MOWIS: A system for building Multimedia Ontologies from Web Information Sources

Vincenzo Moscato, Antonio Penta, Fabio Persia, Antonio Picariello
University of Naples
Dipartimento di Informatica e Sistemistica
via Claudio 21, 80125, Naples
{vmoscato,a.penta,fabio.persia,picus}@unina.it

ABSTRACT

Defining ontologies within the multimedia domain still remains a challenging task, due to the complexity of multimedia data and the related associated knowledge. In this paper, we propose: i) a novel multimedia ontology model that combine both low level descriptors and high level semantic concepts; ii) an automatic construction of ontologies using the Flickrweb services, that provide images, tags, keywords and sometimes useful annotation describing both the content of an image and personal interesting information. Eventually, we describe an example of automatic ontology construction in a specific domain.

1. INTRODUCTION

Nowadays, a lot of repositories containing both multimedia and the related annotations or metadata are publicly available on the web. Such kind of information may be used for an automatically generation of multimedia knowledge, particularly suitable for a variety of applications, such as information retrieval, browsing, data mining and so on.

It is well known in the literature that despite the tons of papers produced about multimedia databases and knowledge representations, there is not yet an accepted solution to the problem of *how to represent, organize and manage multimedia data and the related semantics by means of a formal framework*.
Usually, a multimedia database is described by means of "flat" metadata, the most of the times using a predefined set of metadata (as in mpeg standard), or sometimes using small annotation in natural languages: such kind of structures are substantially inadequate to support complete retrieval by content of image documents.

It is the authors' opinion that there is still a great work to do with respect to the *intensional aspects* of a *multimedia ontology*:

- *what a multimedia ontology is*: is it a taxonomy, or a semantic network of metadata (tags, annotations)?
- *does a multimedia ontology support concrete data*: what is the role of rough data – image, video, audio data– if any?
- *what a multimedia semantic is*: how to define and capture the semantics of multimedia data?
- *how to build extensional ontologies*: once defined a suitable formal framework, can we automatically build the defined multimedia ontologies?

Appears in the Proceedings of the 1st Italian Information Retrieval Workshop (IIR'10), January 27–28, 2010, Padova, Italy.
http://ims.dei.unipd.it/websites/iir10/index.html
Copyright owned by the authors.

Throughout the rest of paper, we will try to give an answer to all the previous cited aspects; in particular the original contribution of this work is: (i) to propose a novel multimedia ontology framework, in particular related to the image domain; (ii) to propose a technique for building ontologies, that operates on large corpora of human annotated repositories, namely the Flickr [7] database, considering both low level image processing strategies and keywords and annotations produced by humans when they store the produced data.

We provide an algorithm for creating image ontology in a specific domain gathering together all this different information. We then provide an example of automatic construction of image ontology and a discussion of the encountered problems and the provided solutions. We concluded that the framework is promising and sufficiently scalable to different domains.

The remaind of paper is organized as follows. Section 2 outlines the related work related to the multimedia ontology topic. In Section 3 the process for building an Image Ontology is described. Section 4 details the system architecture with some implementation issues and a case study for our process is shown in Section 5. In Section 6 some discussions and conclusions are reported.

2. RELATED WORKS

In the last few years, several papers have been presented about multimedia systems based on knowledge models, image ontologies, fuzzy extension of ontology theories.

In almost all the works, multimedia ontologies are effectively used to perform *semantic annotation* of the media content by manually associating the terms of the ontology with the individual elements of the image or video domains [12], thus demonstrating that the use of ontologies can enhance classification precision and image retrieval performance.

Instead of creating a new ontology from the scratch, other approaches [3] extend *WordNet* to image specific concepts, using an annotated image corpus as an intermediate step to compute similarity between example images and images in the image collection.

For solving the uncertain reasoning problems, the theory of fuzzy ontologies is presented in several works, as an extension of ontologies with crisp concepts as in the paper [6], that presents a complete fuzzy framework for ontologies. While in [8], the authors introduce a description logic framework for the interpretation of image contents.

Multimedia semantic papers based on *MPEG-7* [9] are very interesting. The MPEG-7 framework consists of *Descriptors* (*Ds*) and *Descriptor Schemes* (*DSs*) that represent features for multimedia, and more complex structures grouping *Ds* and *DSs*, respectively.

