

# Multimedia web information fusion and analysis

Tao, Jiang

2008

Tao, J. (2008). Multimedia web information fusion and analysis. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/2602>

<https://doi.org/10.32657/10356/2602>

---

Nanyang Technological University

*Downloaded on 24 Aug 2022 19:45:44 SGT*

# Multimedia Web Information Fusion and Analysis



**Tao Jiang**  
**School of Computer Engineering**

A thesis submitted to the Nanyang Technological University  
in fulfillment of the requirement for the degree  
of Doctor of Philosophy

2007

# Abstract

Information on the World Wide Web appears in diverse forms, including text, image, audio, and video. Presented with a wide range of information, an information user often takes great effort to correlate and track online information related to specific topics of interest. Fusion of multimedia information in a unified framework is thus needed for efficiently understanding and further analyzing the semantically related information. This thesis addresses the problem of *multimedia web information fusion and analysis* by presenting an approach for modelling multimedia information in a unified semantic framework, based on which cross-media information analysis and mining is realized.

As multimedia data are heterogeneous in their contents and formats, we employ a strategy for **multimedia information fusion** based on semantics of the data. Specifically, we develop two methods, one using a statistical vague transformation technique and the other employing a self-organizing neural network, to associate web images with related surrounding texts, based on which the semantics of the media objects can be extracted. Our experiments show that the proposed methods can identify associated image and text pairs with good accuracy and outperform a state-of-the-art method for image annotation using a statistical relevance model.

To support cross-media analysis, this thesis develops a **semantic representation schema**, that combines MPEG-7 multimedia description, RDF language specification, and conceptual graph based knowledge representation techniques for modelling multimedia information. In addition, we develop a **semantic metadata extraction** algorithm utilizing a myriad of natural language processing (NLP) techniques to automatically extract concepts and relations from text contents. The extracted concepts are formally represented as bags of WordNet senses, based on which an incremental clustering approach

is applied for organizing the concepts into a taxonomy. The constructed taxonomy, encoded in the form of RDF metadata, is subsequently used for facilitating semantic based multimedia analysis.

For **multimedia analysis**, this thesis presents an algorithm, called GP-Close, for discovering generalized concept-relation association patterns from RDF semantic metadata collection. By adopting the notion of generalization closure, the proposed GP-Close algorithm can eliminate redundant over-generalized patterns during the mining process. We evaluate the GP-Close algorithm on two synthetic data sets and one real-world data set. In addition, a case study is conducted for analyzing an online terror attack document collection. Our experiments show that the proposed method can efficiently identify interesting patterns, effectively remove pattern redundancies, and significantly outperform the existing algorithms in terms of time efficiency.

# Acknowledgments

I would like to express my thanks and gratitude to the following organization and people:

**Nanyang Technological University** for offering me the opportunity to pursue my PhD degree in Singapore, and **Ministry of Education, Singapore** for offering me the scholarship that supported my study at NTU.

I would like to thank my supervisor, **Associate Professor Tan Ah Hwee**, for years of invaluable guidance and encouragement in my study and research. It is him who taught me the way of thinking of problems critically and approaching the solutions with passion. Without his guidance, the thesis would not have been completed.

I would like to thank **Associate Professor Chia Liang Tien** for sharing his ideas and providing valuable suggestions on my research work. I would also like to thank **Professor Wang Ke** for helping me a lot in improving my thinking and writing skills when he was in NTU as a visiting professor.

I would also like to thank **Jiang Xing, Yap Ghim Eng, Woon Kia Yan, Nguyen Luong Dong, Liu Ying**, and **Zhang Hao**, postgraduates and research staffs in the Emerging Research Lab and the Center for Advanced Information Systems for discussions and idea sharing.

In addition, I need to acknowledge **Dr. Rohan Kumar Gunaratna** and **Mr. Ong Teng Kwee** at the S. Rajaratnam School of International Studies (previously known as the Institute of Defence and Strategic Studies) for providing advices on data sources for us to test our developed techniques in real world applications.

I would also like to thank in advance the **thesis examiners** for their insightful comments and suggestions for improving the quality of this thesis.

Special thanks go to **my parents** and **my brother** for all the love, support and sacrifice that they have given me to pursue my PhD degree. Without any of you, all this would not be possible.

Thank you!

# Contents

<b>Abstract</b> . . . . .	i
<b>Acknowledgments</b> . . . . .	iii
<b>List of Figures</b> . . . . .	ix
<b>List of Tables</b> . . . . .	xii
<b>1 Introduction</b>	<b>1</b>
1.1 World Wide Web and Multimedia Information . . . . .	1
1.2 Scenarios of Web Information Service . . . . .	3
1.3 Multimedia Information Systems . . . . .	5
1.4 A Framework for Semantic Based Multimedia Web Information Fusion and Analysis . . . . .	8
1.5 Summary of Contributions . . . . .	11
1.6 Thesis Structure . . . . .	13
<b>2 Related Work</b>	<b>14</b>
2.1 Multimedia Information Fusion . . . . .	14
2.1.1 Multimedia Authoring . . . . .	14
2.1.2 Multimodal Information Fusion for Multimedia Indexing and Re- trieval . . . . .	18
2.2 Semantic Web and Multimedia Modelling . . . . .	20
2.2.1 Semantic Web . . . . .	20
2.2.2 Multimedia Modelling . . . . .	22
2.3 Automatic Multimedia Annotation Extraction . . . . .	23
2.3.1 Automatic Annotation Based on Machine Learning and Statistic Modelling . . . . .	23

2.3.2	Automatic Annotation Based on Multimedia Ontology . . . . .	25
2.4	Multimedia Analysis . . . . .	26
2.4.1	General Framework of Multimedia Data Mining . . . . .	27
2.4.2	Multimedia Data Mining Techniques . . . . .	30
2.5	Summary . . . . .	36
<b>3</b>	<b>Learning Image and Text Associations for Multimedia Information Fu-</b>	
	<b>sion</b>	<b>37</b>
3.1	Introduction . . . . .	37
3.2	Identifying Image-Text Associations . . . . .	40
3.2.1	A Similarity-Based Model for Discovering Image-Text Associations	40
3.2.2	Text-Based Similarity Measure . . . . .	41
3.2.3	Vague Transformation Based Cross-Media Similarity Measure . .	41
3.2.4	Fusion ART Based Cross-Media Similarity Measure . . . . .	50
3.3	Experimental Results . . . . .	56
3.3.1	Data Set . . . . .	56
3.3.2	Performance Evaluation Method . . . . .	60
3.3.3	Evaluation of Cross-Media Similarity Measure Based on Visual Features Only . . . . .	61
3.3.4	Evaluation of Linear Mixture Similarity Model . . . . .	64
3.3.5	Discussions and Comparisons . . . . .	67
3.4	Summary . . . . .	69
<b>4</b>	<b>Semantic Multimedia Content Representation and Metadata Extrac-</b>	
	<b>tion</b>	<b>71</b>
4.1	Semantic Multimedia Content Representation . . . . .	72
4.1.1	Knowledge Representation and Human Brain . . . . .	72
4.1.2	Semantic Modelling of Multimedia Contents Based on Web Texts	73
4.1.3	Integrating RDF Semantic Description and MPEG-7 standard . .	78
4.1.4	A Four-Layer Data Model . . . . .	83
4.2	Semantic Metadata Extraction . . . . .	84
4.2.1	Extraction of Terms and Relations from Sentence Grammar Trees	87



4.2.2	Term Taxonomy Construction . . . . .	88
4.2.3	RDF Encoding . . . . .	95
4.3	Summary . . . . .	95
<b>5</b>	<b>Mining RDF Semantic Metadata for Multimedia Analysis</b>	<b>98</b>
5.1	Association Rule Mining: An Introduction . . . . .	98
5.2	Association Rule Mining Algorithms . . . . .	101
5.2.1	Generalized Association Mining Algorithms . . . . .	101
5.2.2	Closed Itemsets Mining Algorithms . . . . .	102
5.3	Over-Generalization Reduction . . . . .	103
5.3.1	Over-Generalization Problem . . . . .	103
5.3.2	Over-Generalization Reduction . . . . .	105
5.4	GP-Close Algorithm . . . . .	107
5.4.1	Algorithm Overview . . . . .	108
5.4.2	Pruning Child-Closures and Subtrees . . . . .	112
5.4.3	Support Counting . . . . .	115
5.4.4	Complexity Analysis . . . . .	116
5.5	Experiments . . . . .	117
5.5.1	Synthetic Data Sets . . . . .	117
5.5.2	Real World Semantic Web Data Sets . . . . .	120
5.5.3	Effect of Pruning Strategies . . . . .	125
5.5.4	Effect of Specialization-First Sorting . . . . .	125
5.6	Analysis of Terror Attack Documents: A Case Study . . . . .	126
5.6.1	Semantic Metadata Extraction . . . . .	126
5.6.2	Mining Generalized Associations on RDF Metadata . . . . .	128
5.6.3	Analysis of Patterns . . . . .	131
5.7	Summary . . . . .	132
<b>6</b>	<b>Conclusions and Future Work</b>	<b>133</b>
6.1	Conclusion . . . . .	133
6.1.1	Media Fusion . . . . .	133
6.1.2	Multimedia Content Representation and Metadata Extraction . . . . .	134

6.1.3	Multimedia Information Analysis . . . . .	135
6.2	Future Work . . . . .	136
6.2.1	Media Fusion . . . . .	136
6.2.2	Multimedia Content Representation and Metadata Extraction . .	139
6.2.3	Multimedia Information Analysis . . . . .	140
<b>A</b>	<b>List of Publications</b>	<b>141</b>
A.1	Book Chapters . . . . .	141
A.2	Journals . . . . .	141
A.3	Conferences . . . . .	141
	<b>References</b>	<b>141</b>
<b>B</b>	<b>Evaluation of the Similarity Measure for Term Taxonomy Construction</b>	<b>143</b>

# List of Figures

1.1	Forecast of digital camera penetration in mobile phones. . . . .	3
1.2	Multimedia web information related to the Battle of Normandy. . . . .	4
1.3	Multimedia web information for analyzing fashion trends. . . . .	4
1.4	A process of multimedia web information analysis. . . . .	5
1.5	Semantic based multimedia fusion and analysis framework. . . . .	10
2.1	A general process of multimedia authoring. . . . .	16
2.2	An illustration of the multimedia thesaurus (taken from [Tan00]). . . . .	19
2.3	An illustration of the MediaNet (taken from [BCS01]). . . . .	19
2.4	Semantic Web Architecture. . . . .	21
2.5	A general framework of multimedia data mining. . . . .	28
3.1	An associated text and image pair. . . . .	38
3.2	An illustration of implicit associations between visual and textual features among small number of training samples. . . . .	38
3.3	An illustration of cross-media transformation with information bottleneck. . . . .	43
3.4	Visual space projection. . . . .	47
3.5	Bipartite graph of classified text segments and information categories. . . . .	49
3.6	Fusion ART for learning image-text associations. . . . .	52
3.7	Training samples for fusion ART. . . . .	55
3.8	Information gains of clustering the images based on a varying number of visterms. . . . .	60
3.9	Comparison of cross-media models for discovering image-text associations. . . . .	62
3.10	Performance comparison of cross-media models with respect to different training data sizes. . . . .	63

