Ollscoil na Gaillimhe
University of Galway

Provided by the author(s) and University of Galway in accordance with publisher policies. Please cite the published version when available.

Title	A Social Context Enabled Framework for Semantic Enrichment and Integration of Web Videos
Author(s)	Choudhury, Smitashree
Publication Date	2012-09-03
Item record	http://hdl.handle.net/10379/3275

Downloaded 2024-04-16T18:24:21Z

Some rights reserved. For more information, please see the item record link above.

A Social Context Enabled Framework for Semantic Enrichment and Integration of Web Videos

Smitashree Choudhury

A dissertation submitted to the National University of Ireland, Galway for the degree of Doctor of Philosophy

August 2011

SUPERVISOR:

Dr. John G. Breslin

DIRECTOR OF RESEARCH:

Prof. Stefan Decker

Acknowledgement

It is with great pleasure and deep appreciation that I express my profound gratitude to my supervisor Dr. John Breslin for giving me the opportunity to carry out my research in one of the world's leading Semantic Web research institutes - the Digital Enterprise Research Institute (DERI) at National University of Ireland, Galway. His relentless support and constant guidance at every stage of my PhD helped immensely to raise the quality of research during my studies.

I am thankful to Dr. Alexandre Passant, Social Software Unit leader for his valuable guidance and collaboration during my research period. I would like to thanks Prof. Stefan Decker for his great contributions in making DERI a leading research institute in the area of Semantic Web. I am also thankful to Dr. Andreas Harth and Dr. Aidan Hogan for their valuable suggestions and support during my initial research days to bring clarity and depth to many Semantic Web related issues. I cannot forget to extend my thanks to Andrew and Gerard from DERI technical office for their relentless support and to DERI admin staff for their invaluable help and support throughout. Thanks also to all DERIans, for their great spirit of support, collaboration and for making the working environment so pleasant and exciting.

This thesis was supported by Science Foundation Ireland (SFI) under the DERI Lion project (SFI/02/CE1/1131) and Lion 2 (SFI/08/CE/I1380).

I would like to thank all my Indian and Pakistani friends for making my stay so enjoyable by serving excellent food periodically.

Last but not least, I would like to thank my family; this thesis comes to light today because I have been nurtured incessantly with moral support and encouragement from them.

Smitashree Choudhury

Abstract

The automatic inference of video semantics is an important but highly challenging problem whose solution can greatly contribute towards the annotation, retrieval, personalisation and reusability of video on the web. From a semantic annotation and retrieval perspective, this thesis investigates the influence of multiple video contexts for inferring video semantics, specifically aiming to improve video tagging and content description. The objective of the thesis is two- fold 1) formalising the representation of a video and its content via an ontological model and 2) inferring concepts to augment the model. First, a lightweight conceptual model of a video is proposed to describe a video object at four different structural abstractions (video, shot, frame, image region) and four different meta information categories (media, content feature, content semantics and context), and second, we investigate an ensemble of methods to infer video semantics from multiple contextual sources in order to augment the above model.

The study showed that contextual sources positively contribute in understanding the "aboutness" of a video and one can discover many descriptive concepts not originally described by the creator. Experimenting with different contextual sources showed that contexts contributed semantic enrichment is not restricted to document level video annotations, but can go further and be used to localise the detected entities inside the video timeline for a fine grained time-stamped annotation. In all studies we found that a combination of cues results in robust concept detection compared to the cues in isolation. We evaluated our approaches using both quantitative analysis and user based qualitative feedback. The principal benefits of the context based approach over a content based approach is that it is computationally inexpensive, maximises the wisdom of crowds and easily adaptable across domains.

Finally we built an integrated "Annotate, Search and Browse" prototype building over the proposed framework that supports complex structured queries, ontology based concept querying, temporal segment querying as well as the normal keyword search.

Table of Contents

Acknowledgement 1
Abstract 2
Table of Contents 3
List of Publications 8
List of Figures 9
List of Tables 11
Chapter 1 12
Thesis Overview 12
1.1 Motivation 12
1.2 Research Question 17
1.3 Thesis contribution 17
1.4 Thesis Scope 18
1.5 Definition of Terms 19
1.6 Thesis Structure 20
Chapter 2 21
Multimedia Annotation: State of the Art 21
2.1 Introduction 21
2.2 Content Based Media Annotation 22
2.2.1 Content Based Image Annotation 22
2.2.2 Content Based Video Annotation 25
2.2.3 Concept Detection Methodology 26
2.2.4 Challenges from Content Based Annotation 28
2.3 Multimedia Annotation from the Social Web 29
2.3.1 Social Sources of Multimedia Annotation 30
2.3.2 Image Annotation 30
2.3.3 Video Annotation 32
2.4 Conclusion 36
Chapter 3 37
Adding Formal Semantics to the Multimedia Web 37
3.1 Introduction 37
3.2 Semantic Web: Core Concepts 38

3.2.1 Resource Description Format (RDF) 39
3.2.2 RDF Schema (RDFS) 40
3.2.3 Web Ontology Language (OWL) 40
3.2.4 Querying the Semantic Web (SPARQL) 41
3.2.5 Realising Semantic Web Vision: Linked Open Data 42
3.3 Formalisation of Multimedia Object and Contents 43
3.3.1 MPEG-7 44
3.3.2 Multimedia Ontologies Based on MPEG-7 48
3.3.3 MPEG-7 Compliant Ontologies 50
3.3.4 Media Type Ontologies (Images, Audio) 51
3.3.5 Domain Ontologies 54
3.4 Use of Ontologies in the Multimedia Research Community 55
3.5 Rationale for our Proposed Model 56
3.6 Summary and Conclusion 57
Chapter 4 58
Social Semantic Web 58
4.1 Introduction 58
4.2 Social Media Sharing Sites 59
4.2.1 Flickr 59
4.2.2 YouTube 61
4.3 Tagging and Folksonomies 63
4.4 How People Tag? 65
4.5 Why People Tag? 66
4.6 Challenges for Tagging 67
4.7 Tag Semantics 68
4.7.1 Network Representation of Folksonomies for Tag Semantics 70
4.7.2 Semantic Representations of Tag and Tagging 71
4.8 Summary and Conclusion 77
Chapter 5 79
VOW: Lightweight Conceptual Video Model 79
5.1 Introduction 79
5.2 Motivation 80
5.3 Motivating Example 82
5.4.1 Fragment Identification 85
5.5 Related Studies 86

5.6 Abstract Conceptualisation 86
5.7 Description of Reused Schemas and Classes 88
5.7.1 FOAF Ontology 88
5.7.2 Geo Ontology 88
5.7.4 SIOC Ontology 89
5.7.5 SKOS Ontology 89
5.7.6 W3C Media Ontology 89
5.7.7 Event Model 89
5.8 Core Conceptualization 91
5.8.1 Content Space 92
5.8.2 Context Space 93
5.8.3 Important properties 94
5.10 Requirement Satisfaction Revisited 99
Chapter 6 101
Visual Concept Learning Based on Social Knowledge 101
6.1 Introduction 101
6.1.1 Motivation 102
6.2 Tag Learning Approach 102
6.2.1 Visual Concept (Tag) Learning Framework (VCLF) 103
6.2.2 Tag Aggregation and Ranking 104
6.3 Experimental Setup 105
6.3.1 Training 106
6.3.2 Visual Feature Extraction 106
6.3.3 Visual Similarity Metrics 107
6.4 Result 108
Chapter 7 112
Enrichment, Ranking and Integration of Web Videos 112
7.1 Motivation 112
7.2 System Overview 114
7.2.1 Context Analysis and Tag Enrichment Module 115
7.2.2 Tag Space Enrichment 117
7.2.3 Video Tag Ranking 120
7.3 Semantic Integration 122
7.4 Experiment and Results 126
7.4.1 Data Collection 126

7.4.2 Experimental Setup 126
7.4.3 Results 127
Chapter 8 132
Localised Annotation of Event Video with User Tweets 132
8.1 Motivation 132
8.2 What is Twitter? 134
8.3 Linguistic Characterisation of Tweets 135
8.4 Methodology 136
8.4.1 Background Knowledge 137
8.4.2 Data Preparation 137
8.4.3 Ground Truth Annotation 138
8.4.4 Entity Classification 138
8.4.5 Micro-Event Detection 141
8.4.7 Score Identification 144
8.5 Topic to Timeline Alignment 144
8.5.1 Detecting Cues about the Beginning of an Event 145
8.5.2 Detecting Cues about the End of an Event 145
8.6 Evaluation and Results 146
8.7 Related Studies 149
8.8 Conclusion 149
Chapter 9 151
Annotate, Search and Browse Prototype 151
9.1 Introduction 151
9.2 Search Task 152
9.2.1 Keyword Search 152
9.2.2 Structured Query 154
9.4 Result Visualisation 156
9.4.1 Result List 156
9.4.2 Facet Clustering 156
9.5 Tag Recommendation and Annotation Interface 158
9.7 Conclusion 161
Chapter 10 163
Conclusion and Future Work 163
10.1 Summary 163
10.2.5 Role of Social Tagging 167

10.3 Limitations of our work 169
10.4 Future Directions 170
10.4.1 Mapping and Integration of the Video Model to Other Existing Media Models 170
10.4.2 Context Evaluation 170
10.4.3 An Annotation and Video Browsing Tool 171
10.4.4 Social TV 171
Bibliography 172

List of Publications

- [Choudhury & Breslin 2008] Using a Lightweight Multimedia Content Model for Semantic Annotation. In Proceedings of the 8th Annual Conference on Information Technology and Telecommunication (IT&T 2008), Galway, Ireland 2008.
- [Choudhury & Breslin 2009] A Lightweight Web Video Model with Content and Context Descriptions for Integration with Linked Data. In Proceedings of the Semantic Authoring, Annotation and Knowledge Markup Workshop (SAAKM 2009)Workshop at 5th International Conference on Knowledge Capture (K-CAP 2009), Redondo Beach, California, USA, 2009
- [Choudhury, Breslin & Passant 2009] Enrichment and Ranking of the YouTube Tag Space and Integration with the Linked Data Cloud. In Proceedings of the 9th International Semantic Web Conference (ISWC2009), LNCS, Springer 2009.
- [Choudhury & Breslin 2010] Enriching Videos With Light Semantics. In The Fourth International Conference on Advances in Semantic Processing. SEMAPRO 2010
- [Choudhury & Breslin 2010] User Sentiment Detection: A YouTube Use Case. The 21st National Conference on Artificial Intelligence and Cognitive Science (AICS, 2010).
- [Choudhury & Breslin 2010] Focused Concept Detection and Polarity Measurement of Web Videos Using Social Context. In WebSci2010, Web Science Conference 2010.
- [Choudhury & Breslin 2011] Extracting Semantic Entities and Events from Sports Tweets. In Proceedings, 1st Workshop on Making Sense of Microposts, May 2011.

List of Figures

FIGURE 1: A SCREENSHOT FROM YOUTUBE FOR THE QUERY "THE TWILIGHT SAGA ECLIPSE TRAILER". 15
FIGURE 2: A RESULT SCREENSHOT FROM FLICKR FOR THE QUERY “TAJ MAHAL”. 16
FIGURE 3: VISUAL REPRESENTATION OF A VIDEO BEING ENRICHED FROM ITS MULTIPLE CONTEXTS 19
FIGURE 4: REPRESENTATION OF A RDF STATEMENT. 39
FIGURE 5: REPRESENTATION OF RDF TRIPLES. 39
FIGURE 6: THE LINKED DATA CLOUD AS OF SEPT.2010, GENERATED FROM METADATA EXTRACTED FROM THE COMPREHENSIVE KNOWLEDGE ARCHIVE NETWORK (CKAN). THE COLOUR CODING INDICATES DIFFERENT CATEGORIES OF DATASETS. 42
FIGURE 7: SHOWS AN EXAMPLE OF THE SEMANTIC ENRICHMENT OF A PHOTO WITH THE HELP OF LINKED DATA. .. 43
FIGURE 8: MPEG-7 DESCRIPTION SCHEMES. 46
FIGURE 9: A SCREENSHOT OF FLICKR PHOTO PAGE MARKED WITH PHOTO META INFORMATION. 61
FIGURE 10: A SCREENSHOT OF AN INDIVIDUAL VIDEO PAGE ON YOUTUBE WITH VARIOUS META INFORMATION. . 62
FIGURE 11: FLICKR TAGCLOUD. 66
FIGURE 12: CONCEPTUAL MODEL OF TAG ONTOLOGY. 73
FIGURE 13: MOAT DESIGN (IMAGE FROM HTTP://MOAT-PROJECT.ORG/ONTOLOGY). 74
FIGURE 14: CORE CONCEPTUAL MODEL OF SCOT ONTOLOGY (HTTP://WWW.SCOT-PROJECT.ORG/). 75
FIGURE 15: CONCEPTUAL MODEL OF COMMON TAG. 76
FIGURE 16 : SHOWS THE DISTRIBUTION OF MOST WIDELY USED VOCABULARIES IN LOD DATA CLOUD (FIGURE BORROWED FROM HTTP://WWW4.WIWISS.FU-BERLIN.DE/LODCLOUD/STATE/#ITEMS). 81
FIGURE 17: CORE CLASSES AND THEIR ATTRIBUTES. 88
FIGURE 18: EVENT MODEL. 91
FIGURE 19: CONTRIBUTING INFORMATION SPACES. 92
FIGURE 20: SHOT DETAILS. 93
FIGURE 21: IMAGE REGION DETAILS. 93
FIGURE 22: ANNOTATON INTERFACE 97
FIGURE 23: INTERFACE SHOWING THE TAGS SUGGESTED FROM MULTIPLE SOURCES. 97
FIGURE 24: INTERFACE SHOWING THE ANNOTATION FIELD VALUES ENTERED BY THE USER FOR A VIDEO 98
FIGURE 26: SEMANTIC SEARCH RESULT INTERFACE. 99
FIGURE 27: WORK MODULES FOR CLF 103
FIGURE 28: SHOWS THE PROGRESS (FROM LEFT COLUMN TO RIGHT) FROM AN INPUT IMAGE TO THE OUTPUT RECOMMENDED TAGS. 104
FIGURE 29: FEATURE VECTOR OF A TRAINING IMAGE. 107
FIGURE 30: RESULTS FOR TAG RANKING BY FREQUENCY VOTING. 110
FIGURE 31: RESULTS FOR TAG RANKING BY CO-OCCURRENCE. 111
FIGURE 32: WORK FLOW OF THE TAG ENRICHMENT, RANKING AND LINKING PROCESSES. 114
FIGURE 33: TYPES OF CONTEXTUAL SOURCES OF A VIDEO. 117
FIGURE 34: (A) TAG GRAPH FOR A VIDEO AND (B): SPREADING ACTIVATION FROM THE NODE “PLANES” 121
FIGURE 35: SHOWING A INTERLINKING OF DIFFERENT MULTIMEDIA DATA SOURCES IN THE WEB OF DATA. 124
FIGURE 36: VIDEO TAGS IDENTIFIED WITH DBPEDIA URI. 126
FIGURE 37: RESULTS FOR PRE AND POST ENRICHMENT PROCESS. 127
FIGURE 38: RESULTS FOR TAG RANKING (@N). 128
FIGURE 39: : INTER-RATER AGREEMENT OVER THE MOST RELEVANT TAG IN THE FIRST FOUR POSITIONS. 128
FIGURE 40: SYSTEM PERFORMANCE IN TERMS OF TIME TAKEN FOR RELEVANT TAG SUGGESTION FOR 100 QUERIES. 129
FIGURE 41: SHOWS THE OVERLAPPING BETWEEN VARIOUS CONTEXT SPACES FOR ONE VIDEO. 130

FIGURE 42: OVERVIEW OF VARIOUS STEPS FOLLOWED TO EXTRACT THE NAMED ENTITIES AND INTERESTING MICRO-EVENTS WITH THE HELP OF EXTERNAL DOMAIN KNOWLEDGE (FROM WIKIPEDIA AND GAME WEBSITE) AND LINGUISTIC PATTERNS. 137
FIGURE 43: TWEET VOLUMES OVER TIME INTERVALS. 143
FIGURE 44: TWEET AND RETWEET RATIO. 143
FIGURE 45: EXAMPLES OF TWEETS INDICATING HE END OF THE EVENT. 145
FIGURE 46: ENTITIES AND EVENTS EXTRACTED FORM USER TWEETS DURING THE IRELAND VS. ENGLAND MATCH , (BOLD WORDS ARE THE INTERESTING EVENTS OCCURRED DURING THAT TIME INTERVAL). 146
FIGURE 47: RECALL AND PRECISION OF THE PLAYER DETECTION CLASSIFIER. 147
FIGURE 48: EVENT DETECTION PERFORMANCE WITH LINGUISTIC FEATURES ONLY. 147
FIGURE 49: EVENT DETECTION PERFORMANCE WITH ALL FEATURES COMBINED. 147
FIGURE 50: INDIVIDUAL FEATURE PERFORMANCE IN PLAYER CLASSIFICATION. 148
FIGURE 51: (A) PLAYER DETECTION AND (B) EVENT DETECTION IN DATASET DIND. 148
FIGURE 52: SHOWS THE WORKFLOW OVERVIEW OF THE SYSTEM IMPLEMENTED. 152
FIGURE 53: SHOWS THE STEPS OF REFORMULATING A USER QUERY TO CAPTURE SOME OF THE QUERY SEMANTICS. 154
FIGURE 54: SCREENSHOT OF THE QUERY INTERFACE. 156
FIGURE 55: INTERFACE FOR RESULTS FROM A KEYWORD QUERY. 157
FIGURE 56: FACET FILTER SPACE. 157
FIGURE 57: DETAILED META DATA OF A VIDEO. 158
FIGURE 58: SCREENSHOT OF THE SOURCES OF TAG RECOMMENDATION FOR THE VIDEO. 159
FIGURE 59: SCREENSHOT FOR FINAL TAG RECOMMENDATION AND ANNOTATION INTERFACE. 160
FIGURE 60: SCREENSHOT OF THE VCLF WITH SUGGESTED TAGS FOR A KEYFRAME (IDENTIFIED WITH THE BLACK ARROW) 161

List of Tables

TABLE 1: A COMPARISON OF EXISTING MULTIMEDIA VOCABULARIES AND STANDARDS 82
TABLE 2: EXAMPLE SHOWING THE RESULT OF INTRA-LIST CO-OCCURRENCE TAG RANK 105
TABLE 3: LIST OF VISUAL CONCEPTS TAKEN FROM LSCOM AND NUMBER OF TRAINING IMAGES FROM FLICKR. 106
TABLE 4: VISUAL CONCEPTS AND THEIR RANK POSITIONS OBTAINED FROM TWO DIFFERENT METHODS ("V" BY VOTING AND "C" BY CONCEPTUAL CLUSTERING). 109
TABLE 5: FOUR COMPARATIVE TAG SPACES OF RELATED VIDEOS WITH AND WITHOUT FILTERS. 119
TABLE 6: SHOWS THE RESULT OF TAG EXPANSION FROM MULTIPLE SOURCES. 120
TABLE 7: EXAMPLES OF USER TWEET INCLUDING VARIOUS LINGUISTIC VARIATIONS OF WORDS. 136
TABLE 8: FEATURES RELATED TO A PLAYER. 139
TABLE 9: TWEETS WITH THE CONTEXT FEATURE. 140
TABLE 10: MENTIONS OF INTERESTING EVENTS DURING A MATCH. 142
TABLE 11: LINGUISTIC PATTERNS FOR DETECTING MATCH SCORES. 144

Chapter 1

Thesis Overview

1.1 Motivation

At the beginning of this century, due to the emergence of Web 2.0 an explosion of user-generated content, the web as a global information space became too large and started to push the boundaries of human comprehension. A simple query results in millions of documents, making it impossible for anyone to browse them all and select the most relevant ones [Denning, 2006]. Giant search engines and web based systems like Google; Yahoo etc. indexed billions of web pages but continued to struggle at organising and filtering the space in order to give a more user centric and personalised web experience. These search engines are highly efficient at retrieving documents containing some query keywords but fail to give an answer satisfactorily when the query needs an aggregated answer from multiple data sources or is constrained by multiple facets. This scenario becomes worse when we deal with multimedia data where textual descriptions are very sparse and search engines are not trained to analyse non-textual content. The foremost challenge in dealing with multimedia content is the problem of the "semantic gap", that is the difference between the user requirement and machine representation. Semantic gap is defined by Smeulders et al. [2000] *"the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation."*

Video is one of the most efficient and attractive options for capturing real life experiences and events. With the reduced cost for capturing, storing and sharing video, as a medium of communication it has become pervasive and ubiquitous encompassing every aspects of our social and personal life: entertainment, sports, education, politics and even personal life events such as birthdays, marriages etc. Media objects such as photo and videos are the fastest growing content segment on the web, courtesy of sites such as Facebook, Flickr, YouTube, MySpace etc. Gone are the days when creating a video was a time consuming task and needed some level of expertise in video capturing, editing and distributing. According to Manjunath et al. [2002] it has been never been so easy to create multimedia content as it is now. The main reasons are super cheap efficient digital cameras, camera enabled

mobile phones and now smart phones. Internet enabled mobile phones are fast becoming the most preferred communication devices, not limited to phone services alone but rather used to access the internet, to capture high definition photos and videos. Recording, editing and sharing a video on the web can now be accomplished with a single mobile device. Recent statistics give a glimpse of the amount of content we are dealing with:

- Video accounts for 60% of total web traffic volume. Users are uploading 50,000 hours of video every day to YouTube alone. Annually 70,000 terabytes of content are created by TVs and radio channels worldwide and much of it will eventually find its place on the web.

- Apart from the media hosting platforms like YouTube, Vimeo, there are live streaming services like Ustream, Justin tv, Livestream etc. creating huge amount of content on the web each day.

- A third category of content is now coming onto the web is traditional TV, major TV networks are either creating their online streaming facilities or content on demand leading to a huge upsurge of content.

- The popular photo sharing site Flickr claims to contain more than 4 billion photos with an average of 5,000 uploads per minute. Each month, 3.3 million geo tagged photos are uploaded by users to Flickr.

- Statistics from the most popular social networking site Facebook are equally startling. It contains 1.7 billion user photos with 2.2 billion friends are being tagged. The total photo storage in Facebook amounts to 160 terabytes and is increasing by 5 terabytes per month now.

Recently CISCO[1] predicted that in 2015, nearly 91% of internet content will be video. However availability of such content online does not guarantee content discoverability. The existing web infrastructure will soon be overwhelmed due to demands on these services as it was never meant to cope with the amount of content it is facing now. The volume of data poses serious challenges for the task of media retrieval, personalisation, reusability and semantic integration. A simple yet specific query "the twilight saga eclipse trailer" retrieves 5,280 results (Figure 1 (a)) with 20-30 % of redundancy, while the query for photos of "Taj Mahal" (Figure 2) fetches 195,590 photos from a single service (Flickr) alone. There are other similar yet isolated platforms carrying huge loads of content which cannot be shared and integrated with each other under the present web infrastructure. Intuitively, video retrieval will be more efficient if all the videos are annotated with keywords describing the depicted content. The load is not likely to ease unless the technology to structure, annotate, organize and search for media documents evolve as rapidly as the content creation

[1] Cisco Visual Networking Index: Forecast and Methodology, 2008-2013, available from http://www.cisco.com

technologies. The core challenge is to represent the videos in such a way that both humans and machines can understand and make use of the semantics.

The aim of this thesis is to investigate promising approaches for automatic video tag recommendation by utilizing different approaches from the information retrieval community. Performance of any multimedia retrieval engine depends on the efficient annotation of media content and its metadata. Today in 2011, though multimedia documents are more efficiently encoded and transmitted at the fingertip yet the task of media annotation does not get its due priority. Annotation largely remains creator's responsibility, either during production or post-production stage. It is the creator of the content, who decides the textual description of the media according to his understanding and publishes on the web, resulting in a very sparse description and sometimes no description at all. On the other hand content based annotation fared poorly despite of decades of mainstream research [Everingham, 2009]. This leads to the conclusion that there is an urgent need for methods and approaches to automate the media annotation process on the web scale.

To address the problem of information overload, interoperability and lack of semantic integration, the application of Semantic Web technology in describing media data is considered as an alternative approach. Semantic Web reflects a standard set of methods for data representation and sharing across heterogeneous sources. It can be applied to multimedia documents to describe and integrate media semantics into the web of data. Use of The Semantic Web will open up new opportunities in the multimedia processing applications.

The focus of this thesis is the semantic annotation of multimedia objects, especially video, so much of discussions and analysis will be within the framework of semantic descriptions an annotation of a video object. Automatic video annotation aims to create a computational model that will be able to assign terms to video resources describing the depicted content. It provides an alternative to expensive manual annotation. It is alternatively referred as auto tagging, auto captioning, tag propagation and tag suggestion, to name but a few. A semantic annotation is the augmented semantic layer of the media structure as well as its content using some ontological support.

Existing approaches in media annotation primarily focused on three broad schools of research: 1) Manual expert annotation and 2) Machine learning based concept and object detection and 3) Social tagging.

Commercial usage and general reusability of professional media content items can increase if annotated properly. For this reason professional producers and content creators used to hire and employ people to watch the media content items and annotate them with key terms. But this approach will not be scalable and will be extremely expensive when the numbers are in millions.

The second major approach for image and video annotation includes studies of object recognition, concept detection, and scene classification, etc. Researchers in this group (computer vision and speech recognition) used various supervised and unsupervised machine learning algorithms to build individual semantic concepts (car, explosion, policeman, etc.) models using low-level features from visual and audio signals. The learned concepts are then used for indexing and retrieval of the media documents. Decades of research resulted in much stable performance in areas such as scene detection, face detection etc. but less satisfactory in automatic segmentation and object detection. Moreover these studies were based on specific datasets such as COREL image sets [Duygulu et al, 2002] or broadcasting videos from TRECVID [Smeaton, 2005] created by professionals with a well-defined production style. As a consequence, their results cannot be directly translated into different domains such as user videos on YouTube and photos on the web [Yang and Hauptmann, 2008]. Moreover, the concept detection approach alone is simply not feasible at web scale where the number of concepts can be large and intra-concepts variations are difficult to predict.

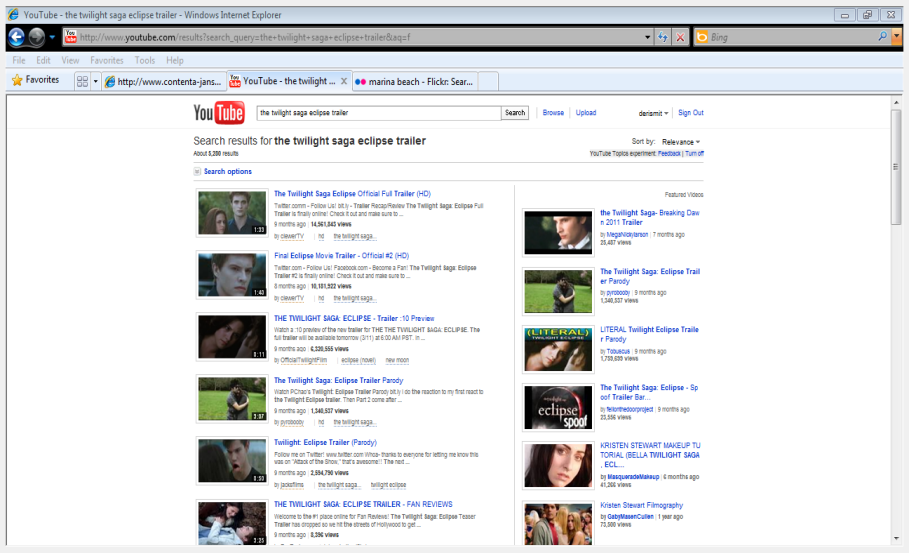


Figure 1: A screenshot from YouTube for the query "the twilight saga eclipse trailer".

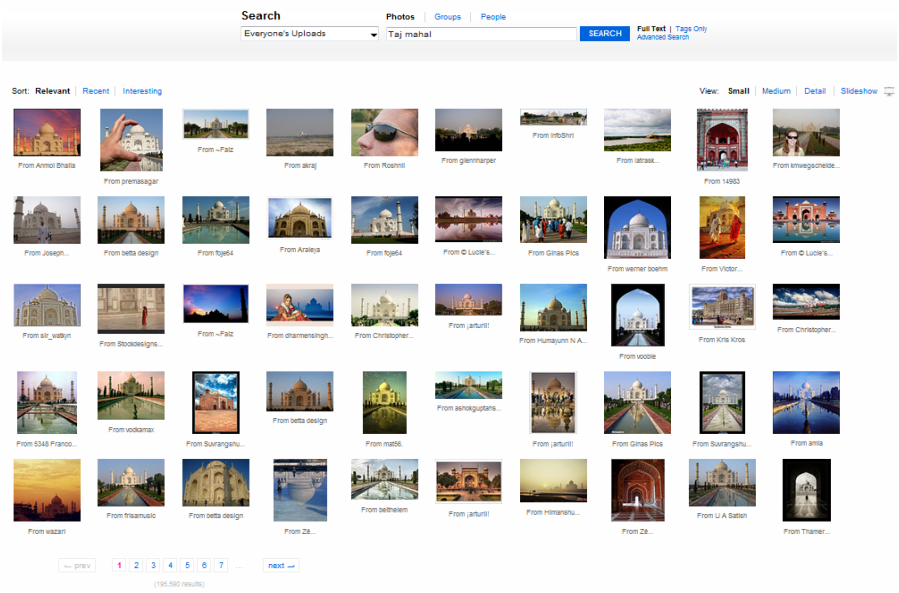


Figure 2: A result screenshot from Flickr for the query "Taj Mahal".

The third major approach is *collaborative and social tagging,* which gained popularity due to the rise in Web 2.0 and media sharing applications. Users of these sites are allowed to tag a media resource with free-form text labels to describe the content. Due to the low cognitive load on users, this approach became hugely popular. Collaborative tagging enabled millions of media objects to be tagged and described with little human effort compared to the previous model based approach. In spite of its popularity, tag based systems are not free from their flaws. Studies showed [Sigurbjörnsson and Zowl, 2008] 64% of Flickr images are tagged with 2 or 3 tags and 48% of tags are not part of the WordNet, making them less accessible. Among others, user tags are noisy, subjective and vague ultimately leading to poor retrieval performance. User tags are not formally structured & thus making semantic interoperability difficult.

The motivation for our research is based on existing gaps that are evident in all of the approaches described above. Automatic content annotation is the ideal solution but looks impractical at this moment. Instead we propose that automatic tag suggestion be based on various media contexts available on the web. Tags with the highest probability score, supported by multiple evidences, will increase the relevance and minimize the subjectivity score of the suggested tag.

This thesis is motivated and based on the following assumptions:

1. Existing Semantic gap between the information need and the discoverability.
2. Lack of semantic description of Web videos for an efficient discovery and content repurposing.

3. Potential benefits of Semantic Web technology for describing and integrating multimedia data to the Web of Data.
4. Complimentary contribution of the Social Web and the Semantic Web in creating a semantically enriched data space.

1.2 Research Question

In this thesis, we focus on semantic description,enrichmant and integration of video object on the web, staring from description to semi-automatic enrichment and integration. Thus, we address the following research question:

How can we semantically model a video and semi-automatically enrich the model from its various Social Web contexts such that it reflects video content description both at document as well as a the segment level?

The question can be broken down as follows:

- How can we formalise a video document reflecting all the fundamental constituent elements and functional aspects: We want to capture the semantics of a vieo document describing the media object and technical attributes, depicted semantic content reflecting the aboutness of the video and the low level temporal and spatial decomposition elements. We need description that fits for computational reasoning about finding videos and its segments based on semantic concepts. Therefore it leads to the next question is how to automatically augment the videos with semantic concepts/tags/entites ?

- How to bootstrap the automatic curation of the conceptual model from existing data? To what extent can the contextual sources contribute towards the understanding of web video objects and its semantic content?

1.3 Thesis contribution

Following are the contributions of this thesis:

- A lightweight conceptual model to describe user video.
- A systematic study of contribution of the Social Web and the Semantic Web in the context of multimedia data.
- Developed and evaluated innovative approaches for video annotation.
    - We present algorithms for suggesting new tags to describe video at the document level
    - Described a unique approaches ro annotae video keyframe
    - We proposed a unique method to annotate video segment with semantic wntities and events

- Finally built a prototype for the video tag suggestion and annotation

1.4 Thesis Scope

We are now in a position to describe the scope and various sub goals (illustrated in Figure 3) of our research in order to achieve the objective of semantic enrichment of the video content for reducing the semantic gap.

- Formal description of the video object: efforts to describe the video object/document with lightweight ontologies, yet covering all facets (content, structure, semantics, usage, etc.).
- Extraction of semantic entities and concepts: efforts to extract semantic entities depicted in the video from multiple contextual sources (visual representation of the scope is illustrated below).
- Relevant ranking of extracted entities: algorithm to filter and rank the suggested content in order to reduce the noise and subjective bias.
- Low computation and cognitive load: All of the above should be computationally feasible with minimum human intervention.
- Creating an integrated semantic space for the video: presents efforts to integrate the described video into a larger context in order to make it more accessible and part of the Semantic Web data cloud following LOD principles.

The above requirements focused our attention towards semantic description and annotation of a video object and leaves out other aspects of multimedia information processing such as multimedia indexing, retrieval etc. beyond our scope.

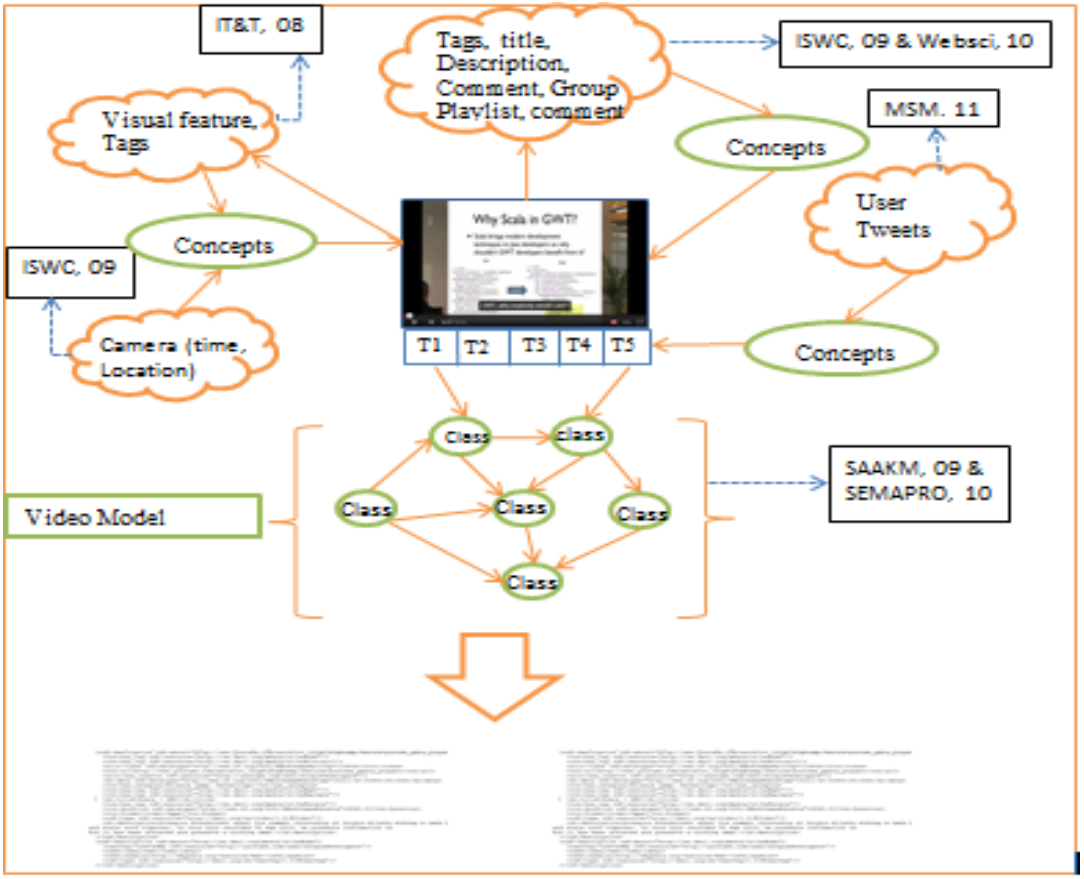


Figure 3: Visual representation of a video being enriched from its multiple contexts

1.5 Definition of Terms

A specific term whose meaning is open to interpretation needs to be defined in the proper context of its usage. Mostly we will describe the intended meaning of the terms during the course of this thesis as and when required, however some omnipresent fundamental concepts are described here to avoid any possible confusion.

Media object and *media document* are intended to be similar and are used interchangeably. The media object may be an image or video within the scope of this thesis. A media object is created by a web user and published on the web either in social network sites (Facebook) or media sharing sites such as YouTube and Flickr.

Context broadly refers to the "*situation or environmental information [Abowd et al., 1999] within which the object or event exists*". In the context of the present research a *media context* refers to the conditions and factors either created as part of the media creation process or factors emerged due to its post publication usage and interaction within a wider community.

The present thesis uses the term content semantics to describe and refer to the linguistic interpretation of video content in contrast to the widely used description of low-level feature descriptions by the computer vision and multimedia retrieval communities.

1.6 Thesis Structure

The next three Chapters (2-4) present three different yet related domains with a detailed review of related works encompassing the fields of multimedia annotation, Semantic Web and its application in multimedia, and social Semantic Web. Being concerned with video annotation, the research acknowledges the prior seminal works in the field of multimedia retrieval (image and video) from various research communities and does not discuss them in greater detail.

- Chapter 2 describes various existing approaches concerning multimedia annotation and retrieval.
- Chapter 3 describes the concept of the Semantic Web and its application in the multimedia domain.
- Chapter 4 describes the notion of Social Semantic Web which includes research attempts to formally describe the user-generated content with Semantic Web technology.
- Chapter 5 illustrates our core conceptual model for describing a video and its content.
- Chapter 6 describes a concept learning framework where tags are predicted based on visual similarity and tag co-occurrence for video keyframes.
- Chapter 7 outlines the social context based enrichment and ranking of video tag space and its integration to the web of data following linked data principles. We explored several ways to identify media content with concept URIs from ontological knowledgebase such as DBpedia[2].
- Chapter 8 outlines an approach to enable the time stamped annotation of sports videos using real time user-generated content from the micro-blogging site Twitter.
- Chapter 9 describes the implemented prototype with its annotation and search modules.
- Chapter 10 concludes this thesis with a discussion of various challenges faced as part of this research work and possible future directions.

[2] http://www.dbpedia.org

Chapter 2

Multimedia Annotation: State of the Art

The chapter discusses the state of the art research studies in the area of multimedia annotation in general and in video annotation in specific. It describes all major existing trends in the domain and compares and contrasts between them to contextualize the thesis motivation.

2.1 Introduction

Multimedia research covers a broad range of topics starting from media retreival, annotation categorisation and spaned across disciplines from computer science, computer vision, natural language processing,linguistics, signal processing, knowledge engineering alike. In this chapter will discuss some background studies in the area of video concept detection and annotation from the perspective of both content based annotation and contributions from the Social Web.

Among all the media types on the Web, video is the fastest growing in its content volume compared to other mediums such as text and audio. Every walks of human life is captured and shared on the web both by amateurs and professionals alike. While humans can intuitively comprehend the intended message, it is not so easy for machines to extract and make sense of the inherent semantics without some expensive content processing capabilities, and efficient reasoning capabilities.

Broadly speaking, interpreting semantics in images and video has remained the domain of the computer vision research community. Research efforts from this community tried to reduce the "semantic gap" by automatically detecting the semantic concepts from media content. In multimedia research, objects, scenes, or events define a controlled set of entities into which the pictures or video at hand can be classified [Gallagher et al., 2009]. An alternative paradigm of image annotation or image tagging allows image labels to be less restrictive, free-form, and unstructured, hence contributing towards a huge amount of user-driven content on the web. This content however come with few limitations, among them; the linguistic irregularities such as polysemy and synonymy are prominent [Cao et al., 2008]. A recently emerging and often co-discussed research issue is that of

multimedia retrieval since retrieval engines derive their bearings from the picture content itself [Arslan & Zimmermann, 2008]. Concept/event/scene detection, or annotation are all essential for filtering, organising, searching and browsing videos from personal as well as professional collections, hence research in multimedia retrieval is heavily tied to the understanding of media semantics [Tsikrika et al., 2009]. While a complete understanding of multimedia semantics is still an open problem [Pavlidis, 2009], meaningful semantic organisation, in its current scope, usually involves the task of categorising concepts based on their visual similarity and other low-level feature clustering.

Recently, an important trend has been noticed within the multimedia and computer vision research community, and thus increased emphasis on the modelling of contextual information [Tsai et al. 2005][Wolf and Bileschi, 2006] for improving the accuracy and coverage of concept detection. A few sources of contextual information are: (i) meta-data captured with pictures or videos [Toyama et al., 2003], (ii) correlations between spatial as well as temporal segments in multimedia data and (iii) patterns in multimedia collections as a whole [Cao et al., 2008], geospatial data for place and event semantics, embedded camera data for scene understanding etc.

2.2 Content Based Media Annotation

Towards the beginning of the last decade, the text search domain sufficiently matured compared to multimedia search. Most of the preliminary works in image and video retrieval were based on text content manually annotated by experts before the computer vision research started playing a major role. Soon image and video retrieval research crossed over into many disciplines i.e. pattern recognition, statistics, databases, signal processing, psychology, artificial intelligence, and was established as a mature research domain which included sub domains such as scene understanding, object detection, people detection, multimedia retrieval and indexing. Hundreds of research publications, tools and applications were built to address the issue of the "Semantic gap".

It is difficult and out of the scope of this thesis to discuss the entire research domain in sufficient detail. We will touch upon some of the representative works both from the image and video domain to get an overview of this area.

2.2.1 Content Based Image Annotation

Traditionally, content-based approaches for image annotation have attempted to bridge the semantic gap problem between low-level feature and high level semantic concepts building individual concept models. For an in-depth review of recent trends and issues in image annotation, readers are referred to the work of [Datta et al., 2008].

Recent trends in image annotation include: 1) concept recognition from visual features, 2) image correlation mining, 3) text and web mining 4) graph based techniques etc. The growth in image annotation research has been tremendous in recent years, not only in terms of the amount of publications and research output, but also in terms of alternative directions adopted to achieve the goal.

Image annotation and concept detection (as a related research domain of image retrieval) is closer to the stated goal of bridging the semantic gap than just providing a ranked list of documents. Annotation allows the image to be searched by text, so an automated concept detection and image annotation method with concept labels for large scale image repositories is considered one of the crucial steps towards an efficient real world image retrieval system. Image understanding is studied through concept detection, and annotation is a subset of concept detection [Datta et al., 2008]. The question is what content features (visual) best describe the concepts or help to identify the depicted concepts in an image.

Traditional image annotation has focused on a visual feature model where each concept is transformed into a model and then is used to detect the presence of the concept in unseen images. Basically, annotation is treated as a detection problem. One of the well-known and widely used benchmarks for this single concept model approach is the Caltech[3] visual concept collection which contains 101 semantic concepts to be detected in images. Various works in this direction can be discussed within the parameters of features used and adopted approaches.

2.2.1.1 Visual Features for Concept Detection

Most content based image retrieval (CBIR) system performs feature extraction as a precursor to concept detection, retrieval and annotation. Once the features are obtained, they are used as the input for subsequent image processing. Research efforts started with identifying global visual features to create concept models. Global features are low dimensional, quick to extract and hence easy to compute similarity but the concept detection performance is expectedly low in the absence of any spatial information.

The second category of features used to build concept models are region based features from image blocks and regions. It involves segmentation of an image and extraction of the segment features such as edge, contour, blobs for identifying concepts in the image. It improves annotation performance at the cost of feature dimension. Moreover it also depends on segmentation efficiency which is still an open research problem.

[3] Caltech dataset available: http://www.vision.caltech.edu/Image_Datasets/Caltech101/

Finally the third feature category used is local features (region based and point based) such as region patches and interesting local points in the image. The most well-known local features are the bag-of-visual-words, Scale Invariant Feature Transform (SIFT) Implementation [Lowe, 2004], SURF [Bay et al., 2008] etc. For similarity calculation, keypoints are clustered and used for matching. Local features are high dimensional in nature and slow in computation but excel in object and concept detection. Recently studies preferred a combination of global and local features to produce a trade-off between performance and computation. Concept detection through model-based supervised classification of single static concepts such as airplane, ship, car, and landscape, sunset has achieved high accuracy in [Vailaya, 2001]. However, only a few prototypes perform this task online and at large scale, e.g. one of such example is the ALIPR server [Li and Wang, 2006] which recommends multiple tags based on the query image similarity.

When visual features are used as the primary source of concept modelling, the annotation process becomes computationally expensive as there is a need for hundreds of concepts and each image has to pass through all the concept models to be detected. Moreover, the detection result became sensitive to feature selection. This led to a new correlation based approache for large scale concept learning.

2.2.1.2 Relation Mining Approaches (Word and Context Relation)

Following the challenges of similarity based only on visual features, researchers focused on exploiting various existing spatial and visual correlations within the image. Visual correlations are mined to identify and reinforce concepts which are positively correlated and highly unrelated concepts are discarded. For example, an image detected with the concept “boat” will reinforce the concept “lake” while discouraging negatively correlated concept. These studies mostly attempted multi label classification as reported in [Kang et al., 2006] who exploited the annotation correlation to propagate concepts; Chen studied multiple instance classification [Chen, 2004] for the categorisation of images into semantic classes. Zhou and Zhang [2006] present two methods for what they term “multi-instance multi-label” learning, where a bag of instances is used to find a set of labels. While one method adds weak classifiers to determine classifications, the other method first performs fuzzy clustering on the instances, and then does constructive clustering while preserving structure information in order to label the image. Parikh et al. [2008] use a segmentation-based approach followed by modelling of each segment’s context, using relative location, scale, and co-occurrence in order to assign multiple labels to the image. Li and Wang [2003] used a two-dimensional multi-resolution Hidden Markov Model for automatic annotation of pictures with words. Many image related research efforts borrowed techniques from the text domain. Duygulu et al. [2002] adopted a word to image translation version in line with the text machine translation. The impression that image annotation is still an open challenge

is constantly reinforced by with ever increasing research efforts in this direction. The association between pixel values and text words has remained elusive despite the factoring in of multiple conditions.

2.2.2 Content Based Video Annotation

Semantic concept detection within a video stream is a precursor for video annotation and therefore for video retrieval. The task of concept detection is to model and predict the presence of one or multiple semantic concepts within the video based on some probabilities and existing knowledge of the concept. A detailed survey of the topic can be found in [Snoek and Worring, 2009]. One of the major research efforts in concept detection is the TRECVID workshop where research groups share their results and experience in detecting concepts from predefined video datasets. Since its beginning in 2001, TRECVID has attracted more than 300 different studies to detect concepts from video using multimodal features. The process pipeline of concept detection involves stages [Snoek and Worring, 2009]:

1. Pre-processing
2. Feature extraction
3. Classification
4. Fusion
5. Modelling correlation between semantic concepts

2.2.2.1 Pre-Processing

This step involves video segmentation into shots (an atomic unit of a video document). Shots can be of fixed temporal length or variable length. Each shot is represented by single or multiple keyframes as a result of an uninterrupted camera action [Davenport et al., 1991]. Shot segmentation is mostly done automatically by comparing successive frames within a shot. Each shot is represented by one or more keyframes and each keyframe is represented with an *n* dimensional feature vector.

2.2.2.2 Feature Extraction

Performance of concept learning heavily depends on feature selection. The features of a shot are intended to be a compact description of a shot that includes visual, audio and text features. Visual features include colour, texture and motion feature vectors at both global and local levels (patch, grid based), bags of visual words and interest point detection. Visual features need to capture the visual variations in the appearance of concepts caused due to editing and lighting conditions (rotation, scale, etc.). These differences are called the sensory gap: the disconnect between the real word representation and the digital representation of a semantic concept [Smeulder et al., 200[b]]. This gap

can be handled with the selection of invariant features [Snoek and Worring, 2009] e.g. SIFT and SURF.

2.2.2.3 Classification

The extracted feature vectors are inputs for concept classification. One of the most popular approaches for this concept prediction is based on Support Vector Machines [Snoek & Worring, 2009], where SVM acts as a binary classifier and used for individual concepts. Some of the other preferred methods are neural networks, generative learning, active learning and other supervised learning algorithms.

2.2.2.4 Fusion

Due to uncertainties in classification performance, many studies opt for multiple feature spaces to represent different attributes of a visual image such as colour, shape, motion, and they later fuse these features to detect the concept more robustly. Multiple features can be extracted from a single mode such as only visual or only audio features, or it may be multimodal such as combination of visual, text and audio features. However, selection of these features and how to fuse them is not trivial. Variations in the mode as well as the values demand feature normalisations. Feature-level fusion may improve the concept detection performance but increases the dimensionality and training time.

Feature-level fusion allows a concatenation of the feature vector before classification whereas the fusion of classifiers takes place at the end of individual classification by multiple classifiers. Kittler et al. advocated classifier fusion for accuracy and efficiency. Similar to feature combinations, combining classifiers is not easy and needs careful consideration of questions such as whether the classifiers should be from a single mode of data or from multiple modes and whether classifiers should be based on the same data or on different training datasets etc. Once the classifiers predict the probabilities, the final probability score is computed either using some weighted or average function. Diversity at multiple levels for example feature selection, learning models, parameter tuning makes the fusion approach relatively expensive for concept detection and hence should be adopted with caution. A comparative study showed that classifier fusion gives better result compared to the feature fusion [Snoek et al., 2005].

2.2.3 Concept Detection Methodology

Concept detection in video mostly starts with supervised learning where labelled data is used to train a concept model and untested data is used for prediction of the label with a probability score $P(w_i,s_i)$.