In particular, the MPEG-7 standard includes tags that describe visual features (e.g. color), audio features (e.g. timbre), structure (e.g. moving regions and video segments), semantic (e.g. object and events), management (e.g. creator and format), collection organization (e.g. collections and models), summaries (e.g. hierarchies of key frames) and, even, user preferences (e.g. for search) of multimedia. In this way the standard includes descriptions of low-level media-specific features that can often be automatically extracted from the different media types.

Unfortunately, MPEG-7 is not currently suitable for describing top-level multimedia features, because: i) its XML Schema-based nature prevents the effective manipulation of descriptions and its use of URNs is cumbersome for the web; ii) it is not open to the web standards for representing knowledge.

Other efforts have been also done in order to translate the semantic of the standard in some knowledge representation languages [11]. All these methods perform a *one to one* translation of MPEG-7 types into OWL concepts and properties.

Finally, a very interesting work reported in [1] has been proposed in order to define a multimedia ontology. The authors try to define a novel multimedia ontology that takes into account the semantic of MPEG-7 standard. They started using some patterns derived from the foundational ontology *DOLCE* [10]. In particular they used two design patterns *Descriptions & Situations* (*D & S*) and *Ontology of Information Objects* (*OIO*). The obtained ontology already covers a very large part of the standard, while their modeling approach has the aim to offer even more possibilities for multimedia annotation than MPEG-7 since it is truly interoperable with existing web ontologies. This approach fits interoperability purposes, but some constraints on the image semantics are introduced.

3. BUILDING AN IMAGE ONTOLOGY

3.1 An Image Ontology Model

An ontology is usually referred as an "explicit specification of a conceptualization" which is, in turn, "the objects, concepts, and other entities that are presumed to exist in some area of interest and the relationships that hold among them".

Stressing its conceptual nature, an ontology may be considered as a *theory* used to represent relevant notion about domain modeling, a "domain" being classified in terms of concepts, relationships and constraint on them.

Let us consider the image domain: as usual in a given media, we detect symbols, objects and concepts; in a certain image we have a region of pixels (symbol) related to a portion of multimedia data; this region is an instance (object) of the certain concept.

In other words, we can detect concepts but we are not able to disambiguate among the instances without some specific knowledge.

A simplified version of the described vision process will consider only two main levels: *Low* and *High*. In fact, the knowledge associated to an image can be easily described at two different levels of analysis:

- *Low level* - the low-level descriptions of raw images;
- *High level* - general or domain-specific image content concepts that can be derivable or less from low-level ones.

It's the author's opinion that an image ontology has to take into account these specific characteristics, as captured by the following definition:

DEFINITION 1 (IMAGE ONTOLOGY). *An Image Ontology is a directed and labeled graph* $(\mathcal{V}, \mathcal{E}, \rho)$*, where:*

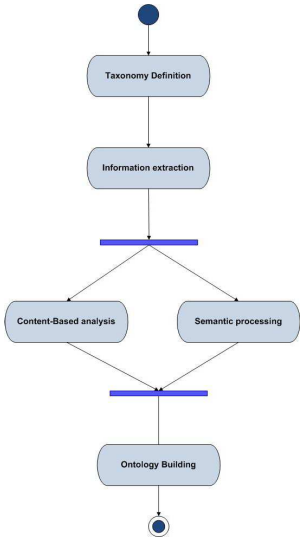

Figure 1: Image Ontology Building Process

1. $\mathcal{V}$ *is a finite set of nodes that can be of different kinds:*
   - low-level nodes $(\mathcal{V}_l)$*, corresponding to an image with the related properties:*
     - *content (e.g. texture, shape, color, objects, etc...) or more enhanced features;*
     - *metadata (e.g. author, title, description, tags, etc...);*
   - high-level nodes $(\mathcal{V}_h)$*, corresponding to* general concepts domain-specific concepts*, or* image content concepts *(that could be associated to low-level nodes);*
2. $\mathcal{E}$ *is a subset of* $(\mathcal{V} \times \mathcal{V})$*;*
3. ρ *is a function that associates to each couple of nodes a label indicating the kind of* relationship *between the two nodes* ρ_s*, and its reliability degree* $\rho_r \in [0, 1]$*:* $\rho : \mathcal{E} \rightarrow \langle \rho_s, \rho_r \rangle$*.*

Depending on the type of relationship in our model, we distinguish among:

- *similarity relationship*: relates between two low-level nodes (images) in function of their similarity degree, exploiting classical algorithms of image matching based on low-level features (e.g. color, texture, shape, etc...);
- *representativeness relationship*: relates between high-level and low-level nodes, containing those content features that better represents the associated concept, by means of clustering or classification algorithms that determine the probability that an image is a valid representative of the concept;
- *semantic relationship*: relates between two high-level nodes (example are those relationships such hypernym/hyponim, holonym/meronym, synonym, retrievable on lexical databases).

