Towards Semantic Multimedia Indexing by Classification & Reasoning on Textual Metadata

Ph. Mylonas, N. Simou, V. Tzouvaras and Y. Avrithis

Department of Electrical and Computer Engineering
National Technical University of Athens
Iroon Polytechniou 9, 157 80 Zographou, Greece
{fmylonas,nsimou,tzouvaras,iavr}@image.ntua.gr

Abstract. The task of multimedia document categorization forms a well-known problem in information retrieval. The task is to assign a multimedia document to one or more categories, based on its contents. In this case, effective management and thematic categorization requires usually the extraction of the underlying semantics. The proposed approach utilizes as input, analyzes and exploits the textual annotation that accompanies a multimedia document, in order to extract its underlying semantics, construct a semantic index and finally classify the documents to thematic categories. This process is based on a unified knowledge and semantics representation model introduced, as well as basic principles of fuzzy relational algebra. On top of that the fuzzy extension of expressive description logic language $\mathcal{SHIN}$, f-$\mathcal{SHIN}$ and its reasoning services are used to further refine and optimize the initial categorization results. The proposed approach was tested on a set of real-life multimedia documents, derived from the Internet[1], as well as personal databases and shows rather promising results.

1 Introduction

During the last decades, many researchers have tried to tackle the problem of identifying the multimedia content that the user desires. In this context, multimedia document categorization has been identified as an increasingly important and cost-effective step towards more efficient manipulation and understanding of multimedia content. Of course, semantic information extraction via knowledge acquisition and identification of high-level semantic features still remains an open research problem. Typical problems that are commonly encountered in textual retrieval systems, such as information overload and mismatch [19], are present in the case of multimedia retrieval, as well [5]. Current state-of-the-art approaches tend to combine multimedia extraction and knowledge evolution in a common framework. This knowledge deployment enhances the robustness of the extraction process, while at the same time the continuous extraction of semantic information is utilized to enrich this knowledge. In order to further reduce the complexity of semantic multimedia analysis, unified modelling and ontology representation of multimedia knowledge is also introduced.

[1] http://www.cnn.com
http://en.wikipedia.org

When dealing solely with the task of thematic categories identification, fast algorithms have been introduced and utilized for browsing of large amounts of multimedia content [7]. Latent Semantic Analysis (LSA) [14], [8], [4] uses Singular Value Decomposition (SVD) to map documents and terms from their standard vector space representation to a lower dimensional latent space, thus revealing semantic relations between the entities of interest. An unsupervised generalization of LSA, probabilistic-LSA (pLSA) [10], which builds upon a statistical foundation, represents documents in a semantic concept space and extracts concepts automatically. Furthermore, pattern recognition and machine learning techniques have also been applied to document classification, such as the fuzzy c-means algorithm [3] in the case of supervised multimedia documents classification. Finally, projections techniques [20] and k-means clustering [17] are proposed to speed up the distance calculations of clustering and their effectiveness are examined in [6]; however, document clustering results are very dependent on the original multimedia content data set.

In the field of multimedia reasoning, a popular family of knowledge representation languages, which gained a great attention during the last decade, is Description Logics (DLs) [2]. DLs are logical reconstruction of the, so called, frame-based knowledge representation languages, with the aim of providing a simple well-established declarative semantics to capture the meaning of the most popular features of structured representation of knowledge. Their well defined semantics, together with their determinable reasoning methods, make them applicable to a variety of applications [2]. On the other hand despite the rich expressiveness of classical DLs, they are insufficient to deal with vague and uncertain information which is commonly found in many real-world applications such as multimedia content. For that purpose a variety of DLs that can handle imprecise information in many flavors like probabilistic [9, 12] and fuzzy [21, 23] have been proposed.

In this paper, we present a hybrid approach composed of two identifiable parts, namely, a text categorization module and a fuzzy reasoning engine. The categorization module is based on the notion of context and common clustering techniques and utilizes fuzzy algebra principles towards thematic categorization of multimedia documents. The fuzzy reasoning engine has been constructed on the basis of DL f-$\mathcal{SHIN}$ [22] which is the fuzzy extension of expressive $\mathcal{SHIN}$ [11]. In the context of this work, the notion of reasoning refers to the automatic derivation and optimization of the text categorization module results, in terms of refined, high-level semantic thematic categorization results through the utilization of the provided (general, domain, structural, etc.) knowledge.

The rest of the paper is organized as follows. The next Section 2 summarizes briefly the utilized mathematical notation used throughout the paper. In Section 3, the textual categorization approach is presented, based on fuzzy algebra principles. In Section 4, the fuzzy knowledge base constructed for multimedia documents categorization is introduced. Finally, in Section 5, some experimental results are demonstrated and Section 6 states our concluding remarks.

2 Knowledge Representation and Notation

The knowledge model utilized within this work is twofold and is based on a set of concepts and relations between them and on a family of logic-based knowledge representation formalisms designed to represent and reason about the knowledge of an application domain in a structured and well-understood way. This knowledge representation allows the establishment of a detailed content description of all kinds of different multimedia documents in a unified way. In the case of still or moving images, the knowledge is specifically constructed to match the specific textual annotations of the corresponding multimedia documents.

