is expected that the availability of such a tool will greatly facilitate the creation of image repositories as islands of the Semantic Web.

Keywords: Ontology, visualization, annotation, Semantic Web, DAML+OIL, ontology storage, ontology-based retrieval.

1. Introduction

More and more researchers have realized that ontologies will play a critical role in the development of the Semantic Web, the next generation Web in which content is not only consumable by humans, but also by software agents. [1,5]. Undoubtedly, images will be major constituents of the Semantic Web, and how to share, search and retrieve images on the Semantic Web is an important but challenging research problem. Unlike other resources, the semantics of an image is implicit in the content of an image. Although this is not a problem to human cognition, it imposes a challenge on image searching and retrieval based on the semantics of image content. Manual annotation of images provides an opportunity to make the semantics of an image explicit and richer. However, different annotators might use different vocabulary to annotate images, which cause low recall and precision in image search and retrieval. We propose an ontology-based annotation approach, in which an ontology is created for a particular domain so that the terms and their relationships are formally defined. In this way, annotators of a particular image domain, say, the Family Album domain, will use the same vocabulary to annotate images, and users will search images guided by the ontology with greater recall and precision.

In [11, 12], we have briefly described *ImageSpace*, an image ontology creation and annotation tool, and our experience of annotating linguistic data using *ImageSpace* for the preservation of endangered languages [13, 17, 18]. This paper extends these results with ontology visualization, the storage of ontologies and annotations in relational databases, and ontology-based information retrieval. In summary, the contributions of this paper are:

- *ImageSpace* supports the functionality of ontology creation. In particular, it facilitates the creation of classes, properties, and relations between classes and relations between properties. It also provides ontology consistency assurance;
- *ImageSpace* provides full support for the standard web ontology language DAML+OIL [2];
- *ImageSpace* supports the visualization of an ontology to enable users to navigate, zoom-in and zoom-out various portions of an ontology.
- *ImageSpace* supports ontology-driven annotation of images.
- *ImageSpace* supports the storage of ontologies and annotations in a relational database.
- Finally, we have developed a simple web-based image retrieval system to search images.

Organization. The rest of the paper is organized as follows. Section 2 describes related work. Section 3 gives a brief primer for the DAML+OIL ontology language. Section 4, section 5 and section 6 present how to create and visualize an image ontology, and annotate images based on the created ontology using *ImageSpace*. Section 7 describes our approach to store ontologies and annotations in relational database. Section 8 gives an overview of a prototype image retrieval system. Finally, Section 9 concludes the paper and presents some future work.

2. Related Work

Extensive research has been conducted on the processing, searching and retrieval of images [6]. Recently, due to the vision of the Semantic Web [1, 5] and the important role of ontologies, there is an increasing interest in ontology-based approaches to image processing and early results show that the use of ontologies can enhance classification precision [8] and image retrieval performance [7].

Numerous ontology creation tools have been developed. Among them, *Protégé* (http://protege.stanford.edu/), developed at Stanford University, and *OntoEdit* [9] are two well-known representatives. While some of these tools provide partial support to DAML+OIL, *ImageSpace* provides full support of this language, and integrate image ontology creation, image annotation and display in one framework. The tool is built particularly with image support in mind and features a user-friendly interface support for image display and ontology-driven annotation capabilities.

Recently, independently and concurrently, *Protégé* has released five publicly accessible plugins that provide capabilities for ontology visualization: *ezOWL*, *Jambalaya*, *OntoViz*, *OWLViz*, and *TGViz*. *ezOWL* supports graphical ontology building. *ezOWL* and *OntoViz* have *ERWin*-like views of ontology classes (rectangles with names) with their properties and restrictions (“attribute” fields in rectangles). *Jambalaya* [10] provides nested interchangeable views and nicely implements three zooming approaches: geometric, semantic and fisheye zooming. *OWLViz*, and *TGViz* have graph-like views of ontologies. *OWLGraph* shares a lot of features with these tools but it provides a richer set of views and layouts. For more details of the features of *Protégé*, the reader is referred to http://protege.stanford.edu/plugins/domain_visualization.html.

3. A Primer on DAML+OIL

DAML+OIL is a semantic markup language for publishing and sharing ontologies on the World Wide Web. It is developed as an extension of XML [14], RDF [15] and RDF Schema (RDF-S) [16] by providing additional constructs along with a formal semantics. DAML+OIL uses 44 constructs (or XML tags) to define ontologies, classes, properties, individuals, data types and their relationships. In the following, we present a brief overview of the major constructs and refer the reader to [2] for more details.

Classes. A class defines a group of individuals that share some properties. A class is defined by *daml:Class*, and different classes can be related by *rdfs:subClassOf* into a class hierarchy. Other relationships between classes can be specified by *daml:sameClassAs, daml:disjointWith*, etc. The extension of a class can be specified by *daml:oneOf* with a list of class members or by *daml:intersectionOf* with a list of other classes.

Properties. A property states relationships between individuals or from individuals to data values. The former is called *ObjectProperty* and specified by *daml:ObjectProperty*. The later is called *DatatypeProperty* and specified by *daml:DatatypeProperty*. Similarly to classes, different properties can be related by *rdfs:subPropertyOf* into a property hierarchy. The domain and range of a property are specified by *rdfs:domain* and *rdfs:range* respectively. Two properties might be asserted to be equivalent by *daml:samePropertyAs.* In addition, different characteristics of a property can be specified by *daml:TransitiveProperty, daml:UniqueProperty,* etc.

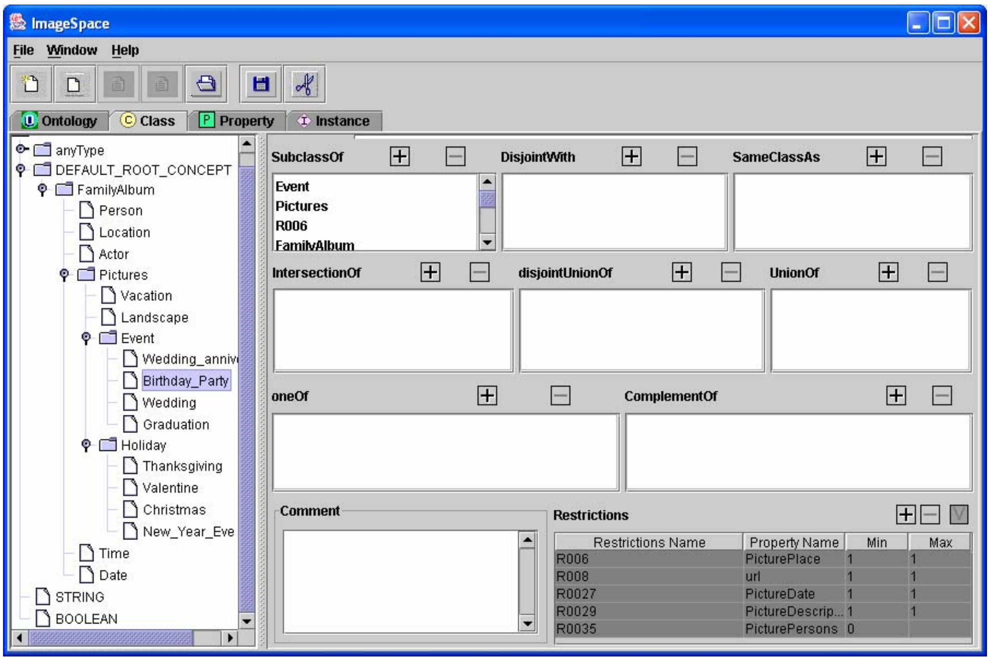
Property restrictions. A property restriction is a special kind of class description. It defines an anonymous class, namely the set of class of all individuals that satisfy the restriction. There are two kinds of property restrictions: *value constraints* and *cardinality constraints*. Value constraints restrict the values that a property can take within a particular class, and they are specified by *daml:toClass, daml:hasClass, etc.* Cardinality constraints restrict the number of values that a property can take within a particular class, and they are specified by *daml:minCardinality, daml:maxCardinality, daml:cardinality*, etc.

Recently, DAML+OIL [2] has been revised into OWL, which is a Web ontology language that has become a W3C recommendation [3].

4. Creating an Image Ontology

ImageSpace provides a user-friendly interface to the user to create image ontologies. Figure 1 shows a snapshot of creating an image ontology *FamilyAlbum*. The four tabs, labeled by *Ontology*, *Class*, *Property* and *Instance,* facilitate the specification of these components and their relationships in a graphical fashion.

As shown in Figure 1, when the *Class* tab is enabled, the left frame displays the class hierarchy, and the right frame shows the relationships of this class with other classes including restriction classes. With this interface, one can easily insert, delete, and update a class. In addition, using the right frame, one can specify the relationships of this class with other classes. At the right-bottom corner of the right frame, is a panel that corresponds to property restrictions, where a user can specify both value constraints and cardinality constraints. Note that those shaded property restrictions are automatically inherited from their parent classes unless they are overridden. Also note that, since a class might have multiple parents, other parent classes are shown in the *SubClassOf* field.



- anyType
- DEFAULT_ROOT_CONCEPT
  - FamilyAlbum
    - Person
    - Location
    - Actor
    - Pictures
      - Vacation
      - Landscape
      - Event
        - Wedding_anniv
        - Birthday_Party
        - Wedding
        - Graduation
      - Holiday
        - Thanksgiving
        - Valentine
        - Christmas
        - New_Year_Eve
    - Time
    - Date
- STRING
- BOOLEAN

Event
Pictures
R006
FamilyAlbum

Restrictions Name	Property Name	Min	Max
R006	PicturePlace	1	1
R008	url	1	1
R0027	PictureDate	1	1
R0029	PictureDescrip...	1	1
R0035	PicturePersons	0	

Figure 1. A snapshot of creating image ontology FamilyAlbum

(http://www.cs.wayne.edu/~shiyong/ontology/FamilyAlbum.daml)

When the user enables the *Property* tab, similarly, the left frame shows the property hierarchy, in which parent-child relationship associates the *subPropertyOf* relations between properties. On the right frame, one can specify the type, domain, range of a property. In addition, one can relate a property to other properies in the fields of *InverseOf* and *SamePropertyAs*. Also note that, since a property might have multiple parents, other parent properties are shown in the *SubPropertyOf* field.

Creating restrictions is a part of the definition of a class. It creates an anonymous class. For example, we define a class *Pictures*. Every instance of a class *Pictures* must have a *PicturePlace*. In this case, we define a restriction. Each restriction must have a property called *onProperty*. In other words, that means the restriction is imposed on that property. We can also define a local range using *toClass*, and *hasClass*, and the number for range (*cardinality*, *minCardinality*, *maxCardinality*). The definition of qualification is a part of the restriction. It has *hasClassQ* and the number for range (*cardinalityQ, minCardinalityQ, maxCardinality*Q). Because restriction is an anonymous class, we represent a restriction with the relation (*subClassOf*, *complementOf*, *unionOf*, *disjoinWith*, *disjointUnionOf*, *sameClassAs*, *intersectionOf*) within the class. For example, when defining a class *Pictures*, *SubClassOf* field contains a restriction on the property *PictureDate*. Its range (*toClass*) is *dateTime* and the number for that range (*cardinality*) is 1. In order to keep the consistency of ontology, we check whether *maxCardinality* is larger than *minCardinality*. If we define the *toClass*, it should not have *hasClass* and qualification; the reverse should agree as well. *Birthday_Party* class (shown on figure 1) has inherited all restrictions from its parent *Pictures*.

The consistency of an ontology is essential and special cares must be taken in order to create a consistent ontology. For example, if Class A is specified as the parent class of Class B, then Class A cannot be in the *complementOf* class list of Class A. *ImageSpace* uses the following four mechanisms to ensure creating only consistent ontologies: (1) *No action.* If an insert, delete or update of a component will violate the consistency of the whole ontology, then the action is canceled with a warning given to the user to indicate the reason of such cancellation. (2) *Cascaded action.* When an offending action occurs, it triggers another or a series of other recovering actions to occur so that the consistency of the ontology is maintained. For example, when a class is deleted, then all references to the class will be deleted as well provided that such cascaded deletion will not cause inconsistency of the ontology. (3) *Using a filter.* To prevent consistency violating action from occurring, a filter is used to restriction the actions that a user can perform. For example, in the *disjointWith* field of a class, a filter is used so that no ancestor classes of this class can be chosen as a class in the *disjointWith* list. (4) *Validation before submission.* This mechanism is used, for example, in the *instance* interface. After an image is annotated, constraints such as cardinality constraints are checked, and if some inconsistencies occur, then the submission is cancelled, with an error message prompted to the user. The submission will not be committed until all constraints are satisfied. A detailed description of all the consistency checks that are performed by *ImageSpace* is beyond the scope of this paper. Interested readers are referred to [4] for details.

5. Ontology Visualization

Ontology visualization plays an important role in understanding and maintaining the structure of large knowledge bases. Ontology creation tools usually have many tabs and dialog windows, because of complex relationships and dependencies among classes, properties, and restrictions. As a result, one of the problems that users experience while navigating large ontologies is disorientation.

We have developed a tool for ontology visualization that can work as a stand alone application as well as an *ImageSpace* plugin. It provides simple and user-friendly interface for graphical navigation through ontology.

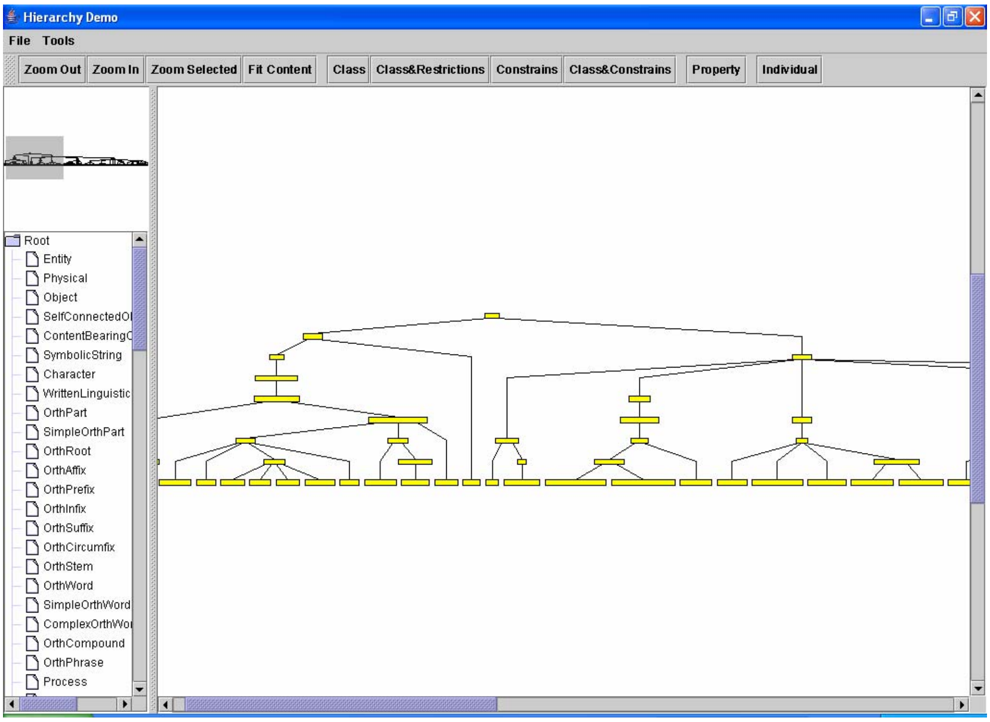


Figure 2. A snapshot of an ontology visualization (class view, hierarchical layout)

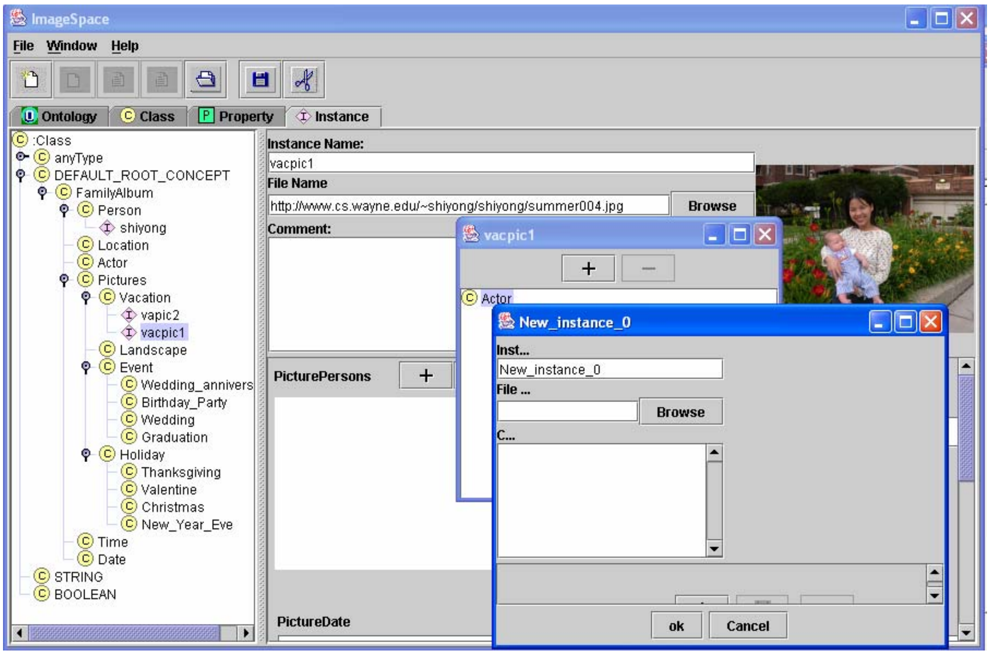
Figure 2 shows a snapshot of a sample ontology visualization. The tool main window has a menu, a toolbar, and 3 frames: left upper frame shows preview of a whole ontology graph; right frame shows main view of an ontology; left bottom frame shows a list of classes. A user can use all 3 frames to navigate an ontology.

The visualization plugin supports the following views/hierarchies: class; class and restrictions; constrains; class and constrains; property; and individual. Various concepts (class, property, individual) have different coloring scheme. In addition, a user can experiment with 3 highly customizable layouts: hierarchical, orthogonal, and organic. Figure 2 shows a class view of an ontology displayed with hierarchical layout.

Finally, we provide support of such common features like zooming (in, out, selected content, frame fitting) and manual layout of graphical primitives.

6. Annotating an Image

One attractive feature of *ImageSpace* is that, it nicely integrates annotation of images into one framework. The *Instance* tab corresponds to this functionality. Figure 3 displays a snapshot of annotating an image using *ImageSpace*. The left frame shows the class hierarchy and instances (shown by I-icons) associated with the classes to which they belong. The interface on the right frame is ontology-driven. In other words, for different ontologies and different classes, the interface will be generated dynamically based on the properties, cardinality constraints specified for the ontology. For example, for the *FamilyAlbum* ontology, the interface will contain fields *PicturePersons*, *PictureDate*, *PictureDescription* (hidden), etc. While *PictureDate* and *PictureDescription* are *DatatypeProperties*, *PicturePersons* is an *ObjectProperty* that relates an image to a list of actors. Here, an *actor* models a particular snapshot of a person in a particular picture. In the example given, there are two actors. The + button on the right of *PicturePersons* field allows a user to pop up a dialogue window to choose from a list of actors, in which the +/- buttons facilitates the insert/delete of actors in this list. This nested dialogue interface greatly facilitates a user to create instances in an on-the-demand fashion. For example, the insert of an actor might require a person to be inserted first, the nesting order of the dialogue windows ensures that a referenced instance is inserted before a referencing instance is inserted. In our example, the *FamilyAlbum* ontology will enable a user to model that two actors, say *Kathleen-actor1* and *Kevin-actor1*, exist in the picture, that these two actors are for persons *Kathleen* and *Kevin*, and *Kathleen-actor1* hugs *Kevin-actor1* in the picture. In this way, an intelligent semantic search such as “*return all the vacation pictures in which Kathleen hugs Kevin*” can be supported.



Instance Name: vacpic1

File Name: http://www.cs.wayne.edu/~shiyong/shiyong/summer004.jpg

Inst... New_instance_0

Figure 3. A snapshot of annotating an image

7. Storing Ontologies and Annotations in a Relational Database

Both ontologies and annotations are saved in a relational database for the support of ontology-driven search of images. We describe our database design in terms of the following tables that we create where primary keys are underlined:

- Ontology(OntologyID, versionInfo, comment)
- Import(OntologyID, importedOntologyID)
- Class(classID, ontologyID, type, label, comment)
- SubClassOf(classID, parentClassID)
- DisjointWith(classID, otherClassID)
- DisjointUnionOf(classID, otherClassID)
- UnionOf(classID, otherClassID)

- SameClassAs(classID, otherClassID)
- IntersectionOf(classID, otherClassID)
- ComplementOf(classID, otherClassID)
- OneOf(classID, instanceID)
- Property(propertyID, ontologyID, type, comment)
- SubPropertyOf(propertyID, parentPropertyID)
- PropertyDomain(propertyID, classID)
- PropertyRange(propertyID, classID)
- SamePropertyAs(propertyID, otherPropertyID)
- InserseOf(propertyID, classID)
- Restriction(restrictionID, onProp, toClass, minC, maxC, C)
- HasClass(restrictionID, classID)
- HasValue(restrictionID, value)
- HasClassQ(restrictionID, classID, minC, maxC, C)
- Instance(instanceID, classID)
- InstanceRelationship(instanceID, propertyID, value)
- DifferentInvividualFrom(instanceID, otherInstanceID)
- SameIndividualAs(instanceID, otherInstanceID)

As an example, consider an image where Kathleen smiles and hugs Kevin, and Kevin cries. An appropriate annotation can be stored in relational tables *Instance* and *InstanceRelationship* which are shown in table 1 and table 2 correspondingly. In practice, for efficiency concerns, we split *InstanceRelationship* table to set of tables with names that correspond to *propertyID* attribute value and with attributes *subject* (corresponds to *instanceID*) and *value*. Thus, the final schema for our example will contain the following tables (instead of *InstanceRelationship*):

- hasActor (subject, value)
- hugs (subject, value)
- hasAction (subject, value)
- hasName (subject, value)
- isSnapshotOf (subject, value)

Table 1. Relational table Instance

instanceID	classID
Kathleen	Person
Kevin	Person
http://www.cs.wayne.edu/example.jpg	Vacation
Kathleen-actor1	Actor
Kevin-actor1	Actor

Table 2. Relational table InstanceRelationship

instanceID	propertyID	value
http://www.cs.wayne.edu/example.jpg	hasActor	Kathleen-actor1
http://www.cs.wayne.edu/example.jpg	hasActor	Kevin-actor1
Kathleen-actor1	hugs	Kevin-actor1
Kathleen-actor1	hasAction	smiles
Kevin-actor1	hasAction	cries
Kathleen	hasName	Kathleen
Kevin	hasName	Kevin
Kathleen-actor1	isSnapshotOf	Kathleen
Kevin-actor1	isSnapshotOf	Kevin

8. Ontology-based Image Retrieval

Based on this database schema presented in the previous section, we have developed a simple web-based image retrieval system to search images. The system provides an interface to allow the user to navigate to images under different categories. In addition, a user can specify a list of “triples” as the search criterion to retrieve images. For example, one can specify a search criterion such as return all the images under the "vacation" category such that:

- Kathleen hugs Kevin, and
- Kathleen smiles, and
- Kevin cries.

The following datalog-style query will retrieve the needed photos where variables are prefixed by a '$':

```
Answer ($instanceID) :-
        instanceOf ($instanceID, Vacation),
        hasActor ($instanceID, $A1),
        hasActor ($instanceID, $A2),
        isSnapshotOf ($A1, $P1),
        isSnapshotOf ($A2, $P2),
        hasName ($P1, "Kathleen"),
        hasName ($P1, "Kevin"),
        hugs ($A1, $A2),
        hasAction ($A1, smiles),
        hasAction ($A2, cries).
```

Finally, query is translated to the following sequence of SQL statements:

- Select all actors for "Kathleen" and store them into *KathleenActor*.
  ```
  SELECT isSnapshotOf.subject
  FROM isSnapshotOf, hasName
  WHERE isSnapshotOf.value = hasName.subject AND hasName.value = 'Kathleen'
  ```
- Select all actors for "Kevin" and store them into *KevinActor*.
  ```
  SELECT isSnapshotOf.subject
  FROM isSnapshotOf, hasName
  WHERE isSnapshotOf.value = hasName.subject AND hasName.value = 'Kevin'
  ```
- Select all "smiling" actors for "Kathleen" and store them into *SmilingKathleenActor*.
  ```
  SELECT hasAction.subject
  FROM KathleenActor, hasAction
  WHERE KathleenActor.subject = hasAction.subject AND hasAction.value = 'smiles'
  ```
- Select all "crying" actors for "Kevin" and store them into *CryingKevinActor*.
  ```
  SELECT hasAction.subject
  FROM KevinActor, hasAction
  WHERE KevinActor.subject = hasAction.subject AND hasAction.value = 'cries'
  ```
- Retrieve all images that satisfy all specified conditions.
  ```
  SELECT H1.subject
  FROM hasActor H1, hasActor H2, Hugs
       SmilingKathleenActor, CryingKevinActor
  WHERE H1.subject = H2.subject AND
          H1.value = SmilingKathleenActor.subject AND
          H2.value = CryingKevinActor.subject AND
          Hugs.subject = SmilingKathleenActor.subject
          AND Hugs.value = CryingKevinActor.subject
  ```

All and only the images that satisfy this criterion will be returned (in our case, http://www.cs.wayne.edu/example.jpg). The reader is referred to [4] for more details about the ontology-driven image retrieval system.

9. Conclusions and Future Work

We have developed *ImageSpace*, an image ontology creation, visualization and annotation tool that fully supports the standard DAML+OIL ontology language and enables the storage of ontologies and annotations in a relational database. Future work includes:

- The development of *MultimediaSpace* that will not only support the annotation of images, but also other multimedia resources such as videos, audios, etc.
- Future version of *MultimediaSpace* will also support OWL, the successor of DAML+OIL.

- The development of graphical ontology building features to support by *MultimediaSpace* and visualization plug-in.
- Better optimization of SQL queries that are generated by image retrieval system.

Bibliography

[1] S. Lu, M. Dong and F. Fotouhi, “The Semantic Web: Opportunities and Challenges for Next-Generation Web Applications”, *International Journal of Information Research*, 7(4), 2002.

[2] F. Harmelen, P. Patel-Schneider and I. Horrocks, “Reference Description of the DAML+OIL Ontology Markup Language”, *http://www.daml.org/2001/03/reference*, March 2001.

[3] S. Bechhofer, F. Harmelen, J. Hendler, I. Horrocks, D. McGuinness, P. Patel-Schneider and L Stein, “OWL Web Ontology Language Reference”, *W3C Recommendation. http://www.w3.org/TR/owl-ref/.* February, 2004.

[4] R. Huang, “ImageSpace: A DAML+OIL Based Image Ontology Creation and Annotation Tool”, master thesis, advisor: Dr. Shiyong Lu, *Department of Computer Science, Wayne State University*. December 2003.

[5] T. Berners-Lee, J. Hendler, and O. Lassila. “The Semantic Web”, *Scientific American*. May 2001.

[6] Y. Rui, T. Huang, and S. Chang, “Image retrieval: current techniques, promising directions and open issues”, *Journal of Visual Communication and Image Representation*, Vol. 10, 39-62, March 1999.

[7] A. Schreiber, B. Dubbeldam, J. Wielemaker, and B. Wielinga, “Ontology-based Photo annotation”, *IEEE Intelligent Systems*, 16(3), pp. 66-74, 2001.

[8] A. Ponnusamy, C. Breen, L. Khan, and L. Wang, “Ontology-based image classification using neural networks”, *in SPIE: The International Society for Optical Engineering*, Boston, MA, USA, July 2002.

[9] Y. Sure, M. Erdmann, J. Angele, S. Staab, R. Studer, and D. Wenke, “OntoEdit: Collaborative Ontology Development for the Semantic Web”, *Proc. of the first International Semantic Web Conference 2002 (ISWC 2002)*, June 9-12 2002, Sardinia, Italia

[10] M. Storey, M. Musen, J. Silva, C. Best, N. Ernst, R. Fergerson and N. Noy, “Jambalaya: Interactive visualization to enhance ontology authoring and knowledge acquisition in Protégé”, *appeared in "Workshop on Interactive Tools for Knowledge Capture", K-CAP-2001*, October 20, 2001, Victoria, B.C. Canada.

[11] R. Huang, S. Lu and F. Fotouhi, "ImageSpace: An Image Ontology Creation and Annotation Tool", in *Proc. of the 19th International Conference on Computers and Their Applications (CATA'2004)*, pp. 340-343, Seattle, WA, USA, March, 2004.

[12] S. Lu, R. Huang, and F. Fotouhi, "Annotating Linguistic Data with ImageSpace for the Preservation of Endangered Languages", in *Proc. of the 19th International Conference on Computers and Their Applications (CATA'2004)*, pp. 193-196, Seattle, WA, USA, March, 2004.

[13] S. Lu, D. Liu, F. Fotouhi, M. Dong, R. Reynolds, A. Aristar, M. Ratliff, G. Nathan, J. Tan, and R. Powell, "Language Engineering For The Semantic Web: A Digital Library For Endangered Languages", *International Journal of Information Research*, vol.9, no.3, April 2004.

[14] T. Bray, J. Paoli, C. Sperberg-McQueen, E. Maler, and F. Yergeau, “Extensible Markup Language (XML) 1.0 (Third Edition) W3C Recommendation 04 February 2004”. 2004.

[15] D. Beckett and B. McBride, “RDF/XML Syntax Specification”, W3C Recommendation 10 February 2004.

[16] D. Brickley, R. Guha, and B. McBride, “RDF Vocabulary Description Language”, W3C Recommendation 10.02.2004.

[17] J. Stefan, R. Reynolds, F. Fotouhi, A. Aristar, S. Lu, and M. Dong, "Evolution Based Approaches to the Preservation of Endangered Natural Languages", in Proc. of the *IEEE International Congress on Evolutionary Computation*, pp. 1980-1987, Canberra, Australia, December, 2003.

[18] W. Grosky, F. Fotouhi, A. Aristar, S. Lu, M. Dong, and R. Reynolds, "A Digital Library for Endangered Languages", the *Nara Symposium for Digital Silk Roads (DSR)*, pp. 85-92, Nara, Japan, December, 2003.

Authors' Information

Shiyong Lu – e-mail: shiyong@cs.wayne.edu

Rong Huang – e-mail: f10272@cs.wayne.edu

Artem Chebotko – e-mail: artem@cs.wayne.edu

Yu Deng – e-mail: yudeng@cs.wayne.edu

Farshad Fotouhi – e-mail: fotouhi@cs.wayne.edu

Department of Computer Science, Wayne State University, Detroit, MI 48202, USA

WEB PAGE RETRIEVAL BY STRUCTURE

William Grosky and Gargee Deshpande

Abstract: Our research explores the possibility of categorizing webpages and webpage genre by structure or layout. Based on our results, we believe that webpage structure could play an important role, along with textual and visual keywords, in webpage categorization and searching.