3.11	A sample set of image-text associations extracted with similarity scores (SC). The correctly identified associated texts are bolded. . . . .	65
3.12	Samples of image annotations using fusion ART. . . . .	67
4.1	Semantic modelling of domain knowledge. . . . .	73
4.2	Two short text documents with similar keywords but of distinct semantic meanings. . . . .	74
4.3	A simplified conceptual graph translated from the first sentence in Figure 4.2. . . . .	75
4.4	Knowledge modelling in the sports domain using RDF. . . . .	76
4.5	XML representation of the RDF metadata in the sports domain. . . . .	77
4.6	A hierarchy of media objects supported by MPEG-7. . . . .	79
4.7	An example of conceptual aspects description using MPEG-7 descriptors, adopted from [Mar04]. . . . .	81
4.8	An extended MPEG-7 ontology. . . . .	82
4.9	Combining domain ontology and MPEG-7 ontology for media object description. . . . .	83
4.10	A four-layer data model in the META system. . . . .	84
4.11	An overview of the procedure for semantic metadata extraction. . . . .	85
4.12	Conversion of text to semantic relations shown in the form of conceptual graph. . . . .	86
4.13	Three cases for merging the new cluster and its most similar cluster. . . . .	93
4.14	Term Taxonomy Construction Using the Sample Terms in Table 4.1. . . . .	94
5.1	A sample market-basket taxonomy. . . . .	99
5.2	An illustration of over-generalization. The itemset $X$ is an over-generalization of the itemset $Y$ . . . . .	105
5.3	The generalization closures of the itemsets $X$ and $Y$ in Figure 5.2. . . . .	105
5.4	A full closure enumeration tree. . . . .	109
5.5	The closure enumeration tree after pruning the non-closed closure $gc(X_1)$ . . . . .	109
5.6	Child-closure pruning. . . . .	113
5.7	Subtree pruning. . . . .	114

5.8	The performance of GP-Close compared with Cumulate and Prutax on the synthetic data sets. . . . .	118
5.9	A sample RDF vocabulary. . . . .	121
5.10	The RDF statement taxonomy inferred from the RDF vocabulary defined in Figure 5.9. . . . .	121
5.11	GP-Close compared with Cumulate and Prutax on the foafPuband ICT-SB data sets. . . . .	123
5.12	Efficiency of using child-closure and subtree based pruning strategy. . . .	124
5.13	Efficiency of the specialization-first sorting compared with the lexical sorting strategy. . . . .	125
5.14	The semantic metadata extraction tool. . . . .	127
5.15	A subset of the extracted term taxonomy. . . . .	128
5.16	Number of Patterns . . . . .	129
5.17	Execution Time . . . . .	130
5.18	Scalability of the GP-Close algorithm. . . . .	130
6.1	Using visterm taxonomy for similarity calculation in visual information space. . . . .	138

# List of Tables

3.1	Cross-media models. . . . .	62
3.2	The average precision scores (%) for image-text association extraction. . . . .	64
3.3	Comparison of the vague transformation and the fusion ART based methods. . . . .	68
4.1	A set of sample terms represented by bags of WordNet senses. . . . .	90
5.1	A sample transaction database. . . . .	104
5.2	Frequent generalized itemsets in sample database ( $minsup=40\%$ ). . . . .	104
5.3	The characteristics of the synthetic data sets. . . . .	118
5.4	The statistics of the RDF vocabularies in the foafPub and ICT-SB data sets. . . . .	122
5.5	The characteristics of the foafPub and ICT-SB data sets. . . . .	122
5.6	Summary of RDF metadata extraction results. . . . .	128
B.1	NSS measure with $WNSD = 3$ . . . . .	144
B.2	Seco et al measure . . . . .	144

# Chapter 1

## Introduction

### 1.1 World Wide Web and Multimedia Information

Following its rapid development in the last decades, *World Wide Web* has become the largest information depository in the world, where a tremendous amount of information covering diverse topics in various languages is published and shared. The Web consists of billions of documents, stored separately in the computers located all over the world. Based on a *hypermedia* paradigm, related Web documents are inter-connected with each other via *hyperlinks*. Web users can explore the Web documents by following these links to find the information fulfilling their needs. However, as the number of the Web document becomes too large, browsing the Web for information is extremely time-consuming. Therefore, web search engines, which employ *information retrieval* techniques, are developed to reduce the information overload for the users. By submitting queries to the search engines, a Web user can obtain a small set of web documents related to his or her information needs in a reasonably short time.

To date, the contents of the web documents are mainly texts. Texts are discrete and symbolic data, which are easy to process and can be converted into other forms of representations, for example bags of words or term feature vectors, based on which large-scale computations can be performed. In addition, the words in the texts are in

CHAPTER 1. INTRODUCTION

---

the form of natural language. Compared with images and videos, it is relatively straightforward to map a word or term to a concept that human can understand. Such favorable characteristics make textual data suitable for the information retrieval task. On one hand, text is an ideal representation to express the users' information needs in queries due to its conformation to conceptual models in the human brains. On the other hand, Web information in texts can be easily matched with the text queries to decide the relevance of the information. Therefore, most of the leading search engines, such as Google, Yahoo, and Baidu, use text-based query interface and text information retrieval techniques for searching the Web.

Though the textual content is still dominant, there is an increasing awareness of the popularity of multimedia information on the World Wide Web. With the prevalence of digital media capturing devices, such as digital recorders, digital cameras and digital videos, people can easily create their own digital visual/audio contents and publish them on the Web. According to the International Data Corporation (IDC, <http://www.idc.com/>), there are 29.8 million digital cameras sold in the USA in 2006, i.e., approximately every ten USA citizens bought a new DC in a year. Moreover, as digital capturing devices keep becoming smaller, many of them can be transplanted into mobile phones, which are considered as a necessity for the people in the information age. According to the prediction of the iSuppli Corporation (<http://www.isuppli.com/>), there will be one billion mobile phones equipped with digital cameras in 2010 (see Figure 1.1). With the combination of the digital capturing devices and mobile phones, multimedia contents could be captured anywhere and anytime.

Whilst the digital capturing devices boost the creation of the multimedia contents, the fast growth of the personal publishing platforms, such as blogs, online albums, and video sharing spaces, is accelerating the delivery of the multimedia information on the Web. For example, today there are ten million users publishing their videos through video-sharing



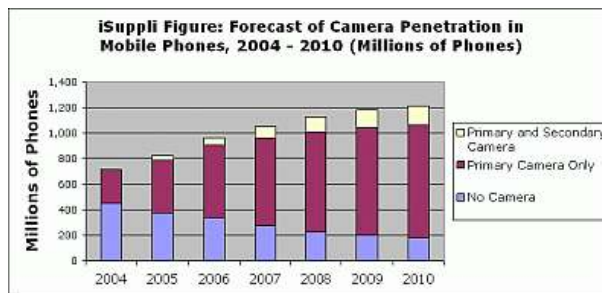


Figure 1.1: Forecast of digital camera penetration in mobile phones.

site YouTube [Inc06]. Google’s online album site, Picasa Web, has ten millions of images uploaded by its users according to an unofficial report <sup>1</sup>. With the boom of 3G mobile networks and mobile web technologies, more multimedia information will be captured and delivered in an online and instant manner. In a foreseeable future, multimedia information will enter into a fast-production and quick-consumption metabolic life cycle.

## 1.2 Scenarios of Web Information Service

As multimedia contents become more active and influential on the Web, collecting and analyzing multimedia online information is becoming an essential task in our daily life. This situation can be illustrated by the following scenarios:

A student, interested in the history of the Second World War, wants to know about the Battle of Normandy. To find the related information, he searches through Google using a query “Battle of Normandy”. Typically, Google returns a ranked list of more than one million web pages which contain plenty of information presented in various media formats (see Figure 1.2). He first reads a text introduction from Wikipedia <sup>2</sup> describing the background and the details of the battle. In the middle of his reading, he may want to see some images which can give him a visual illustration of the battle field or the attack route of the Allied forces. Therefore, he opens some links returned by the Google

<sup>1</sup><http://googlified.com/2007how-many-pictures-are-there-on-picasa-web/>

<sup>2</sup><http://www.wikipedia.org/>

CHAPTER 1. INTRODUCTION

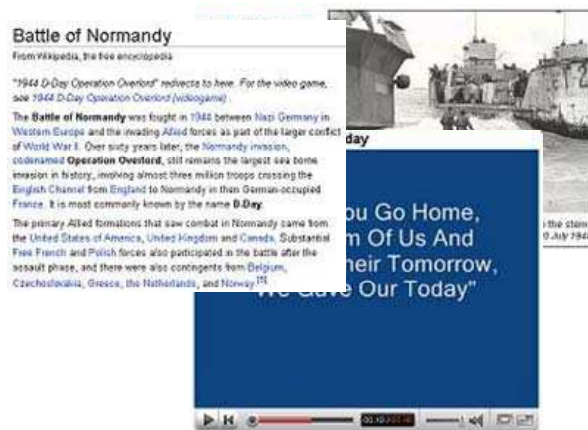


Figure 1.2: Multimedia web information related to the Battle of Normandy.

to look for images related to what he is reading. For obtaining a full impression of the Battle of Normandy, the student needs to collect information from many different web sources and organize pieces of information together. Such work sometimes can be quite tedious.

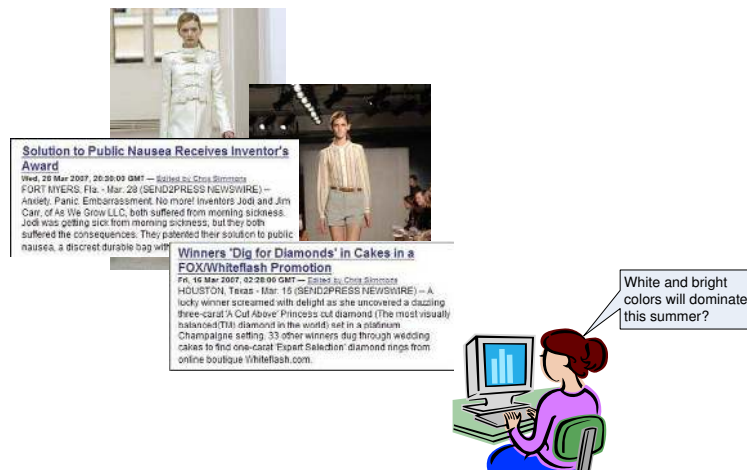


Figure 1.3: Multimedia web information for analyzing fashion trends.

In another scenario, a fashion journalist wants to write a review of the fashion trend for this summer. She browses the fashion-related web sites, online journals, and cloth brand home pages to collect materials, such as photos of fashion shows (images) and

industry reviews (texts). For grasping the trends of the fashionable elements, she aligns the collected visual and textual materials for comparing the designs of different brands and designers.

The above scenarios are examples of cases that we face in our daily life. These scenarios show that acquiring knowledge from multimedia web information typically consists of three stages described as follows (Figure 1.4):

- (i) **Information retrieval** is first performed to obtain a relatively small set of web documents containing multimedia information relevant to users' interested topics.
- (ii) **Information fusion** is then used to assemble discrete, distributed, and heterogeneous multimedia information facets in the retrieved web documents to form an integrated view of concepts or topics.
- (iii) Finally, **information analysis** is conducted to summarize and analyze the collected and collated multimedia information for knowledge distillation, e.g., discover the trends of the fashionable elements based on a design database acquired from the web.

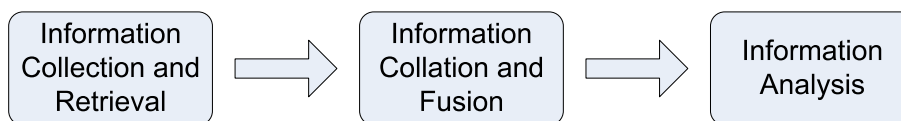


Figure 1.4: A process of multimedia web information analysis.

### 1.3 Multimedia Information Systems

The existing multimedia information systems are designed mainly for three purposes, i.e., multimedia filtering, browsing, and composing. With respect to the different purposes, today's multimedia information systems can be classified into three paradigms,

i.e., multimedia information retrieval, hypermedia, and multimedia authoring. All of the three system paradigms have certain inherent deficiencies which may limit the usages of dynamic and distributed multimedia web information.

**Multimedia information retrieval (MIR)** systems [Sch97] provide the function of multimedia filtering to reduce the information overload on large information collection. When using an MIR system, an information user needs to form a query as a criteria for information filtering. As techniques for feature extraction, data indexing, and similarity measurement are inherently different with regard to various media formats, most MIR systems do not support cross-media information retrieval and can only handle information belonging to a certain media type. In addition, the results returned by the MIR systems are usually discrete information pieces instead of integrated solutions for users' information needs.