2.1 Knowledge formalization for text categorization

Part of the proposed knowledge model in this work is comprised of a set of concepts and relations between them. Nowadays, most relations among real-life concepts are a matter of degree and it is our belief that they are therefore best modelled using fuzzy relations. Herein, we follow such an approach, based on a formal methodology and mathematical notation founded on fuzzy relational algebra [13] and whose fundamental principles are summarized in the following of the current section.

In general, a crisp set S on a given universe U is described by a membership function $\mu_S : U \rightarrow \{0,\ 1\}$. Similarly, a fuzzy set F on S is described by a membership function $\mu_F : S \rightarrow [0,\ 1]$. We may use the well-known sum notation [15] to describe any fuzzy set F, as follows: $F = \sum\limits_i s_i/w_i = \{s_1/w_1, s_2/w_2, \ldots, s_n/w_n\}$, where $i \in N_n$, $n = |S|$ is the cardinality of S, $w_i = \mu_F(s_i)$ and $s_i \in S$.

A fuzzy binary relation on S is defined as a function $\mathcal{R} : S^2 \rightarrow [0,\ 1]$, whereas its *inverse* relation is defined as $\mathcal{R}^{-1}(x, y) = \mathcal{R}(y, x)$. The intersection, union and sup-t composition of two fuzzy relations $\mathcal{R}_1$ and $\mathcal{R}_2$ defined on the same set S are defined as:

$$(\mathcal{R}_1 \cap \mathcal{R}_2)(x, y) = t(\mathcal{R}_1(x, y), \mathcal{R}_2(x, y)) \tag{1}$$

$$(\mathcal{R}_1 \cup \mathcal{R}_2)(x, y) = u(\mathcal{R}_1(x, y), \mathcal{R}_2(x, y)) \tag{2}$$

$$(\mathcal{R}_1 \circ \mathcal{R}_2)(x, y) = \sup_{z \in S} t(\mathcal{R}_1(x, z), \mathcal{R}_2(z, y)) \tag{3}$$

respectively, where t and u are a t-norm and a t-conorm, respectively. The standard t-norm and t-conorm are the min and max functions, respectively.

In the past, approaches based solely on lexical terms typically suffer from the problematic mapping of terms to concepts, as described in [18]. More specifically, as more than one lexical term may be associated to the same concept, and more than one concept may be associated to the same lexical term, the processing of query and index information is not trivial. In order to overcome such problems, we shall work directly with concepts, rather than lexical terms. Thus, in the sequel, we shall denote by $S = \{s_1, s_2, ..., s_n\}$, the crisp set of known concepts. A knowledge representation model may consist of the definitions of these concepts, together with their lexical descriptions, i.e. their corresponding lexical terms, as well as a set of relations amongst the concepts. The objective is to construct a model in which the context determines the intended meaning of each lexical term, and a lexical term used in different context may have different meanings. An initial formal definition of such a model is given by:

$$O = \{S, \{\mathcal{R}_i\}\}, \quad i = 1 \dots n \tag{4}$$

$$\mathcal{R}_i : S \times S \rightarrow \{0,\ 1\}, \quad i = 1 \dots n \tag{5}$$

where O is the knowledge model and $\mathcal{R}_i$ the i-th binary relation amongst the concepts.

In principle, almost any type of binary relation may be included to construct this kind of knowledge representation. However, a decent number of distinct and diverse relations among its concepts must be included for the knowledge model to be highly descriptive. To handle uncertainty issues when modelling real-life information, Akrivas et al. [1] propose the use of fuzzy relations. Thus, in this case, the above presented knowledge model description is modified, as follows:

$$O_F = \{S, \{r_i\}\}, \quad i = 1 \dots n \tag{6}$$

$$r_i = F(\mathcal{R}_i) : S \times S \rightarrow [0,\ 1], \quad i = 1 \dots n \tag{7}$$

where F denotes the fuzzification of the $\mathcal{R}_i$ relations. Based on the r_i relations, we may construct the following combined relation:

$$T = \bigcup_i r_i^{p_i}, \quad p_i \in \{-1,\ 1\}, \quad i = 1 \dots n \tag{8}$$

where the value of p_i is determined by the semantics of each relation used in the construction of T (e.g. order of arguments x, y in Table 1), since some relations may need to be inversed before being used in the construction of T. For the scope of the current paper, focusing on the analysis of multimedia document descriptions, relation T has been generated as described in equation (9), following the approach introduced in [24].

$$T = Tr^t(Sp \cup P^{-1} \cup Ins \cup Pr^{-1} \cup Pat \cup L \cup Ex) \tag{9}$$

As observed in the above equation, a set of fuzzy relations was utilized for the construction of T, whose semantics are defined in MPEG-7 and summarized in Table 1. Based on the semantics of the participating relations, it is easy to see that T is ideal for the determination of the topics that a concept may be related to, as well as for the estimation of the thematic categories identified within a set of concepts in a document.