Keywords: content-based retrieval, genre detection, layout ontologies.

1. Introduction

The amount of data available electronically on the web has increased dramatically in recent years. Users generally retrieve data by browsing and searching by keywords. This is an example of *content-based search*. In this approach, search is based on the words in the heading of the page or the contents of the images displayed on the web pages, or words occurring as meta-data in pages. The overwhelming amount of information on the web requires a powerful search service to render that information accessible and useful. Without such a search strategy, finding a specific web site can be as difficult as finding a book in a library that has no card catalogue and a completely random method of storing its books.

In recent years much research has been done on querying the web. In this research, the web is viewed as a collection of multimedia documents in the form of pages connected through hyperlinks. Unlike most web search engines, the aim here is to provide more database-like query functionality. Also, application of data mining techniques to the World Wide Web, referred to as *web mining*, has been the focus of several research projects and papers. Web mining has been categorized into *web content mining* and *web usage mining*. Web content mining is the process of finding information from the web, whereas web usage mining is the process of mining user browsing histories for access patterns [1].

We believe that it would also be desirable to see the layout of web pages when querying these pages and grouping them according to these layouts. The term *layout* connotes the spatial relationships between the page contents rendered by particular tags. Thus, web pages can be categorized according to their *layout ontology*. The term *ontology* means a specification of a conceptualization, a set of concept definitions. Broadly speaking, an ontology is a description (like a formal specification of a program) of concepts. Each web page has a structured hierarchy of tags that defines the layout ontology for that particular page. It is possible that two different web pages have a similar structure of their tag hierarchy. Then, the layout ontology of these two web pages is said to be the same. Our aim is to categorize web pages according to these structures. Our belief is that pages with similar layout ontologies have somewhat similar semantics, or at least can be categorized as belonging to the same environment. For example, we will present some preliminary experiments that show that pages from different newspapers are more similar in layout ontology to each other than to the layout ontology of commercial sites selling books. This concept can also be used for extracting data from the web depending upon content as well as the structure.

In this paper, we assume that each web page consists of HTML tags. HTML tags can be broadly categorized as *container tags* and *standalone tags*. Container tags can contain other tags inside them but standalone tags are atomic. Because of the containing capacity of some tags, each web page can be represented by a tree structure of its tags. Thus, for each web page, a tree structure of tags can be determined. Our hope is that we get a somewhat different tree structure of tags for each page from different environments.

The remainder of this paper is organized as follows. Section 2 briefly reviews the related literature. Section 3 covers various conceptual details, while Section 4 discusses implementation details and the various technologies used in our experiments. In Section 5, we give the results of some preliminary experiments, while Section 6 gives some concluding remarks.

2. Literature Review

Our hope is that by automatically characterizing the environment of a particular web page, content-based information can more easily be extracted from it. In [2], information from unstructured and semistructured web documents is retrieved from web pages in chunks called *records*. A record is a group of information relevant to some entity. The final goal is to extract information from these records to populate a relational database. The paper describes a heuristic approach to discovering the record boundaries in web documents. It captures the structure of a document as a tree of nested HTML tags and locates the sub-tree containing the records of interest, identifying candidate separator tags within the sub-tree using five independent heuristics, finally selecting a consensus separator tag based on a combined heuristic.

The five heuristics are OM (ontology matching), SD (standard deviation), IT (identifiable separator tags), HT (highest-count tags), and RP (repeating-tag pattern). Each of these heuristics returns one or more candidate separator tags with a measure of certainty attached to each candidate. Finally, they provide a way to combine these individual heuristics to determine a consensus separator tag and hence discover record boundaries.

The technique we exploit in this paper is based on the work of [3]. In this paper, a computational geometry-based spatial color indexing methodology is examined for efficient and effective image retrieval. In this scheme, an image is evenly divided into a number of M*N non-overlapping blocks, and each individual block is abstracted as a unique feature point labeled with its spatial location, its dominant hue and its dominant saturation. For each set of feature points labeled with the same hue or saturation, a Delaunay triangulation is constructed and then a feature point histogram is computed by discretizing and counting the angles produced by this triangulation. The concatenation of these feature-point histograms serves as the image index. This research work has been the motivation for our research.

Related research field to our approach is the research being done on semistructured data. For retrieving web pages by structure, structures of web pages have to be stored and retrieved effectively. For storing semistructured data, paper [4] argues that languages supporting deduction and object-orientation are particularly well-suited, as object-orientation provides a flexible common data model for handling semistructured data. Paper [5] presents the Lorel language designed for querying semistructured data. The main novelties of Lorel are that it makes extensive use of coercion to relieve the user from the strict typing of a query language, which is inappropriate for semistructured data, and that it provides powerful path expressions, which permit flexibility for declarative navigational access.

As against the data model that is underlying [5], [6] argues that semistructured data can be stored in relational format by exploiting the regularities inherent in existing semistructured data instances. The claim is that most of the data will be stored in relational format and future insertions can occur in a self-describing way. In [7], an approach of creating wrappers for storing semistructured data is discussed.

3. Web Page Retrieval by Structure

Motivated by the ultimate goal of automatically computing efficient and effective descriptors which symbolize web page structure, this research has been directed towards the management of information such as the levels of tags comprising a web page, the tag hierarchy, and the area covered by the tags on the web page. As nesting of tags plays important role in defining structure of the web page, dominance of tags is considered for each level.

Hope is that within broad domain of web pages this technique can be used to find the structure of web pages and categorize web pages according to the structure. Further to such categorized web pages, semistructured techniques can be applied for effective content retrieval.

The paper [2] has been the motivation behind this research. This paper examines the use of a computational geometry-based spatial color indexing methodology for efficient and effective image retrieval. In this scheme, an image is evenly divided into number of M*N non-overlapping blocks, and each individual block is abstracted as unique feature point labeled with its spatial location, dominant hue, and dominant saturation. For each set of feature points labeled with the same hue or saturation, a Delaunay triangulation is constructed, followed by computing a feature point histogram realized by discretizing and counting the angles produced by this triangulation. The concatenation of all these feature point histograms serves as the image index.

Following the same concept, we examine the use of a computational geometry-based web page structure analysis for effective web page structure matching. In this scheme, a web page is evenly divided into number of M*N non-overlapping blocks, and each individual block is abstracted as a unique tag that covers the maximum area in that block at its level. For each feature tag selected we get a set of feature points. For each set of feature points labelled with the same tag, we construct a Delaunay triangulation and then compute the feature point histogram as mentioned above. The concatenation of these feature-point histograms serves as our web page descriptor. Web page descriptors are further used to categorize different web pages.

As mentioned previously, we assume in this paper that each web page consists of HTML tags. HTML tags can be broadly categorized as *container tags* and *standalone tags*. Container tags can contain other tags inside them but standalone tags are by themselves. Examples of container tag are the TABLE tag and the PARAGRAPH (P) tag, while examples of standalone tags are the BASE tag and the AREA tag. Because of the containing capacity of the tags, a web page corresponds to a tag tree structure, called a *tag tree*. Not all web pages have similar tag trees. In this paper, we study page layouts to try to categorize web pages semantically.

For our analysis, the level of a tag plays an important role when finding tags covering the maximum area in a block. An example of a web page tag hierarchy is as follows:

```
<HTML>
       <HEAD>
              <TITLE>
              </TITLE>
       </HEAD>
       <BODY>
              <P>
                     <TABLE>
                            <TR>
                                   <TD>
                                   </TD>
                            </TR>
                     </TABLE>
                     <B>
                     <B>
              </P>
       </BODY>
</HTML>
```

In the web page example given above, the <HTML> tag is at the highest level. Nested in the <HTML> tag are tags <HEAD> and <BODY>. Inside the <BODY> tag is a <P> tag and inside the <P> tag is a <TABLE> tag and so on. When we consider the concept of area covered by a tag on a web page, the concept of level plays an important role. In the example given above, the level of tag <HTML> is 1, the level of the <BODY> tag is 2, the level of the tags <TABLE> and <B> are 3, the level of tag <TR> is 4, and so on. When calculating the dominant tag at level 3, both <TABLE> and <B> tags are analyzed to check which tag is covering the maximum area in which block on the page. As tag <TR> is inside the <TABLE> tag, the area covered by the <TABLE> tag on the web page contains the area covered by the <TR> tag. So, for the blocks in which the <TABLE> tag is dominant at level 3, it is possible that in this same block, tag <TR> is dominant at level 4.

Now, each web page consists of tag hierarchy. We consider a few tags as characterizing features F = $\{f_{1,\ldots,}f_k\}$. We believe that the spatial placement and dominance of these various feature tags can be used to characterize the web pages.

The web page is divided into N*M non-overlapping blocks. For each block, at each level, the tag covering the maximum area is found. Then we find for each of the predefined feature tags, which blocks that tag was marked as the predominant tag. We mark the center co-ordinate of all such blocks. The spatial arrangement of these points is an important aspect of our work. As mentioned earlier, we construct a Delaunay triangulation and then compute the feature point histogram by discretizing and counting the angles produced by this triangulation. The concatenation of these entire feature-point histograms serves as our web page descriptor.

It has been shown that histogram intersection is especially suited for comparing histograms for content-based retrieval. Additionally, histogram intersection is an efficient way of matching histograms. The intersection of the histograms W_{query} and $M_{database,}$ each of n bins, is defined as follows:

$$D(W_{query}, M_{database}) = (\sum \min(W_j, M_j)) / \sum W_j$$

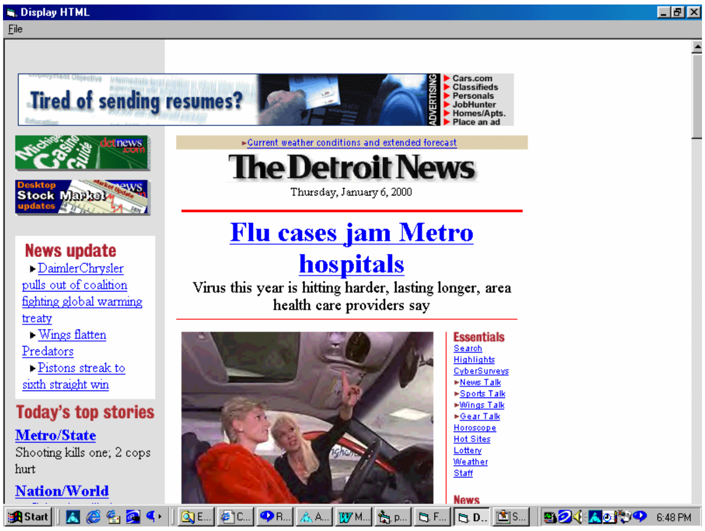
The histogram of a web page characterizes the web page depending upon the placement of tags forming the web page. Thus, the above mentioned formula can be used to check the similarity between the two web pages. If two web pages are similar in structure then the histogram of those two pages are bound to be similar. For such web pages, the above formula returns a value close to 1. Similarly, if two pages are very different in structure then the above formula returns a value close to 0.

4. Implementation

The input to our system is a web page. Our feature representation is extracted from this web page and matched against those extracted from other web page of known semantics. In more detail, we do the following:

1. Our system accepts a URL as input and displays the given web page using the Internet Explorer engine.
2. The web page displayed is analyzed to get all tags on the page with left, top, right, bottom (X1, Y1, X2, Y2) co-ordinates of area covered by each tag on the page. For each tag, the level of nesting is also saved while gathering this data.
3. The page is normalized to size 512 * 512. The original calculated co-ordinates (X1, Y1, X2, Y2) are re-calculated to map to this normalized size.
4. The page is divided into N*M disjoint blocks. The relevant coordinates of each block is calculated.
5. For each block, it is found out that which tag covers how much area.
6. Depending upon the data gathered in step 5, it is found out for each level, for each block, which tag is covers the maximum area.
7. For each of the selected feature tags, the blocks are found in which the tag covers the maximum area. Center X and center Y coordinates of these blocks are written to a file.
8. Histogram program is run on the file and histogram points calculated by the program are read back into the system. The histogram program used to calculate these points is implemented for the two largest angles of each Delauney triangle using 36 bins. Thus, each bin corresponds to 5 degrees.
9. For each web page, descriptor of (36 * Number of feature tags) bins is calculated.
10. When the descriptors of all the web pages of interest are calculated using steps 1 through 9, the distances between these pages and the database pages are calculated.
11. For each query page, the nearest pages are chosen, based on the distances calculated in step 10.
12. The web pages selected in step 11 are analyzed for category information. The most occurring category is chosen. The query page is categorized using this category.

As an example, consider the following web page:



Flu cases jam Metro hospitals

Virus this year is hitting harder, lasting longer, area health care providers say

News update

DaimlerChrysler pulls out of coalition fighting global warming treaty

Wings flatten Predators

Pistons streak to sixth straight win

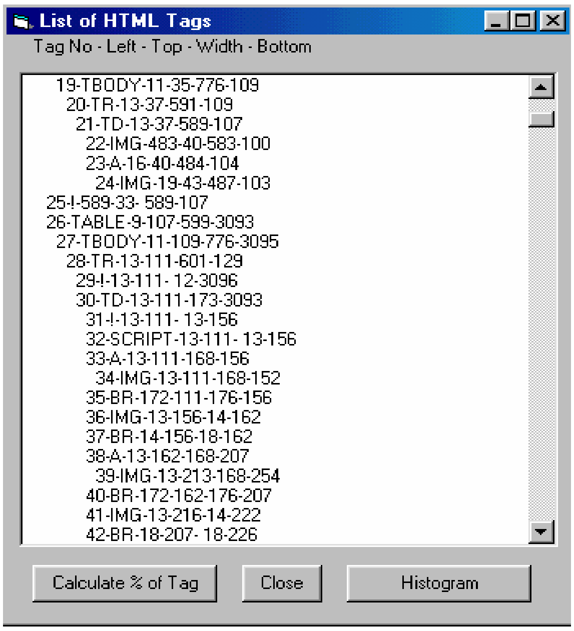
Today's top stories

Metro/State

Shooting kills one; 2 cops hurt

Nation/World

Here is a snapshot of part of the tag tree, along with the coordinates of the rectangular area covered by the rendering of each tag:

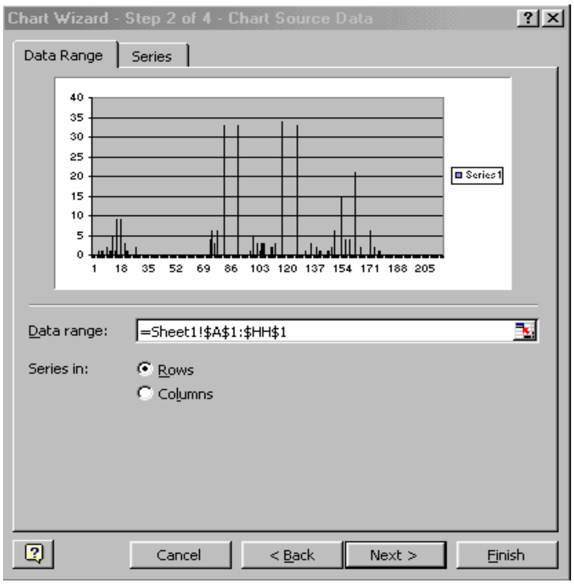


List of HTML Tags

Tag No - Left - Top - Width - Bottom

```
  19-TBODY-11-35-776-109
    20-TR-13-37-591-109
      21-TD-13-37-589-107
        22-IMG-483-40-583-100
        23-A-16-40-484-104
          24-IMG-19-43-487-103
25-!-589-33- 589-107
26-TABLE-9-107-599-3093
  27-TBODY-11-109-776-3095
    28-TR-13-111-601-129
      29-!-13-111- 12-3096
      30-TD-13-111-173-3093
        31-!-13-111- 13-156
        32-SCRIPT-13-111- 13-156
        33-A-13-111-168-156
          34-IMG-13-111-168-152
        35-BR-172-111-176-156
        36-IMG-13-156-14-162
        37-BR-14-156-18-162
        38-A-13-162-168-207
          39-IMG-13-213-168-254
        40-BR-172-162-176-207
        41-IMG-13-216-14-222
        42-BR-18-207- 18-226
```

And here is the resulting histogram for the web page:



Data range: =Sheet1!$A$1:$HH$1

Series in: ● Rows ○ Columns

5. Experimental Results

Our proof-of-concept experiments are carried out on newspaper web pages and e-commerce web pages. Four newspapers and two e-commerce web sites are selected as categories. The categories are: Detroit News,Times of India, Tribune India, Esakal, Amazon.com, and Buy.com.

For each of the newspaper categories, six days of newspaper front pages were analyzed, while from the e-commerce web sites, six web pages were used. Thus, a total of 36 web pages were analyzed.

Initially, we defined a large set of feature tags to ensure a powerful set of independent features for the discrimination of our two classes. This initial set of 52 feature tags were: <A>, <APPLET>, <B>, <BIG>, <BR>, <CAPTION>, <CENTER>, <CITE>, <CODE>, <COL>, <COLGROUP>, <DD>, <DIR>, <DL>, <DT>, <EM>, <FONT>, <FORM>, <H1>, <H2>, <H3>, <H4>, <H5>, <H6>, <HR>, <INPUT>, <LI>, <MENU>, <OBJECT>,

<OL>, <OPTION>, <P>, <PRE>, <SELECT>, <SMALL>, <STRONG>, <SUB>, <SUP>, <TABLE>, <TBODY>, <TD>, <TEXTAREA>,<TH>, <TITLE>, <TR>, <U>, <UL>, <FRAME>, <FRAMESET>, <IMG>, <MAP>, <AREA>.

We also conducted an experiment using a reduced set of tags. For each tag, we calculated a mean descriptor, by computing bin averages over all 36 web pages. We then calculated the deviation of each descriptor from its mean. We only kept those tags with high deviations, as these tags more easily discriminate among the various pages. The tags we kept for this experiment were <FONT>, <STRONG>, and <IMG>.

In all our experiments, we compared each individual web page, using the nearest neighbour approach, to the 35 remaining pages, using both sets of tags. We tried to determine both individualized categories as well as genre categories. The former takes a match as successful only if the two pages came from the same site, while the latter takes a match as successful only if the two pages came from the same genre: newspaper versus e-commerce. Here is the table of our results.

	Individualized Categories		Genre Categories	
	Matches	Failures	Matches	Failures
52 tags	26	10	33	3
3 tags	27	9	33	3

Based on these initial results, it seems that our technique has promise for genre detection.

6. Conclusions

The aim of this research was to analyze the possibility of categorizing webpages and webpage genre by structure or layout. The original insight comes from the fact that many newspaper sites, say, have the same look and feel. Based on our results, we believe that structure could play an important role, along with textual and visual keywords, in webpage categorization and searching.

Bibliography

[1] R. Cooley, B. Mobasher, and J. Srivastava, 'Web Mining: Information and Pattern Discovery on the World Wide Web,' Proceedings *of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97),* November 1997.

[2] D.W. Embley, Y. Jiang, and Y.K. Ng, 'Record – Boundary Discovery in Web Documents,' *Proceedings of the ACM SIGMOD Conference,* 1999, pp. 467-478.

[3] Y. Tao and W.I. Grosky, 'Spatial Color Indexing Using Rotation, Translation, and Scale Invariant Anglograms,' *Multimedia Tools and Applications,* Volume 15, Number 3 (December 2001), pp. 247-268.

[4] B. Ludäscher, R. Himmeröder, G. Lausen, W. May, and C. Schlepphorst, 'Managing Semistructured Data with FLORID : A Deductive Object-Oriented Perspective,' *Information Systems,* Volume 23, Number 8 (1998), pp. 589-613.

[5] S. Abiteboul, D. Quass, J. McHugh, J. Widom, and J.L. Wiener, 'The Lorel Query Language for Semistructured Data,' *International Journal on Digital Libraries,* Volume 1, Number 1 (April 1997), pp. 68-88.

[6] A. Deutsch, M. Fernandez, and D. Suciu, 'Storing Semistructured data in Relations,' *Proceedings of the Workshop on Query processing for Semistructured data and Non-standard Data Formats,* Jerusalem, Israel, January, 1999.

[7] N. Ashish and C. Knobolk, 'Wrapper Generation for Semi-Structured Internet Sources,' SIGMOD Record, Volume 26, Number 4 (December 1997), pp. 8-15.

Authors' Information

William I. Grosky – University of Michigan-Dearborn, Computer and Information Science Department, 4901 Evergreen Road, Dearborn, Michigan 48128, USA; email: wgrosky@umich.edu

Gargee Deshpande – Wayne State University, Computer Science Department, Detroit, Michigan 48202, USA

AN OPTIMAL DISTRIBUTED ALGORITHM FOR ALL-PAIRS SHORTEST-PATH

Saroja Kanchi and David Vineyard

Abstract: In this paper the network problem of determining all-pairs shortest-path is examined. A distributed algorithm which runs in O(n) time on a network of n nodes is presented. The number of messages of the algorithm is O(e+n log n) where e is the number of communication links of the network. We prove that this algorithm is time optimal.

Keywords: distributed algorithm, all-pairs shortest-path, computer network.

Introduction

In this paper we examine the distributed all-pairs shortest-path problem. The all-pairs shortest-path problem is the problem in which the shortest path between every pair of nodes in a network is determined. In the distributed version of the problem, a distributed algorithm is sought such that at the termination of the algorithm, every node knows the shortest path between any two nodes of the network. Floyd published a centralized algorithm [Floyd, 1962], which has been converted into a distributed algorithm by Toueg [Toueg, 1980]. The time complexity of this algorithm is $O(n^2)$ [Toueg, 1980].

The distributed shortest path problem and its variations have been studied because of its many applications. A decentralized algorithm for finding shortest paths in a network was presented by Abraham and Rhodes [Abram, 1978]. A distributed algorithm for finding shortest distances in an undirected graph was presented by Ravichandran, et. al. [Ravichandran, 1986] in which at the termination of the algorithm, each node contains the shortest path between itself and all other nodes. The algorithm described by Chandry and Misra [Chandry, 1982] finds shortest path from a node *i* to node *j* in a directed graph.

Determining topological properties of a network by distributed computation have received considerable attention. A number of papers have covered the topic of finding a minimum weight spanning tree [Awerbuch, 1987], [Korach, 1984], [Garay, 1998]. The problems of leader election, counting, and related problems [Awerbuch, 1987], [Singh, 1995], [Korach, 1984], [Kutten, 1998], [Kanchi, 1993] have also been studied.

In this paper we use the solution for finding a minimum weight spanning tree for finding a time optimal algorithm for the distributed all-pairs shortest-path problem.

There has been no previous distributed algorithm to find the all-pairs shortest-path in a general graph, other than the distributed version of a centralized algorithm given by Floyd [Floyd, 1962]. Therefore the idea of using a spanning tree and the center of the tree to find all-pairs shortest paths is a new element in this algorithm.

Model

The distributed network is considered to be an undirected weighted communications graph $G=(V,E)$, with processors forming the nodes, V, and bidirectional weighted communication links between processors forming the edges, E of the graph. No processor knows the topology of the network. No common memory is shared between processors and there is no global clock. All processors have unique identities from a totally ordered set. No processor knows the identity of any other processor. Each processor knows the links incident to it. For the duration of the algorithm, the network is assumed to be reliable, i.e., there will be no failures of the nodes or links for the duration of the algorithm.

The local computation at any node is assumed to take negligible time compared to the time required to transmit a message along a link. The asynchronous nature of the network permits undetermined communication delays in the delivery of a message. However, for the purpose of determining the time complexity, we assume that each message is delivered in $O(1)$ time along a link, irrespective of the size of the message. The correctness of the algorithm does not depend on this assumption.

The algorithm we present does not depend on any initiator node(s). At any time, one or more nodes may wake up and begin the execution of the algorithm. At the end of the algorithm, all nodes know the shortest path between any two nodes of the network. This data is stored in a square matrix, *D*, where entry *(i,j)* contains the shortest path between the nodes *i* and *j*.

Given any spanning tree *T* of a graph *G*, the edges that are not in *T* are called the *co-tree edges* with respect to *T*.

The size of the set *V* is denoted by *n*. The size of the set *E* is denoted by *e*.

Informal Description of the Algorithm

In this section we describe the algorithm at a high level. The algorithm consists of four steps as described below.

Step I: Finding a spanning tree, *T*, of the weighted graph, *G*:

Initially, all nodes are *inactive*. The first major part of the algorithm is to find a spanning tree, *T*, of the underlying unweighted graph. This can be accomplished by any one of the spanning tree finding algorithms, and we use the algorithm given by Awerbuch [Awerbuch, 1987], which takes time *O(n)*.

The spanning tree algorithm ensures every node can identify the links incident on it as either an edge in the tree *T* or a co-tree edge with respect to *T*.

Step II: Each node determines the identities of its neighbors in the graph *G*:

Each node must determine the identities of its neighbors in graph G. This can be accomplished by each node sending its identity along each link incident to it. The time complexity of this step is *O(1)*. Since each link carries exactly two messages, one from each of the incident nodes to the link, the number of messages is *2e*.

Step III: Determination of the All-Pairs Shortest-Distance matrix *D*:

This step of the algorithm deals with the transmission of distance information in *G* along the tree edges of *T*. Initially, each vertex constructs a local distance matrix that has row and column labels corresponding to the vertex and its neighbors.

Starting at each leaf node, partial distance information is transmitted along the tree edges of *T*. Whenever the partial distance matrix of a neighbor is received at a non-leaf node, new columns and rows are added to the partial distance matrix of that node and existing distance data is updated. When a non-leaf node receives partial distance matrix information from all but one of its neighbors, it becomes a transmitting node and sends its partial distance matrix to the neighbor from which it did not receive a partial distance matrix message.

At the end of this step, exactly one or two nodes, called the saturated node(s) of the tree, would contain Shortest-Distance matrix, *D*, of the entire graph *G*. We will show that the time complexity of this step is *O(n)*.

Step IV: Communicating the All-Pairs Shortest-Distance matrix *D* to every node:

This communication originates at the one or two nodes that are described in Step 3, and messages travel using only tree edges of *T*. This step has complexities of *O(n)* time and *O(n)* number of messages.

Notation Used in the Algorithm

Messages transmitted in this algorithm are of the following three types:

IDENTIFICATION: This type of message is used in Step II, where each node transmits its unique identity to each of its neighbors in the graph *G*.

PARTIAL DISTANCE MATRIX: This type of message in used in Step III, where the partial distance matrix calculated locally at a given node is sent along a single tree edge.

FINAL DISTANCE MATRIX: This type of message is used in Step IV, where the final distance matrix is sent to all the tree neighbors.

The nodes are in one of four states throughout the execution of the algorithm.

INACTIVE: Nodes are in Inactive state prior to the start of the algorithm. Initially all nodes are Inactive.

RECEIVING: Any non-leaf node that is receiving and processing partial distance matrices from other nodes is said to be in Receiving state. A node in Receiving state has not yet transmitted its partial distance matrix.

TRANSMITTING: A node is in Transmitting state if it has received partial distance matrices from all but one of its neighbors (this is trivially true for a leaf node). A node in Transmitting state sends its updated partial distance matrix to one other node from which it did not receive a partial distance matrix.

SATURATED: A node is in Saturated state if has received partial distance matrices from all its neighbors in the tree *T*.

Algorithm

In this section we describe the distributed algorithm for finding the all-pairs shortest-distance matrix.

ALGORITHM (ALL-PAIRS SHORTEST-PATH ALGORITHM)

1. Every node sets its state to Inactive.
2. Construct a spanning tree, *T* of the underlying unweighted graph. Any good asynchronous spanning tree algorithm can be used. The only modification to the spanning tree algorithm, which is required for our algorithm, is that any node with a single neighbor in the tree (a leaf node) must change its state to Transmitting at the end of the spanning tree algorithm. Similarly, any node with more than one neighbor in the tree (an interior node) must change its state to Receiving.
3. Each node *i* determines the identities of its neighbors in *G* and stores identity and distance data in a matrix PD_i. For instance, a node *i* that is adjacent to nodes *j* and *k*, creates entries *(i,j)*, *(j,k)* and *(i,k)* in PD_i. The value of $PD_i[i,j]$, $PD_i[i,k]$ would be the weights of the edges *(i,j)* and *(i,k)* respectively, and the value of $PD_i[j,k]$ would be the sum of weights of the edges *(i,j)* and *(i,k)*. See INITIALIZE_PARTIAL_DISTANCE_MATRIX subroutine below.
4. Determine All-Pairs Shortest Distance Matrix *D* of the graph *G*. Each node's behavior is determined by its state.

```
For each node $i \in V$
    If the state of $i$ is Receiving
        Run the subroutine RECEIVING_NODE_PROCESSING($i$);
    If the state of $i$ is Transmitting
        Run the subroutine TRANSMITTING_NODE_PROCESSING($i$)
```

   As a result, at most 2 transmitting nodes will receive a message from all neighbors and are marked Saturated.
5. Transmit the final All-Pairs Shortest-Distance matrix to every node from a Saturated node. Any Saturated node contains the final all pairs shortest distance matrix *D*. The Saturated node(s) will create a final message consisting of *D* and send this message to all its neighbors in the spanning tree *T*. Any node in the spanning tree that receives *D* will store *D* locally and send *D* to all its tree neighbors except the tree neighbor from which it received *D*.

SUBROUTINE (INITIALIZE_PARTIAL_DISTANCE_MATRIX)

```
1.   For each node $i \in V$
2.        $i$ transmits an Identification message containing its identity along each edge incident at $i$ in $G$
3.        $i$, upon receiving the identities of its $m$ neighbors, creates a distance matrix, $PD_i$, of size
          $(m+1) \times (m+1)$ and assigns the values to $PD_i[j,k]$ as given below.
     3.1. For each $j$, $k \in$ indexes of $PD_i$
     3.2.      If $j == k$ then $PD_i[j,k] \leftarrow 0$.
     3.3.      If $j == i$ or $k == i$ then $PD_i[j,k] \leftarrow$ weight of the edge between $j$ and $k$.
     3.4.      Else $PD_i[j,k] \leftarrow PD_i[j,i] + PD_i[i,k]$.
     3.5. EndFor
```

SUBROUTINE (RECEIVING_NODE_PROCESSING(*i*))

```
1.   Let $Tnbr_i$ be the set of neighbors of node $i$ in Tree $T$ created in Step 2 of the all-pairs shortest-path
     algorithm.
2.   Let $count$ be the number of the partial distance matrices that $i$ has received since it changed state to
     Receiving. Initially count is set to 0.
3.   Let Links_Used be a vector of size $|Tnbr_i|$ of type boolean in which all entries are initialized to False.
4.   While count < $|Tnbr_i| - 1$
5.        Receive message $PD_j$ from neighbor $j$
6.        count++
7.        Link_Used[$j$] $\leftarrow$True
8.        Call ProcessMessage($PD_j$)
9.   EndWhile
10.  Set the state of node $i$ to Transmitting.
```

SUBROUTINE (PROCESSMESSAGE(*PD_j*))

```
1.   For each index k in $PD_j$
2.        if k is not an index of $PD_i$
3.             extend $PD_i$ by one row and one column corresponding to k
4.             For all indexes $m$ in $PD_i$
5.                  Set $PD_i[k, m] \leftarrow PD_i[m, k] \leftarrow \infty$
6.             EndFor
7.             Set $PD_i[k, k] \leftarrow 0$
8.        EndIf
9.   EndFor
10.  For each $k$, $m \in$ indexes of $PD_j$
11.       if $PD_i[k,m] > PD_j[k,m]$
12.            $PD_i[k,m] \leftarrow PD_j[k,m]$
13.  EndFor
14.  For each $k$, $m$, $n \in$ indexes of $PD_i$
15.       if $PD_i[k,m] > PD_i[k,n] + PD_i[n,m]$
16.            $PD_i[k,m] \leftarrow PD_i[k,n] + PD_i[n,m]$
17.  EndFor
```

SUBROUTINE (TRANSMITTING_NODE_PROCESSING(*i*))

```
1.   Node i transmits $PD_i$ to its only neighbor in $T$ from which it has not received a partial distance matrix.
2.   If $i$ receives another partial distance message, say from $j$, then $i$ calls ProcessMessage($PD_j$) and marks
     itself as Saturated.
```

Correctness

In this section we show that the All-Pairs Shortest-Path algorithm produces the correct result.