3.2 The Image Ontology building

The purpose of the image ontology building process (figure 1) is to perform the construction of the graph representing image ontology by a super-visioned approach.

The process is made of:

1. a definition of an initial taxonomy containing few high level nodes (related to the main concepts of a specific domain),

2. an extraction of useful information (images and annotations related to the taxonomy concepts) from several annotated web repositories,

3. a content-based analysis on the row-data and a semantic processing on the related textual annotations,

4. the ontology construction.

3.2.1 Taxonomy definition

Our image ontology building process is domain-oriented. Thus, during this step, it is necessary to define an initial taxonomy containing relevant concepts hierarchy of the considered domain that is represented by a subset of high level nodes.

3.2.2 Information extraction

The main objective of this task is to fetch images and the related textual annotations from web repositories, corresponding to the leaf high-level nodes of the image ontology, and to extract some useful low and high level information. Apposite communication API or web-services are exploited to obtain requested information.

In this paper we used Flickr as web image repository.

3.2.3 Content-Based analysis

The goal of such a task is to obtain a low-level description of images in terms of content features, using classical Computer Vision techniques.

We decided to use the salient points technique - based on the *Animate Vision* paradigm [2] - that exploits color, texture and shape information associated with those regions of the image that are relevant to human attention (*Focus of Attention*), in order to obtain a compact characterization (namely *Information Path*) that could be also used to evaluate the similarity between images, and for indexing issues.

An information path $\mathcal{IP}$=$\langle F_s(p_s;\tau_s)$,$h_b(F_s)$,$\Sigma_{F_s}\rangle$ can be seen as a particular data structure that contains, for each region $F(p_s;\tau_s)$ surrounding a given salient point (where p_s is the center of the region and τ_s is the the observation time spent by a human to detect the point), the color features in terms of HSV histogram $h_b(F_s)$, and the texture and shape features in terms of wavelet covariance signatures Σ_{F_s} (see [2] for more details).

Eventually, apposite super-visioned classification algorithms are exploited to determine content features [2].

3.2.4 Semantic processing

In this task the main objective is to discover textual *labels* that better reflect image semantic using NLP techniques and topic detection algorithms on the textual annotations coming from the considered image repositories. For what Flickr concerns, images usually have three main attached information: i) a *title*, ii) a content *description* and iii) a set of keywords, namely *tags*.

Titles in the majority of the cases contain text that summarizes the content of the images, while in other cases consist of automatically generated text that is not useful in the indexing process. *Descriptions* are very short and usually are not posted for retrieval purposes: they typically contain sentences concerning context of the picture, or user opinion. Finally, *Tags* are simple keywords users are asked (actually they may not insert any tag) to submit, that should describe the context of the image (e.g. among tags for a picture of an "elephant in an African landscape", you will probably see the words 'elephant', 'Africa' and 'landscape').

The simple use of tags does not improve the efficiency of indexing and searching contents. The absence of restrictions to the vocabulary from which tags are chosen can easily lead to the presence of *synonyms* (multiple tags for the same concept), *homonyms* (same tag used with different meaning), and *polysemies* (same tag with multiple related meanings). Also inaccurate or irrelevant tags result from the so called *'meta-noise'*, e.g. lack of stemming (normalization of word inflections), and from heterogeneity of users and contexts: hence an effective use of the tags requires these to be stemmed, disambiguated, and opportunely selected.

To these purposes, information coming from tags could be usefully analyzed in combination with titles and descriptions by suitable NLP technique that overcome the linguistic and semantic heterogeneity of such information, in order to extract a set of *relevant keywords* which more effectively represent image content.

In particular, the semantic processing, which is applied to the textual data attached to a given image can be decomposed into a set of sequential sub-tasks [13]: *meta-noise and named entity filtering*, *linguistic normalization*, *part of speech tagging*, *tokenization*, *word sense disambiguation* and *topic extraction*. Thus, the result of the semantic processing task is a set of *labels* (topics) with an associated *confidence* value - that represents the relative *importance* of the label (with respect to the other ones in the annotations) -, from the set of tags, title an descriptions.