2.2 The Description Logic f-$\mathcal{SHIN}$

The second part of the proposed knowledge model is based on description logics. Description logics (DLs) are a family of logic-based knowledge representation languages that can be used to represent the knowledge of an application domain in a structured

Name	Inverse	Symbol	Meaning
Part	PartOf	P(x,y)	y is part of x
Specialization	Generalization	Sp(x,y)	x is a generalization of y
Example	ExampleOf	Ex(x,y)	y is an example of x
Instrument	InstrumentOf	Ins(x,y)	y is an instrument of x
Location	LocationOf	L(x,y)	y is the location of x
Patient	PatientOf	Pat(x,y)	y is a patient of x
Property	PropertyOf	Pr(x,y)	y is a property of x

Table 1. Fuzzy taxonomic relations used for generation of the proposed knowledge model.

and formally well-understood way. They inherit concept descriptions from their ancestor *description languages* i.e. they describe important notions of the domain using expressions that are build from atomic concepts and atomic roles.

In this section we will present the notation of f-$\mathcal{SHIN}$ which is a fuzzy extension of DL $\mathcal{SHIN}$ [11]. A description language consist of an alphabet of distinct concepts names ($\mathbf{C}$), role names ($\mathbf{R}$) and individual names ($\mathbf{I}$), together with a set of constructors to construct concept and role descriptions. If R is a role then R^- is also a role, namely the inverse of R. The f-$\mathcal{SHIN}$-concepts are inductively defined as follows,

1. If $C \in \mathbf{C}$, then C is a f-$\mathcal{SHIN}$-concept,
2. If C and D are concepts, R is a role and $n \in \mathbb{N}$, then $(\neg C)$, $(C \sqcup D)$, $(C \sqcap D)$, $(\forall R.C)$, $(\exists R.C)$, $(\geq nR)$ and $(\leq nR)$ are also f-$\mathcal{SHIN}$-concepts.

In contrast to crisp DLs, the semantics of fuzzy DLs are provided by a *fuzzy interpretation* [23, 22]. A fuzzy interpretation is a pair $\mathcal{I} = \langle \Delta^{\mathcal{I}}, \cdot^{\mathcal{I}} \rangle$ where the domain $\Delta^{\mathcal{I}}$ is a non-empty set of objects and $\cdot^{\mathcal{I}}$ is a fuzzy interpretation function, which maps an individual name a to elements of $\mathsf{a}^{\mathcal{I}} \in \Delta^{\mathcal{I}}$ and a concept name A (role name R) to a membership function $\mathsf{A}^{\mathcal{I}} : \Delta^{\mathcal{I}} \to [0, 1]$

A f-$\mathcal{SHIN}$ knowledge base Σ is a triple $\langle \mathcal{T}, \mathcal{R}, \mathcal{A} \rangle$, where $\mathcal{T}$ is a fuzzy $TBox$, $\mathcal{R}$ is a fuzzy $RBox$ and $\mathcal{A}$ is a fuzzy $ABox$. $TBox$ is a finite set of fuzzy concept axioms which are of the form $C \equiv D$ called fuzzy concept inclusion axioms and $C \sqsubseteq D$ called fuzzy concept equivalence axioms, where C, D are concepts, saying that C is equivalent or C is a sub-concept of D, respectively. Similarly, $RBox$ is a finite set of fuzzy role axioms of the form $\mathsf{Trans}(R)$ called fuzzy transitive role axioms and $R \sqsubseteq S$ called fuzzy role inclusion axioms saying that R is transitive and R is a sub-role of S respectively. Ending, $ABox$ is as finite set of fuzzy assertions of the form $\langle a : C \bowtie n \rangle$, $\langle (a, b) : R \bowtie n \rangle$, where $\bowtie$ stands for $\geq, >, \leq$ or $<$ or $a \not= b$, for $a, b \in \mathbf{I}$. Intuitively, a fuzzy assertion of the form $\langle a : C \geq n \rangle$ means that the membership degree of a to the concept C is at least equal to n. We call assertions defined by $\geq, >$ *positive* assertions, while those defined by $\leq, <$ *negative* assertions [22].

The main reasoning service for a crisp DL is knowledge base (KB) *consistency*. Additional reasoning services that are *concept satisfiability*, *concept subsumption* and *entailment* are reduced to KB *consistency*. To decide the consistency of a knowledge base, a tableau based algorithm tries to construct a model of it by structurally decomposing the concepts in the knowledge base, thus inferring new constraints on the elements of this model.