Lemma 1 There are at most two Saturated nodes.

PROOF: The algorithm starts at leaf nodes of the tree, and matrices are transmitted to internal nodes. Each internal node, in turn chooses the one node from which it has not received any partial distance matrix as its parent and transmits the partial distance matrix to that node. In this manner eventually the matrices reach the one or two centers of the tree. These centers become the Saturated nodes.

Lemma 2 The shortest distance between any two nodes is known to a Saturated node.

PROOF: We will prove this using induction on the number of edges in the shortest path. Any shortest path consisting of a single edge is known to the Saturated node(s), since every node, by Step 3, creates a partial distance matrix and all these matrices are transmitted eventually to the Saturated node(s).

Assume that if there are fewer than *k* edges in the shortest path between two nodes, then that path is known to the Saturated node(s). Consider two nodes *x* and *y* such that the shortest path *P* between *x* and *y* has *k* edges. Let $P = (x = v_0, v_1, v_2, ..., v_{k-1}, v_k = y)$. Assume that the Saturated node(s) contains a ``path'' *P*' between *x* and *y*, but that the sum of the edge weights of *P*' is greater than the sum of the edge weights of *P*. Then the two paths must differ in at least one edge. Let (v_i, v_{i+1}) be the first edge in *P* that is not in *P*'. Note that v_i could be the same as *x* or v_{i+1} could be same as *y*. But since *P* is the shortest path from *x* to *y*, the path $(x, v_1, v_2, \ldots, v_i)$ is a shortest path from *x* to v_i. Similarly, the path $(v_{i+1}, v_{i+2}, \ldots, v_{k-1}, y)$ is a shortest path from v_{i+1} to *y*. Note that these paths must contain fewer than k edges, since *P* has *k* edges. But by the induction hypotheses the Saturated node contains the shortest path from *x* to v_i and from v_{i+1} to *y* since the number of edges in each of these shortest paths is less than *k*. Also, by Step 4 of the all-pairs shortest-path algorithm, Process_Message combines these two shortest paths to obtain the shortest path between *x* and *y*. Therefore the Saturated node must have the path *P*.

Theorem 1 The All-Pairs Shortest-Path Algorithm guarantees that all nodes in *G* know the all-pairs shortest-paths.

PROOF: By Lemma 1, there are exactly one or two Saturated nodes. By Lemma 2, a Saturated node knows the all-pairs shortest-path matrix *D*. Step 5 of the algorithm is a broadcast of this information to all nodes in the spanning tree, hence in the graph.

Complexity

In this section, we show that Algorithm 1 takes *O(n)* time and *O(e + n log n)* number of messages. Note that the subroutines Initialize_Partial_Distance_Matrix, Receiving_Node_Processing(i), ProcessMessage, and Transmitting_Node_Processing each perform local processing and are thus considered to take *O(1)* time.

Theorem 2 The all-pairs shortest-path algorithm terminates in *O(n)* time.

PROOF: Step 1 of the algorithm takes *O(1)* time. Step 2 of the algorithm, i.e., constructing the spanning tree, takes *O(n)* time. [Awerbuch, 1987]. Step 3 of the algorithm takes *O(1)* time, since each node sends one message on each tree link. Step 4 of the algorithm takes time proportional to the height of the tree with a Saturated node as a root. This is at most *O(n)*. Step 5 takes the same time as Step 4 since the messages travel from the root to the leaves of the tree. The time complexity of the algorithm is dominated by Step 2, and is thus *O(n)*.

Theorem 3 The all-pairs shortest-path algorithm has *O(e + n log n)* bound on the number of messages.

PROOF: The number of messages in Step 2 of the algorithm is *O(e + n log n)* [Awerbuch, 1987]. The number of messages in Step 3 of the algorithm is 2*e* since each edge is used for exactly two IDENTIFICATION messages. The number of messages in Step 4 of the algorithm is *O(n)* because the partial distance matrices are transmitted from leaf nodes to the root of the tree (Saturated node) using only edges of *T*. The spanning tree has *n*-1 edges and exactly one message is sent along each tree edge, thus the number of messages is *O(n)*. Note that if there are two Saturated nodes, the edge between them is used twice. Similarly, the number of messages in Step 5 of the algorithm is *O(n)*. Therefore the number of messages generated by the algorithm is bounded by

O(e +n log n).

Optimality

We claim that our distributed algorithm is time optimal for finding all-pairs shortest-path. This follows since a solution to the leader election problem can be obtained from a solution to the all-pairs shortest-path problem with no additional communication time. For instance, each node can locally elect the node with the highest identity as the leader. Since the time optimal leader election algorithm [Awerbuch, 1987] takes *O(n)* time, our *O(n)* time algorithm for all-pairs shortest-path is also time optimal.

Conclusion

We have developed a distributed algorithm for the all-pairs shortest-path problem which is optimal in time and number of messages. The optimal time is *O(n)*. The optimal number of messages is *O(e + n log n)*.

Bibliography

[Abram, 1978] J.M. Abram and I.B. Rhodes, *A decentralized shortest path algorithm* in Proc. of the 16th Allerton Conf. on Communication, Control, and Computing (Monticello, Ill.), pp. 271-277, 1978

[Awerbuch, 1987] B. Awerbuch, *Optimal distributed algorithms for minimum-weight spanning tree, counting, leader election and related problems*, in Proc. 19th ACM Symp. on Theory of Computing, ACM, New York, pp. 230-240, 1987

[Chandry, 1982] K.M. Chandry and J. Misra*, Distributed computation on graphs: shortest path algorithms*, Comm. ACM 25, pp. 833-837, Nov. 1982

[Floyd, 1962] R. Floyd, *Algorithm 97: shortest path*, Comm. ACM 5, 1962

[Garay, 1998] J. Garay, S. Kutten, and D. Peleg, *A sublinear time distributed algorithm for minimum-weight spanning trees*, SIAM J. Comput., Vol. 27, No. 1, pp. 302-316, February 1998

[Kanchi 1993] S.P. Kanchi and J.L. Kim, *Alternate algorithms for leader election on reliable and unreliable complete networks*, Proc. of the sixth international conf. on parallel and distributed computing and systems p.118-121, October 1993

[Korach, 1984] E. Korach, S. Moran, and S. Zaks, *Tight lower and upper bounds for some distributed algorithms for a complete network of processors*, Proc. of 1985 PODC Conf., Vancouver, BC, pp. 199-207, August 1984

[Kutten, 1998] S. Kutten and D. Peleg, *Fast distributed construction of small k-dominating sets and applications*, Journal of Algorithms 28, pp. 40-66, 1998

[Ravichandran, 1986] A. Ravichandran, S.G. Menon, and R.K. Shyamasundar, *A distributed algorithm for finding the shortest paths in an undirected graph*, Technical Report CS-86-13, Department of Computer Science, Pennsylvania State University, May 1986

[Singh, 1995] G. Singh and A. Bernstein, *A highly asynchronous minimum spanning tree protocol*, Distrib. Comput., pp 151-161, 1995

[Toueg, 1980] S. Toueg, *An all-pairs shortest-path distributed algorithm*, Res. Rep. RC-8327, IBM Thomas J. Watson Research Center, Yorktown Heights, N.Y., 1980

Authors' Information

Saroja Kanchi – Department of Science and Mathematics, Kettering University, 1700 West Third Avenue, Flint, Michigan 48504-4898, USA: skanchi@kettering.edu

David Vineyard – Department of Science and Mathematics, Kettering University, 1700 West Third Avenue, Flint, Michigan 48504-4898, USA: dvineyar@kettering.edu

A TWO LAYERED MODEL FOR EVOLVING WEB RESOURCES

Alfredo Milani and Silvia Suriani

Abstract: In this paper the key features of a two-layered model for describing the semantic of dynamical web resources are introduced.

In the current Semantic Web proposal [Berners-Lee et al., 2001] web resources are classified into static ontologies which describes the semantic network of their inter-relationships [Kalianpur, 2001][Handschuh & Staab, 2002] and complex constraints described by logical quantified formula [Boley et al., 2001][McGuinnes & van Harmelen, 2004][McGuinnes et al., 2004], the basic idea is that software agents can use techniques of automatic reasoning in order to relate resources and to support sophisticated web application.

On the other hand, web resources are also characterized by their dynamical aspects, which are not adequately addressed by current web models.

Resources on the web are dynamical since, in the minimal case, they can appear or disappear from the web and their content is upgraded. In addition, resources can traverse different states, which characterized the resource life-cycle, each resource state corresponding to different possible uses of the resource. Finally most resources are timed, i.e. they information they provide make sense only if contextualised with respect to time, and their validity and accuracy is greatly bounded by time.

Temporal projection and deduction based on dynamical and time constraints of the resources can be made and exploited by software agents [Hendler, 2001] in order to make previsions about the availability and the state of a resource, for deciding when consulting the resource itself or in order to deliberately induce a resource state change for reaching some agent goal, such as in the automated planning framework [Fikes & Nilsson, 1971][Bacchus & Kabanza,1998].

Keywords: Temporal Resources, Dynamic Web, evolutionary resources

Introduction

The basic notion of resource in the semantic web [Berners-Lee et al., 2001] is characterised by a unity of structure, content, and location, i.e. a resource has a structure, which is defined in the ontology, a content, i.e. the actual values of their properties, and a unique location, i.e. an URI [Berners-Lee & Fielding,1998], which uniquely identified it in term of its web location.

In our model a resource is still representing a single entity, but entities can evolve over time with respect to the current value of their contents, and also in their structural and semantic description, in other words, the notion of a resource can be intuitively intended as *the invariant aspects with respect to time of a given URI*. For example our department web page is the same resource, despite of the fact that it is continuously updated, in the content and in the structure since our web server was established in 1995.

Resources can be dated and resources can be updated. For many type of resources it is possible to specify when the information will be update, moreover the resource timestamp also provide a relevant information about its validity.

Consider for example:

a) a web page about the history of the independence war,
b) a personal CV,
c) the news of an online newspaper, and
d) stock exchange prices,

they are all web entities with a different rate of upgrade.

The advantages in explicitly defining the date/update features of a resource are apparent with respect to the trust/validity of the information provided by the resource.

Moreover consider for example a) with respect to b), and assume that these two web resources are not updated since the two years ago. It is clear that the info in a) can be used in any moment (assumed that the source is trustworthy), since we do not expect big new facts about the Independence War, on the other hand, discovering that the personal CV was not updated since two years ago, make this information not valid, thus an hypothetical software agent looking for employee information can decide to look for another CV of the same person or to ask the person to provide an upgraded copy.

Update rate can be estimated for c) and d), online newspapers and stock exchange prices are update at different pace, in the first case the content and structure can completely change after some hours, while in the latter the actual value of the price is the only thing which is likely to change, very rapidly when the stock market is open, and to remain still until next opening, during stock market closing hours. A software agent can exploit this information for its cognitive purposes by browsing the online newspaper by the hours or by the week (e.g. sport news about football matches) and the stock prices by the minutes.

Resource States

The state of a resource is an abstract characterisation of structural properties and actual values, which significantly characterise the resource. Associated with the state, there is the possibility that the resource evolves over time by moving from one state to another, in a transition path, which describes the dynamic evolution of the resource.

In the first instance a resource state is an ontological category which is simply characterised by logical constraints about the values of structural properties provided by the ontology, i.e. different ontological concept which share the same schema but not the same actual values. For example FatMan and SlimMan are instances of the concept of Man, they can be defined by constraints over the values of the properties Weight and Height, the interesting aspect is that the FatMan and SlimMan has a dynamical relationship since an individual can move from one state of another by upgrading its weight, (and less probably its height).

In the most general case, resources can allow structural properties to change, i.e. while moving from one state to another the resource evolve its schema.

It is straightforward to represent the admissible states, and the admissible state transitions of a given resource by a labelled transition diagram in which the label represent conditions or event over web resources or time which trigger the state transition.

Def. LTD for Web resources. A labelled transition diagram for a web resource it is a pair {N, δ}

Where N is a set of nodes representing the states of the web resource, and δ represents the labelled arcs of the diagram, i.e. the state transition function which defines for every pair of nodes $n_1,n_2 \in N$ a condition L, which labels the arc (n_1,n_2), condition L is a condition over web resources (static and dynamic items, operations, web services, conditions over property values etc.) and time conditions (i.e. current date or general date/time functions).

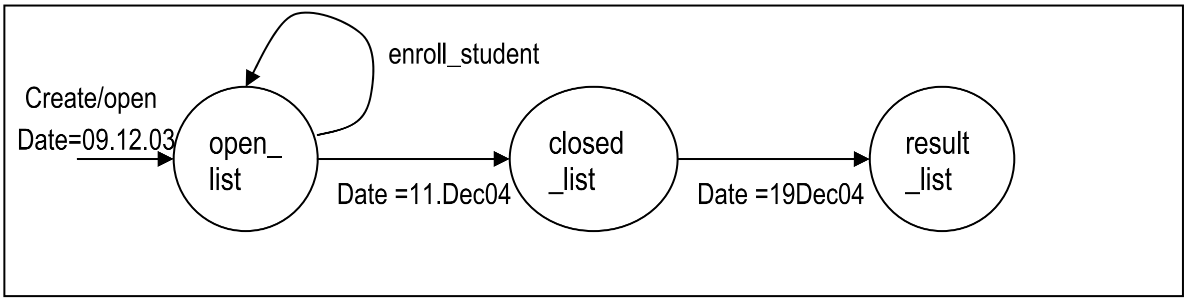


Fig.1 Students List State Transition Diagram

The labelled arc (L, n_1,n_2) denotes the fact that a resource can move from state n_1 to state n_2 when condition L is met, i.e. when the guard is verified, the logical conditions are true or the specified events take place.

By convention (false, n_1,n_2) denotes the fact that no transition is possible between state n_1 and state n_2.

A self-reference loop will represent a resource update; i.e. the resource is remaining in the same state while possibly changing its informative content.

Consider for example the online student enrolment list, exam_list, for the exam code 503 Programming Languages Course which will be held on 12th December 2004, this web resource it is continuously updated since its opening time 3 days before the exam and it is closed the day before and finally it is updated one more time with the list of candidates grades one week after the exam.

The exam_list it is an individual entity despite of the upgrade operations, which are operated on it.

In term of state transitions the evolution of the list can be represented by the state transition network in the figure 1. It is worth noticing that the self-reference loop labelled enroll_student represent the fact that after an enroll_student event.

In this framework the dynamics of resources are represented by appropriate state transition diagrams, which model the resource lifecycle.

A Two Layer Model for Dynamical Web

In order to give an account of the static and dynamic relationships of web resources a two-layer model architecture is proposed in which state transition diagrams are defined over a given ontological network.

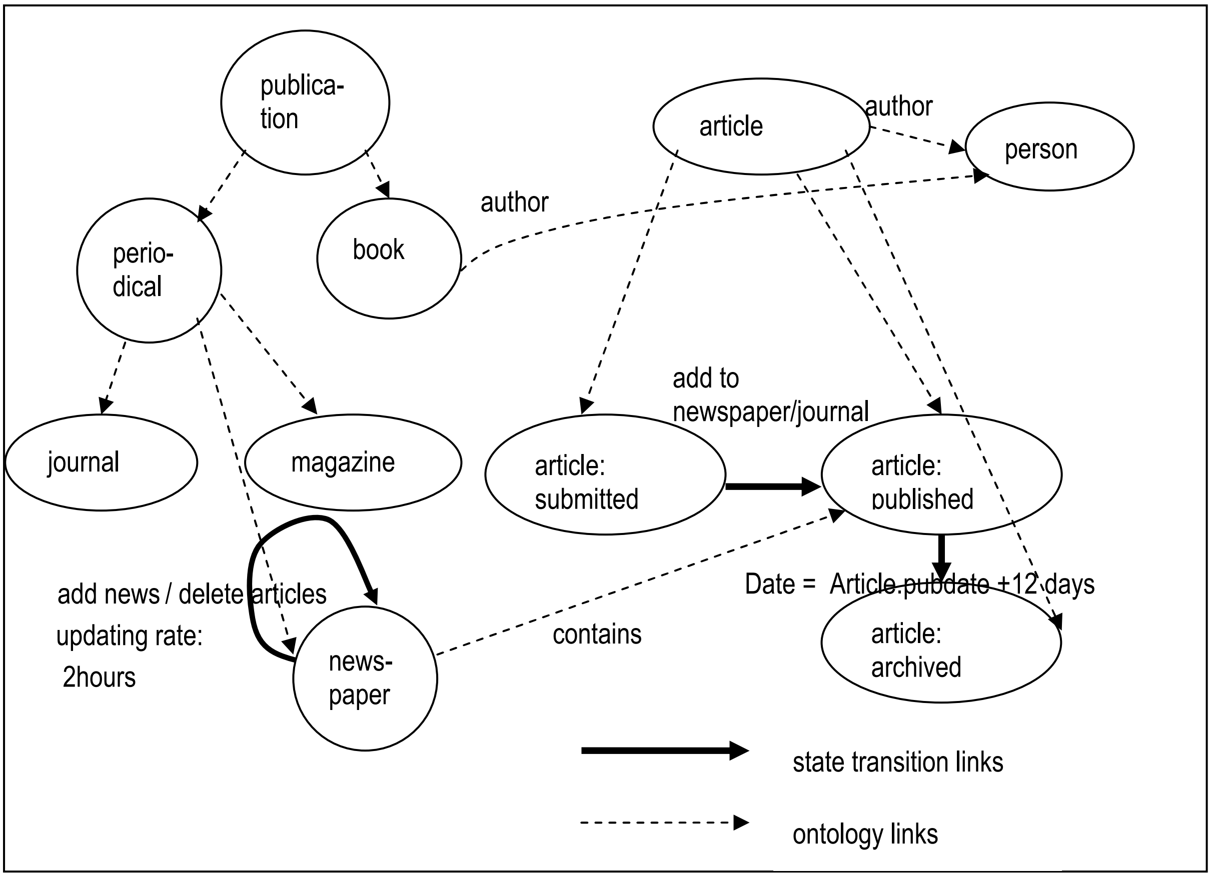


Fig.2 Students List, newspaper and articles state diagrams

For example in figure 2 is represented an online newspaper represented where news change, say, every two hours. In this case the continuous updating of the newspaper resource is represented by a single self reference labelled by operation, i.e. events add/delete news and the conditions that two hours are passed by the previous updated, the newspaper, is related by other ontological concepts (ontological link are represented by dashed lines) such as relationships of type *subclass* with respect to *periodical* and relation *contains* with respect to *published article*.

Articles inherit general concepts such as *author* relationship and can also have a more complex life cycle expressed by a LTD (submitted, published, archived) partially independent from newspaper.

Resource state transitions can be enabled by conditions over other web resources or not web resources.

Reasoning and Acting on Dynamic Resources

State transition diagrams and labelled transition diagrams are formalism popular in the area of process modelling and concurrent system modelling [Gogolla & Parisi-Presicce,1998][Cardell-Oliver et al.,1992].

The labelled transition diagram of a given resource can also be seen as an equivalent deterministic finite state automata DFA, whose transition guards (i.e. the logical conditions) denotes set of symbols in the alphabet defined by the possible binary combination of all atomic conditions. A given guard denotes the set of symbols, which correspond to atomic conditions, which makes the guard itself true.

It would be interesting to investigate the possibility of applying techniques of linear temporal logics [Manna & Pnueli,1991][Vardi,1991] used in circuit testing in order to evaluate LTL queries over a particular state of the dynamical resources.

For example an agent which has found the resource can reason about the truth of LTL modal formula such as Possibly(S) where S is a desired state of the resource, taking appropriate measure, such as *abandoning* the resource if the desired state is unreachable (e.g. the conference submitting deadline is over then transition to state "submitted" is impossible), *waiting* if the state transition is a matter of time or of exogenous events (e.g. wait until tomorrow for the President elections results), inducing the state transition by agent *deliberative action* [Milani & Ghallab,1991](e.g. reserve a ticket in order to buy it), *maintain conditions* which avoid unlikely transitions, consult the resource if the update rate or the type of time validity requires it (i.e. refresh the stock exchange prices, check again the weather forecast service).

Conclusion

The presented preliminary model extends the ontology-based approach to the semantic web, in order to represent the dynamical aspects of web resources, which evolves over time. The classical semantic web hypothesis of web resources as identified by URI is no more valid when web resources evolve; i.e. they assume different states. States of resources are represented by ontological concepts, while labeled transition diagrams are used for description admissible states of resources. The transitions are labeled by the conditions which trigger the state transition, i.e. operation, events or conditions over other resources. The resulting semantic description of web resources consist of a two layered graph which represents both static (i.e. the ontological concepts) and dynamic (i.e. the labeled transitions) relationships among concepts.

The finite state machine model allows to employ powerful reasoning technique in the semantic web graph, as for example LTL (linear temporal logic) in order to make prevision about state of web resources over time. Moreover, this rich knowledge description network can be exploited by web agents who use planning techniques for triggering the desired state transitions.

Bibliography

[Bacchus & Kabanza,1998] Bacchus,F. Kabanza,F., 1998 Planning for temporal extended goals, Annals of Mathematics and Artificial Intelligence, 22:5-27, 1998

[Berners-Lee & Fielding,1998] T. Berners-Lee, R.Fielding, U.C. Irvine, L.Masinter, Uniform Resource Identifiers (URI): Generic Syntax, Request for Comments: 2396, the Internet Society, August 1998

[Berners-Lee et al., 2001] L.Tim Berners Lee, J. Hendler, O. Lassila: The Semantic Web, Scientific American, May 2001

[Boley et al., 2001] Harold Boley, Said Tabet, and Gerd Wagner. Design Rationale of RuleML: A Markup Language for Semantic Web Rules. In International Semantic Web Working Symposium (SWWS), 2001.

[Cardell-Oliver et al.,1992] Rachel Cardell-Oliver, Roger Hale, and John Herbert. An embedding of timed transition systems in HOL. In Higher Order Logic Theorem Proving and its Applications, pages 263--278, Leuven, Belgium, Sept 1992

[Fikes & Nilsson, 1971] R.E.Fikes, N.J.Nilsson. STRIPS: A new approach to the application of theorem proving to problem solving, Artificial Intelligence, 2(3/4), 1971

[Gogolla & Parisi-Presicce,1998] Gogolla, M. and Parisi-Presicce, F., 1998, "State Diagrams in UML - A Formal Semantics using Graph Transformation", Proceedings ICSE'98 Workshop on Precise Semantics of Modeling Techniques (PSMT'98),

[Handschuh & Staab, 2002] S. Handschuh and S. Staab. Authoring and annotation of web pages in CREAM. In The Eleventh International World Wide Web Conference (WWW2002), Honolulu, Hawaii, USA 7-11 May, 2002

[Hendler, 2001] J. Hendler, Agents and the Semantic Web, IEEE Intelligent Systems, vol.16, no.2, Mar./Apr. 2001, pp.30-37.

[Kalianpur, 2001] SMORE - Semantic Markup, Ontology, and RDF, Editor Aditya Kalyanpur A. V. Williams Building College Park, Maryland 20742

[Manna & Pnueli,1991] Manna, Z., Pnueli, A. The temporal logic of reactive and concurrent systems: Specification. Springer Verlag, 1991

[McGuinnes & van Harmelen, 2004] D.L. McGuinness F.van Harmelen, OWL Web Ontology Language Overview W3C Recommendation 10 February 2004

[McGuinnes et al., 2004] D.L. McGuinness, R. Fikes, J. Hendler, and L.A. Stein. ``DAML+OIL: An Ontology Language for the Semantic Web". In IEEE Intelligent Systems, Vol. 17, No. 5, pages 72-80, September/October 2002

[Milani & Ghallab,1991] A.Milani, M.Ghallab eds. "New Direction in AI Planning", IOS Press 1996

[Vardi,1991] M. Vardi. An automata-theoretic approach to linear temporal logic. In F. Moller and G. Birtwistle, editors, Logics for Concurrency, pages 238-266. Springer Verlag, 1996

Authors' Information

Alfredo Milani – Department of Mathematics and Informatics, University of Perugia, Via Vanvitelli, 1, 06100 Perugia, Italy; e-mail: suriani@dipmat.unipg.it

Silvia Suriani – Department of Mathematics and Informatics, University of Perugia, Via Vanvitelli, 1, 06100 Perugia, Italy; e-mail: suriani@dipmat.unipg.it

EFFECTIVENESS OF TITLE-SEARCH VS. FULL-TEXT SEARCH IN THE WEB

Peretz Shoval and Tsvi Kuflik

Abstract: Search engines sometimes apply the search on the full text of documents or web-pages; but sometimes they can apply the search on selected parts of the documents only, e.g. their titles. Full-text search may consume a lot of computing resources and time. It may be possible to save resources by applying the search on the titles of documents only, assuming that a title of a document provides a concise representation of its content. We tested this assumption using Google search engine. We ran search queries that have been defined by users, distinguishing between two types of queries/users: queries of users who are familiar with the area of the search, and queries of users who are not familiar with the area of the search. We found that searches which use titles provide similar and sometimes even (slightly) better results compared to searches which use the full-text. These results hold for both types of queries/users. Moreover, we found an advantage in title-search when searching in unfamiliar areas because the general terms used in queries in unfamiliar areas match better with general terms which tend to be used in document titles.

Keywords: Indexing, Information retrieval, Precision of search results, Search engines, Title search, Web search.

1. Introduction

Search engines generally use the full text of Web pages for searching. The search of full text may be costly in terms of computing resources and time. A possible way to save such resources is by conducting the search on the titles of the documents rather than on their full text. The title of a document is supposed to provide a concise representation of its content. Kwok (1984) used the titles of cited academic publications to improve the indexing of the documents which cite them. He did so by adding, in the indexing process, the content of the cited titles to the content of the documents. Drori (2003) showed that in many cases title terms can be identified by analyzing the content of a document. Belaïd and David (1999), Taniar et al. (2000) and Schenker et al. (2003) used the titles for document representation to help users find relevant documents in search results. Lam-Adesina and Jones (2001) explored the intuitive assumption about the importance of terms appearing in the titles for increasing the weight assigned to such terms while generating document summaries. They generated summaries by extracting sentences out of the documents; sentences containing title terms were scored higher than sentences without title terms.

Obviously, a title of a document cannot provide much detail; it tends to be general and contain general terms, while more specific terms appear in the text itself. Given this, it seems reasonable that a query used for a title-search should include mostly general terms, while a query used for a full-text search should include mostly specific terms.

Research has shown that familiarity of users with the area in which they seek information has an impact on the quality of their search queries. Users who are familiar with the search area know the relevant terminology; therefore it is reasonable to assume that they are able to define precise search queries. On the other hand, users seeking information in areas with which they are not familiar do not know the relevant terminology, and therefore are likely to define imprecise queries that would yield many irrelevant results. With respect to the earlier discussion on generality or specificity of terms appearing in titles vs. the full text, it may be assumed that unfamiliar users would be better of using title-search because they are likely to use general terms in their queries, while familiar users would be better of using full-text search because they are more likely to use specific terms.

The purpose of this study is to compare the effectiveness of title-search and full-text search, and to determine how user familiarity with the search area interacts with the type of search. Section 2 presents related studies on search habits of Web users and the impact of user familiarity with the search area on search results; Section 3 outlines our hypotheses and describes the research; Section 4 presents the results, and Section 5 concludes and suggests further research.

2. Related Studies

Studies on how users behave while searching the Web reveal that they most often tend to define short queries, having an average of 2.35 terms (Jansen et al., 1998), 3.34 terms (Spink et al., 1999) and 2.4 terms (Spink et al., 2001). Web search queries are significantly shorter than queries in classical information retrieval systems, which consist of between 7 to 15 terms (Jansen et al., 2000). Jansen et al. (1998) found that users tend to explore less than three pages of results; the average is 2.21 pages, while half of the users examine only one page, and three quarters examine only two pages or less. Users also tend to perform short search sessions: they pose a query, look at the first page (or two) of results and explore only a few Web sites listed on that page. If they do not find relevant information, they may reformulate the query and repeat the search once or twice, and then abandon the search. On the average they reformulate a query 2.84 times in a search session; two-thirds of the users submit only a single query. These findings indicate how important it is that Web users will get the most relevant information already in the first few pages of the search results. This also explains why a common measure of performance of search engines is "precision at 10" (Jin and Dumais, 2001; Craswell et al., 2001; Eastman, 2002; Plachouras et al., 2003), which means the precision of the 10 top documents (usually presented by search engines in the 1st page of results).

Some search engines have "advanced" search options which allow users to define and run search queries using specific options that extend beyond the "simple" (common) option. For example, an advanced search may enable Boolean operators, or limit the search to specific file types, or to specific attributes of Web documents, such as the title. But users usually do not use these options (Jansen, 1998). Eastman (2002) claimed that the use of advanced options does not improve the search results because the performance of current search engines is good anyhow (as measured by "precision at 10"). The author evaluated the benefit of using advanced search options and found that in 50% of the cases there was no difference in performance between a simple search and an advanced search; in 25% of the cases advanced searches yielded better results than simple searches, and in 25% of the cases simple searches yielded better results than advanced searches. (It should be noted that Eastman's research made no distinction between different advanced search options, so there is no way to discern if any of those options, such as a title-search, is better or worse than another.)

Only a few studies are concerned about the impact of domain knowledge of users on the results of Web searches. Hsieh-yee (1993) found that owing to their domain knowledge, users are familiar with the relevant terminology and hence can define precise search queries. But users who lack domain knowledge need to search for the right terms first, using various tools such as thesauri. Spink et al. (1998) studied the way Web-users judge relevancy of search results. They concentrated on documents that were defined by the users as "partially relevant" and found that the less users knew about the problem at hand, the more items they assessed as partially relevant; and the more they knew, the more items they assessed as relevant. Hoelscher and Strube (1999, 2000) studied the impact of domain knowledge on search performance combined with Web experience. Their subjects were asked to solve information problems using the Web only. They concluded that in order to succeed, users should have both domain knowledge and Web-search experience. Turnbull (2003) surveyed models to determine how users start to search for information in unfamiliar areas. He observed that users usually start by looking for initial information, learn the general concepts of the domain until they gain enough knowledge to enable them to define precise search queries and then evaluate the search results.