3.2.5 Ontology building

As previously discussed, the obtained knowledge in terms of images, low-level characteristics and labels is then merged and translated in the shape of a graph representing image ontology.

In particular, in a first step, all images whose relevant labels are associated with a high confidence value to the high-level nodes, corresponding to the taxonomy leaves, will be represented by apposite low-level nodes; in addition, couple of image nodes, whose similarity (computed by means of the *Information Path Matching algorithm* [2]) is greater than a threshold will be linked by an edge having as reliability degree the related similarity measure.

In the successive step, previous images are clustered by used a *Balanced Expectation Maximazation* algorithm [2] applied in the feature spaces defined by the Information Path descriptors, in order to determine for the high-level nodes the set of images that better could represent the related concepts. Apposite edges (representative relationships) link such nodes with representatives of each cluster.

Eventually, by means of a *Learning Tag Relevance algorithm* [4], topics that are more relevant with respect to the content of images belonging to the same cluster (*winner topics*) are *promoted* to be image ontology high-level nodes. In particular, the tag relevance σ of a generic tag τ of the most significant image (*centroid*) of cluster C is computed by the following formula:

$$\sigma(\tau, C) = \sum_{i=1}^{m} |i_{df}(\tau) \cdot \frac{t_f(\tau.i) \cdot (a+1)}{t_f(\tau) + a \cdot (1 - b + b \cdot \frac{U_i}{\overline{U}})}| \quad (1)$$

where: $t_f(\tau, i)$ is the term frequency of topic τ with respect to the topics of all images belonging to C, $U_i, \overline{U}$ are the number of topics of $i-th$ image of C and the average number of tags related to all images belonging to C respectively, $i_{df}(\tau)$ is the inverse document frequency of τ in C. The winner topics, whose relevance is greater than a threshold, are finally inserted as high-level nodes in the ontology and *linked*, from one hand to the image node that corresponds to the cluster centroid and, from the other one, to those nodes which semantic distance (i.e. Wu/Palmer) is the minimum with respect to the current topic. If it is possible, the new ontology edge is labeled with the type of semantic relationship (e.g. hypernym/hyponim, holonym/meronym, etc...).

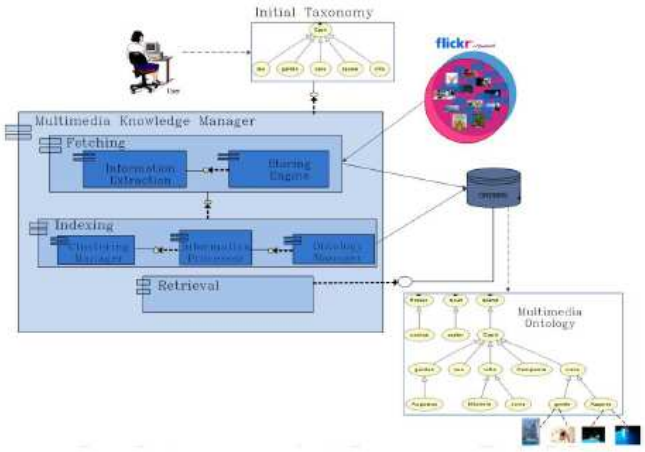

Figure 2: System Architecture

Thus, image ontology can be generated in an incremental way and in correspondence of pick-up operations from the Flickr repository.

4. THE SYSTEM ARCHITECTURE

The system architecture that supports the image ontology building process is shown in figure 2. User generates by an apposite graphical interface an *OWL* file coding the initial taxonomy containing relevant concepts of the considered domain. Such a file is then the input of the *Information Fetching* module that downloads images and the related annotations from the *Multimedia Repository*, using as search keywords the concepts related to leaf nodes of the taxonomy and some filters on users.

A *Storage Engine* module receives such information and stores image annotations (title, description, author, tags, labels, etc...) in a dedicated *RDF Database* and raw data together low-level characteristics in a *Image Database*. Each image is then identified in these databases by an *URI* (*Uniform Resource Identifier*).

Finally, the *Information Extraction* and *Information Processor* analyze both high level information stored into the RDF database and low level information contained into the Image database, in order to generate/update, by means of *Ontology Manager* and of *Clustering Manager*, and in according to the described process, a graph which represents final multimedia ontology.