**Hypermedia** systems [Nel65] are based on the idea of connecting related pieces of information. Therefore, information in a hypermedia system is relatively well-organized. Users explore the information by browsing or navigating from one document to another through certain kinds of links. However, building, maintaining, and updating the links between multimedia objects or documents are usually conducted manually and thus can be very time-consuming.

A **multimedia authoring** system [CC96] supports multimedia content composing, i.e., generating multimedia presentations for users' interested topics based on elemental multimedia components. The multimedia components used for authoring must be pre-annotated and pre-stored in the system's database. The annotation task is also need to be conducted manually and therefore can be very tedious.

Generally, the existing multimedia information systems are not designed to automatically collate and compute discrete and heterogeneous multimedia information. This deficiency is due mainly to two issues. Firstly, there is a great diversity in the multimedia indexing and representation techniques. Typical multimedia indexing methods may

operate in two different levels: (1) low-level perceptual features automatically extracted from raw media objects and (2) high-level concept based features represented by manually assigned keywords. These methods are diverse in two aspects. From the media type aspect, data in heterogeneous media formats have different low-level feature representations. From the information tier aspect, low-level perceptual features and high-level semantic contents are represented and computed separately.

Secondly, though existing user studies [GG92, MS00] show that semantic-level interactions are vital for multimedia information systems to provide practical applications for information users, utilization of the semantic information is quite limited in the multimedia information systems due to the problem of *semantic gap* [SWS<sup>+</sup>00], i.e., there is a lack of efficient methods for automatically extracting semantic contents based on the perceptual information of the raw media objects. Therefore, most existing systems rely on manually assigned keywords for representing the semantics of the multimedia data. However, manually labelling media objects with keywords has several deficiencies. As the meanings of the keywords from natural languages can be vague, there may be a lack of exact correspondences between the keywords and semantic concepts in the multimedia contents. In particular, synonyms and homonyms will cause inaccuracy in the media object indexing and thus low recall and precision in multimedia retrieval.

To solve the diversity problem in the multimedia content representation, many efforts are performed to represent heterogeneous multimedia contents using semi-structured metadata to improve cross-media operability in multimedia applications such as multimedia retrieval. Such efforts result in *MPEG-7* [Mar04], formally named “*Multimedia Content Description Interface*”, an ISO/IEC standard developed by Moving Picture Experts Group (MPEG) for describing the multimedia data and supporting some degree of interpretation of the information’s meaning, which can be processed by a machine or a computer program [Mar04]. MPEG-7 provides users a rich set of standardized tools to

describe properties and contents of multimedia objects. As MPEG-7 focuses on describing data from the perspective of the multimedia research domain using XML language, it lacks certainty in semantic representations [DMH<sup>+</sup>00]. Therefore, recently research efforts propose to employ Resource Description Framework (RDF) [CLS01, W3C] as used in the Semantic Web to build an RDF version of MPEG-7 standard to model multimedia contents with a formal semantics [Hun01, Hun02].

At present, semantic based multimedia content representation methods, which utilize the MPEG-7 and the Semantic Web techniques, usually depend on manual annotation tools for assigning formally defined semantic concepts to media objects [SDWW01, SBC<sup>+</sup>02]. However, manual semantic annotations can be time-consuming, and therefore not suitable for large-scale applications. In addition, the existing methods usually label multimedia data with individual semantic concepts for the task of multimedia information retrieval. Nevertheless, for multimedia information analysis applications, concept details and contexts are indispensable. For example, an image of a fashion show may be labelled with concepts such as “model” and “clothes”. However, detailed information, such as “the model is male” and “the clothes is black”, can be lost. Without such information, information analysis tasks, such as discovery of the fashion design patterns, are unlikely to be performed.

In the next section, we present a framework, which is designed and implemented during my PhD candidature, for addressing the above limitations to achieve cross-media multimedia information fusion and analysis.

## 1.4 A Framework for Semantic Based Multimedia Web Information Fusion and Analysis

Based on the discussion in the previous section, we believe that a system for cross-media information fusion and analysis should have the following characteristics:

- (i) A unified multimedia representation schema for modelling and processing heterogeneous multimedia information.
- (ii) Methods to automatically extract the intermediate multimedia content representations for supporting large-scale multimedia fusion and analysis.
- (iii) Algorithms for multimedia information analysis based on the intermediate multimedia content representations.

To this end, we develop a framework for multimedia web information fusion and analysis, which utilizes a unified semantic representation schema for modelling heterogeneous multimedia contents based on their semantic meanings. Within this framework, multimedia information fusion is first conducted by associating multimedia contents with web texts which describe the semantic meanings of the media objects. Then, the semantic metadata are automatically extracted and encoded based on the unified semantic representation schema. Finally, multimedia analysis is performed by applying data mining algorithms on the extracted semantic metadata collection. An illustration of the proposed framework is shown in Figure 1.5. The functionalities of the framework components are summarized as follows:

- A **unified semantic representation schema** that combines *MPEG-7* multimedia description, RDF language specification, and conceptual graph based knowledge representation techniques for multimedia content modelling. Different from the existing semantic based multimedia systems that label multimedia object with concepts, our method describes semantic multimedia contents using *conceptual graphs* [Sow84, Sow99], a type of semantic networks [Sow91] wherein concepts are interconnected with semantic relations to form conceptual contexts expressing the detailed data semantics. The conceptual graphs are encoded formally in

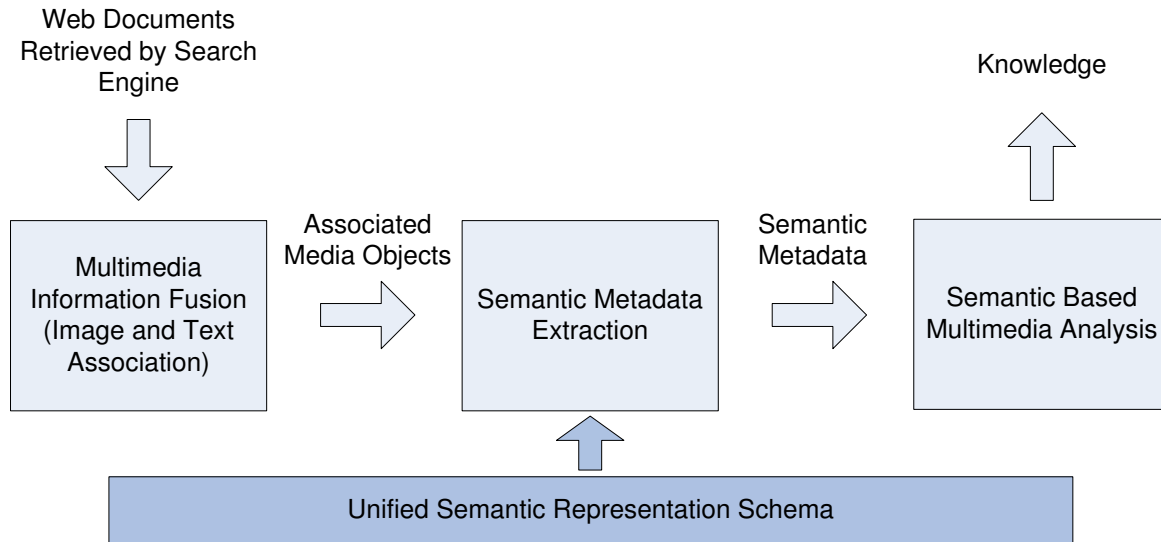


Figure 1.5: Semantic based multimedia fusion and analysis framework.

the RDF language as *semantic metadata*. In addition, the descriptors defined as in the MPEG-7 standard are adopted for representing the information related to the media aspects of the media objects, such as author, creation, and media format information.

- A **multimedia information fusion** module designed to associate media objects with related surrounding texts, which reflect their semantic meanings. This model serves as a basis for extracting semantic metadata to describe the various multimedia contents in a unified form of intermediate representation. Specifically, this thesis focuses on associating web images (including video frames) with surrounding texts. Two methods are developed for discovering the underlying image-text associations in web pages. The first method adopts a vague-transformation technique as used in multilingual retrieval models and domain-specific concepts to indirectly measure the relevance between the images and texts. The second method employs a self-organizing neural network to directly learn the association patterns between the perceptual features of images and the textual features.



- A **semantic metadata extraction module** that automatically generates semantic metadata for the multimedia web contents, including images and videos, based on the associated web texts. A myriad of natural language processing (NLP) techniques are used to analyze the part of speech and grammar structures of the web texts, from which semantic concepts and relations are extracted automatically to form conceptual graph representations of the multimedia semantic contents. The extracted semantic concepts are formally modelled using word senses provided by the *WordNet* lexical database [LC98]. Based on the word sense representations of the semantic concepts, a concept similarity measure is defined and used in a light-weight incremental hierarchical clustering method to build a concept hierarchy. The concept hierarchy is encoded using the RDF Schema [W3C] and serves as a shared vocabulary for constructing RDF semantic metadata.
- For **multimedia information analysis**, an algorithm, called *GP-Close*, is proposed for mining generalized association patterns from RDF semantic metadata consisting of semantic concepts and their relations. Concept hierarchies are also employed for discovering knowledge at various semantic levels. In contrast to the existing generalized association pattern mining algorithms, which tend to generate many *over-generalized patterns* especially from RDF metadata, GP-Close adopts a systematic method for full over-generalization reduction based on the notion of *generalization closure*.

## 1.5 Summary of Contributions

In this thesis, we present a unified framework for addressing several key issues in cross-media web information fusion and analysis. In particular, we present new techniques and algorithms for image-text information fusion, multimedia content modelling, seman-

tic metadata extraction, and RDF metadata mining. The main contributions of this framework are summarized as follows:

- (i) **Media Fusion:** We develop two methods, using a statistical vague transformation technique and a self-organizing neural network method respectively, for visual and textual information fusion. Specifically, the methods discover the underlying associations between images and texts in web pages by linking visual contents with semantic information conveyed by the web texts. The work on learning image-text associations has been reported in [JT06a, JT07].
- (ii) **Semantic Multimedia Content Representation and Metadata Extraction:** We adopt a multimedia content representation schema that combines the MPEG-7, Resource Description Framework (RDF), and conceptual graph techniques for describing and modelling heterogeneous multimedia information in an integrated framework using machine-processable and human-understandable semantic metadata. More significantly, an automatic metadata extraction method is developed for analyzing web texts for semantic concepts and relations based on a set of natural language processing (NLP) techniques. This part of work can be found in [JTW07a].
- (iii) **Media Analysis:** An algorithm, call GP-Close, is designed for discovering generalized association patterns of concepts and relations based on semantic metadata of multimedia objects for information analysis. Extensive experiments have been conducted to show that the GP-Close algorithm can significantly reduce the pattern redundancy and outperform the existing generalized pattern mining algorithms in terms of time efficiency. The various versions of GP-Close algorithm and its applications to web information analysis, including a case study on terrorist information mining, have been published in [JT06c, JT06b, JTW07b, JTW07a].

## 1.6 Thesis Structure

The outline of the thesis is given as follows. We first review the existing techniques and systems that are related to multimedia fusion and analysis in Chapter 2. From Chapter 3 to Chapter 5, the techniques developed and implemented in the META system are presented in detail.

Chapter 3 focuses on the work on discovering image-text association for multimedia information fusion, which is developed to facilitate the automatic semantic metadata extraction. To address this problem, two algorithms based on a similarity based multi-lingual retrieval model and a self-organizing neural network are proposed and evaluated.

Chapter 4 presents the semantic based multimedia content representation techniques and introduces the automatic metadata extraction method which utilizes a myriad of natural language processing techniques to extract concepts and relations from the web texts.