This section will briefly discuss various popular approaches used for video concept detection. One of the popular approaches is Support Vector Machines [Schölkopf and Smola, 2001].

2.2.3.1 Supervised Learning

The majority of concept detection studies focused on a type of machine learning where the label is predicted based on the pre-trained data, called supervised learning. The goal is to get maximum generalisation with minimum training samples and the performance of the model is tested with the data not used for training. Supervised learning needs to strike a balance between poor generalisability due to the large feature space and the issue of "over fitting" due to strict optimisation of the parameters [Jain et al., 2000]. Under such restrictions, Support Vector Machines (SVM) emerged as the popular choice of method for most of the concept detection studies. SVM generates a hyperplane between data points separating two different concept classes. A TRECVid study from university of Amsterdam showed that the SVM based classification performed better in identifying objects and locations but not for persons and character queries[4].

2.2.3.2 Other Approaches

Other methods used for concept detection are neural networks, Gaussian mixture models and Hidden Markov Model (HMM). HMM is a popular technique applied to sequential data for pattern recognition [Rabiner 1989]. It is used to model and explain the present state based on the past sequence. The same model is used to predict the next state in the sequence. Most notable applications of HMM are speech tagging, Part of Speech tagging etc. Yang and Hauptman [2008] exploited the temporal relationship of a concept within or among shots is used to predict its probability. Ebadollahi et al. [2006] used HMM to capture the temporal relationship between static concepts to infer presence of dynamic concepts such as "plane taking off". Authors in [Snoek and Worring, 2005] represent temporal relations between concept detectors using Allen's temporal intervals [Allen, 1983] such as "before", "meets". These concepts and the relations are subsequently used to infer event type concepts such as "goals" in a soccer match. HMM based video analysis and classification reported in [Dimitrova et al., 2000], who proposed to use HMM along with text and face features. Informedia [G.C. Michael,2007] conducted an LDA based multi-modal retrieval by creating a generative topic model to describe the joint distribution of text and visual features in concept detection. Most of the concept detection studies in TRECVID differ in feature space and the selection of the statistical model.

[4] http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html

2.2.3.3 Use of Ontologies

The latest approach in video annotation is the role of ontologies such as WordNet[5], Cyc[6] and ConceptNet[7] in detecting and predicting visual concepts. These linguistics ontologies were extended or connected to visual features to describe concepts and their relationships as well as to facilitate reasoning. Hooges et al. [2003] extended WordNet with limited visual features describing shape, geometry and scenes of the video. Bertini et al. [2005] developed an ontology to describe videos in the Soccer domain. Snoek et al. [2007] manually connected hundreds of concepts to the WordNet synset based on concept similarity. Similarly, Naphade et al. [2006] described their research effort to link LSCOM concepts to the Cyc knowledgebase. Applications of ontologies in all of the above studies mainly focused to increase the accuracy of concept detection [Zha et al., 2007] rather than reasoning, for the prime reason that ontologies are symbolic facts whereas concept detection process carries uncertainties. Zha et al, built an ontology from labelled data and exploited their hierarchical relationships to refine the concept detection further.

2.2.4 Challenges from Content Based Annotation

Unlike text documents, multimedia annotation is a challenging task for several reasons. Traditionally, text; as a human creation, used symbols for interpretation while visual images are the representation of a human's perception of the world which is difficult to represent and describe explicitly. It is still not clear how a human perceives and interprets a visual object. Yet, during the last two decades many ambitious attempts were made to mimic the human vision system and to teach a machine to recognise visual objects with considerable success.

Major search engines and real world applications are yet to adopt these methods as solution. Still state of the art computer vision techniques are far from satisfactory and give only 20%-30% accuracy when tested in a controlled and clean object oriented data sets variations [Everingham et al., 2009]. Some of the challenges faced are:

- Identifying correct feature sets and combinations of features to represent.
- Difficulty in defining models.
- Cross domain application of model.

From the above discussion, it is evident that many learning methods exist for generic as well as specific concept detection with their own advantages and limitations. The selection and combination of methods depends on multiple factors, for example selection of features, fusion techniques, data

[5] http://wordnet.princeton.edu/
[6] http://www.cyc.com/

size, domain of data, training amount, role of context, etc. In most of the cases the role of machine learning especially supervised learning, has been the dominant method of choice. Though the list of concepts used for detection has been extended from 100 concepts to a list of 1000 (LSCOM) concepts, the performance is far from satisfaction. Poor performance of concept detection is also evident from many studies reported in TRECVID 2010, where they reported poor performance of concept detection approach compared to the text based approach which included information sources such as metadata, automatic speech recognition (ASR), external knowledge sources such as Wikipedia and web pages[8]

2.3 Multimedia Annotation from the Social Web

After decades of research, results from content based approaches could not meet the objective of bridging the semantic gap, and moreover, the cost of individual concept models and feature processing proved to be computationally expensive, giving rise to the need for alternative approaches. Challenges from content based media processing combined with the emergence of social and collaborative tagging induced a paradigm shift in the semantic interpretation of multimedia content. User-generated content in combination with multivariate contexts offered a powerful alternative to expert annotation. Novel algorithms were developed in combination with media content and context features to facilitate multimedia processing which includes various sub domains such as categorisation, concept detection, indexing and content retrieval. Figure 3 shows schematic representation of different semantic categories and entities that can be extracted from user-generated content.

While automatic image tagging has received greater attention from both the research and tool developer's communities, very little work has been done on the analysis of the web's video clips, and automatic video annotation and video tag suggestion has remained one of the least explored research areas. Existing content-based concept detection algorithms are not expected to perform well for web videos, for the simple reason that these techniques are heavily dependent on training data and feature selection. Moreover, they are trained on certain domain data produced under professional production environment such as news, documentaries etc. The above assumption was proved correct by Manor et al. [2008] who studied applications of existing content based algorithms in a web video dataset. They attributed the poor performance to poor quality and the immense subject variations of web video. While dealing with web based multimedia content, we commonly find depicted semantic categories such as people, locations, events and concepts answering the core questions of "who, "where", "what"

[7] http://csc.media.mit.edu/conceptnet
[8] http://www-nlpir.nist.gov/projects/tvpubs/tv10.slides/tv10.kis.slides.pdf

and “when”. Many existing related studies have focused on extracting those semantic categories (Figure 3).

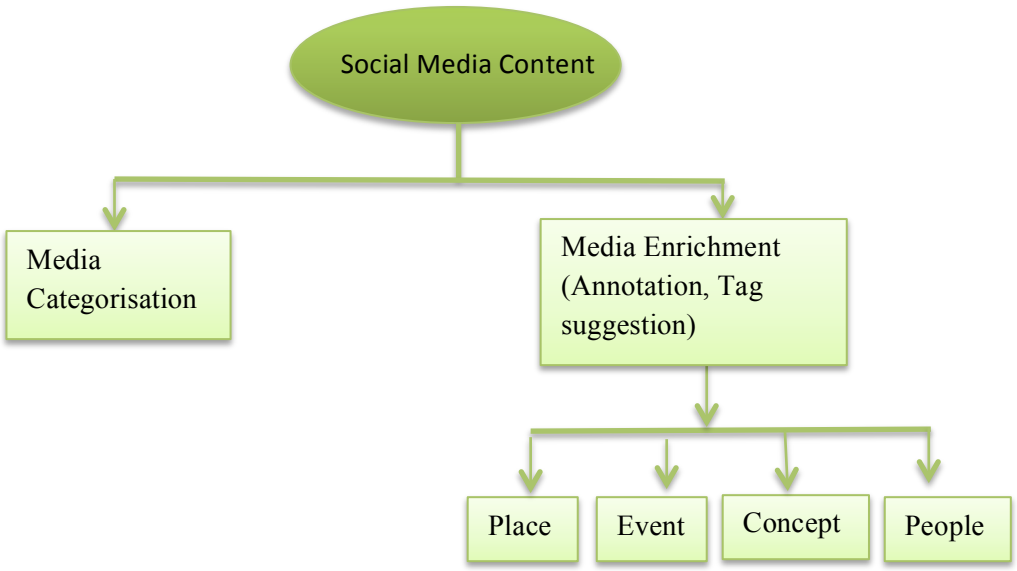


Figure 3: Schematic view of different semantic categories found in multimedia content (both image and videos)

2.3.1 Social Sources of Multimedia Annotation

Social sources of multimedia annotation can be classified into three different categories: 1) distributed web based manual annotation as in the Mechanical Turk [7] and Label Me [Russell et al., 2008] where web users manually annotate resources with labels, 2) annotation through collaborative games, e.g. ESP games [1] and 3) user-generated content in media sharing sites. The sheer volume of annotation and labelling generated by such collective means helps to reduce manual efforts considerably for preparing training data. With some instruction and inter-annotator validation, these data can directly be used for various tasks such as concept modelling and object detection. The next sections will describe the contributions of social web in both image and video annotation.

2.3.2 Image Annotation

Studies related to image concept detection or tag suggestion, explored both supervised (model-based) and semi-supervised (model free) approaches. Supervised or model- based approaches are similar to traditional concept learning studies with the exception of the training source. Instead of creating manual training examples, social media contents were used as the training source [Fergus et al.,2005 and Kennedy et al.,2006]. Despite the high noise level in user-generated content, it proved attractive due to its sheer size, frequency and low computational cost. On the other side, model-free semi-

automatic approaches used statistical techniques such as co-occurrence analysis, neighbour voting and clustering techniques for image tag recommendation and tag ranking [Li and Lu, 2008].

Model-free techniques need basic information such as tags or camera information in order to proceed further in suggesting relevant tags. One such study was reported in [Sigurbjörnsson and Zwol, 2008] where the system required an initial set of tags to perform co-occurrence statistics in order to suggest more tags relevant to the image, and tag relevance was judged using various strategies including majority voting. Authors were assumed to have a set of initial tags and all these tags are supposed to be relevant to the image. A weighted combination of global and personal tag co-occurrence used to annotate Flickr images in [Garg & Webber, 2008]. Weinberger et al. [2008] used a probabilistic framework to model tag co-occurrence and identified tags that disambiguate other tags in temporal, spatial and semantic contexts. Wang et al. [2007] selected candidate tags for recommendation based on the co-occurrences of the existing items. Suggested tags were ranked based on visual similarity using Markov Model.
The second category of method used for image tag suggestion is the nearest neighbour method which identifies visually similar images to propagate tags to the unlabelled images. Visually similar images are presumed to be semantically similar. Tag Suggestr [Kucuktunc et al., 2008] used the initial tags of a user and the visual neighbours to suggest more tags.

Image annotation research is increasingly focusing on exploiting geospatial information created by users and cameras while creating a photo. A number of research studies used the camera generated metadata for image classification and tag suggestion [Cristani et al., 2008]. The ZoneTag tool also used GPS locations to identify related tags from a user's social network [Ahern et al., 2006] and his or her past tagging history to recommend tags.

Event detection in image and video is an interesting research topic for the multimedia community. The event detection process goes beyond object detection by incorporating the temporal aspects of the media into the supervised model. Recent studies of event and scene detection within images used geographical contexts creatively.

Naaman et al in [2004] described one of the earliest work “PhotoCompas” to organise personal photo collections using geographical co-ordinates and time. Similarly, Rattenbury et al. [2007] analysed Flickr tags to extract “place” and “event” semantics. They used the burst features of spatial and temporal data to detect a specific event in a location. Epshtein et al. [2007] used location and orientation information from multiple cameras to detect various geographical objects. According to

Cao et al. [2008] personal geotagged photo collections are small, but collectively they can represent and reflect the correlations between location and time of individual photos.

2.3.3 Video Annotation

Web-based image content has already been studied quite intensively. When it comes to web video content, fewer contributions can be found. Manor et al. [2008] and Schindler et al. [2008] have presented studies on shot boundary detection and categorisation of web video content, and emphasized the difficulty of the domain due to enormous content variance and weakness of user labels. On a positive note, video sharing sites allow its users to upload and annotate their videos with title, tags, descriptions, to categorise under certain pre-defined categories, to interact through comments or ratings and also to create community structures based on theme and interests. All this user-created content may help to classify and categorise video content further. These individual efforts adds to the understanding of the video But as of now, few studies have explored user-generated content in video classification and even fewer to enriching the sparse video tag space. For a segment level annotation, image tagging is relatively easy, as users can draw a bounding box to localize a region of the image and label the selected area with keywords, but for videos, they have to tag a sequence of images. In other words they have to watch the entire video, perform a series of pause-play sessions to tag each frame individually, which discourages users from doing any deep annotation of a video clip and they usually have to settle on few global level tags without any reference to their position in the video timeline. We will discuss some of the studies related to web videos in general and automatic video tagging efforts in specific. The concept of web video and Video on Demand (VOD) was on the web before YouTube, but was mostly confined to dedicated streaming media servers, so many studies relating to the semantic content of user videos were carried out once YouTube became popular but while its videos were still few in number. These studies can be classified under three categories 1) web video statistics, viewing patterns and usage 2) user community and social network, and 3) video content semantics. Content semantics includes applications such as video categorisation, retrieval and video annotation/tag prediction.

2.3.3.1 Web Video Characterisation

During the initial years of YouTube, the major share of research studies concerning user-generated web videos were focused on studying the sharing platform, viewing patterns and video characteristics. Halvey and Keane [2007] explored the social dynamics of the service and found that a majority of users only consume the media, and save videos for later viewing, while a few users had frequent activities. They also found that the social features such as groups, or friendship were being under-

utilised by users. This result was not unexpected as online video sharing was a relatively new concept and YouTube was slowly gaining ground during that period. One such study by Geisler & Burns [2007] reported that a video carries an average number of six tags; 66% of tags do not appear in the title or description which implies the fact that users use the tagging space as extra placeholder for descriptions. Cha et al. [2007] studied large scale video data from YouTube and other video on demand systems to gain an insight into the systems, UGC patterns, the video life cycle, video popularity distribution and other statistical properties. They also studied illegal content uploading in these sites which is more prevalent in popular videos, thereby making the video ranking difficult. Cheng et al..[2008] analysed YouTube characteristics and found significant differences between YouTube and other traditional media sites in terms of usage patterns, video length, popularity, etc. They also observed small world characteristics between related videos indicating a strong correlation to each other.

2.3.3.2 User Community and Social Network

Most media sharing sites owe their popularity to their integrated social networking features where users not only upload but share and form different communities based on their mutual interest. Study of individuals and various interest-based communities they are part of, can reveal not only interesting behavioural patterns but also details on the content evolution. Lai and Wang [2010] studied the impact of external links on video popularity and found that videos from the news, science, and Entertainment categories were most viewed through external links. Gargi et al. [2011] studied the YouTube video graph in order to detect community structures which can be used for community specific topic discovery. The detected communities (clusters) were named based on the titles of the clustered videos. YouTube user behaviours and their interaction patterns were studied in [Benevento et al., 2008] to identify anti-social behaviour, spamming, promotion, etc.

2.3.3.2 Video Categorisation and Tag Recommendation

Research works related to video semantics fall under two groups; video categorisation and tag recommendation.

Currently most of the video sharing sites provide a set of pre-defined categories for user videos. This, leads to inconsistency in category labelling, as different users have different understanding about the content category, and secondly there may be other categories the user want to use for labels but that are not part of the list. Automatic video categorisation can address these issues and dynamically adopt the video categorisation problem based on the video content and the creators input.

Traditionally, genre classification aimed to categorise the videos into different genres such as “movie” and “news. Web video categorisation aims for similar goal: to categorise videos into some specified concepts. However, the dynamics of web video is worth mentioning: first, web videos have high diversity in terms of subject, format, style, and quality compared to professional videos and second, the category specification need to be more flexible, one video can belong to several categories. According to Yang et al. [2007], compared to video genre categorisation, a web video category ontology concerns not only the “genre”, but also the content, such as the categories “animal” and “auto”. Categorisation is different from tag recommendation; categorisation reflects the entire document content unlike tag lists which can be localised related to parts of the video. Yang et al. [2007] studied web video categorisation using multi modal features including textual sources such as title, tag and description, audio features, visual features and concept histograms. They used SVM, Gaussian mixture model and semi-supervised manifold ranking method to categorise web videos. SVM proved to be the best classifier of all. A consensus learning approach for multi label video categorisation was reported in [Chandrasekar et al., 2009]. The authors used multimodal features such as video, audio and user-generated content to build multiple classifiers to predict categories for a test video. The experiment was performed with YouTube videos but they did not report the type of categories they were predicting. Sharma and Elidrisi [2008] claimed 65% classification accuracy for YouTube videos based on user tags.

A concept based video retrieval engine, “TubeTagger” is reported in [Ulges et al., 2009] where semantic concepts from the LSCOM ontology are learned by making use of YouTube clips as a training source. They explored the tag associations to enrich their concept retrieval model. Their study found that some of the concepts cannot be learned properly due to the lack of proper examples. Our assumption is that a pre-defined fixed concept ontology is not the answer to deal with dynamic user-generated content. It should be more of a bottom-up approach to learn and adapt concepts according to the user’s information need and tagging patterns. Work by [Leung et al., 2009] reported on commentary-based video categorisation and concept discovery. They used user comments to generate a list of categories for the music video domain and compared this against tag based categorisation. This study claims to discover user centric categories for the video and various relationships among singers but does not elaborate further on a formal grounding of those findings. Another study exploiting video context is reported by Wu et al. [2009] to categorise videos. The information sources they used are text, related videos and user’s past history. User-generated content results in lot of redundancies e.g. people upload multiple versions of videos resulting in duplicate and near duplicate copies, Study in [Wu et al., 2007] showed that between 15-27% videos have near duplicate videos. Siersdorfer et al. [2009] studied video tag suggestion from near duplicate videos. This approach is more beneficial more in the case of popular videos where multiple users upload videos expressing

different personal perspectives, for example it is difficult to find duplicate videos in the categories of "people and blog" or "science and technology", whereas "music", "entertainment" and "sports" categories may reap the benefits of near duplicate videos. Vallet et al. [2010] studied the impact of external knowledge sources such as DBpedia, Flickr and Google Images to retrieve visual query examples to facilitate video retrieval. They used the underlying semantics of these knowledge sources to reduce the query ambiguity and subsequently increase the retrieval performance. Their findings suggest that structured knowledge sources such as DBpedia improve performance significantly compared to unstructured sources such as Google Images. Video categorisation based on social interaction has been studied by Yew et al. [2011], who considered social interaction surrounding the video as more relevant indicators of the video category compared to the media specific metadata. Wang et al. [2010] suggested YouTube video categorisation using information from related videos, external sources such as web pages and Google search videos. As part of their study they created classifiers for 29 frequent categories and tested this with 80,000 YouTube videos for categorisation. Their results suggested that information from web pages are more relevant and trustworthy compared to related videos or search videos alone. However, combinations of related videos yield best results than any other combinations. Aradhye et al. [2009] reported automatic tag recommendation of YouTube videos based on the audio-visual content of the query video. They learned mapping between audio-visual content and user supplied-tags to predict the labels of unseen videos. Context information from capturing device such as camera is used in image classification, but recently a paper by Arsla et al. [2008] employed spatial and temporal video properties for ranking and search. They used cameras, equipped with GPS sensors and 3D digital compasses to capture and model the "fields of view (FOVs)" relative to a location. An interactive search by a user specifying regions would activate the retrieval process which includes (i) spatio-temporal filtering to get relevant videos, and (ii) ranking videos based on the overlap of query region and FOV.

All the above studies somehow deal with the tag recommendation or categorisation for the entire video not necessarily at the segment level. A few studies were reported about localised and deep annotation of the video. Deep annotation or segment based annotation is required especially for long videos. It will not only assist with easy navigation but it also makes it easy to reuse and share segments. Though shot level annotation of videos is very challenging and needs shot boundary detection before some content similarity computation at video frame level, there are some creative use of social media data to bootstrap the process of fine grained time-stamped annotation. This thesis presents one such approach of annotating event videos using user tweets (described in one of the core chapter).

In one recent study, Ballan et al. [2010] presented a video tag suggestion method and temporal localisation of tags based on frame similarity. This approach is in line with our earlier work, where we

identified tags for video frames following the visual similarity of frames with Flickr tags. Shamma et al., [2009] studied user tweets during the US election campaign to see if tweets can exhibit the structure of broadcast media. Shi et al. [2011] described video segmentation and event detection from synchronous comments based on content similarity and the user's social relationship. These studies are related to one of our studies described in this thesis where we took different domains of video and their corresponding tweets to extract significant entities, topics and micro-events, and aligned them to their corresponding video timeline for temporally localised video annotation.

2.3.4 Video Annotation and Crowdsourcing

Image annotation through crowdsourcing enabled vision community to create massive dataset like Label me, ImageNet, TinyImages etc at a low cost, but that does not hold true for the video dataset because of its dynamic temporal characterstics [Sorokin and Forsyth]. Few works are in the public domain in this direction.Vondrick et.al.[2010] made an attempt to learn best practices for video annotation through using MTurk. They released Vatic (Video Annotation Tool from Irvine, California), an open platform for monetised crowdsource video labelling. Yuen et.al [2009] proposed a web based tool called LabelMe Video to annotate object category, shape, motion and activity information in real-world videos. Dutch Institute for Sound and Vision launched 'Waisda"[9] a video labeling gaming platform to annotate videos with tags.

The above discussion shows that tagging and user-generated metadata has been used in much of the image annotation related research, but not studied well the video domain.

2.4 Conclusion

In this chapter we discussed present trends and challenges of multimedia content annotation from the perspective of both content based approaches and the contribution of the Social Web. We also discussed emerging trends in this research domain due to Web 2.0 factors and user-generated content. Literature in the area shows that image annotation has attracted more attention compared to the video annotation, while interpreting video semantics remain an elusive task. The next chapter will discuss various research efforts in the direction of adding formal semantics to multimedia data.

[9] http://imagesforthefuture.com/en/research/waisda-video-labeling-game-evaluation-report

Chapter 3

Adding Formal Semantics to the Multimedia Web

Focus of this chapter lies in the area of Semantic Web and its use in multimedia data. It discusses various standards and vocabularies to describe multimedia data and documents. It also illustrates some of the non-Semantic Web efforts to describe multimedia semantics.

3.1 Introduction

Explosive growth of media content on the web needs an urgent attention for efficient organisation, storage and retrieval mechanism. Media sharing and social networking sites claim that billions of pieces of media content are being shared and consumed daily. Facebook claims billions of photo views per day while Flickr is assumed to carry nearly 5 billion user photos, the Photobucket photo sharing site claims to have 8,390,190,317 photos and videos as of (April 2011), while YouTube claims to have 50 hours of video uploaded per minute. The existing amount of multimedia content has surpassed the limit for which the present web infrastructure was originally intended and as a result, access to and efficient discovery of this content itself is fast becoming a serious challenge. Unlike text documents, media objects are complex and opaque. Signal level content based search is computationally expensive and seems impractical in the present scenario when the numbers are in their billions. The recent trend of collaborative tagging is one popular approach for annotating media objects, but this shallow annotation does not address the issue of data interoperability across applications. Interoperable content description is key for practical applications dealing with content access, retrieval and sharing.

A number of vocabularies have been proposed reflecting different aspects of a media object ranging from media technicalities, semantic subject, usage and copyright [Boll et al., 2007]. Driven by semantic and interoperability needs, ISO proposed the Multimedia Description Interface, popularly known as MPEG-7 [Chang et al., 2001; Manjunath et al., 2002], specifically aimed for audio-visual documents. MPEG-7 provides an extensive set of description tools for audio-visual content description at various levels including aspects such as semantic content, usage, navigation and content organisation. In spite of its extensive and detailed coverage of multimedia content description, the

adoption of MPEG-7 is not wide spread due to many practical obstacles as documented in [2004] by Ossenbruggen et al. Among them two critical points are the issue of semantic ambiguity due to the syntactic and semantic multiplicity and the use of XML schema as the language of description.

Recently, the Semantic Web is being considered as an alternative approach to make media content (structural and semantic) machine understandable. Towards this goal, the Semantic Web community has initiated a number of efforts for making semantics explicit and for promoting data interoperability using languages such as RDFS and OWL. In an attempt to strike a balance between MPEG-7's flexibility and explicit semantics, researchers have proposed various ontologies based on MPEG-7. In the rest of the sections of this chapter we will discuss some of these initiatives as well as related approaches concerned with different domains.

Before delving into the details of various multimedia standards and ontologies we will give a brief overview of the Semantic Web and related technologies in the next section.

3.2 Semantic Web: Core Concepts

The original idea of the Semantic Web is to enrich the semi-structured and unstructured information on the web with machine understandable descriptions so that the machine can intelligently analyse, discover and reason with the help of background knowledge. The challenge is to facilitate the interoperability of knowledge across data sources and applications. For the Semantic Web, ontologies and ontological languages play a key enabling role. An ontology is described as an "*explicit, formal specification of a shared conceptualisation of a domain of interest*". [Gruber, 93]. Two fundamental characteristics of an ontology are worth mentioning; 1) formal and 2) shared.

> *Formal:* Ontologies are expressed in logic based formal languages so that discovery and analysis can be facilitated through reasoning and rules.
>
> *Shared:* Ontologies are rightly described as shared conceptualisations where social consensus is required for the usage of domain concepts and their relationships, so that the resulting metadata can be interpreted unambiguously.

Wishing for a standard universal ontology in any domain is a difficult task for the simple reason that the meanings of a concept vary according to many socio-cultural factors. There may be many existing ontologies describing similar and often the same domains of knowledge. Instead of attempting to merge everyone under one scaffold, a more desirable solution is to develop rules for the mapping of ontologies. The Semantic Web was first envisioned "*as an extension of the current web where content would be given well defined meanings (metadata) to enable machines to understand and process the content intelligently*" [Berners-Lee, 2001]. Later it included the concept of the "Web of Data" where

web resources are interlinked to each other following some fundamental principles called “Linked Data Principles” [Berners-Lee, 2006]. We will discuss the web of data and linked data later in the chapter. The following section will briefly discuss various core concepts of the Semantic Web.

3.2.1 Resource Description Format (RDF)

The W3C proposed RDF as one of the foundational technologies of the Semantic Web. It is a data model for asserting facts about worldly things in the form of a triple. A triple is a simple assertion in subject-predicate-object format. The triple in (Figure 4) shows an RDF statement asserting that “Video1” has a creator named “John Breslin”, where “Video1” is the subject, “hasCreator” is the predicate and “John_Breslin” is the object. This is the simplistic form of an RDF statement that asserts a fact about a video.

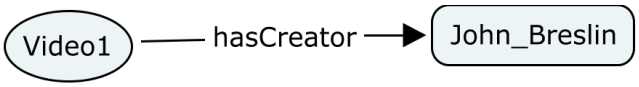


Figure 4: Representation of a RDF statement.

In practice, RDF requires that we use unique identifiers (URI) to describe subjects and predicates while the object may use a URI or any other RDF specified data types.

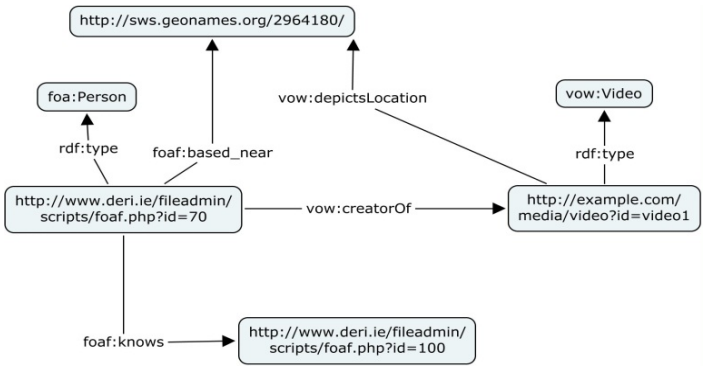


Figure 5: Representation of RDF triples.

In the Figure 5, four nodes are connected with four different labelled arcs (named properties). Nodes represent resources (instance from a class) and arcs are attributes of resources (Video and Person). As per the requirement of RDF, a URI has to be unique to identify any particular resource irrespective of its physical or online location. The use of URI allows statements from different data sources to be interlinked in order to create a graph of statements.

RDF is a schema-less and self-describing graph data model used to assert facts. In other words, domain concepts and their relationships are encoded in a formal language so that a machine can understand it. RDF can be serialised in several formats such as RDF/XML [Beckett, 2004], N3

[Berners-Lee, 2006], N-Triples [Grant and Beckett, 2004] or Turtle [Beckett, 2007]. RDF/XML provides a mechanism to abbreviate long resource URIs with a common namespaces as reusable prefixes. This abbreviation is a syntactical matter and does not influence the abstract RDF model. An RDF statement in itself carries little semantics unless backed by a vocabulary that specifies the meaning of terms. The standard available vocabularies are RDF Schema (RDFS) [Brickley and Guha, 2004] and Web Ontology Language (OWL) [Dean and Schreiber ,2004].

3.2.2 RDF Schema (RDFS)

RDF Schema (RDFS) is a basic ontology language with limited expressivity. It defines terms, their intended use and their semantics. It also describes classes and their hierarchy structure. RDFS defines class properties and property hierarchies, along with their domain and range restrictions. For instance the property "vow:depictsPeople" is a sub property of the "vow:depicts" property. With the help of RDF Schema, a statement like "Video1 vow:depictsPeople JohnBreslin" can be further clarified to say that the property "vow:depictsPeople" is applicable to the Video domain and its values should come from a foaf:Person class. Then we can safely infer that "John_Breslin" is a person and "Video1" is an instance of a video class even though it is not mentioned explicitly. Finally RDFS defines some annotation properties such as rdfs:label and rdfs:comment to describe the resource. Well known schemas developed with RDFS are Dublin Core Terms[10] and Friend-Of-A-Friend [Dan Brickley and Libby Miller 2005]. RDFS is lightweight but lacks many fundamental semantic relationships such as transitivity, symmetry and equivalence, which can be expressed with a more complex yet expressive language called the Web Ontology Language (OWL).

3.2.3 Web Ontology Language (OWL)

The Web Ontology Language (OWL) is another W3C recommended language for ontology representation. It is designed to be used by an application that needs to process the information content rather than presenting it to users. It facilitates greater machine interpretability than XML, RDF and RDFS with increased vocabularies and formal semantics. These vocabularies include semantics for disjoint relationships, cardinality, equality relations, and consistency checking to name a few. OWL has three sublanguages differing in its degree of expressiveness and complexities.

- OWL Lite provides fewer constraints and considered easy to get more tool support due to its low formal complexities.
- OWL DL is based on the foundations of description logic. This language is in between OWL Lite and OWL Full. It provides much more expressiveness yet guarantees computational completeness. Most existing ontologies fall under this category.

- OWL Full is the language with maximum expressiveness with no restrictions on language format, but it guarantees no computation completeness. This makes it hard for tools to be OWL Full complaint.

The choice between OWL DL and OWL Lite depends on the amount of expressiveness the user is required to implement. Recently OWL has graduated into a second version called OWL 2.0 with added flexibility in terms of properties and relationship.

3.2.4 Querying the Semantic Web (SPARQL)

SPARQL is a powerful query language for RDF. There had been many proposals of query languages for RDF data, e.g. RDQL, SeRQL and SPARQL. The W3C Data Access Working Group (DAWG) recommended SPARQL (SPARQL Protocol and RDF Query Language) as the query language for RDF in 2008. It is an SQL-like language for querying RDF graphs via pattern matching. SPARQL can be used to pose queries across diverse data sources, whether the data is stored natively as RDF or viewed as RDF via middleware [Prud'hommeaux and Seaborne, 2008]. It supports patterns such as conjunction, optional, union and value filters. Most of SPARQL queries contain a triple pattern called "basic graph pattern" with variables in place of the subject, object and predicate. It also supports group graph patterns. The example below shows a simple query "find me all the titles of videos uploaded by user John" and its SPARQL representation with the variable "?title" in the object place.

```
PREFIX dc:  <http://purl.org/dc/elements/1.1/>.
PREFIX sioc: <http://rdfs.org/sioc/ns#>

PREFIX rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX vow:  <http://sw.deri.org/media/video/ns#>
SELECT ?title
WHERE{
?s rdf:type vow:Video.
?s dc:title ?title.
?s sioc:has_creator ?user.?user sioc:name 'John'.}
```

SPARQL can match multiple graph patterns and match literals with language. It supports query types such as ASK, DESCRIBE, CONSTRUCT and SELECT. SPARQL 1.0 lacks many basic operations such as aggregation, which were later incorporated in SPARQL 1.1[11] which will soon be a W3C recommendation. Thus SPARQL can process many complex structured queries which are not easily possible in legacy SQL queries.

[10] http://purl.org/terms
[11] http://www.w3.org/TR/sparql11-query/

3.2.5 Realising Semantic Web Vision: Linked Open Data

The idea of the Semantic Web evolved to include the concept of data interlinking and integration as one of the fundamental requirements to realise the vision. In 2006 Tim Berners-Lee proposed the idea of "linked data" as a way forward. In his seminal work, "Design Issues of Linked Data" [Berners-Lee, 2006] he proposed four basic principles to be followed in order to expose and interlink data from various sources:

1. Use URIs as name/the identifier of things (people, object, event, location, etc.)
2. Use http URIs so that it can be located on the web for further relevant information about the resource.
3. Publish the information in a standard format such as RDF
4. Link to other URIs with the resource description so that the chain continues.

Since 2007 the Open Linked Data initiative grew quickly from few datasets to a cloud of more than 200 datasets from multiple domains. It now claims to contain more than 25 billion triples interconnected to each other via 350 million links (Figure 6). The domain of data includes: geographic data, movies (IMDB), life sciences, music (DBTune), e-commerce (Best Buy), and Wikipedia data to name a few. This data cloud uses DBpedia as a central hub which is created out of the structured information present in Wikipedia.

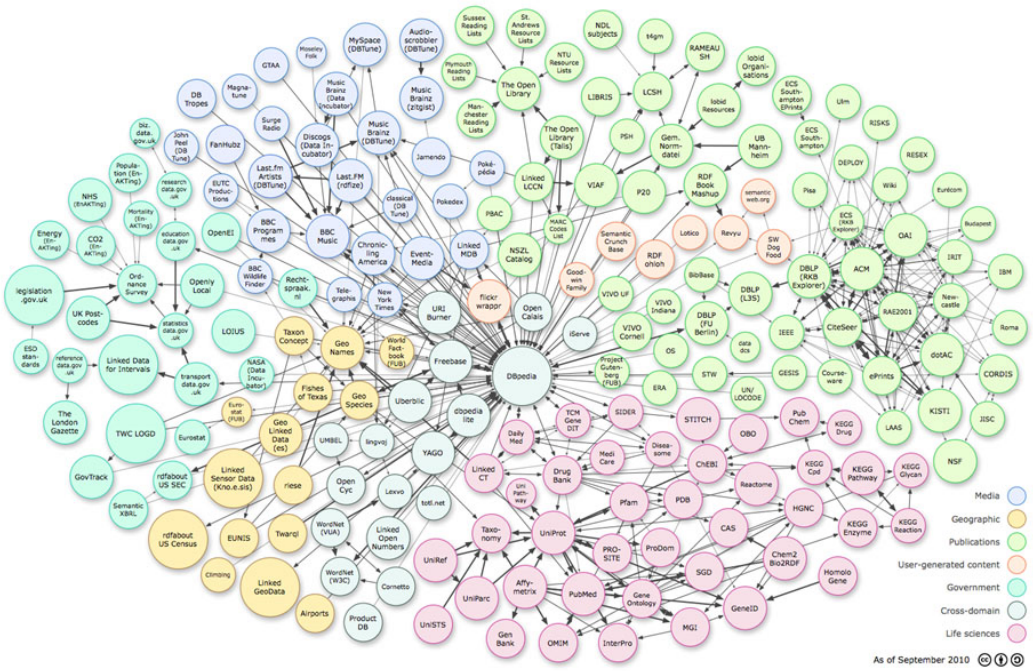


Figure 6: The Linked Data Cloud as of Sept.2010, generated from metadata extracted from the Comprehensive Knowledge Archive Network (CKAN). The colour coding indicates different categories of datasets.

The benefits of this interlinked data cloud are that, it enable complex queries to be answered by traversing semantically rich heterogeneous datasets. To give an example; if a photo (Figure 7) is tagged with a keyword "Taj Mahal", we can use any semantic search service and map the tag to the relevant concept from DBpedia, whose URI is http://DBpedia.org/resource/Taj_Mahal. As per linked data principles, this URI describes relevant information about the concept. If we explore the generic ontological properties of this concept such as rdfs:label, dc:subject, owl:sameAs, redirection, disambiguation, we can get helpful information such as category, type of the concept, geo location of the concept (Taj Mahal) which are not mentioned in the original photo page. Further domain specific exploration will suggest knowledge about its creation, history, etc. A second level link traversal of the page will give ideas about similar and related concepts such as other "Mughal Architecture" and "Islamic Architecture".

Original Tag space

Taj mahal

Enriched Space

DBpedia **URI**
http://DBpedia.org/page/Taj_Mahal

dc:subject
Mughal_architecture (category)
Islamic_architecture (category)
owl:sameAs
freebase:Taj Mahal

http://sws.geonames.org/1255231/

Figure 7: Shows an example of the semantic enrichment of a photo with the help of linked data**.**

In this introductory section we described briefly the concept of the Semantic Web and its various core concepts such as RDF, RDFS, SPARQL and Linked Open Data. In the next section we will discuss in detail the major efforts made by researchers to model multimedia objects and content information using both pre - Semantic Web and the Semantic Web ontologies.

3.3 Formalisation of Multimedia Object and Contents

With the growth of multimedia data both in the user's personal archives and professional digital libraries, content management and access have posed greater challenges making the need for semantic content descriptions essential. Machine readable metadata envisioned as one of the core prerequisite

for intelligent multimedia applications. It goes beyond the idea of facilitating of mere data exchange to content access, sharing and reuse. This led to various initiatives for developing a standard to describe a multimedia object and its semantic content. These efforts can be broadly categorised into two groups; 1) MPEG-7 based ontologies, and 2) MPEG-7 compliance ontologies. We discuss them along with their modelling rationales and approach to domain linking. In general, a multimedia ontology should include descriptions of the media and its composition structure, the semantic content it depicts, the creator and creation situation, usage and privacy, but the primary concern should be to provide linkage from low-level content to high level semantics: bridging the semantic gap. Before going into a detailed discussion of these research efforts we will briefly describe the MPEG-7 and its various tools followed by various formalizations as follows:

- MPEG-7,
- Ontologies based on MPEG-7.
- MPEG-7 complaint ontologies.
- Ontologies for media types.
- Domain ontologies.

Finally, we will give the rational for proposing another framework.

3.3.1 MPEG-7

ISO developed a multimedia content description interface known as MPEG-7. It provides a comprehensive set of tools to describe audio-visual content. It is meant to address a wide variety of media types including: still pictures, graphics, 3D models, audio, speech, video, and combinations of these (e.g., multimedia presentations). Since the objective is to make the standard available for wide range of applications, the specification includes multiple flexible description formats for any single attribute. The main elements of MPEG-7 are as follows:

Description Definition Language (DDL): An extension of XML schema to create new Descriptors and Description Schemes. The MPEG-7 specification opted for XML schema as the preferred language to describe the various syntactic and structural constraints of multimedia data, its multiple and complex data types. The set of MPEG-7 XML Schemas defines 1182 elements, 417 attributes and 377 complex types.

Descriptors are tools, designed to describe low-level audio and visual features such as colour, texture, motion, audio, as well as specific attributes of the audio-visual content such as location, time, quality, and so forth. These features are expected to be extracted automatically with existing algorithms.

Visual Descriptor is a set of descriptors about the colour, shape, texture information in an image or video frame, used for categorisation, filtering and retrieval of images.

Colour Descriptor is one of the most common visual features used by various applications for content level image processing and retrieval. There are six colour descriptors representing different aspects including colour distribution, spatial structure and layout.

Shape Descriptor of an object provides a powerful semantic clue about the presence of an object in the image and is used for similarity matching. Region-based and contour based shapes are used in many applications including handwritten character recognition.

Motion Descriptor describes the object motion and camera motion in a sequence of images in video. Various motion descriptors are: Motion Activity Descriptor, Motion Trajectory Descriptor, Camera Motion Descriptor, and Parametric Motion Descriptor. Motion descriptor is used for many real world applications such as gait detection, activity and event detection in video.

Texture Descriptor refers to the visual pattern of an image region and is described with three descriptors: Region-based Shape Descriptor (RSD), Texture Browsing Descriptor (TBD) and Edge Histogram Descriptor (EHD).

Audio Features Descriptors[12] are divided into generic low-level features and application specific tools. There are 17 temporal and spectral descriptors under the following groups: Basic, Basic Spectral, Signal Parameters, Temporal Timbral, Spectral Timbral and Spectral Basis Representation.

Description Schemes (DS): In MPEG-7 Part 5, the Multimedia Description Scheme (MDS) is described which comprise of a set of Description Tools (Descriptors and Description Schemes) dealing with multimedia entities. It describes the semantics and the relationship between descriptors and their schemes. They are used whenever more than one medium needs to be described (e.g. audio and video.) These Description Tools (Figure 8) can be grouped into five different classes according to their functionality:

- Content description and management: a representation of perceivable information, information about the media features, the creation and the usage of the audio-visual content;
- Content organisation: a representation of the analysis and classification of several audio-visual contents;
- Navigation and access: specification of summaries and variations of the audio-visual content;

[12] http://mpeg.chiariglione.org/standards/mpeg-7/mpeg-7.htm

- User interaction: description of user preferences and usage history pertaining to the consumption of the multimedia material.
- System Tools: A set of tools for the efficient retrieval, transmission and management of the descriptions.

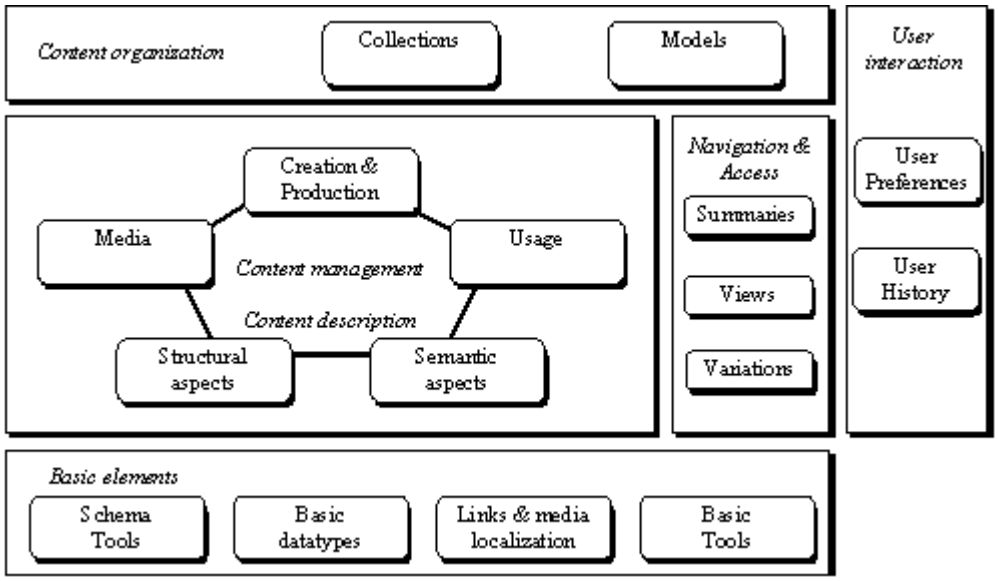


Figure 8: MPEG-7 Description Schemes.[13]

Content Structure

The core element of the content structure tool is the Segment DS. It addresses the description of the physical and logical aspects of audio-visual content. The content structure tool describes the structure of a audio-visual item in terms of segments (video segments, frames, still and moving regions and audio segments). A segment represents an atomic part of an audio-visual item which can be used as an object for content based analysis, processing and querying. It has nine sub classes: Multimedia Segment DS, audio-visual Region DS, Audio-visual Segment DS, Audio Segment DS, and Still Region DS, Still Region 3D DS, Moving Region DS, Video Segment DS and Ink Segment DS. Therefore, the classes may have both spatial and temporal properties.

Semantic Description

The MPEG-7 **Semantic DS** tool describes the multimedia content in terms of event, objects, time and agents. Events can be perceived as a situation occurring at a certain time, involving objects, people and places as the participating elements. These entities can have interrelated properties and states

[13] http://mpeg.chiariglione.org/standards/mpeg-7/mpeg-7.htm#2.9_MPEG-7_Profiles

among themselves. In MPEG-7 the participants, background, context, and all the other information that makes up a single narrative are referred to as a "narrative world"[14]. The "Semantic DS" is derived from the "SemanticBag DS", which is an abstract class representing any kind of collection of semantic entities and their relationships. Some specialised concepts are the "Object", "AgentObject", "Event", "SemanticPlace", "SemanticTime", "SemanticState" and "Concept DS".

3.3.1.1 Limitations of MPEG-7

MPEG-7 is the most comprehensive standard for multimedia content description with hundreds of elements and attributes, making it complicated and in practice, difficult to implement in many real world applications. Besides its complexities, MPEG-7 suffers from some major obstacles in terms of semantic data interoperability and integration [. Dasiopoulou et al.,, 2009].

1. XML Schema provides little support for expressing semantics and leads to ambiguous content description.
2. MPEG-7 does not incorporate formal semantics, and as a result, interoperability remains an issue both at the syntactic and semantic level.

Increasingly it was realised that in order to reduce the ambiguities and achieve semantic interoperability across applications and data sources, the semantics of MPEG-7 need to be explicitly formalised using some standard representation language. XMLSchema is lacking in this regard as the scope for explicit semantics is restricted whereas the Semantic Web standards such as RDFs and OWL fills the gap.

With the vision of making multimedia semantics explicit, the Semantic Web community started a number of research initiatives to formalise MPEG-7 in an ontological representation. The resulting multimedia standards and specifications can be broadly categorised either as generic or domain dependent ontologies. Some of the generic ontologies are either more semantically structured and complex, conforming to the MPEG-7 specification or lightweight predicate based vocabularies (Dublin Core, W3C Ontology for Media Resource[15]). We will describe some of them below, starting with MPEG-7 based ontologies, followed by lightweight ontologies, and finally some domain ontologies.

[14]http://mpeg.chiariglione.org/standards/mpeg-7/mpeg-7.htm
[15] http://www.w3.org/TR/mediaont-10/

3.3.2 Multimedia Ontologies Based on MPEG-7

To address the interoperability issues, research efforts have been started to translate MPEG-7 into an ontological representation but care was taken to preserve MPEG-7's original schema and flexibilities. Two such prominent methodologies include the proposals by Hunter et al. [2001] and Tsinaraki et al. [2004]. Besides these two approaches (which aim to provide a framework for interoperable MPEG-7 compliant multimedia metadata), there are some other research attempts to describe MPEG-7 partly or fully, will also be discussed in this chapter.

3.3.2.1 Harmony Project

One of the initial initiatives to formalise the MPEG-7 standard started within the Harmony project by Hunter [2001]. The RDF Schema (RDFS)-based ontology, which later was adopted in DAML and finally converted to OWL, was proposed to formalise the structural and localisation tools of the MPEG-7 MDS, as well as the visual description part. It also includes description for production, creation, media information and usage. In the process of translation it preserves the MPEG-7 schema.

Different MPEG-7 segments are modelled as specialised classes based on the content types. In order to preserve the flexibility of the MPEG-7 schema, segment types are treated as multimedia content types too, allowing multiple semantic interpretations of a defined semantic entity. As a result, the ambiguities present in MPEG-7 are propagated, incurring serious implications on the conceptual clarity and subsequent management of the produced descriptions [Dasiopoulou et al., 2009]. Hunter's MPEG-7 ontology has been utilised for semantic analysis and annotation of fuel cell and pancreatic cell images [Hollink et al., 2005]. Linking to a domain ontology is done through the ABC ontology, which is an upper-level ontology describing abstract concepts (time, place, event), developed within the Harmony international digital library project.

3.3.2.2 AceMedia Ontology

AceMedia[16] project proposed two ontologies to describe media content structures and low-level visual features. The ontology for the media content structure is called the "Multimedia Structure Ontology (MSO)" and the ontology for visual features is called the "Visual Description Ontology (VDO)", which is reused in our proposed model to describe the low-level features of various segments.

The class hierarchy in MSO is similar to the approach adopted in the ABC Harmony project except for a few exceptions. The concept "MultimediaConent" is the superclass with Audio, Image, Video, Multimedia, "AudioVideo" as its specialised classes. The second major concept, Segment, is a

[16] http://www.acemedia.org

subclass of the class "MultimediaContent" and specialises into different segment classes based on time and space decomposition and content types. Linking to a domain ontology is achieved through upper ontologies like DOLCE[17] and Annotation Ontology developed within the project. It also follows the same pattern as of the Harmony project (keeping the MPEG-7 flexibilities), it therefore inherits similar semantic ambiguities due to the multiple interpretations. It has been used in image and video analysis as well for annotation processes.

3.3.2.3 VDO

VDO is the Visual Descriptor Ontology developed within the AceMedia project describing MPEG-7 visual parts (descriptors). It consists of a set of visual descriptors conforming to the specification of MPEG 7 in order to represent image and video data. This model consists of four upper level concepts (VisualDescriptor, MetaConcept, Feature and Region). VisualDescriptor specialises several descriptors such as colour, texture, shape and motion descriptions, while the "Region" class links to the visual descriptor through the "hasDescriptor" property. The concept "Feature" represents various parameters used for extracting the descriptions. This ontology contains 61 classes and 237 properties.

3.3.2.4 The SmartWeb Ontology

The SmartWeb[18] project developed an MPEG-7 based ontology [Vembue et al., 2006] to support annotation of multimedia content based on the SmartSUMO upper ontology, which is a combination of DOLCE and SUMO[19] ontology. It includes structural, localisation, media and low-level feature descriptions. They treat "MultimediaContent" and "Segment" as two different classes, where the decomposition of a segment is carried out based on spatial/temporal aspects and content type. This ontology was used in the soccer domain for structural and low-level feature descriptions.