3. The Research

Our research hypotheses are as follows: Web users who are familiar with the research area are able to define search queries that yield high quality (high "precision at 10") results, whether the search is in full-text or in the title only; but in title-search the number of results (documents) they get is smaller than in full-text search because the precise query terms which they tend to use fit less with the more general terms used in titles. Contrarily, Web users who are not familiar with the search area are able to define search queries that yield purer (less precise) results; but in title-search the number of results they get is bigger than in full-text search because the more general terms which they tend to use fit more with the general terms used in titles.

To test the hypotheses we conducted Web searches with thirty-four subjects, all 4th-year students of Information Systems Engineering having several years of computer usage and Web search experience. Each of the participants was asked to define two search queries: one in an area familiar to him/her, and the other in an unfamiliar area. We used the Google search engine (Google, 2003) to run the queries. Google enables users to limit the search to the title field of Web documents. Title field search uses the content of the field enclosed by HTML title tags. (Even documents not in HTML, e.g. PDF files, can be viewed because Google automatically generates HTML versions of such files as it crawls the Web (Notess, 2001; 2002.))

Each user ran each of his two search queries twice: One was a "simple search", i.e. search of the full text, and the other was "advanced search", with the option of searching only in the title field. After conducting each search, the user evaluated the top 10 results appearing in the first page of results (or less, in cases when there were less than 10 results). For this, the user had to access the linked Website, read at least its first page, and decided whether or not it is relevant. The users' decisions were recorded for further analysis.

4. Results

The results of the searches are presented in Table 1. The rows detail the 34 cases (users). The columns show the four search scenarios: full-text search in a familiar area, title-search in a familiar area, full-text search in an unfamiliar area, and title-search in an unfamiliar area. Every column is sub-divided into two: one presents the "precision at 10" and the other presents the number of search results. "Precision at 10" is calculated by counting the number of relevant results (as determined by the user) divided by 10 or by the number of results in cases when there were fewer results. In the following sections we discuss the results according to three issues: a) precision of results; b) number of results; and c) length of queries.

Table 1: Search results

Unfamiliar Area				Familiar Area				User
Title		Full-Text		Title		Full-Text		
# of results	Precision at 10	# of results	Precision at 10	# of results	Precision at 10	# of results	Precision at 10	
0	no data	> 10	0.6	0	no data	> 10	0.9	1
> 10	0.3	> 10	0.2	0	no data	> 10	0.8	2
> 10	0.8	> 10	0.8	0	no data	> 10	0.9	3
0	no data	> 10	0.7	0	no data	> 10	0.8	4
1	1	> 10	0.6	3	0.667	> 10	0.7	5
> 10	0.8	> 10	0.9	> 10	0.8	> 10	0.9	6
> 10	0.6	> 10	0.7	0	no data	> 10	0.7	7
> 10	0.6	> 10	0.6	> 10	0.6	> 10	0.4	8
> 10	0.3	> 10	0.7	0	no data	> 10	0.9	9
0	no data	> 10	1	0	no data	> 10	0.7	10
2	1	> 10	0.5	0	no data	> 10	0.3	11
> 10	0.8	> 10	0.8	0	no data	> 10	0.9	12
0	no data	> 10	0.4	0	no data	> 10	0.5	13
9	0.889	> 10	0.9	10	0.9	> 10	0.8	14
> 10	0.2	> 10	0.2	0	no data	> 10	0.2	15
> 10	0.3	> 10	0.3	2	0.5	> 10	0.5	16
0	no data	> 10	0.9	0	no data	> 10	0.9	17
0	no data	> 10	0.8	0	no data	> 10	0.9	18
> 10	0.3	> 10	0.5	> 10	0.8	> 10	0.4	19
> 10	0.4	> 10	0.5	0	no data	> 10	0.4	20
> 10	0.8	> 10	0.6	> 10	1	> 10	1	21
> 10	0.7	> 10	0.5	2	0.5	> 10	0.7	22
0	no data	> 10	0.7	0	no data	> 10	1	23
> 10	0.8	> 10	0.8	2	0.7	> 10	0.9	24
> 10	0.7	> 10	0.7	> 10	1	> 10	0.7	25
1	1	> 10	0.7	> 10	0.9	> 10	1	26
> 10	1	> 10	1	0	no data	0	no data	27
0	no data	> 10	0.6	0	no data	> 10	1	28
> 10	1	> 10	1	0	no data	> 10	1	29
0	no data	> 10	0.8	0	no data	> 10	0.3	30
0	no data	> 10	0.7	2	1	> 10	1	31
> 10	0.9	> 10	1	> 10	0.5	> 10	0.4	32
3	1	> 10	0.8	> 10	0.7	> 10	0.6	33
> 10	0.8	> 10	0.7	> 10	0.8	> 10	0.9	34

4.1 Precision of Results

Table 2 presents the average "precision at 10" (as based on the values presented in Table 1) for the four scenarios.

Table 2: Average precision

Title	Full-text	
0.76	0.73	Familiar area
0.71	0.68	Unfamiliar area

As can be seen, title-search yielded better results compared to full-text search in both the familiar and unfamiliar areas. We can also see that search in familiar areas yielded better results compared to search in unfamiliar areas. However, t-tests of differences between the averages of the familiar and unfamiliar areas, for both full-text and title-searches, revealed that the differences are not significant (p=0.111 and p=0.409, respectively). Similarly, t-tests of differences between the averages of the full-text and title-searches, for both the familiar and unfamiliar areas, also revealed that the differences are not significant (p=0.248 and p=0.151, respectively). At any rate, it is important to note that the results of the full-text search are **not better** than those of the title-search.

4.2 Number of Results

Table 3 shows the number of results in all cases. The columns represent the four scenarios; each column is sub-divided into two, distinguishing between the number of cases with 10 or less results, and the number of cases with more than 10 results.

Table 3: Number of results

Unfamiliar Area				Familiar Area			
Title		Full-Text		Title		Full-Text	
# of cases with >10 results	# of cases with ≤ 10 results	# of cases with >10 results	# of cases with ≤ 10 results	# of cases with >10 results	# of cases with ≤ 10 results	# of cases with >10 results	# of cases with ≤ 10 results
20	4	34	0	9	6	33	0

As can be seen, in full-text search there are more than 10 results in all the cases (except for one search in a familiar area search where no results at all were obtained). In title-search the results are different: when searching in a familiar area only 15 cases yielded results, and in only in 60% of them (9) the number of results exceeds 10. When searching in an unfamiliar area, more (24) cases yielded results, and in 83% of them (20) the number of results exceeded 10. These results tell us that full-text search yields a redundancy of results regardless of the level of user familiarity with the search area. On the other hand, title-search yields less results, sometimes too few. However, title-search in an unfamiliar area provided more results than title-search in a familiar area. The reason for the difference may be, as hypothesized, that users in familiar areas are able to defined precise queries yielding good results in any case (full-text as well as title-search); but because their queries are specific, using precise terms, their title-search yield less results (because title terms tend to be more general). Contrarily, users in unfamiliar areas use more general query terms, which correlate better with the general terms used in titles, and therefore they obtain more results.

4.3 Lengths of Queries

The average length of the search quarries was 3.29 terms for an unfamiliar area, and 2.94 terms for a familiar area. These lengths are similar to Web query lengths reported earlier. In order to better understand the differences in the results between the two cases of the **title-search**, we analyzed the lengths of queries by comparing the differences between lengths of queries which yielded results and queries which did not yield results. Table 4 shows the query lengths, distinguishing between searches in familiar and unfamiliar areas.

Table 4: Length of queries

Title-search in Unfamiliar Area		Title-search in Familiar Area		
Results	No results	Results	No results	
2.79	4.5	3.06	4.53	Average
1.14	1.18	0.93	1.23	Standard dev.

For search in familiar areas, the queries that yielded **no results** are 50% longer than those that yielded results (4.53 compared to 3.06 terms). For search in unfamiliar areas, the difference is even greater, being about 60% longer (4.5 terms compared to 2.79 terms). T-tests reveal that the differences in the query lengths are significant (p<0.00 for both types of queries).

The lengths of queries that yielded results are within the range of query lengths reported in the literature (3.34 according to Spink et al., 1999; and 2.35 according to Jansen et al., 2000). But queries that yielded no results are substantially longer. Hence, the reason for fewer results in title-search can be explained by their length, because of the excessive number of specific terms. While specific/detailed queries are good if one wants to reduce the number of irrelevant results and seeks high precision in full-text search, they seem not to be so good in title-search, because – as said - a title consists of a small number of general terms, which do not correlate with the terms in detailed queries. Hence, queries used in title-search should be shorter if the user wants to get a substantial amount of results.

5. Conclusions

Using Google as a search engine, we showed that search queries provide highly precise results, regardless of whether a familiar or unfamiliar area is being searched. The results support the hypothesis that Web users searching in a familiar area are able to define precise search queries that yield high quality results. For searching in an unfamiliar area, this result contradicts the hypothesis (as we were expecting low precision). But the more interesting result is that **search in the title filed yielded results which are not worse, and sometimes even better, than searching in the full text**. Hence, a lot can be saved in the ways searches are performed and indexes are constructed: searching and indexing of Web pages can be based on titles of documents rather than on their full text.

Although a title-search yielded fewer results, this does not present a problem since users usually do not examine more than one or two pages of search results. At any rate, if a query yields too few results, it may be too specific and include too many terms, so the user can revise the query accordingly.

As any other experiment, this too has limitations, such as the small number of search queries and the use of one search engine only. The results of this study should be further validated by using more queries, different types of users, different search areas, and more search engines (besides Google).

Bibliography

Belaïd, A. & David, A. (1999). The use of information retrieval tools in automatic document modelling and recognition. Proc. of the 10th International Workshop on Database & Expert Systems Applications, 522-526.

Craswell, N., Hawking, D. & Griffiths, K. (2001). Which search engine is best at finding airline site home pages? CSIRO Mathematical and Information Sciences TR01/45.

Drori, O. (2003). Identifying the subject of documents in digital libraries automatically using frequently occurring words – study and findings. Proc. of the 3rd International Workshop on New Development in Digital Libraries, NDDL 2003, 3-12.

Eastman, C. (2002). 30,000 hits may be better than 300: precision anomalies in Internet searches. Journal of the American Society for Information Science and Technology, 53 (11), 879-882.

Google. www.google.com, (accessed September, 2003).

Hoelscher, C. & Strube, G. (1999). Searching on the Web: two types of expertise. Proc. of the 22nd Annual Int'l ACM SIGIR Conference on Research and Development in Information Retrieval, 305-306.

Hsieh-yee I. (1993). Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. Journal of the American Society for Information Science, 44 (3), 161-174.

Jansen, B., Spink, A., Bateman, J. & Saracevic, T. (1998). Real life information retrieval: a study of user queries on the Web. SIGIR Forum, 32 (1), 5-17.

Jansen, B., Spink A. & Saracevic T. (2000). Real life, real users and real needs: a study and analysis of user queries on the Web. Information Processing and Management, 36, 207-227.

Jin, R. & Dumais, S. (2001). Probabilistic combination of content and links. Proc. of the 24th Annual Int'l ACM SIGIR Conference on Research and Development in Information Retrieval, 402-403.

Kwok, K. (1984). A document-document similarity measure based on cited titles and probability theory, and its application to relevance feedback retrieval. Proc. of the 7th Annual Int'l ACM SIGIR Conference on Research and Development in Information Retrieval, 221-231.

Lam-Adesina, A. & Jones G. (2001). Applying summarization techniques for term selection in relevance feedback. Proc. of the 24th Annual Int'l ACM SIGIR Conference on Research & Development in Information Retrieval, 1-9.

Notess, G. (2001). Tracking title search capabilities. http://www.onlinemag.net/OL2001/net5_01.html.

Notess, G. (2002). Review of Google. http://www.searchengineshowdown.com/features/google.

Plachouras, V., Ounis, I., Amati, G. & Van Rijsbergen. (2003). University of Glasgow at the Web track of TREC 2002. Proc. of the 11th Text Retrieval Conference, TREC 2002.

Schenker, A., Last, M., Bunke, H. & Kandel, A. (2003). Graph representations for Web document clustering. Proc. of the 1st Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA'2003).

Spink, A., Graisdorf, H. & Bateman, J. (1998). From highly relevant to not relevant: examining different regions of relevancy. Information Processing and Management, 34 (5), 599-621.

Spink, A., Bateman, J. & Jansen, B. (1999). Searching the Web: a survey of EXCITE users. Internet Research: Electronic Networking Applications and Policy, 9 (2), 117-128.

Spink, A., Wolfram, D., Jansen, B. & Saracevic, T. (2001). Searching the Web: the public and their queries. Journal of the American Society for Information Science and Technology, 52 (3), 226-234.

Taniar, D., Jiang, Y., Rahaya, J. & Bishay, L. (2000). Structured Web pages management for efficient data retrieval. Proc. of 1st Int'l Conference on Web Information Systems Engineering (WISE'00), 2097-2104.

Turnbull, D. (2003). Augmenting information seeking on the World Wide Web using collaborative filtering techniques. http://donturn.fis.utoronto.ca/research/augmentis-toc.html (accessed June 2003).

Authors' Information

Peretz Shoval – Department of Information Systems Engineering, Ben-Gurion University, Beer-Sheva 84105, Israel; e-mail: shoval@bgumail.bgu.ac.il

Tzvi Kuflik – Department of Management Information Systems, Haifa University, Haifa, Israel;
e-mail: tsvikak@mis.hevra.haifa.ac.il

EMPIRICAL METHODS FOR DEVELOPMENT AND EXPANDING OF THE BULGARIAN WORDNET

Pavlina Ivanova, George Totkov, and Tatiana Kalcheva

Abstract: Some basic points from the automated creation of a Bulgarian WordNet – an analogue of the Princeton WordNet, are treated. The used computer tools, the received results and their estimation are discussed. A side effect from the proposed approach is the receiving of patterns for the Bulgarian syntactic analyzer.

Keywords: Empirical Methods in NLP, WordNet

1. Introduction

WordNet is developed in the Princeton University [2,4] as a lexical database of English. The first multilingual database to realize such approach is *EuroWordNet* (EWN) ([11], [12]) consisting of eight European languages. The monolingual databases are related to the *Princeton WordNet* (PWN) (and in this way to each other) via an interlingual index (ILI).

The *Bulgarian WN* (BWN) has been developed as a cooperative task involving the Plovdiv University and the Department for Computer Modelling of Bulgarian Language at the Bulgarian Academy of Sciences (DCMB). The work is part of an EC funded project (IST-2000-29388) *BalkaNet* [7] for the creation of a multilingual lexical database (like EWN) for 6 Balkan languages (Bulgarian, Greek, Romanian, Serbian and Turkish, Czech).

2. Forming of a BWN

The main stages in the automatic creation of a BWN (A_BWN) are presented in [8]. We discuss further the tools and the results received in this process – namely the extraction of synsets from an English-Bulgarian dictionary (EBD) and the receiving of A_BWN.

Our starting point is the transformed EBD [6] with more than 160,000 entries. Each different meaning of an English word is placed on a different row. Each row contains the English word (entry) and its translation equivalents (TE) in Bulgarian. A link is added (where it was possible) between the EBD rows and the PWN synsets (via the ILI) [9].

Each TE row may contain Bulgarian words and phrases separated with the following signs: comma, colon, semi-colon, full stop, slash and brackets. In order to receive the different synonyms from a TE row we had to differentiate the punctuation marks used as 'separators' from the ones marking some orthographical rule. E.g. in the translation of "*anticipant*" – *"човек, който чака, чакащ"*, the first comma is not a separator while the second one is.

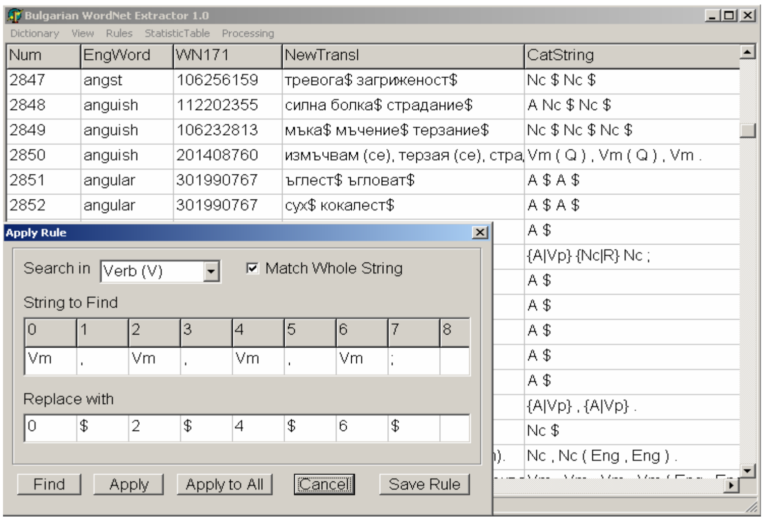
A special tool *BWN Extractor* (BWNE) is designed for the solving of the problem. The BWNE was created to extract almost automatically meaningful rules for forming Bulgarian synsets corresponding to PWN. In the first place, the Bulgarian words in TE rows were processed by Bulgarian Morphological Analyzer BulMorph 2.0 [10] in order to get a list of their morphological characteristics (MC). As a result we received a string-pattern in which every Bulgarian word from the TE row was replaced with a special symbol(s) coding its MCs (e. g. N denotes a noun, A – adjective, V – verb, D – adverb, Vm – a verb in indicative mood, Va – the verb 'be', Vp – participle, Nc – common noun, Q – particle, etc.) The morphological alternatives (ambiguities) are separated with '|' and the results from the robust morphological analysis [10] are marked with the sign '^'.

Table 1 presents syntactic patterns (SynP), obtained with BWNE and ordered according to their frequency in the processed TE.

SynP	Noun	Verb	Adjective	Adverb	Total
Nc	10134	11	45	2	10192
Nc , Nc	4123	5	8	0	4136
A	59	2	3749	25	3835
Vm	20	2463	24	1	2508
A , A	13	0	2460	11	2484
Vm , Vm	11	2215	8	2	2236
A Nc	2070	2	36	1	2109
Nc , Nc , Nc	1009	0	2	0	1011
Nc|Vn	913	0	5	0	918

Table 1. The first 9 syntactic patterns received by BWNE

What this statistics shows is that, for example, when the TE row of an English word consists of two nouns separated by comma (Nc, Nc) in 4123 of 4136 cases (more than 99.6%) the English word is also a noun and the corresponding two Bulgarian words (nouns) are two synonyms. Only the cases when the part of speech (POS) does not match are questionable and need to be marked by expert using BWNE.



Num	EngWord	WN171	NewTransl	CatString
2847	angst	106256159	тревога$ загриженост$	Nc $ Nc $
2848	anguish	112202355	силна болка$ страдание$	A Nc $ Nc $
2849	anguish	106232813	мъка$ мъчение$ терзание$	Nc $ Nc $ Nc $
2850	anguish	201408760	измъчвам (се), терзая (се), стра	Vm (Q) , Vm (Q) , Vm .
2851	angular	301990767	ъглест$ ъгловат$	A $ A $
2852	angular	301990767	сух$ кокалест$	A $ A $
				A $
				{A|Vp} {Nc|R} Nc ;
				A $
				A $
				A $
				A $
				A $
				{A|Vp} , {A|Vp} .
				Nc $
			).	Nc , Nc (Eng , Eng) .

Apply Rule

Search in Verb (V) ☑ Match Whole String

String to Find

0	1	2	3	4	5	6	7	8
Vm	,	Vm	,	Vm	,	Vm	;	

Replace with

0	$	2	$	4	$	6	$	

Figure 1. The Rule Editor window of the BWNE

The rules for the separation of the synonyms are based on the automatically received SynP. Moreover, BWNE provides a special *Rule Editor*. Figure 1 shows the creation of a rule to be applied on all rows corresponding to an English entry defined as 'verb'.

The functional capabilities of the *Rule Editor* are: a) automatically synthesizes rules, starting with the most likely ones; b) allows additional editing of the automatically synthesised rules; c) represents all the rows in EBD corresponding to the processed SynPs in *View* mode; d) allows changes in the respective rows in EBD in *Edit* mode; e) gives possibility for successive processing of rows from EBD (one by one or in group) in *Apply* mode; f) provides *Save Rule* mode, etc.

Experiments show that approximately 45,000 rows (TE) from the initial EBD can be automatically processed with the first 100 synthesized rules. The next 3,000 rules process additional 20,000 rows. In this way about 65,000 rows of EBD are almost automatically processed with 3,300 rules. The extracted synonyms form A_BWN, containing about 42,000 Bulgarian synsets linked to the corresponding English synsets in PWN.

Table 2 presents 15 of the 3,300 automatically synthesized rules. Each rule consists of three parts: *POS* of the entry for whose TE a given rule is applied; *Left side* containing the (searched) string-pattern and *Right side* defining the replace string – a sequence of numbers (position of the Left side components) separated by the sign '$'. E.g. rule 15 means that 4 synonyms will be extracted in all the TE rows (for which the ILI corresponds to a 'verb') matching the pattern ***verb1/verb2 noun1/noun2.*** The four extracted synonyms (separated by '$') are as follows: ***verb1 noun1 $ verb1 noun2 $ verb2 noun1 $ verb2 noun2$.***

Note that in rules 9-12 the comma is not (always) a separator. Its role depends from the POS of the entry – a comma followed by a relative pronoun (Pr) is a separator when the corresponding POS is A (rule 1) but it isn't when the POS is N (rule 10).

№	POS	Left Side	Right Side
1.	A	A , A , Pr Pp Vm	1 $ 3 $ 5 6 7 $
2.	A	D {A|Vp} , A	1 2 $ 4 $
3.	A	A ; R A Nc	1 $ 3 4 5 $
4.	A	Vp , A , A , R A Nc	1 $ 3 $ 5 $ 7 8 9 $
5.	D	D , D , R Pd {A|Nc}	1 $ 3 $ 5 6 7 $
6.	D	R A Nc / Nc	1 2 3 $ 1 2 5 $
7.	N	A / Vp Nc	1 4 $ 3 4 $
8.	N	A Nc , {An|D} Nc , {An|D|Nc}^	1 2 $ 4 5 $ 7 $
9.	N	An Nc , Vp R A Nc	1 2 3 4 5 6 7 $
10.	N	Nc , Pr Vm / Vm	1 2 3 4 $ 1 2 3 6 $
11.	N	Nc , R Pr Q Vm Nc	1 2 3 4 5 6 7 $
12.	V	Vm (Nc , Nc , {Nc|Np}) ; Vm	1 2 3 4 5 6 7 8 $ 10 $
13.	V	Vm (Q) , Vm (D)	1 3 $ 1 $ 6 7 8 9 $
14.	V	Vm , Nc Va R	1 $ 3 4 5 $
15.	V	Vm / Vm Nc / Nc	1 4 $ 1 6 $ 3 4 $ 3 6 $

Table 2. Rules for the extraction of Bulgarian synonyms

3. Evaluation of the A_BWN

In order to validate the A_BWN we used BWN prototype[1]. The presented result is for an A_BWN consisting of 39,109 Bulgarian synsets and containing 9,936 (common) ILI with the BWN prototype.

Let denote the number of the common literals (different words and phrases in a synset) with E, the number of the A_BWN literals –with F and the number of the A_BWN literals in the intersection – with P[1]. In order to estimate the A_BWN we use two measures:

[1] The prototype, containing 15,007 Bulgarian synsets, is created (manually) by the DCMB experts.

$$\text{Precision} = \frac{P}{F} \text{ and } \text{Recall} = \frac{P}{E}.$$

The number of literals in the BWN prototype is 18,520 and in the A_BWN – 21,302. The average number of literals in a synset is 1.864 and 2.144 respectively. The number of literals common to A_BWN and the BWN prototype is 9,449. The number of synsets common to A_BWN and the BWN is 9,936. The average number of common literals in a synset is 0.951. The *Recall* is 51.02% and the *Precision* is 44.36%.

The new synsets in A_BWN (more than 33,000 additional ILI) give opportunity for further expanding of the BWN prototype.

4. Receiving of Syntactic Patterns

A side effect of the proposed approach is the receiving of syntactic patterns for 4 phrase types in Bulgarian: NP (noun phrase), VP (verb phrase), AP (adjective phrase) and AdvP (adverbial phrase). For example Table 3 presents the first 4 (applied) rules for A (see Table 2).

№	POS	Right Side
1.	A	A $ A $ Pr Pp Vm $
2.	A	D {A|Vp} $ A $
3.	A	A $ R A Nc$
4.	A	Vp $ A $ A $ R A Nc$

Table 3. The (applied) rules 1-4 from Table 2

In fact the received SP for the structure of AP in Bulgarian:

AP := A | Pr Pp Vm | D {A|Vp} | Pr Q Vm | R A Nc | Vp

has to be checked by expert.

The first 10 SP (with greatest frequency) are presented in Table 4.

The experiments show that in this way we define some meaningful rules for the structure of NP, VP, AP and AdvP. The most frequent patterns are most likely to produce correct rules. Using the proposed approach we received 1762 syntactic patterns for the Bulgarian phrases: 1470 for NP, 175 – AP, 169 – VP and 79 – AdvP.

№	SyntacticPattern	NP	VP	AP	AdvP	Total
1.	Nc	**10744**	3	26	2	10775
2.	A	57	0	**3786**	14	3857
3.	A Nc	**3553**	1	0	3	3557
4.	Vm	10	3435	34	1	3480
5.	{Nc|Vn}	**1328**	4	3	0	1335
6.	Vm Q	0	**958**	2	0	960
7.	Vp	39	0	**887**	7	933
8.	Nc^	**871**	1	1	0	873
9.	Vm R Nc	1	**782**	0	0	783
10.	Vm Nc	2	**725**	0	0	727
Total		26028	8187	7336	902	42453

Table 4. The first 10 syntactic patterns

[1] The literals that don't match the literals in the BWN prototype are not necessarily "incorrect".

5. Perspectives

A method for improvement of Bulgarian Synonym Dictionary (BDS) and removing logical discrepancies in synonym rows is described in [3, 9]. The next step to be done is the expanding and correction of the synsets in A_BWN using the improved synsets from regular BDS [5].

A tool analogous to the *Split/Merge* program [9] is under development. The main features of the tool are: a) displaying all the synsets from A_BWN and BDS, in which a chosen word (or phrase) takes part; b) choice of an A_BWN synset to be processed; c) finding the BDS rows which are closest to the chosen synset [9].

The method for extracting syntactic patterns can be applied to *other lexical resources*, for example to Bulgarian Thesaurus [1]. Additional MCs (number, gender, definiteness, etc.) can be used for synthesis of more precise syntactic rules.

The receiving of precise syntactic patterns can be used for the almost automatic creation of a *Bulgarian computer grammar* (including thousands of syntactic rules). The creation of the computer grammar is a crucial step towards the development of a syntactic analyzer of Bulgarian texts.

References

1. Andrejchin L. (ed.), Bulgarian Explanatory Dictionary. Sofia, Nauka i Izkustvo, 1999 (in Bulgarian).
2. Fellbaum C. (ed.), WordNet: An Electronic Lexical Database. The MIT Press, Cambridge, London, England, 1998.
3. Ivanova P., Totkov G., Automated Improving and Forming Synsets on Conventional (non computer based) Synonym Dictionaries, Proceedings of the International Conf. Automation and informatics'2002, Sofia, 33-36.
4. Miller G., R.Beckwith, C. Fellbaum, D. Gross and K.Miller, Introduction to WordNet: an on-line lexical database. In: International Journal of Lexicography 3(4), 1993, accessible at ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps.
5. Nanov L, A. Nanova, Bulgarian Synonym Dictionary. Sofia, Hejzal, 2000 (in Bulgarian).
6. Rankova M., T. Atanasova, I. Harlakova. English-Bulgarian Dictionary. Izd. Nauka I izkustvo, Sofia, 1990.
7. Stamou S., K. Oflazer, K. Pala, D. Christoudoulakis, D. Cristea, D. Tufiş, S. Koeva, G. Totkov, D. Dutoit, M. Grigoriadou, BALKANET: A Multilingual Semantic Network for the Balkan Languages, Proceedings of the International Wordnet Conference, Mysore, India, 21-25 January 2002, 12-14.
8. Totkov G., Towards Building Bulgarian WordNet: Language Resources and Tools, Proceedings of the ICT&P'03, Sofia, 2003, 31-40.
9. Totkov G., P. Ivanova, Iv. Riskov, Automated Improving and Forming WordNet Synsets on Conventional (non computer based) Synonym and Bilingual Dictionaries. in A. Narin'iyani (ed.), Comp. Ling. and its Applications, Proc. of the Int. Workshop DIALOGUE'2003, Protvino, June 2003.
10. Totkov G., R. Doneva, Bipartite Finite State Transducers as Morphology Analyser, Synthesizer, Lemmatizer and Unknown-Word Guesser, Proc. of 2nd Intern. Seminar „Computer Treatment of Slavonic Languages" SLOVKO'2003, Oct. 24-25, 2003, Bratislava (in print).
11. Vossen P. (ed.), EuroWordNet General Document. EuroWordNet (LE2-4003, LE4-8328), Final Document, 1998, 108p.
12. Vossen P. Building a multilingual database with wordnets for several European languages. http://www.hum.uva.nl/~ewn, 1999.

Authors' Information

Pavlina Ivanova – pavlina@pu.acad.bg

George Totkov – totkov@pu.acad.bg

Tatiana Kalcheva – selinashery@abv.bg

Plovdiv University, 4 Tzar Asen str., 4000 Plovdiv, Bulgaria.

MULTIPLIERLESS DCT ALGORITHM FOR IMAGE COMPRESSION APPLICATIONS

Vassil Dimitrov and Khan Wahid

Abstract: This paper presents a novel error-free (infinite-precision) architecture for the fast implementation of 8x8 2-D Discrete Cosine Transform. The architecture uses a new algebraic integer encoding of a 1-D radix-8 DCT that allows the separable computation of a 2-D 8x8 DCT without any intermediate number representation conversions. This is a considerable improvement on previously introduced algebraic integer encoding techniques to compute both DCT and IDCT which eliminates the requirements to approximate the transformation matrix elements by obtaining their exact representations and hence mapping the transcendental functions without any errors. Apart from the multiplication-free nature, this new mapping scheme fits to this algorithm, eliminating any computational or quantization errors and resulting short-word-length and high-speed-design.