Chapter 5 is dedicated to metadata analysis. In particular, an algorithm, called GP-Close, for mining semantic relation patterns from RDF-like metadata is described. Experimental evaluations and a case study of applying the proposed method to a terrorist domain data collection are presented.

Finally, Chapter 6 summarizes the work done and discusses how this work could be taken further.

# Chapter 2

## Related Work

In this chapter, we review the state of the art in research fields related to the topic of multimedia information fusion and analysis. In particular, we shall focus on four research areas, i.e., multimedia information fusion, multimedia content modelling in the Semantic Web, automatic multimedia annotation, and multimedia analysis.

### 2.1 Multimedia Information Fusion

Research in multimedia information fusion takes two primary directions. Multimedia authoring is concerned with the integration of multimedia data for generating multimedia presentation. Another direction is fusion of multimodal low-level information for indexing and retrieval. Both fusion approaches are not designed for the purpose of multimedia information analysis aiming at discovering compact and useful knowledge from large multimedia data collection.

#### 2.1.1 Multimedia Authoring

Multimedia authoring research aims at developing tools for helping users to generate multimedia presentation. A general process of multimedia authoring typically contains four stages as shown in Figure 2.1:

CHAPTER 2. RELATED WORK

---

- (i) Multimedia object collection: Multimedia objects related to a particular topic are selected from a closed multimedia repository or an open multimedia data source, such as the World Wide Web. The selected multimedia objects are treated as the raw material for building the multimedia/hypermedia presentation of the user's interested topic.
- (ii) Semantic structure generation: The semantic relations of the selected multimedia objects are determined. For example, given a text paragraph introducing "the biography of Vinci" and a picture depicting "the portrait of Leonardo da Vinci", their relation can be identified as "the image can *illustrate* the text paragraph". These semantic relations form the semantic structure of the selected multimedia objects.
- (iii) Presentation (spatial-temporal) structure generation: Based on the semantic structure of the selected multimedia objects, a set of spatial-temporal relations between these multimedia objects are generated, which compose of a multimedia presentation structure. Generation of spatial-temporal relations usually depends on pre-defined presentation templates which define a set of rules to convert the semantic relations to the spatial-temporal relations. For example, a rule may be defined as "the textual introduction of an artist should be displayed on the right of a portrait image of the artist".
- (iv) Multimedia presentation generation: According to the spatial-temporal structure of the multimedia objects, a multimedia presentation will be rendered and generated.

Since the last decade, massive efforts have been conducted in this area and a series of multimedia authoring tools [vRJMB93, BH95, RHB99] and standards [Har98, Hos98] have been developed. However, most of the existing tools and standards, including the

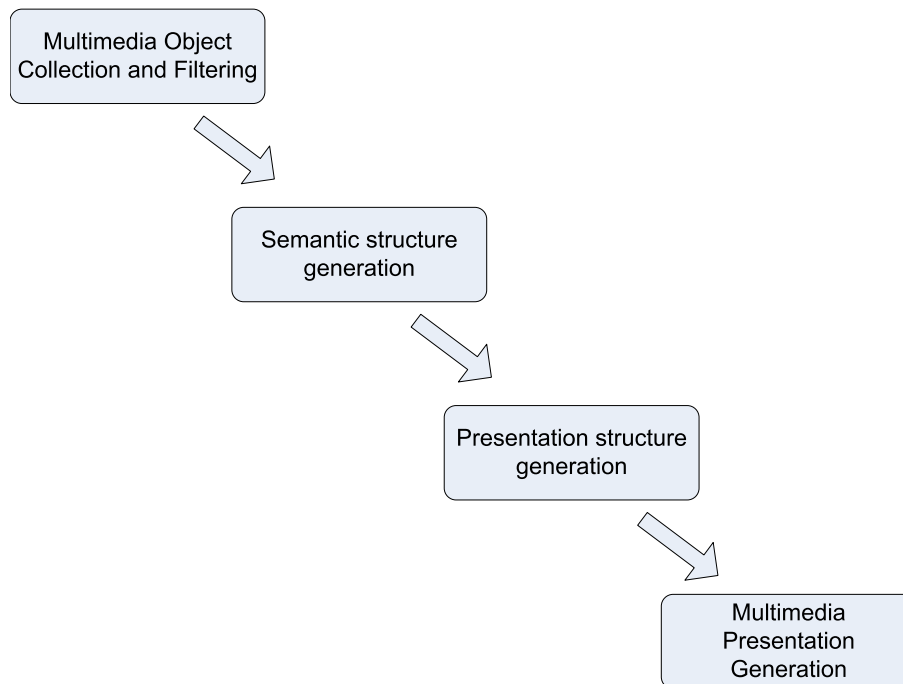


Figure 2.1: A general process of multimedia authoring.

powerful GRiNS and the well-know SMIL (Synchronized Multimedia Integration Language) standard proposed by World Wide Web Consortium (W3C) [Hos98], focused on developing a set of constructs and interfaces for users to build the presentation structures of multimedia objects, there is a lack of automation in the multimedia authoring process which limited the applications of those developed authoring tools and standards [RHB99]. In particular, most of the authoring stages, such as selecting multimedia objects, determining their semantic relations, and building the presentation structures, need to be carried out manually by human authors and based on their personal experiences.

Seeing the deficiencies of the existing work, many research efforts have been recently conducted to develop techniques to facilitate the multimedia authoring automation process. In particular, semantic based techniques were considered important for tackling the problems that the most of the existing multimedia authoring systems encountered. The use of the semantics has two advantages for the multimedia authoring automation. Firstly, with the identified semantics, the meanings of multimedia objects are much

CHAPTER 2. RELATED WORK

---

clearer and therefore can be more easily, accurately, and precisely retrieved. Specifically, the techniques developed in the Semantic Web research are also helpful for the multimedia authoring system designers to automate the multimedia object collection phrase. Secondly, the semantics can be leveraged to determine the relationship between the multimedia objects. For example, an image of Microsoft's Seattle Headquarter can be seen as a visual illustration for a text description of Microsoft. Such semantic relations can be identified based on a set of rules that are defined by human experts or learned using machine learning methods. Currently, most of the state-of-the-art multimedia authoring systems employ ontology (i.e., a specification of a conceptualization for defining concepts and relations for knowledge sharing [Gru93, Gru95]) and semantic modelling techniques for automatically generating multimedia presentations.

Many state-of-the-art authoring systems used domain templates to generate meaningful presentations of semantic information. DArtbio [BKRR01] used presentation plans (a kind of genre-specific templates), expressed in terms of Rhetorical Structure Theory (W.C. Mann, et al. 1989), to create presentations of artist biography based on a domain model of art history.

In [SHSG02], an ontology-based multimedia e-document construction and delivery method was introduced. Artequakt [KAH<sup>+</sup>02, KAH<sup>+</sup>03] uses human authored biography templates to create textual artist biography. The templates contain query patterns (which determine the types of information needed in the biography) for retrieving pieces of information from a knowledge base. The information of this system is collected by using NLP techniques to extract information from text. Subsequently, the extracted information is used to enrich the domain ontology and the domain knowledge base.

In [GBvOH03], the authors introduced an approach to create multimedia representations of semantic knowledge in Semantic Web. This approach uses genres, each of which was associated with a set of rules (like templates), to order, group, and prioritize multimedia information.

Falkovych et al. presented the SampLe system which supports a human-controlled five-step process of production of multimedia presentations [FNvOR03, FNvOR04, FN06]. Ontology is also used in this work for generating metadata and material selection.

Most of the existing multimedia authoring systems, either manual or automatic based, depend largely on manually created annotations for multimedia material collection and integration due to the semantic gap problem. Therefore, such systems are less scalable and usually concentrate on specific domains.

### **2.1.2 Multimodal Information Fusion for Multimedia Indexing and Retrieval**

The purpose of multimodal information fusion is to associate low-level media information from multiple channels in an integrated model for disambiguating the meaning of the information in individual modals. An initial work in this area is presented in a multimedia retrieval system, MAVIS 2 [Tan00]. In MAVIS 2, a Multimedia Thesaurus is adopted to integrate multimodal information. The proposed Multimedia Thesaurus is a semantic network which is composed of interconnected abstract concepts, each of which can be illustrated by a set of terms and a set of multimedia objects. An illustration of the “car” concept in the Multimedia Thesaurus is shown in Figure 2.2. In MAVIS 2, the Multimedia Thesaurus is used for query expansion to improve the multimedia retrieval results. For example, upon receiving a query, such as the term “automobile”, MAVIS 2 first finds a terms matched with “automobile” so as to identify the “car” concept and then all terms and media objects related to the “car” concept can be used to expand the query to find more relevant multimedia information.

Another work in this area is known as MediaNet presented in IMKA (Intelligent Multimedia Knowledge Application) system [BSC00, BCS01]. MediaNet is also a semantic network. However, it is different from the Multimedia Thesaurus [Tan00] in combining perceptual information and semantic concepts and representing semantic and perceptual



## CHAPTER 2. RELATED WORK

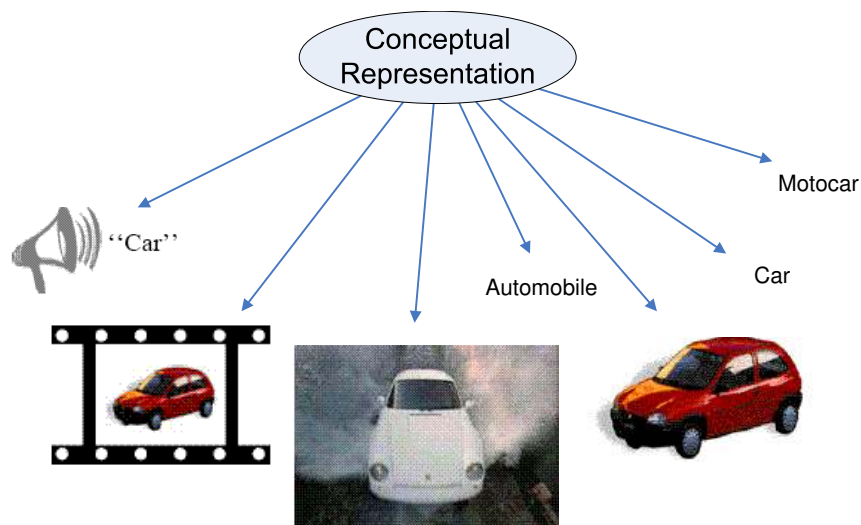


Figure 2.2: An illustration of the multimedia thesaurus (taken from [Tan00]).

relationships between concepts. MediaNet encodes a rich set of relationships between the network concepts, which are derived from the lexical database WordNet [Mil95], and utilizes perceptual feature similarities to weigh the strength of the relationships. In addition, concepts and relationships in MediaNet are not only represented by positive examples but also illustrated by negative examples. An illustration of the MediaNet is shown in Figure 2.3.

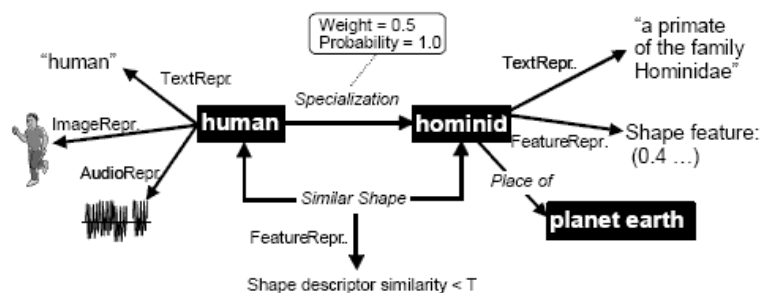


Figure 2.3: An illustration of the MediaNet (taken from [BCS01]).

In the Informedia project <sup>1</sup>, multimodal information fusion is widely used for video indexing [DW03, DPW03, DH04]. For disambiguating the visual contents in video frames,

<sup>1</sup><http://www.informedia.cs.cmu.edu/>

textual information, such as speech transcripts, is extracted and related to the video frames according to the video time-lines. The associations between textual and visual information are learned using various statistical models, including machine translation [DW03] and statistical correlation matrix [DH04]. The learnt models can be used for labelling new video clips for supporting automatic indexing and semantic based retrieval.