3.3.2.5 The Boemie MPEG-7 based ontology

In the context of the Boimie project, two multimedia ontologies were designed to capture MPEG-7 semantics (both structure and audio-visual descriptions). These two ontologies are: 1) Multimedia Content Ontology (MCO) and 2) Multimedia Descriptor Ontology (MDO). In order to capture the semantics they re-engineered the structural and localisation descriptions by introducing some distinct classes. MCO consists of two major classes: 1) MultimediaContentType (Audio, Video, Image, etc.) and 2) SegmentType (VideoSegment, AudioSegment, StillRegion, etc.).

[17] http://www.loa-cnr.it/DOLCE.html
[18] http://www.smartweb-projekt.de/start_en.html
[19] http://www.ontologyportal.org/

MDO is the ontology used to describe the low-level content features of MPEG-7. Linking to domain knowledge is realised through a connection between the content segment and a domain concept.

3.3.2.6 MPEG-7 Rhizomik

The Rhizomik ontology attempted a fully automatic translation of the complete MPEG-7 Schema to OWL [García and Celma, 2005]. The translation was based on a generic XML Schema to OWL mapping, called XSD2OWL, in combination with XML to RDF (XML2RDF) mapping. The end result is an OWL DL ontology covering the entire MPEG-7 specification. It strictly preserves MPEG-7 semantics through its class hierarchies as a result inheriting most of the semantic ambiguities as well. Linking to a domain ontology is part of the ontology, as it translates the semantic description tools as well. However it requires some complicated and tedious mapping for domain linking [Dasiopoulou et al., 2009]. This ontology is used in music retrieval for digital rights management and e-business domain.

3.3.2.7 The DS-MIRF MPEG-7 Based Ontology

The DS-MIRF framework [Tsinaraki et al., 2007] made a one-to-one manual translation of the full MPEG-7 MDS and the system tools into an OWL DL ontology. The translation from XML Schema to an OWL ontology is made explicit with an additional OWL DL ontology which holds the mapping from XML to OWL classes and vice versa. Unlike the Rhizomik ontology, it provides a straightforward linking of domain concepts to ontology classes. This ontology has been used in the sports domain (soccer and Formula 1).

3.3.3 MPEG-7 Compliant Ontologies

3.3.3.1 Core Ontology for Multimedia (COMM)

The Core Ontology for Multimedia [Arndt et al., 2007] is one recent approach for formalising MPEG-7 descriptions. It is an OWL DL based formalisation covering the structural, localisation and media descriptions of MDS. It also includes descriptions about algorithms and their input-output parameters for content analysis. COMM extends two DOLCE design patterns: (1) Description and Situation, and (2) an ontology for Information Object, to act as the core foundation. The extended patterns include 1) a decomposition pattern for structural and localisation aspects, 2) a content annotation pattern to attach metadata to segments, 3) a media annotation pattern for physical realisation of the multimedia content, and finally 4) the semantic annotation pattern which is meant to link the multimedia entity with domain knowledge through an annotation pattern. COMM has been used in the industrial domain.

3.3.3.2 Multimedia Metadata Ontology (M3O)

Multimedia Metadata Ontology (M3O) [Saathoff and Scherp, 2010] has recently been proposed for annotating rich structured multimedia presentations. It is a kind of Meta model to integrate existing multimedia metadata standards and formats. It is based on the DOLCE+DnS Ultralight (DUL) foundational ontology. The description logic based DUL provides some generic classes such as event, object and information. To represent various formats, M3O proposed six design patterns: 1) Decomposition Pattern, 2) Annotation Pattern (low-level description, non-visual technical details, authorship or semantic annotation), 3) Information Realisation Pattern (distinction between information object (image) and information realisation (jpeg)), 4) Data Value Pattern, 5) Collection Pattern and 6) Provenance Pattern. In order to align any multimedia format with M3O these patterns need to be specialised. M3O described a manual alignment of COMM, EXIF and the Media Ontology with itself through a four step iterative process.

3.3.4 Media Type Ontologies (Images, Audio)

3.3.4.1 Media Resource Ontology

The Media Annotation Working Group (MAWG)[20] from W3C initiated an effort to provide a core set of descriptive properties of media object on the web. The objective is cross-community media data integration on the web. According to W3C, the vocabulary is mostly targeted towards media resources available on the Web, as opposed to media resources that are only accessible in local archives or museums. It recognised 20 formats (MPEG7, IPTC, Dublin Core, EXIF, VRA, DIG35[21], etc.) and specifies the syntactic and semantic level mapping between various schema attributes. It's basic properties (28 properties) include elements to describe various media aspects; identification, creation, content description, relational, copyright, distribution, fragments and technical properties. Media resources here can denote both the abstract concept of media as well as a specific instance. For the sake of simplicity, the Media Resource Ontology does not make distinctions between different levels of abstraction that exist in some formats. In addition, the ontology is accompanied by an API that provides uniform access to all its elements. The ontology also defines mappings from properties to various existing schemas and recommends ways to provide API methods for data access. An example: the property createDate from XMP can be mapped to the property DateCreated from IPTC. The API will then define an access method that will return values either from XMP or IPTC metadata. Regarding some important classes, it is worth mentioning that a MediaResource can be one or more images and/or one or more audio-visual MediaFragment. By definition, in the model, an audio-visual MediaResource is made up of at least one MediaFragments. A MediaFragment is the equivalent of a

[20] http://www.w3.org/2008/WebVideo/Annotations/
[21] URL http://xml.coverpages.org/FU-Berlin-DIG35-v10-Sept00.pdf.

segment or in some standards like NewsML-G2[22] or EBUCore[23] , a part. At the same time, a MediaFragment is composed of one or more media components organised in tracks (separate tracks for captioning/subtitling or signing if provided in a separate file): audio, video, captioning/subtitling, signing. There could be other types of tracks like a 'data' track, etc.

3.3.4.2 DIG35

The Digital Imaging group proposed DIG35, an XML-based specification for recording and describing image metadata. The specification aimed to provide in-depth meta-information on the image, its creation date and time, focus distance, content and subject, light level, GPS location, etc. DIG35 aimed to promote extensibility and interoperability between imaging devices. IBBT Multimedia Lab[24] (University of Ghent) in the context of the W3C Multimedia Semantics Incubator Group developed a full OWL Schema covering the entire DIG35 specification.

3.3.4.3 Mindswap Image Region Ontology (MIRO)

The MindSwap Image Region Ontology[25] is an OWL Full ontology which models concepts and relations covering various aspects of the digital media domain (Image, Segment, Video, Video Frame, etc.). The objective of the ontology is to describe what is depicted within various types of digital media, including image and videos [Halaschek and Schain, 2005]. The ontology defines concepts including image, video, video frame, shot, region, as well as relations such as depicts, segmentOf, hasRegion, etc. MIRO consists of 14 classes and 12 object properties.

3.3.4.4 EXIF

Exchangeable File Format (EXIF)[26] is a specification for image file formats used by digital cameras, smartphones and scanners. It is a standard for storing interchange information inside the image file to encourage interoperability across imaging devices. Its use by camera manufacturers is almost universal. The nature of metadata covered in EXIF describes camera settings, capture conditions, time (digitised, modified), image data (pixel composition, pixel data, pixel aspect, thumbnail), copyright, creators (artist). Most smartphones and some cameras are now available with built in GPS-receivers that stores the photo location within the EXIF header. An RDF representation of the EXIF format is

[22] http://www.iptc.org/cms/site/
[23] http://tech.ebu.ch/publications/tech3293
[24] http://www.mmlab.be
[25] http://www.mindswap.org/2005/owl/digital-media
[26] EXIF 2.2. Specification by JEITA, April 2002. Available at http://www.exif.org/Exif2-2.PDF.

reported where IDF is a single class and all the tags are its properties. In spite of its immense popularity among camera manufacturers and its near universal presence in imaging devices, the specification has some major limitations: 1) It is applicable to JPEG and TIFF image formats, but not other formats, 2) It allows camera manufacturers to add many proprietary tags not included in the specification, as a result leading to the same interoperability problem it aimed to address.

3.3.4.5 PhotoRDF

PhotoRDF[27] is an W3C note for describing and retrieving digitised photo with (RDF) metadata. It describes metadata using various existing schemas such as Dublin Core, EXIF to describe various photo related metadata (creator, date, coverage etc.), technical descriptions (camera, film) and content related descriptions using dc:subject.

Other image related metadata specifications are such as XMP and IPTC can be embedded in the image header as-key value pairs.

3.3.4.6 VRA Core

The Visual Resource Association (VRA) consists of many American universities, galleries and art institutes. These institutes often maintain large collections of (annotated) slides, images and other representations of works of art. The association has defined the VRA Core Categories[28] to describe their archived collections. The VRA Core is a data standard for the description of works of visual culture as well as the images that document them. The last release version is VRA Core 4.036 and consist of 19 descriptors for three types of objects: work (vra:Work), collection of works and/or images (vra:Collection) and finally an Image (vra:Image). The VRA Core 3.0 elements were designed to facilitate the sharing of information among visual resources collections about works and images. A work is a physical entity (unique event or object of cultural production) that exists, has existed at some time in the past, or that could exist in the future (e.g., painting, composition, an object of material culture). An image is a visual representation of a work in part or in whole (a digital image of an artwork, a photograph of a building). A visual resources collection may own several images or a group of works.

3.3.4.7 ID3

ID3 is a metadata container used for the MP3 audio file format. Like EXIF, it is a set of attributes with key-value pairs. It is considered the de facto metadata storage format for MP3 audio files. An extended version of ID3 was released to include textual metadata such as title, artist and album fields.

[27] http://www.w3.org/TR/photo-rdf/
[28] http://www.vraweb.org/projects/vracore4/

3.3.5 Domain Ontologies

3.3.5.1 Movie Ontology

The Movie Ontology (MO)[29] is an OWL based specification which provides a controlled vocabulary to semantically describe movie related concepts such as "movie", "actor "director". The objective is to standardise movie representation across databases and on the web. It has 39 object properties and 68 classes.

3.3.5.2 IMDB Movie Ontology

IMDB Movie Ontology is based on the IMDB database and tries its best to keep its original semantics while describing various concepts related to a movie. The root class is "IMDBObject" which has three subclasses : "Movie", "AwardsAndNominations" and "ScheduleAndLocation". The Movie class has 20 major processes described as sub classes ("Acting", "Advertising", "cinematography", "Art", "Composition" etc.). The Movie concept has two generic subclasses called "MoviePerson" to describe people related to a movie and "MovieCompanyInformation" to describe companies and organisations related to the movie. The AwardsAndNomination class describes awards and nominations received by people related to the movie while "ScheduleAndLocation" represents the timing and locational information with four more specialised concepts.

3.3.5.3 Music Ontology (MO)

The Music ontology[30] is an attempt to link all music related information (Artist, Albums, Tracks, performances, arrangements etc.) together in order to facilitate user queries and music discovery. It describes music along with its production workflow, performance and release events. The Music Ontology depends on the Event ontology, the OWL Time ontology and the Timeline ontology to describe various temporal attributes. MO is influenced by the ABC ontology, and the FOAF ontology. It has 138 classes and 267 object properties. It also includes an audio feature ontology describing audio signal and its characteristics. MO is used extensively in various Linked Data applications including the resources available on DBTune.

3.3.5.4 Kanzaki's Music Vocabulary

Another music related vocabulary or ontology has been described by Kanzaki[31] for modelling classical music and performances. It defines classes (categories) and their attributes for musical works, events, instruments and performers. Kanzaki vocabulary makes a distinction between musical

[29] http://www.movieontology.org/
[30] http://musicontology.com/
[31] http://www.kanzaki.com/ns/music

works (e.g. Opera) from a performance events (Opera_Event), or works (String_Quartette) from performer (StringQuartetEnsemble in this vocabulary).

3.3.5.5 SIMAC Ontology

Music Recommendation Ontology[32] used FOAF profiles to recommend music to a user. It describes concepts such as artist, music title, as well as audio content properties such as tonality, rhythm, mode tempo, intensity, etc. This ontology is a part of the system "Foafing the Music" [Celma, 2006] which aims to recommend music based on a personal profile and listening habits. It also describes a mapping to the MusicBrainz ontology within the MPEG 7 specification.

Some other music related ontologies that are worth mentioning include; 1) The MusicBrainz community produced a small RDF vocabulary (Musicbrainz RDF Schema) based on MusicBrainz data. The class structure includes concepts for artists, albums, tracks. 2) The Nepomuk ID3 Ontology [33]is mainly focused around the concept of an audio item (as in ID3).

3.3.5.6 BBC Programme Ontology

BBC Programme ontology[34] aims to provide a simple vocabulary for describing programmes and its schedules. It covers brands, series (seasons), episodes, broadcast events, broadcast services, etc. It reuses various existing ontologies such as Event, Music, FRBR, Dublin Core, SKOS, Timeline, and FOAF to describe various programmes and their interactions. It has 40 classes and 48 properties.

3.4 Use of Ontologies in the Multimedia Research Community

The Use of ontology for knowledge representation and reasoning is slowly gaining ground within the core multimedia community. There are some studies of the use of ontologies in domains such as video events and activity detection, and image annotation. Leslie et al. [2007] proposed a two-layer artistic concept ontology, which includes the visual characteristics such as colour, texture in the first layer and artist name, painting style, art period on the second layer for the annotation of paintings. Holink et al. [2005] defined a set of rules in Semantic Web Rule Language (SWRL) for image annotation. Saathoff and Staab [2008] modelled image segments with visual descriptors and associated semantic concepts through manual labelling. They exploited context information to reduce the label ambiguities by using constraint reasoning techniques. An ontological framework was developed by Liu [2004] for semantic interpretation of the image content in natural scene domain. He used it on top of the

[32] http://foafing-the-music.iua.upf.edu/music-ontology
[33] http://www.semanticdesktop.org/ontologies/nid3/
[34] http://www.bbc.co.uk/ontologies/programmes/2009-09-07.shtml

MPEG-7 standard for efficient image query and retrieval. Akdemir et al. [2008] followed general ontology design principles and adapted them in domains of human activity, bank and airport surveillance videos to detect interesting events. Francois et al. [2005] proposed a formal language to describe events using Allen's formal logic. They used the approach for detecting events in surveillance videos. One recent study by Scherp et al. [2009] defined a formal event model to allow interchange of information between different event-based systems. It uses causal relationships between events, and interpretations of the same event by different people. Neumann and Moller [2006] used description logic to reason over an aggregated view of various scene parts and detect semantic concepts in the scene. Bai et al. [2007] defined a soccer ontology and applied temporal reasoning for event annotation in soccer videos. Snidaro et al. [2007] combined simple events to detect complex events in security videos. Ballan and Bertini [2010] used SWRL rules to detect events in the soccer domain such as "goal", "players' movement" etc. Bertini et al. [2008] proposed a framework for semantic video annotation using the Pictorially Enriched Ontology Model. The framework learnt spatio-temporal SWRL rules to detect events. They validated their study with TRECVID 2005 data.

3.5 Rationale for our Proposed Model

We have discussed various existing ontologies intended for multimedia content description focusing on four main topics 1) media document representation, 2) media content description, 3) media structure description, and 4) semantic content description. A review of the literature reveals two schools of thought prevailing while formalising multimedia ontologies: 1) studies concerned with preserving the comprehensiveness of the MPEG-7 schema with its flexible descriptions at the cost of conceptual clarity, and 2) the Semantic Web focused initiatives: these concerned about formal semantics and reasoning at the cost of coverage. This thesis assumes that these two rigid positions resulted in a comparatively low uptake of multimedia ontologies. This assumption is further validated by recent work [Fergal, 2009] where the author made a survey about the usage of multimedia ontologies in the Semantic Web data. They crawled an RDF dataset of 1.1 billion triples (ISWC 2008 billion triple challenge) and the statistics showed that the most frequently used ontologies describing any multimedia objects were FOAF, RSS, EXIF, PhotoRDF, DCMI and Mindswap (MIRO).

Motivated by the above scenario we proposed a lightweight conceptual model for video on the web to represent video and its semantic content. At the time of the writing we could not find an ontology describing user-generated videos on the web describing all the four core components (Media, low-level content, media structure and domain subject linking) described above. Though W3C working

group on media ontology proposes a predicate centric vocabulary with a set of core properties to describe video on the web, it does not address many issues such as low-level content descriptions which are inseparable in any multimedia related models. Besides these points, our proposed model takes video description a step further by considering it as a social object on the web. Web videos are created and shared for a purpose other than personal use and become an object of interaction between several individuals and group actors. These online social interactions and community activities contribute and shape towards the higher semantics of the video object more details of our model is described in Chapter 5.

3.6 Summary and Conclusion

In this chapter we discussed the need for an ontological representation of a multimedia object and its content in view of emerging challenges on the web. We have discussed various models and ontologies aimed at describing multimedia content. Our discussion started with MPEG-7 and its challenges in the absence of formal semantics. The next section described various Semantic Web based ontologies aimed to translate MPEG-7 standard fully or in part, followed by some specific media type ontologies and a few domain ontologies. Lastly, we briefly described the rationale for our proposed video model, as described in the Chapter 5.

Chapter 4

Social Semantic Web

This chapter of the thesis discuses the topic of Social Semantic web and their complimentary role to create a more semantically enriched and integrated data space on the Web.

4.1 Introduction

The Social Semantic Web has been a subject of interesting research recently. The concept of the Social Semantic Web subsumes developments in which social interactions on the Web led to the creation of explicit and semantically rich knowledge representations. The Social Semantic Web can be seen as a web of collective knowledge systems, which are able to provide useful information based on human contributions and subsequently improves with larger participation and interactions [Gruber, 2006].

The Social Web is part of the existing web and is characterised by social interaction among its users and communities. It mostly refers to social network oriented sites such as Facebook or MySpace and content sharing sites such as Flickr, or YouTube powered by social networking features. According to Parameswaran & Whinston [2007], these social computing systems, designed to allow users to share information, have seen a dramatic rise in popularity in recent years. Such systems facilitate collective actions and interactions among users and communities through the exchange of rich multimedia information. The Social Web revolves around some core concepts: identity, reputation, presence, relationship, groups, conversation and sharing[35]. Some well-known multimedia content oriented social computing platforms are Flickr, YouTube and Lastfm, while Delicious is a social bookmarking service and Facebook is a primarily a social networking system. Continued participation in communities is critical to these services [Burke et al. 2009; Butler, 2001; Chiu, Hsu, &Wang, 2006; Koh, Kim, Butler, & Bock, 2007].

[35] http://connollyshaun.blogspot.com/2008/05/7-key-attributes-of-social-web.html

Before discussing the details of user motivation for participation and contributions on The Social Web, we will briefly describe these social computing platforms and their characteristic features in the next section.

4.2 Social Media Sharing Sites

By social media sharing sites, we primarily mean to discuss those online services (media sharing and social networking sites) that are focused towards multimedia content such as photo, videos, music etc. There are many popular media sharing sites with some similar functionality such as content sharing, user interaction, social connection and content retrieval. Hundreds of photo, video and music sharing sites are present on the web catering to millions of users. Since most of them provides many similar services to their users, our discussion will primarily base on the services and usage patterns of most popular media sharing sites such as Flickr and YouTube.

4.2.1 Flickr

Flickr, a popular and one of the earliest social photo management and sharing web application started in 2004. It allows its users upload, share photos, and build interest-based communities. Its content and community structure received much research attention from multiple domains including computer vision and multimedia retrieval. Since its launch, it has acquired more than 40 million users (Nov, 2009) and more than 5 billion photos (March, 2010).

4.2.1.1 Content Organization by User

Every user can upload photos and videos and annotate them with a title, description and tags to categorise and describe the content. The Flickr system extracts the temporal and location information of the photo from the photo header and makes this available as the photo meta information. Social tagging feature of Flickr allows other users to tag a photo, which is not the case in many other photo sharing sites. Users can do some content level annotation by adding a note inside a photo. Its users can further categorise and cluster his own photos in various sets with proper set descriptions. This meta information is used by the system to categorise, locate and recommend photos and videos. Users can have some control over the access of their content by choosing the options of “public”, “private” and “family and friends”.

4.2.1.2 Social Features

One of the reasons for the popularity of these social computing platforms is their social networking features. It allows users to form groups based on shared interests and thematic content. The social

interaction pattern available in most social media sharing sites (including Flickr), are user comments and ratings attached to an individual photo or a group. People can give their comments, rate and bookmark a photo/video for their personal use as well as for recommendation. Recently most content-driven online services have begun to users to share their content beyond the service through external services such as Facebook, Twitter etc.

4.2.1.3 Search and Browsing

All social media sharing sites also acts as media search engine for their storage contents. Flickr provides basic keyword search over textual content and advanced filtered search mechanisms to discover and browse user photos. The results interface shows photos organised based on relevance, recency and interestingness. The detailed photo page includes all the Meta description (title, date, tags, description) of the photo, right information and details of the sets and groups to which the photo belongs. Search can be based on media content type such as photo, video, illustration and screenshot. Flickr provides API-level access to upload and search its content programmatically. However the Flickr interface does not support any serious recommendation strategy to recommend either tags or content to the user.

4.2.1.4 Privacy and Content Access

Flickr's privacy settings uses the "public" (visible to everyone) option as a system default while allowing the user to have some degree of access control for family and friends.

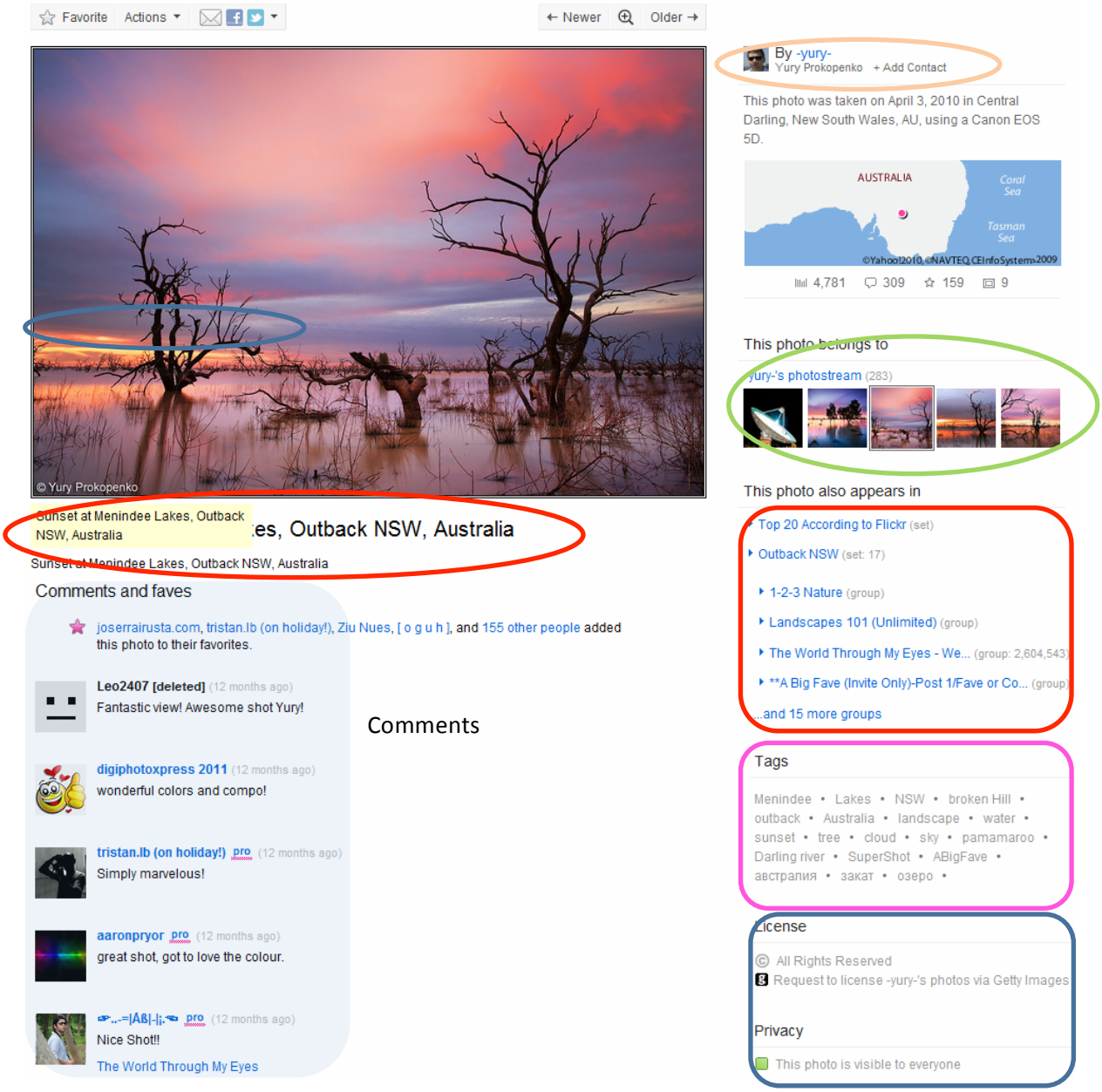


Figure 9: A screenshot of Flickr photo page marked with photo Meta information.

4.2.2 YouTube

For video sharing, YouTube is the most popular video sharing site on internet. According to the recent statistics, YouTube claims that 34 hours of video are being uploaded per minute. Started in 2005, YouTube now accounts for 60% of video traffic on the web and has crossed the 2 billion views per day mark.

4.2.2.1 Content Organization by User

Users in YouTube can create (via webcam), upload, share and annotate videos, add title, tags and descriptions to their videos. YouTube videos are categorised with a flat category structure of 15 categories leaving much space for improvement as most of the videos may well belong to multiple categories or be best described in a hierarchical category structure. YouTube users can add captions, subtitles and time-stamped annotations to their videos. Basic content (visual and audio) editing have been rolled out recently. Users can bookmark and create playlists of videos for better organisation of the content.

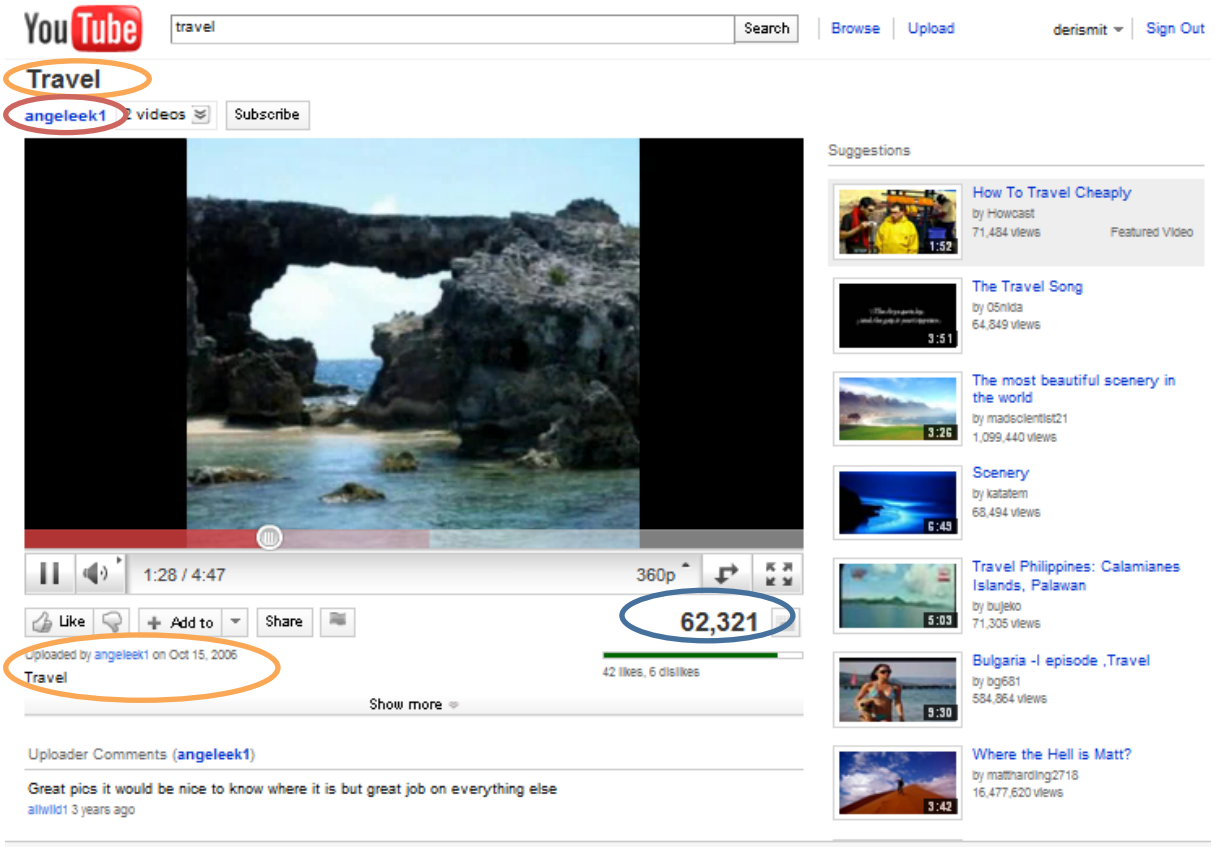


Figure 10: A screenshot of an individual video page on YouTube with various meta information.

4.2.2.2 Social Features

Modern video sharing sites include social features such as user comments about a specific video, voting and rating of the videos in terms of "like/dislike". It allows users to create a contact network through subscription and friendship relationships. Sharing of videos are done through embedding HTML snippets which includes the video URL or external sharing through social networking services such as Facebook, Twitter. Community features of YouTube allow its user to create and join groups

of like-minded users, or theme-based groups such as “Extreme sports” where all the videos of this nature are grouped together.

4.2.2.3 **Search and Browsing**

Search and browsing of content in YouTube follows the normal search paradigm of keyword search and advanced filtered search with constraints on date, view counts, rating, relevance and categories. It also provides users with a browse interface listing popular videos in different categories. However the search mechanism is purely text based and entirely depends on the amount of information provided by the video creator. YouTube recommends a list of videos based on the presently viewed video as well as based on user profiles and viewing history. It also provides API level access to its data and various usage statistics.

4.2.2.4 Privacy and Content Access

The privacy setting of YouTube videos is by default public, but users can change the setting to mark them as entirely private or accessible through links. A user can restrict the sharing and embedding of a video through its upload settings. However YouTube has its proprietary mechanism to detect and delete copyrighted materials or objectionable content.

4.3 Tagging and Folksonomies

The emergence of collaborative tagging enables users to assign a freely chosen keyword to describe any resource. Users select keywords based on their preferences and on a subjective understanding of the resource and its content. Collective contributions of the user’s view represent a common understanding of the resource from a community perspective and is called as “folksonomy” [Mathes, 2004]. A tag is a keyword or term assigned to a piece of information (such as a web page, music, and video) as its description. This kind of metadata helps to describe an item and allows it to be discovered by search engines. Tags are generally created personally by the creator or by its viewer, depending on the hosting system, hence the process of creating and labeling tags to a resource is called “tagging”. Aggregation of tags to a resource from multiple users is called “folksonomy”; a term coined by Thomas Vander Wal, to describe the taxonomy generated by folks by means of social tagging.

Tagging as an information management practice is not new, but has been in use for decades in libraries and other information management spaces. Documents in libraries were organised by assigning keywords for future search and navigation. These organization efforts were carried out by

librarians or authors [Rowley, 1985]. In contrast, social tagging or collaborative tagging is a recent web specific phenomena where users assign keywords called tags to digital resources such as photos, web documents, videos or pieces of music. Collaborative tagging proved useful in the absence of any central organising authority. Some of the early adopters of social tagging were Del.cio.us, Flickr. These services allowed its users to tag resources with keywords and use the same tags to navigate other publicly available content within the service. Most content-driven online services now include tagging as one of their integrated features, not only for content organisation but also as a tool for search and browsing.

A folksonomy is often compared and contrasted with a taxonomy. Taxonomy is a pre-defined hierarchical structure (e.g. the Dewey Decimal Classification for libraries, computer directory systems) where each term is associated with one category which leads to an automated structuring of content that allows search with different level of specificity. Whereas tagging is a non-hierarchical, flat and inclusive structure that consists of keywords and continuously evolves with usage and time, it contains no explicit hierarchy, or any specified parent-child or sibling relationships between terms. All the tags are considered at the same level. One of the defining characteristics of folksonomy systems is that it promotes a democratic and distributed classification system where multiple users can tag without a central authority of control.

Being uncontrolled and unstructured, folksonomies inherit some fundamental flaws (discussed in the next section) but there are many important benefits which are worth mentioning in order to understand the utility of these social platforms. Six major benefits are discussed in one of the first paper on folksonomies [Mathes, 2004] 1) browsing and filtering: tagging facilitates browsing of information as opposed to looking for it in a search engine. The fundamental difference is that the user has the knowledge of the information space within which she/he is navigating. 2) Desire lines: this is described as the digital pathway of the users which reflects an individual's vocabulary space from their tag usages [Merholz, 2004]. 3) Low cognitive cost: there is less effort required by allowing tags to be created by any users as opposed to an expert annotation based on taxonomy, 4) Asymmetric communication and feedback loop: immediate feedback is one of the reasons why tagging works. When user uses a term to tag, she/he gets all the resources with the same tag, thereby getting the chance to include tags used by others or refine his/her own tag space. 5) Personal and social utility and finally 6) unexpected use: unexpected uses of tags include the formation of interest groups and interest based participatory social networks to communicate messages and organise events.

4.4 How People Tag?

To understand social tagging we need to understand some basic questions e.g. What are the tagging practices? What are the typical tag distributional statistics? What is the quality of tags in terms of content semantics? Fabian et al. [2008] described that 50% of the tagging words are not known to the dictionary (WordNet); this may not be true for all social tagging platforms but shows the importance of filtering the informative tags from the non-informative ones. This study also showed that common nouns and proper nouns are higher in frequency compared to other tag categories such as verbs, adjectives. Major tag related research studies included tagging platforms such as Deli.cio.us, Flickr, weblogs, MovieLens, YouTube, CiteuLike, Lastfm and Amazon. Literature shows that, despite the benefits of tagging in recall and retrieval, people are lazy taggers. From our dataset of more than 100000 Flickr photos randomly collected, 20% of the photos do not contain a proper title, but are instead given titles which make little semantic sense; Examples of such titles are IMG_001, DSC012, etc. Users uploading in batches (photos taken at one event) give the same title and tags to the whole batch of photos even if the visual content is clearly different. Studies by Kirk et al. [2006] showed that people are often interested in event based organisation of photos.

Regarding the type of tags, the trend has not changed significantly. In 19 2004, the 150 most popular tags from Flickr included common subjects of photos: cat, friends, dog, sky, sea, park, kids, garden, baby, building, flower, sculpture, city and vacation. Over 25% (41 out of 150) of the tags were proper place names like cities or countries. Colour names and years were also mentioned as popular tags. A recent most popular tag cloud from Flickr (as shown in Figure 11) lists tags such as, proper names locations like Japan, London, top camera names such as Nikon, Canon. Popular tags also include tags for events such as wedding, party, music, etc. Many of the popular tags from 2004 (cat, dog, sea, sky, baby) are now down in the list of the most 145 popular tags. The top tags for places occupy the same 25% (34/145) percentage share.

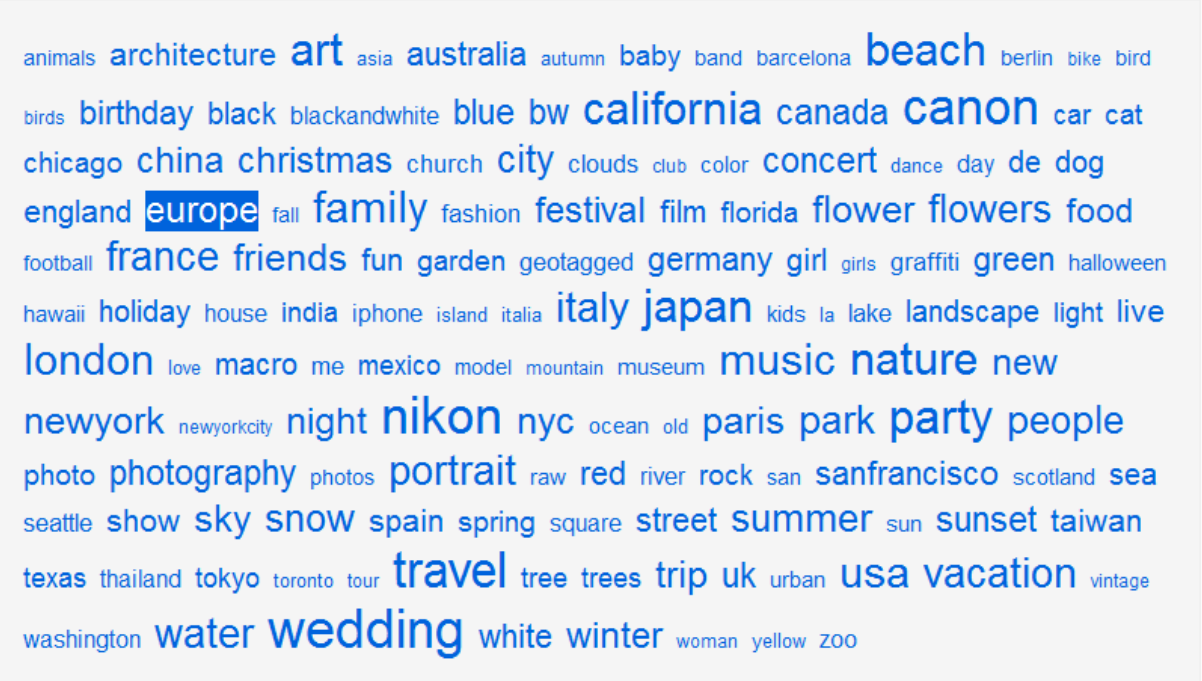


Figure 11: Flickr Tagcloud.

The top tags above show a practice of soft categorisation of content. Many tags are semantically related and can be grouped together such as (travel, trip, and vacation), (flower, flowers, nature, tree, and trees, and garden).

4.5 Why People Tag?

Social tagging differs from expert annotation in terms of motivation and objectives. The prime motivation behind tagging is cited as organisation and retrieval of resources. However many studies showed that besides personal motivation, social dimensions greatly influence tagging behaviour. Kustanowitz and Shneiderman [2005] classified social motivation of tagging in three categories: "family and friends", "colleagues & neighbours" and "citizens & markets". Ames et al., [2007] described users tagging motivation in a two dimensional scale: 1) function and 2) sociality. "Function" implies organisation and communication while "sociality" implies utility for self and community, family and friends. Marlow et al., [2006] described tagging motivation in terms of organisation, retrieval, contribution, sharing, attention seeking, play, competition, opinion expression and self-presentation. One of the first studies about user motivation in tagging was carried out by Golder [2006], who identified seven functions of tag i.e. identifying what and who, identifying what it is, who owns it, refining categories, quality and characteristics, self-reference and task organisation. Korner et al. [2009] described user motivation in terms of categorisation and description. Categorisers are the users who use tags to create a navigation space for the resources and categorise resources based on some shared high level characteristics. They use a stable vocabulary with less redundant tags

while the describers are those who describe the content in the resource for the purpose of efficient retrieval. A Describer's tag set is more diverse and dynamic in nature, contains lots of synonyms and infrequent tags. User motivation has a direct impact on the usability of tags [Nov and Ye, 2010].

4.6 Challenges for Tagging

Tagging, despite its wide acceptability, poses certain inherent drawbacks when used for content description, search and navigation. The flaws can be broadly classified under two classes 1) natural language specific limitations includes the problem of polysemy, synonym and basic level variation [Golder et al., 2006], and 2) social media specific limitations (slangs, lexical variations, community specific tags etc.).

Polysemy: A polysemous word is one that has multiple related senses. For example, "window" may refer to a hole in the wall, or to the pane of glass that resides within it [Pustejovsky 1995]. A multiple sense query dilutes the search result by returning many irrelevant results due to its related yet different senses. It is similar to homonym where a word has multiple yet unrelated senses e.g. "apple" may refer to either a fruit name or to the computer company.

Synonym: Synonym problems arise when multiple words have similar meanings. It poses a greater problem when people use different terms to tag the same resource. As a result searching with one version of the word will miss the results tagged with other synonym words. For example a photo may be tagged with "NYC" or "NY" to describe "New York City". The available options are either that a complex query has to be formulated in order to capture all possible synonyms and the corresponding resources or that one is prepared to miss out an unknown amount of relevant results.

Basic level variation: This is the discrepancy in the choice of words depending on the expert knowledge of the user. While describing items, users select a keyword that varies in degrees of specificity, starting from a very general word like "animal" to a specific word "Siamese", while the term "cat" falls in between. While all three terms are meant to be semantically related, an expert of "cats" will select a specific term to describe them, while normal users will prefer the universally accepted terms like "cat" [Tanaka & Taylor, 1991].

Compound tags: Most social tagging systems were originally made for single word tags. Though multi-term tags can be put inside quote (Flickr) or delimited with comma, people are still used to the practice and use compound tags, where two or more words are concatenated together without any white space. For example the tag "night-time" is intended to indicate "night time", "5photosaday" is originally meant to be "5 photos a day". These tags are difficult to disambiguate in any dictionary. It is difficult to expect that any user will anticipate and use advanced queries with compound tags for

searching. Many event tags are created periodically by specific communities, for example, conference tags ("eswc2011" the tag for the International Semantic Web Conference 2011) are frequently used by users while tagging conference related resources.

Social media specific tags: The interaction style in social media and social networking sites are largely informal in nature which also reflects the type of tags created by users. They use many culture specific words, slang words, abbreviations and acronyms to describe contents.

Adjective and vague tags: Users use many subjective and vague tags such as adjective to describe content e.g. funny, crappy, great video, wonderful shot etc. instead of content specific tags.

Contextual tags: Users also use many contextual tags expressing place, time, device (camera, phone) where the content is presumed to be implicitly embedded such as "view from my window", "after breakfast") which are difficult to process and ensure what exactly is visible. These kinds of tags are specifically meant to communicate with friends and families known to the tagger.

Misspelling, common spelling variations of a tag: tags with only numbers (09, 08), or with common usage of singular and plural nouns such "flower" and "flowers".

Owing to these problems, user-generated tags cannot be readily used as a substitute for expert annotation. Studies in [Kennedy at al., 2007] reported that almost 50% of user-generated tags are not of much use for general public. They need to be filtered, aggregated and validated for extracting semantics. The next section will describe various efforts attempted to address the issue of tag semantics.

4.7 Tag Semantics

With the increasing popularity of tagging and large folksonomies, a new problem of the "semantic disconnect" between user-generated content and their intended meanings has become apparent. Due to the linguistic problems inherent in tags as discussed above, sharing and aggregation of tags has proved difficult across systems and tools, leading to multiple data silos. These difficulties were mainly due to the absence of formal structure of tags and the corresponding systems.

The idea of tag semantics started with the realisation of the fact that 1) tag patterns become stable after some time [Golder and Huberman, 2006], 2) tag distribution follows a power law distribution [Halpin et al., 2007], 3) feedback encourages tagging (popular tags have higher probability of being tagged). Hence various efforts started to extract concepts and the relationship implicit in the

folksonomy. Researcher, from various communities stressed that Web 2.0 and the Semantic Web can complement each other and benefit from each other's strengths. It was emphasised that a folksonomy approach to building an ontology is a more practical and obvious approach for users compared to the traditional rigid ontology engineering approach which need more time and expertise. Moreover, a folksonomy based approach can quickly adapt to the changes and evolve with the community needs.

Based on the existing literature, several studies were carried out to learn conceptual hierarchies from folksonomy following three major approaches, which include:

- Statistical methods (co-occurrence analysis).
- Machine learning methods (clustering and association rule mining).
- Graph and network analysis over folksonomy.

One school of thought is based on studying statistical patterns of tags, users, resources and their usages. This group prefers a bottom up approach and calls it emergent semantics. The study of the statistical patterns of human word usage for semantic interpretation is called statistical semantics [Furnas et al., 1984]. The term statistical semantics was coined by Weaver [1955] for a task of machine translation. The underlying assumption was that word sense disambiguation should be based on co-occurrence statistics of the context words near the target word. In other words, Firth [1957] maintained that "a word is characterised by the company it keeps". One fundamental aspect of statistical semantics is that it considers the wisdom of crowds is close to the truth when compared to a single individual, a view which perfectly blends with social tagging systems where multiple users tag a single resource giving a sense of collective intelligence.

Cattuto et al. [2007] studied the statistical properties of tag associations by means of co-occurring tags in all resources. Their assumption is based on the hypothesis that the semantic contexts of a tag will show certain self-organising hierarchical structures between high ranked and low ranked tags. Their result showed a power law distribution of co-occurring tags where high frequency tags ("web") are more general whereas low frequency tags ("ajax") are more specific. They concluded that "the rich get richer" theory exists in social collaborating systems where popular tags have more chances of being selected by new users compared to non-popular tags. Schimtz [2006] tried to induce an ontology from Flickr tags using a subsumption model. His objective was to present a faceted ontology as a substitute for a folksonomy which can retain both the simplicity of a folksonomy and the accuracy of ontology, and he found facets such as location, interesting points. Rattenbury et al. [2007] extracted tag semantics especially place and event semantics from Flickr photos using temporal and spatial patterns of tags. They detected events and places using a combination of burst detection and Scale-Structure Identification which computes the cluster similarity at various spatial scales. Location and interest point detection was reported in [Moxley et al., 2009] by mining geo-referenced

coordinates and co-occurrence statistics of Flickr tags. Specia and Motta [2007] tried to integrate folksonomy and Semantic Web technology to extract ontological concepts and relations from Flickr tags. Their cluster based approach revealed clusters of concept related tags. A relationship between tags was discovered using the Semantic Web search engine Swoogle[36]. However they reported that due to the sparseness of Semantic Web data, mapping to ontological concepts is poorer compared to a WordNet mapping. Benz et al. [2008] combined iterative clustering and tag networks to get an incremental taxonomy structure.

Another method used for ontology learning is "association rule mining" or frequent set mining introduced by Agrawal et al. [1993]. In market basket analysis, transactions are analysed to detect the sets of items bought together. Here tags are considered as items and tag assignments are the transactions used to mine the rules. Schimtz et al. [2006] used association rule mining on a tripartite hypergraph of folksonomy and derived rules such A→B which reads as users who assign tags from A often assign tags from B to the same resource. They also cited rules for subsumption relation as in taxonomic structure. Lin et al. [2009] employed a low-support association rule mining to extract ontological concepts and relations from the fruit knowledge domain.

4.7.1 Network Representation of Folksonomies for Tag Semantics

There are many approaches where a folksonomy is represented as a graph or network. Lambiotte [2005] proposed a tripartite graph model to represent a folksonomy. This network consists of three finite sets of nodes (user, tag and resource). The resulting network can be presented as a graph where edges run between the resource (r_i) and user (u_i) passing through tag (t_i). The graph represents $U=\{u_1,u_2,\ldots u_n\}$ as a set of users, $R=\{r_1,r_2\ldots r_n\}$ as a set of resources (web pages, photos, videos) and $T=\{t_1,t_2,\ldots,t_n\}$ as a set of tags. Each user u_i can be represented with an $R_n x T_n$ matrix where the cell of matrix will take the frequency of tag t_i for resource r_i, and a similar matrix can be created for each tag and resource separately. In order to avoid the complexities of computing on a tripartite graph, the authors suggested some projection methods thereby converting this tripartite graph into normal bipartite and then a unipartite graph. A projected bipartite network will represent a user in vector space for both resources and tags separately, while a unipartite network will be based on the similarity between nodes of one kind (u_i, u_j). These network structures were then exploited to build concept taxonomy out of flat folksonomy. Mika [2007] proposed a model based on a similar tripartite model of actor (A), concept (C) and instance (I) network to detect community based semantics from large folksonomies. His model was based on an implicit assumption that meaning is necessarily depends on agents of a community. Thus the traditional bipartite model of an ontology with concepts and instances is given a third dimension of users creating a set of annotations T =AxCxI. Network

[36]http://swoogle.umbc.edu/

analysis of these graph structures (bipartite and one mode graphs) clearly showed the evidence of emerging semantics. His result showed that concepts within a cluster are more specialised compared to the concepts between clusters. The result confirmed that concepts emerging out of the actor and resource graph (G_{ar}) are of high quality compared to the concept and instance graph (G_{ci}), supporting the assumption of community contributed semantics.

Cattuto et al. [2008] proposed the semantic grounding of tag relatedness in social bookmarking systems. Their study proposed several tag similarity measures to study folksonomy structure. They proposed three tag context similarity measures (where a tag is represented by a user vector, tag co-occurrence vector and a resource vector), co-occurrence similarity and Folk rank [Hotho et al., 2006] to study the tags and their structure. Results of their study showed that tag context and resource context similarity were able to extract synonyms and sibling relations, whereas the co-occurrence measure suggested more general tags suggesting a generalisation and specialization hierarchy. They adopted a WordNet based semantic grounding approach to establish tag to concepts mapping. Their study is based on Del.icio.us tags and they found that only 60% of the total tags could be mapped to WordNet.

A later study by [Benz et al. 2008] published a working group summary about analysing tag semantics across a collaborative tagging system. They studied tag data from both Delicious and Flickr and found that not-so-substantial overlapping between the two systems may be due to the difference in domains under analysis. They found that tag context similarity gives better results in a narrow folksonomy like Flickr, but the result is not encouraging at global scale.

In all of the above studies, the input is the folksonomy and the expected output is either a set of concepts and their relationships, or a simple taxonomy. Most of the relations extracted are hierarchical "is-A" relationship in nature. Some studies reported "synonym" detection but detecting specific and functional relationships proved difficult. This may be due to the noisy nature of tag data. We assume that tag data need efficient pre-processing steps in order to avoid the noise and get some semantic structure out of it.