These reasoning services are also available by f-$\mathcal{SHIN}$ together with greatest lower bound queries which take the advantage of the fuzzy element. The tableau algorithm for f-$\mathcal{SHIN}$ was presented by Stoilos et al [22]. In this paper fuzzy reasoning engine FiRE[2] that is based on f-$\mathcal{SHIN}$ is used for text categorization. In simple words the reasoning services can be presented as queries. Hence, fuzzy entailment queries ask whether an individual participates in a concept in a specific degree. Subsumption queries on the other hand ask whether a concept is sub-concept of another concept, e.g. $FootBallCategory \sqsubseteq Sports$. Finally, since a fuzzy $ABox$ $\mathcal{A}$ might contain many positive assertions for the same individual (pair of individuals), without forming a contradiction, it is in our interest to compute what is the best lower and upper truth-value bounds of a fuzzy assertion. The concept of *greatest lower bound* of a fuzzy assertion w.r.t. Σ was defined in [23]. Greatest lower bound ask for the degree of participation of an individual in a concept.

3 Semantic Indexing and Textual Categorization

An integrated handling of concepts and relations derived from the textual metadata that accompany every multimedia document is only possible with the utilization and construction of a *semantic index*. This process is based on the fuzzy knowledge model described in Section 2.1. After the semantic index construction, multimedia documents are classified to topics (e.g. *sports*, *politics*, etc.) through a fuzzy clustering of the concepts associated to each document. In the framework of our current work, the semantic index that links concepts to multimedia documents is constructed in two steps. Initially, all available multimedia documents are mapped to concepts along with a degree of confidence, where this mapping includes a semantic interpretation of their lexical terms (originating from the accompanying textual annotation). On a second step, semantic indexing (associated concepts) of each document is further analyzed in order to estimate the degree to which the given document is related to each one of a predefined set of topics.

3.1 Semantic topic categorization

The semantic indexing procedure refers to the construction of the semantic index, i.e. an association between multimedia documents and concepts, obtained through analysis of the associated textual annotation (in case of raw multimedia content) or the actual textual terms (in case of textual documents). In a further analysis process, each document is analyzed to detect associated topics. This is achieved by applying a meaningful clustering approach on the concepts associated to a document according to their common meaning. In a more mathematical manner, the set to be clustered for document d

[2] FiRE can be found at `http://www.image.ece.ntua.gr/~nsimou` together with installation instructions and examples.

is its support S_d, defined as:

$$S_d = \{s \in S : \mathcal{I}(s, d) > 0\} \tag{10}$$

where $\mathcal{I}$ represents the semantic index (directory) and is a fuzzy relation between documents and concepts, i.e. $\mathcal{I}(s, d)$ represents the degree of membership of concept s in document d. Letting C_d be the set of clusters detected in d, each cluster $c \in G_d$ is a crisp set of concepts or in other words: $c \subseteq S_d$. As this approach alone is not sufficient for efficient text categorization (as we need to support documents belonging to multiple distinct topics by different degrees and at the same time retain the robustness and efficiency of the clustering approach), without any loss of functionality or increase of computational cost, we adopt the cluster fuzzification methodology proposed in [16]. Thus, we shall identify the fuzzy set of topics W_d related to document d.

Since in our work the number of topics that may be encountered in a multimedia document is not known beforehand (although the overall number of possible topics Y is considered to be available), we do not know the number of clusters as input. Thus, traditional partitioning clustering methods are inapplicable [15]. The same applies, for instance, to the use of a supervised clustering method which allows one concept to belong to two or more clusters, like fuzzy c-means [3], because it also requires the number of concept clusters as input, i.e. it uses a hard termination criterion on the amount of clusters. As a result, the counterparts of partitioning methods, i.e. *hierarchical clustering* methods, are the only ones suitable for the current task. In general, they are divided into agglomerative and divisive methods, but since the former are more robust and are more widely studied and applied, they will be used herein. Their general structure, adjusted for the needs of the problem at hand, is as follows:

1. When considering document d, turn each concept $s \in S_d$ into a singleton, i.e. into a cluster c of its own.
2. For each pair of clusters c_1, c_2 calculate a compatibility indicator $K(c_1, c_2)$, which is also referred to as cluster similarity, or distance metric.
3. Merge the pair of clusters that have the best K. Depending on whether this is a similarity or a distance metric, the best indicator could be selected using the min or the max operator, respectively.
4. Continue at step 2, until the termination criterion is satisfied. The termination criterion most commonly used is the definition of a threshold for the value of the best compatibility indicator K.

The two key points in the above hierarchical clustering approach are the identification of the clusters to merge at each step, i.e. the definition of a meaningful metric for K, and the identification of the optimal terminating step, i.e. the definition of a meaningful termination criterion. in our work, the clustering process terminates when the concepts are clustered into sets that correspond to distinct topics. Therefore, the termination criterion is a threshold on the selected compatibility metric. The output is a set of clusters C_d, where each cluster $c \in C_d$ is a crisp set of concepts $c \subseteq S_d$.

3.2 Cluster fuzzification and topics' extraction

The above clustering method determines successfully the count of distinct clusters that exist in S_d, but in principle is inferior to partitioning approaches in the following senses:

(i) it only creates crisp clusters, i.e. it does not allow for degrees of membership in the output, and (ii) it only creates partitions, i.e. it does not allow for overlapping among the detected clusters. However, in real-life a concept may be related to a topic to a degree other than 1 or 0, and may also be related to more than one distinct topic. In order to overcome such problems, fuzzification of the clusters is carried out based on the fuzzification methodology described in [16].