The existing efforts in multimodal information fusion mainly focus on building representative prototypes of low-level perceptual features for a predefined set of high-level semantic concepts and relations to facilitate multimedia retrieval. It is different from our task of learning the association between images and free web texts wherein such a predefined set of concepts and relations does not exist.

## 2.2 Semantic Web and Multimedia Modelling

### 2.2.1 Semantic Web

Observing the current situation of the World Wide Web, whose content is mainly texts in natural languages which is not understandable for machines, Tim Berners-Lee brings forward his vision of Semantic Web [BL01] aiming to describe the web resources using human-understandable and machine-processible representation languages. It is an effort to explicitly describe the meaning of the web documents in the form of *semantics*, an alternative of the machine-processible metadata in the context of the Semantic Web, so that computer programs and software agents can find, share, and integrate web information more easily.

Figure 2.4 shows an architecture of the Semantic Web which includes the following components:

- **Universal Resource Identifier (URI)** is a global naming schema for linking a web resource with a unique id.

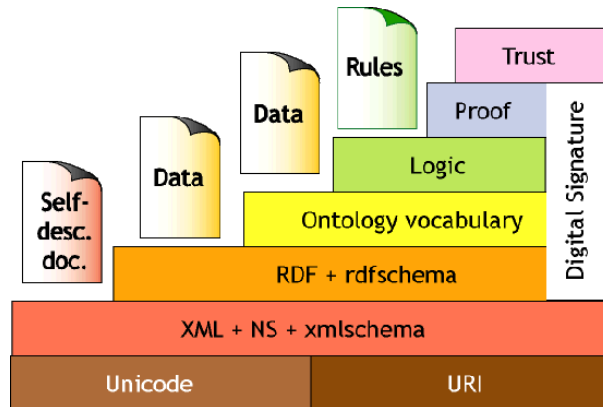


Figure 2.4: Semantic Web Architecture.

- **XML (Extensible Markup Language)** and **XML Schema** offer a standard language syntax to allow users to define their own data structures for information exchanging on the Web.
- **Resource Description Framework (RDF)** and **RDF Schema (RDFS)** provide a data model to describe the semantic meaning of web resources in terms of concepts and relations. Whilst RDF is mainly a language specification concerning syntax and rules of semantic representations, RDF Schema addresses the problem of vocabulary definition which provides a set of concepts and relations for describing web resources.
- An **ontology vocabulary** is a conceptualization of knowledge for a particular domain which can be formulated using RDF Schema. With a defined ontology, computer program and software agent can share and interchange information using a common conceptual model.
- Based on the semantics built upon RDF and ontologies, **logic**, **proof**, and **trust** mechanisms are needed to provide inferencing rules, reasoning capability, and accessibility and credibility controls.

Among the various layers of the Semantic Web, RDF and ontology are the foundations of modelling data semantics. The basic element of RDF is *RDF statements* (or *RDF relations*), each consisting of a subject, a predicate, and an object. For simplicity, we use a triplet of the form  $\langle \text{subject}, \text{predicate}, \text{object} \rangle$  to express an RDF statement. At the semantic level, an RDF statement could be interpreted as “the subject has an attribute whose value is given by the object” or “the subject has a relation with the object”. Whilst RDF statements form a model for describing knowledge, ontologies, which can be defined by RDF Schema or its extensions (e.g. OWL and DAML), provide a set of concepts and relation types that can be used to construct RDF statements.

### 2.2.2 Multimedia Modelling

As there is an increasing amount of multimedia contents available on the Web, modelling multimedia contents using machine-processible and human-understandable metadata has been investigated as an approach for bridging the semantic gap and improving the multimedia retrieval performance. MPEG-7, formally known as “Multimedia Content Description Interface”, developed by Moving Pictures Expert Group (MPEG) is an initial standard for this purpose. However, MPEG-7 is designed based on XML, which may cause ambiguities in semantic modelling, reusing, and interoperability as data with the same semantics can be modelled in different forms of XML representations. With the rising of the Semantic Web, many research efforts have been conducted to incorporate the semantic representation techniques in the Semantic Web into the existing MPEG-7 standard for formally modelling multimedia contents.

Jane Hunter first proposed to build an MPEG-7 ontology for describing multimedia content based on RDF schema [Hun01]. Subsequently, a further practice of combining MPEG-7 and CIDOC-CRM (CIDOC object-oriented Conceptual Reference Model) models into a single ontology for describing and managing multimedia in museums was

presented [Hun02]. The result schema could provide a richer set of vocabularies for annotating museum multimedia data. However, the detailed description of those multimedia data (e.g. what story does this film tell) was still described in a plain text form. In [HSL<sup>+</sup>06], another effort that combines RDF and CIDOC-CRM ontology is presented. Similar to Jane Hunter's work, the detailed description of the media contents are represented in free text form. Note that those multimedia modelling approaches are mainly designed for the task of multimedia retrieval. Therefore, representing main topics and concepts of multimedia objects in semantic formulation of RDF and leaving detailed contents in free texts can be suitable for the retrieval purpose. However, for the purpose of multimedia information fusion and analysis, such a representation may not be useful due to incomprehensibility of the free texts for computer programs.

## 2.3 Automatic Multimedia Annotation Extraction

Most mature tools for creating semantic multimedia annotation depend on an ontology-guided human labelling basis. A thorough discussion of the problems and issues of ontology-based semantic annotation of images is presented by Schreiber et al [SDWW01, SBC<sup>+</sup>02]. However, manual semantic annotation is considered too labour-intensive to support scalable applications. Automatic semantic multimedia annotation therefore becomes essential for realizing the blueprint of semantic based multimedia computing. The existing automatic multimedia annotation techniques typically fall into two classes, namely (1) machine learning or statistical modelling based annotation and (2) multimedia ontology based annotation.

### 2.3.1 Automatic Annotation Based on Machine Learning and Statistic Modelling

Recently, many research efforts have been conducted on image semantic annotation using statistical modelling approaches [CMC05]. Barnard and Forsyth [BF01] proposed

CHAPTER 2. RELATED WORK

---

a method to encode the correlations of words and visual features using a co-occurrence model. The learned model can be used to predict words for images based on the observed visual features. Another work [JLM03] by Jeon et al. presented a cross-media relevance model for annotating images by estimating the conditional probability of observing a term  $w$  given the observed visual content of an image. In [DBdFF02], Duygulu et al. showed that image indexing could be considered as a machine translation problem to find the correspondence between keywords and image regions. Experiments conducted using IBM translation models illustrated promising results. In [LW03], Li and Wang presented an approach that trained hundreds of two-dimensional multiresolution hidden Markov models (2D MHMMs), each of which encoded the statistics of visual features in the images related to a specific concept category. For indexing an image, the 2D MHMM that generates the image with the highest probability will be chosen.

Most of the existing machine learning and statistic based automatic multimedia annotation methods have the deficiency that they usually assign a fixed number of keywords or concepts to a media object. Therefore, there will inevitably be some media objects assigned with unnecessary annotations and some others assigned with insufficient annotations. More recently, some machine learning algorithms are proposed for multi-label learning which aims to annotate data samples, such as images, with a various number of labels. In [ZZ06], Zhou and Zhang proposed two solutions to convert multi-instance multi-label learning into multiple traditional supervised learning tasks. However, when the label or keyword set is very large, the efficiency and scalability of applying such methods need to be further explored. In [ZZ07], another k-Nearest Neighbor (kNN) based method is proposed for assigning multiple labels to a data sample based on the statistic information of the labels of its  $k$  nearest neighbors. This method showed very promising results on several real-world multi-label learning applications. Unfortunately, how the  $k$  value will influence the number of labels assigned to a data sample is not clearly stated in [ZZ07].

Another deficiency of the most existing methods lies in that model evolution is not well supported. Specifically, after the machine learning and statistical models are trained, they are difficult to update. In addition, the above methods usually treat the semantic concepts in media objects as separate individuals without considering relationships between the concepts for multimedia content representation and annotation.

### 2.3.2 Automatic Annotation Based on Multimedia Ontology

Different from the machine learning and statistical modelling approach, some other efforts consider building multimedia ontologies for multimedia annotation. An advantage of the ontology based automatic annotation is that the multimedia ontology is easy to expand and update.

Bloehdorn et al. presented an infrastructure of the aceMedia project that combines DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering) core ontology and a visual descriptor ontology, based on RDFS MPEG-7 ontology, to support automatic content annotation [BPS<sup>+</sup>04]. The combined ontology contains both domain knowledge and low-level audio-visual features which forms prototype instances for describing domain concepts. Multimedia annotation can thus be conducted by matching the multimedia content with the existing visual-audio prototypes. Based on the infrastructure presented in [BPS<sup>+</sup>04], a multimedia annotation tool, M-OntoMat Annotizer [PBS<sup>+</sup>06], has been developed for initializing and enriching the multimedia ontologies and performing annotation tasks.

Another ontology based annotation tool, named Multimedia Ontology Manager (MOM) was presented in [BBT06]. MOM creates multimedia ontology by assigning video sequences as instances of concepts in a linguistic ontology, and adopts an unsupervised Fuzzy C-Means algorithm to group the instance video clips into clusters. Centers of the clusters in low-level feature space are computed as visual representations of the semantic

concepts. The created ontology is encoded in Web Ontology Language (OWL), which can be seen as an extension of RDF Schema by adding more logic constructs for supporting reasoning capability. A new video clip can then be automatically annotated by assigning it into a similar video cluster in the ontology. After the new video clip is assigned into an existing cluster, the visual concept representation corresponding to that cluster will be updated. In this way, the multimedia ontology can evolve over time.

A defect of the current multimedia ontology based annotation lies in that though automatic multimedia annotation can be achieved with the use of the multimedia ontology, building such a multimedia ontology still involves a lot of human assistances. In addition, the above method still tends to link multimedia objects with individual concepts without modelling concept relations. In fact, detecting semantic relations based on low-level media features is much more complicated than concept detection as much more background knowledge and context information needs to be considered.

## 2.4 Multimedia Analysis

For analyzing multimedia information, various techniques are developed in the field of multimedia data mining (MDM). Data mining (DM) is “the exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules” [BL97]. Traditionally, data mining has been applied to well-structured data whose formats are formally defined using data feature fields, such as those data in large relational databases.

Multimedia data mining can be seen as an extended area of DM for analyzing media data of singlemodal or multimodal for finding useful patterns [PK06]. The discovered patterns can be used for either improving the traditional multimedia applications, such as multimedia indexing and retrieval, or supporting real life decision making, such as



improving the traffic control systems based on the traffic jam patterns extracted from surveillance videos.

MDM and DM have some significant differences. First of all, multimedia data are unstructured by nature. No well-defined data feature is provided to indicate their meanings or content information. Preprocessing is necessary to arrive at feature representations of media objects. And usually, interpretation of the feature representations is the most difficult task. For example, an image depicting fire and another one containing a red flag may be both dominated by red colors. Computer programs can hardly discriminate their differences. Such ambiguities cause the semantic gap between the low-level perceptual features and the high-level semantic concepts. Secondly, multimedia data are inherently heterogeneous. Even images in the same format may have different dimensions of the pixel matrices and audio files related to the same music may have different sampling frequencies. Those heterogeneities are resulted from the differences in their creation processes and media capture devices. All these special characteristics impose new issues and challenges on multimedia data mining.

### **2.4.1 General Framework of Multimedia Data Mining**

Typically, multimedia data mining can fit into the general data mining framework which consists of four major stages: (1) data collection, (2) data preprocessing, (3) pattern discovery, and (4) pattern visualization and interpretation. An illustration of this MDM framework is shown in Figure 2.5.

