Fuzzy Relational Knowledge Representation and Context in the Service of Semantic Information Retrieval

Manolis Wallace and Yannis Avrithis
Image, Video and Multimedia Systems Laboratory
Department of Computer Science
School of Electrical and Computer Engineering
National Technical University of Athens (NTUA)
E-mail: {wallace,iavr}@image.ntua.gr

***Abstract*— In this paper we follow a fuzzy relational approach to knowledge representation. With the use of semantic fuzzy relations we define and extract the semantic context out of a set of semantic entities. Based on this, we then proceed to the case of information retrieval and explain how the three participating contexts, namely the context of the query, the context of the document and the context of the user, can be estimated and utilized towards the achievement of more intuitive information services.**

I. Introduction

Information retrieval algorithms and systems are generally partitioned in two distinct classes: text and multimedia retrieval. The former are focused on the handling of textual terms of the available documents and of the query [1], while the latter typically attempt to match visual and audio features of a sample query document to those of the documents of the database [2]. Unfortunately, neither approach attempts to treat the documents and the user query at a semantic level.

An important step in the direction of semantic, knowledge – based information retrieval has already been made, with the definition of the *semantic entity* [9]; this corresponds to what we might call a concept, object or event, and aims to replace terms and keywords Once the semantic entities in a textual or multimedia document are detected, a uniform approach to their semantic handling can be followed. Of great importance is, as well, the construction of ontologies, which constitute an attempt to describe the relations between real life entities, in a conceptual level [3].

In this work we extend on the ideas presented in the field of ontologies as to include fuzzy degrees of membership in the utilized semantic relations. Such relations, and especially the partially ordering ones, can be used to define and extract the context of a set of semantic entities. This is then utilized in order to estimate the context of the document, the context of the user and the context of the query, which are all very important in the process of intelligent information retrieval.

The structure of the paper is as follows: in section II we present the semantic fuzzy relations utilized in this work, and in section III we use them in order to define and extract the context of a set of semantic entities. Based on this, in sections IV, V and VI we extract and use the context of the query, the document and the user, respectively. In section VII we present some simple results from the application of the presented methodologies and in section VIII we list our concluding remarks.

II. The fuzzy semantic relations

Ontologies are an attempt for modelling real world entities. They define the entities, from simple objects to abstract object classes and concepts, using their textual, visual and other descriptors, as well as the relations between them. Although any type of relation may be included in the definition of a new ontology, all utilized relations are practically either ordering (taxonomic) or symmetric (compatibility) relations. Moreover, all relations are crisp.

Compatibility relations have traditionally been exploited by information retrieval systems for tasks such as query expansion. They are ideal for the description of similarities of various natures, but fail to assist in the determination of the context of a set of semantic entities; the use of ordering relations is necessary for such tasks [4]. Thus, a main challenge of intelligent information retrieval is the meaningful exploitation of information contained in taxonomic relations of an ontology.

It is well understood that relations among real life entities are always a matter of degree, and are, therefore, best modelled using fuzzy relations. Ontological taxonomies, on the other hand, are crisp in principle. Thus, they fail to fully describe real life concepts, and are limited to α-cuts of the desired relations. This is a very important drawback, that makes such relations insufficient for the services that an intelligent information retrieval system aims to offer.

In [5], the utilization of fuzzy relations for the description of the relation among real life entities is introduced. In this section, we present a few commonly encountered semantic relations that can be modelled as fuzzy ordering relations, and propose their combination for the generation of meaningful, fuzzy, quasi-taxonomic relations. Based on such relations, in

TABLE I

THE FUZZY SEMANTIC RELATIONS

Symbol	Name
Sp	Specialization
Ct	Context
Ins	Instrument
P	Part
Pat	Patient
Loc	Location
Ag	Agent

the following sections we will explain how the context of the user, the document and the query may be estimated and utilized.