4.7.2 Semantic Representations of Tag and Tagging

Tags are simple and effective but not suitable to share across tools and applications due to its informal nature. The reason for tags not being suitable to share semantically is that it comes out of a user's personal space with a specific purpose and context independent of other users and without any shared agreement. The same tag may not be used by other users for the same purpose or for the same resource; he/she may use a different tag which reflects his the view about the resource. This incompatibility in user's perceptions about tag meanings makes the tag in its original form (keywords)

semantically incompatible. The answer to this issue is the "formal conceptualisation of folksonomy" [Gruber, 2005]. An ontology for a folksonomy is not meant to organise information but rather expose the implicit semantics inherent in user-generated tag with its context and purpose. Based on this assumption, the first conceptualization of folksonomy was the "Tag Ontology".

According to TagCommons[37] any tag formalisation should address the following:

- "Tagging" is an assertion (activity) by a tagger (user/agent) who uses a label to describe a tagged item (resource).
- "Tagger": describes an agent who creates tags for a resource
- "TagLabel" is a textual description of the "tagged item". It may be a word, phrase.
- "TaggedItem" is a resource identified by a universal identifier and tagged by any tagger.
- "TagSource" is the environment where the tagging originates.
- "TagAssertionTime" is the temporal attributes of a tagging activity.
- "TagRatings" is a voting mechanism to rate the tags or to filter the spam tags.

Based on these requirements, we describe some of the conceptual models designed to describe tagging activities in a collaborative system.

4.7.2.1 Tag Ontology

The Tag Ontology started with the objective to share tags semantically. It gave two use cases as its requirements for such ontology:

1) Collaborative tagging across multiple application
2) Collaborative filtering based on tagging

Core Conceptualization

The core concept of this ontology is "Tagging" which accounts for the full environment of social tagging [Newmanet al., 2005] including perspectives from users and applications. From a user's perspective, "Tagging" is an activity where an object is labelled with a tag (Tagging (object, tag)). This is fine in a closed world but to enable collaborative filtering the notion of "Tagger" is required, hence it became a concept with three relations i.e. Tagging (object, tag, Tagger). Different applications and data sources use tags for various purposes and in different contexts. To avoid tag silos and to make tags interoperable at the application level one more relation of "source" (e.g. a site like Flickr) added to the tuple would make it: Tagging (object, tag, tagger, and source).

[37]http:// tagcommons.org

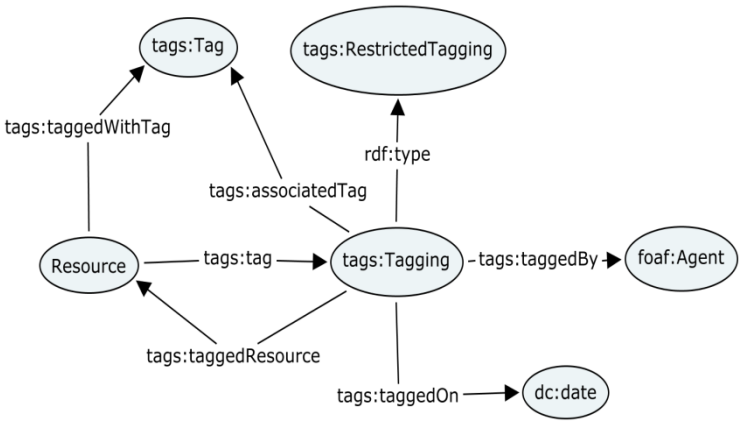


Figure 12: Conceptual model of Tag Ontology.

Figure 12 shows the core model of the Tag ontology. Tagging is a subclass of skos:Concept, it has "tags:Tag" property to associate it with a resource, which is identified with a resource URI. A "Tagger" is a foaf:Agent. Tagging can be described with a tags:taggedOn property to indicate its time of tagging. The relationship between tags is expressed via properties like tags:equivalentTag and tags:relatedTag.

4.7.2.2 Tagging Ontology (TagOnt)

The Tagging Ontology proposed by Torben Knerr [2006], is one of the first formal descriptions of tags and tagging produced in order to make published tag data interoperable. It also followed Gruber's conceptual model as its core with a few extensions. TagOnt has a root class "Tagging" which consists of a "Tag" created by a "User" (a user may be a person or group). Tagging is described within a service domain ("ServiceDomain") which is equivalent to the "Source" concept of Newman's tag ontology. The exception of this ontology is that it defines a "hasType" property to describe the media type and "hasVisibility" property to regulate access (private, public, etc.) to the tag data.

4.7.2.3 Meaning of a Tag (MOAT)

Another lightweight ontology described in [Passant, 2008] is used to represent meanings of a tag. It encapsulates tag meaning with both global and local scope. Global meaning includes a list of all meanings available to a tag in the folksonomy. It defines a Tag object extended from the Tag class of the Tag Ontology and is related to a "Meaning" class through the "hasMeaning" property. Tag's local scope signifies the meaning in a local context within a tagging session. This local meaning is modelled using the "tags:RestrictedTagging" class with a "moat:tagMeaning" property. Figure 13 shows a visual representation of MOAT ontology.

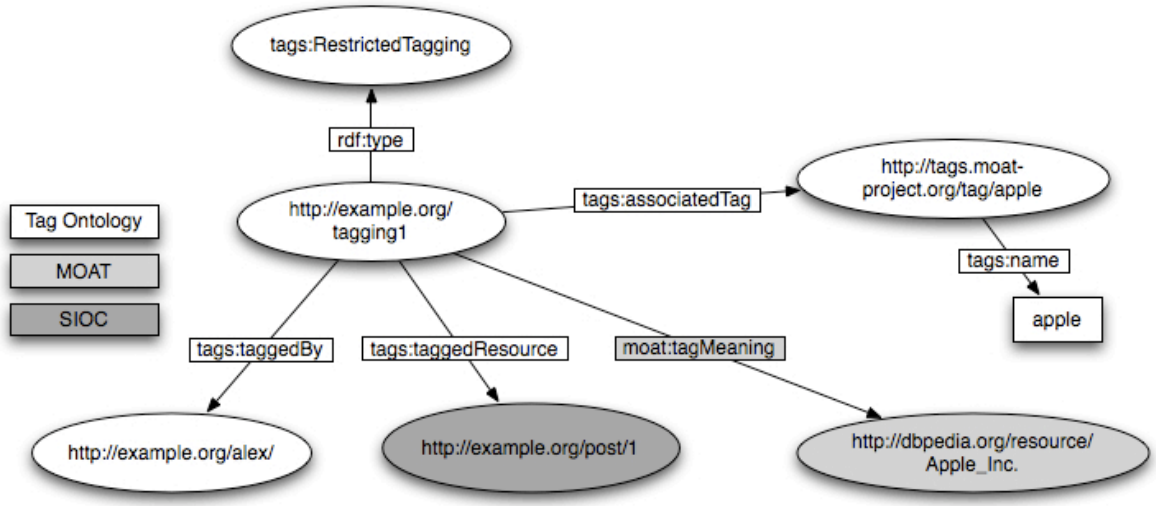


Figure 13: MOAT design (Image from http://moat-project.org/ontology).

4.7.2.4 SCOT

The objective of the SCOT (Figure 14) ontology is to describe the structure and semantics of a collection of tags and to represent social relations across different sources. Its focus is to describe the collaborative tagging process and represent the folksonomy features such as user group, frequencies, tag co-occurrences etc. This lightweight framework integrates concepts from various existing vocabularies such as SIOC, FOAF and SKOS, besides the core scot:Tag class, it also uses scot:TagCloud to describe the collection of tags. Its folksonomy model is formalised as follows:

Folksonomy: (*tag set*, *user group*, *source*, *occurrence*, *Tagging*)

This folksonomy model represents the collective tagging process by a group of users with a set of tags where each tag is represented by Gruber's conceptual tagging model:

Tagging: (*object*, *tag*, *tagger*).

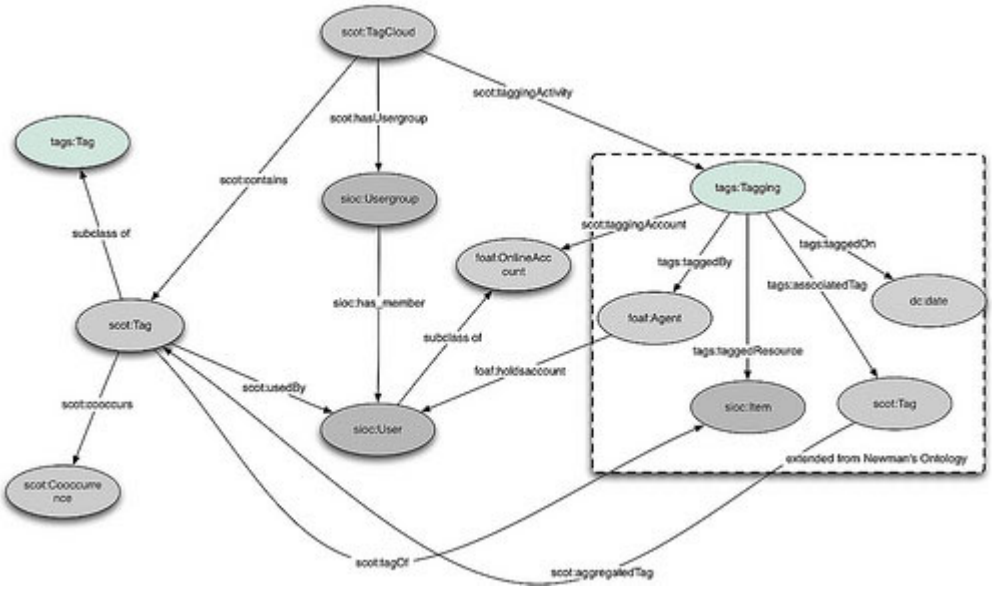


Figure 14: Core conceptual model of SCOT ontology (http://www.scot-project.org/).

SCOT also provides ways to model a tag with its language specific occurrences such as spelling variations, acronyms, singular and plural relations by using properties such as scot:spellingVariant and scot:delimited. It also represents several folksonomy characteristics such as scot:hasUsergroup, scot:createdBy, scot:contains, and scot:taggingActivity.

4.7.2.5 Common Tag

Common Tag[38] is an open tagging format to make content more connected, discoverable and engaging. It is defined in the RDFa[39] format to embed structured data within HTML. As with other formats it identifies a tag with a URI. It has four classes Tag, AuthorTag, ReaderTag and AutoTag. AuthorTag, ReaderTag and AutoTag are subclasses of the class "Tag" based on the methods of tag creation. A tag created by the author of the resource is called "AuthorTag" while a tag given by the user of the resource is called a "ReaderTag" and automatically created tags are classified as "AutoTag".

[38] http://commontag.org/Home
[39] http://www.w3.org/TR/rdfa-syntax/

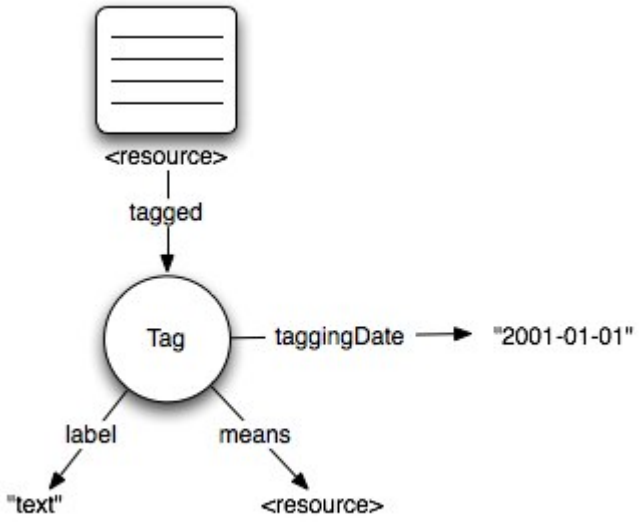


Figure 15: Conceptual model of Common Tag.

Figure the 15 shows the core model of Common Tag where the meaning of a tag is expressed through "means" property of the "Tag" class.

4.7.2.6 Upper Tag Ontology (UTO)

Upper Tag Ontology (UTO) [Ding, 2009] is focussed on the structure of tagging activities to facilitate modelling of tagging data, allowing data integration from multiple bookmarking sources and the alignment of tagging ontologies. The core is based on Gruber's conceptual model. Core classes consist of concepts: "Tag", "Tagging", "Object", "Tagger", "Date", "Source", "Vote" and "Comment". Except for the "Comment" class, all other concepts are included in some or other of the models previously described. Comment is attached to an "Object" or "Tag" via "Tagging". Its claim about modelling the tagging structure so that it can integrate other tag sources is not very clear and it does not say what difference it has from the other models.

All of these conceptual models address the issues of representing individual tags, user and resources, while tagging as a process also gets almost unanimous agreement among all the models. Some models like SCOT and UTO are also focused towards collaborative tagging activities with folksonomy features. A detail study of tag model comparisons is reported in [Kim et al., 2008] which includes all the conceptual model excluding the Common Tag model and the Upper Tag Ontology.

4.8 Summary and Conclusion

This chapter gives an overview of the Social Web and its related features and practices, especially a core feature of the Social Web called “tagging” or “folksonomy”. We restricted our discussion to tagging, why people tag and how they tag. A major portion of the chapter is focused on tag semantics and various approaches to extract ontological concepts and relations from user tags. Finally we described some of the efforts made in the direction of representing tag and folksonomies using Semantic Web technology.

Section II

Core Chapters

Chapter 5

VOW: Lightweight Conceptual Video Model

The research question addressed in this chapter is "How to design a conceptual model to describe "Video" object on the Web encapsulating all the major facets of multimedia object description and yet keeping the model lightweight for both humans and machines?"

This work have been published in proceedings of the 8th International Conference on Information Technology and Telecommunication (IT&T, 2008) and the Fourth International Conference on Advances in Semantic Processing. SEMAPRO 2010.

5.1 Introduction

Machine readable metadata is envisioned as one of the core requirements of intelligent data access, sharing and reuse. Knowledge representation techniques are implicitly designed to represent the world in its textual format. A video object presents unique challenges in terms of representation; having three different layers of abstraction, i.e. video as the medium, visual or audio features used to communicate the semantics, and the depicted concepts itself. The richness of the media and the enormity of content volume on the web make it imperative that the multimedia content should be optimally described to increase the content interoperability and reusability. Various models and standards have been proposed in this regard, but for a seamless semantic level content management framework, the description should be machine understandable and not merely machine processable. In this scenario, Semantic Web technology can play its role to capture, describe, and manage multimedia semantics by making it formal and explicit. There are a few Semantic Web-based media ontologies proposed prior to this work; details of these efforts are described in Chapter 3.

In this chapter we will discuss the motivation for a video model, design issues and core concepts of our lightweight video model to support an integrated semantic annotation of video content in terms of depicted objects, events, people, time and place. The framework is designed to be lightweight (easy to understand for humans, yet computationally inexpensive to implement) while uncovering the

maximum gamut of semantics encoded in a video. In the present work, we propose a vocabulary of concepts for video content While designing the model, we tried our best to reuse existing vocabularies such as Dublin Core, FOAF, SIOC, and SKOS.

5.2 Motivation

Usefulness of lightweight ontologies compared to the full-fledged ontologies is well documented in literature [reference]. Moreover, usage of ontologies in the Web further supports our view that lightweight ontologies are easy to learn and adopt. Some of the widely used ontologies on the web are Dublin Core, Foaf, SKOS and Sioc ontologies to describe document, people,Concepts and online communities respectively. Despite the existence of many full-fledged highly expressive ontologies such as Cyc [reference], these lightweight ontologies are adopted rapidly by the community. Proposed VOW models falls in between the continuum of taxonomy to full fledged formal ontology. Choice of a lightweight ontology guided by three primary conditions described below.

5.2.1 Existing Usage Statistics

Usefulness and popularity of any existing schema can be inferred from it's real world usage . To get an overview of such statistics where media data especially videos are semantically described we explored the open linked data cloud statistics[40]. Fig 16 shows the most popular vocabulary in the semantic web data cloud. From the diagram it is clear that media data occupies relatively smaller share compared to the amount f media content available on the Web and whatever media related data available are coming from music domain (MusicBrainz, DBTune, BBC Music etc.), IMDB data, BBC programme with others. Out of 295 datasets contributing to the cloud, only two datasets use mpeg-7 vocabulary constituting less than 1% of the total cloud. Moreover the most popular schemas used are Dublin Core, FOAF, SKOS, Geo validates the hypothesis that lightweight schemas are widely adopted by the community compared to the full-fledged ontology with increased expressiveness.

[40] http://www4.wiwiss.fu-berlin.de/lodcloud/state/#terms

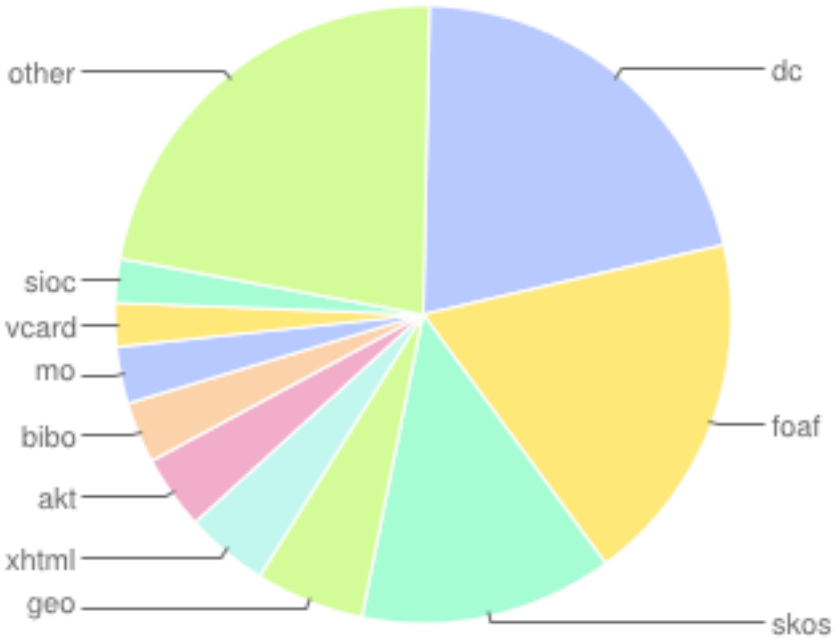


Figure 16 : Shows the distribution of most widely used vocabularies in LOD data cloud (Figure borrowed from http://www4.wiwiss.fu-berlin.de/lodcloud/state/#items).

5.2.2 Easy to Create

Lightweight ontologies are easy to create but risks the chances of disagreement on the concepts and relations, in order avoid this limitation we adopted a data driven approach to create the class names and identify the relations between them. By means "data driven" we meant to emphasise that the conceptual model is designed based on the available social web data created by users to describe their video on the web, usage pattern and user activities. As a result, a web video is considered as a social object used for interaction among various community members and groups through sharing, rating and comments. Being data driven, the model adopted an evolutionary and bottom-up approach as its design principle and started with basic attributes of a video object, its subject content, creators, rights and usage property. The advantage of this approach is directly influenced by the advantage of the social web e.g. "collective intelligence" in other words community agreement. On the other hand, formal ontologies may increase automation but are very complex to create and needs time and efforts of knowledge engineers for the process.

5.2.3 Ontology Structure Comparison

Structure of ontology in terms of its classes and object relationships reflects the amount of expressiveness it tries to capture while satisfying the major ontological requirements. Besides semantic interoperability, modularity and extensibility, a multimedia ontology needs to cover the

aspects of separation of concern, content decomposition and linking to domain knowledge. The table 1 shows a comparative analysis of the existing ontologies in terms of number of classes and their object properties while satisfying two major requirements of content structure semantics and linkage to domain knowledge.

Table 1: A comparison of existing multimedia vocabularies and standards

Ontology	Number of Classes	Number of Object properties	Content Structure semantics	Linking to domain
COMM	40	10	X	X
MPEG-7 MDS	69	38	X	X
MPEG-7 Tsinakari	420	175	X	X
M3O	126	129	X	X
Media Resource Ontology	14	55	X	X
MPEG-7 Rhizomik	814	580	X	X
VOW	**13**	**29**	**X**	**X**

5.3 Motivating Example

From a collection of conference and lecture videos, John has to select all the videos where the guest speaker speaks about the topics of the "Social Web and Semantic Web". There are nearly 100 different presentations and keynote videos available. John needs only videos from highly ranked speakers (ranked through user ratings or by means of a high number of views) between the years 2005-2010. At the same time, John wants to retrieve the different publications of each speaker from the web. In the present form, the video library has some basic metadata about each video such as title, date, presenter's name, conference name, conference venue and some keyword description. With a keyword search interface John gets almost all the videos where the textual description includes the query terms "Social Web" and "Semantic Web", and the list includes all the speakers on the topic. From the initial list John has to manually browse through all the videos in order to make sure he gets

only the videos satisfying his needs. Now he has a much smaller subset of videos but conference and lecture videos are normally of long duration , so he has to either watch all the videos for the entire time duration in order to get the segments where the speakers speak about his topic of interest, or manually seek through the timeline in order get the required segments. In both scenarios the entire process is time intensive. To get the publications John has to open another search session in some search engine or other digital library and formulate a new query; only if both the videos and publications were annotated with some unique topic identifiers a semantic search engine could pull them together and present them in an integrated view.

It would have been better if John could have given a query to retrieve only the segments of videos where the topic was discussed. Even though some video libraries are now creating some video chapters from the speech transcription, posing a segment query is still not possible. Moreover John's query has two more facets of "rank of the speaker" and "time constraint".

Such complex queries can be answered only if videos are described with some conceptual model combined with domain ontology; in this case a conference presentation ontology.

- To retrieve the segments, a video needs to be described with its structural components. It can be decomposed into smaller shots and then single or multiple shots can be combined as a semantic segment depicting semantically related topics. Each segment needs to be identified with a fragment URI describing the temporal location in the timeline.
- Time constraints can be filtered from the video document level descriptions.
- To make the videos interoperable across applications we can annotate the topic value with semantic concept URIs from DBpedia.
- Ranks of speakers is more application oriented and depends on users' perceptions. Yet this kind of trust attribute can be inferred from other existing data such as "number of views" or "ratings" etc.

Based on the above motivating example, we have proposed a lightweight video conceptual model which satisfies not only its document level description but also the content level description with its structural decomposition.

5.4 Design Issues

Formalising ontology engineering processes is tedious and time consuming. There are several studies providing guidelines from an ontology engineering perspective. McGuinness [2003] suggested to ontology building should start with an established model and extend it when necessary. The

Ontoclean project [Guarino and Welty, 2002] provides various guidelines to design an ontology but lacks a strategy to evaluate the steps. Fensel et al. [2002] focused on building a content driven ontology creation tool called "On-To-Knowledge" which is suitable for representing, share and access knowledge from online documents. Fernandez et al., [1998] suggested a methodology to create ontology from scratch, but each phase focuses on capturing the formalisation process which is complex and subjective. Ontometric [Tello and Perez, 2004] suggests measures to compare and select existing ontologies but it demands the user to specify the analysis requirements before doing any analysis on the ontological structure thereby making the more complex. Recently the NeON methodology[41] describes guidelines to reuse an ontology instead of building from scratch. While these approaches describe various steps to be followed while designing and developing an ontology, none of them are comprehensive enough to be treated as gold standard. As a result, much ontologies are either started from the scratch with individual efforts resulting in multiple models of the same domain or no formal model at all. We followed the step wise approach suggested by Noy and. McGuinness [2001] to create the conceptual model.

- *Domain and Scope of the Ontology*: At this stage the coverage and usage of the model is determined. It involves answering questions such as what type of queries the model can answer.
- *Reusing Existing Ontology*: To ensure maximum reusability and interoperability, we searched for existing ontologies and their concepts and checked if we can utilise them without creating a new concept.
- *Term Definition*: This stage demands that one outlines various terms for classes. The terms should reflect its intended usage in the model.
- *Class Hierarchy*: Class hierarchy is one of the most painstaking processes in a model engineering process and will influence the complexities of a model. Consistency of the class structure has to be continuously verified by an expert or with consistency checking tools.
- *Property Definition*: After the class definition we need to define their attributes and relationship to other classes.
- *Restriction on Domain and Range Value*: Finally we outline the restrictions to put on domain and range values on various class attributes.
- *Instance Creation* is the last stage of the ontology engineering process when individual instances of various classes are created to create a knowledgebase.

For designing the proposed core model and implementing the requirements we carefully focussed on a few design principles. While satisfying various ontological requirements as described in Hunter and

[41] http://oa.upm.es/3879/

Armstrong [1999], it should adhere to the principle of "Keeping the model short and simple (KISS)": that means there has to be less hierarchical nesting of classes.

The first and foremost design issue is to make the model *Lightweight:* Wider adoption of various ontologies largely depends on its complexities in terms of hierarchies of its classes and properties. This hypothesis is supported by the wide adaptation of lightweight ontologies such as FOAF, SIOC, and Dublin Core on the web [Fergal,2010]. Lightweight schemas are easy to understand and quick to learn by non-experts.

The second design issue we are concerned with is separation of the low-level content description from the semantic content description at the level of different structural components such as shot, frame and image region.

The third design issue is based on the notion that semantics of media is determined by its contextual relationship. One such source is social context, hence the model should reflect the video object as a *social object* linked to other social situations and interaction patterns. As a consequence it should be able to reflect the dynamics caused by social actions in an online community. Such social actions may be annotation by users other than the creator of the video.

5.4.1 Fragment Identification

Media segments should be localised, whether it is spatial region in an image or temporal region (shot and frames) in a video. Until recently there was no agreed-upon framework to represent media segments. Localisation (temporal and spatial) of media is one of the crucial requirements for representing and querying multimedia content. Recently W3C[42] started a working group to define a media fragment URI agnostic to media format. It is specified in the dimension of temporal, spatial, track and named fragment. As described in the W3C media fragment URI specification document, "the aim is to enhance the web infrastructure for supporting the query and retrieval of subparts of time-based web resources (audio, video), as well as the automated processing of such subparts for reuse. Example uses are the sharing of such fragment URIs with friends via email, sharing of a TV programme segment through mobile, the automated creation of such fragment URIs in a search engine interface, or the annotation of media fragments with RDF." An example of such fragment URI is: "http://www.example.org/video.ogv#t=60,100".

[42] http://www.w3.org/TR/media-frags/

5.5 Related Studies

Various multimedia ontologies both Semantic Web-based, and XML-based, ontologies are discussed in Chapter 3 in greater detail. To recapture the multimedia ontology scenario, we will briefly discuss the most representative works in this section.

Work on a Semantic Web-enabled multimedia ontology started when researchers realised the problems with a lack of formal semantics and interoperability in the MPEG-7 standard. In order to get the best of both approaches, namely the description flexibility of MPEG-7 and the formal semantics of Semantic Web, researchers started to formalise a multimedia ontology complying with MPEG7. Hunter [2001] started the first initiative to formalise parts (structural and localisation tools and visual descriptors) of MPEG7 in RDFS later, integrated with the ABC upper ontology, while Tsinaraki et al. [2004] converted the entire MPEG7 to an OWL DL ontology. Garcia and Celma [2005] described the whole MPEG7 ontology by means of automatic mapping from XSD to OWL. All the above studies were attempts to produce one-to-one transformation of MPEG7 modules. But the fundamental challenge of interoperability remained unresolved, leading to the design of a more Semantic Web oriented ontology called COMM [Arndt et al., 2007] based on DOLCE upper ontology. Ontologies focusing on MPEG7 compliance wanted a full coverage of the standard with less semantics while Semantic Web led studies focused more on reasoning at the cost of flexibility. The tussle between the two groups resulted in low uptake of any standards for the ever increasing amount of multimedia content.

Recently W3C has recommended a media annotation ontology which is a collection of frequently used media properties from various existing vocabularies.

With the presence of these ontologies, the question arises, why another model? We proposed our model keeping in mind that, despite some serious efforts in the modelling of multimedia object and content, user uptake is rather slow or not visible. Secondly our proposed model is extremely lightweight and focused towards web videos.

5.6 Abstract Conceptualisation

Our objective is to create an enriched semantic description of the media object with all its depicted content and relevant contextual properties. This process of semantic augmentation will not only promote integration and interlinking of the media with the rest of the web of data but also, will facilitate complex and structured query answering leading to more relevant knowledge discovery. Our first step to meet the objective is to design an all-encompassing multimedia object model for data

representation. This thesis will mostly focus on video modelling as opposed to other media objects such as image and audio. The description of the model should cover the following concepts,

- Video: A multimedia object/document extended from the root class “MediaObject”.
- Segment: A subclass of the generic MediaObject that can be specialised further into temporal, spatial and spatio-temporal segments.
- Annotation: A unique user-generated object attached to a media segment with textual description.
- Location: A spatial concept to indicate the geographical location.
- Person: Describes individual and answers the who question
- Time: A temporal entity used to describe the temporal attributes such as duration.
- Event: Describes various events captured within the media as well as the events depicted through a timeline.
- Concept: describes abstract concepts and objects in the real world.
- User: Describes an online identity, be it of a person, group or organisation.
- Site: A virtual container of the media object.
- Device: A capturing device, whether a phone, digital camera or any other device.
- Descriptor: An abstract class for low-level feature descriptions.

This model will serve as a generic layer and can be coupled with the domain knowledge for in-depth semantic description, e.g. a video can be described as depicting the event called “DERI research talk”, but with no details about the event itself. All the event details can be queried through an“Academic Event ontology”. The end result of the model will be an integrated and interlinked semantic eco-space for media object.

Before we go into the details of the model, the section next will briefly discuss the reused schemas and their specific concepts. There are some well-established ontologies to describe people, documents, time, events, and location which we selected for use in our model.

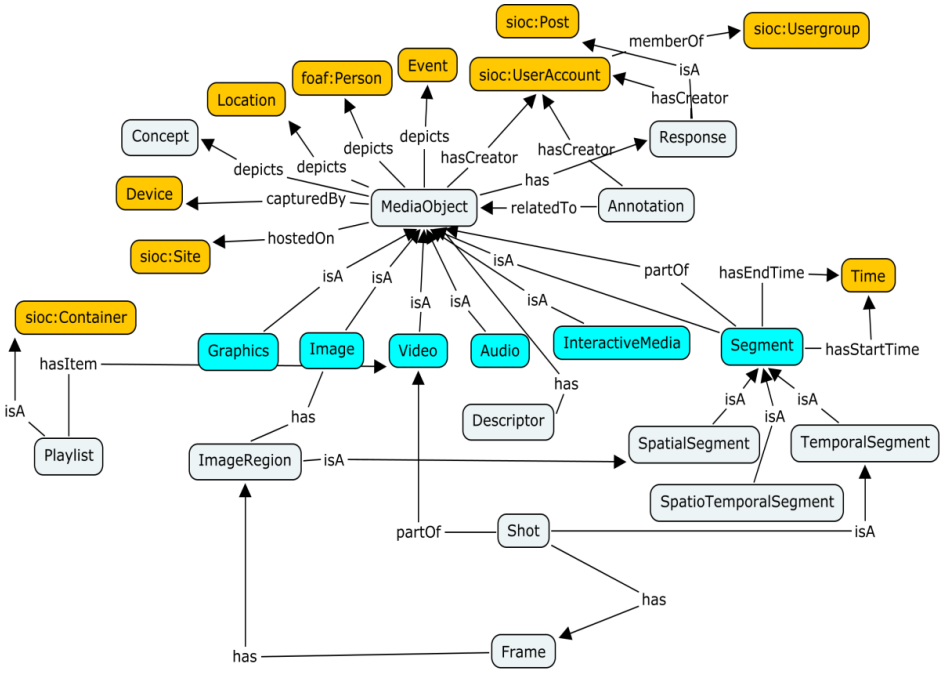


Figure 17: Core classes and their attributes.

5.7 Description of Reused Schemas and Classes

This section describes various existing schemas reused as part of the proposed conceptual model of web video. Figure 17 above illustrates some core classes and their relationships.

5.7.1 FOAF Ontology

The Friend of a Friend (FOAF) ontology is so far one of the most popular and widely used ontologies on the web today. It aims to create a web of machine readable pages describing people, social links between them and the things they create and do. We used the foaf:Person and foaf:Agent concept to represent various facts involving people; properties such as “depictedPerson” takes a foaf:Person as its value, while a video also links to a foaf:Person through the foaf:maker property.

5.7.2 Geo Ontology

This ontology describes places as geographical features with coordinates such as longitude, latitude. This Geo ontology extends the well-established World Geodetic System 1994 (wgs84) which is now widely used by various applications and tools to describe geospatial properties, hence it is a good choice for maximum interoperability. The model uses a recorded location as a “SpatialThing” class with two attributes of longitude and latitude.

5.7.3 Time Ontology

The OWL Time ontology defines two temporal concepts i.e. Interval and Instant. Both of these classes are subclasses of the class called “TemporalEntity”. This model aimed to capture the temporal content and properties. In our model a segment can be described with the properties “hasBegining” and “hasEnd” where the range will be an instance of “time:Instant”.

5.7.4 SIOC Ontology

Semantically Interlinked Online Community (SIOC) is an initiative to enable the integration of online community information. It uses Semantic Web technology to represent rich data from the Social Web. Recently it has got a significant boost in terms of user and community adoption for both commercial and open applications. We used sioc:Post, sioc:UserAccount, sioc:topic and sioc:Container to describe various social aspects and their influence for media content understanding.

5.7.5 SKOS Ontology

Simple Knowledge Organisation System (SKOS) is a common data model for sharing and linking knowledge systems such as thesauri and classifications via the web. A topic or subject can be linked to a skos:Concept. The value of the property “depictsConcept” is a skos:Concept.

5.7.6 W3C Media Ontology

Ontology for Media Resource from W3C, is a set of core properties to describe any media resources on the web. These properties span across content, technical, rights and other global attributes which we adopted for our study.

5.7.7 Event Model

Linear media objects, specifically videos, are mostly (though not all) event centric. The representation and recognition of events in video is important for semantic content-based browsing and retrieval. An event may be a simple event or a complex one through the combination of simple events. A few event models exist in the literature, each developed while keeping an application domain in view. All these event models agree upon some fundamental components such as time, space, agent, change of state, etc. The question we asked was: can we adopt one of them? The model should be able to capture all the required important facets of an event as depicted in a video stream. Recently published research

suggested a core model LODE[43]: which is an attempt to capture the common properties of all major event models. Therefore we decided to start with LODE. Before adopting any model, we need that to discuss the requirements need to be satisfied by any event, model. It covers some fundamental aspects of any event whether there be past historical events such as a war, natural disaster, or planned future events such as concerts or conference talks. These fundamental facets of an event are described below;

1. **Event and Time:** The link between Event and time is established through three different properties 1) duration 2) hasStartTime and 3) hasEndTime, "duration" takes an instance of OWL:Time object as its range value, while the "hasStartTime" and "hasEndTime" properties accept xsd:dateTime literals. The reason for two different kinds of temporal attributes is to enjoy both the simplicity of xsd:dateTime literal and the reasoning power of an OWL:Time object where the instances are the time intervals.

2. **Event and Place:** Another fundamental characteristic of an event is the place or location where the event takes place. Event and place can also be linked via two different properties e.g. "hasLocation" and "hasRegion". The attribute "hasLocation" accepts "geo:SpatialThing" as its value whereas "hasRegion" may have a string literal due to the absence of precise geo-coordinates.

3. **Event and Agent:** The relationship between an Event and Agent is captured with "particiatingAgent" property, used to describe multiple roles played by one or more agents before and during the course of an event. The agent may be a creator, organiser or sponsor of the event. The Agent class is further specialised through the "hasRole" property.

4. **Event and Object:** An event should able to describe the participating objects as the factors influenced during the event, which are described via the property "participatingObject". A talk event may involve a related PDF document or a video recording.

5. **Event Structure:** An event may be a simple event or a complex event which in turn consists of many sub events. These sub events occur sequentially or in a temporally overlapping space. This structural aspect of an event needs to be captured in the model through "hasSubevent" property.

6. **Casual Relation**: The creation of one event may affect and lead to another event as a consequences, this cause and effect relationship should be formalised in any event model.

[43] http://linkedevents.org/ontology/

7. **Event Documentation:** An event description should be able to identify its documentation aspects. It must be recorded, published or mentioned somewhere in some format. An event of a musical concert is captured through recorded video, published as news and sold in DVDs. These physical realisations are an integral part of any event discovery mechanism. In the context of multimedia data modelling, this link between the content and the event is essential and expressed through the property "vow:documentedIn", whose range will be a media URL or a web page location.

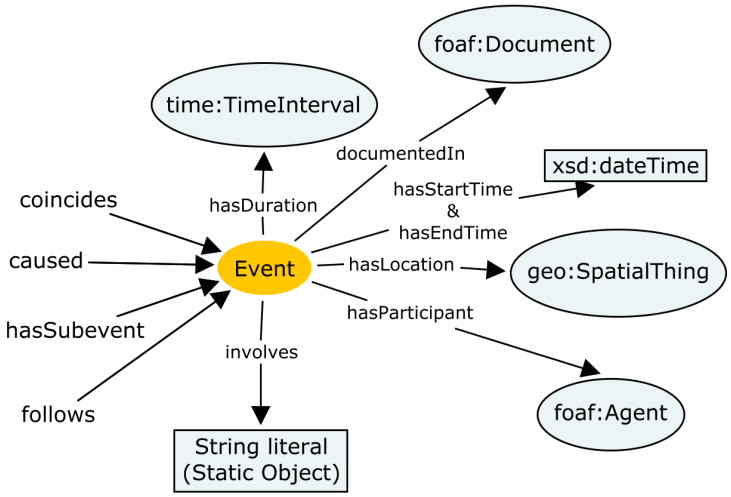


Figure 18: Event Model.

We have followed the OWL Event ontology[44] with few extensions including the documentation aspect of it.

For the rest of the important concepts such as video, image, etc. we could not locate a vocabulary satisfying all our needs so we decided to design a lightweight conceptual model to describe a video on the web.

5.8 Core Conceptualization

The core video model we are going to describe is aimed at capturing video information aggregated over three contributing information spaces (Figure 19) i.e. content, context and document (media, technical). However, it should be noted that the focus of this thesis is the content and context descriptions of the model, as the editorial or document level properties which are mostly reused from different existing media ontologies (Dublin Core, W3C) which are well understood.

[44] http://motools.sourceforge.net/event/event.html

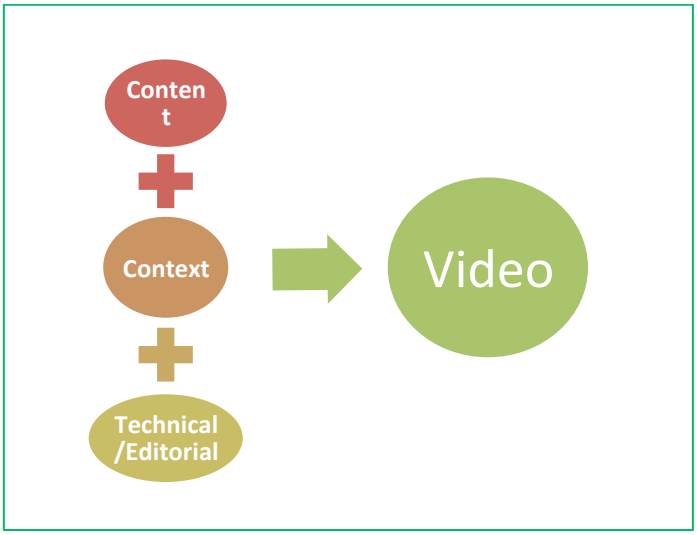


Figure 19: Contributing information spaces.

5.8.1 Content Space

This space carries all the classes and properties concerned with the structural decomposition of a media object as well as the visual and audio content signal descriptions. MediaObject is the super class of Video, Audio, Image and Segment class.

Descriptor is a superclass of visual, aural, texture, shape, motion and other low-level descriptors. Visual content description can be expressed through MPEG-7 descriptors such as edge histogram, color structure descriptor etc. It is attached to a segment or image region via the "vow:hasDescriptor" property. This simple class contains two properties "descriptorName" and "descriptorValue", however it can be further extended to include an input parameter and their possible values if necessary. One example of the "Descriptor" instance is given below:

```
ImageRegion_1 vow:hasDescriptor        descriptor_1

        descriptor_1 rdf:type vow:Descriptor

        descriptor_1 vow:descriptorName "Edge Histogram"

        descriptor_1 vow:descriptorValue "4,5,2,0,0,3,1,2,..6"
```

Segment is a spatio-temporal entity extended from the MediaObject class, and links to other MediaObjects through the "part_of" relationship. There can be a spatial segment such as ImageRegion or a temporal segment such as Shot.

Shot is an uninterrupted image sequence of a camera recording and hence treated as an atomic unit for video analysis. It is a temporal decomposition of a video with non-uniform length. It is a subclass of Segment and is linked to a video by a "part_of" relationship. A Shot can be automatically segmented out of a video using various algorithms. Each Shot can be annotated at multiple levels such as audio, visual and textual content.

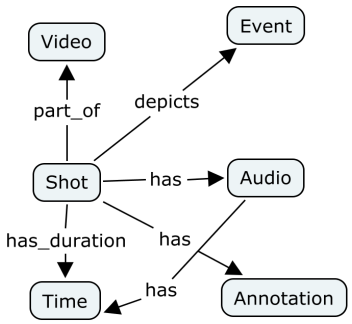


Figure 20: Shot details.

Frame is also a subclass of Segment and holds a "part_of" relationship with Shot. One or more Frames are automatically selected to represent a Shot depending on various criteria. A Frame can also be described with an ImageRegion which may be the entire Frame or a small part of it.

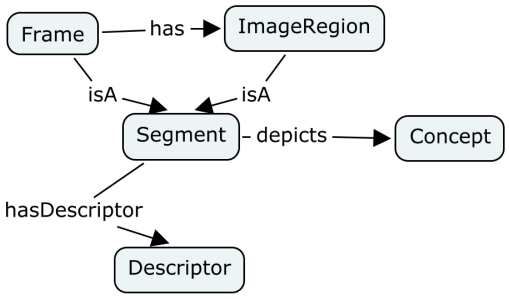


Figure 21: Image Region details.

ImageRegion is spatial region within an image which can be defined with a bounding box and its four co-ordinates. A MPEG 7 compliant descriptor can be used to describe an ImageRegion with its low-level features (Figure 20).

5.8.2 Context Space

Annotation is considered as a subclass of sioc:Post. It can be attached to an entire media object or a part of it via the "relatedTo" predicate. It is created by a sioc:UserAccount and carries a text description via a "dc:description" attribute. A subtitle or a transcript can be attached to the media

through this property. The relation between a media item and annotation is qualified with a creator of the annotation, thereby providing a provenance factor and restricting queries by specific annotators.

```
Video(v1) vow:contains Segment(s1)
        .s1 vow:hasAnnotation Annotation(a1)
                .a1 vow:hasCreator sioc:UserAccount(ua1)
                .a1 dc:description "Annotated text or subtitle".
```

Sioc:UserAccount is a virtual identity of a member in online community. It is connected through the "hasCreator" property to a Response, MediaObject, Group, Playlist and Annotation class. A user account can be linked to a foaf:Person through the property sioc:account_of.

Response (either text or video) is a feedback or a comment on a media object or a part of it added by various user accounts. A Response can be attached to an entire video or to a segment of a video with time stamps. It is a subclass of sioc:Post created by a sioc:UserAccount.

Set/Gallery/Playlist/Thematic_Group is a group of media objects sharing a common theme. For example, in Flickr photos are grouped based on their content, pohtographic properties or based on events. A Set is created by a user. It may have a title, descriptions and some user tags. Members of a set are media objects.

5.8.3 Important properties

depicts (Concept/Event/Person)

The property "vow:depicts" is a general property to connect between media segments and semantic concepts. This property can have four specific sub properties i.e. vow:depictsPerson, vow:depictsEvent, vow:depictsLocation and vow:depictsConcept. This attribute of MediaObject is one of the major connecting links from media content to the semantic content. The range of objects includes event, people, location, object or any abstract concept identified with its URI.

contains/part_of

The "vow:contains" property denotes the containment relationship between various media components, e.g. a video can contain a shot and a shot can contain a frame which in turn contains image regions. The inverse property of this property is "part_of".

relatedTo/has Annotation

A “MediaObject” is related to a subtitle or caption or any user annotation through the “relatedTo” and “hasAnnotation” properties whose range value is from “MediaObject” and “Annotation” respectively. Annotation has a “hasCreator” property to reflect the provenance aspect.

hasDescriptor

A segment is linked to its low-level content descriptions through the property “hasDescriptor”. The range of the property is an instance of a Descriptor class. A descriptor can be a type of visual descriptor or aural descriptor depending on the type of segment. The descriptors can be further specialised into colour, shape, and motion descriptors.

5.9 VOW Implementation

The work is envisioned to be used as a model to describe video documents published or to be published on the Web created both by professionals and amateurs. Comprehensive semantic description of video is non-trivial and rare on the Web. According to the usage statistics from the Linked Open Data cloud, multimedia related data includes sources such as MusicBrainz, DBTune, Lastfm and IMDB etc. A small portion (.68%) of the entire data is described with mpeg-7 schema along with other vocabularies. This clearly shows that user videos and videos on the social web are mostly ignored while the amount of video content is rapidly outnumbering the other content medium.

This proposed model and approach of populating the model is a ground-up approach of semantic modelling and description of a video. The model can be used individuals as well as organisations to describe their collections with minimal efforts. The scope of the model is to describe a video with its essential properties covering technical, low-level content structure, usage and subject depiction. Any description of domain knowledge can be linked through various properties such as “hasTag”,”hasCategory”, “depictsEvent” and “depictsConcept” whose range of values include semantic concept form other ontologies or thesauruses. The generic model is well suited to all kinds of video(videos from broadcasting house, an advertisement video or a TedTalk video) and can be easily extended and glued to domain ontologies to describe a domain specific knowledge e.g. full length movie or a sports video both at the document level as well as at the segment level. It also provides scope for time-based annotation at the shot and frame level. Detecting shots boundary and selecting representative frames are implementation details need to be addressed at the system level. However, the task of extracting video shots and frames has further simplified and accessible with the emergence of new standard HTML 5 which allows manipulation of video with client side scripting.The annotation interface supporting the proposed model is described below.

Producing an all-encompassing vocabulary covering every aspect of digital media from all possible domains is non-trivial. This is the precise reason for multiple proprietary schemas used to describe the video collections by many audio-visual archives, which makes semantic interoperability even more difficult. There are two possible scenarios of using the model to describe a video to be published on the web or videos already published on the web.

1. Personal Video: A simple form based tool (e.g. prototype) can be provided to annotate a user video with basic details such as title, descriptions, date etc. The system can then take this description and extract related information from various sources to enrich the video. It also keeps track of the user activities to provide a set of personalised tags for annotation.
2. Archive video collections: Video collection from the archive can map their basic video details with the model classes and attributes such as title, duration, format, frame size, keywords etc. and the system can extract related information and semantic entities from the web for content enrichment.

However there will be cases where the annotation accuracy and quality is of prime importance therefore, Social Web alone cannot be the sole source of knowledge due to its inherent noise and uncertainities. In such cases, expert annotation can be combined with the system recommendation to achieve the desired result.

5.9.1 Annotation tool

Annotation interface is part of the prototype described in chapter 9.This prototype starts with a search interface where the user can search for videos with keywords. The videos are presented with its existing details (Annotation interface) as well as recommended tags from multiple contexts. The user has a chance to select the tags if appropriate and describes the video segments with time stamps. The system then annotates the video and its segments using the proposed conceptual model. The annotation interface (Figure 23) is simplified as much as possible. It provides annotation field with the labels, similar to the classes of VOW.

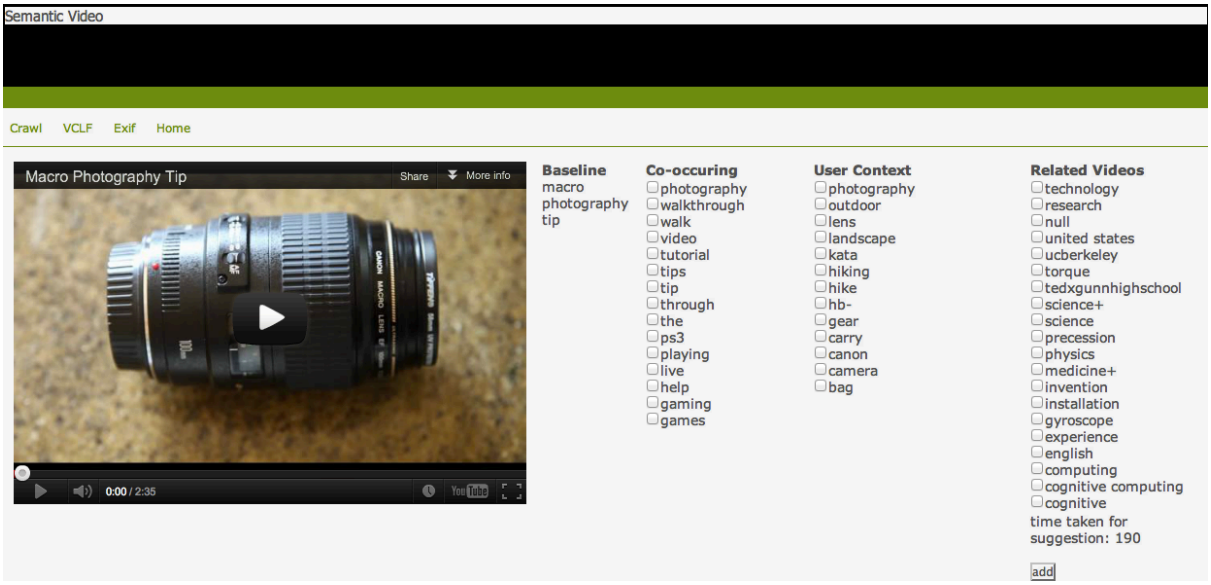


Figure 22: Annotation interface

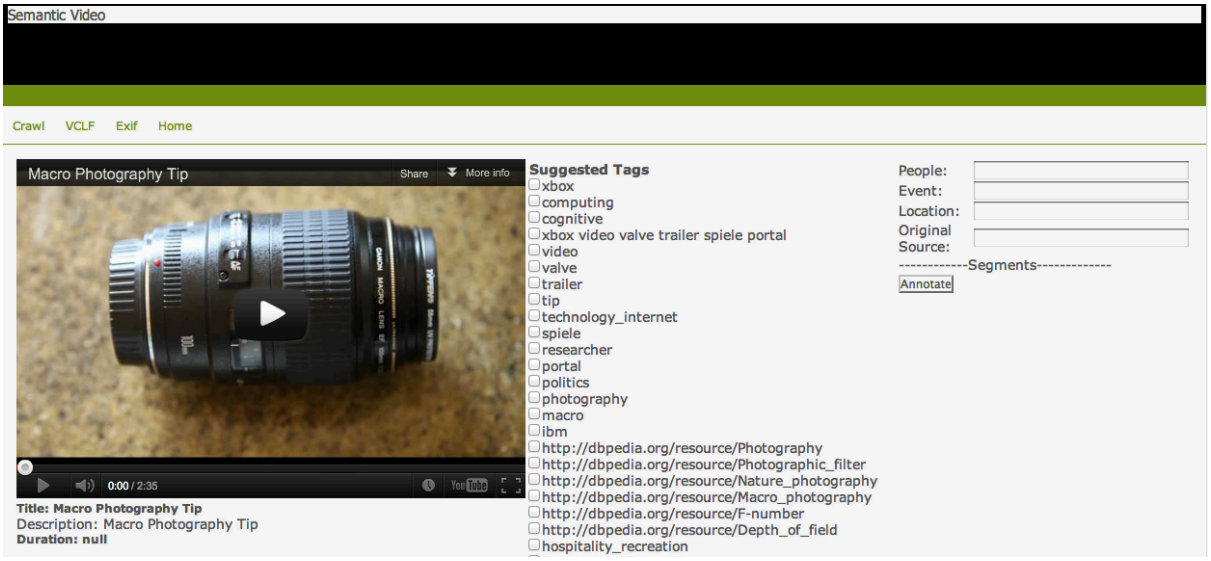


Figure 23: Interface showing the tags suggested from multiple sources.