The data collection stage involves gathering and filtering, from the internet or a multimedia database, a collection of multimedia objects that will be used for data mining. As multimedia objects are unstructured, data filtering sometimes takes more efforts than choosing proper data feature fields of relational databases in traditional data mining. In particular when collecting data from the internet, filtering out noisy data, such as advertisement and duplicated images, is a challenging task.

CHAPTER 2. RELATED WORK

---

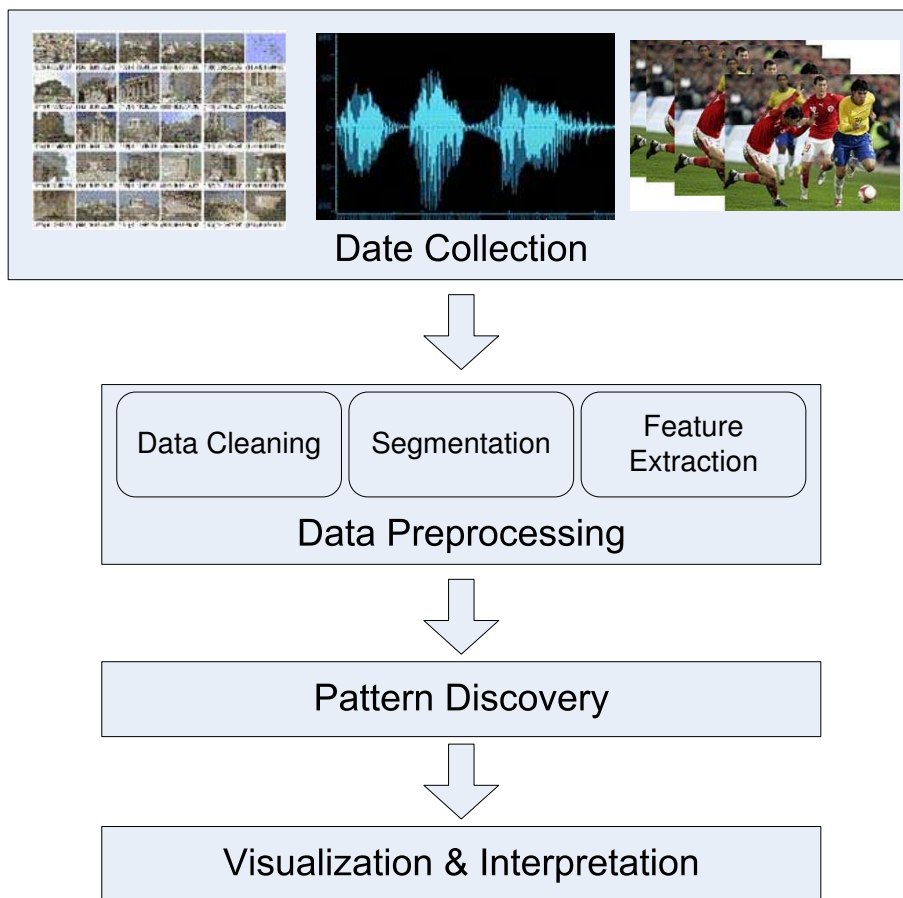


Figure 2.5: A general framework of multimedia data mining.

CHAPTER 2. RELATED WORK

---

The data preprocessing stage may include three subtasks, namely data cleaning, segmentation, and feature extraction. Data cleaning is to clean the multimedia data collected from different sources and with varying data qualities. For example, image processing algorithms can be applied to an image collection to reduce the noise or cut down the unwanted edges in original images so that they can be easily processed by the subsequent segmentation and feature extraction stages. Note that a multimedia object is usually a combination of multiple semantic parts. For example, a photo may include a human in the front and a church in the background or a news video may contain several news events. As the data related to different concepts are mixed up, the media objects are usually difficult to interpret as a whole. Therefore, spatiotemporal segmentation is needed to break the multimedia objects into parts whose characteristics can be captured more easily and meaningfully. Then feature extraction schema and tools need to be chosen to extract features to reflect the contents of the media objects. Multimedia data features are inherently different with respect to various data types. For example, images and video frames can be represented by visual features, such as color, edges, shape, and texture. Audio can be represented by features that lie in either temporal or spectral domains, e.g. pause rate, short-time energy, normalized harmonicity, fundamental frequency, frequency spectrum, bandwidth, spectral centroid, spectral roll-off frequency and band energy ratio [PK06].

The core of the MDM process is the pattern discovery stage, where the patterns and trends buried in the large multimedia data collection are unveiled. Technologies from various disciplines can be applied for this purpose, such as statistical analysis, machine learning, artificial intelligence, and pattern recognition. Therefore, the discovered patterns and trends are also encoded in different forms and models, such as statistical rules, feature vectors, or probability density formulas.

At the visualization and interpretation stage, the discovered patterns or trends will be presented to allow the users to evaluate whether the patterns are useful for solving

real world problems. If not, the user needs to go back to adjust the data preprocessing or pattern discovery methods to extract more useful patterns. Therefore, like traditional data mining, multimedia data mining is an iterative process.

## 2.4.2 Multimedia Data Mining Techniques

To date, most multimedia data mining techniques have concentrated on overcoming the semantic gap between the low-level perceptual information and high-level semantic concepts. Specifically, many existing techniques focus on discovering concepts from multimedia data and detecting events/structures of multimedia objects.

Moreover, visual data mining attracts much more research efforts than audio data mining. The reason is that the semantics of audio data, with the exception of speech data, can be quite ambiguous. For example, it is hard to describe what emotion a music really expresses. In addition, different people may have different explanation about the same music. Due to the domination of the visual data mining techniques in MDM, our introductions in the following sections focus on techniques for mining images and videos.

The efficiency of a visual data mining technique usually lies in two aspects, i.e., *feature extraction* and *mining methodology*. A feature extraction method determines whether the content of the media objects would be represented meaningfully. It serves as the basis to build the media information space from which the knowledge or patterns will be discovered. Given a meaningful representation of the media information, a data mining method determines the types and qualities of the discovered patterns. Therefore, the descriptions of the multimedia data mining techniques below focus on these two aspects.

### 2.4.2.1 Multimedia classification

Classification is a data mining technique used to predict group or class membership for data instances [Han05]. For example, classification can be used to predict whether a person is infected by dengue disease. Given a set of groups or classes, traditional classification

techniques usually train a prediction model, called classifier, for each group/class. In the multimedia domain, classification has been used for annotation and indexing purposes, i.e., predicting whether certain semantic concepts appear in media objects. Therefore, the multimedia classification and the multimedia annotation are closely related. In fact, when the number of the keywords used for annotation is small, the multimedia annotation can be seen as a multimedia classification task, for which the traditional classification techniques can be applied. However, when the number of the annotation keywords is very large, training a classifier for each keyword is not efficient and practical, and different techniques and strategies need to be explored for the multimedia annotation. Therefore, in this thesis, we discriminate the multimedia classification from the multimedia annotation and refer it as a technique for classify the multimedia data instances into a small number of groups or classes.

An early work in the multimedia classification area is to classify the indoor-outdoor scenarios of the video frames [YW95]. In this work, a video frame or image is modelled as sequences of image segments, each of which is represented by a set of color histograms. A group of one-dimension Hidden Markov Models (HMM) are first trained to capture the patterns of image segment sequences and then used to predict the indoor-outdoor categories of new images. Another effort for classifying indoor-outdoor images were presented in [PLSH<sup>+</sup>99], wherein a classification approach that combines image and text features was presented. A classical TF\*IDF (term frequency – inverse document frequency [SM86]) weighting method is used for forming term vectors for representing texts. For image portion, each image is divided into a  $8 \times 8$  grid of image segments, each of which is characterized by an HSV color histogram. Image segments are then grouped into clusters, whose centroids are considered representing different image objects. As these image objects can form a visual vocabulary for describing images like using terms to depict texts, they are also known as *visterms*. For generating visterm vectors to

represent the images, an OF\*IIF (object frequency multiply inverse image frequency) based method, similar to the TF\*IDF method, is proposed for weighting the importance of a visterm for an image. Classification of a new image is done by assigning the image into the category whose training images have largest average similarity with it.

Different from the above methods focusing on categorizing images into very abstract classes (indoor/outdoor), many recent efforts aim to classify and annotate images with relevant concrete concepts. In [SCS01], a decision tree is used to learn the classification rules that associate color features, including global color histograms and local dominant colors, with semantic concepts such as sunset, marine, arid images, and nocturne. In [BB97], a learning vector quantization (LVQ) based neural network is used to classify images into outdoor-domain concepts, such as sky, road, and forest. Image features are extracted via Haar wavelet transformation. Another approach using vector quantization for image classification was presented in [MS03]. In this method, images are divided into a number of image blocks. Visual information of the image blocks is represented using HSV colors. For each image category, a concept-specific codebook is extracted based on training images. Each codebook contains a set of codewords, i.e., representative color feature vectors for the concept. New image classification is performed based on finding most similar codewords for its image blocks. The new image will be assigned to the category whose codebook provides the most number of the similar codewords. This work further shows that color features are suitable for identifying *mass noun entities* [BB97], such as grass, forest, and sky, which do not have definite boundary. For the *count noun entities*, such as cars, with concrete boundaries, shape-based features are needed for image categorization.

#### 2.4.2.2 Mining multimedia association rules

Association rule mining (ARM) is originally used for discovering association patterns in transaction databases. An association rule is an implication of the form  $X \Rightarrow Y$ , where

CHAPTER 2. RELATED WORK

---

$X, Y \subseteq \mathcal{I}$  (called itemsets or patterns) and  $X \cap Y = \emptyset$ . In the domain of market-basket analysis, such an association rule indicates that the customers who buy the set of items  $X$  are also likely to buy the set of items  $Y$ . Mining association rule from multimedia data is usually a straightforward extensions of ARM in transaction databases.

Zaiane et al. presented a multimedia mining system, MultiMediaMiner [ZHL<sup>+</sup>98, ZHLH98]. It collects web images through an image crawling and processing engine, C-BIRD. For each downloaded image, C-BIRD extracts a set of descriptors and features, such as file name, URL, surrounding keywords, image size, and image visual features. Visual features used in C-BIRD include RGB color histogram, most frequent color (MFC) vector, and most frequent orientation (MFO) vector. All extracted web image descriptors are stored in a multimedia data cube so that association rules can be extracted at multi-levels of various dimensions. A sample association rule extracted by the proposed algorithm in [ZHLH98] may like “*if image is related to sky, it is blue with a possibility of 73%*”. The same authors of MultiMediaMiner later proposed an Apriori-based algorithm, MaxOccur, for mining associations of recurrent items with spatial relationships from large image collections [ZHZ00]. A progressive refinement strategy is used to first quickly discover approximate association patterns at a coarse resolution level and then eliminate false positives by verifying them at a finer resolution level. A sample rule extracted by MaxOccur may like “*if an object  $X$  is horizontally next to an object  $Y$ , the object  $Y$  is vertically next to an object  $Z$  with a possibility of 10%*”.

Besides the work that focus on discovering association patterns among the features or objects within images, many efforts were also made to extract the association between low-level visual features and high-level semantic concepts for image indexing and annotation. Ding et al. [DDP02] presented a pixel based approach to deduce associations between pixels’ spectral features and semantic concepts. In [DDP02], each pixel is treated as a transaction, whilst the set ranges of the pixel’s spectral bands and auxiliary concept

labels (e.g. crop fields) are considered as items. Then pixel-level association rules of the form “Band 1 in the range [a, b] and Band 2 in the range [c, d] are likely to imply crop yield  $E$ ”. However, Tesic et al. [TNM03] pointed out that using individual pixel as transaction may cause the loss of the context information of surrounding locations which are usually very useful to determine the image semantics. This motivated them to use image and rectangular image regions as transactions and items respectively. Image regions are first represented using Gabor texture features and then clustered using self-organizing map (SOM) [Koh01] and learning vector quantization (LVQ) to form a visual thesaurus. The thesaurus is used to provide the perceptual labelling of the image regions. Then, the binary spatial relations among the image regions (e.g., “region a is below region b”) are tabulated in spatial event cubes (SECs), based on which higher-order association rules are determined using Apriori algorithm [AS94]. For example, a third-order itemset is in the form of “If a region with label  $u_j$  is a right neighbor of a  $u_i$  region, it is likely that there is a  $u_k$  region on the right side of  $u_j$ ”. More recently, Teredesai et al. [TAKG06] proposed a framework to learn multi-relational visual-textual associations for image annotation. Within this framework, keywords and image visual features (including color saliency maps, orientation and intensity contrast maps) are extracted and stored separately in relational tables in a database. Then an FP-Growth algorithm [HPYM04] is used for extracting multi-relational associations between the visual features and keywords from the database tables. The extracted rules, such as “Intensity=4, Yellow  $\rightarrow$  EARTH, GROUND”, can be subsequently used for annotating new images.