The specialization relation Sp is a fuzzy partial ordering on the set of semantic entities. $Sp(a,b) > 0$ means that the meaning of a includes the meaning of b; the most common form of specialization is sub – classing, i.e. a is a generalization of b. The role of the specialization relation in knowledge – based retrieval is as follows: if a document refers to the meaning of entity b, then it is also related to a, since b is a special case of a. Still, there is no evidence that the opposite also holds; it is obvious that the specialization relation contains important information that can not be modelled in a symmetric relation. The context relation Ct is also a fuzzy partial ordering on the set of semantic entities. $Ct(a,b) > 0$ means that b provides the context for a or, in other words, that b is the thematic category that a belongs to. Other relations considered in the following have similar interpretations. Their names and corresponding notations are given in table I.

Fuzziness of the aforementioned relations has the following meaning: High values of $Sp(a,b)$, imply that the meaning of b approaches the meaning of a, in the sense that when a document is related to b, then it is most probably related to a as well. On the other hand, as $Sp(a,b)$ decreases, the meaning of b becomes "narrower" than the meaning of a, in the sense that a document's relation to b will not imply a relation to a as well with a high probability, or to a high degree. Summarizing, the value of $Sp(a,b)$ indicates the degree to which the stored knowledge shows that an occurrence of b in a document implies relation to a. Likewise, the degrees of the other relations can also be interpreted as conditional probabilities or degrees of implied relevance.

A last point to consider is the transitivity of the relations presented above. It is obvious that if b is a specialization of a and c is a specialization of b, then c is a specialization of a. This implies that the specialization relation is transitive. A similar argument can be made for the other relations, as well. Still, the form of transitivity used cannot be $\sup - \min$ transitivity, but one relying on a subidempotent norm. Therefore, we demand that the presented relations are $\sup - t$ transitive, where t is an Archimedean norm.

Given the large number of semantic relations we have defined, it is obvious that the available knowledge is partitioned among them, and thus each one of them alone does not suffice for the offering of intelligent information services. Based on the relations r_i of table I we may construct the following semantic relation, thus accumulating all stored knowledge in one relation:

$$T = Tr^t(\bigcup_i r_i^{p_i}), p_i \in \{-1, 1\}, i \in 1 \ldots n \tag{1}$$

where $Tr^t(A)$ is the $\sup - t$ transitive closure of relation A; the transitivity of relation T was not implied by the definition, as the union of transitive relations is not necessarily transitive. The transitive closure is achieved using the methodology presented in [8].

It is important to point out that there is no such thing a a correct choice of values for parameters p_i. Different choices lead to the generation of relations that are optimal for different tasks. For example, relation

$$T_{TC} = Tr^t(Sp \cup C^{-1} \cup Ins \cup P \cup Pat \cup Loc \cup Ag) \tag{2}$$

is ideal for the definition of semantic user profiles as well as for the thematic categorization of documents, while relation

$$T_Q = Tr^t(Sp \cup P^{-1}) \tag{3}$$

is ideal for context sensitive query expansion.

III. THE CONTEXT OF A SET OF SEMANTIC ENTITIES

In general, term context refers to whatever is common among a set of elements. In this work, where the elements are semantic entities, term context may refer to the common meaning of a set of entities. The fact that relations T described in the previous section are (almost) ordering relations allows us to use them in order to define, extract and use the context of a set of semantic entities. Relying on the semantics of the T relations, we define the context $K(s)$ of a semantic entity $s \in S$ as the set of its descendants in some relation T:

$$K(s) = T_{\leq}(s) \tag{4}$$

Relation T is assumed to be reflective, so that $s \in K(s)$. Assuming that a set of entities $A \subset S$ is crisp, i.e. all considered entities belong to the set with degree one, the context of the group, which is again a set of semantic entities, can be defined simply as the set of their common descendants.

$$K(A) = \bigcap_i K(s_i), s_i \in A \tag{5}$$

Obviously, as more entities are considered, the context becomes narrower, i.e. it contains less entities and to smaller degrees:

$$A \supset B \implies K(A) \subseteq K(B) \tag{6}$$

When the definition of context is extended to the case of fuzzy sets of semantic entities, this inequality must still hold. The satisfaction of the following is also an obvious constraint:

- $A(s) = 0 \implies K(A) = K(A - \{s\})$, i.e. no narrowing of context.