Keywords: DCT, Image Compression, Algebraic Integers, Multiplier-less Architecture

Introduction

The Discrete Cosine Transform (DCT) is the core transform of many image processing applications for reduced bandwidth image and video transmission including JPEG and MPEG standards. Several algorithms and architectures have been proposed to optimize DCT implementations using 1-D and 2-D algebraic integer (AI) encoding of the DCT basis functions, where both single and multidimensional AI schemes have been used which allow low-complexity and parallel architectures [Dimitrov, 1998][Dimitrov, 2003]. In all of these previous encoding techniques, conversion from the output of the 1-D DCT algorithm has been required, even if the DCT is being used in a separable 2-D DCT computation. Recently, we have introduced a new algebraic integer encoding technique which, along with a previously published scalar quantization algorithm [Arai, 1988], removes the need for conversion to binary at the end of the first 1-D DCT [Wahid, 2004]. But here in this paper, we present an extensive analysis of that idea of 2-D error-free algebraic integer encoding, in terms of computational complexity and mathematical precision required to implement the algorithm. Here, we also show that this new algorithm provides a considerable reduction in hardware for the separable 2-D DCT computation, and also a reduction in hardware for a stand-alone 1-D DCT computation.

The final conversion step, where we convert the algebraic integer numbers to fixed-precision (FP) binary, may generate some rounding errors but these errors are only introduced at the very end of the transformation process, not distributed throughout the calculation, as is the case for a finite-precision binary implementation. This 2-D algebraic integer quantization not only reduces the number of arithmetic operations, but also reduces the dynamic range of the computations. This scheme can also be extended for the error-free computation of the Scaled Inverse-DCT [Arai, 1988].

Algebraic Integer Quantization (AIQ)

Algebraic integers are defined by real numbers that are roots of monic polynomials with integer coefficients [Dedekind, 1996]. As an example, let $\omega = e^{\frac{2\pi j}{16}}$ denote a primitive 16th root of unity over the ring of complex numbers. Then ω is a root of the equation $x^8 + 1 = 0$. If ω is adjoined to the rational numbers, then the associated ring of algebraic integers is denoted by $Z[\omega]$. The ring $Z[\omega]$ can be regarded as consisting of polynomials in ω of degree 7 with integer coefficients. The elements of $Z[\omega]$ are added and multiplied as polynomials, except that the rule $\omega^8 = -1$ is used in the product to reduce the degree of powers of ω to below 8. For an integer, *M*, $Z[\omega]_M$ is used to denote the elements of with coefficients between $-\frac{M}{2}$ and $\frac{M}{2}$.

The idea of using algebraic integers in DSP applications was first explored by Cozzens and Finkelstein [Cozzens, 1985]. In their work, the algebraic integer number representation, in which the signal sample is represented by a set of (typically four to eight) small integers, combines, if necessary, with the Residue Number System (RNS) to produce processors composed of simple parallel channels [Games, 1989]. In their procedure, AI representation was used to approximate complex input signals. People have found various applications of algebraic integer in Coding theory such as, algebraic integers can produce exact pole zero cancellation pairs that are used in recursive complex finite-impulse response, frequency sampling filter designs [Meyer, 2001].

Apart from the low-complexity error-free computation of DCT and IDCT, our group has also introduced algebraic integer coding to compute the ‘cas’ function of the Discrete Hartley Transform [Baghaie, 2001] and the basis functions of the Discrete Wavelet Transform [Wahid, 2003]. In case of DWT, using 2-D AI encoding technique, not only have we achieved significant improvement in quality of reconstructed image but the hardware is also greatly reduced. Like the DWT, the application of the DCT and the IDCT is also in the field of image and video compression for low bandwidth transmission, and so the enhancement of 2-D AI encoding technique for these transforms is quite necessary and timely in this regard. Another advantage of using AI scheme to these discrete-valued transforms is that greater accuracy can be achieved using fewer bits than necessary with a conventional two’s complement approach.

Discrete Cosine Transform (DCT)

For a real data sequence $x(n)$ of length *N*, the DCT is defined as follows:

$$F(k) = 2\sum_{n=0}^{N-1} x(n)\cos\left[\frac{(2n+1)k}{2N}\pi\right];\ 0 \le k \le N-1 \tag{1}$$

The Inverse DCT (IDCT) is also defined as:

$$x(n) = \frac{1}{N}\sum_{k=0}^{N-1} \overline{F}(k)\cos\left[\frac{(2n+1)k}{2N}\pi\right];\ 0 \le n \le N-1 \tag{2}$$

Where, $\overline{F}(k) = \begin{cases} \dfrac{F(0)}{2} & k=0 \\ F(k) & otherwise \end{cases}$

The 2-D DCT and IDCT can also be found by extending the above 1-D equations.

AI Encoding of the Classical DCT: Both 1-D and 2-D AI encoding have been applied to classical DCT by our group [Dimitrov, 1998][Dimitrov, 2003]. In order to better understand the concept of 2-D algebraic integer quantization to DCT-SQ algorithm, here we will provide a quick review of AI encoding to classical DCT. Taking $z_1 = 2\cos\frac{\pi}{16}$ and $z_2 = 2\cos\frac{\pi}{4}$ and considering the 2-D polynomial expansion, $f(z_1, z_2) = \sum_{i=0}^{K}\sum_{j=0}^{L} a_{ij} z_1^i z_2^j$, we can exactly represent all cosine angles without error as shown in Table 1.

Table 1: 2-D AI representation of the cosine functions for 8-point DCT

2*cos*(0.π/16)	$\begin{bmatrix} 2\,0\,0\,0 \\ 0\,0\,0\,0 \end{bmatrix}$	2*cos*(4.π/16)	$\begin{bmatrix} 0\,0\,0\,0 \\ 1\,0\,0\,0 \end{bmatrix}$
2*cos*(1.π/16)	$\begin{bmatrix} 0\,1\,0\,0 \\ 0\,0\,0\,0 \end{bmatrix}$	2*cos*(5.π/16)	$\begin{bmatrix} 0\,3\,0\,-1 \\ 0\,1\,0\quad 0 \end{bmatrix}$
2*cos*(2.π/16)	$\begin{bmatrix} -2\,0\,1\,0 \\ 0\,0\,0\,0 \end{bmatrix}$	2*cos*(6.π/16)	$\begin{bmatrix} 2\,0\,-1\,0 \\ -2\,0\quad 1\,0 \end{bmatrix}$
2*cos*(3.π/16)	$\begin{bmatrix} 0\,-3\,0\,1 \\ 0\quad 0\,0\,0 \end{bmatrix}$	2*cos*(7.π/16)	$\begin{bmatrix} 0\,-1\,0\,0 \\ 0\,-3\,0\,1 \end{bmatrix}$

The use of multidimensional AI provides us with a variety of advantages. First of all, we are in a position to choose the best representation from a computational viewpoint, such that the equivalent representation scheme is as sparse as possible. Secondly, there are 24 possible combinations of applicable pairs of parameters, which makes this encoding scheme extremely flexible. Thirdly, the final reconstruction can be accomplished by making use of systolic architectures for polynomial evaluations. This technique allows reduction of the degree of polynomial expansion by a factor of two (compared to 1-D encoding where the degree of polynomial is 7 [Dimitrov, 1998]) and consequently speeds up the final reconstruction step by a factor of 2. A 2-D 8x8 DCT IP core based on this technique has recently been designed and fabricated. The core size of this chip is 1.8mmX1.2mm, the latency is 80 clock cycles, the power consumption is 4.8mW and the overall throughput is 75 mega-pixels/seconds [Jullien, 2003]. A micrograph of the chip is shown in Figure 1.

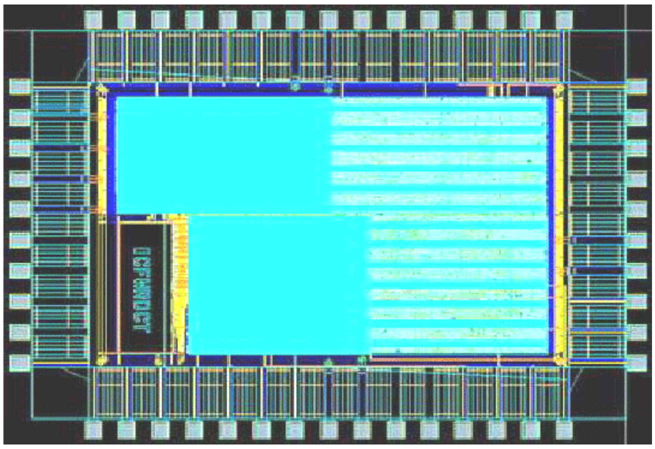
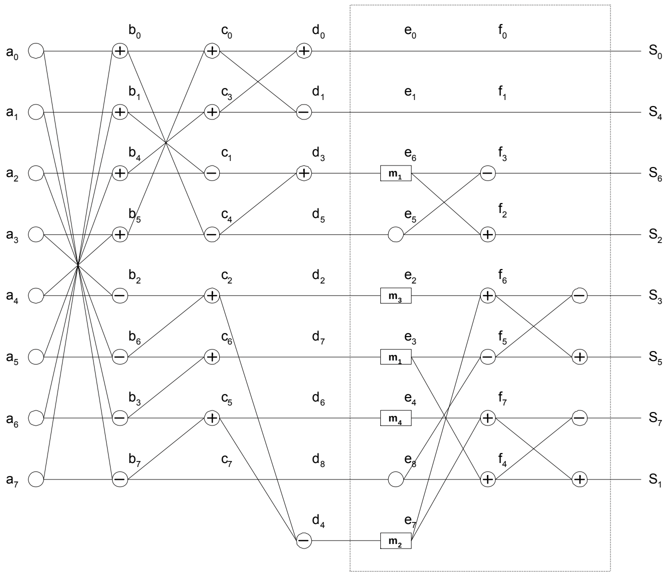


Figure 1: Micro-graph of AI-based DCT chip

Scaled DCT Algorithm

The DCT-SQ (sequential quantization) algorithm proposed by Arai et. al. [Arai, 1988] is presented as follows:



b_0 c_0 d_0 e_0 f_0
a_0 S_0
b_1 c_3 d_1 e_1 f_1
a_1 S_4
b_4 c_1 d_3 e_6 f_3
a_2 m_1 S_6
b_5 c_4 d_5 e_5 f_2
a_3 S_2
b_2 c_2 d_2 e_2 f_6
a_4 m_3 S_3
b_6 c_6 d_7 e_3 f_5
a_5 m_1 S_5
b_3 c_5 d_6 e_4 f_7
a_6 m_4 S_7
b_7 c_7 d_8 e_8 f_4
a_7 S_1
d_4 e_7
m_2

Figure 2: Signal Flow Graph of Arai Algorithm

where $\{a_i\}$ are input elements, $\{S_i\}$ are scaled DCT coefficients, and fixed multipliers are given by eqn. (3).

$$\{m_1, m_2, m_3, m_4\} = \left\{\cos\frac{4\pi}{16}, \cos\frac{6\pi}{16}, (\cos\frac{2\pi}{16} - \cos\frac{6\pi}{16}), (\cos\frac{2\pi}{16} + \cos\frac{6\pi}{16})\right\} \quad (3)$$

Proposed AI Encoding

The outlined area in Figure 2, (with a hardware cost of 5 multiplications and 10 additions) is where our new algebraic integer mapping will be used. Not only will we reduce the hardware count but we will also produce error-free results based on the exact representation of the basis function multipliers.

1-D Algebraic Integer Encoding: Let $z = \sqrt{2+\sqrt{2}}$ and consider the polynomial expansion:

$$f(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3 \tag{4}$$

Since, $\cos\frac{2\pi}{16} = \frac{\sqrt{2+\sqrt{2}}}{2}$, $\cos\frac{4\pi}{16} = \frac{\sqrt{2}}{2}$ and $\cos\frac{6\pi}{16} = \frac{\sqrt{2-\sqrt{2}}}{2}$, we can represent $\{m_1, m_2, m_3, m_4\}$ (from eqn. (3)) exactly (infinite precision) with the integer coefficients (scaled by 2) as shown in Table 2.

Table 2: 1-D error-free multiply encoding

	a_0	a_1	a_2	a_3
m_1	-2	0	1	0
m_2	0	-3	0	1
m_3	0	4	0	-1
m_4	0	-2	0	1

Note that the multiplication between any real number and these coefficients can now be implemented with at most 2 shifts and 1 addition. This reduces the 5 multiplications and the 10 subsequent additions to only 9 AI additions. So, the total number of addition required to perform 1-D DCT is 30. We also note that there is no longer a precision problem since the AI encoding provides an exact representation. The flow graph of Figure 2 can now be implemented as shown in Figure 3.

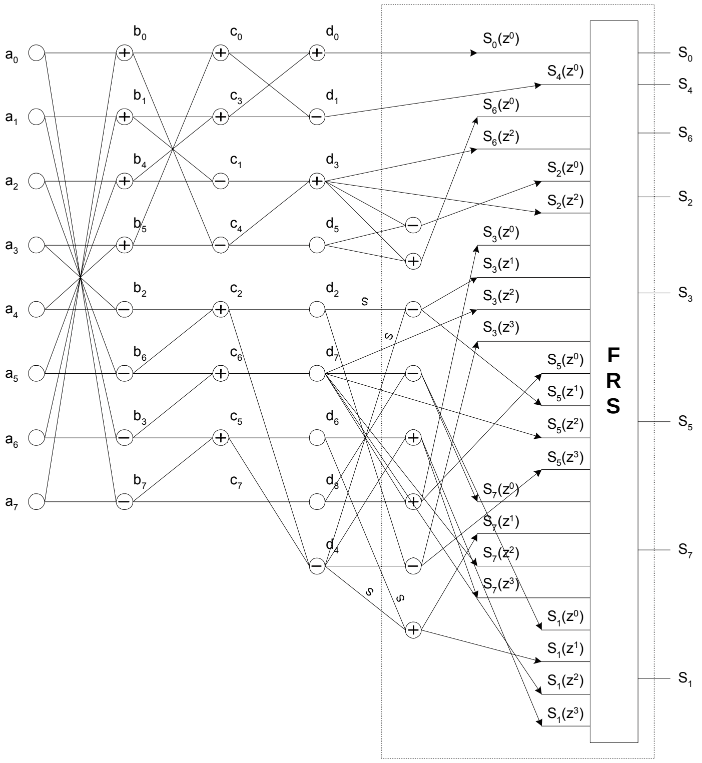


Figure 3: 1-D DCT (1-D error-free encoding)

The real numbers of $f(z)$ form a ring which may be denoted as $Z[\sqrt{2+\sqrt{2}}]$. Addition in this ring is component-wise and multiplication is equivalent to a polynomial multiplication modulo $z^4 - 4z^2 + 2 = 0$.

2-D Algebraic Integer Encoding: Applying a 2-D algebraic integer scheme to this algorithm results in a more sparse representation and more flexible encoding compared to previous techniques [Dimitrov, 2003]. For this encoding, the polynomial is expanded into 2 variables:

$$f(z_1, z_2) = \sum_{i=0}^{K}\sum_{j=0}^{L} a_{ij} z_1^i z_2^j \qquad (5)$$

Here we choose *K*=1 and *L*=1 to guarantee error-free encoding. For the most efficient encoding (i.e., to obtain the most sparse matrix), we have found the following: $z_1 = \sqrt{2+\sqrt{2}} + \sqrt{2-\sqrt{2}}$ and $z_2 = \sqrt{2+\sqrt{2}} - \sqrt{2-\sqrt{2}}$. The corresponding coefficients (scaled by 4) are encoded in the form of $\begin{bmatrix} a_{00} & a_{10} \\ a_{01} & a_{11} \end{bmatrix}$ as shown in Table 3.

Table 3: 2-D error-free multiply encoding

m 1	$\begin{bmatrix} 0\ 0 \\ 0\ 1 \end{bmatrix}$	*m 3*	$\begin{bmatrix} 0\ 0 \\ 2\ 0 \end{bmatrix}$
m 2	$\begin{bmatrix} 0\ 1 \\ -10 \end{bmatrix}$	*m 4*	$\begin{bmatrix} 0\ 2 \\ 0\ 0 \end{bmatrix}$

We have therefore, mapped the multiplier transcendental functions without any error and with very low complexity. Note that in the encoding of all four multipliers, a_{00} requiring only 3 independent parallel channels, as a result, the flow graph in Figure 2 can now be implemented as in Figure 4.

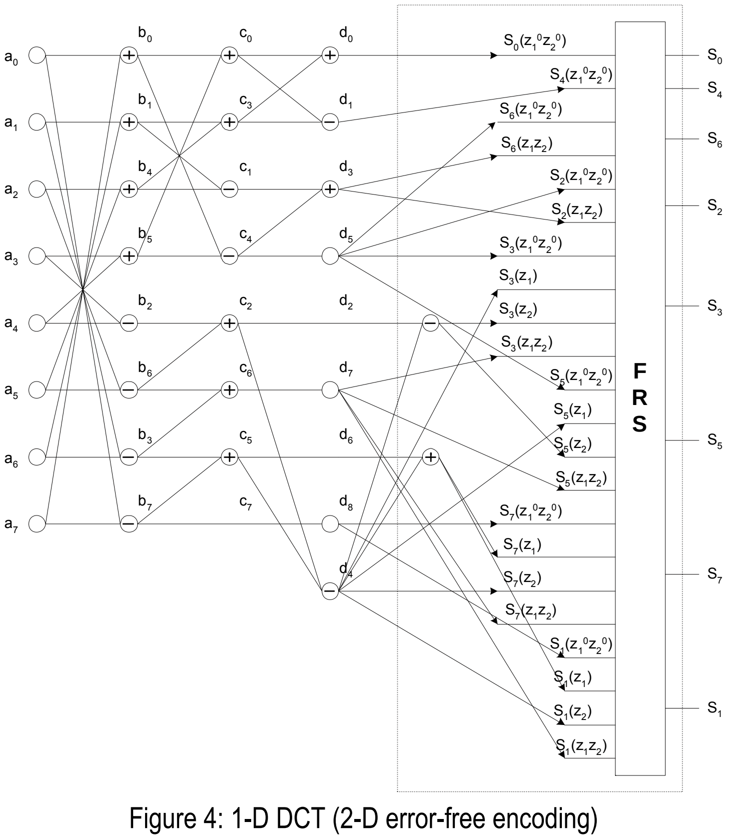


Figure 4: 1-D DCT (2-D error-free encoding)

The outlined area in Figure 4 contains only 2 adders and a Final Reconstruction Stage (FRS), where we finally map to a binary output, is performed as shown in Figure 5 and 6.

Final Reconstruction Step: For the computation of 2-D DCT, we need to recover the integer part of the result and the most significant bit of the fractional part, to allow correct rounding. By applying the 10-bit mappings of Table 4 to the 2-D AI representations of the inputs to the FRS stage, we reduce the hardware cost of the entire outlined area to 5 adders. A considerable reduction from the 5 multiplications and 10 adders of the architecture of Figure 2. Hence a total of only 24 adders is required to perform 8-point 1-D DCT.

For the final reconstruction, we can use Horner's rule [Knuth, 1981]. In that case, eqn. (4) and eqn. (5) can be re-written as:

$$f(z) = ((a_3 z + a_2)z + a_1)z + a_0 \tag{6}$$

$$f(z_1, z_2) = (a_{11} z_1 + a_{01})z_2 + a_{10} z_1 \tag{7}$$

Now, taking different bit-lengths, and using Booth encoding, we can easily find the errors for different substitution precision. The signed-digit encoding errors (%) for different word lengths are provided in Table 5.

Table 4: FRS for different encoding scheme

Scheme	*Parameter*		*FRS*
1-D	z	10 bits	$2-2^{-2}-2^{-5}+2^{-8}$
		12 bits	$2-2^{-2}-2^{-5}+2^{-8}$
2-D	z_1	10 bits	$2+2^{-1}+2^{-3}-2^{-6}$
		12 bits	
	z_2	10 bits	$1+2^{-4}+2^{-6}+2^{-8}$
		12 bits	$1+2^{-4}+2^{-6}+2^{-8}$

Table 5: Bit encoding errors (%)

No of bits	*1-D*	*2-D*	
	z	z_1	z_2
8	2.16×10^{-3}	1.40×10^{-3}	3.94×10^{-3}
10	5.57×10^{-5}	1.40×10^{-3}	3.33×10^{-4}
12	5.57×10^{-5}	3.10×10^{-4}	3.33×10^{-4}
14	5.57×10^{-5}	3.36×10^{-5}	4.86×10^{-6}

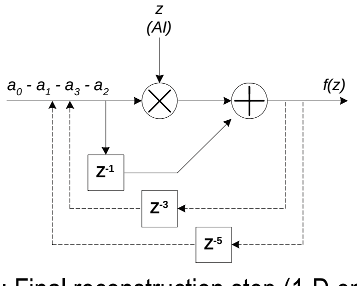


Figure 5: Final reconstruction step (1-D encoding)

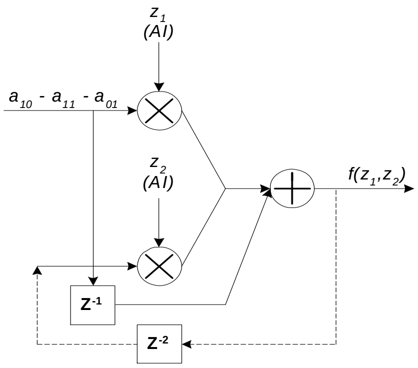


Figure 6: Final reconstruction step (2-D encoding)

Comparisons

In Table 6, we compare the computational complexity of previously published AI-based DCT encoding with the proposed scheme. In all cases, the new 2-D AI encoding scheme has the least number of computations. In Table 7, we present a comparison between some other published 2-D DCT architectures and the proposed algebraic integer approach. Taking the additions as the main computational block, the new multidimensional algebraic integer-quantization based architecture clearly has the lowest hardware count. Also remember the fact that all AI computations are performed without any error.

Table 6: Hardware complexity for different AIQ schemes

Algorithm	Degree of Polynomial	Additions	Shifts	Multiplications	Total Additions
1-D AI-based Chen DCT [Dimitrov, 1998]	7	6	9	0	156
2-D AI-based Chen DCT [Dimitrov, 2003]	7	3	4	0	132
Proposed 1-D AIQ	3	1	2	0	30
Proposed 2-D AIQ	2	0	1	0	24

Table 7: Comparison between different 8-point 2-D DCT

Algorithm	Multiplications	Additions
DCT-SQ [Arai, 1988]	80	464
Chen DCT [Chen, 1977]	256	448
Distributed DCT [Shams, 2002]	0	672
Proposed 1-D AIQ	0	480
Proposed 2-D AIQ	0	384

Conclusions

In this paper, we have introduced a new encoding scheme to compute both 1-D and 2-D DCT and IDCT which effectively reduces the overall arithmetic operations and allows multiplication-free, parallel, and very fast hardware implementation. Except for the final reconstruction stage, the complete 2-D DCT and IDCT can be implemented without error. The use of integers in the encoding scheme also results exact reconstruction. This idea of using algebraic integer scheme can be easily generalized to other algorithms when it is necessary to use real algebraic numbers of special form. The future work is directed towards the VLSI implementation of this approach for 2-D DCT and IDCT.

Acknowledgements

The authors are very thankful to Mr. Minyi Fu for his work on the VLSI implementation of the the 2-D AI algorithm shown in Figure 1. The authors also wish to thank the Alberta Informatics Circle of Research Excellence (iCORE), the Natural Sciences and Engineering Council (NSERC) of Canada, the Micronet and Gennum Corporation for their financial support for this research. The authors are also indebted to the Canadian Microelectronics Corporation (CMC) for providing the hardware and software infrastructure used in the development of this design.

Bibliography

[Arai, 1988] Y. Arai, T. Agui and M. Nakajima, "A Fast DCT-SQ Scheme for Images", Transactions of Institute of Electronics, Information and Communication Engineers, vol. E71, no. 11, pp. 1095-1097, 1988.

[Baghaie, 2001] R.Baghaie and V.Dimitrov, "Systolic Implementation of Real-valued Discrete transforms via Algebraic Integer Quantization", International Journal on Computers and Mathematics with Applications, vol. 41, pp. 1403-1416, 2001.

[Cozzens, 1985] J. H. Cozzens and L. A. Finkelstein, "Computing the Discrete Fourier Transform using Residue Number Systems in a Ring of Algebraic Integers", IEEE Transactions on Information Theory, vol. 31, pp. 580-588, 1985.

[Chen, 1977] W. Chen, C. Smith and S. Fralick, "A Fast Computational Algorithm for the Discrete Cosine Transform", IEEE Transactions on Communications, vol. COM-25, no. 9, pp. 1004-1009, 1977.

[Dedekind, 1996] Richard Dedekind, "Theory of Algebraic Integers", Translated and introduced by John Stillwell, 1996.

[Dimitrov, 1998] V. S. Dimitrov, G. A. Jullien and W. C. Miller, "A New DCT Algorithm Based on Encoding Algebraic Integers", IEEE International Conference on Acoustics, Speech and Signal processing, pp. 1377-1380, 1998.

[Dimitrov, 2003] V. Dimitrov and G. A. Jullien, "Multidimensional Algebraic Integer Encoding for High Performance Implementation of the DCT and IDCT", IEE Electronics Letters, vol. 29, no. 7, pp. 602-603, 2003.

[Games, 1989] R.A.Games, D.Moulin, S.D.O'Neil and J.Rushanan, "Algebraic Integer Quantization and Residue Number System Processing", IEEE International Conference on Acoustics, Speech and Signal processing, pp. 948-951, May 1989.

[Jullien, 2003] M. Fu, G. A. Jullien, V. S. Dimitrov, M. Ahmadi and W. C. Miller, "The Application of 2D Algebraic Integer Encoding to a DCT IP Core", Proceedings of the 3rd IEEE International Workshop on System-on-Chip for Real-Time Applications, vol. 1, pp. 66-69, 2003.

[Knuth, 1981] D. Knuth, "The Art of Computer Programming", vol. 2 - Seminumerical Algorithms, 3rd edition, Addison Wesley, 1981.

[Meyer, 2001] U. Meyer-Base and F. Taylor, "Optimal Algebraic Integer Implementation with Application to Complex Frequency Sampling Filters", IEEE Transactions on Circuits and Systems -II: Analog and Digital Signal processing, vol. 48, no. 11, pp. 1078-82, 2001.

[Shams, 2002] A. Shams, W. Pan, A. Chidanandan and M. Bayoumi, "A Low Power High Performance Distributed DCT Architecture", Proceedings of IEEE Annual Symposium on VLSI, pp. 21-27, 2002.

[Wahid, 2003] K. Wahid, V. Dimitrov, G. Jullien and W. Badawy, "Error-Free Computation of Daubechies Wavelets for Image Compression Applications", IEE Electronics Letters, vol. 39, no. 5, pp. 428-429, March 2003.

[Wahid, 2004] Vassil Dimitrov, Khan Wahid and Graham Jullien, "Multiplication-Free 8x8 2D DCT Architecture using Algebraic Integer Encoding", IEE Electronics Letters, vol. 40, no. 20, pp. 1310-1311, 2004.

Authors' Information

Vassil Dimitrov – Dept. of Electrical and Computer Engineering, University of Calgary, Calgary, Alberta, Canada; e-mail: dimitrov@atips.ca

Khan Wahid – Dept. of Electrical and Computer Engineering, University of Calgary, Calgary, Alberta, Canada; e-mail: wahid@atips.ca

APPLICATIONS OF NONCLASSICAL LOGIC METHODS FOR PURPOSES OF KNOWLEDGE DISCOVERY AND DATA MINING[1]

Vladimir Jotsov, Vassil Sgurev, and Adil Timofeev

Abstract: Methods for solution of a large class of problems on the base of nonclassical, multiple-valued, and probabilistic logics have been discussed. A theory of knowledge about changing knowledge, of defeasible inference, and network approach to an analogous derivation have been suggested. A method for regularity search, logic-axiomatic and logic-probabilistic methods for learning of terms and pattern recognition in the case of multiple-valued logic have been described and generalized. Defeasible analogical inference and new forms of inference using exclusions are considered. The methods are applicable in a broad range of intelligent systems.

Introduction

The classical binary logic is related to formalizing strictly correct (formal) arguments. Still the object field that is the background for the basic concepts and conclusions possesses an incomplete, inaccurate, contradictory, and frequently variable information [1-7]. So there is a necessity to use and develop new non-classical methods for formalizing intelligent processes and information technologies.

At present we have a mighty big variety of different non-classical logics [2,3,7]. Yet the methods for application of these logics in tangible problems are poorly developed. Besides the potential of these logics (e.g. the K-valued logics) does not perfectly satisfy the necessities that originate during the elaboration of intelligent systems and technologies.

The statistical approach to data analysis and making optimal decisions remains popular at present. However it requires a representativeness of the output data, and is not functioning in knowledge-poor environments. Practically the training data sets from which the knowledge is found and the intelligent decisions are formulated are very limited and therefore they are not statistically representative.

This paper describes methods of application for multiple-valued and probabilistic logics to solutions of intelligent systems' problems (particularly, to problems of machine learning and search of regularities on an example of three-valued logics). Some approaches to the creation of conclusions are used: inference by analogy, logic-axiomatic and logic-probabilistic methods, and modeling of network flows. It has been shown that the application of non-classic logic tools allows a significant widening both of the application area and also of the theoretical basis for development even in such a developed area as inference using exceptions – the notion defeasible inference is used below.

The suggested methods allow the cooperation between logic and probabilistic approaches and also to obtain preferences from each of them.

1. Basic Characteristics of Defeasible Inference

Let the unity of classes V is comprised by the subsets S_1, S_2, ... and $S \in V$. Every subset of type S includes elements $x_{s:1}$;, $x_{s;2}$, ..., that form a new model. The original set S is related to one of the classes $S_i \in V$. The final result from the analysis S is idenitified with one of the classes S_i in U. The output is an answer of the type V_s = (T; F;?) with three values: “true”, “false” and “uncertainty”. In the case with an answer V_s =? or V_s=F the set S may be identified with more than a single known class S_{i1}, S_{i2}, ... (i1≠ i2 ...). The answer V_s=T is received if and only if the examined class S coincides with S_i.

Amongst the classes S_i there exists an interdependence of the type "ancestor - successor", (e.g. S_i – an ancestor of S_{i1}). Thus it is possible to form simple types of semantic nets – with one type of relation. It is necessary to note that the elements $x_{si1;1}$, $x_{si1;2,}$ ... produce the differences between the class S_{i1} and the other successors of the

[1] The work is partially supported by Grant no. NIP917 of the Ministry of Science and Education – Republic of Bulgaria.

common ancestor S_i. All differences that appear in the comparison process of S_{i1} with other classes that are not direct successors of S_i are determined after the application of the heredity mechanism.

The conclusion (response) V_s=T is formed when for all ancestors S_i and also for S_{i1} the corresponding conjunction terms are of the following form: $A_1^{'}\Lambda A_2^{'}\Lambda..A_n^{'}$, where A_k is x_k or ¬ x_k; A_k' may coincide with A_k or it may include (using a disjunction) A_k and analogical terms for other variables.