Figure 23 shows the annotation interface with fields for various concepts and segments. The video is also presented with multiple temporal segments with start and end time. Though, the segmentation follows a simple time based approach for calculating number of segments, a more efficient interface can be created based on shot detection technique (not implemented in this prototype). The second figure (figure 22) shows the recommended tags for annotation. there are overlapping suggestion which is aggregated and ranked on the backend once the user selects the appropriate concepts to annotate and finally the last interface shows the generated annotation but in property value pairs

where as a RDF file is generated and indexed on the back. The generated RDF is re-indexed for the semantic search module described later.

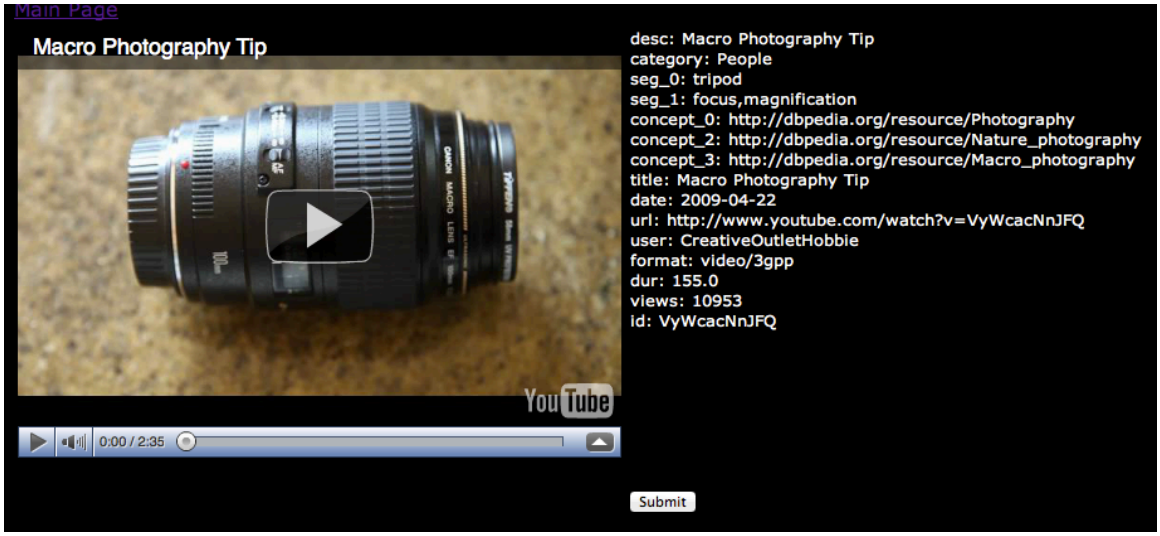


Figure 24: Interface showing the annotation field values entered by the user for a video

5.9.2 Semantic Search Interface with Example

Search module consists of two types of search, 1) keyword and 2) semantic search (figure 54 and 26). The keyword search presents its results in a faceted interface exposing the underlying data and while structured queries can be posed through a sparql interface. This interface presents few examples of structured query in order to give an idea to form complex queries. However, it is my understanding that pure sparql interface of little or no use to users other than experts. We need to develop a more user-friendly interface that will guide to understand the underlying data structure and build the query in a visual way in order to maximise the usefulness of any semantically structured data.

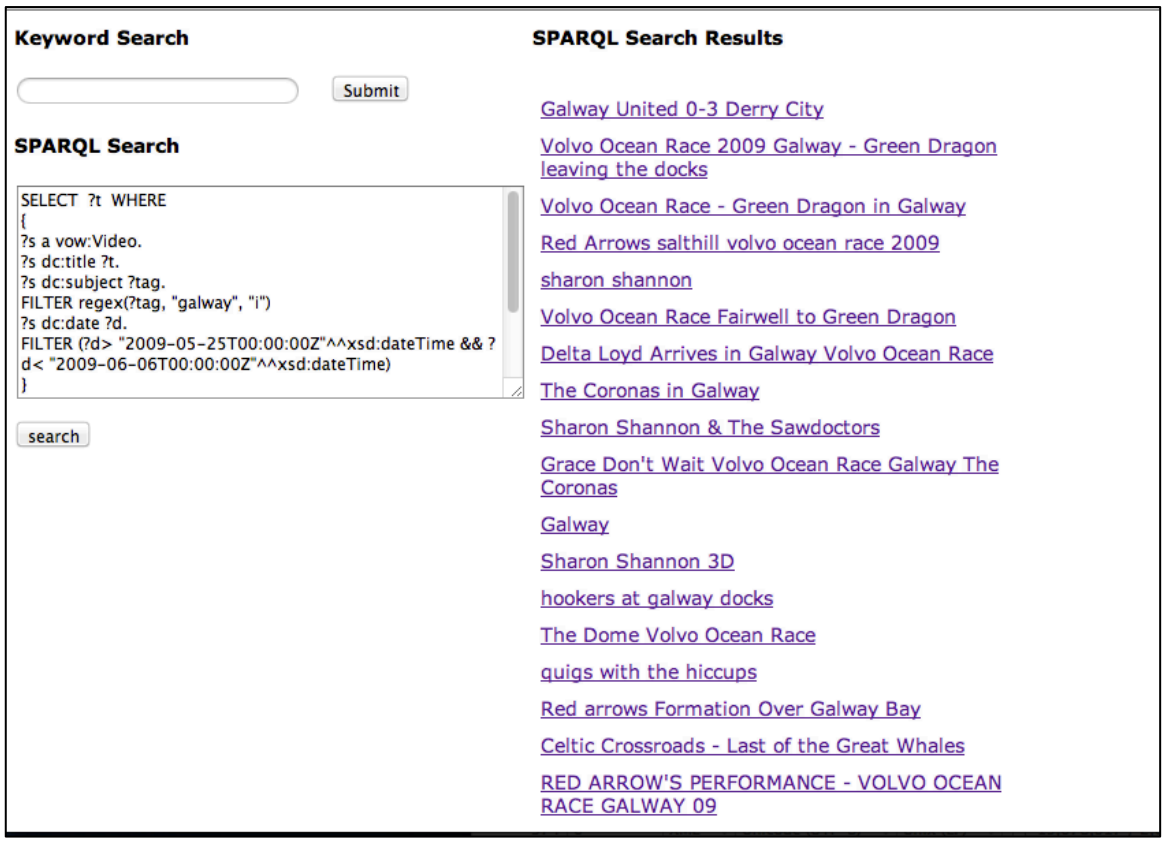


Figure 25: Semantic search result interface.

5.10 Requirement Satisfaction Revisited

We will discuss here whether the multimedia ontology requirements mentioned in the work by Hunter and Armstrong [1999] are satisfied with the proposed model.

Interoperability: Interoperability among existing vocabularies is a core requirement for wide acceptance of any ontology. The ontology has been formalised in OWL-DL, which makes it available for interoperation with other widely used schemas. This model has extensively reused various classes from other vocabularies wherever necessary and sometimes adds redundancies to cover a wider community, for example, a user-generated tag can be described in various ways: using *dc:subject* (Dublin Core) and as a *sioc:topic* (SIOC).

Ease of Use: The proposed model mainly focuses on maximising the automatic extraction of metadata either from the content module or from the context module. Users on the web produce most of the social context metadata related to a video over time. Moreover the terms used to describe the model concepts are mostly self-explanatory.

Representation of Content Structure: Our proposed model describes the structure of the media at different abstraction levels, both temporally as well as spatially. Shots and Frames are temporally segmented, and ImageRegion describes content in spatial dimensions with proper fragment identification.

Modularity and Extensibility: Our proposed model is an aggregation of different modules such as Time, Location, and Event. Most of the entities are described as a basic class with scope for further extension. The model can be extended with media types such as 3D models, graphics, and sketches as specialised classes of media object.

Separation of Concern: Content representation and separation of concern is achieved through predicates defined between spatial or temporal segments and their depicted semantic concepts.

Multimedia Reasoning: An ontological model should describe the reasoning scope. Multimedia reasoning includes reasoning at both the content and semantic levels. Neuman et al. [2006] described that an aggregated composition of parts constrained with spatial and temporal relations will facilitate concept reasoning such as object configuration, occurrences, events and scene interpretation. Our proposed model provides ample scope for the extraction and description of such spatial and temporal segments that can be used for reasoning.

We presented the core aspects of the model which includes both low-level feature description and semantic annotation. It satisfies structure decomposition of a media object (video) both in temporal and spatial dimensions. The information realisation description is achieved through global description of the media in terms of its type, format, length size, etc.

5.11 Revisiting the Research Question

This chapter describes a proposed conceptual model for Video on the Web (VOW) encompassing three major information spaces i.e. content, document and context. It captures all crucial aspect of a video document describing its subject content, media properties as well as its relationships to various contextual sources. It satisfies the major requirements of a multimedia ontology including separation of concern, modularity etc. Since the model is about the Web videos the description scope also extends into the social context space where a video is shared, interacted with through user given tags and comments, thematic groups, playlists to name a few online social activities putting the video as the target object.

The second research objective of this work is to keep the model lightweight for both machines and humans. Since the model is drawn from the social web content and online activities surrounding video objects, the classes and its relations are largely based on the existing practices making it easier to understand for annotation and querying tasks. Whereas the document properties are borrowed from existing and widely used vocabularies. The model has only 13 classes and 29 relationships.

Chapter 6

Visual Concept Learning Based on Social Knowledge

In this chapter, we argue that cross media data can be creatively explored and leveraged for mutual benefits for e.g. content of images can be used to understand the contents of a video . We propose a cross media .visual concept learning framework to annotate video keyframes from Flickr image tags.

This work have been published in proceedings of the 8th International Conference on Information Technology and Telecommunication (IT&T, 2008).

6.1 Introduction

Visual concepts are used as one of he basic units to describe depicted semantic content of a visual object (image and video). Mining and automatic detection of visual concepts in image and video has been the subject of intense research for multimedia and computer vision community for years. Decades of core and interdisciplinary research have studied this problem from multiple perspectives with multiple domain applications. A more detailed description of such efforts is in Chapter 2. With the emergence of Web 2.0, multimedia content creation and sharing has become much easier. Users are willingly annotating and uploading photos and videos to many content sharing sites such as Flickr and YouTube. Apart from their primary objective of content organisation and discovery, social tags are now being used innovatively by various researcher communities to address many long standing research problems. In one such attempt, in this thesis ,we leveraged the user created images and their textual annotations from the photo sharing site Flickr to learn various visual concepts depicted in video frames.

Unlike images, video content is more complex and computationally expensive to process. Any concept learning in video primarily depends on various combinations of computer vision signal processing, statistical and machine learning algorithms. Heavy dependence on manually annotated training data limits its applicability to the potentially unlimited array of concepts that exists in real world. It cannot be expected to learn every concept for an automatic concept annotation and tag prediction. Moreover unlimited visual variations of the same concept are observed on the web whereas existing methods are trained on a relatively controlled and professionally-created content set such as news videos, documentaries or Corel photo sets. The scarcity of training samples and availability of user tags opens new research alternatives to learn concepts. Authors in [Sigurbjörnsson et al., 2008] suggested a method for image tag recommendation based on tag co-occurrence, but they did not consider any visual feature.

In this chapter we present a framework to learn concepts for video annotation based on visually similar images found in Flickr[45]. Our approach makes use of user tags associated with these images as candidate concepts which are then filtered, aggregated and ranked for final recommendation. Experimental result shows the potential of the approach for conceptual video annotation in a much lightweight fashion compared to existing trends.

6.1.1 Motivation

With large scale media content being uploaded and shared publicly domain constantly, it is difficult to manage, organise and annotate with pure content-based approaches. On the other hand, huge amount of images and videos are being freely annotated and published on the web, and these can be leveraged to learn various visual concepts and facilitate semi-automatic media content annotation.

6.2 Tag Learning Approach

With the above motivation, we made two intuitive assumptions: 1) Visually similar images share similar semantics; and 2) When two similar images carry the same tag created by different users, and then the tags are likely to be more relevant to describe the content.

Based on the above assumptions we proposed a neighbour voting tag propagation strategy unlabelled images (key frames of a video). Unlike many other studies [collective knowledge], our approach does not require an initial set of tags to start with, rather we are opting for a bimodal learning framework exploring the visual content and using user created tags to learn a concept.

[45] This work was presented and published in IT&T prceedings.

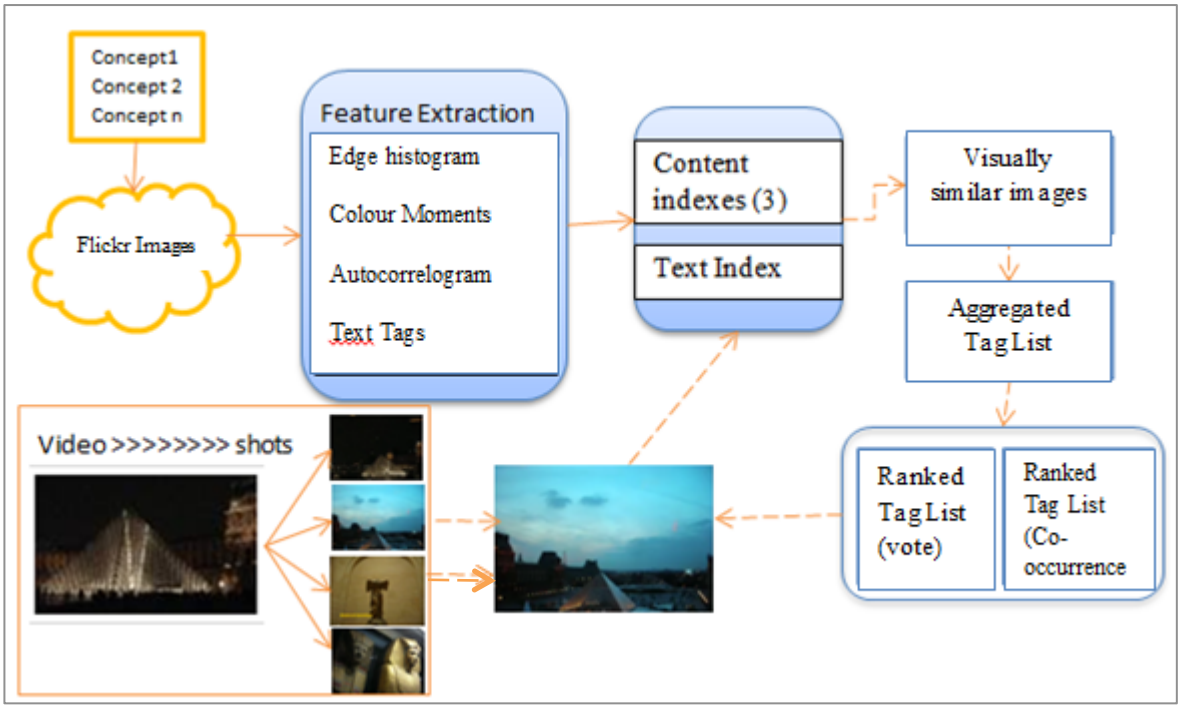


Figure 26: Work modules for CLF.

6.2.1 Visual Concept (Tag) Learning Framework (VCLF)

Figure 27 shows an overview of the system. Given an image (video frame) with no tags or any textual description, an ordered list of candidate tags is generated based on visually similar nearest neighbour images. This list then becomes the input for an aggregation and voting process which finally produces the ranked list of n suggested concepts or tags for the query image. Consider the example given in Figure 28, where the input image needs some tags to describe it and gets five visually similar images are retreived with their tags. The five nearest images constitute a set of candidate tags used towards the final list. An aggregated and then finally ranked (majority voting) tag list shows the top 10 tags for suggestion. The result is clearly not very descriptive as the tags in 5th, 6th and 7th position are of little use whilethe tag in 8th and 10th position should move up in the list.

Original Image	**Nearest Neighbour**	**Tags from Nearest Neighbour**	**Aggregated Tags**	**Ranked Concepts (Voting)**
		Paris summer 2005 august holiday Louvre	paris summer 2005	Paris louvre night pyramid holiday august

		night	august holiday louvre night Europe wonderful shot pyramid reflection travel architecture palace 2006 light	2005 europe travel reflection architecture museum palace
		Louvre Paris Wonderful shot pyramid night		
		Louvre museum night Paris Europe palace		
		Reflection pyramid Louvre museum Paris travel		
		Paris architecture night Light 2006		

Figure 27: Shows the progress (from left column to right) from an input image to the output recommended tags.

6.2.2 Tag Aggregation and Ranking

When we get the list of candidate tags from all the results, an aggregation and ranking step is required for the final recommended tags. To rank the tags we adopted two approaches 1) majority voting, and 2) concept clustering based on tag co-occurrence.

Majority Voting: Frequency based majority voting computes the tag frequency from each result and recommends the top *m* tags as the final list; *m* in our case is 10. The vote strategy computes a score for tag c ∈ C where a vote for c is cast by means of its inclusion in the result list.

$$voteScore(c) = \begin{cases} 1, & if\ c \in C \\ 0, & otherwise \end{cases}$$

A list of recommended tags *R* is obtained by sorting the tags by frequency. A relevance score for a tag, c_r is computed by adding the number of votes in each of the result sets:

$$relevantScore(c) = \sum_{r=R} voteScore(c)$$

Rank by Co-occurrence: The second ranking approach we adopted is the rank by co-occurrence. Unlike the suggested approach in [Sigurbjornsson and Zwol, 2008] where each tag was individually extended with its co-occurring tags, we computed an intra-list co-occurrence calculation. An intra-list co-occurrence calculation is computed over the candidate tag list *C*. There is no effort to add extra tags beyond the candidate tag list; rather it clusters the tags within the list based on their co-occurrence value. Finally the cluster is ranked based on the aggregated frequency of the individual tags within the cluster. The table (2) summarises one such intra list co-occurrence example and the resulting conceptual grouping of tags.

Table 2: Example showing the result of intra-list co-occurrence tag rank

Candidate List	**grouped by Co-occurrence**	**Ranked tags**
Paris	Paris, Europe - 6	Louvre, museum, pyramid
Summer	light, reflection – 2	Paris, Europe
2005	Louvre, museum, pyramid - 8	night
August	travel, holiday, summer - 3	travel, holiday, summer
Holiday	palace, architecture - 2	light, reflection
Louvre	night - 4	palace, architecture
Night	2006 - 1	2006
Europe	2005 - 1	2007
Wonderful shot	August - 1	August
Pyramid	Wonderful shot - 1	Wonderful shot
Reflection		
Travel		
Architecture		
Palace		
2006		
light		

6.3 Experimental Setup

In the following experimental setup we evaluate our neighbour voting based tag recommendation for unlabelled images / video key frames. The task is defined as follows: given an image, the system will retrieve a ranked list of highly-probable tags describing the image content. The next two sections (6.3.1 and 6.3.2) will describe the training and test phase while the evaluation will be described in result section.

6.3.1 Training

For the training phase we collected 10 concepts from the LSCOM ontology. LSCOM is a visual concept ontology developed by [Naphade et al., 2006] for the video concept detection research effort. It consists of 1000 visual concepts (mostly abstract upper level concepts) and is mainly used by the participants of TRECVID workshop. The lists of concepts are described below. We used these concepts to download images from Flickr. Each downloaded image is tagged with the concept and other user-generated tags. For the study we collected 5000 images with their metadata (author and tags) covering all 10 concepts and their associated terms.

User generated tags, though relevant, contain lots of noise which may propagate if used unfiltered. Again there may be very few tags describing the visual content. In order to create noise-free training models, we manually checked each image and its tag description and filtered out many low-informative words such as tags related to cameras, and photography. We also filtered the image tags with normal English language stopwords. Secondly, wherever needed we manually add relevant tags to the image so that concept learning can be more reliable? Now we have a set of concepts *C*, a set of images *I* and a set of image tags *T*. The next step is to extract visual features for each image and create content based index of the images so that it can be used to compute visual similarity. For each image we computed three global feature sets to describe the visual characteristics.

Table 3: list of visual concepts taken from LSCOM and number of training images from Flickr.

Concept	Associated words	Number of images
Emergency Vehicle	ambulance, ems, hospital	497
Police	Car, street and cop	456
Protest	Demonstration, cop, crowd	477
Explosion	Bomb, fire, nature	434
Louvre museum	sculpture, art, pyramid	532
Ship	Boat, sea, water	500
Airplane/helicopter	Plane, sky, airport, jet, aircraft	500
Bridge	Water, boat, city	500
Night	Light, city, moon, sky	500
Harbour	Sea, port, boats	491

6.3.2 Visual Feature Extraction

1. Edge Histogram: The MPEG-7 descriptor “Edge Histogram”, *h* represents spatial edge distribution in the image. The elements in the vector of 80 dimensions are used to obtain semi global and global histograms to improve the matching results. It is suggested as a good candidate for image similarity with non-uniform edge distribution.

2. Grid Colour Moments: Grid Colour moments have been widely used to model the distribution of colour in the image. Prior studies show the first three moments of each 3x3 grid for three colour channels outperforms histogram based comparisons. The three colour moments (Mean, standard deviation and skewness) represent a compact characterisation of the colour distribution for the region.
3. Colour autocorrelogram: Colour correlogram describe the spatial relationship between colours [Huang et al., 1997]. An autocorrelogram expresses the probabilities of colours to re-occur in a certain distance. We preferred a small distance of 4 (pixels in the neighourhood), so that local spatial correlations of identical colours are represented are described in the HSV (Hue, Saturation and Value) colour space. In total, each colour correlogram results in a 256-dimensional feature vector.

Each image is now represented with three visual feature vectors and one textual feature set as shown in Figure 29. The text features acts as labels and the visual features are used to train three separate indexes using the Lire library. Lire [Mathias and Chatzichristofis, 2008] is an image analysis and indexing library which uses the Lucene[46] text engine to index the image features. The system contains three different indexes for three feature class so that it can return three different result sets for one query image.

	Tags: Paris, summer, 2005, august, holiday, Louvre, night
	Edge Histogram: 5, 2, 1, 0, 0, 5, 7, 3, 3, 2, 1, 1, 4, 6, 3, 3, 1, 2
	Colour moments: 90.878,45.385, 0.336,87.45,32,09,0.312
	Autocorrelogram: 3.7777,1.9921,0,0,0,0,0076, 1,0009,0,0,0

Figure 28: Feature vector of a training image.

6.3.3 Visual Similarity Metrics

Similarity and distance calculations between two images D(A,B) depend on the type of features used to represent them. For Edge Histogram we followed the MPEG-7 prescribed [Won et al., 2002] distance calculation..

$$D(A,B) = \sum_{i=0}^{79} |h_a(i) - h_b(i)| + 5\sum_{i=0}^{4} h_A^g(i) - h_B^g(i)| + \sum_{i=0}^{64} |h_A^S(i) - h_B^S(i)|$$

[46] http://lucene.apache.org/java/docs/index.html

This equation takes the semi-global edges and global edges, both computed from its original local histogram and computes a vector based distance calculation to measure the similarity between two images A and B.

For Colour Moments and Autocorrelogram we followed the L1 based Euclidian distance between two images (A, B).

$$(a_1 - b_2)2 + (a_2 - b_2) + (a_3 - b_3)2 \ldots + (a_n - b_n)2 = \sqrt{\sum_{i=1}^{n} (a_i - b_i)2}$$

6.3.4 Test Collection

- We collected two videos for each query concept. The average duration of the videos is 90 seconds.
- Each video then segmented into shots following a simple shot boundary detection algorithm of histogram comparison between frames. To keep the computation simple, each shot was represented by a single key frame (middle).
- The next step involves computing visual similarity against the training model created earlier. We restricted our result to five unique visually close neighbours. We made it clear that the neighbour should belong to different authors in order to avoid any bias. The reason for such a conditional filter is that, in Flickr, users tend to follow similar tag patterns when uploading multiple or thematically similar photos.

- For each query we get three sets of results from three different classifiers.
- Aggregate and rank the top *m* tags with majority voting and co-occurrence.
- Suggest the recommended list

6.4 Result

As part of the evaluation, top ten recommended tags and their ranking positions were analysed. The assessors were asked to judge the presence of the concept in the list. They were given the main query concept and its associated terms to help in the evaluation process. The assessors were asked to judge if the query concept was present in the" top 5" positions, or "in between the 6^{th} and 10^{th} positions", or as a third option "missing" altogether from the list. The metrics used to evaluate the performance are:

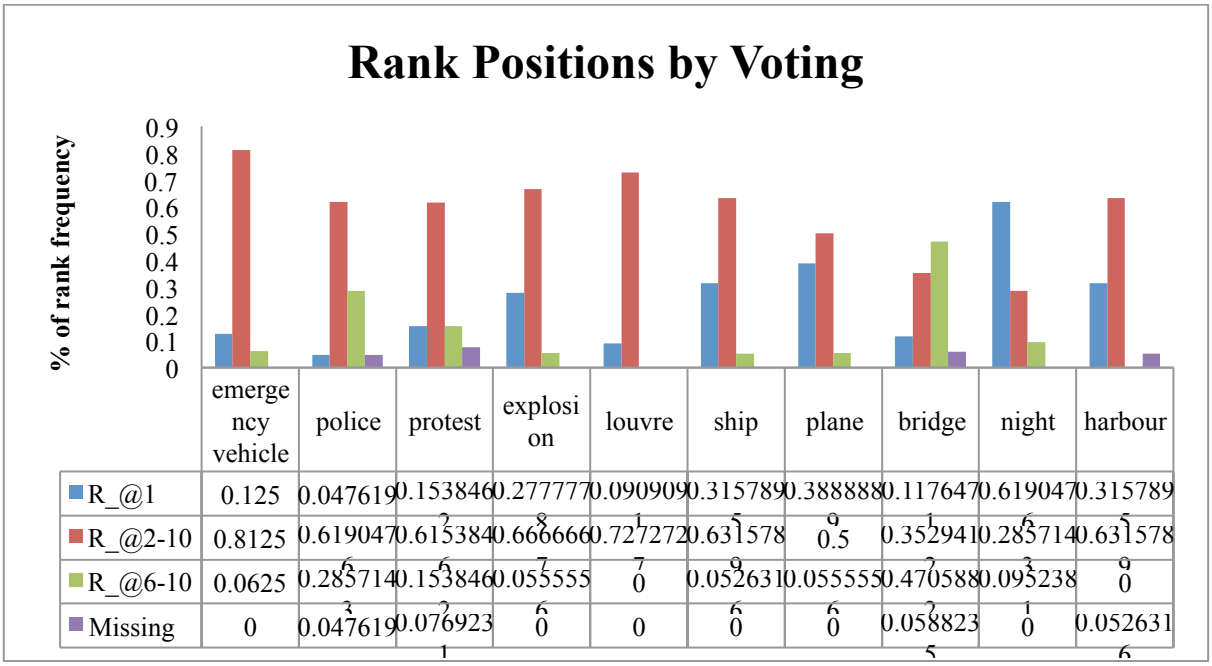
- "Average precision of the query concept": it measures the system performance in terms of the average ranking of the most relevant tag across all test images.
- success @ 1(tags occurring at position number 1): number of times the most relevant concept comes in at the top position in the suggested list.

- Success @ 2-5(tags occurring within 2nd and 5th position): indicates the system's ability to retrieve the relevant tag within the top five tags.
- Success @ 6-10 (tags occurring within 6th and 10th position): reports that the relevant tag was placed between the 6th and 10th position.
- Missing: Amount of time the most relevant is missing from the recommended tag list.

Table 4 shows the number of test images (these images are the key frames extracted from the video shot) for each query and their classification in three different rank categories by voting method and. Some of the result from the evaluation is discussed below.

Table 4: Visual concepts and their rank positions obtained from two different methods ("V" by voting and "C" by conceptual clustering).

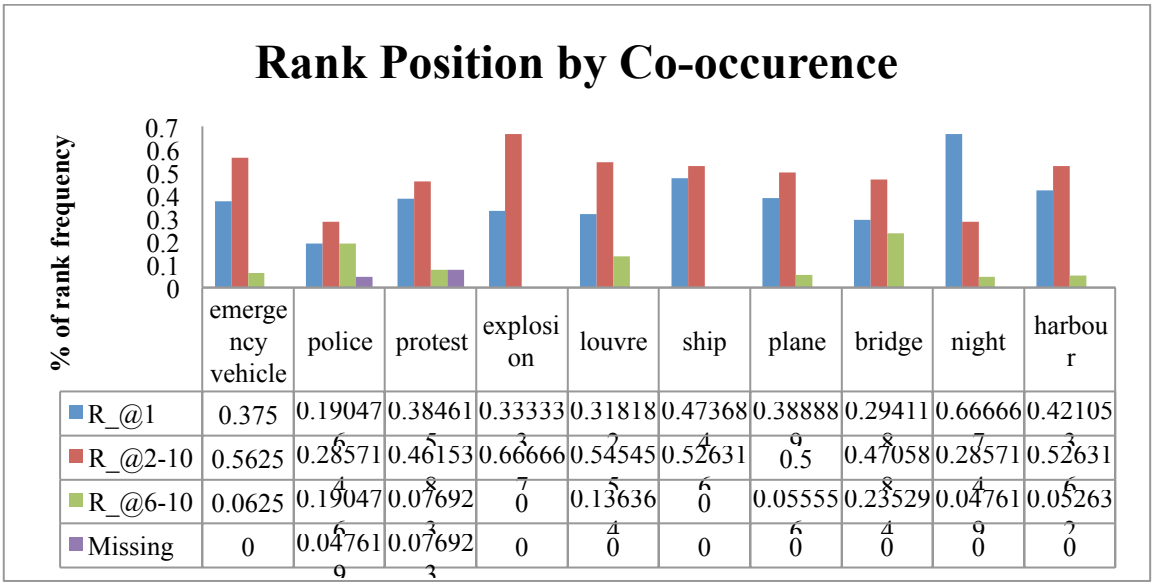
Concept	Number of images for testing	Rank 1		Rank 2-5		Rank 6-10		Missing	
		V	C	V	C	V	C	V	C
Emergency Vehicle	16	2	6	13	9	1	1	0	0
Police	21	1	4	12	12	6	4	1	1
Protest	13	2	5	8	6	2	1	1	1
Explosion	18	5	6	12	12	1	0		0
Louvre museum	22	2	7	16	12	4	3	0	0
Ship	19	6	9	12	10	1	0	0	0
Airplane/helicopter	18	7	7	10	9	1	1	0	0
Bridge	17	2	5	6	8	8	4	1	0
Night	21	13	14	6	6	2	1	0	0
Harbour	19	6	8	12	10	0	1	1	0



	emergency vehicle	police	protest	explosion	louvre	ship	plane	bridge	night	harbour
R_@1	0.125	0.0476196	0.1538462	0.2777778	0.0909091	0.3157895	0.3888889	0.1176471	0.6190476	0.3157895
R_@2-10	0.8125	0.6190476	0.6153846	0.6666667	0.7272727	0.6315789	0.5	0.3529411	0.2857143	0.6315789
R_@6-10	0.0625	0.2857143	0.1538462	0.0555556	0	0.0526316	0.0555556	0.4705882	0.0952381	0
Missing	0	0.0476196	0.0769231	0	0	0	0	0.0588235	0	0.0526316

Figure 29: Results for tag ranking by frequency voting.

- The average ranking of the most relevant concept @ position1 is .2452.
- Results here shows that by following the voting strategy, most relevant tags (in our case the query concept) is often relegated to the rank position between 2 and 5. The reason for such a result may be due to the very nature of frequency voting where the most frequently occurring tags move up the list. For example, in the case of "Louvre Museum" (Figure 22) it has been observed that the tag "Paris" comes to the top position. To counter this effect we may consider various semantic relationships that exist between the recommended tags. By exploiting the tag relationship we may group the semantically related tags into a more coherent cluster, and vote for the best semantic cluster instead of a single tag.
- Present evaluation showed that in more than 82% of the cases the query concept comes within the top 5 tags and this coverage becomes 95% when we consider the entire tag list.



	emergency vehicle	police	protest	explosion	louvre	ship	plane	bridge	night	harbour
R_@1	0.375	0.190476	0.384615	0.333333	0.318182	0.473684	0.388889	0.294118	0.666667	0.421053
R_@2-10	0.5625	0.285714	0.461538	0.666667	0.545455	0.526316	0.5	0.470588	0.285714	0.526316
R_@6-10	0.0625	0.190476	0.076923	0	0.136364	0	0.055556	0.235294	0.047619	0.052632
Missing	0	0.047619	0.076923	0	0	0	0	0	0	0

Figure 30: Results for tag ranking by co-occurrence.

- Co-occurrence based ranking showed an improvement in R@1 where the average improved from .2452 to .3846.
- 86% of time the most relevant concept appeared within the top five positions.

6.5 Research Question Revisited

This chapter of the thesis discussed a framework for visual concept learning from cross media sources i.e. from Flickr images to video concepts. Given a global feature set of a key frame, the system retrieves visually similar images from Flickr to consider the corresponding social tags as the candidate concepts. Further filtering, aggregation and ranking leads to a final set of recommended tags or concepts. Experiment with a subset of LSCOM vocabulary shows the potential of the approach need to be explored further with a fraction of computational expense compared to the available TRECVID evaluation efforts, which needs extensive manual annotation of videos. Since the user created images on the Social Web are tagged with keywords, despite its noisy characteristics, UGC significantly reduced the annotation effort and serve as the training data for the proposed system.

The experiment result showed that there are some concepts can be categorised as good concepts (museum, ship/boat, beach) compared to the other concepts such as protest, emergency vehicle etc. This result is in line with TRECVID 2008 results where "ship", "airplane", "city", "night" are considered good visual concepts to be detected with the help of social web whereas concepts such as "emergency vehicle", "two people" [ref for trecvid 2008] are considered poorly classified concepts. The reason for such results are not yet clear and need further investigation.

Chapter 7

Enrichment, Ranking and Integration of Web Videos

This chapter discusses automatic tag suggestion from social contexts and semantic enrichment of user videos on the Web. The research question in this chapter is part of the overall objective of this thesis: "Bridging the semantic gap with the contributions from the social web". This chapter explores multiple contextual sources of information to infer and suggest tags to describe video content. It also describes the semantic enrichment of content using the linked data principles.

Work of this chapter has appeared in proceedings of the International Conference on Semantic Web 2009 (ISWC 2009).

7.1 Motivation

Contexts are considered crtitical to understand the content in its totality. After exploring the correlations in visual similarity and using user-generated content to learn visual concepts to label video frames, this chapter is focusing purely on social context data and their possible contribution to enrich and expose the semantics of video content. It also describes our effort to semantically enrich and integrate web videos into the Semantic Web data cloud called "Linked Open Data" with the help of DBpedia concepts[47].

A simple query on YouTube will return thousands of videos, making it impossible for the user to browse all and even filter the most relevant ones. There are two directions to address the issue: 1) semantically enhance and enrich the content space to answer the complex queries; or 2) manually

[47] Work of this chapter presented and published in the ISWC 2009 and 2010 SEMAPRO proceedings.

cluster videos into categories and subcategories so that users can navigate through them. Clearly the approach of the second option is very difficult and challenging when the numbers are in billions and amount of data is in terabytes though the objective is desirable.

Research into the semantic enrichment of multimedia content is heavily tilted towards the image domain where we find a large number of studies focused on how to enrich the semantic description of user photos on the web. The video domain has been largely neglected partially due to the inherent complexities added by its temporal dimension. For image analysis, one photo can be considered as one document while one video of 1 minute duration will be made up of 1800 images (30 FPS over 60 seconds). Though not all the frames are analysed, multiple representative frames (images) are still required to make the video content descriptive. Our objective is to concentrate on various contextual sources of information in a video document to enhance and enrich its content description space. All the contexts are not equal and each needs careful scrutiny for its relevance to the video, since these sources bring lots of noise and need efficient filtering mechanisms to cancel the noise and focus on the depicted content. Though various video sharing sites allow users to describe their videos with title, descriptions and multiple tags, users normally add very few tags (5-6 average) making the content description insufficient. The tags also include many vague ("funny", "wonderful", "from my balcony") and more abstract tags such as "nature", "last summer trip" which do not speak much about the content itself. These abstract tags may be useful for the user in terms of personal content organisation and memory but add little value to the content discovery process.

Also, there are some other challenges observed in the user-generated content space: 1) tags are given by a single user (creator) reflecting a very subjective understanding of the content, while different users may give different view of the same content, and 2) in the absence of any control in selecting keywords, a user's choice of words are influenced by language, culture, community and various other factors making the tags and content description less reliable. The non-intuitive tagging interface may add to the above problems, with many content sharing sites providing a very naïve interface when it comes to tagging. The instructions to the user differ from site to site adding to the confusion. Some sites advocate separating tags by a "comma" while others ask users to separate tags by spaces. Flickr asks that a multi term tag should be enclosed in quotes, which users often forget. For example, they may enter the tag "city of light", as three tags ("city" "of" "light") without the quotes. All the above problems reduce the usability of user-created metadata as a true semantic description of the video content.

The question here is what can be done to enrich the description space so that it can reflect the depicted content of the video in a manner closer to reality? Technically we can process the media content and learn the semantic concepts (both visual and audio) using some of the existing methods developed by

the respective communities, but the problems of such approaches are well evident (described in Chapter 2). Alternatively, we focused on leveraging multiple sources of context information surrounding the video and aggregating the information. The sources of information can be categorised as document context, user context, device context and the emerging social context. This chapter describes various strategies to explore these sources of information in order to enrich the original video description.

An aggregation of multiple sources definitely increases the amount of texts available to represent the semantic content of the video, but it gives no guarantee for a qualitative description. Instead the chances of noise are more if due attention is not given to the relevance of the information source. To increase the tag relevance we opted for and will discuss two different tag ranking algorithms: majority voting and spreading activation.

Various related studies in this direction are described in details as part of the background research section (Chapter 2).

7.2 System Overview

In this section, we will give a detailed description of the tag expansion and ranking module we have built. We begin with a general overview of the different modules, followed by an explanation of the tag filtering and expansion step. Then we will describe the tag graph creation process, and finally we will describe the tag ranking methods by means of spreading activation over the tag graph.

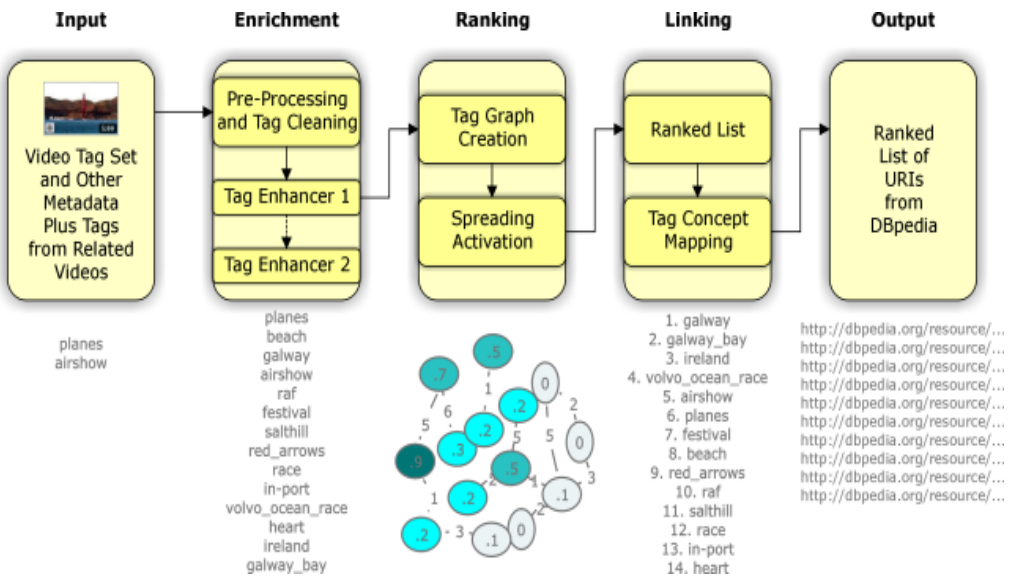


Figure 31: Work flow of the tag enrichment, ranking and linking processes.

Figure 32 shows the normal work flow of the system: (1) context analysis and tag expansion; (2) tag ranking; and (3) concept mapping and linking to the Semantic Web.

The enrichment module takes the original tag space of a video as its input and augments it with information from the title and description fields of the video. It also takes other context fields such as playlists, thematic groups, geographical locations, date, time etc. to enrich the original tag space. The output of this module is a partially ranked tag list, which becomes the input of the next ranking module.

The tag ranking module takes an aggregated tag set to determine its relevance in relation to the video. The relevance rank of the tag is obtained through a node ranking mechanism called "spreading activation", where the tag node is ranked based on the connection strength it receives from its neighbour nodes.

The final module is the Integration module which aims to interlink and integrate the video content to the Semantic Web of data. It maps the ranked tags to DBpedia concepts. DBpedia is considered as a central node in the LOD cloud (the DBpedia nucleus), and linking to DBpedia also allows one to reach other datasets, thanks to the network effect of this project.

7.2.1 Context Analysis and Tag Enrichment Module

Before any details of the three major system modules are given, we need to discuss the data pre-processing and cleaning.

7.2.1.1 Data Pre-Processing

We will begin with a description of our pre-processing step. User-generated tags consist of three broad categories of tags: functional tags (meaningful and mostly single keywords), noisy tags, and compound or emerging tags. Compound or emerging tags are those tags consisting of two or more keywords without any white space such as "friendsoftheearth", "iswc11" (used for friends of the earth, ISWC 2011 respectively). There are other categories of tags which are subjective or opinionated tags, and reflect a user's view point rather than the video content, for example, "funny", "wonderful", "watch this", etc. In our system, we excluded subjective tags, tags made of only numbers (2009, 07 etc.), non-English tags, and tags describing usernames for the purpose of this study. However, there can be some difficulty with compound tags as these tags are not common words listed in any thesauri or dictionary, and further work must be performed to identify meaningful tags from these composite sets. The textual content from the video title and descriptions are subject to the same kind of pre-processing described above, including stop word removal. The stop word list includes common English stop words and some YouTube specific words such as "video",

"YouTube", "hdquality" etc. as these are some of the high frequency words observed in YouTube with little information content.

7.2.1.2 Co-Occurrence Matrix Generation

Tag co-occurrence is one of the key enablers towards creating a more comprehensive semantically related tag space for the video. Co-occurrence statistics used in many previous studies are based on a simple assumption that tags co-occur together when they are semantically related to each other. The semantic relation may be a type of synonym relation or a sub-sumption relation or any other functional relationship. In this experiment we used co-occurrence statistics in two stages; 1) as a means to enhance the video tag space and 2) as the weight between two tag nodes of the video tag graph (described in the tag ranking section). A tag co-occurrence matrix *M* is generated from the video data *V* $\{v_1, v_2 \ldots v_n\}$. The tag co-occurrence matrix *M* is a *nxn* matrix where *n* is the total number of unique tags found in the video collection *V*. We define the co-occurrence between two tags (t_i,t_j) as the number of videos labeled using both tags. This is the raw co-occurrence score without any consideration for the total tag frequency, and hence is less meaningful. A normalised co-occurrence score between two tags will reflect their relationship more realistically. The co-occurrence relation can be symmetric or asymmetric in nature. A symmetric co-occurrence is computed using the equation below:

$$cs(t_i, t_j) = \frac{|t_i \cap t_j|}{|t_i \cup t_j|}$$

This co-efficient takes the number of intersection between two tags, divided by the union of the two tags. Symmetric co-occurrence gives high rank to equivalent tags such as "web2.0", "Web 2.0" compared to other related tags. Moreover the distance is the same irrespective of its direction (distance between Paris (t_i) and Eiffel tower (t_j) is the same as the distance between Eiffel tower and Paris). In asymmetric co-occurrence relation, using the frequency of one of the tags in the entire collection does the normalisation:

$$cs(t_i, t_j) = \frac{|t_i \cap t_j|}{t_i}$$

Asymmetric co-occurrence captures how often the tag (t_i) co-occurs with tag (t_j) and normalised by the frequency of the tag t_i if the origin tag is t_i. Take for example the co-occurrence strength between

two tags “ocean” and “ship”: this is the intersection of “ocean” and “ship” divided by the total number of occurrences of the tag “ocean” in the collection V.

7.2.2 Tag Space Enrichment

In order to facilitate semantic description and retrieval of the video, enhancing and enriching the sparse content space is crucial. With the assumption that multiple information spaces connected to a video explicitly or implicitly determine the video semantics and its message, we started to explore various possible connected sources. These sources of information are created in three different stages: 1) during the creation of the video, 2) during publishing, and finally 3) during usage and sharing. Information can come either from a document context, a device context, and user context or from the emerging social context. Figure 33 shows an overview of possible contextual sources contributing towards the video.

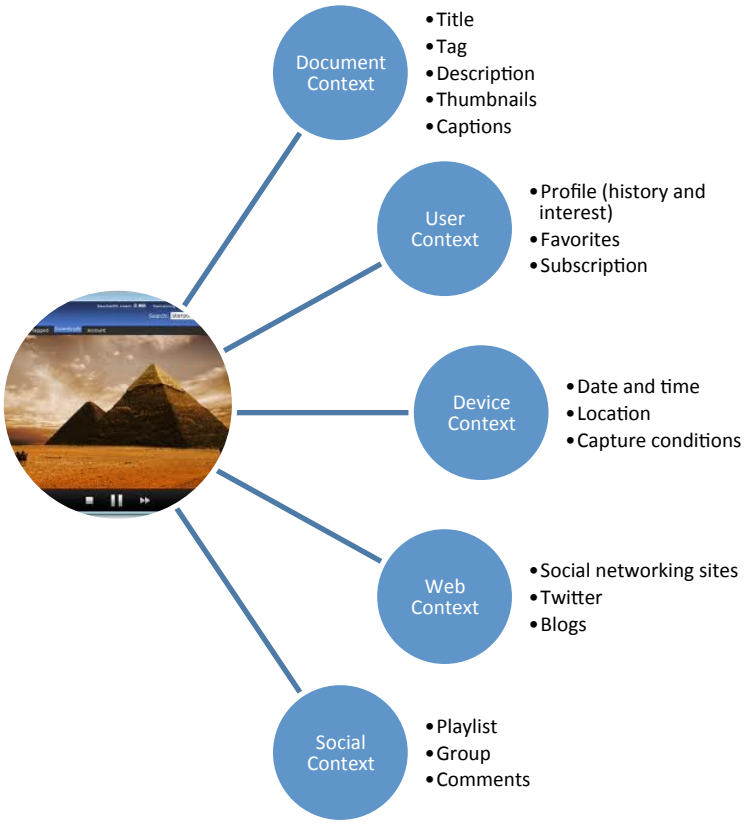


Figure 32: Types of contextual sources of a video.

- Document Context includes those information fields directly related to the video and created during uploading of the video. These fields include title, tag space and description and captions if included. It may also include the location information when the user manually enters the location data.

- Device context gives out information about the time of the video shooting, its quality, size, bit rate sampling rate, date and time information, and capturing conditions such as lighting, scene mode etc. Newer cameras also have built in GPS to record the location information.
- User context reflects a user's interest (favorites, subscriptions), past tagging history and his/her social network.
- Web context includes any information source available on the web beyond the media sharing platform, whether in blogs, wikis, Twitter or on any other social networking sites. It also includes the semantically related videos to the video in question. The present web is full of redundant data. Many videos are duplicated and near duplicate copies are published by multiple users. This content redundancy is a powerful source of tag propagation.
- Social context is the emerging and dynamic source of information about the video it surrounds. Once the video is shared it generates text comments from users, and users often respond with another video related to the original video. Videos become part of the thematic groups and playlists of other users.

Textual contexts such as video titles, descriptions and categories are used to rank the tag weights and sometimes add extra tags that are missing in the tag space itself. To avoid noise propagation, weights are added to different sources, e.g. the "title" field can have a higher weight compared to the "description" field.

A *playlist* named "extreme sailing" can include videos whose tag space is more compact and clustered than the general "sailing" tag space. Playlist and group structures where videos and users are members can propagate tags to the individual video items [Choudhury et al., 2009].