In particular, a fuzzy classifier $C_c : S \rightarrow [0,\ 1]$ is constructed, that measures the degree of correlation of a concept s with cluster c and is used to expand the detected crisp clusters to include more concepts. A concept s should be considered correlated with cluster c, if it is related to the common meaning of the concepts in c. Of course, as not all clusters are equally compact, a meaningful correlation measure needs to be determined, according to the characteristics of the cluster in question. This measure should fulfill the following properties:

- $C_c(s) = 1$, if the semantics of s imply it should belong to c, e.g. $C_c(s) = 1, \forall s \in c$
- $C_c(s) = 0$, if the semantics of s imply it should **not** belong to c.
- $C_c(s) \in (0, 1)$, if s is neither totally related, nor totally unrelated to c.

Finally, cluster c is replaced by the fuzzy cluster $c_f : c_f = \sum\limits_{s \in S_d} s/C_c(s)$, using again the *sum notation* [15] for fuzzy sets.

At this point, we should also stress out the fact that the above described process of fuzzy hierarchical clustering has been based on the crisp set S_d, thus ignoring any potential fuzziness that may reside in the semantic index. The incorporation of this kind of information into the clusters' calculation process, is an open research field and part of our ongoing research work. In this paper, the ultimate goal remains to identify the fuzzy set of topics related to document d, initially through the calculation of the set of topics $W(c)$ related to each cluster c. The set of topics that correspond to a document is the set of topics that belong to any of the detected clusters of concepts that index the given document. More formally: $W_d = \bigcup\limits_{c \in G_d} W(c)$, where $\bigcup$ is a fuzzy co-norm and C_d is the set of clusters that have been detected in d. It is easy to see that W_d will be high if a cluster c is detected in d, and additionally, if the cardinality of c is high and the degree of membership of s in the context of the cluster is also high (i.e., if the topic is related to the cluster and the cluster does not consist of misleading concepts). Of course, the set of topics that correspond to a document d are derived from the *a priori* known set of all possible topics Y.

4 The Fuzzy Knowledge Base

As the second step of our approach, a fuzzy knowledge base is used for the categorization of multimedia documents. A fuzzy knowledge base Σ consists of a fuzzy $TBox$, a fuzzy $RBox$ and a fuzzy $ABox$. $TBox$ and $RBox$ introduce the terminology, i.e the vocabulary of the application domain, while $ABox$ contains the assertions about named individuals in terms of this vocabulary. As it will be further illustrated within this section, our current domain of interest is *news* and its multimedia documents can be categorized to various categories according to their content and their length.

The semantic-based refinement of the textual categorization approach presented in Section 3, initializes the ABox of the fuzzy knowledge base. The semantic indexing algorithm evaluates the multimedia documents producing a set of concepts, which exist in the document, together with a degree $w \in [0, 1]$ that the higher it is the more synonyms of the specific concept exist in the document. This kind of information is represented as positive assertions in the constructed fuzzy knowledge base using the document as the individual participating in the concept to the given w. Since the amount of the utilized concepts during the previous step of textual categorization forms a very large vocabulary (150.000 concepts derived from the Wordnet[3] dataset), a reduction for the creation of the fuzzy knowledge alphabet was necessary. This reduction was made selecting a representative set of concepts for every category. The set of concepts that comprise our fuzzy knowledge base alphabet is as follows:

$Concepts = \{$*Event, Goal, Goalkeeper, Football, Hands, Penalty, Competition, Extra, General, Intercept, Team, Audience, Turf, Modern, Player, Today, Air, Sport, Score, Touch, Kick Grounding, Crossbar, League, Winner, Include, Vary, Numbers, Role, Injury, Times, Primary, Shirt, Arms, Defense, Contraction, Local, Objective, Favorite, Speaking, Amateur, Individual, Professional Simple, Wearing, Behavior, Female, Male, Footwear, Whole, Tournament, Country, Squad, Nation, Year, Quarter, Shoots, Rebounds, Points, Timeout, Block, Slam-dunk, Dig, Spike, Serve, Set, Ace, Government, Law, Minister, Election, Parliament, Price, Profit, Bank, Commercial, Cost, Fall, Buy, Loss, Rate, Rise, Student, Term, Funding, Academy University, School, Administration, Plus, Basic, Money, Budget, Billion, Week, Holiday, Month, Choice, Sampling, Nomination, Overall, Gender, Analyst, Chief, Part, Identity, Fact, Rival, Defection, Contest, Sharp, Assertion, Public, Politics, Leading, Organizer* $\}$

Additionally, the number of paragraphs of every document are considered as an information concerning the length of the document and they are represented by the number restriction constructors of f-$\mathcal{SHIN}$. The set of roles is as follows:

$Roles = \{has - paragraph\}$.