Let rules of a Horn type describe some domain:

$$B \leftarrow \underset{i \in I}{\Lambda} A_i. \tag{1}$$

During the usage of a binary logic in the referred rules if at least a single variable A_i is not "true" then the truth of B is indefinite i.e. B may mean "true" or "false". In the case when the corresponding exclusion from the conjunction (1) of the rule is based on the inclusion of a term with any A_k ($k \in I$) then the inference procedure changes. In the case if the exclusion E (C, A_k) and C is true and A_k is false then the right side of rule B may be true (as an exception).

The extended inference models with exclusions were introduced and generalized in formalized ones in [9,10] in the following form.

$$\frac{B \leftarrow \overset{z}{\underset{i=1}{\Lambda}} A_i, C, E(C, A_k), \neg A_k \leftarrow C}{B \leftarrow A_1 \wedge A_2 \wedge \ldots A_{k-1} \wedge \neg A_k \wedge \ldots A_z} \, , \tag{2}$$

$$\frac{C,\ B \leftarrow \overset{z}{\underset{i=1}{\Lambda}}\ \mathrm{A}_i\ , E(C, A_k)}{B \leftarrow A_1 \wedge \ldots A_{k-1} \Lambda A_{k+1} \Lambda \ldots A_z} \, , \tag{3}$$

$$\frac{C,\ B \leftarrow \overset{z}{\underset{i=1}{\Lambda}}\ \mathrm{A}_i,\ \ E(C, A_k)}{B \leftarrow A_1 \wedge \ldots A_{k-1} \Lambda (A_k \mathrm{V} C) \Lambda A_{k+1} \ldots A_z} \tag{4}$$

It is clear from formulas (2)-(4) that the exclusions are a kind of special-rules inclusions with their effective fields. The interpretation of formula (2) is based on the following: if there exists an exclusion E(C, A_k) that is related to one of the rules with a conclusion B and A_k is its effect then the conjunct A_k must be replaced by ¬A_k. In the case when C is not "true" then the corresponding replacement is impossible. The application of the Modus Ponens rule means that the relation between B and ¬A_k leads to a formal logical contradiction.

Therefore the formation of exclusions of the type E(C, A_k) may lead to a contradictory result that is provoked by an incompleteness in the description of the object field. In the case when C is true then the exclusion E(C, A_k) includes this meaning in the conjunct A_k to defeat the meaning of the last conclusions. The result is that A_k is replaced by C because the test of its meaning does not influence the output. In the case when C is true then the corresponding conjunct A_k is directly replaced by C.

Rules of type (1) are united in systems:

$$\begin{cases} B_1 \leftarrow \underset{i \in I}{\Lambda} A_{1i}. \\ B_2 \leftarrow \underset{j \in I}{\Lambda} A_{2j}. \\ \ldots \end{cases} \tag{1A}$$

$$\begin{cases} B_1 <- \underset{i \in I}{\Lambda} A_{1i}. \\ B_2 <- \underset{j \in I}{\Lambda} A_{2j}. \\ \ldots \end{cases} \tag{1B}$$

In the general case the causal-effective relation may be realized using non-classical operations of successions that are denoted '<-' in the paper and B_i may be presented as combinations of sophisticated logical relations (see formula (1B)).

The usage of exclusions (2) up to (4) may be applied also in systems (1A) or (1B); in the general case it reflects the interrelations between different parts of the causal-effective relations influenced by a new information (an exclusion that is attached to one or other group of relations). The new information may influence the mutual relation between the elements of rule (1) or of systems (1A); (1B). In this case the relations of a causal-effective type are defeated or they are strengthened due to an additional information that is contained in the exclusions. The rest of the paper does not include versions (1A) and (1B) because in the majority of our practical applications it is sufficient to confine ourselves to rules (1) thus the algorithmic complexity of the used combination of methods is significantly lowered. By their nature the presented exclusions are an enlarged version of defeasible inferences that is widely used in the intelligent systems. It is a difference from the classical inference with exclusions that in the presented work it is possible not only to exclude the exclusion A_k that is contained in and tailored to the rule but also that we may include in the rule a new formula e.g. $\neg A_k$ in formula (2) or an interrelation between A_k and C in (4). The research also includes versions of formulas using a non-classical negation ~, versions with exclusions of implications influenced by exclusions, etc.:

$$\frac{B \leftarrow \overset{z}{\underset{i=1}{\Lambda}} A_i C, E(C, A_k), \sim A_k \leftarrow C}{B \leftarrow A_1 \wedge A_2 \wedge \ldots A_{k-1} \wedge \sim A_k \wedge A_{k1} \wedge A_{k+1} \wedge \ldots A_z}, \qquad (2A)$$

$$\frac{C,\ B \leftarrow \overset{z}{\underset{i=1}{\Lambda}}\ A_i\ , E(C, A_k)}{A_1 \wedge \ldots A_{k-1} \wedge A_{k+1} \wedge \ldots A_z}, \qquad (3A)$$

$$\frac{C,\ B \leftarrow \overset{z}{\underset{i=1}{\Lambda}}\ A_i\ , E(C, A_k)}{A_1 \wedge \ldots A_{k-1} \Lambda A_k \Lambda A_{k+1} \Lambda \ldots A_z}, \qquad (3B)$$

where A_{k1} is an additional condition for the transition from $\sim A_k$ to $\neg A_k$. The investigation includes schemas with multi-argument exclusions $E(C,A_k,A_l,\ldots A_s)$ that lead to the simultaneous change of several parts of the rule. The introduced method leads to three basic results: the truth of parts of the rule is altered influenced by the exclusion (if the conditions for activation of the exclusion are enabled), formulas are included in or excluded out of the rule or the rule itself is defeated as it is shown in (3A) or (3B). The results from the research led to a great number of inference versions with exclusions; a part of them is included in our bibliography list.

As it was already shown we introduced a generalized concept of defeating that is based on the following facts. Object scope modeling is a dynamic process. In the act of scope-field completion by the system the old relations between separate parts of the knowledge and/or between different knowledge may be eliminated, changed or their effect may be redirected. This is accomplished influenced by the new knowledge that complete or correct the primary existing knowledge or the interrelations in it. The processes are formalized in the following way.

We did a research of the situations that appear after the addition of new knowledge to the existing knowledge basis and we grouped them in 11 basic groups. Let P is the part of the new knowledge that influences one or more formulas (e.g. see (1) up to (4)).

I. P 'nullifies' A_k: it defeats its relation to the conclusion B. As a result of the defeat A_k has a meaning of 0 and no matter whether it is true or false the true of the conclusion does not change.

$$\frac{B \leftarrow \overset{z}{\underset{i=1}{\Lambda}} A_i, P}{B \leftarrow A_1 \wedge A_2 \wedge \ldots A_{k-1} \wedge A_{k+1} \wedge \ldots A_z, \neg(B \leftarrow \overset{z}{\underset{i=1}{\Lambda}} A_i)},$$

where in difference with defeasible inference schemes the first rule format existing before the appearance of P becomes false.

II. This is an extreme version of the situation from group I when all the atoms in the antecedent are defeated. Now rule (1) turns into a fact: B ←.

$$\frac{B \leftarrow \overset{z}{\underset{i=1}{\Lambda}} A_i, P}{B, \neg(B \leftarrow \overset{z}{\underset{i=1}{\Lambda}} A_i)},$$

III. P changes the true of A_k from true to false or v.v.

$$\frac{B \leftarrow \overset{z}{\underset{i=1}{\Lambda}} A_i, P}{B \leftarrow A_1 \wedge A_2 \wedge \ldots A_{k-1} \wedge \neg A_k \wedge \ldots A_z, \neg(B \leftarrow \overset{z}{\underset{i=1}{\Lambda}} A_i)},$$

IV. P defeats the existing meaning of A_k and increases it to 1. The meaning of the other parts of the antecedent of (1) duly drops down to 0. Independently on the way (conjunctively or disjunctively) they are related to A_k in this situation they are defeated by the antecedent of rule (1).

$$\frac{B \leftarrow \overset{z}{\underset{i=1}{\Lambda}} A_i, P}{B \leftarrow A_k, \neg(B \leftarrow \overset{z}{\underset{i=1}{\Lambda}} A_i)},$$

V. P redirects the relation between the rule and the other knowledge in the domain.

The causal-effective relations are not exhausted by the classical implication and the next example will show that even by formal means it is possible to present different causal-effective relations. Let us have the following two rules:

R_1: B ← A; R_2: N ← M.

Let both rules initially be related to the object X. Let also after the appearance of the new set of conclusions P R_1 is related to Y and R_2 to the former object X. In this case the first rule is preserved but its effect is redirected to another object.

For example it is known that by nature a disease is provoked either by a virus or by a bacteria. However let us have a case when a patient manifests simultaneous symptoms of an illness both from a virus and from a bacteria. The sequent investigation (P) shows that the symptoms of a virus-provoked disease are related to the patient's throat and that the bacterial symptoms are related to the patient's lungs. The redirecting of the conclusion that contradicts to the rule from the example and the discovery of the second disease solve the problem from this example. It is possible to redirect whole rules as an analogy to the presented example.

VI. P breaks or amplifies the relation between the rule and the other knowledge in the domain.

The difference with the previous situation V now is either the elimination of the existing relations or the addition of new relations between the existing rules. The very rules are preserved at that.

For example every chess-player must have a good physical condition so that he/she can present himself/herself well in the tournaments. If however the 'examined' chess-player is a computer program – this is the effect from the new information P – then the already said does not at all concern this program.

VII. P influences the conclusion from one or from a group of rules: from R_1: B ← A into R'_1: B^* ← A. In this way the old conclusion P is defeated or it is replaced by the new one B^*.

$$\frac{B \leftarrow \overset{z}{\underset{i=1}{\Lambda}} A_i, P}{B^* \leftarrow \overset{z}{\underset{i=1}{\Lambda}} A_i, \neg(B \leftarrow \overset{z}{\underset{i=1}{\Lambda}} A_i)},$$

VIII. The appearance of P changes the antecedent of the examined rule (1). It imports a new atom on the place of A_k, before or after the chosen one A_k. In the last two cases the new atom is conjunctively or disjunctively related to A_k, e.g.

$$\frac{B \leftarrow \overset{z}{\underset{i=1}{\Lambda}} A_i, N(P,J)}{B \leftarrow A_1 \wedge A_2 \wedge \ldots A_{k-1} \wedge J \wedge \ldots A_z, \neg(B \leftarrow \overset{z}{\underset{i=1}{\Lambda}} A_i)},$$

This situation can be named specifying the antecedent as a result from the new information P.

IX. R_1 is replaced by R_2 influenced by P:

R_1: B ← A; R_2: N ← Q.

The difference from the previous situation here is in the provoked by P complete replacement of the rule in accordance with the a priori defined concepts.

$$\frac{B \leftarrow \overset{z}{\underset{i=1}{\Lambda}} A_i, P}{N \leftarrow Q, \neg(B \leftarrow \overset{z}{\underset{i=1}{\Lambda}} A_i)},$$

X. We have a situation from I to IX but the obtained consequences may not be used in the antecedents of the other rules. The reasons for similar constraints are different e.g. limiting an insecure information along long chains of rules, etc.

XI. The atoms of the investigated rule (1) remain the same but some of the logical operations are changed affected by P, e.g.

$$\frac{B \leftarrow A_1 \wedge A_2 \wedge \ldots A_{k-1} \wedge \neg A_k \wedge \ldots A_z, N(P,J)}{B \leftarrow A_1 \wedge A_2 \wedge \ldots A_{k-1} \wedge \sim A_k \wedge \ldots A_z, \neg(B \leftarrow \overset{z}{\underset{i=1}{\Lambda}} A_i)},$$

A characteristic example of a similar situation is the transformation of the strong classical negation '¬' into a weak paraconsistent negation '~'.

Let us discuss the following illustrative example. On principle it is not possible that a single man is a teacher and a student at the same time. Let us denote that 'John is a teacher' by the variable Q. Then it will not be an error if we denote that 'John is a student' by ¬Q.

This is valid in the prevailing number of situations but it is inapplicable on condition (P) that John is a student in one subject in one school but he is a teacher in other subject in other e.g. sports school. After the advent of the new information P it is not possible to say that 'John is a student' is ¬Q; now it is correct to use the weak negation and ~Q will lead to a contradiction only in the cases when definite conditions hold – in the example the conditions are the subject for teaching and also the location for teaching.

The described situations from I to XI present a research for the influence of the new information P over different parts and relations between existing conclusions. In the majority of the cases the discussed situations may be used contemporary mechanisms for defeasible inference. The difference is just in the fact that P totally changes the existing a priori situation. But if P replaces the literal in the first argument in the exception E(C, A_k) then the exclusion does not change the action progress for the existing up to the advent of P things and it adds to them a new scheme that is activated if and only if when P is false. The present chapter does not contain formalizations of all the possible realizations of the situations from I to XI because the number of their combinations in all the possible realizations is too great.

We propose the application of inference by analogy to increase the effectiveness of searching. This method is viewed in details in [8-10].

2. Analogical Inference Using the Defeasible Schemes

Graph models and network flows play an important role in intelligent systems. Let graph G(N, U) has a set of arcs U and a set of nodes N. It is shown in [10] that the inference by analogy may be presented as a network flow on a graph. The geometric interpretation of this presentation is depicted in fig. 1

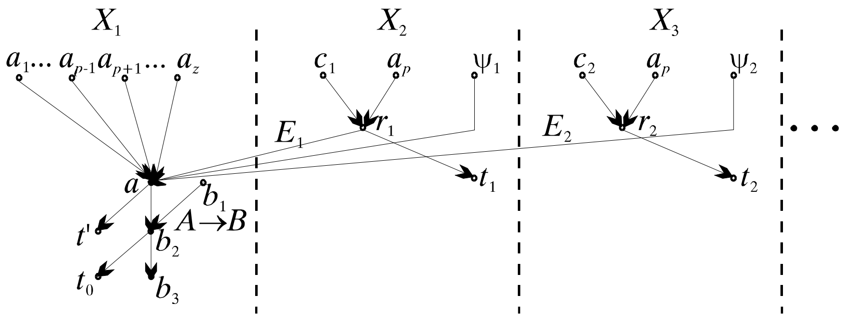


Fig. 1.

Fig. 1 is separated in different regions by dotted lines. Each region contains data corresponding approximately to a single object Xj. The set S contains all elements A_i of the exclusion E and ψ_i. The set of conclusions T contains all t_v, t_0 and t'. Then the following interrelations hold for all X,y∈N :

$$f(y,X)-f(X,y)=\begin{cases} v(a_j), \text{ iff } y\in S, \\ 0, \quad \text{ iff } y\notin S,T, \\ v(t_k), \text{ iff } y\in T. \end{cases} \tag{5}$$

Here A_i corresponds to the stream function $f(a_i,a)$ and C_v, A_p and $E(C_v,A_p)$ to the functions $f(c_i,r_i)$, $f(a_p,r_i)$, $f(r_i,a)$ respectively. The function $f(\psi_i,a)$ has an initial value of 1 if $\psi(X_j,X_1)<T$ or a value of 0 otherwise. Some corresponding exclusions $E(C_v,A_p)$, i≥j may be included in the knowledge about the object X_j. The pairs of input arcs are disjunctively connected in the nodes r_i and they are conjunctively connected in the nodes *a* and b_2. The functional dependency v has the following form:

$$v(a_j)=f(a_j,a),\ v(tj)=f(a,t_j). \tag{6}$$

The conjunction of all A_i and ψ_i is denoted with A and it corresponds to the arc (a,b_2). The implication A→B is a set of arcs (b_1,b_2) and the result of the inference $f(b_2,b_3)$ possesses a meaning of truth B. Then the inference by analogy may be presented by the following system of equalities and inequalities [6-10]:

$$f(a,b_2)-f(a_ia)\le 0;\ i=1,...z, \tag{7}$$

$$f(a,b_2)-f(r_j,a)\le 0;\ j=1,...n, \tag{8}$$

$$f(a,b_2)-f(r_j,a)\le 0;\ j=1,...n, \tag{9}$$

$$f(r_j,a)-f(c_j,r_j)\ge 0;\ j=1,...n, \tag{10}$$

$$2f(r_j,a)-f(c_j,r_j)-f(a_p,r_{1j})=0;\ j=1,...n, \tag{11}$$

$$(2n+z-1)f(a,b_2)-\sum_{i=1}^{z-1} f(a_i,a)-\sum_{j=1}^{n} f(r_j,a)-\sum_{j=1}^{n} f(\psi_j,a)=0\,, \tag{12}$$

$$2f(b_2,b_3)-f(a,b_2)-f(b_1,b_2)=0, \tag{13}$$

$$f(r_j,a) = 0 \ \text{ or } \ 1, \tag{14}$$

$$f(x,y)\ge 0; \quad (x,y)\in U, \tag{15}$$

$$f(r_j,t_j)\le 1, \tag{16}$$

$$f(a,t')\le 2n+z-2, \tag{17}$$

$$f(b_2,t_0)\le 1.$$

In this way the problem of inference by analogy is reduced to a problem of linear programming with a goal function of the kind:

$$\sum_{(x,y)\in D} f(x,y) \to max \qquad (18)$$

with constraints (7) – (17). This problem is viewed in details in [10].

The purposes of the analogical defeasible reasoning are two: check-up significance of the selected set of hypotheses and knowledge acquisition by analogy. The idea of the considered scheme of reasoning is to transfer such knowledge from the base into the goal of the transformation that this proposition reduces the significance of the considered part of formula to zero. Before the transfer, the propositions have to pass 'filters'. After filtration of wrong or insignificant information, the resulting information is applicable for the defeasible reasoning or elsewhere.

It follows from the scheme above that the elaborated by us defeasible analogy uses one of the already presented defeasible inferences combined with the inference by analogy with a goal defeating or confirming intermediate results – hypotheses that are inferred by analogy.

3. Logic Derivation in Problems of Search for Regularities and Pattern Recognition

Let us suppose that the information about some plant area has been defined in the form of a database that is interpreted as a learning sample for search (extraction) of logic regularities connecting these data. Let the set Z={X_i,Y_i}$^{m}_{i=1}$ is some database (learning sample) and the data are connected by an unknown dependence of the kind:

$$Y = f(X), \qquad (19)$$

where X and Y are multiple-valued predicates. It is required to define a dependence (regularity) (19) on the learning database Z of power m.

First let us see a case of coding for the learning sample by two-valued predicates. In this case the initial object area may be described by rules of productions like

$$\Lambda^{n}_{j=1} x_{ij} \rightarrow y_i, j = 1,..., m \ . \qquad (20)$$

Every rule of production is an implication, so it may be presented as a perfect disjunctive normal form (PDNF). In the case of a two-valued logic rule the transformation of implication to DNF is executed by the formulas:

$$A \rightarrow B = \neg A V B \rightarrow \qquad (21)$$

Therefore in the case of knowledge coding by two-valued predicates every suggestion may be presented by the rule of production and transformed to PDNF like

$$V^{n}_{i=1} x_i^{\sigma_{ij}} V y_i \ , \qquad (22)$$

where $\sigma_{ij\,i}$ is equal 0 or 1.

Further it is required to unite all formulas for the learning sample in a single logic function or system for functions, giving one-valued interpretation of the initial object area. Thus the unknown dependence Y=f(X) may be reconstructed simultaneously on the learning sample Z.

Any logic function that is written in the kind of PDNF may be reduced. Therefore the system of logic knowledge may be also reduced as a rule. Then reducing PDNF corresponding to the logic function may be interpreted as minimizations of the initial database.

We suggest the following algorithm for the PDNF reduction with an account of the object-area speciality:

1. If DNF has single-letter disjunctions x and $\neg x$, DNF is generally significant;

2. If some variable is in DNF with one sign, then delete all disjunctions containing this variable (this variable is non-informative);

3. If DNF has some single-letter disjunction x, then execute the following actions:

а) delete all disjunctions of the kind x Λ ... (rule of absorption);

б) substitute disjunctions of the kind ¬x Λ s ... on disjunctions of the kind s Λ p ... .

As a result of such reduction we obtain “the strongest” logic rules, describing the initial object area.

The described method may be used for learning of concepts (classes) in problems of pattern recognition. The synthesized concepts may be interpreted as axioms of classes (patterns) $A_k(\omega)$ in the object area defined by the learning database. Then the problem for pattern recognition is reduced to a search of a logic derivation using the Robinson method for resolutions or the Maslov back method [11].

The problem of identification for an image $\hat{\omega}$ of k-th class (pattern) on the complex image ω with a logic description D(ω) is reduced to a formula derivation:

$$D(\omega)\rightarrow\exists\hat{\omega}A_k(\omega)\,,\ \hat{\omega}\in\omega \tag{23}$$

The meaning of this formula is in the following: a complex image ω with a logic description D(ω) contains an image $\hat{\omega}$ of k-th class on which the axiom $A_k(\hat{\omega})$ is true. It allows to identify automatically and localize (select) the image of k-th class (pattern) on a complex image containing images (patterns) from M different classes $S_1, S_2, \ldots, S_M$.

Multiple applications of the logic-axiomatic method with every k=1,2,…,M allow to recognize (classify) all images of all classes, located on the complex image [11].

4. Multiple-Valued and Probabilistic Logics in Problems for Learning and Search of Regularities

The described method for search of logic regularities may be generalized on a case of multiple-valued coding for back samples and a search for multiple-valued regularities. The use of multiple-valued logics is complicated by the ambiguity of interpretation for functions of negation, implication, etc. Therefore let us discuss the most general variant in a case of use of three-valued logic.

Let a set of values for truth has the kind {0 1 2} with the following interpretation:

x=0 – false, *x*=1 – nonsense (indefinite), *x*=2 – truth.

Then let us introduce the concept of inversion as ¬*x*= 1V0, i.e. negation of truth may be either false or nonsense. This concept is defined by Table 1. This definition of inversion provides the inclusion of all possible interpretations of inversion in different logics.

Table 1

X	¬X
0	1V2
1	0V2
2	0V1

Some functions of three-valued logics are introduced. The most important of them are the characterizing functions, defined in the following way:

$$I_i(x)=\begin{cases}k-1, & if\ \ x=i,\\ 0, & if\ \ x\neq i,\end{cases} \tag{24}$$

$$J_i(x)=\begin{cases}1, & at\ \ x=i,\\ 0, & at\ \ x\neq i,\end{cases} \tag{25}$$

The main rules of operation with these functions have the kind:

$$I_\sigma(x)I_\tau(x)=\begin{cases}I_\sigma(x), if\ \sigma=\tau,\\ 0, \qquad if\ \sigma\neq\tau,\end{cases} \tag{26}$$

$$\sigma \wedge \tau = \min(\sigma,\tau), \qquad \sigma \vee \tau = \max(\sigma,\tau). \tag{27}$$

Let us use also a two-valued analogue of implication in the discussed three-valued logic, i.e.

$$A \to B = \neg A \mathrm{V} B = I_0(A) \mathrm{V} I_1(A) \mathrm{V} B \tag{28}$$

The form (28) as a negation is an extension that includes in itself a series of possible implications of a three-valued logic. Such a wide definition of main functions of logic is convenient for modelling intelligent systems in cases when it is not possible to describe intelligent processes by some concrete multiple-valued logic.

Let us return to the solution of the initial problem in the terms of the three-valued logics. Also let every line in the learning sample be described by rules of production:

$$\Lambda^{n}_{j=1} X_{ij} \to Y_i, \; i = 1,\ldots, m \;. \tag{29}$$

Then the analogue of PDNF will be the following function of three-value logic:

$$I_0(I_l(x)) = \begin{cases} I_1(x).. \vee I_{l-1}(x) \vee I_{l+1}(x).. \vee I_m(x), & if \quad x = l, \\ 0 \quad if \;\; x \neq l & . \end{cases} \tag{30}$$

$$V^{m}_{j=1} I_0(I_i(x)) \vee Y_i, \tag{31}$$

Because every regularity (knowledge) corresponding to the learning sample may be written in the kind of the suggested function of three-value logic, we want to have the possibility to present all regularities, forming the database, by a function or a system function of three-valued logic.

Single-value correspondence is easy to obtain if, for example, we multiply logically the rules of productions. It corresponds to discussions of the following type: we know partial (local) rules and thus we know all local rules (regularities) determining the global knowledge base built by the learning sample.

As a result we will obtain a three-valued function that determines the desired regularity. This function can be obtained if we use an adapted version of a reducing algorithm for multiple-valued logic as follows:

1. If some variable is in DNF with one sign ($I_j(x)$, j=const, in all disjunctions), then delete all disjunctions, containing this variable (this variable is non-informative);

2. If DNF has some single-letter disjunction $I_j(x)$, then execute the following actions:

   a) delete all the disjunctions of the kind $I_j(x)$ Λ ... (rule of absorption);

   b) substitute disjunctions of the kind $I_i(x)$ Λ s... ($i \neq j$) by the disjunctions of the kind s Λ p ... .

The result of the algorithm is a multiple-valued function built by the initial learning sample, characterizing it by a single value and giving a set of the most significant rules (regularities) defining the initial knowledge area.

By the addition of a new rule of production (new knowledge) we check if a given rule may be derived from the already existing ones or not. If it is possible to derive this rule then the function remains the same. Otherwise the knowledge base shall be enlarged adding a new rule (regularity) by a multiple-valued logic multiplying of the existing function and a new production written in the kind of a multiple-valued PDNF.

The other method of learning for concepts and search of multiple-valued regularities on defined databases is based on local-optimal logic-probabilistic algorithms [12,13]. It provides automatic synthesis, optimization (by precision) and complexity minimization for knowledge bases in terms of multiple-valued predicates with a non-defined valuation by learning databases. It allows the interpretation and the realization of synthesized knowledge (regularities) in the form of three-layer or multi-layer neural networks of a polynomial type with a self-organizing architecture [14,15].

Conclusions

An approach is introduced for inference by analogy based on three-valued logics and network flows. The approach is oriented at applications in systems of artificial intelligence and maintenance of decision making. The discussion fixes the peculiarities and the general characteristics of different types of inference by analogy.

A method is elaborated where the suppressed proof is formalized as a network flow. This approach reduces the problems of logic programming to the corresponding problems of linear programming.

The difference is investigated between the logics of the type ‘knowledge about changing knowledge’ and logics using different types of exclusions (defeasible inference).

The multiple-valued logic approach may be applied to solutions of learning problems and regularities searches in databases permitting the identical description of the object area, to structural analyses of the initial information, to reductions of it and to its changes by a measure of forming a new knowledge that is not derived from the initial data.

Logic-axiomatic and logic-probabilistic learning methods for concepts and pattern recognition have been generalized on a case of a multiple-valued logic. It is shown that synthesized concepts and recognizing rules may be realized in the kind of multiple-valued neural networks of a polynomial type and used in systems of intelligent and neural control [13-14].

The work is done under a partial support of the grants no. NIP917 of the Ministry of Science and Education – Republic of Bulgaria, RFBR N03-01-00224 - Russia, and the grant RHSF N03-06-12019v-Russia.

Bibliography

[1] P.Dung, P. Moncarella, Production systems with negation as failure, IEEE, Trans/ Knowledge and Data Engineering, vol. 14, no.2, pp. 336-352, April 2002.

[2] N. Rescher, Many Valued Logics, Basil Blackwell, Oxford 1979.

[3] Яблонский С.В. Введение в дискретную математику. Москва. Высшая Школа, 2001

[4] V. Gladun, Process of new knowledge formation. Pedagog 6, Sofia, 1994.

[5] Z. Markov, Inductive Machine Learning Methods, TEMPUS JEN 1497, Softex, Sofia, 1996.

[6] V. Jotsov, Knowledge discovery and data mining in number theory: some models and proofs, Proc. Methods and Algorithms for Distributed Information Systems Design. Institute for Information Transmission Problems of RAS, 1997, pp.197-218.

[7] K. Kowalski, Logic for Problem Solving, North-Holland Inc., Amsterdam, 1979.

[8] Halpern, J., M. Rabin, A logic to reason about likelihood, Artificial Intelligence, 1987, no. 32, pp. 379-405.

[9] V.Sgurev, V.Jotsov, Some characteristic features of the application of three-valued logical schemes in intelligent systems, Proc. First International IEEE Symp. 'Intelligent Systems' (T. Samad and V. Sgurev, Eds.), Vol. I., Varna, Bulgaria, September 10-12, 2002, pp. 171-176.

[10] V.Sgurev, V.Jotsov, Some defeasible inference methods by using network flows, J. Problems of Engineering Cybernetics and Robotics, 2001, no. 51, pp. 110-116.

[11] A.Timofeev, T.Kossovskaya, Logic-Axiomatical Method for Knowledge Representation and Recognition in Intelligent Manufacturing Systems, Elsevier, Amsterdam – Oxford – New York, 1990, pp. 3-6.

[12] A.Timofeev, Z.Shibzuchov, The Synthesis and Optimization of Knowledge Bases Based on the Local-Optimization Logic-Probabilistic Algorithms. Int. Journal of Information Theories and Applications, 1995, vol. 3, no. 2.

[13] A.Timofeev, R.Yusupov, Evolution of Intelligent Control in Adaptive Systems. Int. Journal of Adaptive Control and Signal Processing, 1992, vol. 6, pp.193-200.

[14] Тимофеев А.В., Шеожев А.В., Шибзухов З.М. Синтез нейросетевых архитектур по многозначному дереву решений. – Нейрокомпьютеры: разработка и применение, 2002, № 5-6, с. 44-49.

[15] Timofeev A.V. Polynomial Neural Networks with Self-Organizing Architecture, Int. Journal on Optical Memory and Neural Networks, 2004, v.1.

Authors' Information

Vasil Sgurev, Vladimir Jotsov – Institute of Information technologies of the Bulgarian Academy of Sciences;
P.O.Box 161, Sofia 1113, Bulgaria; sgurev@bas.bg, jotsov@ieee.org
Adil Timofeev – Saint-Petersburg Institute for Informatics and Automation of RAS,
199178, Saint-Petersburg, 14-th Line, 39; tav@iias.spb.su

XML PRESENTATION OF DOCUMENTS USED FOR DATA EXCHANGE IN MANAGEMENT OF GENETIC RESOURCES

Lina Yordanova and Vladimir Dimitrov

Abstract: In the global strategy for preservation genetic resources of farm animals the implementation of information technology is of great importance. In this regards platform independent information tools and approaches for data exchange are needed in order to obtain aggregate values for regions and countries of spreading a separate breed. The current paper presents a XML based solution for data exchange in management genetic resources of farm animals' small populations. There are specific requirements to the exchanged documents that come from the goal of data analysis. Three main types of documents are distinguished and their XML formats are discussed. DTD and XML Schema for each type are suggested. Some examples of XML documents are given also.

Keywords: XML, document's format, data exchange

Introduction

The farm animals' genetic diversity is endangered and many breeds and lines extinct every year. World Watch List [Loftuse, 1992] with endangered breeds becomes longer as well as the list of lost breeds forever. The conservation of farm animals' genetic resources needs a sustainable management of small populations in each country and region. This is connected with the establishment of information systems for data collecting, maintaining individual records for the animals and relevant data processing for population analysis.