"Related Videos" in YouTube are those videos that are considered similar to the original video in some aspects. YouTube provides a related video feed for each video. It is not known on what basis YouTube ranks the relatedness of a video, and sometimes the results are unexpected. Moreover, YouTube feeds cannot be filtered with complex queries such as "*give me the videos where relatedness is based on a shared tag space, should be from the same geographical location, and must be within a certain time range, and filter these for unique users*" without a lot of work, so we decided to generate a list of related videos for each video from our own data set. The related videos are judged based on mutual content information in the tag space.

Querying for a Related Video Set: Given a query video $v_i \in V$, and its enhanced tag list ET(v_i)={et_1,et_2,..,et_n} where et_i is the i^{th} tag of the enhanced tag list of video v_i, the system returns a list of related videos RV(v_i){rv_1,rv_2,..,rv_n}. Relatedness of a video is determined through the approach of "Find similar documents" and computed using cosine similarity between two text vectors. Once the RV is determined, the related tag set can be obtained directly by aggregating the tags from each of the

related video. But sometimes the list of related videos may be large enough to add noise to the candidate tag list so we restricted the size of related video list to 10.

Querying for a related Video Set with Spatio-temporal filter: The second approach to query related videos is more complex. It is based on the assumption made by [Rattenbury, 2007] that event specific content shows a burst in terms of time and location. We adopted a space and time normalisation criteria in selecting the related videos. To explain this, if videos share a time and place value with the original video, they are ranked higher in relatedness. The intuitive explanation for this is that videos from the same place and same time are more likely to capture the same events and content. Videos from Galway (a geographical area) from 30-05-2009 to 01-06-2009 are more likely to contain the events "Salthill air show" and "Volvo Ocean Race", so accordingly there is a definitive pattern of high-frequency tags such as "Salthill", "air show", "red arrows", "beach", "Volvo Ocean Race", etc. Table 5 shows the comparative tag spaces of related videos with and without time and space filters.

Spatial context is information regarding the geo-location where the video has been recorded or the place that the content describes which can be extracted from the geo coordinates. *Temporal context* is the time of video recording (not publishing). This contextual information not only expands the initial tag space, but it also adds weights to the tags. The intermediate list of tags is the input for the final phase of tag expansion and recommendation based on tag co-occurrence. Table 6 shows the first phase of tag expansion.

Table 5: Four Comparative tag spaces of related videos with and without filters.

Original Video Tags	*Related Video Tags Without Filter*
Planes, Air show, Galway	Planes, Air show, red_arrows, Volvo Ocean Race, Galway, Ireland, Panasonic, NV-GS330, NV, GS, 330, NV-GS
Original Video Tags	*Related Video Tags With Time and Space Filters*
"MOV04687", "Galway", "Ireland", "air show"	"heart", "festival", "raf", "in-port", "race", "beach", "galway_bay", "salthill", "Volvo Ocean Race", "red arrows", "beach"

Table 6: shows the result of tag expansion from multiple sources.

Title	Planes Salthill Galway	"Planes", "beach", "Galway", "Air show", "raf", "festival", "salthill", "red arrows", "race", "in-port", "volvo_ocean_race", "heart", "Ireland", "galway_bay"
Description	same planes!!	
Tags	Planes, air show	
Related Videos	"heart", "festival", "raf", "in-port", "race", "beach", "galway_bay," "salthill", "Volvo Ocean Race", "red arrows", "beach"	
Geo-location	Galway, Ireland	

7.2.2.1 Partial Tag Ordering

Knowing the characteristics of social tagging, it is not surprising that the candidate tags $ET(v_i)$ obtained from the above process is an enhanced tag set for the video v_i but contains lot of noise. We need to estimate the degree of relevance of each tag in the candidate tag set to determine the most relevant and informative tag while pruning the noisy tags. Global tag ranking is based on the assumption that a relevant tag will be voted from multiple sources. Effectively it boils down to a majority voting scheme of the tag $t_i \in ET(v_i)$,, a ranked tag list is obtained by sorting the tag frequency in descending order according to their voting score.

7.2.3 Video Tag Ranking

Tags are entered in a random order with no relevance ranking in relation to the video. The visible tag order is nothing but the user's input sequence. Unordered tags greatly reduce their own usability in tag based retrieval and tag recommendation scenarios.

In this work tag ranking is ascertained in two stages. In the first stage, this is done by means of the enrichment process where multiple sources vote for a tag to be suggested in the final list, however the ordering is agnostic to any semantic tag relation between themselves, and as a result, there are possibilities that many related tags are included in the list but scattered in different position. So a second stage implementing a spreading activation [Quillian, 1968] method is used to bring the related tags closer and make the description semantically more coherent.

The section below will describe the detection of ranked nodes in the graph, beginning with an overview of the tag graph creation, and then describing spreading activation over the graph to rank the nodes.

7.2.3.1 Tag Graph Creation

Given a video $v \in V$ and an extended tag set $XT(v_i)= \{t_1, t_2, \ldots, t_n\}$, we create a local graph of tags. The tag graph is a directed weighted graph with tags as nodes and the links between nodes are weighted edges. The edge weight is an asymmetric correlation based on tag co-occurrence. If the correlation value is less than a threshold (τ) the tags are not considered connected. The co-occurrence relation is calculated as per equation 1. Figure 24(a) shows a tag graph for the video.

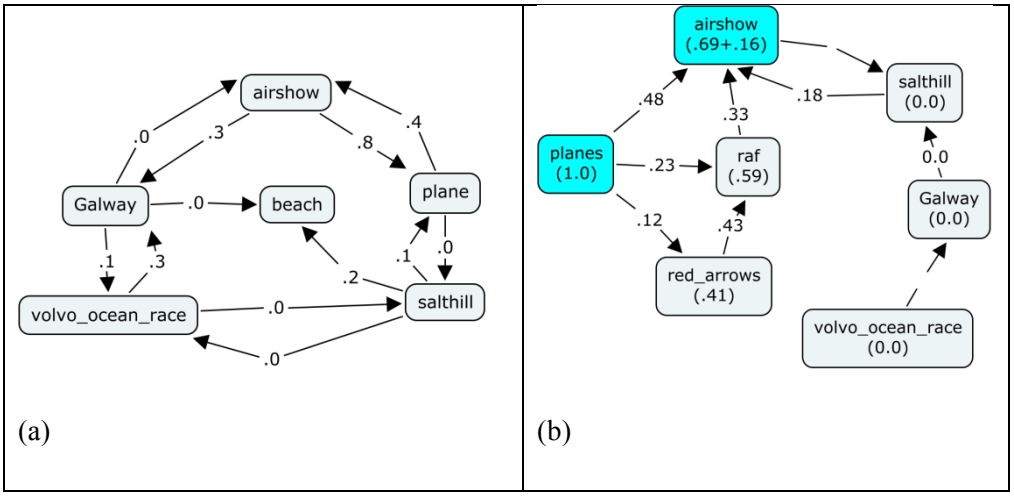


Figure 33: (a) Tag graph for a video and (b): spreading activation from the node "planes".

7.2.4 Spreading Activation

Once a graph is created we can manually browse for the node with the highest link to find the highest relevant node, but there are some algorithms that exists which can simulate such a process. One such algorithm is spreading activation. Spreading activation has been used in many information retrieval and associative network processing tasks. We adopt the constraint spreading activation to exploit implicit relations that exist between tags. The spreading action starts with a node n_i of the graph G and traverses to its connected nodes in order to activate them and identify the most related node. The initiated node n_i starts with some energy e(0,1) and a decay value d∈[0,1] and continues traversing and activating nodes until the node's energy content decays beyond a threshold t∈[0,1]. There can be single or multiple initiating nodes. The spreading activation thus can be formulated as:

$$A(n_j)=(s(n_i)+w_{ij})(1\text{-}d)$$

Where $s(n_i)$ is the input score of the source node n_i, n_j is the target node waiting for activation, and w_{ij} is the edge weight between tag t_i and tag t_j. The activation works in three phases: (1) initialisation, (2) activation (3) inference or ranking. Initialisation process includes the creation of the graph and initialising the energy of each node and their associations. The second phase includes a query where the nodes are activated, and the last phase is the inference stage, in our case the ranking of nodes.

Initialisation: Initialisation of nodes and associations are set. The association strength w_{ij} between n_i and n_j is initialised using the co-occurrence similarity discussed above.

Activation: Each node has an activation strength (A_i) of 0 except the query node which has the activation strength A_i=1.0. Thus in each iteration the energy spread to the connected neighbors are controlled by a decay value $\propto$ which adjusts the network strength and stops at a certain point when the energy is less than a pre-defined threshold. Each path is visited once. Initial activation scores are given in the following

$$A_i^{(0)} = 0$$

$$A_i^{(t+1)} = A_i^{t} + \propto (w_{ij})$$

Where A_i^t is the activation value of node *i* at time *t* and w_{ij} is the association strength between two nodes.

Inference/Ranking: This is the final stage where we have to infer the ranking of nodes. The activation value of a node n_i is the weighted sum of its contributing nodes, and the final ranking is the sorted list of activated nodes. Node strengths below thresholds is discarded.

The activity level of a node indicates the probability of relevance. Experimentally we set up the decay factor to be .85. Once the energy propagation starts, the node spreads its energy to the connected nodes and the receiving amount of energy is a function of relationship strength and the decay ($\propto$) energy factor. Figure 34 (b) shows the activation process starting with the node "plane" and spreading to three nodes "air show", "raf", "red arrows" (this node again spreads and contributes to the "airshow" node).

7.3 Semantic Integration

In this section we introduce our last module; the integration of a video document and its content into the web of data following Linked Data principles. Content in social media sites (both media content and textual metadata) often remains locked within the system significantly reducing its interoperability and reusability value. There is no scope for this data to be integrated with the rest of the web of data unless it is explicitly designed by the hosting service. Moreover, the metadata is not

formally defined either with the support of any ontology or knowledge base which can be used to validate this data. For example, a video annotated with "Einstein" is not linked to the semantic concept of "Albert Einstein" defined in various knowledge base such as Wikipedia, Freebase or DBpedia.

On the other hand, the current Linked Data cloud is growing in size through the addition of more and more heterogeneous datasets from diverse domains. While one of the objectives of the Linked Data initiative is to create a web of data by connecting and interlinking various data sources, thus providing a global graph that can be traversed to discover new knowledge by clients, it is unfortunate that the fastest growing data segment of the web (multimedia data) is mostly out of its reach. Most of the interlinked structured data in the Linked Data cloud is coming from text documents. Very limited multimedia data source are hooked into the LOD cloud. These sources include the IMDB movie database, DBTune, Lastfm music data, etc. More and more multimedia data needs to be linked and integrated to the Semantic Web cloud to increase its reusability and interoperability.

This rich Linked Open Data cloud (LOD) gives each object and concept a unique identifier (URI) which is referenceable and linkable on the web, such that they can make reference to each other irrespective of the vocabulary used. There may be syntactic variations in the usage of tags, in such cases; multiple tags with a common semantics are used to label videos. If we can disambiguate these variations and link to one identifier, this makes retrieval much easier.

To address this problem a solution is to disambiguate each tag to an ontological concept identified by its own URI. Since tags are simple uncontrolled keywords, they inherit the same IR-related problems of synonyms and polysemy. A robust disambiguation method is needed for direct tag-to-concept matching. In the present study, our tag-to-concept opts for a text similarity measure between a tag text and the literal values of either the resource URI local name (e.g. "Semantic_Web" of http://dbpedia.org/page/Semantic_Web) or rdfs:label property. The final output of the framework is a set of RDF triples describing the video and its contextual metadata with the support of a video model and various existing lightweight ontologies such as Dublin Core, SIOC, MOAT, FOAF, etc.

Linking data creation involves three steps:

1. Identify the resource or the part of the resource to be linked to an external data source
2. Identify the URI (in this case the concept URI)
3. Identify the property of linking

The resource we want to link is the video object and its various structural segments and the keywords describing the semantic content of the video. To identify the URI we need to use various lookup services which generally provide a series of URIs from different datasets. We used DBpedia as the

ontological backbone to describe the semantic content of a video. Finally we need to identify the property in our vocabulary to be used as a linkage property, in our case the vow:depicts property will link the video to a DBpedia concept.

7.3.1 DBpedia

DBpedia is a Semantic Web gateway which collects data from Wikipedia. Wikipedia articles consist mostly of free text, but also contain different types of structured information, such as info-boxes, categories, images and links to external Web pages. Much of this structured information is indexed by DBpedia, which serves as a basis for enabling sophisticated queries against Wikipedia content. As of January 2011, the DBpedia dataset[48] describes more than 3.5 million "things", of which there are 364,000 persons, 462,000 places, 99,000 music albums, 54,000 films, 16,500 video games, 148,000 organisations, 148,000 species and 5,200 diseases. These 3.5 million things are described in up to 97 different languages; there are 1,850,000 links to images and 5,900,000 links to external web pages; there are also 6,500,000 external links into other RDF datasets, and 632,000 Wikipedia categories are referenced.

7.3.2 Linked Data Creation

Once the tags are cleaned and ranked, they need to be mapped to their respective concept URIs in order to be interoperable across applications and data sources. For example, there are videos of the "Volvo Ocean Race" on YouTube as well as on other media sharing sites such as Vimeo.

Automatic mapping from tags to concepts is desirable, but challenging due to multiple contexts of the concept. Studies in MOAT[49] showed that users are willing to do this manually when they realise the benefits of such an effort, for instance, if they get advanced browsing or querying features.

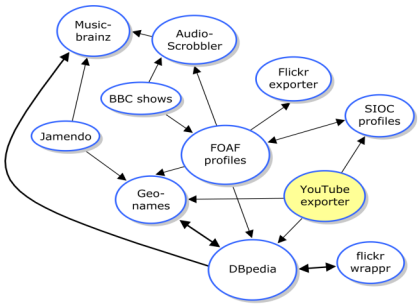


Figure 34: Showing a interlinking of different multimedia data sources in the web of data.

In the present study, since we have a limited domain, the number of mappings is quite small so we used the semantic indexing engine Sindice[50] to query DBpedia and select the most appropriate URIs.

[48] http://blog.DBpedia.org/category/dataset-releases/

Links can also be added to user accounts (SIOC) and locations obtained from the YouTube API (Figure 35).

7.3.3 Tags-to-Concept Mapping

In order to map the tag to a concept, the tag has to be matched with entities (classes or instances) of DBpedia. We presume that cross-resource mapping to other sources from the LOD initiative (such as Freebase) can easily be adapted. Depending on the context, some particular datasets may also be considered, e.g. a genes database when dealing with medical videos. Here, we will briefly describe our approach for tag-to-concept matching. Once the tags are finalized, we use a two-step process for assigning concept identifiers. In the initial step, we lookup the WordNet dictionary for a match and follow some simple heuristics:

1. If the tag matches with a WordNet noun, and if there is only one matching synset, we select the corresponding WordNet URI in DBpedia (Figure 36).
2. If there are more than one WordNet synset, we send the tag and its context tags to a similarity module to compute the cosine similarity between the current tag context and already-existing tag URIs.

If the tags are not found in WordNet we opt for a DBpedia match. We feed the tags to the semantic indexing engine Sindice to look for resources. Sindice presently looks for the rdfs:label and property to find the URI. Once we get the top *k* URIs for the query, the user can select the URI manually or else it is fed into another disambiguation module where URIs can be contextually disambiguated. In order to match with the DBpedia URI format we do some morphologic transformations on the tag label.

- All the characters of the tag except the first one, which is forced to upper case, are converted to lower case.
- If the keyword is a compound noun, all the characters except the first characters of the keyword tokens are converted to lower case (e.g., “new york” is transformed into “New York”) and the first characters are converted to title case.
- We change the multi term label blank spaces to underscores “_”.

---

[49] http://www.*moat-project.org/*
[50] http://sindice.com/

Figure 35: Video tags identified with DBpedia URI.

7.4 Experiment and Results

7.4.1 Data Collection

We have collected 3,990 YouTube videos. All video metadata including the metadata of related videos was collected through the YouTube API. We collected videos of specific categories such as “skiing”, “sailing” and “cricket”. The data includes video tags, dates, places (if available), titles, descriptions and group tags (if available). The total number of unique tags is more than 11,900, which includes many misspellings, number tags, co-joined tags, and subjective as well as meaningless tags. There are 2,261 distinct users in the data set. On average, one user has less than two videos. Since users tag differently depending on their background and expertise, we can assume a relatively heterogeneous tag source. We did a preliminary filtering of tags by removing stop words, tags with usernames and number tags. Though the tag list is far from clean, this reduces a lot of noise.

7.4.2 Experimental Setup

Given a video with some tags or a title or a description, our objective is to suggest an enriched ranked list of tags reflecting the depicted semantic content of the video. To evaluate the approach we need assessment of both the enrichment module and the tag ranking module.

We randomly selected 100 videos from the larger set to explore potential benefits and problems. Three users familiar with the topics were asked to evaluate the new content description in terms of extended tag space and rank the tag lists of these videos on a scale of 1 to 4.

7.4.3 Results

7.4.3.1 Tag Enrichment

To evaluate the system performance in the tag enrichment task and its ability to describe the depicted content, we performed a two-step assessment: "pre and post tag enrichment". The users were asked to watch the video and rank both the original tag list and the enriched tag list on a three-point scale of (1) sufficient for the content, (2) okay, but can be improved, and (3) insufficient. The comparative results are in Figure 37.

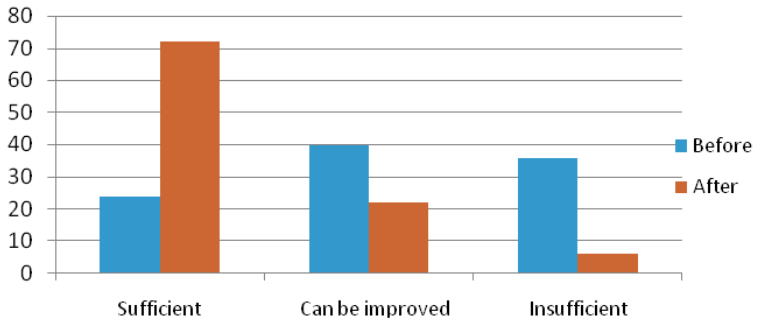


Figure 36: Results for pre and post enrichment process.

The result showed (Figure 37) that the tag enrichment process considerably increased the tag description. Still, 28% of the videos need improvement. This shortcoming of insufficient description may be attributed to many factors: extremely sparse textual data compared to the length of the video, need for more stringent and efficient tag cleaning or perhaps due to noise propagation from the different non-weighted contextual sources used.

7.4.3.2 Tag Ranking

For the tag ranking evaluation we used randomly selected 100 videos. Three users familiar with the topics were asked to rank the system-suggested tag lists for these videos on a scale of 1 to 4: "most relevant", "relevant", "partially relevant" and "irrelevant", according to their depicted content. The objective is to find out how many times the most relevant tag occupy different positions (success@n) in the final recommended tag list. The evaluated positions were 1- 4 of the list (Figure 38).

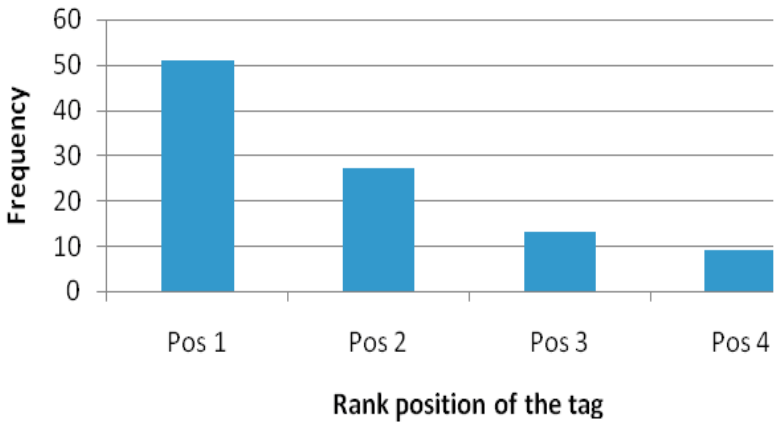


Figure 37: Results for tag ranking (@n).

In order to reduce the subjective bias of the assessor we conducted an inter-annotator agreement and considered the tag's rank when two out of three assessors agreed on it.

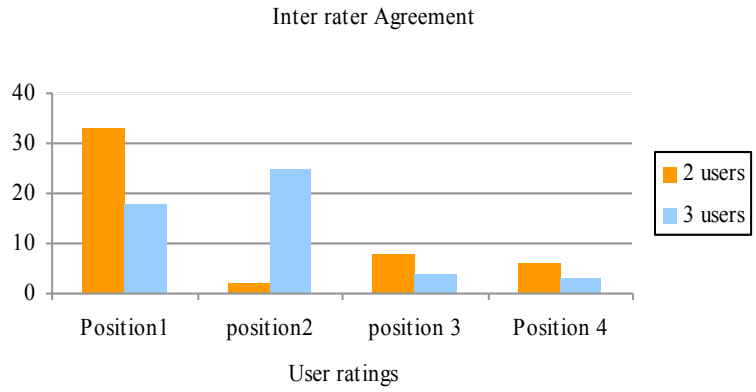


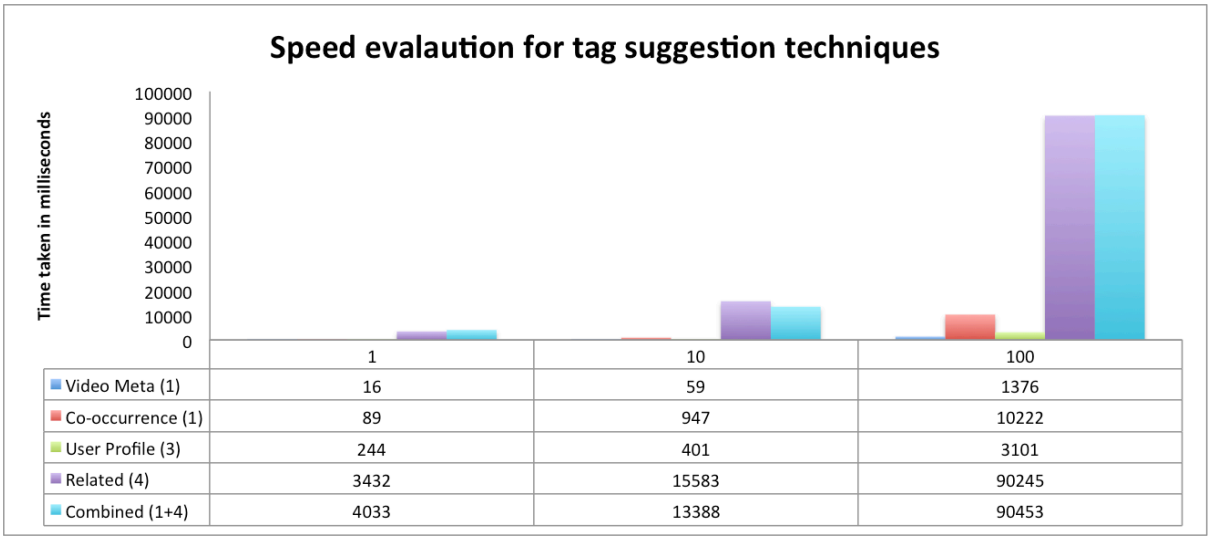
Figure 38: : Inter-rater agreement over the most relevant tag in the first four positions.

Though relatively small, the user feedback gave some interesting results. Figure 39 shows that the most relevant tag came in at the top position 51 times (where a minimum of two users have agreed), whereas the top tag came in the second position, 27 times. Therefore, for almost 80% of the time, the top-ranked tag was in either of the first two positions. Figure 39 shows the inter-rater agreement over the tag relevance in the first four positions.

7.4.4 System Performance

Speed is critical for real-time tag annotation of videos. Efficiency of such approach is determined not only by the system accuracy but also with scalability of the approach to implement in any real world scenario. To get an insight, we describe the time taken by various tag suggestion techniques

individually as well by the combined approach. The performance is tested with 100 query videos on a MAC (2.3 GHz Intel Core i5) of four GB RAM.



	1	10	100
Video Meta (1)	16	59	1376
Co-occurrence (1)	89	947	10222
User Profile (3)	244	401	3101
Related (4)	3432	15583	90245
Combined (1+4)	4033	13388	90453

Figure 39: System performance in terms of time taken for relevant tag suggestion for 100 queries.

The figure 40 shows the time taken for tag suggestion and ranking in terms of milliseconds for 1, 10 and 100 queries. It is evident that the tag suggestion from related videos take comparatively more time (3 seconds per query) than other approaches. This is because the system is live querying to the YouTube data source through its API. In case of a local data source it is much more cheaper and faster. In case of combined approach such as video Meta data and related videos, the time is slightly higher (4 seconds per query) than the "related video" technique alone.

The process can be more efficient when the computation is distributed or used in one of the efficient parallel computing infrastructure.

7.4.5 Overlapping in Context Space

In many cases, context spaces overlap each other in terms of suggested tags. The three major sources of intersection are social contexts, user context and related videos. Social context includes tags title, description, playlists etc., user context includes tags from his own videos, watched history as well as subscribed videos. Here in this work, we extracted the interest profile of a user from his uploaded videos. Below is an example of tag intersection between these sources for a video about "Volvo ocean race". However, this thesis could not cover any empirical study on the amount of overlap between different sources. Out of total 47 tags, 27 tags are overlapping in any two or three sources (figure 41).

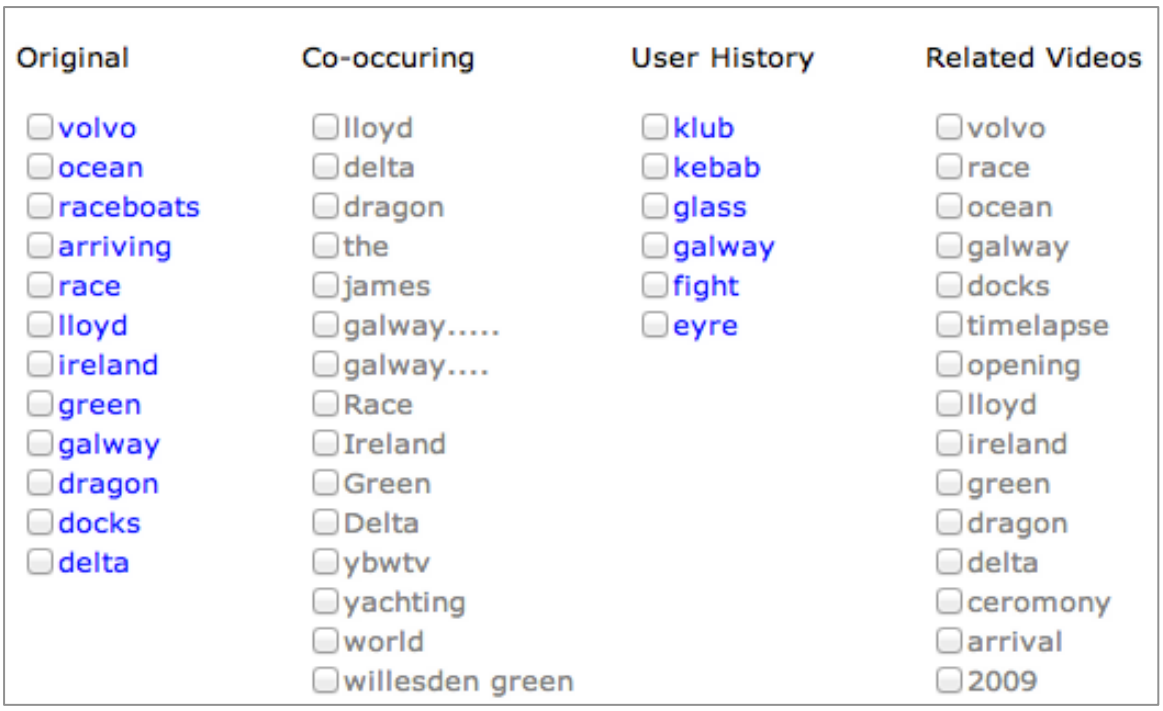


Figure 40: Shows the overlapping between various context spaces for one video.

7.5 Research Question Revisited

In this chapter we discuss yet another experimental use case of contextual contribution for video content annotation. The system implemented with three primary modules of enrichment, ranking and integration.

Social context of a video is defined as the environment where the video is published, shared and interacted within a community. It is reflected through many online user activities such as user comments, tagging, rating, and sharing the link, bookmarking as well as thematic categorisation. Exploring the above knowledge sources provide useful content to enrich the original video content without any low-level content processing and concept classification. Given the initial set of information from a user video (title, tags and description), the system applies a stack of techniques to enhance and enrich the content description. It used contextual information such as tags, time, location, playlist, and thematic groups etc. to gauge the subject content of the query video. An experimental evaluation showed that 72 out of 100 videos showed improvement in terms of qualitative content enrichment to the satisfaction of users while 28 remained insufficiently described. Even though the proposed approach is only applicable for a global level video description, it can act as a reliable bootstrap technique for further enrichment and can feed into the system for an improved search result.

The advantages of the proposed algorithm for tag enrichment are: (1) multiple sources can make the tags more reliable for content description; (2) the subjective ambiguities are reduced; (3) the method is scalable since it does not require any domain specific model training; and (4) it can evolve with tag usage.

Chapter 8

Localised Annotation of Event Video with User Tweets

After the conceptual model and two experimental evaluations for video content enrichment at the global level, the next logical focus is on the lsegment level ocal annotation of the video

This chapter explores the hypothesis that "***real-time web data can be leveraged for time-stamped automatic annotation of videos with semantic entities and interesting events***". We tested the research hypothesis in sports domain.

This work has been published in proceedings of the Making Sense of Microposts workshop at ESWC 2011.

8.1 Motivation

Microblogs and microposts are comparatively recent development in the Social Web landscape. Though informal and short, the richness of these sources and the accompanying metadata not only eases many of the hard core unresolved issues of multimedia content processing but also opens up novel perspectives to address the issues in a multi-dimensional space.

Micro-blogging sites such as Twitter[51], Tumblr[52] and Identi.ca[53] have become some of the preferred communications channels for online public discourse. All of these sites share common characteristics in terms of their real-time nature. Major events and issues are shared and communicated on Twitter before many other online and offline platforms. Recently it has been observed that Twitter users discuss and converse during live events when these events are televised or streamed on the internet. The flurry of user activities during any event depends on the perceived popularity of the event, even sometimes making the event hashtag a worldwide trending topic (such as was the case for Oscar awards, World Cup football, etc.) An observation of the data showed that users discuss the events very enthusiastically with minute-by-minute accounts of the event. This rich and collective textual

[51] http;//www.twitter.com/
[52] http://www.tumblr.com/

stream, if not utilised, can be lost in the short term. Some event organisers have started taking a keen interest both in promoting and evaluating the events through user tweets. Researchers are using tweets in general for various useful real-world applications, e.g. product marketing, sentiment mining, event detection, disaster response measurement, etc. Based on prevailing trends, we made an assumption that this content can be leveraged to address a more complex problem called localised video annotation: i.e. fine grained annotation of video content at the shot and frame level.

Video content annotation is still a computationally expensive process. Automated computer vision and signal processing approaches are either not scalable or are domain dependent. Recent trends in social tagging have enabled users to tag videos with descriptive keywords but research with Flickr tag show that less than 50% of users tags are about the media content [Kennedy, 2007], resulting in insufficient and subjective annotation. Moreover, web based video tagging is mostly restricted to the document level annotation given by a single user.

With our assumption, we believed that these user-generated content items if filtered, processed and analysed can bootstrap the process of time stamped (localised) annotation of the event videos. Extracting useful information from this constant stream of uninterrupted but noisy content is not trivial.

The extraction of useful content such as entities, events and concepts needs to address many conventional IR-related issues as well as some Twitter-specific challenges. Nevertheless, the results can be useful in many real-world application contexts such as trend detection, content recommendation, real-time reporting, event detection, behavioural and sentiment analysis, to name a few. In the present study, we tried to detect named entities and interesting micro-events from user tweets created during a live sports event (a cricket match). The application of these results aims to augment sports video.

This chapter[54] describes our attempt to classify the tweets mentioning the named entities and interesting micro-events occurring during a live game. Despite knowing that the content generated during an event includes discussions and opinions about the event, detecting the discussed entities and interesting sub-events is challenging. As an example, consider a tweet "O'Brien goes ARGH!!!" which actually means that a player called (surname) O'Brien got out. Manual observation says that this tweet contains one named entity (the player's name) and one interesting event (getting out), but text processing applications fail to detect them due to the lack of context rules. We propose various approaches including linguistic analysis, statistical measures and domain knowledge to get the best possible result. For instance, instead of simple term frequency measures, we represent each player and

[53] http://www.identi.ca/
[54] Published in "Making sense of Microposts" 2011 workshop prceedings.

possible interesting events with features drawn from multiple sources and further strengthen their classification score with various contextual factors and user activity frequency (tweet volume). Different steps include:

- Detecting named entities based on various feature sets derived from tweets and with the help of background knowledge such as event websites and Wikipedia.
- Developing a generic framework to detect interesting events which can be easily transferred to other sports events.

8.2 What is Twitter?

Twitter is a popular microblogging and social networking site. Since its launch in 2006, it reached a tipping point after SXSW in 2007 and the popularity of Twitter has grown multiple times since then. The amount of content Twitter now generates has crossed the one billion posts per week mark from around 200[55] million users, covering all possible topics including politics, entertainment, technology and even natural disasters like earthquakes and tsunamis.

The huge popularity of Twitter is due to three factors 1) simplicity, 2) its real time nature, and 3) ubiquity. It allows people to communicate and share *simple text* messages of up to 140 characters. By default the tweets are public and anyone can follow another user without seeking their permission (with a few exceptions) if their interests converge. The flexible follower and following framework lowered the community building threshold with many influential people and celebrities having thousands and sometimes millions of followers. They can update, reply to, address and mention each other breaking the traditional communication barrier. Phenomena like Twitter are shrinking the world smaller and smaller while increasing the size of the emerging social networks. The second most important feature is its instantaneous characteristic: *real time.* Twitter has become a useful platform for reporting real-world events such as breaking news, public mobilisation during political movements (Iran election, Egypt and Tunisia protests), and emergency communications during natural disaster (the Haiti earthquake, the Japan Tsunami). Solis [2010] said “news are no longer breaks, it tweets”. Finally the third popular factor is its *ubiquitous presence*,. Access to the microblogging service is no more restricted to an internet-enabled personal computer; rather it has clients for almost all handheld mobile devices allowing users to communicate anytime, anywhere, thereby greatly reducing the latency between event occurrence and event reporting.

[55] http://www.readwriteweb.com/archives/twitter_confirms_it_has_passed_200_million_account.php

8.3 Linguistic Characterisation of Tweets

Tweets have a maximum 140 characters which includes various content segments such as:

- Hashtag: This is a way to organise content around a topic, event or theme. The word is prefixed with a hash character (#) e.g. *#fail, #www2011*. It can be used as a topic/tag of the message.

- Mention: A user may mention about another user with a "@" character followed by the username to indicate that the message is addressed to the mentioned user is about him/her, e.g. "*@EllnMllr: Is this any way to set the country's spending priorities?"*

- Retweet (RT): This is a Twitter way of endorsing or forwarding an interesting or useful message e.g. *RT @kurtis_williams: Mindshare's Sentiment Analysis Symposium coverage.*

- External resource: It can also contains links to external sources such as news, video, photos, etc.

Due to the length restriction on content size, users generally follow one of two approaches to communicate: either break the message and communicate in chunks, thereby sending a sequence of messages in a short time period, or make the message short and compact using words creatively. Users prefer the second approach as it saves time and reduces the chance of the message being overlooked. But this approach comes with its follies.

- Many frequently used terms are shortened such as "pls" for "please", "forgt" for "forgot", etc. We need a special dictionary to understand this constantly-evolving Twitter-specific language format.
- There is a lack of standard linguistic rules and grammar. Due to the space constraint, language rules are avoided when necessary, and as a result conventional information extraction techniques do not work as expected.
- The use of slang words, abbreviations and compound hashtags are community driven rather than based on any dictionary or knowledge base.

These social media specific terms may not be difficult for humans to read and comprehend, but this is not the case for language processing applications. Extracting semantics from these short messages are not trivial. Existing natural language processing (NLP) techniques such as named entity recognition (NER), speech tagger, etc. are trained on well-formed textual documents such as news wire, blogs and articles, where rich word context becomes an important feature to determine the word meaning.

However tweet messages are accompanied by lots of metadata such as date and time, user location, links which can be leveraged for making sense of the tweet content.

Table 7 shows some of these examples to make the point clear.

Table 7: Examples of user tweet including various linguistic variations of words.

Desktop is for feasting on information, Mobile is for snacking. #adtech
At least my provocative tweet brought **U** back **:)**RT @samadk: MLKing **wld** disagree=Break unjust laws.pay the necessary punishment
@Scobleizer #cfoundry The company **w/** the best service wins **<** That is what open [not that source **thingy**] was originally meant to allow happen

8.4 Methodology

Our goal is to build classifiers which can correctly detect the players' named entities and the interesting micro-events within a sports event. We started by crawling tweets during the time of the cricket matches using the Twitter API. Since tweets can be crawled with keywords, we collected some related keywords and various hashtags (ICC cricket world cup, #cwc2011, cwc11, cricket, etc.) as a seed query list. Despite our filtered and focused crawling, many users use the popular hashtags and keywords to spam the stream to get attention. Including these tweets due to the mere presence of hashtags or keywords may bias the analysis, so a further round of de-noising is performed following a few simple heuristics as described below:

1. Messages with only hashtags.
2. Similar content, different user names and with the same timestamp are considered to be a case of multiple accounts.
3. Same account, identical content are considered to be duplicate tweets.
4. Same account, same content at multiple times are considered as spam tweets.

Using the above heuristics, we were able to identify and remove 1923 tweets from the dataset of 20,000 tweets. Our goal is not to eliminate all noise but to reduce it as much possible in order to get a proportionally higher percentage of relevant tweets. The next step is to divide the datasets into two parts (D_F and D_{GT}). D_{GT} is manually annotated and D_F is used for feature extraction. Each event and entity is considered as a target class and is represented with a feature vector.

Once the players are represented with the feature vector, the next step is to classify each tweet to say whether it contains any mention of a player or not. If the classification is positive, then matching is performed based on the player's full name. Each player is considered as a target class. Let $P = \{p_1, p_2, \ldots, p_n\}$ be a set of players and let $FV(p_i)$ be a set of features used to represent the player. Let $M = \{m_1, m_2, \ldots m_n\}$ be a set of tweets belonging to a single game. We then train the classifier:

$$f(p_i, m_i) = 1 = if\ m_i\ makes\ reference\ of\ a\ player\ p_i, 0$$
$$= if\ m_i\ does\ not\ make\ any\ reference\ p_i.$$

where p_i is the player's feature and m_i is the input tweet. Similar classification is performed for the micro-event detection task.

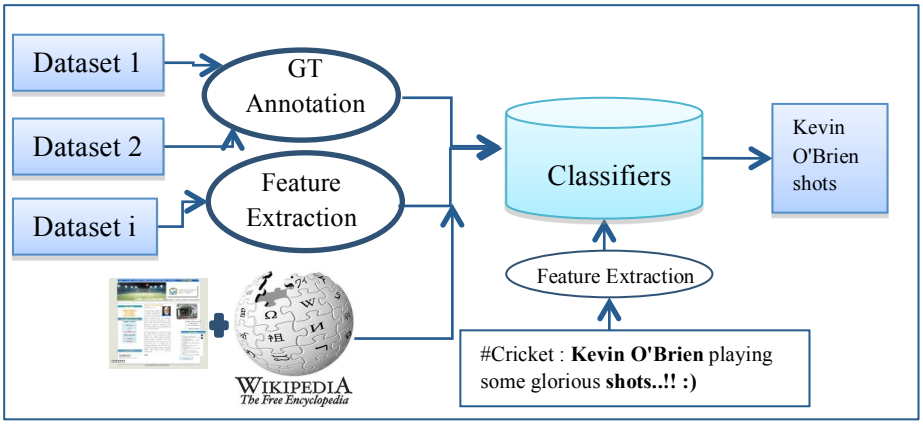


Figure 41: Overview of various steps followed to extract the named entities and interesting micro-events with the help of external domain knowledge (from Wikipedia and game website) and linguistic patterns.

8.4.1 Background Knowledge

Since the main event (a game between two teams) is a pre-scheduled event, we obtained the background knowledge - in terms of the team names, venue, date, starting time, duration, and player details (names) - from the game website. We also collected various concepts common to cricket games from Wikipedia as a list of context features. The list consists of domain terms such as "crease", "field", "wicket", "boundary", "six", "four", etc. All of this background information was collected manually.

8.4.2 Data Preparation

After cleaning and removing the spams and duplicates, we prepared the data for the experimental setup. To evaluate the proposed approach we used the three sets of data i.e. the entire collection of 20,000 tweets for feature extraction (D_F) , a subset that is selected from D_F for ground truth annotation (D_{GT}) which consist of 2000 tweets belong to a single match, and an independent dataset (D_{ind})

collected from a match held at a later date. In the absence of any specific evaluation benchmark we decided to evaluate against the human annotation.

8.4.3 Ground Truth Annotation

Three student colleagues who understood the game and were keenly following the event were asked to annotate a dataset (D_{GT}) of 2000 tweets. To maintain the quality and consistency of the annotations we gave the annotators a list with the players' names and the team names collected from the ICC World Cup website. Annotators were asked to read and label the tweets in two stages: 1) player labelling and 2) event labelling.

Player labelling involves a decision on whether the tweet contains any of the players from either team and if yes, what are the names. The annotator has to either label the tweet with detected player names ("yes") or with a "no". To make a "yes" label, the tweet must satisfy two conditions:

1. The names must be part of the list provided,
2. The players in question must be a part of the game during that time interval.

The logic for the second condition is to avoid the number of false positives. There may be many instances where users may be discussing one famous player of the team even if they are not in action at that moment, and there may be cases where a new user has just joined in the stream and started discussing immediate recent events involving other players.

Event labelling annotators were asked if the tweets were about any interesting event and to give one of the three labels "yes", "no" and "other". When they think the tweet is about an event that happened just recently, it will be labelled as "yes", and they describe the type of event, while the "other" label is given when there is confusion and the content is not explicit. To make the annotation more reliable we computed the inter annotator agreement for both types of labelling (player and micro-event). We considered that two out of three annotators had to agree for a label. The results showed that all three agreed on labels in 86% of cases while agreement between two occurred 94% of the time.

8.4.4 Entity Classification

We developed player classifiers which capture a few general characteristics and language patterns from the tweets. It is well established that feature selection is one of the crucial steps for classification performance. To capture the characteristics of players and events as completely as possible, we made a thorough observation of training data to identify the distinguished features. Each feature is given a binary score of 1, 0. Below we describe various feature based methods to classify tweets.

8.4.4.1 Naive Method with Players Full name

The first method is the simplest of all and can act as a baseline method. A player's feature vector is represented by their full name. The full name is part of the list obtained as part of the background knowledge gathering process. With this method a tweet "Tim Bersnan looks not so happy" will be classified as a player positive tweet. The shortcomings of this method are quite obvious as less than 30% of tweets mention a player with their full name and the rest come with different variations.

8.4.4.2 Method with Name Variations

To address the shortcomings of the first method we modified the player representation with features of name variations as our second method for classification. The variations include a first name, last name, and players' initials. Many players are now addressed and mentioned via their Twitter account, so the presence of a player's username (@<player>) or hashtag (#<player>) are also counted as Twitter-specific features *("firstname_only", "lastname_only", "initials", "initial_plus_lastname", "twitter_handle" and "player_hashtag"*) as additional features. One more feature which we considered to be useful was the nickname of the player. However, since correlating nicknames to player names proved difficult, we could not include that feature. Table 8 shows a few examples of the feature subset.

Table 8: Features related to a player.

Player	**Name-Related Feature**
Kevin Peterson	<Kevin Peterson, Peterson, KP, #peterson>
Sachin Tendulkar	<Sachin Tendulkar, Sachin, Tendulkar, SRT>

It follows the same scoring as the first method and marks the feature as 1 and 0. Results showed that this method substantially increased the recall but lowered the precision due to many false positives. The reason for such false positives is the presence of many other player names not strictly relevant to the game at that point in time.

It is worth noting that both approaches above are not sufficient for our purpose, as one suffers from low recall and the other is low in precision. To annotate the videos we need better recall to cover the full video and better precision for efficient segment retrieval. Moreover the above two approaches are missing the temporal relevance of a detected entity. The detected entity has to be in action during the time of the game; in other words the detected entity should be associated with some game relevant actions during that particular time. The above two methods detect players who are both in and out of action. Since our goal is to associate the entities to a particular segment of the video, the above result will not be optimal. Identifying the present action can be done using various present tense

identification techniques such as part of speech tagging, the presence of verbs like “is”, “are” or identifying words ending with “ing” e.g. “fielding”, “bowling” in combination with a player’s name in context etc. Data observation showed, due to the latency between the action on the field and the user’s tweeting activity, many of the on-going field activities are reported in the past tense. Therefore we opted for a presence of game concept approach, where game related concepts are associated with the detected players in order to ensure that the player is in action.

8.4.4.3 Domain (Game) Concept within the Entity Context Method

To address the issues arising in the above two methods, we adopted a third approach called “Domain Concept within the Entity Context”. This method is based on the assumption that the presence of domain concepts within a certain distance of a detected player entity will validate the entity relevance. This method takes the above feature set and enriches it with game related concepts. These concepts are manually prepared from the Wikipedia listing of various rules and action concepts of the game. This method works as follows: once a player is detected from the tweet we scan the tweet with a contextual window of four terms to detect the presence or absence of a game related concept. Examples of such occurrences are given below in Table 9. If we find these rules existing in the message, the feature score becomes 1. Experimental result showed that this context feature has the strongest discrimination power among all other features.

Table 9: Tweets with the context feature.

#Cricket : **Kevin O'Brien** playing some glorious **shots..!! :)**
@slbry **- Mooney** smokes another one **over mid-wicket. Four !! :) #cwc2011**
First **SIX** of tournament for **Afridi!!!** #cwc2011

The context feature gives an improvement over the other methods by filtering out the active players from the non-active players.

As tweets are highly informal, capitalisation is infrequent, but when it does occur we count it as a feature and score accordingly. Finally, a player’s feature vector looks like:

$$FV(p_i) = (\text{“full name”, “firstname_only”, “lastname_only”, “initials”, “initial_plus_lastname”, “twitter_handle” , “player_hashtag”, context_word, and capitalisation})$$

8.4.5 Micro-Event Detection

The second phase of the work is to identify interesting micro-events within main event. Interesting event detection has multiple user applications besides fine-grained segment based annotation and retrieval of video content. Take for example a cricket game which is 4 hours for a short version of the game and 9 hours for a longer version of the game. No one would like to see the entire game if it was originally missed, so a good video highlight generator can identify the interesting points and present them to the user. Presently video highlights are manually edited and recompiled with the help of professionals in post-production facilities. Many researchers attempted to automate the process with expensive content techniques. With automatic interest point identification, this process can be more dynamic and created on demand.

An event is defined as an arbitrary classification of a space/time region. We target events which are expected to occur during a certain time frame (i.e. the match duration), but location is not an issue here as we know the venue of the match and we are not interested in fine-grained locational information such as field positioning inside the stadium. We made a few assumptions regarding an event's characteristics, namely that (1) they are significant for the results of the game, and (2) many users (the audience) will be reacting to these events via their tweets. The methodology options available for detecting game-related micro-events from tweets are: (1) statistical bursty feature detection; and (2) feature-based event classification. We combined both approaches to get the best possible result.

8.4.5.1 Event Feature Selection

Interesting events that arise during a game are not pre-scheduled, but there is the possibility that these events can occur at any moment of time during the game. We manually selected these events from the Wikipedia "Rules of Cricket" pages. There are two broad categories ("scoring runs" and "getting out") and 12 sub-categories of micro-events. Through our observation of tweets, we saw that most tweets referred to the "out" event by itself while not bothering too much with the specific "out" types such as "bowled", "LBW" or "run-out", though they are occasionally mentioned. Based on this, we restricted our classification task to three major possible events, i.e. "out", "scoring six", and "scoring four". Each event is represented with a feature vector which consists of keyword features related to the event.

Keyword Variations: An event is represented by various key terms related to the event. The logic of including such variations is that users use many subjective and short terms to express the same message - "gone", "departed", "sixer", "6", etc. – when caught up in the excitement of the game. These features are again extracted from the D_F dataset.

Linguistic Patterns: Like the player classifiers, the event classifier also includes contextual features and linguistic patterns to detect the events. The presence of such a pattern gets a score of 1 for the feature, otherwise 0. A few of the examples are shown below:

Table 10: Mentions of interesting events during a match.

#sixer from **#kevinobrien** for #ireland against #england #cricket
Kevin O'Brien OUT ! Ireland 317/7 (48.1 ov) #ENGvsIRE #cricket #wc11
Crap **O'Brien goes** ARGH!!!

8.4.6 Tweet Volume and Information Diffusion

We cannot say from a single tweet that an event has occurred. In order to make our detection reliable, we take crowd behaviour into account. Based on the assumption that interesting events will result in a greater number of independent user tweets, we computed two more features to add to the event feature vector: (1) the tweet volume; (2) the diffusion level. Tweet volume is the level of activity while the event is being mentioned, taken during a temporal interval tm_i where $i = \{1 \dots n\}$ and the duration of each tm_i is two minutes (can be any duration depending the requirement). We used a two-minute interval for simplicity but it can be of any temporal size. If the number of messages is higher than a threshold of average plus 1 SD, we mark the feature as 1, otherwise set it to 0.