- $A(s) = 1 \implies K(A) \subseteq K(s)$, i.e. full narrowing of context.
- $K(A)$ decreases monotonically with respect to $A(s)$.

Taking these into consideration, we demand that, when A is fuzzy, the "considered" context $\mathcal{K}(s)$ of s, i.e. the entity's context when taking its degree of participation to the set into account, becomes low when the degrees of taxonomy are low and the degree of participation $A(s)$ is high. Therefore:

$$cp(\mathcal{K}(s)) \doteq cp(K(s)) \cap (A(s) \cdot S) \tag{7}$$

where cp is an involutive fuzzy complement, and $\cap$ and $\cup$ correspond to a t-norm and a t-conorm which are dual, with respect to cp. By applying de Morgan's law, we obtain:

$$\mathcal{K}(s) \doteq K(s) \cup cp(A(s)) \tag{8}$$

Then the set's context is easily calculated as follows:

$$K(A) = \bigcap_i \mathcal{K}(s_i), s_i \in A \tag{9}$$

Considering the semantics of the utilized T relation and the process of context determination, it is easy to realize that when the entities in a set are highly related to a common meaning, the context will have high degrees of membership for the entities that represent this common meaning. Therefore, the height of the context $h(K(A))$ may be used as a measure of the semantic correlation of entities in set A. We will refer to this measure as *intensity* of the context.

IV. The Context of The Query

Ideally, a user query consists of keywords, each one of which corresponds to a single semantic entity. In that case, the interpretation of the query int o semantic entities is simple and straightforward. In some cases though, this is not true, as some words can be matched to more than one semantic entity. It is left to the information system to utilize knowledge in order to correctly decide which semantic entity was indeed implied by the user. In this task, the context of the query can provide the required information.

However, we have defined the context of a set of semantic entities, not he context of a set of keywords. Thus, the detection of the query context cannot be performed before the query interpretation is completed. Therefore both tasks, query interpretation and context detection, must be performed simultaneously.

Let the textual query contain the textual terms t_i, $i = 1, 2, \ldots, N_Q$. Let also t_i be the textual description of semantic entities s_{ij}, $j = 1, 2, \ldots, N_{Si}$. Then there exist $\prod_{i=1}^{N_Q} N_{Si}$ distinct combinations of semantic entities that may be used for the representation of the user query.

The most intuitive approach to solving the query interpretation problem is by assuming that out of all the possible interpretations of each textual term, the one truly implied by the user is the one that is most related to the other terms of the query. Thus, out of all the candidate queries the one that has the most intense context is selected:

$$q = q_i \in Q : h(K(q_i)) \geq h(K(q_j)) \forall q_j \in Q \tag{10}$$

$$Q = \{q_k\}, k = 1, 2, \ldots, \prod_{i=1}^{N_Q} N_{Si} \tag{11}$$

Once the query interpretation has been completed, query expansion enriches the query in order to increase the probability of a match between the query and the document index. The presence of several semantic entities in the query during the query interpretation defines a context, which may be used to direct the expansion process.

More formally, we replace each semantic entity $s_i \in q$ with a fuzzy set of semantic entities $X(s_i)$; we will refer to this set as the expanded semantic entity. In a context – sensitive query expansion, the degree of significance, x_{ij}, of the entity s_j in the expanded semantic entity $X(s_j)$ is dependent on the relevance of s_j to the query, on the weight $w_i = q(s_i)$, and on the degree of the relation $T(s_i, s_j)$. We define the measure of relevance of semantic entity s_j to the query as:

$$h_j = max(\frac{h(T(s_j) \cap K(q))}{h(K(q))}, c(h(K(q)))) \tag{12}$$

The fuzzy complement c in this relation is Yager's complement with a parameter of 0.5. Considering now the initial entity's importance in the query and the degree to which the initial and the candidate entity are related, we have

$$x_{ij} = h_j \cdot q(s_i) \cdot T(s_i, s_j) \tag{13}$$

V. The Context of The Document

In order to be able to treat all documents in a uniform manner, it is important to be able to map them to some common space. The mapping to abstract but semantic thematic categories is an intuitive way of solving this task.