The management of genetic resources requires separate subsystems to exchange data or to send to a centre where aggregate values to be obtained for the region or country of keeping given farm animals population. It is possible separate system nodes to use different operating systems or database management systems. This could make difficult the data exchange between them. Therefore, they need of platform independent tools what do not restrict their communications. The implementation of XML standard could be a successful approach nevertheless the target area is very complicated and there are many possible options for used documents definition.

The current work is a part of an environment for developing information system for managing small population of farm animals. The first implementation of XML standard in the environment is connected with definition of a XML format for database model and creating implementation tools for it's utilising [Yordanova, 2003].

The subject of current paper is to suggest XML formats for determined main types of documents used in the management of genetic resources. The analysis of all used documents restricts the discussion to three types:

Documents connected with data streams

Documents for data exchange in population analysis

Documents for data exchange in other kinds of data analysis.

XML Format of an Auxiliary Data Stream Document

The main type of documents exchanged in management of genetic resource is connected with data streams populating the database. A data stream is a document containing records of a same format. They could be repeated records for one animal or one record per each animal in a group. Such documents contain variety of concrete data elements and that is why their representation with a generic structure is difficult without high degree of abstraction. What we can do is to reach common XML format for the description of any data stream. If we ignore the concrete contain of the documents the result could be a very simple document tree with a root element *stream* and its descendent *dataelement*. The suggested set of elements, even a simple one, will be enough for representation the structure of any document of a data stream.

The DTD of an auxiliary data stream is given in the listing 1. The root element *stream* is considered with an attribute for its name. The element *dataelement* would be well characterized with set of attributes *name*, *type* and *description*. This element could be at least once in a separate document of such type. Although it is impossible to have only one element in a document on practical reasons “once” could be accepted conditionally.

Listing 1. The DTD of XML format of a data stream

```
<DOCTYPE stream [
<!ELEMENT stream(dataelement+)>
<!ATTLIST stream                    Name        CDATA        #REQUIRED>
<!ELEMENT dataelement               EMPTY>
<!ATTLIST dataelement               Name        CDATA        #REQUIRED
                                    Type        (CHAR|HUGEINT|BIGINT|SMALLINT|
DATE|TIME|TIMESTAMP|SMALLFLOAT|BIGFLOAT|BOOL)   "CHAR"
                                    description CDATA        #REQUIRED> ]>
```

In the XML Schema of a data stream the elements *stream* and *dataelement* are defined as Complex type and the attribute *type* has Simple type with enumerated values.

After definition of above XML format for the description of a data stream we must discuss the way of its usage. One possibility is such XML file to be attached to the document connected with the data stream in order to describe its structure. Then the application programs of the system could use it as a dictionary for data within the stream. They also could generate a set of commands inserting the data into the database. This seems to be a generic solution applicable to all possible data streams with different structure.

As a second possibility we consider the conversion of the data stream documents to the XML format that must be a solution of a separate information system. A separate XML format reflecting the structure of a given document could be developed and implemented there.

An example of a XML file containing a description of a data stream is given in listing 2. This one describes the data stream named Semen from the information system “Cryo” [Groeneveld, 2002].

Listing 2. A XML file with description of a data stream

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE stream SYSTEM "file:///home/lina/teza/potok.dtd">
<stream name="Semen">
   <dataelement description="The external id for the bull"
             name="ext_animal"/>
   <dataelement description="The external id for the reporting unit"
             name="unit"/>
   <dataelement description="The date of semen delivery"
             name="delivery_dt" type="DATE"/>
   <dataelement description="The date of semen production"
             name="production_dt" type="DATE"/>
   <dataelement description="The number of doses total"
             name="no_doses" type="SMALLINT"/>
   <dataelement description="The number of straws per dose"
             name="no_straws" type="SMALLINT"/>
   <dataelement description="The type of straws"
             name="type_straws"/>
   <dataelement description="Quality mark- motility post freezing %"
             name="mot_post" type="SMALLINT” />
   <dataelement description="Quality mark- after collection %"
             name="mot_after" type="SMALLINT” />
   <dataelement description="Semen certificate identification"
             name="certificate_id" type="SMALLINT"/>
</stream>
```

The XML file from the listing 2 is created and validated through defined DTD with XML editor what could be done also by any program application working with XML standard.

The documents of a data stream could be obtained on various approaches. Nevertheless which approach is used the documents of data streams have the logical structure defined above. The defined XML format of their description can be used in document exchange. In consequence data from corresponding XML document of a data stream could be inserted into the database via middleware software.

There are practical cases in management of genetic resources with manual filling in paper documents and than converting in any electronic form. The most common situation is conversion to a comma separated value format. Technical tools produce files of the same format often. It is possible also that the concrete document to be converted to its own XML format.

The documents of data streams are connected in general with the GUI forms for inserting and manipulating data [Yordanova, 2000]. The GUI forms in the environment are created according to them. The description of a data stream does not contain the access actions to the database elements in order to insert data from the stream.

We should consider that the data streams are mainly connected with primary data collecting. It is very seldom the data they contain to be retrieved from another database but if this is the case then the approaches from next chapters are applicable.

XML Formats of Documents for Population Analysis

The data exchange in management of genetic resources covers different groups of data depending on the purpose of their analysis.

The population analysis needs of data about the animal origin. Such analysis requires obtaining of individual inbreeding coefficients, effective population size and other genetic parameters that the manager of breed conservation program could choose. Then the data exchange between the center and peripheral nodes must include: the identification of the animal, the identifications of its parents, gender and birth date or birth year. This set of data is a minimum, enough for calculation the genetic parameters for population analysis.

We define the XML format of document containing data for animal origin and its individual identification via DTD (listing 3) and XML Schema.

Listing 3. The DTD of a document for data exchange in population analysis

```
<DOCTYPE pedigree [
<!ELEMENT pedigree(animal+)>
<!ATTLIST pedigree                    name    CDATA    #REQUIRED>
<!ELEMENT animal(birthdt, sire, dam)>
<!ATTLIST animal                      ext_id  CDATA    #REQUIRED
                                      unit    CDATA    #REQUIRED
                                      gender  (F | M)  #REQUIRED>
<!ELEMENT birthdt(#PCDATA)>
<!ELEMENT sire(#PCDATA)>
<!ATTLIST sire                        ext_id  CDATA    #REQUIRED
                                      unit    CDATA    #REQUIRED>
<!ELEMENT dam(#PCDATA)>
<!ATTLIST dam                         ext_id  CDATA    #REQUIRED
                                      unit    CDATA    #REQUIRED> ]>
```

The root element is called *pedigree* and its sub element is *animal*. The sub element is defined to be at least once in the document. It has attributes *ext_id*, *unit* and *gender* and sub elements *birthdt*, *sire* and *dam*. The elements *sire* and *dam* should have the same attributes like the element *animal*. In all cases the attribute *ext_id* means an external identification of an animal depending on the unit that reports the animal. Here it is not convenient to use for *ext_id* type ID, because in common case it is not unique. Two units can report two animals with the same identification. That requires the couple of elements (*ext_id, unit*) to be unique.

The element or attribute connected with information about the breed to which the animal belongs is not included. This is done because the population analysis supposes that the data collecting concerns animals from the same breed. If it is necessary one could mark the breed into the name of the document or to add an element breed. For the given example the breed name is stored in the attribute *name* of the root element (listing 4). The example document is for population analysis according the XML definition explained above. The data is for two family couples. Four animals (the progeny of the families) are included. The document is checked for validation with corresponding XML schemas through program applications that use DTD or XML Schema.

Listing 4. An example for data exchange in population analysis

```
<?xml version="1.0" encoding="UTF-8"?>
 <!DOCTYPE pedigre SYSTEM "file:///home/lina/teza/pedigre.dtd">
 <pedigre name="minipigs">
   <animal ext_id="1677" gender="F" unit="12">
         <birthdt>23.12.1998</birthdt>
         <sire ext_id="3456" unit="12">04.04.1997</sire>
         <dam ext_id="2345" unit="12">12.03.1996</dam> </animal>
   <animal ext_id="1698" gender="F" unit="12">
         <birthdt>23.12.1998</birthdt>
         <sire ext_id="3456" unit="12">04.04.1997</sire>
         <dam ext_id="2345" unit="12">12.03.1996</dam> </animal>
   <animal ext_id="1701" gender="M" unit="12">
         <birthdt>3.11.1998</birthdt>
         <sire ext_id="5003" unit="12">12.08.1996</sire>
         <dam ext_id="4312" unit="12">12.03.1996</dam> </animal>
   <animal ext_id="1702" gender="F" unit="12">
         <birthdt>3.11.1998</birthdt>
         <sire ext_id="5003" unit="12">12.08.1996</sire>
         <dam ext_id="4312" unit="12">12.03.1996</dam> </animal>
   </pedigre>
```

XML Formats of Documents for Other Kinds of Analysis

The common structure for documents exchanged in the management of genetic resources with the goal to perform other kinds of analysis contains mandatory animal identification and its one or multytrait measurements. A possible generic XML scheme of such document is given in the listing 5 with DTD.

Listing 5. The DTD of a document for data exchange in other kinds of analysis

```
<DOCTYPE data [
<!ELEMENT data(animal+)>
<!ELEMENT animal(trait+)>
<!ATTLIST animal              ext_id  CDATA      #REQUIRED
                              Unit    CDATA      #REQUIRED>
<!ELEMENT trait(measurement+)>
<!ELEMENT measurement EMPTY>
<!ATTLIST measurement         Date    CDATA      #REQUIRED
                              Value   CDATA      #REQUIRED
                              Type    CDATA      #REQUIRED> ]>
```

The root element named *data* is consisted by at least one sub element *animal*. It has attributes *ext_id* and *unit*s as well as one sub element trait meet more than once. The element *trait* connects any investigated trait to an animal. It is possible a document to have data about more traits that complete a process. One animal could be measured many times for a trait. That is why it is appropriate to have a sub element *measurement* repeated many times. The measurement is characterized with attributes or sub elements *date*, *value* and *type*. The measurements type requires from the application programs to maintain with external coding for different types of measurements.

The animal identification is given with attributes *ext_id* and *unit*, which means maintaining external identification for both objects, *animal* and *unit*. This requires from the software to obtain a new or to retrieve existing internal identification from the database. The last one is used according the system supporting unique identification for the animals. About reporting unit it is most possible to have only second situation.

A XML document obtained on the defined format is given in the listing 6. The document is validated according the XML format definition. It contains weight measurements of two animals.

Listing 6. An example of XML document with data for the trait weight

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE data SYSTEM "file:///home/lina/teza/data.dtd">
<data>
   <animal ext_id="6768" unit="11">
       <trait name="weight">
            <measurment date="12.03.2000" type="bdw" value="0.3"/>
            <measurment date="12.05.2000" type="wdw" value="0.6"/>
            <measurment date="14.06.2000" type="rw" value="1.9"/>
            <measurment date="13.07.2000" type="us" value="2.4"/>
       </trait>
   </animal>
   <animal ext_id="4546" unit="11">
       <trait name="weight">
            <measurment date="14.04.2000" type="bdw" value="0.2"/>
            <measurment date="15.06.2000" type="wdw" value="1.7"/>
            <measurment date="14.07.2000" type="us" value="2.6"/>
       </trait>
   </animal>
</data>
```

For the documents of data streams and for the data exchange it is preferable to get data automatically. Retrieving data from the database is common for the documents for data exchange. Very seldom in small populations' management data is generated automatically from computer systems with measure tools. Then generated files are usually in CSV format. Their transformation to the XML format is not a difficult task. Better solution is the XML format defined for a data stream to be used as a meta form of such CSV file. The middleware software can operate with data according the description of the document structure in XML format.

Middleware Software for Data Exchange

In current work the XML documents are used for data transfer after generating or parsing by dynamic program applications. The main processes discussed here are:
Parsing and retrieving data from a XML document in order to be inserted into the database
Creating a XML document retrieving data from the database.

The related program codes use the XML formats of the documents in data exchange defined above.

The program code for inserting data into the database from a XML document

Let the XML document that is going to be parsed and analyzed in order to insert data into the database has the structure from listing 3. Let the target database has the conceptual database scheme for small populations management [Yordanova, 2000], where the main relations are named TRANSFER, UNIT and ANIMAL. Then the algorithm for parsing the documents and inserting the data into the database includes the next steps:

```
Begin                         (#begin
DBI connection                (# Connection to the database
Objects and variables         (# Declaration of objects and variable
SQL statements                (# Definition of SQL statements -
Foreach $row                  (# For each animal retrieving of:
  Sire/ID, sire/unit             (# db_sire from TRANSFER(ext_id, unit)
  Dam/ID, dam/unit               (# db_dam from TRANSFER(ext_id, unit)
  Animal/unit                    (# db_unit from UNIT via unit
  get_next_val(sequence_name)    (# new db_animal identification
  INSERT into ANIMAL, TRANSFER   (# Inserts in TRANSFER and ANIMAL
end foreach                   (# End of the cycle
Db disconnect                 (# Disconnect the database
End                           (#End
```

If any parent does not have internal database identification then a new one has to be obtained in the relation TRANSFER and recorded into the relation ANIMAL. The inserts will not be done if there is another animal with the same values of (ext_id, unit). The released program is a Perl code and uses the module XML::XPath of Matt Sergeant that implements the XPath standard and allows fast search and parsing elements of XML document via a tree of document's nodes.

The program code for retrieving data from database

The function of the code is connected with the XML format for population analysis (listing 3). The algorithm is separated to two main steps:

1. Getting via Query a set of tuples, containing the external identification of all active animals and their parents as well as their reporting units, birth dates and gender.
2. Recording data from the tuples in XML document elements.

This code uses Perl&XML module XML::Writer that allows creating of XML document via defined objects.

Conclusion

The usage of XML standard makes the data exchange in management genetic resources much more flexible and platform independent. There are a lot of program applications that work with many operating systems and could facilitate implementation of defined XML formats of documents for data exchange. The user could create the XML documents containing the description of data streams using: 1) XML editors that apply XML declarations DTD or XML Schema; 2) program applications that includes processing and validating XML documents through schema.

The other XML documents for data exchange in all kinds of data analysis for small populations could be created and used via briefly presented here program codes.

The defined XML formats could be extended and could become a base language for data exchange in management of genetic resources.

Bibliography

[Groeneveld, 2002] E. Groeneveld, L.Yordanova and S.J. Hiemstra. An Adaptable Management Support System for National Genebank, 7th World Congress on Genetics Applied to Livestock Production, August 19-23, 2002, Montpelier, France, Session 26, Management of genetic diversity, 513-516, 2002.

[Loftuse, 1992] R. Loftuse and B. Scherf, World Watch List for Domestic Animal Diversity, FAO Rome, 1992.

[Yordanova, 2000] L. Yordanova, E. Groeneveld. The implementation Strategy of Information System with Existing Data Streams, Vortrastagung der Deutschen Gesellschaft fuer Zuchtungkunde e. V. (DGfZ) u. der Gesellschaft fur Tierzucht Wissenschaft (GfT), A28, 2000.

[Yordanova, 2003] L. Yordanova, E. Groeneveld, Vl. Dimitrov. A XML Approach for Support DB Modeling. In: Proceedings of the Thirty Second Spring Conference of the Union of Bulgarian Mathematicians, 297-302, 2003.

Authors' Information

Lina Yordanova – Thracian University, Department of Informatics, Mathematics and Physics, Stara Zagora, 6000, Studentsko gradtche, Bulgaria; e-mail: lina@uni-sz.bg

Vladimir Dimitrov – Sofia University, Faculty of Mathematics and Informatics, Sofia, Bulgaria; e-mail: cht@fmi.uni-sofia.bg

XML EDITORS OVERVIEW AND THE CHALLENGE TO BUILD XML-ORIENTED EDITOR FOR MEDIAEVAL MANUSCRIPT DESCRIPTIONS[1]

Pavel Pavlov

Abstract: The paper presents an overview of XML and software tools for its use, with an emphasis on XML editors. Based on the experience of two Bulgarian projects on preparing electronic descriptions of mediaeval manuscripts from the 1990es, we define the following requirements to the editor used for manuscript cataloguing: minimum elements on the screen; arrangement of elements according to the practice in the subject domain; supplying default values whenever this is possible; supplying possible values in combo boxes whenever this is possible; and ease of data entry (in Bulgarian with possibility to enter mediaeval text fragments in Old Cyrillic). These requirements were taken into account for the development of a specialized editor, XEditMan, which is presented in the article. Currently, 200 descriptions of manuscripts are available which were entered and edited using XEditMan. The average time for data entry with the editor is about three times less than the time spent in previously used software tools in Bulgaria.

Keywords: XML, XML editors, mediaeval manuscript cataloguing, XEditMan.

Introduction

The interest to digitisation of scientific and cultural heritage has been considerably growing in the last decades. The electronic access to cultural heritage is one of the priority areas of the European Commission. The Cultural Heritage Applications Unit of the Information Society Directorate General of the European Commission promoted the priorities in the field through a document known as *The Lund Principles* which put emphasis on *making visible and accessible the digitised cultural and scientific heritage of Europe*; *coordination of efforts*; *development of a European view on policies and programmes*, as well as of *mechanisms to promote good practice in a consistent manner* [Lund Principles, 2001]. Currently, most large institutions from the cultural sector are taking measures to make their collections available online. The first step in this direction is to provide access to cataloguing information about the holdings in a specific collection. In Bulgaria, this still is not done for the manuscript collections of any repository.

One of the recognised approaches on world wide scale is to use XML to present data on manuscripts. In this paper, we first give a brief overview on XML and tools, which allow its use. Then we present the experience of two Bulgarian projects in the field of manuscript cataloguing and formulate several basic requirements to a specialised editor for entering data on mediaeval Slavonic manuscripts and present our work in this direction.

These requirements were taken into account for the development of XEditMan, an XML editor for mediaeval manuscripts. The use of the editor is illustrated. One basic advantage is higher accuracy of entered data and better time characteristics (about three times faster data input compared to previously used tools).

Overview: XML and Various Types of Tools

XML (eXtensible Markup Language) is an open standard developed by the W3C (World Wide Web Consortium). It has two interconnected applications: web presentation and data exchange. One distinguished feature of XML is that it separates the encoding of data from program logic and user interface code. This leads to platform independence and reusability of resources.

XML is based on the Standard Generalized Markup Language (SGML), an ISO standard which puts the basics of many developments in the field of electronic transmission of documents through defining tag sets forming the DTD (document type definition) which are used to mark-up the electronic text and allow easy processing [ISO,

[1] The research presented here is partially supported by the project KT-DigiCult-Bg (FP6) and by the ICT Agency in Bulgaria.

1986]. SGML was designed in 1986 and was oriented towards work with large collections of documents, not towards the Web. The DTD practice of SGML was expanded in XML in order to offer more data types and allow easy building of hyperdocuments. HTML was another (earlier than XML) successor of SGML designed for visualization of documents on the Web, but its orientation to present the document layout leads to limitations on the presentation of data for processing, not just for display.

An XML application unifies several parts stored separately: data, structure, presentation and program access.

The XML data are stored as the actual XML document (it is usually called the document instance). This document contains data and mark-up elements.

The second element is a file, which contains the rules defining the specific XML document's elements, attributes, entities, and processing instructions. In the beginning, following the SGML principles, a DTD file served this purpose. Later XML Schema specification started to be used in order to solve several shortcomings of the DTD: it is too restrictive to introduce new elements and does not offer support for a number of data types. XML Schema allows creating both simple and complex data types and associating them with new element types. Thus specialists working in various fields and preparing specific documents may define the structure of their documents with a great freedom.

XSL (Extensible Stylesheet Language) is the part, which ensures presentation. It allows one to render XML elements using formatting objects. For example, CSS (Cascading Style Sheets) outputs documents in HTML format. XSLT (XSL Transformation), outputs XML document into text, HTML, or any mark-up language.

The last component is called DOM (Document Object Model) which allows accessing data encoded in XML through programs. Thus data can be separated from the specific platform.

There are several types of XML-oriented tools. The *XML editors* are used to create structured documents. From the point of view of the user, it is important to have an easy and understandable interface for entering data. The task of collecting manuscript descriptions in XML format inevitably raises the question how the data will be entered. Other types of tools are necessary basically for the IT staff, such as software for the *creation of DTD's or XML Schemas*, parsers for *validating XML* files (applications which check the documents against the DTD or the Schema); parsers for *parsing XSLT* (they prepare XML documents for presentation as text, HTML or PDF by applying the XSLT stylesheet language). Technical staff may also need a specialised editor for speeding the *creation of XSLT stylesheets*. For the work on manuscript catalogue descriptions, most important are the editor and the parser. We provide below some explanation and examples of tools from the various categories.

Editors

Editors allow users to create and edit XML documents using the proper DTD. XML editors often provide additional functionality, for example validation of the document against the DTD or schema. To facilitate the user in his/her work, editors rely on two basic methods:

- Use of colours to distinguish elements, attributes, and text, etc. for easy reading.
- Providing clickable lists of possible elements and attributes at the current cursor point in the document. These lists usually are located in the left pane of the editor window.

Popular professional XML editors are XMetaL® 4[1], xmlspy® 2004[2], NoteTab Pro[3]. Free editors are XMLCooktop[4], Bonfire Studio 1.4[5], NoteTabLight[6], Xeena[7], Xerlin[8] etc. The illustration on Fig. 1. shows a snapshot from xmlspy® 2004.

[1] http://www.sq.com/, last visited on 25 April 2004.

[2] http://www.xmlspy.com, last visited on 25 April 2004.

[3] http://www.notetab.com, last visited on 25 April 2004.

[4] http://www.xmlcooktop.com/, last visited on 25 April 2004.

[5] http://www.nzworks.com/bonfire/download.asp, last visited on 25 April 2004.

[6] http://www.notetab.com, last visited on 25 April 2004.

[7] http://www.alphaworks.ibm.com/tech/xeena, last visited on 25 April 2004.

[8] http://www.xerlin.org/, last visited on 25 April 2004.

Validating Parsers

Usually, the professional XML editors contain a built-in validator. Some are internal to the editor and others use a separate piece of software.

XSLT Parsers

The XSLT parsers play the role of formatting engines. They output data most often in HTML, text, PDF. They are sometimes part of the XML editor.

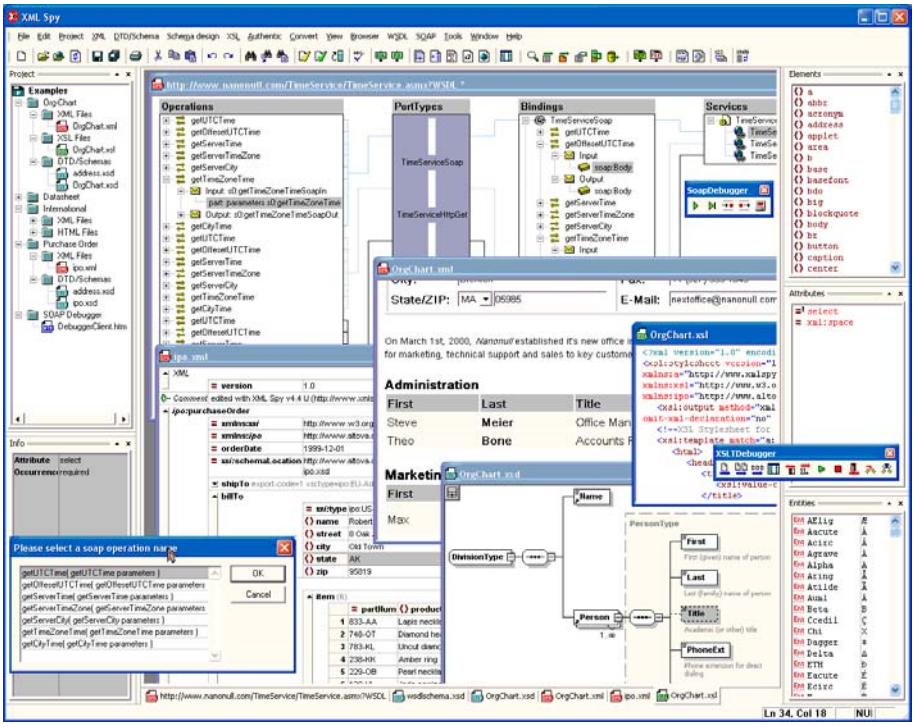


Fig. 1. Snapshot from *xmlspy® 2004, source:* http://www.xmlspy.com

The Bulgarian Experience in Preparing XML Editor for Mediaeval Manuscripts Descriptions

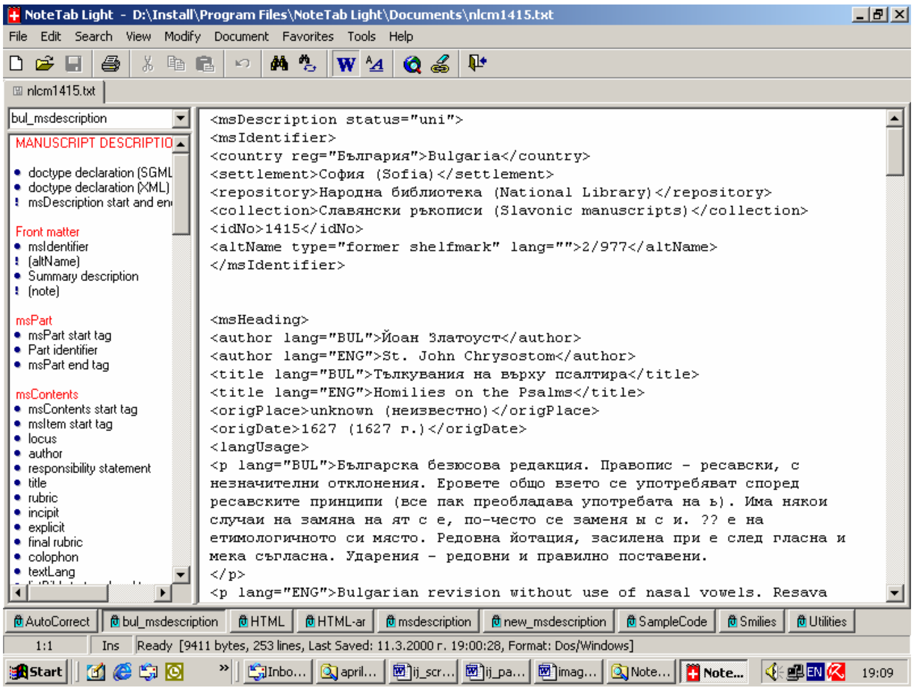
The idea to use a markup language for manuscript descriptions goes back to the 1990es. With the advent of mark-up languages, a team in Bulgaria suggested in 1994-95 a structured description of manuscript data built as an extension of Text Encoding Initiative [TEI] of that time. A project called *The Repertorium of Old Bulgarian Literature and Letters* was started as "…an archival repository capable of encoding and preserving in SGML (and, subsequently, XML) format of archeographic, palæographic, codicological, textological, and literary-historical data concerning original and translated medieval texts represented in Balkan Cyrillic manuscripts" [Repertorium], [Miltenova et al., 2000]. This is a typical repository project aimed to answer researchers' (not librarians') needs. The computer model based on SGML is discussed in [Dobreva, 2000]. Currently there are 300 manuscript descriptions, which should be made available on the project website[1].

In the late 90es, the National Library "St. Cyril and St. Methodius" and the Institute of Mathematics and Informatics became associated members of the MASTER project *(Manuscript Access through Standards for Electronic Records)* supported by the EC [MASTER]. Within this project, a TEI-conformant DTD for mediæval manuscripts was developed with the ambition to answer the needs of all repositories in Europe, and software for making and visualising records on manuscripts. The MASTER standard (may be with small revisions) was adopted by the TEI in May 2003.

[1] On April 25, 2004 there was still a message that link is disabled for file update.

In the Repertorium project, data were entered through Author/Editor software product of SoftQuad Company, a predecessor of HoTMetaL and currently available XMetaL editors. In the data entry process, users were seeing all elements from the description on the screen (surrounded by the SGML delimiters, e.g. `<P> </P>`) which formed long list spread on several screens. This was not very convenient, if we also add that the appearance of elements followed the structure of the DTD, which is not the same as the sequence of elements natural for the people working with mediaeval manuscripts. The organization of work was oriented towards one specialist working on one description, which produced results of different quality in the group of almost 10 specialists working on the descriptions [Dobreva, Jordanova, 2000]. The description data were entered in English which made them usable by English language speakers. To enter fragments of Old and Middle Bulgarian texts a designated font was created, and in data entry the LANG attribute was assigned to elements containing text in Old or Middle Bulgarian while for all other languages was supposed that they contain texts in English.

The experience of the pilot catalogue descriptions within the MASTER project was different in two directions: the data were entered in both Bulgarian and English with the idea that this will serve larger research community, and the editor used for the tests was NoteTabLight[1] (see Fig. 2). To enter data on both languages, elements were repeated with including of the LANG attribute showing the language of the data entered within the specific element.



```
<msDescription status="uni">
<msIdentifier>
<country reg="България">Bulgaria</country>
<settlement>София (Sofia)</settlement>
<repository>Народна библиотека (National Library)</repository>
<collection>Славянски ръкописи (Slavonic manuscripts)</collection>
<idNo>1415</idNo>
<altName type="former shelfmark" lang="">2/977</altName>
</msIdentifier>


<msHeading>
<author lang="BUL">Йоан Златоуст</author>
<author lang="ENG">St. John Chrysostom</author>
<title lang="BUL">Тълкувания на върху псалтира</title>
<title lang="ENG">Homilies on the Psalms</title>
<origPlace>unknown (неизвестно)</origPlace>
<origDate>1627 (1627 г.)</origDate>
<langUsage>
<p lang="BUL">Българска безюсова редакция. Правопис - ресавски, с
незначителни отклонения. Еровете общо взето се употребяват според
ресавските принципи (все пак преобладава употребата на ь). Има някои
случаи на замяна на ят с е, по-често се заменя ы с и. ?? е на
етимологичното си място. Редовна йотация, засилена при е след гласна и
мека съгласна. Ударения - редовни и правилно поставени.
</p>
<p lang="ENG">Bulgarian revision without use of nasal vowels. Resava
```

1:1 Ins Ready [9411 bytes, 253 lines, Last Saved: 11.3.2000 г. 19:00:28, Format: Dos/Windows]

Fig. 2. Example of data entry of manuscript descriptions in NoteTabLight

Unlike the Author/Editor interface, in this case all available elements can be seen on the left pane of the window, and the person who enters data should click on the specific element which is needed. This led to high number of erroneously located elements and a heavy workload on editing the descriptions.

The experience from both projects stimulated us to formulate the following requirements to a specialized editor:

- The number of elements visible on the screen should be the minimum possible. Lengthy lists of elements confuse users who are specialists in mediaeval studies or library cataloguing who normally would enter the data. This also slows down the process of data entry and leads to mistakes.
- The sequence of appearance of the elements should follow the logic of the subject domain, not of the XML DTD.

[1] http://www.notetab.com/ntl.php, NoteTabLight – free editor offered as an alternative to the commercial professional editor NoteTab Pro.