### 2.4.2.3 Multimedia clustering

Clustering is a data mining technique that divides data into groups of similar objects [Ber02]. Data object groups are called clusters which help people to understand the “structure” of the data set.



CHAPTER 2. RELATED WORK

---

A heuristic is that media objects sharing similar semantics are likely to be perceptually similar and vice versa. Therefore, some research efforts utilize the clustering techniques to facilitate semi-automatic knowledge discovery processes, i.e., clustering is first carried out to automatically identify the groups of perceptually or semantic similar media objects, and then users manually analyze the extracted clusters to identify useful knowledge. Lamirel et al. presented a method that adopts a self-organizing map (SOM) [Koh01] to organize art images into clusters based on the textual features extracted from image descriptions [LDK00]. The image clusters can then be visualized using a topographic interface where two clusters that are similar in the textual feature space will appear close. With the assistance of the proposed topographic interface, users are able to extract new knowledge, such as semantic correlations between different art themes. Uehara et al. presented another effort of using SOM for semi-automatic knowledge discovery [UES<sup>+</sup>01]. Different from the work in [LDK00] which uses textual features to measure the image similarity, Uehara et al. extracted both textual and visual features, i.e., HSV color histograms and Mallat transformation coefficients, for calculating the Euclidean distances between images. Based on the Euclidean distances, images are clustered using an SOM. The smaller the Euclidean distance between the two images, the closer they are on the SOM, and the more similar they are in terms of perceptions and semantics. By visualizing the image clusters on the SOM, implicit knowledge, such as a design pattern like “flag images of Asian countries usually contain moons and stars”, can be identified by human users.

Besides the above applications, multimedia clustering is also used as an important technique for mining concepts in multimedia data collections. For example, Stan and Sethi proposed to use a hierarchical clustering method to extract concepts from an image collection [SS01] where each image is associated with a set of visual features (i.e., dominant HSV colors and their spatial layout) and textual features (i.e., keywords). Images

are first hierarchically clustered according to their visual features by recursively applying a variation of k-means clustering. For each resultant cluster, representative visual feature components (i.e., features with small variations) are identified. Then frequently occurred keywords are extracted to represent the cluster's semantics. In a related work, Stan and Sethi presented a visualization method that uses multidimensional scaling to visually present the image clusters to users, who can then interactively select proper image clusters for annotating new images [SS03].

## 2.5 Summary

In this chapter, we provide a survey of the existing research related to the topic of multimedia information fusion and analysis. Specifically, we reviewed four areas in detail, i.e., multimedia information fusion, multimedia modelling in the Semantic Web, automatic multimedia annotation, and multimedia analysis. In summary, the existing techniques, designed for traditional multimedia authoring, retrieval, and mining, lack the ability to model and automatically extract the semantic meanings of heterogenous multimedia data in a unified framework for cross-media information analysis. Starting from the next chapter, we shall present a suite of techniques developed as part of our proposed multimedia web information fusion and analysis framework for addressing the limitations of the existing approaches.

## Chapter 3

# Learning Image and Text Associations for Multimedia Information Fusion

### 3.1 Introduction

In this chapter, we focus on the problem of learning to identify relations between multimedia components, in particular, *image and text associations*. An image-text association refers to a pair of image and text segment that are semantically related to each other in a web page. A sample image-text association is shown in Figure 3.1. Identifying image-text associations enables a system to link visual contents of images (or video frames) with semantics conveyed in the web texts automatically.

Note that learning image-text association is similar to the task of automatic annotation [CMC05] (see Chapter 2) but has important differences. Whereas image annotation concerns annotating images using a set of predefined keywords, image-text association links images to free text segments in natural language. The methods for image annotation are thus not directly applicable to the problem of identifying image-text associations.

A key issue of using a machine learning approach to image-text associations is the lack of large training data sets. However, learning from small training data sets poses the new challenge of handling implicit associations. Referring to Figure 3.2, two associated

## CHAPTER 3. LEARNING IMAGE AND TEXT ASSOCIATIONS FOR MULTIMEDIA INFORMATION FUSION



Figure 3.1: An associated text and image pair.

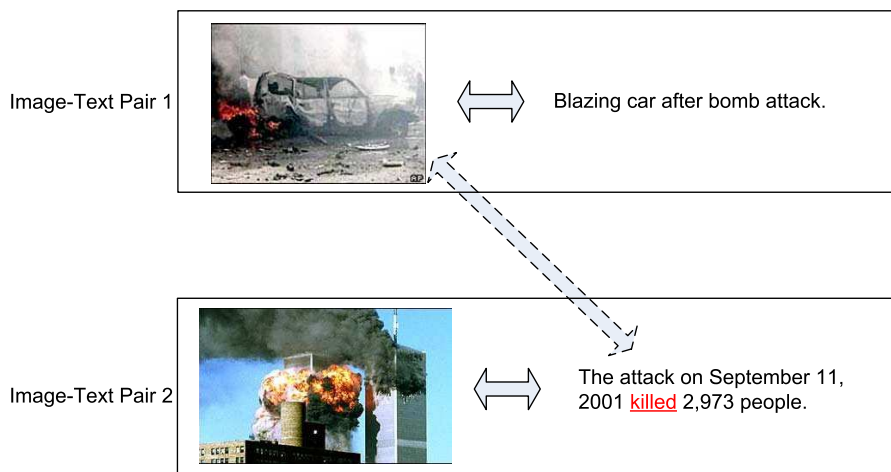


Figure 3.2: An illustration of implicit associations between visual and textual features among small number of training samples.

image-text pairs (I-T pairs) share partial visual (smokes and fires) and textual features (“attack”) but also have different visual and textual contents. As the two I-T pairs are actually on similar topics (describing scenes of terror attacks), the distinct parts, such as the visual content (“damaged car window”) of the image in I-T pair 1 and the term “killed” (underlined) in I-T pair 2, could be potentially associated. We call such domain association patterns, which do not explicitly appear in the training data samples but can be obtained by combining multiple data samples, *implicit associations*. We can imagine

that the smaller the data set is, the more domain association patterns will be missed in the individual data samples and the more implicit associations exist. Therefore, we need an algorithm which is capable of generalizing and combining the data samples to induce the missing implicit associations.

In this chapter, we present two methods, following the multilingual retrieval paradigm [OD96, Man98] for learning image-text associations. The first method is a textual-visual similarity model [JT06a] with the use of a statistical vague transformation technique for extracting associations between images and texts. As vague transformation typically requires large training data sets and tends to be computationally intensive, we employ a set of domain-specific information categories for *indirectly* matching the textual and visual information at the semantic level. With a small number of domain information categories, the training data sets for vague transformation need not be large and the computation cost can be minimized. In addition, as each information category summarizes a set of data samples, implicit image-text associations can be captured (see Section 3.2.3 for more details). As information categories may be inconsistently embedded in the visual and textual information spaces, we further employ a *visual space projection* method to transform the visual feature space into a new space, in which the similarities between the information categories are comparable to those in the textual information space. Our experiments show that employing visual space projection can further improve the performance of the vague transformation.

Considering that indirectly matching the visual and textual information using an intermediate tier of information categories may result in a loss of information, we develop another method based on an associative neural network model called fusion ART [TCG07], a direct generalization of Adaptive Resonance Theory (ART) model [CG87b] from one feature field to multiple pattern channels. Even with relatively small data sets where *implicit associations* tend to appear, fusion ART is able to automatically learn a set

of prototypical image-text pairs and therefore can achieve a good performance. This is consistent with the prior findings that ART models can efficiently learn useful patterns from small training data sets for text categorization [HTT03]. In addition, fusion ART can directly learn the associations between the features in the visual and textual channels without using a fixed set of information categories. Therefore, the information loss might be minimal.

The two proposed models have been evaluated on a multimedia document set in the terrorist domain collected from the BBC and CNN news web sites. The experimental results show that both vague transformation and fusion ART outperform a baseline method based on an existing state-of-the-art image annotation method known as Cross-Media Relevance Model (CMRM) [JLM03] in learning image-text associations from a small training data set. We have also combined the proposed methods with a pure text-matching based method matching image captions with text segments. We find that though the text based method is fairly reliable, the proposed cross-media methods consistently enhance the overall performance in image-text associations.

## 3.2 Identifying Image-Text Associations

### 3.2.1 A Similarity-Based Model for Discovering Image-Text Associations

The task of identifying image-text associations can be cast into an *information retrieval* (*IR*) problem. Within a web document  $d$  containing images and text segments, we treat each image  $I$  in  $d$  as a query to find a text segment  $TS$  that is most *semantically related* to  $I$ , i.e.,  $TS$  is most similar to  $I$  according to a predefined similarity measure function among all text segments in  $d$ . In many cases, an image caption can be obtained along with an image in a web document. Therefore, we suppose that each image  $I$  can be represented as a visual feature vector, denoted as  $\mathbf{v}^I = (v_1^I, v_2^I, \dots, v_m^I)$ , together with a

textual feature vector representing the image caption, denoted as  $\mathbf{t}^I = (t_1^I, t_2^I, \dots, t_n^I)$ . For calculating the similarity between an image  $I$  and a text segment  $TS$  represented by a textual feature vector  $\mathbf{t}^{TS} = (t_1^{TS}, t_2^{TS}, \dots, t_n^{TS})$ , we need to define a similarity measure  $sim_d(\langle \mathbf{v}^I, \mathbf{t}^I \rangle, \mathbf{t}^{TS})$ .

To simplify the problem, we assume that, given an image  $I$  and a text segment  $TS$ , the similarity between  $\mathbf{v}^I$  and  $\mathbf{t}^{TS}$  and the similarity between  $\mathbf{t}^I$  and  $\mathbf{t}^{TS}$  are independent. Therefore, we can calculate  $sim_d(\langle \mathbf{v}^I, \mathbf{t}^I \rangle, \mathbf{t}^{TS})$  with the use of a linear mixture model as follows:

$$sim_d(\langle \mathbf{v}^I, \mathbf{t}^I \rangle, \mathbf{t}^{TS}) = \lambda \cdot sim_d^{tt}(\mathbf{t}^I, \mathbf{t}^{TS}) + (1 - \lambda) \cdot sim_d^{vt}(\mathbf{v}^I, \mathbf{t}^{TS}). \quad (\text{Eq. 3.1})$$

In the subsequent sections, we first introduce our method used for measuring the textual similarities between image captions and text segments (Section 3.2.2). Then, two cross-media similarity measures based on vague transformation (Section 3.2.3) and fusion ART (Section 3.2.4) are presented.

### 3.2.2 Text-Based Similarity Measure

Matching between text-based features is relatively straightforward. We use the cosine distance

$$sim_d^{tt}(\mathbf{t}^I, \mathbf{t}^{TS}) = \frac{\sum_{k=1}^n t_k^I \cdot t_k^{TS}}{\|t^I\| \|t^{TS}\|} \quad (\text{Eq. 3.2})$$

to measure the similarity between the textual features of an image caption and a text segment. The cosine measure is used as it has been proven to be insensitive to the length of text documents.