For what implementation issues concerns, we notice that: (i) the initial taxonomy is generated by a JAVA desktop application that uses *Protégé* API; (ii) Flickr has been chosen as the multimedia repository; (iii) the Information Fetching module has been implemented as a JAVA application that exploits Flickr API; (iv) the RDF and Image Database have been realized by *Sesame* and *PostegreSQL* DBMS, respectively; (v) the Information Fetcing and Indexing packages have been implemented by apposite JAVA packages.

5. A CASE STUDY

This section describes a case study for our image ontology building process. In particular, we have built an ontology pertinent to *Capri*, a wonderful Italian island of the Sorrentine Peninsula, on the south side of the Gulf of Naples. A set of experts of natural and cultural attractions of Capri provided as initial taxonomy a graph reported containing the most relevant concepts in terms of high level nodes for the considered domain.

We used Flickr [7] as multimedia repository of annotated images. Flickr is one of the most popular web-based tagging system, that allows human participants to annotate a particular resource, such as web pages, blogs, images, with a freely chosen set of keywords, or tags, together with a short description of the content.

This kind of system has been recently termed *folksonomy* [5], i.e. a folk taxonomy of important and emerging concepts within user groups. The dynamic nature of these repositories assures the richness of the annotation; in addition, they are quite accurate, because they are produced by humans that want to share their images and the experience they have had, using tags and an annotation process.

The Flickr repository has been queried using as search keywords the *logical AND* between concepts reported in the leaf nodes of the taxonomy and the one corresponding to the root node and exploting some filters on user *id*s, in order to retrieve images really belonging to the domain. Each retrieved image undergoes a content-based analysis to determine the low-level description – i.e. the *IP* (*Information Path*) and content features. Moreover, in a first step we estimated similarity existing between each couple of different images by comparing their IPs by means of the *image path matching algorithm* [2].

All images belonging to the same concept are then clustered into different groups, which contain images that are more similar among themselves. We used as clustering procedure the *BEM algorithm* [2], that is recursively invoked to dynamically determine more fitting clusters without knowing a-priori the number of clusters themselves (that is usually proportional to number of images related to the current concept). Then we selected for each cluster the representative image as the closest one to all the other images of the cluster, and a suitable *representation probability* is associated to each representative image on the base of minimum and average distances.

The process is iterated for each taxonomy leaf concept and the ontology is incrementally built: images belonging to different topics could be linked on the base of their similarity values allowing to *merge* the multimedia knowledge in a unique graph. Thus, the more relevant tags are propagated in the ontology and linked to the other nodes.

We report in figure 3 a step by step complete example of the generation of *Capri* ontology.

6. CONCLUSION

In this paper we have addressed the problem of building a multimedia ontology in an automatic way using annotated image repositories. Our work differs from the previous papers presented in the literature for different reasons. First, we propose a notion of multimedia ontology, described by means of a graph and particularly suitable for managing the different levels of semantics of images. In addition, we obtain a dynamic generation of image ontologies using tags and annotations already produced by users in their social web networks.

Further works will be devoted to produce experimental results to evaluate the effectiveness of the produced ontologies with respect to other approaches by means of different criteria: *class match measure*, *density measure*, *semantic similarity measure*, *betweenness measure*.