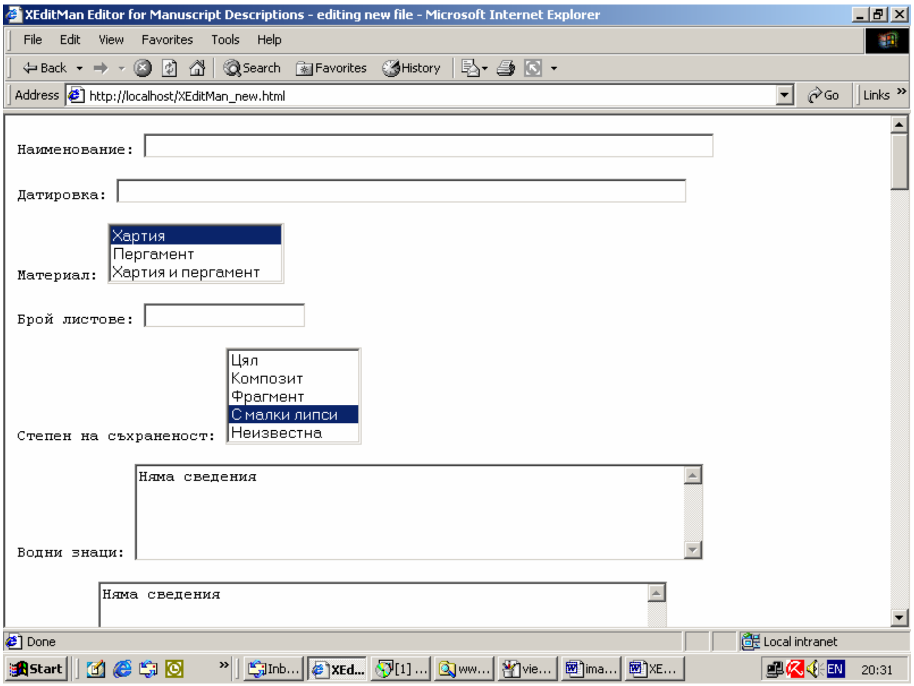
- Quite often, the value of element is “No information” (this is because in some cases there is no information on the matter since these descriptions are based on preliminary research work on the manuscripts). To avoid multiple entry of this phrase, the value can be supplied in advance and changed by the person who enters data whenever this is needed.
- There are several elements where the values are chosen from a list: for example, names of repositories, cities, values of attributes for language, etc. To avoid errors, combo boxes with possible values could be supplied.
- Ease of entering data written in Old Cyrillic script.
- Interface in modern Bulgarian (thus specialists who enter the data see names of elements which are familiar to them, and do not have to become acquainted in details with the DTD itself.

Taking these considerations into account, we decided to create a specialized editor, which takes into account these requirements in its interface. The decision to create a home-made editor was taken after the consideration of possibilities to adapt existing commercial editors. Since the left pane with all elements listed and the alphabet encoding could not be solved satisfactorily, we decided to create a tool which could be easily installed on a computer with a running Microsoft Internet Explorer browser and Internet Information Server.

XEditMan: A XML Editor for Mediaeval Manuscripts Descriptions

XEditMan is actually a set of tools: editor for new document, editor for existing document and a visualisator. The editor is currently oriented towards the use of the MASTER DTD for manuscript descriptions adopted by TEI[1].

Data Entry: The New Document Editor and the Editor for an Existing Document



Наименование:

Датировка:

Материал: [✓ Хартия] Пергамент Хартия и пергамент

Брой листове:

Степен на съхраненост: Цял Композит Фрагмент [✓ С малки липси] Неизвестна

Водни знаци: Няма сведения

Няма сведения

Fig. 3. XEditMan: Data entry interface

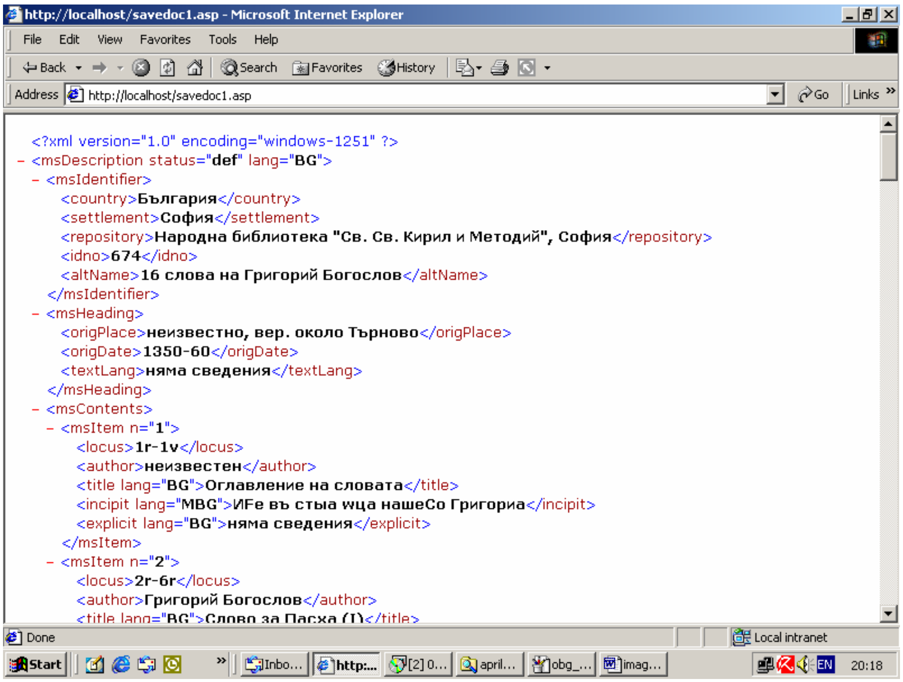
The editor of new document is used to enter data arranged in the order, which is natural for the subject domain. In two cases repetitive elements are possible: description of scribes and description of texts appearing in the manuscript. In these cases, during the first entry the user supplies the data on the first scribe (respectively, text)

[1] The relevant materials can be found on http://www.tei-c.org.uk/Master/Reference/, last accessed on April 25, 2004.

and the total number of scribes (texts). Then when the description is opened with the editor of existing texts, the respective number of elements appear in the window and make possible the entry of the information on the other scribes (respectively texts). Fig. 3 presents part of the data entry window, in which we see several types of elements: with no value; with supplied values, and combo boxes for choice of possible value.

The first two fields on Fig. 3, name and date, are typical fields for direct data entry. The third and fifth elements, material and manuscript status, are supplied with combo boxes containing possible values. In the last two elements (the visible one is a watermark) the value “No information” is entered by default. If there is no information about the element, the specialist who enters data does not have to bother with writing this text again and again.

After the data are entered, the users clicks a button “Save the description” which generates the XML document conformant to the MASTER project DTD (see Fig. 4); now all element identifiers appear according to the DTD.


http://localhost/savedoc1.asp - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://localhost/savedoc1.asp
Go Links
Done
Local intranet
Start Inbo... http:... [2] 0... april... obg_... imag...
EN 20:18

```
  <?xml version="1.0" encoding="windows-1251" ?>
- <msDescription status="def" lang="BG">
  - <msIdentifier>
      <country>България</country>
      <settlement>София</settlement>
      <repository>Народна библиотека "Св. Св. Кирил и Методий", София</repository>
      <idno>674</idno>
      <altName>16 слова на Григорий Богослов</altName>
    </msIdentifier>
  - <msHeading>
      <origPlace>неизвестно, вер. около Търново</origPlace>
      <origDate>1350-60</origDate>
      <textLang>няма сведения</textLang>
    </msHeading>
  - <msContents>
    - <msItem n="1">
        <locus>1r-1v</locus>
        <author>неизвестен</author>
        <title lang="BG">Оглавление на словата</title>
        <incipit lang="MBG">ИFе въ стыа wца нашеСо Григориа</incipit>
        <explicit lang="BG">няма сведения</explicit>
      </msItem>
    - <msItem n="2">
        <locus>2r-6r</locus>
        <author>Григорий Богослов</author>
        <title lang="BG">Слово за Пасха (I)</title>
```

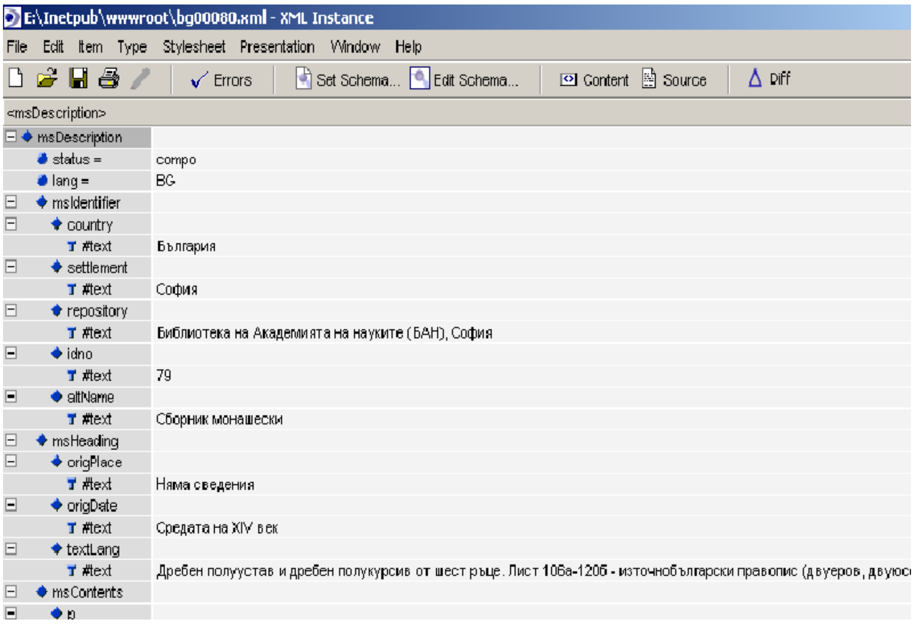
Fig. 4. A Sample of XML Document which is being saved after data were entered in XEditMan

The generation of this document is done in a way, which guarantees successful validation. This organization of work combines easy data entry and DTD-conformant result. For this reason, the editor does not include an internal validator. It is suggested to use a commercial editor for validation purposes and for cases where the interface of the editor does not support too specialized cases appearing sometimes in manuscript descriptions, like quoting within the content of specific element. We made experiments with the use in such cases of TurboXML editor (see Fig. 5). The work on XEditMan was done with the idea to cover the mass case of data entry on manuscripts. In very specific cases which appear rarely (like nesting quotes, bibliographical references and corrections to the Old Bulgarian texts), but would require too many complications in the interface, specialists who are familiar with the DTD could enter data using commercial editors which arrange the document as it is saved in XML format.

Data Visualisation

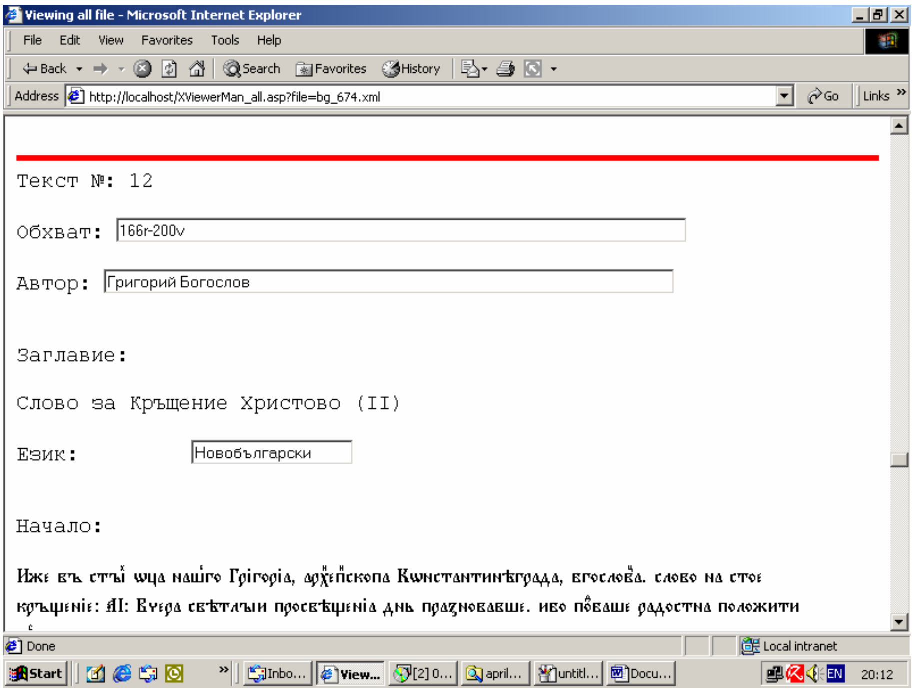
After the data are entered, they can be visualized with the help of another component of the editor. There are two modes of visualization: visualization of the complete document, which is demonstrated on Fig. 6, or visualization of selected elements from the description.

To make possible further processing of data on sets of manuscript descriptions, we are currently working on a program interface, which would extract data from XML descriptions into a database. This would provide tools for group queries.


E:\Inetpub\wwwroot\bg00080.xml - XML Instance
File Edit Item Type Stylesheet Presentation Window Help
Errors Set Schema... Edit Schema... Content Source Diff
<msDescription>

msDescription	
status =	compo
lang =	BG
msIdentifier	
country	
#text	България
settlement	
#text	София
repository	
#text	Библиотека на Академията на науките (БАН), София
idno	
#text	79
altName	
#text	Сборник монашески
msHeading	
origPlace	
#text	Няма сведения
origDate	
#text	Средата на XIV век
textLang	
#text	Дребен полуустав и дребен полукурсив от шест ръце. Лист 106а-120б - източнобългарски правопис (двуеров, двуюс
msContents	
p	

Fig. 5. An Example of Description Prepared in XEditMan Visualised in TurboXML editor


Viewing all file - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://localhost/XViewerMan_all.asp?file=bg_674.xml
Go Links

Текст №: 12

Обхват: 166r-200v

Автор: Григорий Богослов

Заглавие:

Слово за Кръщение Христово (II)

Език: Новобългарски

Начало:

Иже въ стъı ѡца нашего Грiгорiа, архıепископа Кѡнстантинѣграда, бгослова. слово на стое крьщенiе: аI: Вчера свѣтлыи просвѣщенiа днь празновавше. ибо пѣваше радостна положити

Done
Local intranet
Start Inbo... View... [2] 0... april... untitl... Docu... EN 20:12

Fig. 6. An example of Visualized Data

Conclusion

The paper presented a brief overview of XML and the current trends in developing tools for its use. It formulated several basic requirements for the development of a specialized editor on mediaeval manuscripts, which guarantee faster and more accurate data entry.

It also presented the experience of the author in designing XEditMan, a specialized editor for manuscript descriptions. XEditMan was tested in the Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences and is now used to enter data on Bulgarian manuscripts stored in Bulgaria. Two hundred descriptions are already available by the date of preparation of the paper (25 April 2004) based on the catalogue [Ikonomova et al., 1982].

This work is made as part of the current project Knowledge Transfer for the Digitization of Cultural and Scientific Heritage to Bulgaria, coordinated by the Institute of Mathematics and Informatics and supported by the Framework Programme 6 of the European Commission.

The basic idea is to provide in the next months a set of 800 manuscript descriptions which form about 1/10 of the manuscripts stored in Bulgaria. The first group of manuscripts, which was chosen, consists of Bulgarian manuscripts.

This work is extensible in two ways – more manuscripts could be added to the collection, and more data could be supplied at a later stage. For this reason, we believe that this initiative will contribute to the more adequate presentation of the cultural heritage of Bulgaria.

Bibliography

[Dobreva, 2000] M. Dobreva, A Repertory of the Old Bulgarian Literature: Problems Concerning the Design and Use of a Computer Supported Model, In: A. Miltenova, D. Birnbaum (eds.), Medieval Slavic Manuscripts and SGML: Problems and Perspectives, Sofia, Academic Publishing House, 2000, pp.. 91-98.

[Dobreva, Jordanova, 2000] M. Dobreva, M. Jordanova, *Some Psychological Aspects of Computer Modeling of Complex Objects,* In: A. Miltenova, D. Birnbaum (eds.), Medieval Slavic Manuscripts and SGML, Problems and Perspectives. Prof. M. Drinov Academic Publishing House, Sofia, 2000, pp. 295–310.

[Ikonomova et al., 1982] A. Ikonomova, D. Karadzhova, B. Christova, Bulgarian Manuscripts from 11 to 18 cent., stores in Bulgaria. Vol. 1. Sofia, 1982. (In Bulgarian – Икономова, А., Д. Караджова, Б. Христова. Български ръкописи от XI до XVIII век, запазени в България. Своден каталог, том I, НБКМ, София, 1982.)

[ISO, 1986] International Organization for Standardization, *ISO 8879: Information processing – Text and office systems - Standard Generalized Markup Language (SGML),* Geneva, ISO, 198б.

[Lund Principles, 2001] http://www.cordis.lu/ist/ka3/digicult/lund_principles.htm—*eEurope: creating cooperation for digitisation (Lund Principles)*

[TEI] http://www.tei-c.org/—*Text Encoding Initiative Website*

[MASTER] MASTER, http://www.cta.dmu.ac.uk/projects/master/, website of the MASTER project.

[Miltenova et al., 2000] A. Miltenova, D. Birnbaum (eds.), *Medieval Slavic Manuscripts and SGML: Problems and Perspectives*, Sofia, Academic Publishing House, 2000, 372 pp.

[Repertorium] *Repertorium*, http://clover.slavic.pitt.edu/~repertorium/index.html — website of the *Repertorium of Old Bulgarian Literature and Letters*

Author Information

Pavel Pavlov – Sofia University, Faculty of Mathematics and Informatics, Assistant Professor; 5 J. Bouchier Blvd, Sofia, Bulgaria; e-mail: pavlovp@fmi.uni-sofia.bg

USING WEB SITES EXTERNAL VIEWS FOR FUZZY CLASSIFICATION

Georgi Furnadzhiev

Abstract: In the paper a fuzzy sets implementation into web sites classification is considered. Web sites external features are addressed and the possibility to use them for the classification is proved. An example with five different categories classification is given.

Keywords: web mining, fuzzy sets, classification

Introduction

There are more than 8.10^9 Google indexed web pages in the World Wide Web. Finding relevant information is very difficult. Searching information is a main problem. When we find many results, it is a good idea to classify them.

Using web search engines we can choose: result file type, language, domain, etc. Often we receive a message "This web site is added to directory X in category (ies)..." in the result list. This directory contains qualitative, but very small subset of all web sites in the world, and for most results, we do not have any information about their types. This makes a big part of our result uncategorized. We can group them by region or language, for example, but not regarding their content. It will be good if we can, using a web crawler or metasearch engine, to specify a web site type from a given set at least. Other useful opportunity will be to classify uncategorized part of the result list of our search query. It is not the same using Google to find word "accommodation" in science conferences' web sites or travel agencies' web sites.

Automatically web sites categorization provides two main advantages for the end user. He or she can search information in specific group of web sites at first. The result will be always classified at second. The user will receive grouped results and it makes easier finding relevant information. A search engine supporting automatically web sites classification will allow more flexible queries or well arranged results.

The main part of realised web sites classification is based on the internal document representation and structure. The authors never forget the web page is a text document, and the web site is a text documents structure. The main efforts are addressed to find relevant text and structural features to the web site.

But in the other hand users can classify the web sites without knowledge in web development. The web directories editor can categorize a new site without reading meta-tags information or finding other web sites linked to the regarded one. Here we use external web sites views for their classification.

Web Sites Classification

Unknown objects classification is a main part of machine learning and data mining research. When we classify a set of objects we need

- formal object and classes descriptions
- classification model
- training set and training mechanism
- rules adding unknown objects into a class

There are created many automatical classification approaches, based on artificial neural networks, decision trees, genetic algorithms, etc. The classification process follows the steps:

- Model choice.
- Training. We use a relatively small and labelled subset, called training set. The labels mean belonging to a class. Based on this training set, we construct the classifier.
- Unknown objects classification.

Web resources classification is an application of traditional data mining techniques in respect to the specific area. ([8], [9]) The datasets contain web sites. All web sites classification could be possible, if we have a good ontology describing the current state of the art. However, it is a very difficult activity. We have to know at least the current situation in the entire web. A rational idea is to have a specific sub-ontology and use it to decide on the particular problems.

Talking about web sites classification, we have to keep in mind two main arguments.

- The hyperlinks between web sites do not reflect on their types. The authors are not obligated to relate their web sites to any other ones.
- The most adequate web sites description approach is using quality data. We can detect features or count them only.

There are many realised approaches to determine web site type or automatical construction of web directories. In general, we can find two very popular directions – adapted for web documents text mining techniques ([2]) and web structure mining techniques ([5], [6], [7]). In first case, authors prefer to weight different parts of the web sites or the web pages, and in the second – to use web structure in general. There are examples for domain specific classification ([3], [4]).

Here we try to prove how the type of web sites affects their external features. We try to find how the content influences the external view.

Web Sites External Features

Firstly we have to define what *an external feature* is. Every web site can be considered from two points of view:

- Internal – this is the site structure, meta tags, technologies, formal languages used in site creation, etc
- External – this is the visible part of the site

For example, when the user clicks “Sign in”, it could be a button, or (GIF or JPEG) image or text hyperlink in the different cases. It can start a script, written in some formal language, providing one and the same semantics.

When talking about links here, we mean external views of the same web site’s links.

Web Sites Features and Fuzzy Sets

The fuzzy sets [1] are good mechanism for describing the features of the web sites classes. There are not any formal models for the web sites creation and the authors are not obligated to include anything. Moreover, main purpose in the web is to be distinctive. However, content and specific area has an effect on the language, structure, representation of the data, etc. We can expect similar content to be presented in similar ways. From this point of view we cannot say a given feature is specific for a web sites class or not, but we can define a relative belonging into a set of features describing the class. That makes the fuzzy web sites description very relevant. We can regard the web site like a fuzzy set of features.

In other hand a given web site can belong into different categories. Sometimes the site category is not exactly defined. In these cases it might be as well to use fuzzy belonging into a web sites category. It is possible to regard the web sites categories as fuzzy set of web sites.

Semi structured nature of the web make the fuzzy models very useful for its explanation.

How to Prove

To define a fuzzy set describing a class, we need to discover a relatively small training set of web sites and their descriptions. For every member of this set we have to find the features contained in them at first, and compare the given results at second. With a simple comparison and counting, we find relatively belonging into a set of features for this class (and this small set). This makes our results as accurate as our training set is representative.

We need to prove whether our fuzzy sets are relevant or not. Of course, the initial fuzzy set is not enough for the classes’ description. It is possible to find one or more elements for all classes, but we have to find the specific ones. In a formal model if we have the classes $C_1 \dots C_n$, and T_i i= 1 … n are the fuzzy sets given from the first step, we actually are interested in sets

$$T_i \setminus \bigcup_{j \neq i} T_j \, (*)$$

for every i= 1, …, n. Here we can use the equation

$$A \setminus B = A \bigcap \overline{B}$$

where A and B are arbitrary sets. This representation will help us to apply definitions for the section and the union of fuzzy sets. If for every i = 1, …, n all of the sets (*) are not empty and are not the universal set, we can say we have found lists of features describing given classes.

This model is temporary because of the temporary nature of the web. It is exact for the training set only, not for all web sites in the world, belonging into the classes. Moreover, it provides correlations among the given classes, but not among all classes, which could exist in the world around. To improve the model correctness and accuracy we have three ways:

- Using carefully selected and relatively big training sets
- Frequently testing the training set for changes and actualise the features database and sets (*)
- Using model for classification web sites with “expected” types

How to Classify

The next task is to find a rule for unknown web site evaluation. A natural approach is to consider every web site description like a fuzzy set too and find all of the distances between this description and the fuzzy sets, associated with the classes. An uncategorized web site belongs to a class, if and only if, the distance between the site and the class is the smallest. The distance can be defined in many different ways. Actually, this is clustering web with preliminary defined cluster centres. In our works, we compare the Hamming and the Euclidean metrics. The metrics choice can be automated. It is necessary to have program applying two or more metrics or similarity functions. In the second case, the system must prove how similarity is bigger. The system can simultaneously follow two criteria:

1. Better total correctness, and in case they are equal -
2. The web sites distribution after the test. The statistical dispersion is good measure there.

For metrics choice, the same training set can be used.

In other hand we can use similarity function for fuzzy classification. It makes our model fuzzy in general. It is not difficult to see that

$$n(x,y) = \frac{1}{1 - d(x,y)}$$

when d is a given metric is a similarity function. If the objects x and y are equal, $n(x,y)$=1.

Objects and Classes Descriptions

Web sites’ descriptions in this model are simple. For every one we define a vector $V_i(v_{i1}, v_{i2}, \dots, v_{in})$ where v_{ij}=1 if the feature j is found in the site, and v_{ij}=0 if the feature is not found in the site.

If we have m web sites belonging into a given class, we define the vector T with components:

$$t_j = \tfrac{1}{m} \sum_{k=1}^{m} v_{kj}$$

It is not difficult to see that

- vector *T* defines a fuzzy set
- if we have two or more classes and mark them with T_i their vectors, then $D_i = T_i \setminus \bigcup_{k \neq i} T_k = T_i \bigcap \overline{(\bigcup_{k \neq i} T_k)}$ is a fuzzy class descriptor for every *i*.

Experiment

We made experiments with 100 web sites from five following types

T1. University web sites
T2. Newspaper web sites
T3. International unions web sites
T4. Governmental web sites
T5. Parliament web sites

We used 20 web sites by class. Their first nontrivial pages have been considered. Here by *nontrivial page* we mean the first page after simple Enter page. We used Yahoo! Directory for finding representative for all world training sets, from different languages, countries and continents with respect of their relative distribution. When we described these web sites, we obtain 127 different features. We count the features found into the classes. We compare the classes by (*) and obtain classes descriptors. Here we give elements *x* with $\mu(x) \geq 0.5$ for every class. Here $\mu(x)$ is the fuzzy set characteristic function.

T1: University web sites (30 elements with nonzero value of $\mu(x)$)

Feature	Belonging
Link "Alumni"	0.70
Link "About university"	0.70
Link "Structure"	0.65
Link "Events"	0.65
Link "Library"	0.60
Link "Researches"	0.60
Link "Students"	0.60
One colour background	0.55

T2: Newspaper web sites (60 elements with nonzero value of $\mu(x)$)

Feature	Belonging
Link "News"	0.60
Link "Sport news"	0.60
Link "Archives"	0.50
Link "Advertising"	0.50

T3: International unions web sites (52 elements with nonzero value of $\mu(x)$)

Feature	Belonging
Link "About us"	0.55

T4: Governmental web sites (57 elements with nonzero value of $\mu(x)$)

Feature	Belonging
Link "Searching"	0.50

T5: Parliament web sites – 45 elements with nonzero value of $\mu(x)$ but $\mu(x) \leq 0.45$ for all. The first five are

Feature	Belonging
Language choice	0.45
Link “Contacts”	0.35
Links to institution’s documents	0.35
Link "News"	0.35
One colour background	0.35

In our tests, we compare the Hamming and the Euclidean metrics and test them with 100 random selected web sites – by twenty for class. The results are given in the following tables

Euclidean metrics						
	1	**2**	**3**	**4**	**5**	Correctness
T1	17		1		2	85 %
T2		18	1		1	90 %
T3			17		3	85 %
T4			4	16		80 %
T5			2	2	16	80 %
Total	17	18	25	18	22	84 %

Hamming metrics						
	1	**2**	**3**	**4**	**5**	Correctness
T1	18				2	90 %
T2		14	1		5	70 %
T3			12		8	60 %
T4			1	8	11	40 %
T5			1		19	95 %
Total	18	14	15	8	45	71 %

Here “Correctness” is the percent of web sites correctly added into their class’ sets. Based on the results we can say the Euclidean metrics is better and can recommend it.

Less correctness for some types we can explain with classes’ similarity. Distance matrix between classes is as follow

T1	0,00				
T2	2,48	0,00			
T3	1,99	1,71	0,00		
T4	2,04	1,82	1,04	0,00	
T5	1,94	1,80	0,94	1,03	0,00
	T1	**T2**	**T3**	**T4**	**T5**

when the Euclidean distances between classes descriptors are given. These descriptors are obtained in the training phase and depend on training set only. As we can see, the best results in metrics tests we obtain for most “isolated” classes.

In our example we used the smallest distance between web site and web site class. In fuzzy web sites classification we can talk about biggest similarity.

Conclusions

Based on the experiment results we can say this approach have acceptable correctness for further studies and applications. The best results are observed for less similar classes. The main weak points are similar classes' areas. The approach is applicable to most general web sites categories.

It is a good idea to test the approach in a similar web sites classification. There are many huge categories in the web directories. We can apply the approach for subcategories creation. We can expect similar classes, but the web sites are similar too.

Other result is fuzzy sets are suitable mechanism for web sites classes description and study.

The results manifest how important the web site structure is. The most of the described features are external representation of this structure. It is prove in practise the proposition web sites are independent objects for classification.

References

1. Zadeh L., Fuzzy sets, Information and control, Vol. 8, 1965 (338 – 353)
2. Pierre J., On the Automated Classification of Web Sites. Linköping Electronic Articles in Computer and Information Science, Vol. 6(2001): nr 0. http://www.ep.liu.se/ea/cis/2001/000/. February 4, 2001
3. Ardo A., T. Koch, and L. Nooden. The construction of a robot generated subject index. EU Project DESIRE II D3.6a, Working Paper 1 1999. http://www.lub.lu.se/desire/DESIRE36a-WP1.html
4. Kock T., A. Ardo. Automatic classification of full-text HTML documents from one specific subject area. EU Project DESIRE II D3.6a, Working Paper 2 2000. http://www.lub.lu.se/desire/DESIRE36a-WP2.html
5. Attardi G., A. Gulli, and F. Sebastiani. Automatic Web Page Categorization by Link and Context Analysis. In Chris Hutchison and Gaetano Lanzarone (eds.), Proceedings of THAI'99, European Symposium on Telematics, Hypermedia and Artificial Intelligence, 105-119, 1999.
6. Cho J., H. Garcia-Molina, and L. Page. Efficient crawling through URL ordering. In Computer Networks and ISDN Systems (WWW7), Vol. 30, 1998.
7. Rennie J., A. McCallum. Using Reinforcement Learning to Spider the Web Efficiently. Proceedings of the Sixteenth International Conference on Machine Learning, 1999.
8. Han, J. and Chang, K. C.-C. Data Mining for Web Intelligence, IEEE Computer, Nov. 2002
9. M. N. Garofalakis, R. Rastogi, S. Seshadri, K. Shim, Data Mining and the Web: Past, Present and Future, Proceedings of WIDM99, Kansas City, U.S.A., 1999.

Author Information

Georgi Furnadzhiev - Institute of Mathematics and Informatics, BAS, Information Research Department; Acad. Georgi Bonchev St., Block 8, Sofia 1113, Bulgaria; e-mail: furnadjieff@math.bas.bg

TABLE OF CONTENTS