The effective extraction of implicit knowledge from the explicit one requires an expressive terminology, which is able to define higher concepts. The categories in which the documents are classified are defined by the terminology that is illustrated in table 2. As it can be observed, the categories: *Business*, *Education* and *Politics* are subcategories of *News* and similarly categories: *FootballCategory BasketBallCategory* and *VolleyBallCategory* are sub-categories of *Sports*.

Sport sub-categories have been defined in a similar manner. Their defining axioms require that a document consist of some necessary concepts and one of some optional concepts. The necessary concepts for each of them are the concept *Team* along with the winning criterion. (i.e. Points in the case of basketball and volleyball and goals in the case of football) The optional concepts consist of the special terms used for each sport. Someone may wonder why the concept *Football* is not one of the essential keywords for the categorization of a document in football category. The answer is that many documents may describe a football game but without necessarily using this concept.

On the other hand, news subcategories due to their abstract content were defined in a more general way. Hence, *Politics* and *Business* categories are defined by axioms which

[3] http://wordnet.princeton.edu/

$$
\begin{aligned}
\mathcal{T} = \{\mathsf{BasketBallCategory} &\equiv \mathsf{Team} \sqcap \mathsf{Points} \\
&\qquad \sqcap(\mathsf{Quarter} \sqcup \mathsf{Shoots} \sqcup \mathsf{Rebounds} \\
&\qquad\qquad \sqcup\mathsf{Timeout} \sqcup \mathsf{Block} \sqcup \mathsf{Slam-dunk}), \\
\mathsf{VolleyBallCategory} &\equiv \mathsf{Team} \sqcap \mathsf{Points} \\
&\quad \sqcap(\mathsf{Dig} \sqcup \mathsf{Spike} \sqcup \mathsf{Serve} \sqcup \mathsf{Set} \sqcup \mathsf{Block} \sqcup \mathsf{Ace}), \\
\mathsf{FootBallCategory} &\equiv \mathsf{Team} \sqcap \mathsf{Goal} \\
&\quad \sqcap(\mathsf{Football} \sqcup \mathsf{Penalty} \sqcup \mathsf{Goalkeeper} \sqcup \mathsf{Score}), \\
\mathsf{Sports} &\equiv \mathsf{FootBallCategory} \\
&\quad \sqcup\mathsf{BasketBallCategory} \sqcup \mathsf{VolleyBallCategory}, \\
\mathsf{Politics} &\equiv (\mathsf{Government} \sqcup \mathsf{Law} \sqcup \mathsf{Minister} \sqcup \mathsf{Parliament}) \\
&\qquad \sqcap(\mathsf{Administration} \sqcup \mathsf{Election} \sqcup \mathsf{Public}), \\
\mathsf{Business} &\equiv (\mathsf{Price} \sqcup \mathsf{Profit} \sqcup \mathsf{Bank} \sqcup \mathsf{Commercial} \sqcup \mathsf{Cost}) \\
&\qquad \sqcap(\mathsf{Fall} \sqcup \mathsf{Buy} \sqcup \mathsf{Loss} \sqcup \mathsf{Rate} \sqcup \mathsf{Rise}), \\
\mathsf{Education} &\equiv \mathsf{Student} \sqcup \mathsf{School} \sqcup \mathsf{Term} \\
&\qquad \sqcup\mathsf{Funding} \sqcup \mathsf{Academy} \sqcup \mathsf{University}, \\
\mathsf{News} &\equiv \mathsf{Politics} \sqcup \mathsf{Business} \sqcup \mathsf{Education}, \\
\mathsf{Bulletin} &\equiv (\geq 1\mathsf{has-paragraph}) \sqcap (\leq 1\mathsf{has-paragraph}), \\
\mathsf{NewSummary} &\equiv (\geq 2\mathsf{has-paragraph}) \sqcap (\leq 3\mathsf{has-paragraph}), \\
\mathsf{ShortArticle} &\equiv (\geq 3\mathsf{has-paragraph}) \sqcap (\leq 6\mathsf{has-paragraph}), \\
\mathsf{Article} &\equiv \leq 6\mathsf{has-paragraph}\}
\end{aligned}
$$

Table 2. Knowledge Base ($TBox$)

require the occurrence of two concepts. Each of these concepts belongs to a wider set of keywords related to the thematic category though on their own they are insufficient. Consider for example a document which includes concept *Election*. This concept alone is not necessarily included in a document that describes a political event. On the contrary, if this concept is included in a document together with one of the concepts *Government*, *Law*, *Minister*, *Parliament* then this document possibly describes a political event. The *Education* category is less descriptive than the others since every occurrence of some domain-relative concepts, within the document, classify the document in this category.

Finally, four axioms have been defined according to the length of the document. The documents which have exactly one paragraph are categorized as *Bulletin*. A document is categorized as *NewsSummary* if it has 2 or more than 2 and 3 or less than 3 paragraphs, as *ShortArticle* if it has 3 or more than 3 and 6 or less than 6 paragraphs and as *Article* if it has 6 or more paragraphs.