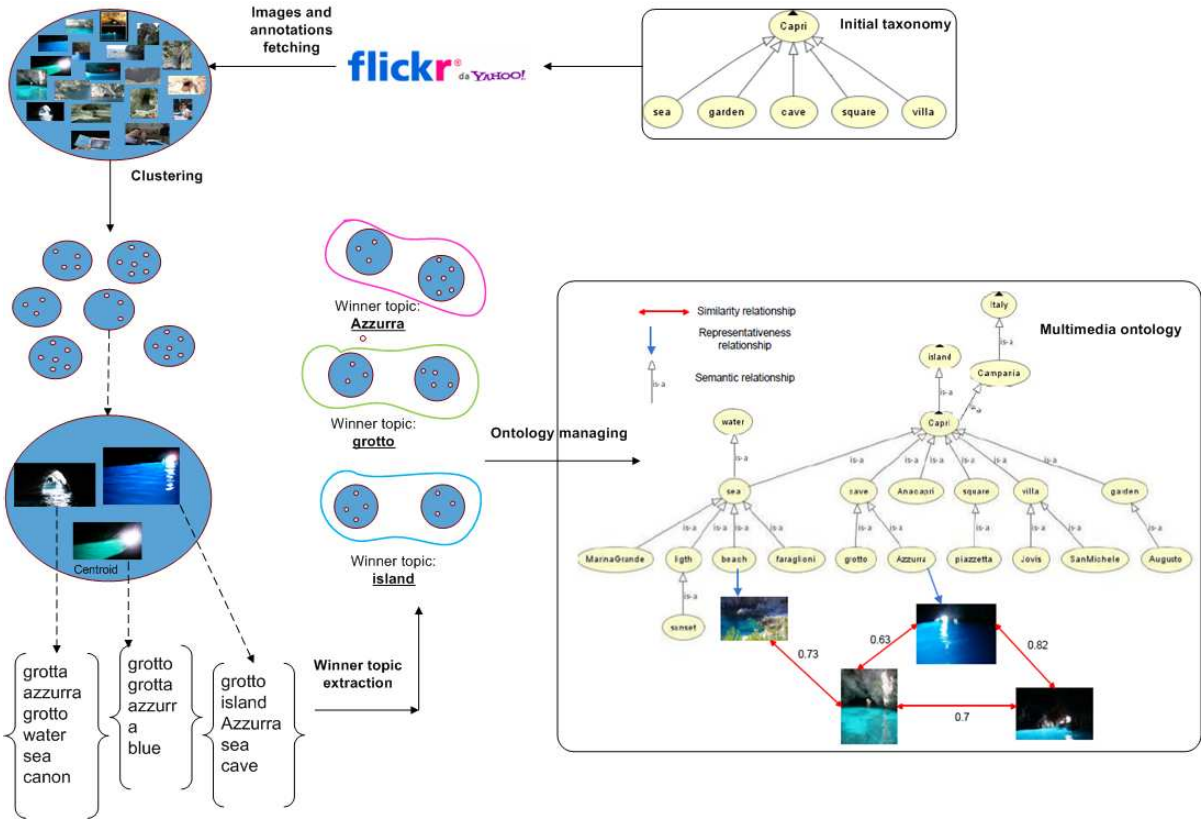

Figure 3: Bulding of the *Capri* Ontology

7. REFERENCES

[1] R. Arndt, R. Troncy, S. Staab, and L. Hardman. Adding formal semantics to mpeg-7: Designing a well-founded multimedia ontology for the web. Technical report, University of Koblenz, 2007.

[2] G. Boccignone, A. Chianese, V. Moscato, and A. Picariello. Context-sensitive queries for image retrieval in digital libraries. *Journal of Intelligent Information Systems*, 31(1), 2008.

[3] Y. Chang and H. Chen. Approaches of using a word-image ontology and an annotated image corpus as intermedia for cross-language image retrieval. In *Proceedings of Cross-Language Evaluation Forum*, 2006.

[4] S. Golder and A. Hubemann. Usage patterns of collaborative tagging systems. *Information Science*, 2006.

[5] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury. How flickr helps us make sense of the world: context and content in community-contributed media collections. *ACM Multimedia*, 2007.

[6] C. Lee, Z. Jian, and L. Huang. A fuzzy ontology and its application to news summarization. *IEEE Transactions on Systems, Man and Cybernetics,*, 35:859 – 880, 2005.

[7] K. Lerman and L. Jones. Social browsing on flickr. *CoRR*, abs/cs/0612047, 2006.

[8] R. Maller and B. Neumann. Ontology-based reasoning techniques for multimedia interpretation and retrieval. In Springer, editor, *Semantic Multimedia and Ontologies*, pages 55–98. Springer London, 2008.

[9] B. Manjunath, P. Salembier, and T. Sikora. *Introduction to MPEG-7: Multimedia Content Description Interface*. 2002.

[10] C. Masolo and et al. The wonderweb library of foundational ontologies (wfol). Technical report, WonderWeb Deliverable 17, 2002.

[11] J. V. Ossenbruggen, F. Nack, and L. Hardman. That obscure object of desire: Multimedia metadata on the web, part 2. *IEEE Multimedia*, 12:54–63, 2005.

[12] G. Stamou and et al. Multimedia annotations on the semantic web. *Multimedia, IEEE*, 13:86 – 90, 2006.

[13] D. Trieschnigg and W. Kraaij. Tno hierarchical topic detection report at tdt 2004. *Proceedings of Corpus Linguistics 2005*, 99(7):1–8, 2004.