In this process, a number of issues, such as the following, have to be considered:

- A semantic entity may be related to multiple, unrelated thematic categories.
- A document may be related to multiple, unrelated thematic categories.
- The indexing of a document may have been created in an automated manner. Thus, existence of random, and therefore misleading semantic entities cannot be excluded.

Before actually extracting thematic category information from the set of semantic entities that are related to a document d via the semantic index, in order to support the possibility of existence of multiple distinct topics in a single document, the entities that are related to it needs to be clustered to groups, according to the topics they are related to.

Not knowing beforehand the count of the distinct topics to which the document is related, we apply an agglomerative clustering algorithm [7]. The two key points in hierarchical clustering are the identification of the clusters to merge at each step, i.e. the definition of a meaningful measurefor the distance between clusters, and the identification of the optimal terminating step, i.e. the definition of a meaningful termination criterion.

When clustering semantic entities, the ideal similarity measure is one that quantifies their semantic correlation. We have already defined such a measure in section III; it is the height of their common context. Therefore, the merging of clusters will be based on this measure.

$$d(c_1, c_2) = h(K(c_1 \cup c_2)) \tag{14}$$

The process of merging should terminate when the entities are clustered into sets that correspond to distinct topics. We may identify such sets by the fact that their common contexts will have low, if not zero, intensity. Therefore, the termination criterion shall be a threshold on the intensity of the common meaning, i.e. a threshold on the selected compatibility measure.

At the end of this process, each cluster c is described by the crisp set of semantic entities that belong to it. Using those, we may create a fuzzy classifier, i.e. a function C_c that will measure the degree of correlation of a semantic entity s with the cluster c.

$$C_c : S \to [0, 1] \tag{15}$$

Obviously, a semantic entity should be considered correlated with c, if it is related to the common meaning of the semantic entities in c. Therefore, the quantity

$$C_c(s) = \frac{h(K(c \cup \{s\}))}{h(K(c))} \tag{16}$$

is used. Using such classifiers, we may expand the detected crisp partitions, as to include more semantic entities, as follows: partition c is replaced by cluster

$$c' = \sum_{s \in I(d)} s/C_c(s) \tag{17}$$

Obviously $c' \supseteq c$.

Thematic categories are semantic entities that have been selected as having a special meaning for the system; more formally:

$$TC \subset S \tag{18}$$

This simplifies the process of automatic thematic categorization: The thematic categories that are contained in the context of a cluster of semantic entities are obviously thematic categories that are related to the whole document. Clusters that do not have a high cardinality probably only contain misleading entities, and therefore need to be ignored in the estimation of the thematic categorization of the document. The notion of “high cardinality” is modelled with the use of a “big” fuzzy number L. $L(a)$ is the truth value of the preposition “the value of a is high”.

$$R_{\mathrm{TC}}(c) = w(K(c) \cap TC) \cdot L(|c|) \tag{19}$$

where w is a weak *modifier* [6]. It is easy to see that a thematic category t is detected if a cluster c, whose context contains t, is detected in the document, and additionally the cardinality of c is high (i.e. the cluster is most probably not comprised of misleading entities) and the degree of membership of t in the context of c is high.

VI. The Context of The User

The context of the user in information retrieval is defined by the user’s preferences. In the extraction of the preferences from the accumulated history of user feedback, issues similar to those related to the thematic categorization of documents need to be considered. Specifically, one needs to consider that

- A user may be interested in multiple topics.
- Not all topics that are related to a document in the usage history are necessarily of interest to the user.

These issues are tackled using similar tools and principles, as the ones used to tackle the corresponding problems in content analysis. Thus, once more, the basis on which the extraction of preferences is built is the context. The common topics of documents are used to cluster documents and cluster cardinalities are considered in order to determine which documents are indicative of a preference of the user and which exist in the usage history coincidentally.