5 A Use Case Scenario

In order to demonstrate the efficiency of the proposed methodology followed in our work, we have developed a use case scenario applied on a set of multimedia documents, derived from the *sports* and *politics* domains. The scenario comprises of nine documents $(d_1, ..., d_9)$, the set of concepts S that comprise the fuzzy knowledge base alphabet presented in Section 4 and relation T, defined on S, as in subsection 2.1. The overall categorization results for the 9 documents are presented in Table 3. The left side of each document column indicates the textual categorization results, whereas the right side of each column presents the results obtained after the application of the proposed semantic reasoning approach.

Topic	Documents																	
	d_1		d_2		d_3		d_4		d_5		d_6		d_7		d_8		d_9	
sports	0.88	0.69	0.86	0.67	0.94	0.56	0.88	0.52	0.91	0.71								
football		0.68		0.67		0.51		0.54		0.71								
politics	0.85				0.91		0.92		0.91		0.88	0.67	0.93	0.73	0.93	0.72	0.93	0.76
business														0.62				
education	0.89		0.89				0.89		0.89							0.71	0.89	0.68
news							0.97											

Table 3. Document categorization results. Values below 0.10 are omitted.

In order to clarify the process of thematic categorization, we analyze further the specific steps for document d_2. Document d_2 is a sequence of video shots from a news broadcast. Due to the diversity of stories presented in the sequence, the concepts that are detected and included in the semantic index are quite unrelated to each other. The semantic indexing of document d_2 is given by Table 4 and at the same time acts as the fuzzy concept assertions for the forthcoming fuzzy reasoning approach. Considering the described hierarchical clustering and cluster fuzzification process, we compute the following fuzzy clusters of concepts for the document (clusters with extremely low cardinality are omitted):

$$\begin{aligned} c_{475} &= game/0.11 + play/0.01 \\ c_{474} &= fair/0.20 + union/0.09 \\ c_{473} &= score/0.05 \end{aligned} \tag{11}$$

Based on the methodology presented in Section 3, we identify the topics *sports* and *education* related to d_2, as described in the corresponding column of Table 3. We observe that the proposed methodology identifies the existence of more than one distinct topic in the document.

The set of concepts included in the document together with a degree of importance belong to the set $Concepts$ defined in Section 4 as the alphabet of our fuzzy knowledge base and the assertions declared for this document are illustrated in Table 4. The way that the information regarding the length of the document is represented requires special attention. The number of paragraphs in a multimedia document can be easily evaluated (d_2 has 3 paragraphs). So d_2 crisply participates in a complex concept that requires at least 3 and at most 3 paragraphs (i.e exact 3) using the role has-paragraph of our defined alphabet.

Concept	Degree
Touch	0.5
Kick	0.5
Grounding	0.5
Goal	0.75
Penalty	0.67
Crossbar	0.67
Team	0.67
League	0.75
Winner	0.67
Include	0.75
Vary	0.75
Score	0.5
$(\geq 3\mathsf{has}-\mathsf{paragraph}) \sqcap (\leq 3\mathsf{has}-\mathsf{paragraph})$	1.0

Table 4. Fuzzy concept assertions for d_2

For the classification of the document in a thematic category the greatest lower bound reasoning service is used. Greatest lower bound queries are performed for all the defined thematic categories. The thematic categories in which the document d_2 is classified are presented in Table 5. Since d_2 participates in concepts *Team* with degree 0.67, *Goal* with degree 0.75 and *Penalty* with degree 0.67 is classified in category *FootballCategory* with degree 0.67 that is the lower allowed bound for participation in it. It is furthermore classified in category *Sports* with the same degree since *FootballCategory* is sub-category of *Sports*. Concerning the length-categories, d_2 participates in all of them but with different degrees. Though this participation might seem strange, it is normal and can be explained due to the fuzzy properties of f-$\mathcal{SHIN}$. Furthermore, as it can be observed d_2 participates is *NewSummary* which is the closest length-category to d_2, with degree 1 according to the greatest lower bound query for the specific length-category.

6 Conclusions

In this paper, a reasoning-based approach for multimedia document categorization was presented. A multimedia document is evaluated by the semantic indexing algorithm producing a set of concepts, which exist in the document, together with a degree that

Concept	Degree
FootBallCategory	0.67
Sports	0.67
Bulletin	0.25
NewSummary	1.0
ShortArticle	0.75
Article	0.67

Table 5. Document Classification for d_2

the higher it is the more synonyms of the specific concept exist in the document. These results together with the number of paragraphs for each multimedia document comprise the fuzzy $ABox$ of a fuzzy knowledge base build for the domain of *news* using DL f-$\mathcal{SHIN}$. This knowledge base is used to classify a multimedia document in a category with a degree of participation by using the greatest lower bound reasoning service of f-$\mathcal{SHIN}$ that takes the advantage of the fuzzy element.