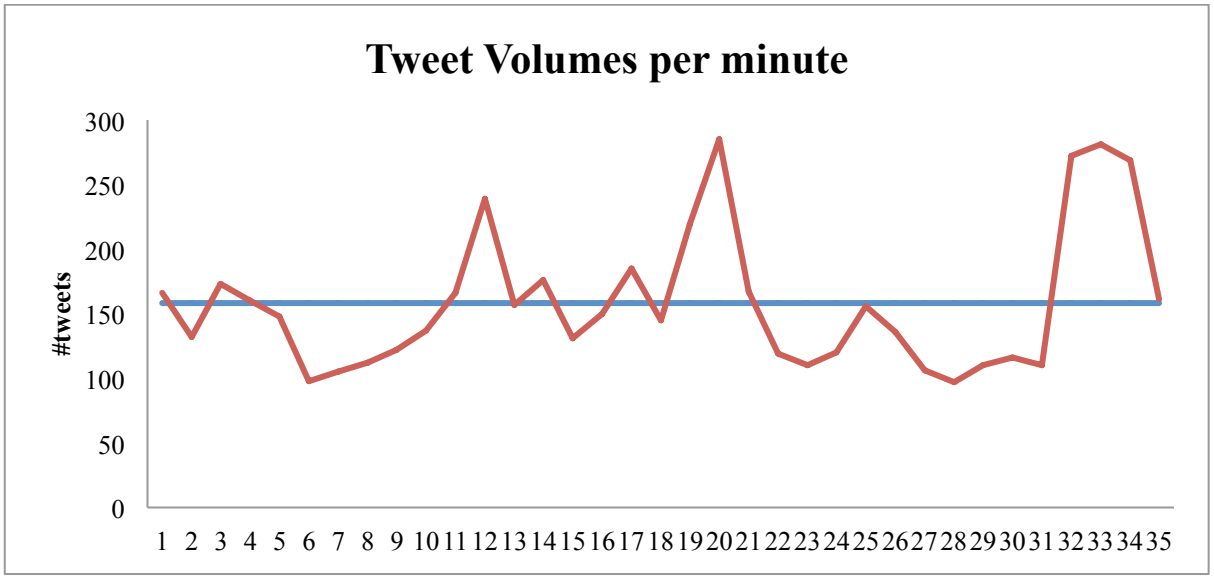


Figure 42: Tweet volumes over time intervals.

The second feature is the level of information diffusion that takes place during the time interval tm_i. It is presumed that more and more users will be busy sharing and communicating the event through their own tweets rather than reading and forwarding others. This means that there will be less retweets (RTs) during the event interval compared to the non-event intervals. This assumption has been confirmed from our observations of the data that the immediate post-event interval has a lesser number tweets than the non-event intervals. The same assumption is also proved in the study [Sakaki et al., 2010]. The feature is marked the same way as the tweet volume feature.

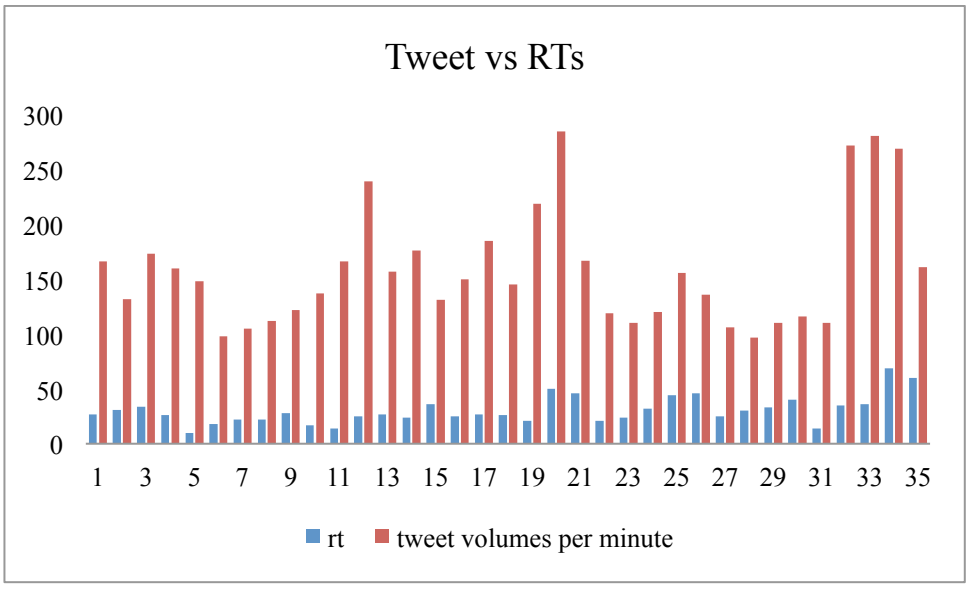


Figure 43: Tweet and Retweet ratio.

8.4.7 Score Identification

The third module is the score identification. It seems that score identification is the simplest task out of the three tasks. The feature dataset showed only seven different frequent patterns of expressing the scores of the game at any point of time. Any pattern may include the team name, number of overs, the number of balls bowled, the number of runs made by two players on the crease and the number of wickets down at that time. Various score patterns and their statistics are listed in the table 11.

Discrepancies in score reporting are observed within short time intervals because of multiple users reporting. We decided to take the latest (or the greater score number) as the final score, e.g if one tweets reports 86/2 in (20.5ov) Ireland and another user reports 89/2 in (21 ov), we take the latter as the right score for that time interval. By following these simple term patterns we could correctly detect 89% of score patterns and the rest are missed due to different linguistic patterns not trained before.

Table 11: Linguistic patterns for detecting match scores.

Patterns	Frequency in %
55<runs>/3<wicket> in 18.0 Overs	12
ENG<team> 59<runs>/3<wicket> after 19 overs	21
England<team> 56<runs>/3<wickets> in 18 overs	9
86<runs>/2<wicket> in (20.5ov) Ireland<team>	11
South Africa<team> 288<runs>-7<wicket>	19
RSA 260/6 (45.5 Ovs)	13
SA<team> 223<runs>-4 <wicket>after 40.3<over>	8
others	7

8.5 Topic to Timeline Alignment

Having described ways to extract interesting topics and entities from the tweets, we are now ready to explore ways to realign these entities and micro-events with the event video timeline. By aligning them to the video timeline, we will be able to index and retrieve video segments based on those entities. Given that topics are detected with a time stamp attached to them, our primary task is to establish a localised relation between two timelines.

Topics to Timeline (T2T) alignments are slightly tricky because we cannot directly map both the timelines. We have to know when users started updating their statuses, and whether the tweets indicate something about the "start of the event". It cannot be done linearly even if we have tweets starting at the same time as the event. The tweet timeline will have some latency compared to the video timeline. We have to find the latency and adjust the tweets accordingly. From existing tweets, it has been observed that users started tweeting about the event as and when it appears with an average latency period of 20-40 seconds. Again it differs from event to event. The latency is more at the beginning of the event and reduces quickly as the event unfolds for the sports event. The initial latency is as short as 20 seconds and people started tweeting about the movement of players, the position of the players, etc. almost near real time. Before aligning the topics, we tried to explore whether user tweets also reflect something about the beginning and end of the event.

8.5.1 Detecting Cues about the Beginning of an Event

Users start tweeting as soon as the event begins. They even tweet while anticipating the beginning. There are some textual cues evident in this regard which can be used to identify the beginning of an event in user tweets. Some of the cues are "starting", about to start", "started", "beginning".

8.5.2 Detecting Cues about the End of an Event

Similar to the beginning, the end of the event can also be detected with the presence of textual cues such as "over", "finished", "end", and in the sports context the most important cues are "win", "beat", "victory". marking these as the boundaries, we turn next to the content processing.

JSNFRMN_MU	16:58	Ireland won! England... hang your heads! #cwc2011
alandooney	16:58	HOORAY #cwc2011
wrighty1707	16:58	Can't expect to win the #cwc2011 bowling garbage like this! England are an embarrassment!!
pranav_teegs	16:58	Truly inspirational victory for Ireland. I rate it right up there along with indo pak clashes #cwc2
KerrillT	16:58	Maggie Thatcher, Winston Churchill, Lady Diana..... Your boys took one hell of a beating :) #c
B12Geezer	16:58	Well done Ireland #cwc2011
adeelatwork	16:58	winnnnnnn #cwc2011

Figure 44: Examples of tweets indicating he end of the event.

To align the detected topics and entities we followed a simple linear technique, using a temporal window of 2 minutes interval and assigning the detected entities to that interval. The size of interval is fixed here but can be set experimentally. It can also be a variable time interval based on the semantic relatedness of the detected entities, which we have not explored yet.

The example (Figure. 46) shows four different shots from the last 12 minutes of the game between Ireland vs. England. The left most columns show all the detected entities from the last 12 minutes of the game and the rest shows four time intervals of 2 minutes duration. The entity localisation method

is somewhat crude but it gives a good approximation. The results are discussed in the evaluation section. Further study is required for better alignment.

	00:00-2:00	2:00-4:00	6:00-8:00	10:00-12:00
England Ireland #cricket O'Brien Kevin #kevinobrien **Four** Won Mooney Bresnan catch drop run out yardy	**Four**, Kevin O'Brien, John Mooney, Tim Bresnan	Michael Yardy, "**catch drop**"	**Out**, Kevin O'Brien, Mooney	John Mooney, **Four, Won**, Ireland

Figure 45: Entities and events extracted form user tweets during the Ireland vs. England match , (bold words are the interesting events occurred during that time interval).

8.6 Evaluation and Results

Our evaluation started with the dataset D_{GT} which is manually labelled both for players and interesting events. We first ran the players classifier and the results are shown in Figure 47. The objective of the evaluation is to judge the effectiveness of the proposed approaches to detect players' named entities and game-related micro-events against the manually-annotated datasets D_{GT} and D_{ind}. We also tested the weight of various features in the classification (positive) and found that a combination of any name feature with the context feature (game-related term) is the best performing feature compared to any other combinations (Figure 50).

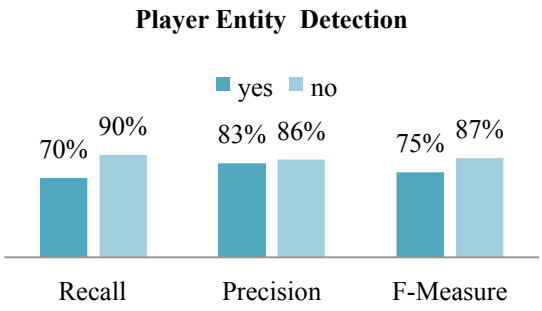


Figure 46: Recall and precision of the player detection classifier.

Like the player classifier, we ran the same evaluation for micro-event detection but in two different stages: (1) classification with only linguistic features, and (2) classification with all features. With linguistic features only (Figure 48), recall is expectedly low at 70% and precision is 74%. This may be due to the noise in tweets. Many event-related keywords are also used in normal conversations like "out", "over", etc.

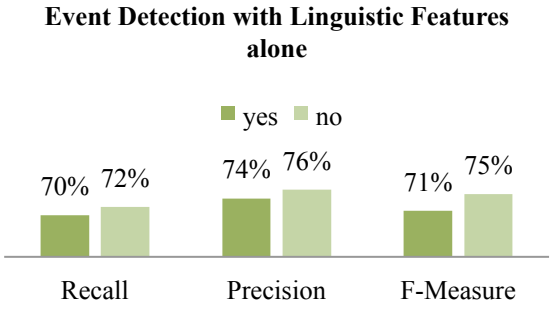


Figure 47: Event detection performance with linguistic features only.

However, when we included the tweet volume and information diffusion level scores, both recall and precision further increased to 86% and 85% respectively, as shown in Figure 49.

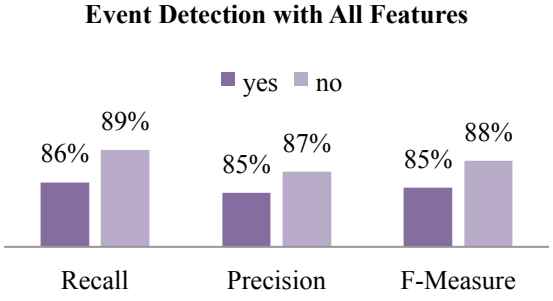


Figure 48: Event detection performance with all features combined.

The results show that irrespective of any features, performance for the "no" labels are always better than for the "yes" labels. We assume this result may be due to the greater number of negative samples available in the data compared to the positive samples.

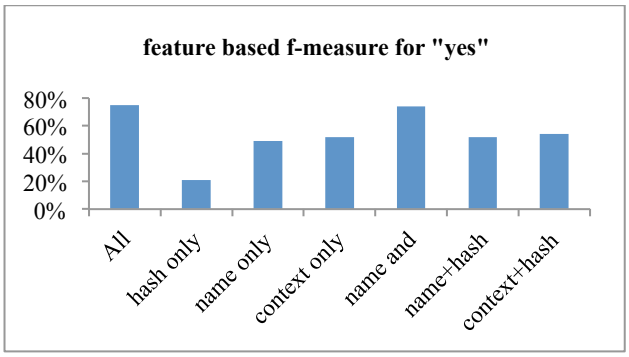


Figure 49: Individual feature performance in player classification.

One question we were interested in answering was can the classifiers be used on other data which is independent of the training and the testing data? To explore this proposition, we ran the classifier on the independent dataset D_{ind} collected from a different game involving two different teams (England vs. Ireland). For this experiment, we tagged the content with part-of-speech tagging using the Stanford NLP tagger[56]; in the feature space, we replaced the player's name with a proper noun placeholder. A summary of the results for both players and event detection is shown in Figure 51 (a) and 51 (b).

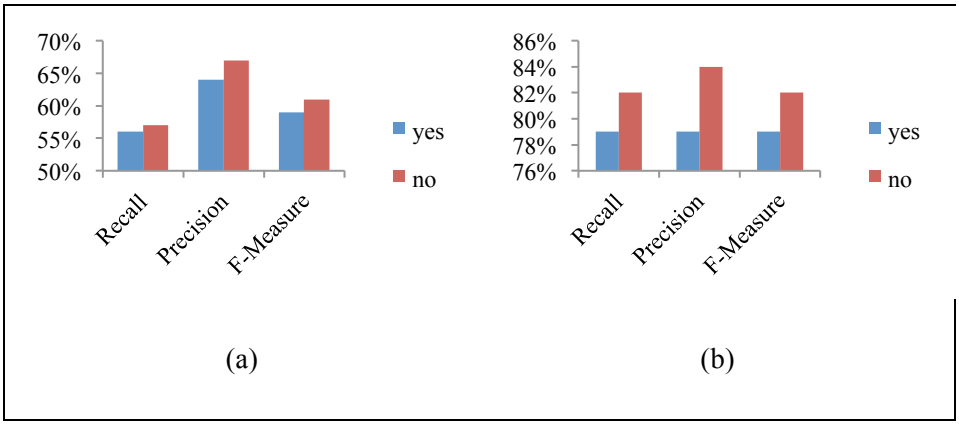


Figure 50: (a) Player detection and (b) event detection in dataset Dind.

As expected, the player classifier scored poorly compared to the event classifier, as the player classifier is heavily dependent on the players' names and their variations. Even if we replace the names with proper noun placeholders, many player mentions are only by first or last name, and other names could not be identified as proper nouns by the part-of-speech tagger. However, the event detection results are good, and the F-measure is above 80% as the features are more generic in nature.

[56] http://nlp.stanford.edu/links/statnlp.html

8.7 Related Studies

Twitter is one of the most popular social media sites with hundreds of thousands of users sending millions of updates every day. It provides a novel and unique opportunity to explore and understand the world in real time. In recent years, many academic studies have been carried out to study issues such as tweet content structures, user influence, trend detection, user sentiment, the application of Semantic Web technologies in microblogging [Passant et al.,2008], etc. Many tools exist for analysing and visualizing Twitter data for different applications. For example, [Jansen et al., 2009] analyses tweets related to various brands and products for marketing purposes. A news aggregator called "TwitterStand" is reported in [Sankaranarayanan et al., 2009] which captures breaking news based entirely on user tweets.

The present study addresses the research question of identifying named entities mentioned in microblog posts in order to make more sense of these messages. Therefore, the focus of our discussion in this section will be on various related studies concerning entity and event recognition in social media scenarios, especially in microblogs. Finin et al., [2010] attempted to perform named entity annotation on tweets through crowdsourcing using Mechanical Turk and CrowdFlower. Similar research in [Rowe and Stankovic, 2010] reported an approach to link conference tweets to conference-related sub-events, where micro-events are pre-defined as opposed to the sports domain where interesting events unfold as and when the event proceeds. Researchers in [Sakaki et al., 2010] built a classifier based on tweet features related to earthquakes and used a probabilistic model to detect earthquake events. Authors in [Sriram et al., 2010] used content-based features to categorise tweets into news, events, opinions, etc. Tellez et al. [Tellez et al., 2010] used a four-term expansion approach in order to improve the representation of tweets and as a consequence the performance of clustering company tweets. Their goal was to separate messages into two groups: relevant or not relevant to a company. We have adopted many lightweight techniques to identify named entities and micro-events during a sports event so that we can later use these results to address existing problems related to conceptual video annotation.

8.8 Conclusion

This chapter described a third social media based multimedia content annotation approach, where user-generated content is used to identify semantic entities and micro-events in sports videos. It also described various feature patterns and their utility in detecting the entities. We started with a filtered crawling process to collect tweets for cricket matches. We arranged three datasets (D_F, D_{GT}, D_{ind}); D_{GT} is a subset of D_F. D_{GT} and D_{ind} are manually annotated with player names and "yes" or "no" for

players and events respectively, while D_F was used to extract the feature set. Classifiers built on these features were able to detect players and events with high precision. It also described various linguistic patterns to identify the score at any particular stage of the game. The generic features of our event detection classifier were applied to an independent dataset (D_{ind}) with positive results. Our future work includes transferring the algorithm to other sports areas as well other domains such as entertainment, scientific talks and academic events.

Next chapter will discuss the research prototype built for the video tag suggestion and annotation purpose.

Chapter 9

Annotate, Search and Browse Prototype

This chapter discuses a prototype implemented as part of this thesis. It includes a search-browse-annotation paradigm to demonstrate various modules for annotation (underlying the conceptual video model), tag suggestion, search and browse: that includes both keyword and semantic search. It also briefly describe the evaluation intrface of the Visual Concept Learning Framework described in chapter 6.

9.1 Introduction

As part of the research approach, building tools to exhibit the reserch process and its evaluation will validate the thesis proposition. In this chapter we present an overview of a user interaction prototype, supporting various functionalities including recommending tags for the annotation of video, and also querying for and browsing semantically enriched video data.

The increasing availability of semantically structured data demands better interface design to support more expressive queries that address complex information needs. The interface needs to present and allow the users a flawless naviagtion cum search mechanism, moreover supporting natural language queries. Many of the semantic web enabled applications provide a SPARQL endpoint to query the underlying data. Though SPARQL, is rich query language for RDF data, it demands a knowledge of the query syntax and the data schema on the part of the information seeker. As an alternative approach, Keyword interfaces in combination with faceted browsers are popularly used as a solution to this problem. Some existing semantic search engines such as SWSE, Sindice, Swoogle and Watson adopt keyword based interfaces to query data and graph-based approach to explore the connections between nodes that correspond to keywords in the query. Despite the maturity of keyword-based interaction models, a few challenges are still evident 1) resolving the query term expressed by a user but not present in the index documents/facts, 2) presenting a more user friendly interface to represent a SPARQL query efficiently. Other Semantic Web based applications such as cultural archives (E-Culture, CHIP) and digital libraries opted for faceted browsing paradigm to expose their data. The prototype described in this thesis adopts a similar approach of faceted navigation to query and visualise the videos an its metadata.

The present prototype is based on the data collected from the web with the help of the web APIs and transformed to RDF data using the described model (VOW) and other related schemas to present a semantic description of the video. Since most APIs only support global level video descriptions and do not yet provide access to the closed caption files (if available), we have document level description in the repository. The RDF data then indexed in a Jena[57] RDF store backed with a MySQL database and the literals are stored in a Lucene based text index.

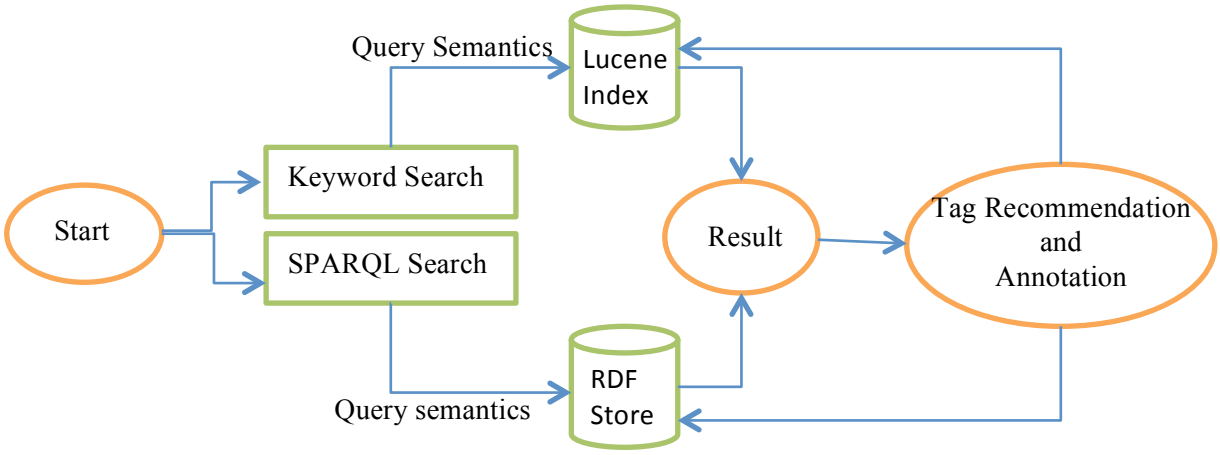


Figure 51: Shows the workflow overview of the system implemented.

9.2 Search Task

The search module allows functionality for both simple term based query and complex queries via keyword search and a SPARQL endpoint respectively. Results are presented as target objects of the query depending on the type of object in the subject position. A query for "Semantic Web" may retrieve the video, user, playlist objects if matched in the object literals.

9.2.1 Keyword Search

For keyword based search, the user has to enter query terms in the search box; he/she also can select the filtering options for getting results for videos matching the keyword, or for users or channels/playlist for the term. In the absence of any filtering, the system returns a list of videos relevant to the query term.

[57] http://jena.sourceforge.net/

9.2.1.1 Query Semantics

User queries are simple keywords where the intention of the user is not explicit. In the absence of the user intention, retrieved result captures all possible senses of the query term putting relevance of the query at risk. The concept of semantic search can fully implemented if we match the intended semantics of the user query with the explicit semantics of the existing knowledge and asserted data. Our prototype tries to capture the query semantics with the help of the user as well as with background knowledge such as WordNet, Wikipedia. Once the user submits a query term/phrase, a query expansion module induces three different functionalities to capture the essence of the term (described in Figure 53):

9.2.1.2 Lexical analysis

Lexical analysis of the term requires the need to determine the number of terms included in the query and create an alternative version of the query term:

- By replacing whitespaces with underscores, e.g. “Semantic Web” will become “semantic_web”.
- By deleting the whitespace between terms and creating a compound term “semanticweb”.

These basic lexical analyses are performed due to the nature of free form user content found in social media sites.

9.2.1.3 Synonym and hyponym lookup

Recall and relevance of the result can be increased if the query term is understood with its semantic relationships. There may be many semantic relationships, e.g general/specific, part-of relationship, relatedness, synonym and antonym. Some of them are more useful and semantically rich compared to others such as the hyponym relation where concept A is a hyponym of concept B. In such cases, concept A is a more general than concept B. Using an example to make the point: A query for “astronomy” will include all the results where the text “astronomy” is evident but it will overlook those videos tagged only with the term “astrodynamics” as the terms do not match and the system is not knowledgeable about the existing semantic relationship between them.

To capture these semantics, user queries are tokenised into terms and processed to detect the presence of nouns and other parts of speech with the help of WordNet. Once the nouns are identified we look for any synonyms and hypernyms available and append them to the original query term to form the expanded query.

9.2.1.4 Wikipedia lookup

WordNet is lexical database consisting of more than 200,000 thousand word sense pairs, but fails to identify the terms that evolve out of community practice such as online interaction and discussion, or the concepts resulting from the latest technical advancements and their usage in contemporary society. On the other hand, Wikipedia has emerged as one of the world's largest encyclopaedias consisting of more than 19 million articles of which 3.7 million are in English, much larger than the WordNet. Since Wikipedia is a collaborative encyclopaedia it continues to grow and includes new concepts with time. We used Wikipedia as a background knowledgebase to look for concepts and collected its redirect labels as part of the query expansion. The final query is a set of terms/key phrases with weight scores assigned to each or some of the elements.

9.2.1.5 Term Co-occurrence Analysis

Co-occurrence analysis of a term reveals its co-relation to other terms not listed as part of the above knowledgebase. There may exist some functional relationship between terms which can be exploited through some ontological learning or rule mapping, and these term co-relations can be extracted from the co-occurrence statistics, e.g. "theory of relativity" can easily extract the concept of "Albert Einstein" because of their high co-occurrence.

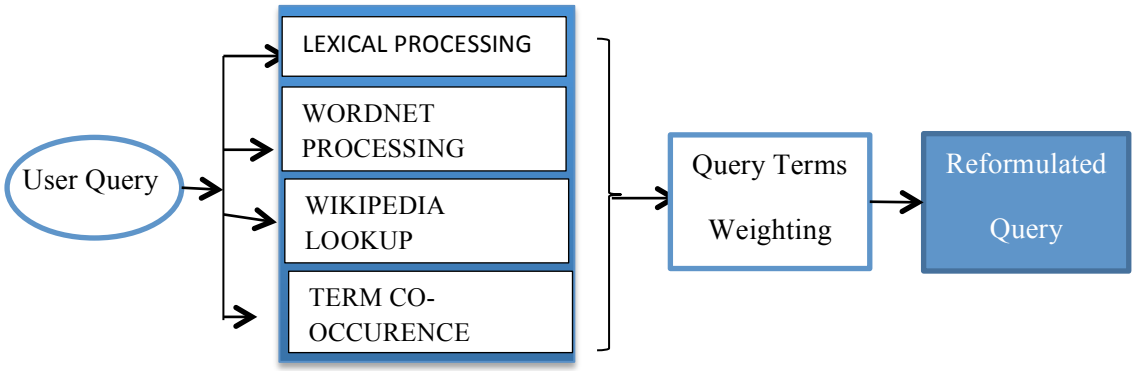


Figure 52: Shows the steps of reformulating a user query to capture some of the query semantics.

9.2.2 Structured Query

Structured data facilitates complex queries, not feasible in normal keyword based search. W3C recommended the use of SPARQL, a SQL-like language to access and query Semantic Web data. Though many Semantic Web applications provide a SPARQL end point to access the underlying data, its benefits are largely constrained due to the absence of an efficient query interface. Normal users are not expected to learn the SPARQL syntax for data access, hence some alternative visual approaches

are adopted by the community such as faceted browsing, data visualisation in tables and graph format. Faceted browsing presents the underlying information space to the user for further navigation and query reformulation.

There are various Semantic Web browsers available for structured data and linked data, which follow a tree traversal paradigm, as followed by the Tabulator [Berners-Lee et al., 2006] interface. A Semantic Web browser is typically implemented to render RDF resources and allows the user to navigate from resource to resource through internal (links within a RDF graph) or external links. Other similar interfaces are Disco [Bizer and Gauss, 2007], and Zitgist[58] where browsing follows a tree structure. Semantic Web Search engine interfaces like SWSE, Watson, and Swoogle start with a keyword based interface and follows a similar resource navigation paradigm to visualise and render the structured data.

Research in the area of user interface design in the context of Semantic Web will make or break the adoption of Semantic Web by users beyond the community. However, since this is not directly within the scope of this thesis, we will only focus on the interaction design we have adopted in this prototype. Below is a list of queries that can be answered by the query engine:

1 List the users who created videos in the topic "Semantic Web".

2 Video titles uploaded by "DERIGalway" (user) with the topic "linked data" in the year 2010.

3 List of friends (within a user's social network) who commented on the video "Video1".

4 List the video segments depicting John Breslin.

9.3 Query Interface

The query interface for the structured query task in our prototype is a SPARQL interface which allows the user to submit a SPARQL query; in fact we consider that a SPARQL interface alone is of little use for non-experts and therefore, greatly reduces the potential benefits. A proper user-friendly interface with SPARQL as the backend is the needed and will be part of our future work. However the interface gives some examples of SPARQL queries and their meanings for user guidance (Figure 54).

[58] http://dataviewer.zitgist.com/.

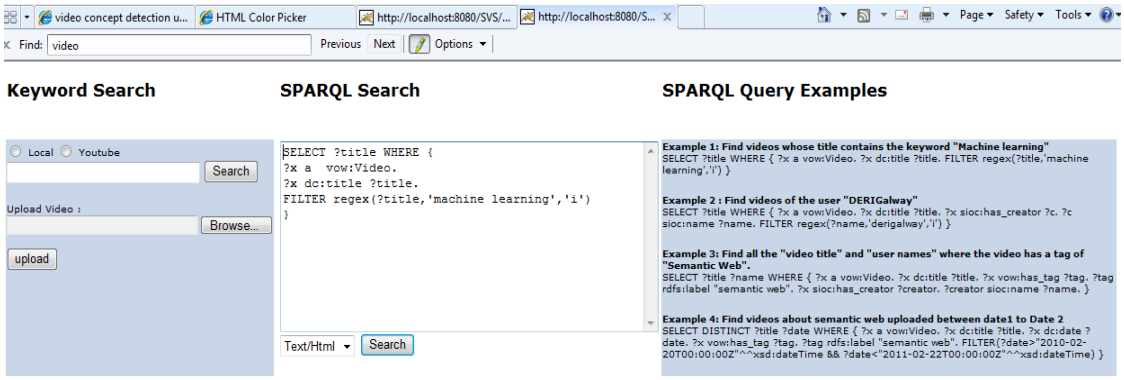


Figure 53: Screenshot of the query interface.

9.4 Result Visualisation

We describe the interface design in which we combined browsing and faceted filtering as part of the interaction model. It also includes implicit query building through facet filtering.

9.4.1 Result List

The core of the interface is the result list. It collects a list of resources (videos as objects in a keyword search or RDF resources in SPARQL search) .

9.4.2 Facet Clustering

The main interaction of the user rests on the result page combined with faceted filter space (on the right hand side). The method of facet clustering helps the user to get a sense of the underlying information space [Yee and Swearingen, 2003]. The basic idea of facet filtering is to segment the data space into conceptual units, for example, clustering all the videos belongs to a particular user. Apple's iTune[59] interface is a good example of facet filtering where the music items can be filtered based on artist, albums and other metadata. Similarly, Exhibit is another faceted interface for graph data structures where the user can filter the result through the facets. Other examples of faceted browsers are mspace [schraefel and Wilson, 2007] and Longwell[60].

Our approach follows a similar method of filtering and restricting the results through facets. Facets are automatically computed and displayed based on the query result and clustered based on the statistics of the predicates like dc:subject and vow:has_creator. This filter space gives users a broad overview of how the information is organised and the data distribution.

[59] A. Inc. itunes. http://www.apple.com/itunes/.
[60] Longwell. http://simile.mit.edu/wiki/Longwell.

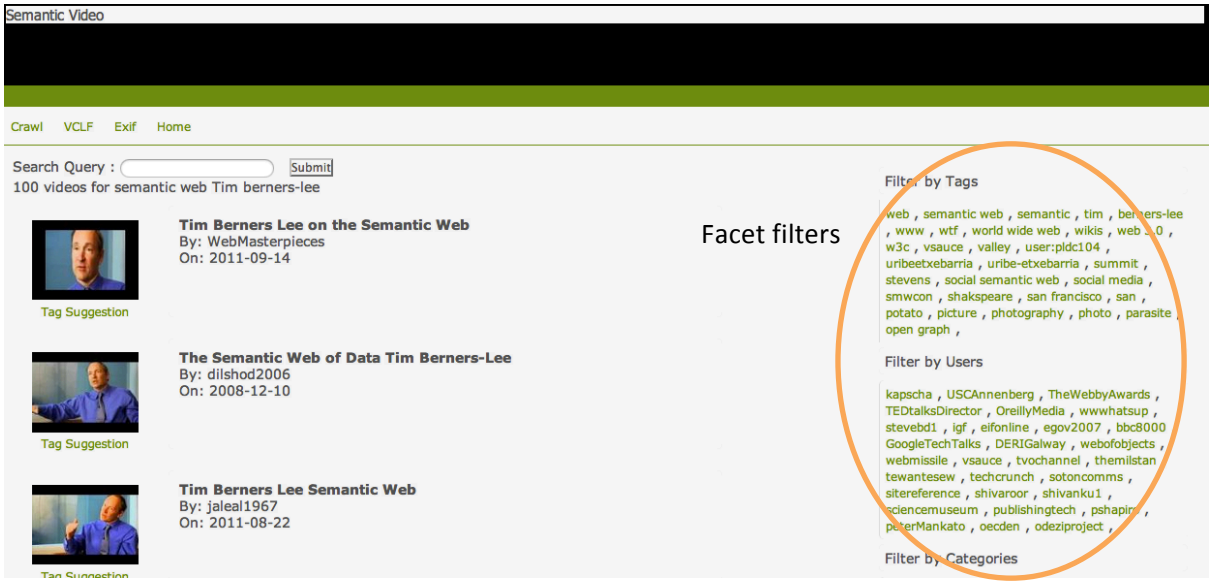


Figure 54: Interface for results from a keyword query.

For example a list of users on a particular topic like “machine learning” can be clustered based on their style (expert user (guru), academics, etc.), interest, or foaf:based_near values. The user can drill down into the result space through the selection of any facet presented, thereby restricting the result space and re-computing the new facets. Figure 48 below illustrates this facet filtering and query building efforts.

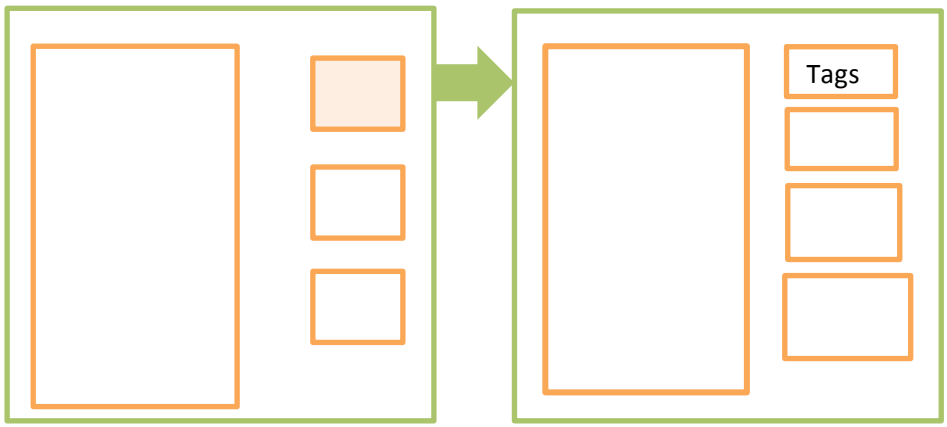


Figure 55: Facet filter space.

Selecting a single object will lead to the details (Figure 57) of the object with its attribute values. Other related objects such as topic, creator, and playlist are connected through various attributes.

Figure 56: Detailed meta data of a video.

The video detail page serves a dual purpose of tag recommendation and annotation, which we describe in the next section.

9.5 Tag Recommendation and Annotation Interface

Tag recommendation and annotation is accomplished in two alternative ways:

1) Through tag suggestions for videos uploaded by the user.
2) Through tag suggestions for the videos already existing in the database so that they can be enriched further beyond their original descriptions.

Tag suggestion for a video comes from multiple sources (figure 58) and can be ranked together to give a final list of tags (figure 59) . The user can verify the sources and select the tag if deemed relevant. Recommendation sources are grouped under two broad categories: 1) based on the title, descriptions, tag co-occurrences and other video level meta information, 2) suggestions from similar neighbouring videos and tags from the user profile or user history. These tags are ranked and presented as one final list to the user. Different ranking mechanisms were explored and discussed in Chapter 7.

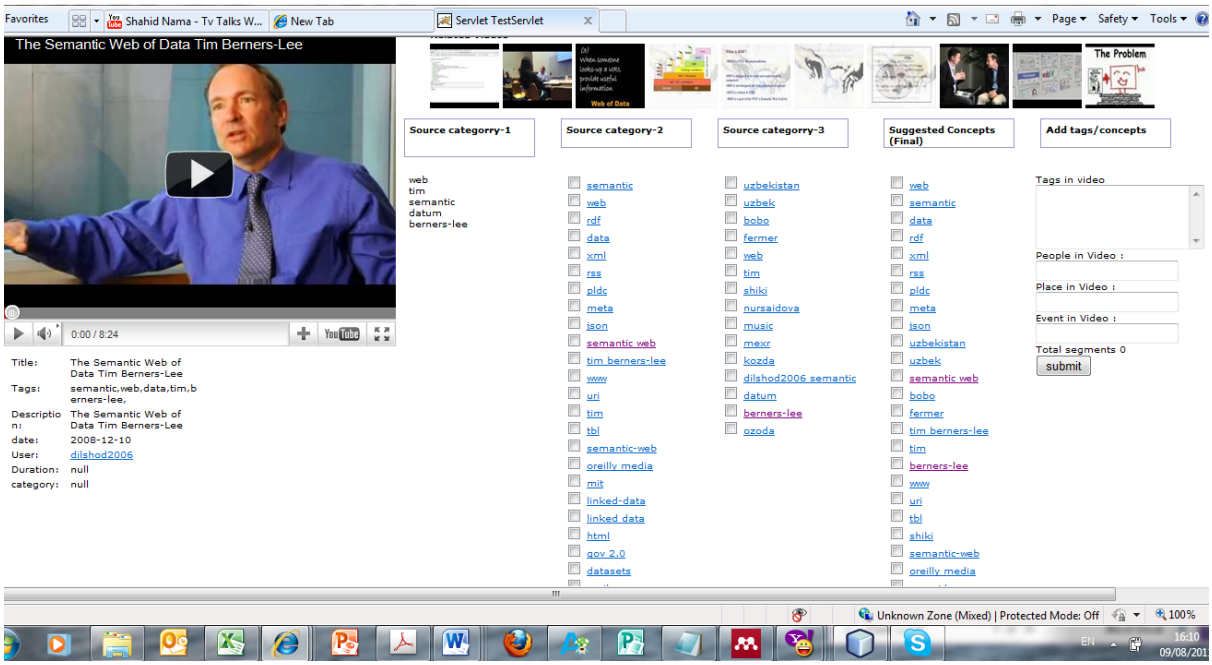


Figure 57: Screenshot of the sources of tag recommendation for the video.

If the user is not satisfied with the suggestion, he/she can always enter his or her own tags and concepts, people or event tags to describe the video both at the global level as well as the segment level. The present system offers only fixed interval temporal segments. However a user controlled segmentation and automatic shot detection module will be part of the future work. Once the user selects or adds any new tags for the video, the system takes all the data and tries to resolve the tags to the DBpedia concept URIs and then transforms the entire description of the video into RDF for re-indexing.

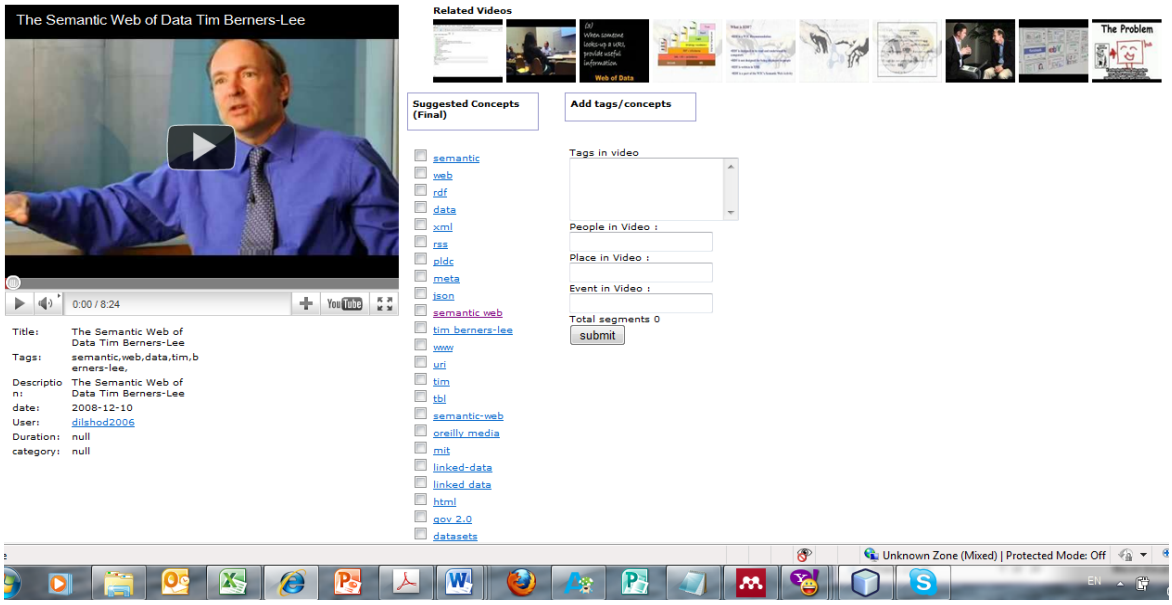


Figure 58: Screenshot for final tag recommendation and annotation interface.

Searching for a user leads to a user page where, besides seeing user uploaded videos, we can find the user profile in terms of tag usage, their social network by means of contacts, user interests, and other user information based on the user object and its associated attributes.

This simple search, browse and annotate user interface is a research prototype developed incrementally as part of the research work. Along the way we realised many challenges in designing and data handling especially for the structured data. Though interface design for a keyword based search engine is sufficiently matured, the same cannot be said about the interfaces for Semantic Web data. We cannot find an implemented user interface to both query and visualise the data in a visually appealing and efficient manner. Discussion of interface design is beyond the scope of this thesis. We have adopted existing practices of querying the Semantic Web through keyword-matching and visualising through facets. Facet-based query reformulation, while it accommodates some of the query semantics, it cannot capture fully what SPARQL can do.

9.6 Visual Concept Learning Framework (VCLF)

This prototype also includes the evaluation module for the visual concept learning framework (VCLF). For VCLF, all the test videos (20) are listed with their original tags and the key frames

1. Each video represented with a set of key frames and their corresponding time stamp. Key frames are detected offline using a shot boundary detection algorithm
2. Clicking a frame will lead to a set of visually similar images and their corresponding tags from Flickr.

3. Ranked list of tags selected for key frame level annotation
4. Aggregation over all the frames constitutes a global description. Of the test video

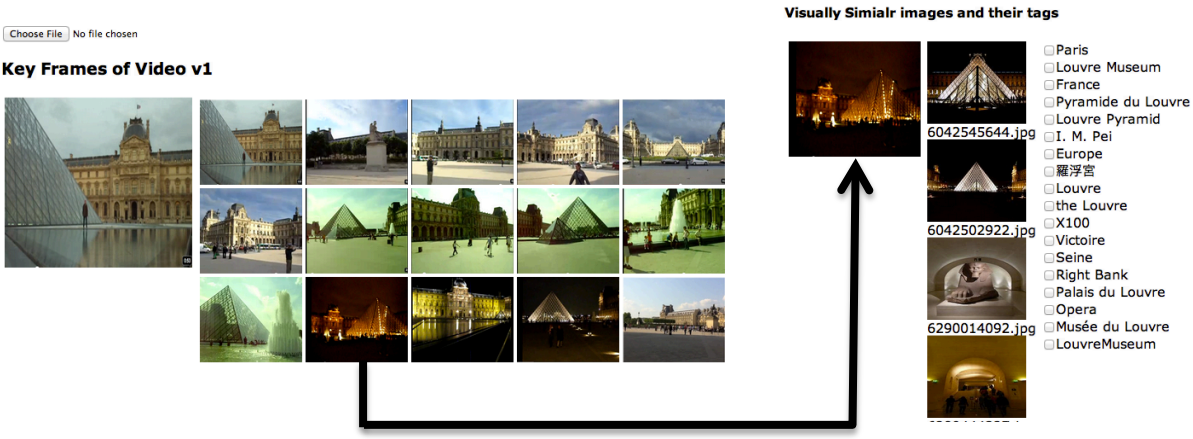


Figure 59: Screenshot of the VCLF with suggested tags for a keyframe (identified with the black arrow)

Figure 60. is a screenshots from VCLF , the left figure describes a video with a set of key frames and the right part of the figure shows four visually similar images to a query keyframe with a list of ranked tags where the concept “Louvre museum comes at the second position”.

9.7 Conclusion

In this final chapter, we discussed briefly the research prototype implemented for the tag recommendation and annotation purpose. This prototype consists of two modules: 1) recommendation and annotation and 2) search and browse. The tool was designed to make recommendations and annotations as a feedback loop where the existing video is re-described and re-indexed based on the system’s suggestion and user acceptance.

Section III

Conclusion and Future Work

Chapter 10

Conclusion and Future Work

10.1 Summary

Inferring video semantics is a complex and challenging research problem. Unlike text documents, multimedia documents are complex and their depicted content is opaque, hence these documents need explicit description. This problem is more acute in the case of video compared to other mediums due to the intermingling of multiple mediums into one. One of the objective of this thesis is to produce a semantic description of web video which is understandable to both humans and machines, hence we described a core conceptual model to represent the video object. The next logical question is, how to populate the ontology model to create a knowledge base? Manual annotation is not feasible at web scale while automatic population of an ontology may be utopian given present status of the existing technology. The ideal solution is to adopt a middle ground, a **"machine suggests and human decides"** approach, where the system suggests possible descriptions in the form of tags/concepts with a certain degree of relevance ranking for the final judgment.

To realise the above goal, the present thesis investigates the hypothesis that content generated on the Social Web, both during the video production, and its usage are valuable contributors interpretation of video semantics, hence should be maximally utilised using multitude of computational methods.

This thesis is motivated by two main research challenges faced by the multimedia research community: 1) the lack of formal description for a media object especially the video and its content; 2) the lack of an automatic means to facilitate and enrich the semantic description of a media object in general, and video specifically. Our thesis starts with two broad research questions: *1) How can we semantically describe the web videos, reflecting core facets such as media properties, links to the subject domain and media content structures, while maintaining a lightweight framework for quick implementation, 2) To what extent the Social Web and multiple contextual sources of a video can contribute towards automatic enrichment and understanding of its depicted semantics, henceforth reducing the existing "Semantic gap"?* As contributions, this thesis proposes 1) a lightweight video framework to semantically describe, and integrate web videos into the structured semantic web for greater interoperability and 2) experimental investigation and evaluation of multiple approaches for semantic enrichment.

The thesis consists of three broad sections: 1) background studies, 2) core implementation, and finally 3) the conclusion and future works.

Background section includes related studies in areas of multimedia annotation, Semantic Web and multimedia semantics followed by the topic of Social Web and its semantics.

To model web videos, we mainly focused on reusing several existing vocabularies to describe the media object, its low-level content features as well as its semantic linkage to domain ontologies and knowledgebases. Departing from existing practices, we treated video on the web as a social object embedded in and gathering its strength from different contexts (rather than an atomic object as per the existing literature). Contextual sources such as user profile (a user and its social network), device context (capturing device and its metadata) and its large social context (in and outside the hosting platform) are rich sources that can be used to infer the implicit semantics. The proposed model focused on three core functionalities to describe a user video on the web:

- Video as a media object (with its structure)
- Content (both low-level and semantic)
- Context (captured through multiple sources)

To evaluate and populate the model with data we showed several application scenarios using videos from the web.

Understanding video content primarily boils down to detecting the presence of semantic concepts (objects, people, events etc.) within a video shot or frame. Existing approaches to address this problem are heavily tilted towards the application of machine learning algorithms using combinations of low-level features of a video. With extensive training these approaches perform well in some selected datasets such as broadcast videos and studio recorded videos, translation of these models to web videos has proved difficult due to the ovious reasons of increased complexities in data quaity.

Based on the above scenario, one of the prime objectives of the present dissertation is to investigate automatic concept detection in video from multiple contextual and complimentary sources exists on the web. We adopted a stepwise approach to investigate combinations of context features related/linked/referred to a video in order to detect and infer semantic concepts (entities, events etc.). To our understanding, the context of digital artefact can be defined as the situation or environment that influences or is influenced by it (video). From these perspectives, contexts for video documents can be multiple, created either during the process of production or during usage and sharing of the video. These context sources may be classified into these broad categories of:

- Data acquisition context(device)
- User context (creator)

- Document context (texts/resources surrounding the video)
- Temporal context (time)
- Spatial context (location)
- Social context (online activities and communities)
- External domain contexts (web, calendars etc.)

To study these sources of information and their contributions towards inferring depicted semantics, we conducted three different experiments where the Social Web and the corresponding user-generated content acted as the knowledge source which needed to be mined in order to extract the relevant pieces of information.

Following the conceptual model, we described various implementations to generate semantic content descriptions of a video both at the document level and at the segment level. Three experiments (Chapter 6, 7 and 8) positively reinforces our hypothesis that social web is an important knowledge source, contineously evolving and providing new opportunities in solving long standing research problems. Concept learning framework discussed in Chapter 6 shows the usefulness of cross media sources where tags for videos (YouTube) are learned from the tags of images (Flickr).

Chapter 7 illustrates another Social Web based enrichment approach, where multiple social contexts were mined to semantically enrich and rank web videos. This study took account of multiple information sources (created during the time of publishing and during various interactions with the video by the online community) to generate and rank video tags. The approach also considered that a combination of spatio-temporal filters facilitates the clustering of videos based on events and thereby generates more reliable event-related tags which can then be ranked using various relevance measures. This chapter also describes the last requirement of the framework: "*semantic integration into the web of data*". Semantic enrichment of video data may help us to understand the video as an atomic unit, but it will not address the issue of interoperability and we still end up with videos in data silos. Reusability and interoperability of data increases when it is described in the context of other similar and relevant data sources. To address the above issues, we adopted linked data principles to describe the video data and curate the conceptual model.