What is common among two documents d_1, d_2 i.e. their common topics, can be referred to as their common context. This can be defined as

$$K(d_1, d_2) = R_{TC}(d_1) \cap R_{TC}(d_2) \tag{20}$$

A metric that can indicate the degree to which two documents are related is, of course, the height of their common context. This can be extended to the case of more than two documents, in order to provide a metric that measures the similarity between clusters of documents:

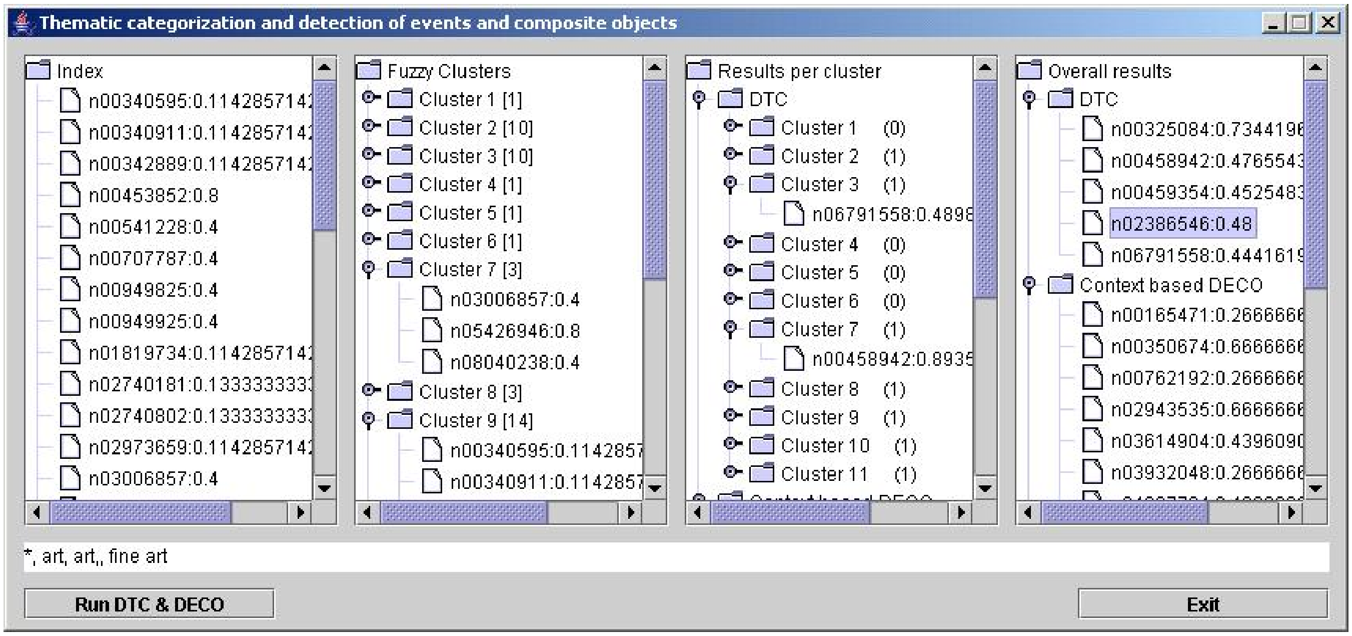
$$d(c_1, c_2) = h(K(c_1 \cup c_2)) \tag{21}$$

$$K(c) = h(\bigcap_{d \in c} R_{TC}(d)) \tag{22}$$

If H^+ is the set of documents for which the user has indicated preference then we proceed as follows: using the distance metric presented above we apply an agglomerative clustering algorithm on documents of set H^+, in order to detect the distinct topics that interest the user. Out of each detected cluster we extract the corresponding interests as follows:

$$U^+(c) = K(c) \cdot L(|c|) \tag{23}$$

During searching, all retrieved documents are compared to the interests in the user’s profile, and are re-ranked according to the degree of relevance that they have to the known preferences. This re-ranking is intense when the context of the query is not intense, and vice versa. Thus, when the query contains sufficient information in order to describe the exact topic of the search the results remain unaltered, while when the query context is vague, information from the user profile is utilized in order to remove some of the uncertainty and enhance the system’s response.