This approach was evaluated on various multimedia documents producing reasonable results, presenting by that way its strong potential, although limitations that will be improved in future works are also present. Firstly, a more descriptive representation of the multimedia documents is considered to be feasible. Such a representation could include an enriched alphabet and also the division of a document into paragraphs with different extraction of concepts for each paragraph. This representation would allow thematic categorization of every paragraph and then an improved overall document categorization according to them.

Additionally, a very promising approach would be the parallel use of an image-video analysis algorithm together with the semantic indexing algorithm of our approach for a multimedia document. The extracted semantics would provide rich information that could then be used by the fuzzy reasoning engine FiRE for the extraction of implicit information regarding the multimedia document under consideration.

7 Acknowledgement

This research was partially supported by the European Commission under contracts FP6-001765 aceMedia and FP6-027026 K-Space. The expressed content is the view of the authors but not necessarily the view of the projects as a whole.

References

1. Akrivas, G., Stamou, G., Kollias, S.: Semantic association of multimedia document descriptions through fuzzy relational algebra and fuzzy reasoning. IEEE Transactions on Systems, Man, and Cybernetics, part A **34(2)** (2004)
2. Baader, F., McGuinness, D., Nardi, D., Patel-Schneider, P.: The Description Logic Handbook: Theory, implementation and applications. Cambridge University Press (2002)

3. Benkhalifa, M., Bensaid, A., Mouradi, A.: Text categorization using the semi-supervised fuzzy c-means algorithm. Proceedings of the 18th International Conference of the North American Fuzzy Information Processing Society - NAFIPS (1999) 561–565
4. Berry, M.W., Dumais, S.T., O'Brien, G.W.: Using linear algebra for intelligent information retrieval. SIAM Review **37(4)** (1995) 177–196
5. Bimbo, A.D.: Visual Information Retrieval. Morgan Kaufmann Publishers (2001)
6. Burgin, R.: The retrieval effectiveness of five clustering algorithms as a function of indexing exhaustivity. Journal of the American Society for Information Science **46(8)** (1995) 562–572
7. Cutting, D., Karger, D.R., Pedersen, J.O., Tukey, J.W.: Scatter/gather: A cluster-based approach to browsing large document collections. Proceedings of the ACM/SIGIR (1992) 318–329
8. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by latent semantic analysis. Journal of the American Society of Information Science **41(6)** (1990) 391–407
9. Heinsohn, J.: Probabilistic description logics. In: Proceedings of UAI-94. (1994) 311–318
10. Hofmann, T.: Probabilistic latent semantic indexing. Proceedings of the 22nd ACM-SIGIR International Conference on Research and Development in Information Retrieval (1999) 50–57
11. Horrocks, I., Sattler, U., Tobies, S.: Reasoning with Individuals for the Description Logic $\mathcal{SHIQ}$. In MacAllester, D., ed.: CADE-2000. Number 1831 in LNAI, Springer-Verlag (2000) 482–496
12. Jaeger, M.: Probabilistic reasoning in terminological logics. In: Proceedings of KR-94. (1994) 305–316
13. Klir, G.J., Yuan, B.: Fuzzy Sets and Fuzzy Logic: Theory and Applications. Prentice-Hall (1995)
14. Landauer, T., Foltz, P., Laham, D.: An introduction to latent semantic analysis. Discourse Processes **25** (1998) 259–284
15. Miyamoto, S.: Information clustering based on fuzzy multisets. Inf. Process. Manage. **39**(2) (2003) 195–213
16. Mylonas, P., Vallet, D., Fernandez, M., Castells, P., Avrithis, Y.: Ontology-based personalization for multimedia content, 3rd European Semantic Web Conference - Semantic Web Personalization Workshop, Budva, Montenegro, 11-14 June 2006 (2006)
17. Sahami, M., Yusufali, S., Baldonado, M.Q.W.: Real-time full-text clustering of networked documents. In: AAAI/IAAI. (1997) 845
18. Salembier, P., Smith, J.R.: Mpeg-7 multimedia description schemes. IEEE Transactions on Circuits and Systems for Video Technology **11(6)** (2001) 748–759
19. Salton, G., Mcgill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, Inc., New York, NY, USA (1986)
20. Schütze, H., Silverstein, C.: Projections for efficient document clustering. In: SIGIR '97: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM Press (1997) 74–81
21. Stoilos, G., Stamou, G., Tzouvaras, V., Pan, J.Z., Horrocks, I.: Fuzzy OWL: Uncertainty and the Semantic Web. In: Proc. of the OWL-ED 2005)
22. Stoilos, G., Stamou, G., Tzouvaras, V., Pan, J.Z., Horrocks, I.: The fuzzy description logic f-shin. In: a, International Workshop on Uncertainty Reasoning For the Semantic Web (2005) (2005)
23. Straccia, U.: Reasoning within fuzzy description logics. Journal of Artificial Intelligence Research **14** (2001) 137–166
24. Wallace, M., Akrivas, G., Mylonas, P., Avrithis, Y., Kollias, S.: Using context and fuzzy relations to interpret multimedia content. Proceedings of the 3rd International Workshop on Content-Based Multimedia Indexing (CBMI) (2003)