Both studies in Chapter 6 and Chapter 7 illustrate how the Social Web and user-generated content can help to augment the video document as a whole.

Finally Chapter 8 attempts to infer semantic entities and interesting events in sports video from the corresponding micro-blog posts created during the recorded video (sports event). Results showed that even simple computation over these rapidly evolving user-generated content items can detect semantic entities and events with recall of 86% and precision of nearly 85%. These detected entities

are can easily aligned with their corresponding temporal locations in the video timeline allowing both temporal as well as semantic segmentation of the video.

Finally, there are many issues and reservations observed during the research work are worth mentioning and discussed breifly in the next section before the future work section.

10.2 Research Observations

10.2.1 Ground Truth Annotation

Constructing ground truth for video annotation is tricky compared to any other type of document annotation for the simple reason that the user has to spend certain minimum amount of time (equal to the video duration) to watch the video and understand the subject content. We followed a similar approach of manual annotation of videos to evaluate the system performance in suggesting tags. TRECVID challenge also adopts a collaborative annotation[61] approach for hours of video data (professional sources) by multiple users over many months of a time to prepare for the challenge whereas for the internet video dataset they depend on the textual content described by the video creator. This clearly shows the difficulties in annotating video data even for a limited dataset.

10.2.2 Participants

The objective of the evaluation was to judge the quality of the tags suggested by the system against the baseline of user-generated tags created during the publication of the video on the web.

The initial experimental evaluation was done with colleagues from my work place. Being in the domain of computer science and web research they are more or less aware of the concept of social media, tagging, user generated concepts etc. and their age ranges from 24 to 31. Participants were asked to do the evaluation within a week's time. In terms of representativeness of participants, this group is not very diverse in terms of their age and background, which may bias the result. In order to avoid any bias, an inter annotator agreement was computed. Though the user study was small and less representative in terms of actual web users, who vary widely in terms of age, demography and cultural background, videos of known topics and issues are viewed similarly, especially when the participants were instructed to judge on the depicted content "subject content" of the video. For example a video about "sailing, race" will be more or less perceived as a video in sports category and express their views in "race" related terms, Though there may be some content describing video quality, camera details or location of the shooting. This is the precise reason we evaluate the results not only against the baseline of tags provided by the creator, but also against a set of manually ground truth.

[61] http://mrim.imag.fr/tvca/

10.2.3 Evaluation

When developing an evaluation metric for our results, it is worth mentioning that our approach is designed explicitly for generating annotations and is not synonymous with video retrieval as in many tasks of TRECVID. Rather than computing a list of videos for a particular keyword, we generate an ordered list of most likely concept/tag for each query video. As such, the standard retrieval measure of precision-recall is not appropriate for our task. Instead, we considered a closely related measure: for an annotation list position called P@n.

10.2.4 Why top n tags?

Top n tags are evaluated based on two realistic prepositions as evident in the data:

1. Average number of tags fond in user videos –From a comparative analysis of more than 40,000 video descriptions we found a range of 3-51 with an average of 14 tags per video including noise, opinion oriented tags, duplicates and lexical variations etc.
2. Existing literature [Sigurbjörnsson and R. van Zwol.] in automatic tag suggestion recommends this evaluation measure of P@n (precision at n position) which is the probability of finding a tag/concept among the top N recommended tags in our case it is 1,5 and 10. TRECVID studies also pursue a similar approach of mean average precision (MAP) for their concept detection task.

10.2.5 Role of Social Tagging

Most video sharing sites allow users to tag videos at the document level with a list of keywords. Adding scene level annotation is computationally expensive and labour intensive. As the amount of videos increase, these approaches of manual annotation become difficult to scale

Recently some of the services such as YouTube, Vimeo have adopted two major strategies to annotate videos at the scene level e.g. automatic speech transcription and user created textual annotation. While automatic speech transcription is restricted to specific video types (educational videos) the rest of videos are left for user's discretion to upload a text file with time stamp and text content so that they can be aligned to the video timeline.

In our proposed approach, systems can well explore the internal as well as external UGC such as comments, tweets, hyperlinks, and user photos to link and understand user videos. Some of the possibilities include, users tweet about movies, concerts sports, conferences during the live events, users comments on videos, hyperlinks in the content, topical grouping of videos etc. Media sharing systems can make use of the data redundancy to their benefits and describe media content with the

help of near duplicate videos. With increasing computing power (cloud computing) and the popularity of HTML 5, representative frames from a video can be grabbed on the fly and compared with other visual content for an in depth content description.

In the present thesis, role of social tagging was explored to annotate video segments and key frames in two scenarios, 1) leveraging social tags from Flickr while recommending tags based on visually similar images, 2) exploiting user comments for entities and events for event videos. Both the use cases are easily implementable in fact some of the event organisers have already started to utilise real time user content to get event feedbacks. Though, use of tweets to annotate videos for retrieval purpose have not yet been seriously worked on, but the flexibility of the approach in addressing such a complex problem makes it an attractive proposition.

Evaluation against a state-of–the–art baseline is an ideal scientific approach, but unfortunately in social video domain, a benchmark dataset for tag prediction is missing till now. As a consequence we took the user provided tags as the baseline to compare with.

10.2.6 Tag Suggestion: A Way to Bridge the Semantic Gap

This thesis is an attempt to address the issue of "Semantic gap" exists in multimedia domain by exploiting various contextual sources of information. The problem of "Semantic gap" is addressed through automatic tag/concept detection from the video context. Context of a video implies the information sources surrounding a video, where the video or its related attributes are directly or indirectly discussed, described or referred. Definition of the context also includes the information about the creator and capturing device of the video.

Various experimental studies conducted as part of the thesis work showed that there is no "one-fits-all" approach to address the issue of semantic gap. While there are multiple contextual sources exist for a web video, their relevance may vary depending on the application domain for e.g our visual concept learning framework works better with static concept /object oriented videos or travel domain videos where video key frames can be matched against user photos uploaded on the web. Whereas event related videos can be best understood from user activity streams (discussions, opinions and comments). Educational and topical videos can best understood by exploiting user comments, interest profile and their social network. Despite, some limitations, existing contextual sources contribute valuable content description to a web video at the document level. Three different work with the similar research objective supports the above hypothesis.

1. 72 out of 100 videos could be sufficiently described from the existing social contexts compared to the 22 of original data.
2. 16 out of 20 videos could be enriched compared to the baseline of 9.

3. 5 out of 5 sports video could be described at the scene level (with an 2 minutes interval) compared to the baseline of none (no videos were described at the scene level).
4. Scene level description achieved more than 80% of precision in detecting relevant entities and sub-events.
5. Aggregation of tags from all the frames and temporal intervals over the timeline constitute the document level description.

10.3 Limitations of our work

A post work critical observation shows that challenges came from as many sources as the contributions. We faced many problems in striking a trade-off between domain coverage and ontological complexities. We outline some of the challenges we faced throughout the work:

- As evident in almost all prior research work, modelling an audio-visual domain is complex while satisfying all the requirements and maintaining a lightweight framework. We faced some of these challenges while modelling the media content structure and its class hierarchies. Conceptual clarity in any ontology will result in an efficient model from a knowledge engineering perspective but substantially increases the hierarchal structure of the model leading to more verbose descriptions. On the other hand, simple models increase the chances of semantic ambiguity while doing any complex reasoning. We intentionally bargained with reasoning aspects in favour of making an all-encompassing lightweight framework.
- The second major constraint was that we wanted to make the semantic enrichment process as parameter free as possible so as to propose some generic yet satisfying options. Hence the selection of relevant features is critical to the performance of content enrichment. Filtering and de-noising user-generated data is a challenge in itself. Social media specific content adds to the complexity in terms of natural language processing and entity extraction.
- All the sources are not equal some are more relevant concerning the video content compared to the others, so a proper context weighting mechanism is worth exploring in possible future work.
- Though external contextual sources give a good glimpse about the overall semantic content of the video, low-level content processing is still required to support fine grained annotation.
- Despite many expected benefits, it is a realistic assumption that our Proposed approach of modelling and semantic enrichment from social web may face challenges when the

deployment varies in size and domain. It is worth exploring the difference between the real world requirements and the proposed solution in future.

10.4 Future Directions

CISCO predicted that in 2015 almost 91% of all internet data will be in video. This is an astonishing amount of data which demands various proactive research efforts in the multimedia data domain. While storage may not be a problem in coming years, data analysis, discovery, reusability and data interoperability will be the prime concerns.

As future perspectives, we purpose the following directions and extensions of the present work.

10.4.1 Mapping and Integration of the Video Model to Other Existing Media Models

Lightweight ontologies have proved their popularity (DublinCore, FOAF, SIOC, SKOS etc.) on the web both for their design flexibilities and easy implementation. They can be easily integrated with other existing schemas to enhance the scope of data interoperability across data sources and applications. Further work should focus on creating a middleware framework that can map and integrate the VOW model with other existing media models in use, promoting data interoperability and enabling cross querying of multiple ontologies.

10.4.2 Context Evaluation

The role of context in interpreting meaning is well established in natural language processing and text document understanding, but its influence has so far been understudied for multimedia content understanding. Existing multimedia context-oriented studies are restricted to the visual context to recognise an object or a scene. However in view of Web 2.0 paradigms, various dynamic contexts are created, especially contexts emerging out of online social interaction processes (social networks, user comments, shared links, event calendars, etc.), and these have not been studied well. These new sources need to be explored and evaluated for their usability and reliability in different domains.

As part of future efforts we would like to see a study of the dynamics of a user's online activities and their social networks to understand the nature and type of media objects they create, share and interact with. Studying user profiles, analysis of social networks and time-based interest graphs can help us not only to categorise user videos in a subject hierarchy but also to improve personalised recommendation.

10.4.3 An Annotation and Video Browsing Tool

We implemented a prototypical tool for video annotation and browsing. It needs to be converted into a fully functional application by enhancing its various sub modules of tag prediction, model population and linked data creation, as well as the search module which includes mechanisms for both keyword and structured queries.

10.4.4 Social TV

One recently emerging research topic is Social TV: the integration of the Social Web and television for a more personalised and interactive television experience. It will be interesting to study online user activities and interactions in the context of various television programmes to recommend personalised programme guides. On the other side, more and more televisions are being produced as smart televisions with internet enabled social features embedded within. These social features will allow users to carry out live interactions and sharing of watching habits, preferences sharing and programme recommendations within their social network or co-watching community.

Bibliography

1. Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (CHI '04). ACM, New York, NY, USA, 319-326. DOI=10.1145/985692.985733 http://doi.acm.org/10.1145/985692.985733
2. S. Ahern, M. Davis, D. Eckles, S. King, M. Naaman, R. Nair, M. Spasojevic, and J. Yang ZoneTag: Designing Context-Aware mobile media capture to increase participation. In In: Proceedings of the Pervasive Image Capture and Sharing: New Social Practices and Implications for Technology Workshop (PICS 2006) at the Eighth International Conference of Ubiquitous Computing (UbiComp 2006).
3. Umut Akdemir, Pavan Turaga, and Rama Chellappa. 2008. An ontology based approach for activity recognition from video. In *Proceedings of the 16th ACM international conference on Multimedia* (MM '08). ACM, New York, NY, USA, 709-712. DOI=10.1145/1459359.1459466 http://doi.acm.org/10.1145/1459359.1459466
4. James F. Allen. 1983. Maintaining knowledge about temporal intervals. *Commun. ACM* 26, 11 (November 1983), 832-843. DOI=10.1145/182.358434 http://doi.acm.org/10.1145/182.358434
5. Richard Arndt, Rapha\&\#235;l Troncy, Steffen Staab, Lynda Hardman, and Miroslav Vacura. 2007. COMM: designing a well-founded multimedia ontology for the web. In *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference* (ISWC'07/ASWC'07), Busan, South Korea.
6. Morgan Ames and Mor Naaman. 2007. Why we tag: motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI conference on Human factors in computing systems*(CHI '07). ACM, New York, NY, USA, 971-980. DOI=10.1145/1240624.1240772 http://doi.acm.org/10.1145/1240624.1240772
7. Amazon mechanical turk. http://www.mturk.com/mturk
8. Hrishikesh Aradhye, George Toderici, Jay Yagnik, "Video2Text: Learning to Annotate Video Content," Data Mining Workshops, International Conference on, pp. 144-151, 2009 IEEE International Conference on Data Mining Workshops, 2009
9. Sakire Arslan Ay, Roger Zimmermann, and Seon Ho Kim. 2008. Viewable scene modeling for geospatial video search. In *Proceedings of the 16th ACM international conference on Multimedia*(MM '08). ACM, New York, NY, USA, 309-318. DOI=10.1145/1459359.1459401 http://doi.acm.org/10.1145/1459359.1459401
10. Liang Bai, Songyang Lao, Gareth J. F. Jones, and Alan F. Smeaton. 2007. Video Semantic Content Analysis based on Ontology. In *Proceedings of the International Machine Vision and Image Processing Conference* (IMVIP '07). IEEE Computer Society, Washington, DC, USA, 117-124. DOI=10.1109/IMVIP.2007.44 http://dx.doi.org/10.1109/IMVIP.2007.44
11. Lamberto Ballan, Marco Bertini, Alberto Del Bimbo, and Giuseppe Serra. 2010. Semantic annotation of soccer videos by visual instance clustering and spatial/temporal reasoning in ontologies. *Multimedia Tools Appl.* 48, 2 (June 2010), 313-337. DOI=10.1007/s11042-009-0342-4 http://dx.doi.org/10.1007/s11042-009-0342-4
12. Herbert Bay, Andreas Ess, Tinne Tuytelaars, Luc Van Gool, "SURF: Speeded Up Robust Features", Computer Vision and Image Understanding (CVIU), Vol. 110, No. 3, pp. 346--359, 2008
13. Tim Berners-Lee (2006-07-27). "Linked Data—Design Issues". W3C. Retrieved 2010-12-18.
14. Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web: Scientific american. Scientific American, May 2001.
15. Tim Berners-Lee, Yuhsin Chen, Lydia Chilton, Dan Connolly, Ruth Dhanaraj, James Hollenbach, Adam Lerer, and David Sheets In: *Proceedings of the 3rd International Semantic Web User Interaction* (2006) .
16. Bizer, C., Cyganiak, R., Heath, T. (2007). How to publish Linked Data on the Web. Retrieved June 20, 2010, http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/

17. Bizer, Chris and Gauß, Tobias .Disco - hyperdata browser , 2007 at http://www4.wiwiss.fu-berlin.de/bizer/ng4j/disco/ accessed on 12/March 2010.

18. John G. Breslin, Andreas Harth, Uldis Bojars, Stefan Decker. Towards Semantically Interlinked Online Communities. 2nd European Semantic Web Conference, Heraklion, Greece, May 29 to June 1, 2005, pp. 500-514. LNCS 3532.

19. D. Brickley and R. V. Guha. RDF vocabulary description language 1.0: RDF Schema.Technical report, W3C, February 2004. URL http://www.w3.org/TR/rdf-schema/.

20. Marco Bertini, Alberto Del Bimbo, and Carlo Torniai. 2005. Automatic video annotation using ontologies extended with visual information. In *Proceedings of the 13th annual ACM international conference on Multimedia* (MULTIMEDIA '05). ACM, New York, NY, USA, 395-398.
21. Marco Bertini, Alberto Del Bimbo, and Giuseppe Serra. 2008. Learning ontology rules for semantic video annotation. In *Proceedings of the 2nd ACM workshop on Multimedia semantics* (MS '08). ACM, New York, NY, USA, 1-8.
22. S. Boll, T.Burger, O. Celma, C. Halaschek-Wiener, E. Mannens and, R. Troncy (2007) Multimedia vocabularies on the SemanticWeb. In: W3C incubator group report, 24 July 2007.
23. Christopher J. C. Burges. 1998. A Tutorial on Support Vector Machines for Pattern Recognition.*Data Min. Knowl. Discov.* 2, 2 (June 1998), 121-167. DOI=10.1023/A:1009715923555 http://dx.doi.org/10.1023/A:1009715923555
24. Caltech dataset available: http://www.vision.caltech.edu/Image_Datasets/Caltech101/
25. M.Cristani, A.Perina, Umberto Castellani,and V.Murino. ***Geo-located image analysis using latent representations.*** In proceedings of IEEE Computer Vision and Pattern Recognition (CVPR) 2008, Anchorage, Alaska
26. Oscar Celma and Xavier Serra. 2008. FOAFing the music: Bridging the semantic gap in music recommendation. *Web Semant.* 6, 4 (November 2008), 250-256.
27. Ciro Cattuto, Dominik Benz, Andreas Hotho, and Gerd Stumme. 2008. Semantic Grounding of Tag Relatedness in Social Bookmarking Systems. In *Proceedings of the 7th International Conference on The Semantic Web* (ISWC '08, Springer-Verlag, Berlin, Heidelberg, 615-631.
28. C. Cattuto, A. Baldassarri, V. D. P. Servedio, and V. Loreto. Emergent community structure in social tagging systems. In Proceedings of the European Confeence on Complex Systems, Dresden, Germany, October 2007.
29. Liangliang Cao, Jiebo Luo, and Thomas S. Huang. 2008. Annotating photo collections by label propagation according to multiple similarity cues. In *Proceedings of the 16th ACM international conference on Multimedia* (MM '08). ACM, New York, NY, USA, 121-130.
30. Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. 2007. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement* (IMC '07). ACM, New York, NY, USA, 1-14.
31. S.F.Chang, T. Sikora, and A. Puri. Overview of the MPEG-7 standard. IEEE Transactions on Circuits and Systems for Video Technology, 11(6):688–695, 2001.
32. Smitashree Choudhury, John G. Breslin, and Alexandre Passant. 2009. Enrichment and Ranking of the YouTube Tag Space and Integration with the Linked Data Cloud. In *Proceedings of the 8th International Semantic Web Conference* (ISWC '09), Springer-Verlag, Berlin, Heidelberg, 747-762.
33. X. Cheng ,C. Dale and J.C. Liu . "Statistics and Social Network of YouTube Videos " Quality of Service, 2008. IWQoS 2008. 16th International Workshop on In Quality of Service, 2008. IWQoS 2008. 16th International Workshop on (June 2008), pp. 229-238.
34. Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. 2008. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.* 40, 2, Article 5 (May 2008), 60 pages. DOI=10.1145/1348246.1348248 http://doi.acm.org/10.1145/1348246.1348248

35. G. Davenport, T. G. A. Smith, and N. Pincever, “Cinematic principles for multimedia,” *IEEE* Computer Graphics & Applications, vol. 11, pp. 67–74,1991.
36. Stamatia Dasiopoulou, Vassilis Tzouvaras, Ioannis Kompatsiaris, and Michael G. Strintzis. 2010. Enquiring MPEG-7 based multimedia ontologies. *Multimedia Tools Appl.* 46, 2-3 (January 2010), 331-370.
37. M. Dean and G. Schreiber. OWL Web Ontology Language reference. recommendation,W3C, February 2004. URL http://www.w3.org/TR/2004/REC-owl-ref-20040210/.
38. P.J. Denning,.(2006) Infoglut. Communications of the ACM*,* 49 (7).
39. Gregory D. Abowd, Anind K. Dey, Peter J. Brown, Nigel Davies, Mark Smith, and Pete Steggles. 1999. Towards a Better Understanding of Context and Context-Awareness. In *Proceedings of the 1st international symposium on Handheld and Ubiquitous Computing* (HUC '99), Springer-Verlag, London, UK, 304-307
40. N. Dimitrova, L. Agnihotri, and G. Wei. Video classification based on HMM using text and faces. European Signal Processing Conference, Tampere, Finland., 2000.
41. S. Dasiopoulou, V. Tzouvaras, I. Kompatsiaris, and M.G. Strintzis. Enquiring MPEG-7 based multimedia ontologies. Multimedia Tools and Applications. 2009;46(2-3):331-370. Available at: http://www.springerlink.com/index/10.1007/s11042-009-0387-4.
42. Y. Ding. Upper Tag Ontology for Integrating Social Tagging Data, Journal of the American Society for Information Science and Technology, Vol. 9999, No. 9999., 2009.
43. P. Duygulu, Kobus Barnard, J. F. G. de Freitas, and David A. Forsyth. 2002. Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. In *Proceedings of the 7th European Conference on Computer Vision-Part IV* (ECCV '02), Springer-Verlag, London, UK, UK, 97-112.
44. S. Ebadollahi, L. Xie, S.-F. Chang, and J. R. Smith, “Visual event detection using multi-dimensional concept dynamics,” in Proceedings of the IEEE International Conference on Multimedia & Expo, pp. 881–884, Toronto, Canada,2006.
45. Boris Epshtein, Eyal Ofek, Yonatan Wexler, and Pusheng Zhang. 2007. Hierarchical photo organization using geo-relevance. In *Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems* (GIS '07). ACM, New York, NY, USA.
46. R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from googles image search. ICCV, 2005.
47. P.M. Fergal, “Context-Aware Photograph Annotation on the Semantic Web” (PhD Thesis), National University of Ireland, Galway, 2010.
48. Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. Annotating named entities in Twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*(CSLDAMT '10). Association for Computational Linguistics, Stroudsburg, PA, USA, 80-88.
49. J.R. Firth (1957). A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*, pp. 1–32. Oxford: Philological Society. Reprinted in F.R. Palmer (ed.), Selected Papers of J.R. Firth 1952-1959, London: Longman (1968).
50. Alexandre R. J. Francois, Ram Nevatia, Jerry Hobbs, and Robert C. Bolles. 2005. VERL: An Ontology Framework for Representing and Annotating Video Events. *IEEE MultiMedia* 12, 4 (October 2005), 76-86.
51. G W. Furnas, T K. Landauer, L M. Gomez, and S T. Dumais. 1984. Statistical semantics: analysis of the potential performance of keyword information systems. In *Human factors in computer systems*, John C. Thomas and Michael L. Schneider (Eds.). Ablex Publishing Corp., Norwood, NJ, USA 187-242.
52. Nikhil Garg and Ingmar Weber. 2008. Personalized, interactive tag recommendation for flickr. In*Proceedings of the 2008 ACM conference on Recommender systems* (RecSys '08). ACM, New York, NY, USA, 67-74.

53. Gallagher, D. Joshi, J. Yu, and J. Luo (2009) Geo-location inference from image content and user tags. In Proceedings of the IEEE Workshop on Internet Vision (CVPR).
54. R. Garcia., O. Celma. Semantic Integration and Retrieval of Multimedia Metadata.In: 5th International Workshop on Knowledge Markup and Semantic Annotation (2005) Galway, Ireland.
55. U. Gargi, W. Lu, V. Mirrokni, and S. Yoon. Large-Scale Community Detection on YouTube for Topic Discovery and Exploration. In Proceedings of the AAAI Conference on Weblogs and Social Media 2011, Barcelona, Spain*, 2011*.
56. Gary Geisler and Sam Burns. 2007. Tagging video: conventions and strategies of the YouTube community. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries* (JCDL '07). ACM, New York, NY, USA, 480-480.
57. Scott A. Golder and Bernardo A. Huberman. 2006. Usage patterns of collaborative tagging systems. *J. Inf. Sci.* 32, 2 (April 2006), 198-208.
58. T. Gruber, (2005). TagOntology - a way to agree on the semantics of tagging data. Presentation to Tag Camp, Palo Alto, CA, October 29, 2005. http://tomgruber.org/writing/tagontology.htm.
59. Thomas R. Gruber. 1993. A translation approach to portable ontology specifications. *Knowl. Acquis.* 5, 2 (June 1993), 199-220.
60. T. Gruber (2006). "Where the Social Web Meets the Semantic Web". Keynote presentation at ISWC, The 5th International Semantic Web Conference, November 7, 2006
61. Nicola Guarino and Christopher Welty. 2002. Evaluating ontological decisions with OntoClean.*Commun. ACM* 45, 2 (February 2002), 61-65.
62. C. Halaschek-Wiener and A. Schain. A flexible approach for managing digital images on the semantic web. In Proceedings of the 5th Knowledge Markup and Semantic Annotation Workshop, volume 185 of SemAnnot05, 2005.
63. Harry Halpin, Valentin Robu, and Hana Shepherd. 2007. The complex dynamics of collaborative tagging. In *Proceedings of the 16th international conference on World Wide Web* (WWW '07). ACM, New York, NY, USA, 211-220.
64. Martin J. Halvey and Mark T. Keane. 2007. Exploring social dynamics in online media sharing. In*Proceedings of the 16th international conference on World Wide Web* (WWW '07). ACM, New York, NY, USA, 1273-1274.
65. X. W, A. Hauptmann A, and C. Ngo (2007) Practical elimination of near- duplicates from web video search. In: Int'l conference on multi- media
66. L. Hollink, S. Little, and J. Hunter. 2005. Evaluating the application of semantic inferencing rules to image annotation. In *Proceedings of the 3rd international conference on Knowledge capture*(K-CAP '05). ACM, New York, NY, USA, 91-98.
67. Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. 2006. Information retrieval in folksonomies: search and ranking. In *Proceedings of the 3rd European conference on The Semantic Web: research and applications* (ESWC'06), Springer-Verlag, Berlin, Heidelberg, 411-426.
68. Hoogs, J. Rittscher, G. Stein, and J. Schmiederer, Video content annotation using visual analysis and a large semantic knowledgebase. in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 327–334, Madison, USA, 2003.
69. J. Hunter (2001) Adding multimedia to the Semantic Web: building an MPEG-7 ontology. In: Proc the first Semantic Web working symposium, SWWS'01, Stanford University, California, USA.
70. Jane Hunter and Liz Armstrong. 1999. A comparison of schemas for video metadata representation. *Comput. Netw.* 31, 11-16 (May 1999), 1431-1451.
71. J. Huang, Z. Liu, Y. Wang, Y., Chen, and E. Wong. Integration of multimodal features for video scene classification based on HMM. In IEEE Workshop Multimedia Signal Processing (MMSP-99) Copenhagen, Denmark, 1999.
72. Jing Huang, S. Ravi Kumar, Mandar Mitra, Wei-Jing Zhu, and Ramin Zabih. 1997. Image Indexing Using Color Correlograms. In *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)* (CVPR '97). IEEE Computer Society, Washington, DC, USA.

73. Chunneng Huang, Tianjun Fu, and Hsinchun Chen. 2010. Text-based video content classification for online video-sharing sites. *J. Am. Soc. Inf. Sci. Technol.* 61, 5 (May 2010), 891-906.
74. David F. Huynh, David R. Karger, and Robert C. Miller. 2007. Exhibit: lightweight structured data publishing. In *Proceedings of the 16th international conference on World Wide Web* (WWW '07). ACM, New York, NY, USA, 737-746.
75. Anil K. Jain, Robert P. W. Duin, and Jianchang Mao. 2000. Statistical Pattern Recognition: A Review. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 1 (January 2000), 4-37.
76. Ramesh Jain and Arun Hampapur. 1994. Metadata in video databases. *SIGMOD Rec.* 23, 4 (December 1994), 27-33. DOI=10.1145/190627.190638 http://doi.acm.org/10.1145/190627.190638
77. Bernard J. Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. 2009. Twitter power: Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci. Technol.* 60, 11 (November 2009), 2169-2188.
78. J. Jeon, V. Lavrenko, and R. Manmatha. 2003. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (SIGIR '03). ACM, New York, NY, USA, 119-126.
79. Feng Kang, Rong Jin, and Rahul Sukthankar. 2006. Correlated Label Propagation with Application to Multi-label Learning. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2* (CVPR '06), Vol. 2. IEEE Computer Society, Washington, DC, USA, 1719-1726.
80. M. Kankanhalli and Y. Rui, "Application potential of multimedia information retrieval," *Proceedings of the IEEE*, vol. 96, pp. 712–720, 2008.
81. Lyndon S. Kennedy, Shih-Fu Chang, and Igor V. Kozintsev. 2006. To search or to label?: predicting the performance of search-based automatic image classifiers. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval* (MIR '06). ACM, New York, NY, USA, 249-258.
82. Hak Lae Kim, Simon Scerri, John G. Breslin, Stefan Decker, and Hong Gee Kim. 2008. The state of the art in tag ontologies: a semantic model for tagging and folksonomies. In *Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications* (DCMI '08). Dublin Core Metadata Initiative 128-137.
83. David Kirk, Abigail Sellen, Carsten Rother, and Ken Wood. 2006. Understanding photowork. In*Proceedings of the SIGCHI conference on Human Factors in computing systems* (CHI '06), Rebecca Grinter, Thomas Rodden, Paul Aoki, Ed Cutrell, Robin Jeffries, and Gary Olson (Eds.). ACM, New York, NY, USA, 761-770.
84. Josef Kittler, Mohamad Hatef, Robert P. W. Duin, and Jiri Matas. 1998. On Combining Classifiers.*IEEE Trans. Pattern Anal. Mach. Intell.* 20, 3 (March 1998), 226-239.
85. T. Knerr. Tagging ontology - towards a common ontology for folksonomies, 2006. Available at: http://code.google.com/p/tagont/ (accessed January 2011).
86. C. Körner. Understanding the motivation behind tagging. ACM Student Research Competition - Hypertext 2009, July 2009.
87. J. Kustanowitz , B. Shneiderman (2005) Motivating annotation for personal digital photo libraries: Lowering barriers while raising incentives. Tech rep, University of Maryland.
88. Sare Gul Sevil, Onur Kucuktunc, Pinar Duygulu, and Fazli Can. 2010. Automatic tag expansion using visual similarity for photo sharing websites. *Multimedia Tools Appl.* 49, 1 (August 2010), 81-99.
89. Kunfeng Lai and Dan Wang. 2010. Towards understanding the external links of video sharing sites: measurement and analysis. In *Proceedings of the 20th international workshop on Network and operating systems support for digital audio and video* (NOSSDAV '10). ACM, New York, NY, USA, 69-74
90. Renaud Lambiotte and Marcel Ausloos. 2006. Collaborative tagging as a tripartite network. In*Proceedings of the 6th international conference on Computational Science - Volume Part III*(ICCS'06), Springer-Verlag, Berlin, Heidelberg, 1114-1117.
91. Liza Leslie, Tat-Seng Chua, and Jain Ramesh. 2007. Annotation of paintings with high-level semantic concepts using transductive inference and ontology-based concept disambiguation. In*Proceedings of*

the 15th international conference on Multimedia (MULTIMEDIA '07). ACM, New York, NY, USA, 443-452.

92. Till Quack, Bastian Leibe, and Luc Van Gool. 2008. World-scale mining of objects and events from community photo collections. In *Proceedings of the 2008 international conference on Content-based image and video retrieval* (CIVR '08). ACM, New York, NY, USA, 47-56.
93. Huairen Lin, Joseph Davis, and Ying Zhou. 2009. An Integrated Approach to Extracting Ontological Structures from Folksonomies. In *Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications* (ESWC 2009 Heraklion), Springer-Verlag, Berlin, Heidelberg, 654-668.
94. S. Liu, L.T. Chia, and S. Chan, "Ontology for nature scene image retrieval," In Proceedings OTM confederated international conferences, Oct. 2004
95. Q. Li, S. Lu. Collaborative tagging applications and approaches. IEEE Multimedia 15(3), 14{21(2008).
96. X.R.Li, C.G.M. Snoek, and M.Worring. Learning tag relevance by neighbour voting for social image retrieval. In Proc. of the ACM International Conference on Multimedia Information Retrieval (2008).
97. Jia Li and James Z. Wang. 2008. Real-Time Computerized Annotation of Pictures. *IEEE Trans. Pattern Anal. Mach. Intell.* 30, 6 (June 2008), 985-1002.
98. Jia Li and James Z. Wang. 2003. Automatic Linguistic Indexing of Pictures by a Statistical Modeling Approach. *IEEE Trans. Pattern Anal. Mach. Intell.* 25, 9 (September 2003), 1075-1088.
99. uhong Liu, Ouri Wolfson, and Huabei Yin. 2006. Extracting Semantic Location from Outdoor Positioning Systems. In *Proceedings of the 7th International Conference on Mobile Data Management* (MDM '06). IEEE Computer Society, Washington, DC, USA.
100. Ken-Hao Liu, Ming-Fang Weng, Chi-Yao Tseng, Yung-Yu Chuang, and Ming-Syan Chen. 2008. Association and Temporal Rule Mining for Post-Filtering of Semantic Concept Detection in Video.*Trans. Multi.* 10, 2 (February 2008), 240-251.
101. David G. Lowe. 2004. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vision* 60, 2 (November 2004), 91-110.
102. Lozano-Tello and A. Gomez-Perez. ONTOMETRIC: A method to choose the appropriate ontology. Journal of Database Management, 15(2):1{18, April-June 2004.
103. Mathes. Folksonomies - cooperative classification and commu nication through shared metadata, December 2004. Available at: http://www.adammathes.com/academic/computer-mediated-communication/ folksonomies.html .
104. Manjunath, P. Salembier, and T. Sikora. Introduction to MPEG-7: Multimedia Content Description Language. Wiley, 2002.
105. Mathias Lux and Savvas A. Chatzichristofis. 2008. Lire: lucene image retrieval: an extensible java CBIR library. In *Proceedings of the 16th ACM international conference on Multimedia* (MM '08). ACM, New York, NY, USA, 1085-1088.
106. L. Mcguinness. Ontologies Come of Age. MIT Press, 2003..http://www.ksl.stanford.edu/people/dlm/papers/ontologies-come-of-age-mit-press-(with-citation).htm.
107. G.C. Michael. Carnegie Mellon University traditional informedia digital video retrieval system. In *Proceedings of the 6th ACM international conference on Image and video retrieval* (CIVR '07). ACM, New York, NY, USA, 2007.
108. Peter Mika. 2007. Ontologies are us: A unified model of social networks and semantics. *Web Semant.* 5, 1 (March 2007), 5-15.
109. Florent Monay and Daniel Gatica-Perez. 2003. On image auto-annotation with latent space models. In *Proceedings of the eleventh ACM international conference on Multimedia* (MULTIMEDIA '03). ACM, New York, NY, USA, 275-278.
110. Fergal Monaghan and David O'Sullivan. 2006. Automating Photo Annotation using Services and Ontologies. In *Proceedings of the 7th International Conference on Mobile Data Management* (MDM '06). IEEE Computer Society, Washington, DC, USA.

111. Emily Moxley, Jim Kleban, Jiejun Xu, and B. S. Manjunath. 2009. Not all tags are created equal: learning Flickr tag semantics for global annotation. In *Proceedings of the 2009 IEEE international conference on Multimedia and Expo* (ICME'09). IEEE Press, Piscataway, NJ, USA, 1452-1455.
112. M. R. Naphade and T. S. Huang, "A probabilistic framework for semantic video indexing, filtering and retrieval," IEEE Transactions on Multimedia, vol. 3, pp. 141–151, 2001.
113. Milind Naphade, John R. Smith, Jelena Tesic, Shih-Fu Chang, Winston Hsu, Lyndon Kennedy, Alexander Hauptmann, and Jon Curtis. 2006. Large-Scale Concept Ontology for Multimedia. *IEEE MultiMedia* 13, 3 (July 2006), 86-91.
114. Mor Naaman, Yee Jiun Song, Andreas Paepcke, and Hector Garcia-Molina. 2004. Automatic organization for digital photographs with geographic coordinates. In *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries* (JCDL '04). ACM, New York, NY, USA, 53-62.
115. B. Neumann, and R. Moeller. On scene interpretation with description logics. In: Cognitive Vision Systems: Sampling the Spectrum of Approaches, Lecture Notes in Computer Science, vol. 3948, pp. 247–278. Springer (2006).
116. R. Newman, D.Ayers, and S. Russell. Tag ontology, December 2005. Available at: http://www.holygoat.co.uk/owl/redwood/0.1/tags/
117. Oded Nov and Chen Ye. 2010. Why do people tag?: motivations for photo tagging. *Commun. ACM*53, 7 (July 2010), 128-131.
118. N. F. Noy, & D. L. McGuinness,. Ontology Development 101: A Guide to Creating Your First Ontology *Stanford Knowledge Systems Laboratory and Stanford Medical Informatics,* 2001
119. Jacco van Ossenbruggen, Frank Nack, and Lynda Hardman. 2004. That Obscure Object of Desire: Multimedia Metadata on the Web, Part 1. *IEEE MultiMedia* 11, 4 (October 2004), 38-48.
120. Parikh, C. Zitnick, and T. Chen. From Appearance to Context-Based Recog- nition: Dense Labeling in Small Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8, June 2008.
121. Passant, T. Hastrup, U. Bojars, and J.G. Breslin, "Microblogging: A Semantic Web and Distributed Approach", The 4th Workshop on Scripting for the Semantic Web (SFSW 2008) at the 5th European Semantic Web Conference (ESWC '08), Tenerife, Spain, 2008.
122. Passant and P. Laublet. Meaning of a tag: A collaborative approach to bridge the gap between tagging and linked data. In Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW2008), Beijing, China, April 2008.
123. Theo Pavlidis. 2009. Why meaningful automatic tagging of images is very hard. In *Proceedings of the 2009 IEEE international conference on Multimedia and Expo* (ICME'09). IEEE Press, Piscataway, NJ, USA, 1432-1435.
124. Pigeau and M. Gelgon. 2005. Building and tracking hierarchical geographical \& temporal partitions for image collection management on mobile devices. In *Proceedings of the 13th annual ACM international conference on Multimedia* (MULTIMEDIA '05). ACM, New York, NY, USA, 141-150.
125. M.R. Quillian. Semantic memory. In Semantic Information Processing, M. Minsky (ed.), MIT Press, 227-270 (1968).
126. Lawrence R. Rabiner. 1990. A tutorial on hidden Markov models and selected applications in speech recognition. In *Readings in speech recognition*, Alex Waibel and Kai-Fu Lee (Eds.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA 267-296.
127. Tye Rattenbury, Nathan Good, and Mor Naaman. 2007. Towards extracting flickr tag semantics. In *Proceedings of the 16th international conference on World Wide Web* (WWW '07). ACM, New York, NY, USA, 1287-1288.
128. J. Rowley, (1995). Organizing Knowledge. 2nd Ed. Brookfield, VT: Gower.
129. J. Rowley. The wisdom hierarchy: representations of the dikw hierarchy. Journal of Information Science, 33(2):163-180, April 2007. URL http://jis.sagepub.com/ cgi/reprint/33/2/163.
130. M. Rowe, and M. Stankovic, "Mapping Tweets to Conference Talks: A Goldmine for Semantics", Proceedings of the 3rd International Workshop on Social Data on the Web (SDOW 2010) at the 9th International Semantic Web Conference (ISWC 2010), 2010.

131. Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. 2008. LabelMe: A Database and Web-Based Tool for Image Annotation. *Int. J. Comput. Vision* 77, 1-3 (May 2008), 157-173

132. Carsten Saathoff and Steffen Staab. 2008. Exploiting Spatial Context in Image Region Labelling Using Fuzzy Constraint Reasoning. In *Proceedings of the 2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services* (WIAMIS '08). IEEE Computer Society, Washington, DC, USA, 16-19.

133. David A. Shamma, Lyndon Kennedy, and Elizabeth F. Churchill. 2009. Tweet the debates: understanding community annotation of uncollected sources. In *Proceedings of the first SIGMM workshop on Social media* (WSM '09). ACM, New York, NY, USA, 3-10.

134. Ansgar Scherp, Thomas Franz, Carsten Saathoff, and Steffen Staab. 2009. F--a model of events based on the foundational ontology dolce+DnS ultralight. In *Proceedings of the fifth international conference on Knowledge capture* (K-CAP '09). ACM, New York, NY, USA, 137-144.

135. Alan F. Smeaton. 2005. Large scale evaluations of multimedia information retrieval: the TRECVid experience. In *Proceedings of the 4th international conference on Image and Video Retrieval* (CIVR'05), Springer-Verlag, Berlin, Heidelberg, 11-17.

136. Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. 2000. Content-Based Image Retrieval at the End of the Early Years. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 12 (December 2000), 1349-1380.

137. Lucia Specia and Enrico Motta. 2007. Integrating Folksonomies with the Semantic Web. In*Proceedings of the 4th European conference on The Semantic Web: Research and Applications*(ESWC '07), Springer-Verlag, Berlin, Heidelberg, 624-639.

138. L. Snidaro, M. Belluz, and G. Foresti. Domain knowledge for surveillance applications. In:Proc. of Int'l Conference on Information Fusion (2007).

139. Jagan Sankaranarayanan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Sperling. 2009. TwitterStand: news in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (GIS '09). ACM, New York, NY, USA, 42-51.

140. G. Schindler, L. Zitnick, and M. Brown, "Internet Video Category Recognition," in Proc. First Internet Vision Workshop, Jun. 2008, pp. 1–7

141. Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. 2010. Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*(SIGIR '10). ACM, New York, NY, USA, 841-842.

142. P. Schmitz. Inducing ontology from Flickr tags. In Proc. of the Collaborative WebTagging Workshop (WWW '06). 2006. (Cited on pages 33, 35 and 36)

143. Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web* (WWW '10). ACM, New York, NY, USA, 851-860.

144. G. Furnas, T. Landauer, L. Gomez, S. Dumais (1984) Statistical semantics: analysis of the potential performance of keyword information systems. Proc SIGCHI Conf Hum Factors Comput Syst, 187–242

145. Arjan T. Setz and Cees G. M. Snoek. 2009. Can social tagged images aid concept-based video search?. In *Proceedings of the 2009 IEEE international conference on Multimedia and Expo*(ICME'09). IEEE Press, Piscataway, NJ, USA, 1460-1463.

146. Börkur Sigurbjörnsson and Roelof van Zwol. 2008. Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th international conference on World Wide Web* (WWW '08). ACM, New York, NY, USA, 327-336.

147. Schmitz, A. Hotho, R. Jäschke, G. Stumme. Mining Association Rules in Folksonomies. In Data Science and Classification, pp. 261–270. Springer, 2006. (Cited on pages 35, 36, 77 and 78).

148. Bernhard Scholkopf and Alexander J. Smola. 2001. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA.
149. Cees G. M. Snoek, Marcel Worring, and Arnold W. M. Smeulders. 2005. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia* (MULTIMEDIA '05). ACM, New York, NY, USA, 399-402.
150. G.M. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring. 2007. Adding Semantics to Detectors for Video Retrieval. *Trans. Multi.* 9, 5 (August 2007), 975-986.
151. G.M. Snoek and M. Worring. 2005. Multimedia event-based video indexing using time intervals.*Trans. Multi.* 7, 4 (August 2005), 638-647. DOI=10.1109/TMM.2005.850966 http://dx.doi.org/10.1109/TMM.2005.850966
152. Cees G. M. Snoek and Marcel Worring. 2009. Concept-Based Video Retrieval. *Found. Trends Inf. Retr.* 2, 4 (April 2009), 215-322. DOI=10.1561/1500000014 http://dx.doi.org/10.1561/1500000014
153. Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. 2000. Content-Based Image Retrieval at the End of the Early Years. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 12 (December 2000), 1349-1380.
154. M.C. schraefel, Max Wilson, Alistair Russell, and Daniel A. Smith. 2006. mSpace: improving information access to multimedia domains with multimodal exploratory search. *Commun. ACM* 49, 4 (April 2006), 47-49.
155. Carsten Saathoff and Ansgar Scherp. 2010. Unlocking the semantics of multimedia presentations in the web with the multimedia metadata ontology. In *Proceedings of the 19th international conference on World wide web* (WWW '10). ACM, New York, NY, USA, 831-840.
156. Chrisa Tsinaraki, Panagiotis Polydoros, and Stavros Christodoulakis. 2007. Interoperability Support between MPEG-7/21 and OWL in DS-MIRF. *IEEE Trans. on Knowl. and Data Eng.* 19, 2 (February 2007), 219-232.
157. C. Tsinaraki, P. Polydoros, and S. Christodoulakis, S.: Interoperability support for Ontology-based Video Retrieval Applications. In: 3rd International Conference on Image and Video Retrieval (CIVR), pp. 582–591 (2004).
158. J. Tanaka, and M. Taylor, (1991). Object Categories and Expertise: Is the Basic Level in the Eye of the Beholder? Cognitive Psychology 23(3). 457-482.
159. V. S. Tseng, Ja-Hwung Su, Jhih-Hong Huang, and Chih-Jen Chen. 2008. Integrated Mining of Visual Features, Speech Features, and Frequent Patterns for Semantic Video Annotation. *Trans. Multi.* 10, 2 (February 2008), 260-267.
160. TRECVID, http://www-nlpir.nist.gov/projects/trecvid/.
161. TagCommons-Project. Ontologies vs. formats vs. schema vs. apis, March 2007. http://tagcommons.org/2007/03/02/ontologies-vs-formats-vs-schema-vs-apis/ (accessed 8 May 2009).
162. C.M. Tsai, A. Qamra, and E. Chang (2005) Extent: inferring image metadata from context and content. In Proceedings of IEEE ICME.
163. Fernando Perez-Tellez, David Pinto, John Cardiff, and Paolo Rosso. 2010. On the difficulty of clustering company tweets. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents* (SMUC '10). ACM, New York, NY, USA, 95-102.
164. Theodora Tsikrika, Christos Diou, Arjen P. de Vries, and Anastasios Delopoulos. 2009. Image annotation using clickthrough data. In *Proceedings of the ACM International Conference on Image and Video Retrieval* (CIVR '09). ACM, New York, NY, USA.
165. Kentaro Toyama, Ron Logan, and Asta Roseway. 2003. Geographic location tags on digital images. In *Proceedings of the eleventh ACM international conference on Multimedia* (MULTIMEDIA '03). ACM, New York, NY, USA, 156-166.
166. G. Toderici, H. Aradhye, M. Pasca, L. Sbaiz, and J. Yagnik. Finding meaning on youtube: Tag recommendation and category discovery. In CVPR, 2010.
167. Vailaya, M. A.T. Figueiredo, A. K. Jain, and Hong-Jiang Zhang. 2001. Image classification for content-based indexing. *Trans. Img. Proc.* 10, 1 (January 2001), 117-130.

168. S. Vembu, M. Kiesel, M. Sintek, and S. Bauman (2006) Towards bridging the semantic gap in multimedia annotation and retrieval. In: Proc workshop on Semantic Web annotations for multimedia (SWAMM), Edinburgh, Scotland
169. David Vallet, Iván Cantador, and Joemon M. Jose. 2010. Exploiting external knowledge to improve video retrieval. In *Proceedings of the international conference on Multimedia information retrieval*(MIR '10). ACM, New York, NY, USA, 101-110.
170. http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html.
171. Y. Wang, Z. Liu, and J. Huang. Multimedia content analysis using both audio and visual clues. IEEE Signal Processing Magazine, 17(6):12–36, 2000.
172. Kilian Quirin Weinberger, Malcolm Slaney, and Roelof Van Zwol. 2008. Resolving tag ambiguity. In*Proceedings of the 16th ACM international conference on Multimedia* (MM '08). ACM, New York, NY, USA, 111-120.
173. W. Weaver (1955). Translation. In W.N. Locke and D.A. Booth (eds.), Machine Translation of Languages, Cambridge, MA: MIT Press. ISBN 0-8371-8434-7.
174. WordNet. http://wordnet.princeton.edu/.
175. Lior Wolf and Stanley Bileschi. 2006. A Critical View of Context. *Int. J. Comput. Vision* 69, 2 (August 2006), 251-261.
176. C.Wang, F. Jing , L. Zhang, and H.J. Zhang (2007) Content-based image annotation refinement. In proceedings of Computer Vision and Pattern Recognition, 2007. CVPR '07
177. C. S. Won, D. K. Park, and S.-J. Park, "Efficient use of MPEG-7 edge histogram descriptor", *ETRI J.*, vol. 24, no. 1, pp.23 - 30 , 2002.
178. Lei Wu, Linjun Yang, Nenghai Yu, and Xian-Sheng Hua. 2009. Learning to tag. In *Proceedings of the 18th international conference on World wide web* (WWW '09). ACM, New York, NY, USA, 361-370.
179. Rong Yan, Jelena Tesic, and John R. Smith. 2007. Model-shared subspace boosting for multi-label classification. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '07). ACM, New York, NY, USA, 834-843.
180. Jun Yang and Alexander G. Hauptmann. 2008. (Un)Reliability of video concept detection. In*Proceedings of the 2008 international conference on Content-based image and video retrieval*(CIVR '08). ACM, New York, NY, USA, 85-94.
181. Ka-Ping Yee, Kirsten Swearingen, Kevin Li, and Marti Hearst. 2003. Faceted metadata for image search and browsing. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (CHI '03). ACM, New York, NY, USA, 401-408.
182. Jude Yew, David A. Shamma, and Elizabeth F. Churchill. 2011. Knowing funny: genre perception and categorization in social video sharing. In *Proceedings of the 2011 annual conference on Human factors in computing systems* (CHI '11). ACM, New York, NY, USA, 297-306.
183. Z.H. Zhou and M.L. Zhang. Multi-instance multi-label learning with application to scene classification. In Proceedings of Neural Information Processing Systems, 2006.
184. Zheng-Jun Zha, Tao Mei, Zengfu Wang, and Xian-Sheng Hua. 2007. Building a comprehensive ontology to refine video concept detection. In *Proceedings of the international workshop on Workshop on multimedia information retrieval* (MIR '07). ACM, New York, NY, USA, 227-236.
185. L. Zelnik-Manor, S. Zanetti, and P. Perona, "A WalkThrough the Web's Video Clips," in Proc. First Internet Vision Work- shop, Jun. 2008, pp. 1–7.
186. Sorokin, A., Forsyth, D.: Utility data annotation with amazon mechanical turk. Urbana 51 (2008) 61820
187. Yuen, J., Russell, B., Liu, C., Torralba, A.: LabelMe video: Building a Video Database with Human Annotations. (2009)
188. Carl Vondrick, Deva Ramanan, and Donald Patterson. 2010. Efficiently scaling up video annotation with crowdsourced marketplaces. In *Proceedings of the 11th European conference on Computer vision: Part IV* (ECCV'10), Springer-Verlag, Berlin, Heidelberg, 610-623.