Thematic categorization and detection of events and composite objects
Run DTC & DECO
Exit

```
Index
  n00340595:0.114285714
  n00340911:0.114285714
  n00342889:0.114285714
  n00453852:0.8
  n00541228:0.4
  n00707787:0.4
  n00949825:0.4
  n00949925:0.4
  n01819734:0.114285714
  n02740181:0.133333333
  n02740802:0.133333333
  n02973659:0.114285714
  n03006857:0.4

Fuzzy Clusters
  Cluster 1 [1]
  Cluster 2 [10]
  Cluster 3 [10]
  Cluster 4 [1]
  Cluster 5 [1]
  Cluster 6 [1]
  Cluster 7 [3]
    n03006857:0.4
    n05426946:0.8
    n08040238:0.4
  Cluster 8 [3]
  Cluster 9 [14]
    n00340595:0.1142857
    n00340911:0.1142857

Results per cluster
  DTC
    Cluster 1    (0)
    Cluster 2    (1)
    Cluster 3    (1)
      n06791558:0.4898
    Cluster 4    (0)
    Cluster 5    (0)
    Cluster 6    (0)
    Cluster 7    (1)
      n00458942:0.8935
    Cluster 8    (1)
    Cluster 9    (1)
    Cluster 10    (1)
    Cluster 11    (1)

Overall results
  DTC
    n00325084:0.7344196
    n00458942:0.4765543
    n00459354:0.4525483
    n02386546:0.48
    n06791558:0.4441619
  Context based DECO
    n00165471:0.2666666
    n00350674:0.6666666
    n00762192:0.2666666
    n02943535:0.6666666
    n03614904:0.4396090
    n03932048:0.2666666
```

*, art, art,, fine art

Fig. 1. Application of thematic categorization

VII. Results

In figure 1 we present an implementation of the thematic categorization methodology described in section V. In the first column, the IDs of the objects detected in a multimedia document are presented. In our ontology each one of these IDs is related to a textual description, a set of keywords, and in some cases a set of audiovisual descriptors. For example, ID n02386546 is related to keywords ”art” and ”fine art”, as can be seen at the lower part of the application. In the second column the entities have been clustered and in column 3 thematic categorization information is extracted from each cluster, without yet considering cardinality.

Finally, in column 4 results are summarized for all clusters, while also considering cluster cardinality. Not all thematic categories detected in distinct clusters participate in the overall result; findings that correspond to clusters of small cardinality have been ignored, as they are possibly misleading.

Figure 2 demonstrates the results of a user query corresponding to the keyword “politics”. Thematic categorization has been performed beforehand for all documents in the database. The search process starts with query interpretation, which in this example is trivial, and continues with query expansion. Matching between the indexing of the documents and the expanded query entities is performed and the matching documents are re-ranked based on the degree of relevance to the user preferences.

VIII. Conclusions

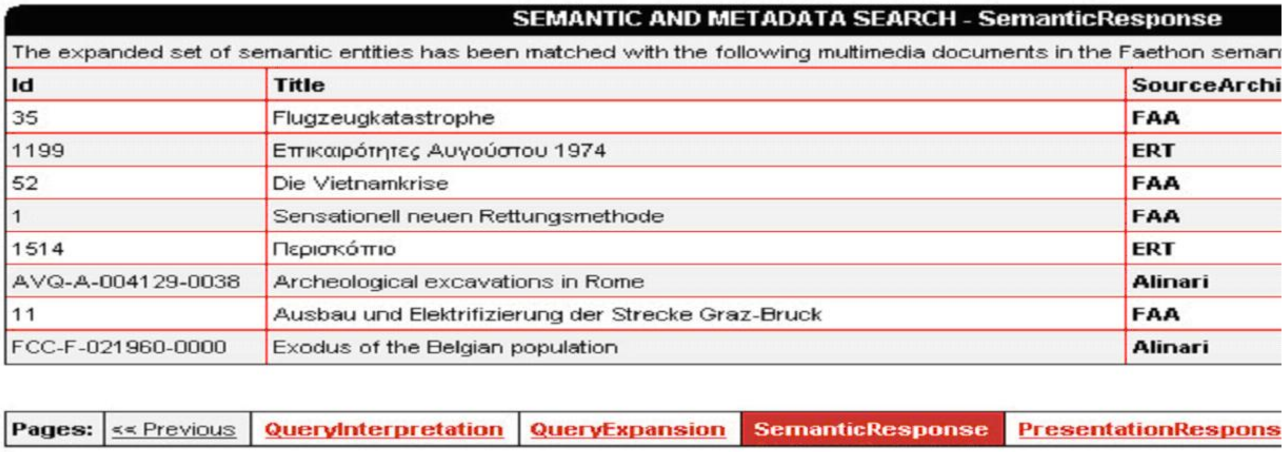
In this paper we have extended on the crisp relations defined in ontologies and followed a fuzzy relational approach to knowledge representation. Using this knowledge, we have defined and extracted the semantic context of a set of semantic entities, i.e. their common meaning. This allows us to follow a unified approach to intelligent information retrieval, both for textual and multimedia documents.

Based on the definition we provided for context, we explained how the three sources of information participating in the process of retrieval may be treated in a semantic manner. Specifically, we have utilized the context of the query in order to assist in the processes of query interpretation and query expansion, we have utilized the context of the document in order to drive the process of thematic categorization and we have extracted the user context from the usage history in the form of a sequence of fuzzy sets of thematic categories.

All of the above can easily be integrated in a single information system, thus providing enhanced searching and browsing services. Some results from such a system are also provided.

Acknowledgements

This work has been partially funded by the EU IST-1999-20502 FAETHON project.



SEMANTIC AND METADATA SEARCH - SemanticResponse

The expanded set of semantic entities has been matched with the following multimedia documents in the Faethon seman

Id	Title	SourceArchi
35	Flugzeugkatastrophe	FAA
1199	Επικαιρότητες Αυγούστου 1974	ERT
52	Die Vietnamkrise	FAA
1	Sensationell neuen Rettungsmethode	FAA
1514	Περισκόπιο	ERT
AVQ-A-004129-0038	Archeological excavations in Rome	Alinari
11	Ausbau und Elektrifizierung der Strecke Graz-Bruck	FAA
FCC-F-021960-0000	Exodus of the Belgian population	Alinari

Fig. 2. Ranked multimedia documents retrieved for a user query with the keyword “politics”

REFERENCES

[1] Baeza-Yates, R.A., Ribeiro-Neto, B.A., Modern Information Retrieval, ACM Press / Addison-Wesley, 1999.

[2] Ciocca G. Schettini R., A relevance feedback mechanism for content-based image retrieval, Information Processing and Management 35(5):605–632, 1999.

[3] A. Maedche, B. Motik, N. Silva and R. Volz, MAFRA - An Ontology MApping FRAmework in the Context of the SemanticWeb. Proceedings of the Workshop on Ontology Transformation at ECAI2002, Lyon, France, July 2002.

[4] G. Akrivas, M. Wallace, G. Andreou, G. Stamou and S. Kollias, ”Context - Sensitive Semantic Query Expansion”, Proceedings of the IEEE International Conference on Artificial Intelligence Systems (ICAIS), Divnomorskoe, Russia, September 2002

[5] G. Akrivas G. and G. Stamou, Fuzzy Semantic Association of Audiovisual Document Descriptions, Proc. of Int. Workshop on Very Low Bitrate Video Coding (VLBV), Athens, Greece, Oct. 2001

[6] G. Klir and B. Yuan, Fuzzy Sets and Fuzzy Logic, Theory and Applications, New Jersey, Prentice Hall, 1995.

[7] S. Theodoridis and K. Koutroumbas, Pattern Recognition, Academic Press, 1998.

[8] Wallace M., Kollias S., “Computationally efficient incremental transitive closure of sparse fuzzy binary relations”, Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Budabest, Hungary, July 2004.

[9] ISO/IEC JTC 1/SC 29 M4242, Text of 15938-5 FDIS Information Technology – Multimedia Content Description Interface – Part 5 Multimedia Description Schemes, October 2001.