### 3.2.3 Vague Transformation Based Cross-Media Similarity Measure

Measuring the similarity between visual and textual features is similar to the task of measuring relevance of documents in the field of multilingual retrieval for selecting documents

in one language based on queries expressed in another [OD96]. For multilingual retrieval, transformations are usually needed for bridging the gap between different representation schemes based on different terminologies.

An open problem is that there is usually a basic distinction between the vocabularies of different languages, i.e., word senses may not be organized with words in the same way in different languages. Therefore, an exact mapping from one language to another language may not exist. This is known as the *vague problem* [Man98], which is even more challenging in visual-textual transformation. For individual image regions, they can hardly convey any meaningful semantics without considering their contexts. For example, a yellow region can be a petal of a flower or can be a part of a flame. On the contrary, words in natural languages usually have a more precise meaning. In most cases, image regions can hardly be directly and precisely mapped to words because of the ambiguity. As vague transformations [BFL<sup>+</sup>88] [SB96] [Man98] have been proven useful in the field of multilingual retrieval, in this chapter, we borrow the idea from statistical vague transformation methods for cross-media similarity measure.

### 3.2.3.1 Single-Direction Vague Transformation

In our cross-media information retrieval model described in Section 3.2.1, the images are considered as the queries for retrieving the relevant text segments. Referring to the field of multilingual retrieval, transformation of the queries into the representation schema of the target document collection seems to be more efficient [Man98]. Therefore, we first investigate a *single-direction vague transformation* of the visual information into the textual information space.

A drawback of the existing methods for vague transformation, such as those presented in [BFL<sup>+</sup>88] and [SB96], is that they require a large training set to build multilingual thesauruses. In addition, as the construction of the multilingual thesauruses requires



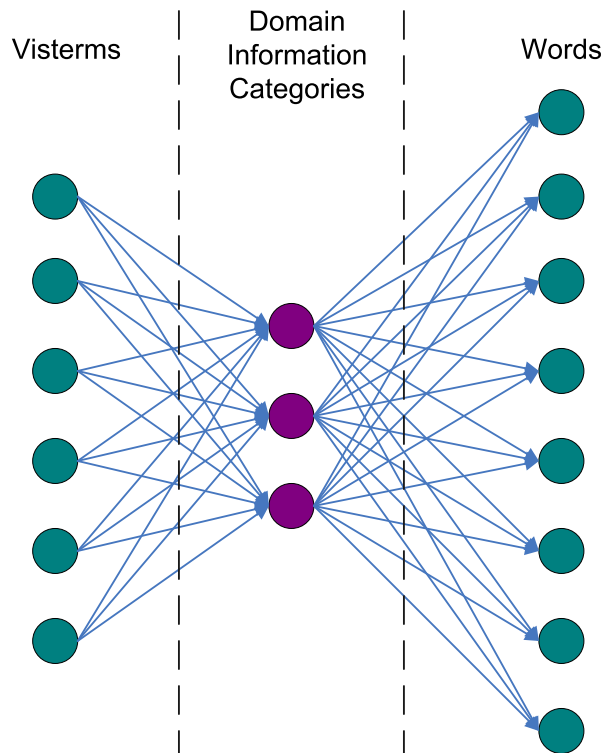


Figure 3.3: An illustration of cross-media transformation with information bottleneck.

calculating an association factor for each pair of words picked from two languages, it may be computationally formidable. To overcome these limitations, we introduce an intermediate layer for the transformation. This intermediate layer is a set of domain information categories which can be seen as another vocabulary of a smaller size for describing domain information. For example, in the terror attack domain, information categories may include *Attack Details*, *Impacts*, and *Victims* etc. Therefore, our cross-media transformation is in fact a concatenation of two sub-transformations, i.e., from visual feature space to domain information categories and then to textual feature space (see Figure 3.3). This is actually known as the *information bottleneck* method [TPB99]. For each sub-transformation, as the number of domain information categories is small, the size of the training data set for thesaurus construction need not be large and the construction cost can be affordable.

## CHAPTER 3. LEARNING IMAGE AND TEXT ASSOCIATIONS FOR MULTIMEDIA INFORMATION FUSION

In addition, given an associated pair of image and text, their contents may not be an exact match. For example, the associated text segment and image in Figure 3.1 are all related to a bombing event, but there is no direct correspondence between the words of the text segment and the image content. However, we believe that such a pair of text segment and image can always be mapped in terms of general domain information categories.

Based on the above observations, we build two thesauruses in the form of transformation matrices, each of which corresponds to a sub-transformation. Suppose the visual space  $\mathcal{V}$  is of  $m$  dimensions, the textual feature space  $\mathcal{T}$  is of  $n$  dimensions, and the cardinality of the set of high-level domain information categories  $\mathcal{C}$  is  $l$ . Based on  $\mathcal{V}$ ,  $\mathcal{T}$ , and  $\mathcal{C}$ , we define the two following transformation matrices:

$$M^{\mathcal{V}\mathcal{C}} = \begin{pmatrix} m_{11}^{\mathcal{V}\mathcal{C}} & m_{12}^{\mathcal{V}\mathcal{C}} & \cdot & \cdot & m_{1l}^{\mathcal{V}\mathcal{C}} \\ m_{21}^{\mathcal{V}\mathcal{C}} & m_{22}^{\mathcal{V}\mathcal{C}} & \cdot & \cdot & m_{2l}^{\mathcal{V}\mathcal{C}} \\ \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & & \cdot \\ m_{m1}^{\mathcal{V}\mathcal{C}} & m_{m2}^{\mathcal{V}\mathcal{C}} & \cdot & \cdot & m_{ml}^{\mathcal{V}\mathcal{C}} \end{pmatrix} \quad (\text{Eq. 3.3})$$

and

$$M^{\mathcal{C}\mathcal{T}} = \begin{pmatrix} m_{11}^{\mathcal{C}\mathcal{T}} & m_{12}^{\mathcal{C}\mathcal{T}} & \cdot & \cdot & \cdot & m_{1n}^{\mathcal{C}\mathcal{T}} \\ m_{21}^{\mathcal{C}\mathcal{T}} & m_{22}^{\mathcal{C}\mathcal{T}} & \cdot & \cdot & \cdot & m_{2n}^{\mathcal{C}\mathcal{T}} \\ \cdot & \cdot & & & & \cdot \\ m_{l1}^{\mathcal{C}\mathcal{T}} & m_{l2}^{\mathcal{C}\mathcal{T}} & \cdot & \cdot & \cdot & m_{ln}^{\mathcal{C}\mathcal{T}} \end{pmatrix}, \quad (\text{Eq. 3.4})$$

where  $m_{ij}^{\mathcal{V}\mathcal{C}}$  represents the association factor between the visual feature  $v_i$  and the information category  $c_j$ ; and  $m_{jk}^{\mathcal{C}\mathcal{T}}$  represents the association factor between the information category  $c_j$  and the textual feature  $t_k$ . In our current system,  $m_{ij}^{\mathcal{V}\mathcal{C}}$  and  $m_{jk}^{\mathcal{C}\mathcal{T}}$  are calculated by

$$m_{ij}^{\mathcal{V}\mathcal{C}} = P(c_j|v_i) \approx \frac{N(v_i, c_j)}{N(v_i)} \quad (\text{Eq. 3.5})$$

and

$$m_{jk}^{\mathcal{C}\mathcal{T}} = P(tm_k|c_j) \approx \frac{N(c_j, tm_k)}{N(c_j)}, \quad (\text{Eq. 3.6})$$

where  $N(v_i)$  is the number of images containing the visual feature  $v_i$ ;  $N(v_i, c_j)$  is the number of images containing  $v_i$  and belonging to the information category  $c_j$ ;  $N(c_j)$  is the number of text segments belonging to the category  $c_j$ ; and  $N(c_j, tm_k)$  is the number of text segments belonging to  $c_j$  and containing the textual feature (term)  $tm_k$ .

For calculating  $m_{ij}^{\mathcal{V}\mathcal{C}}$  and  $m_{jk}^{\mathcal{C}\mathcal{T}}$  in Eq. 3.5 and Eq. 3.6, we build a training data set of texts and images that have been manually classified into domain information categories (see Section 3.3 for details).

Based on Eq. 3.3 and Eq. 3.4, we can define the similarity between the visual part of an image  $\mathbf{v}^{\mathbf{I}}$  and a text segment represented by  $\mathbf{t}^{\mathbf{TS}}$  as  $(\mathbf{v}^{\mathbf{I}})^T M^{\mathcal{V}\mathcal{C}} M^{\mathcal{C}\mathcal{T}} \mathbf{t}^{\mathbf{TS}}$ . For embedding into Eq. 3.1, we use its normalized form

$$\text{sim}^{\mathcal{V}\mathcal{T}}(\mathbf{v}^{\mathbf{I}}, \mathbf{t}^{\mathbf{TS}}) = \frac{(\mathbf{v}^{\mathbf{I}})^T M^{\mathcal{V}\mathcal{C}} M^{\mathcal{C}\mathcal{T}} \mathbf{t}^{\mathbf{TS}}}{\| ((\mathbf{v}^{\mathbf{I}})^T M^{\mathcal{V}\mathcal{C}} M^{\mathcal{C}\mathcal{T}})^T \| \| \mathbf{t}^{\mathbf{TS}} \|}. \quad (\text{Eq. 3.7})$$

### 3.2.3.2 Dual-Direction Vague Transformation

Eq. 3.7 calculates the cross-media similarity using a single-direction transformation from visual feature space to textual feature space. However, it may still have the vague problem. For example, suppose there is a picture  $I$ , represented by the visual feature vector  $\mathbf{v}^{\mathbf{I}}$ , belonging to a domain information category *Attack Details*, and two text segments  $TS_1$  and  $TS_2$ , represented by the textual feature vectors  $\mathbf{t}^{\mathbf{TS}_1}$  and  $\mathbf{t}^{\mathbf{TS}_2}$ , belonging to the categories of *Attack Details* and *Victims* respectively. If the two categories *Attack Details* and *Victims* share many common words (such as *kill*, *die*, and *injure*), the vague transformation result of  $\mathbf{v}^{\mathbf{I}}$  might be similar to both  $\mathbf{t}^{\mathbf{TS}_1}$  and  $\mathbf{t}^{\mathbf{TS}_2}$ . To reduce the influence of common terms on different categories and utilize the strength of the distinct words, we consider another transformation from the word space to the visterm space. Similarly, we define a pair of transformation matrices  $M^{\mathcal{T}\mathcal{C}} = \{m_{kj}^{\mathcal{T}\mathcal{C}}\}^{n \times l}$  and  $M^{\mathcal{C}\mathcal{V}} = \{m_{ji}^{\mathcal{C}\mathcal{V}}\}^{l \times m}$ , where  $m_{kj}^{\mathcal{T}\mathcal{C}} = P(c_j | tm_k) \approx \frac{N(c_j, tm_k)}{N(tm_k)}$  and  $m_{ji}^{\mathcal{C}\mathcal{V}} = P(v_i | c_j) \approx \frac{N(v_i, c_j)}{N(c_j)}$  ( $i = 1, 2, \dots, m$ ,  $j = 1, 2, \dots, l$ ,

and  $k = 1, 2, \dots, n$ ). Here,  $N(tm_k)$  is the number of text segments containing the term  $tm_k$ ;  $N(c_j, tm_k)$ ,  $N(v_i, c_j)$ , and  $N(c_j)$  are same as those in Eq. 3.5 and Eq. 3.6. Then, the similarity between a text segment represented by the textual feature vector  $\mathbf{t}^{\text{TS}}$  and the visual content of an image  $\mathbf{v}^{\text{I}}$  can be defined as

$$\text{sim}^{\text{TV}}(\mathbf{t}^{\text{TS}}, \mathbf{v}^{\text{I}}) = \frac{(\mathbf{t}^{\text{TS}})^T M^{\text{TC}} M^{\text{CV}} \mathbf{v}^{\text{I}}}{\| ((\mathbf{t}^{\text{TS}})^T M^{\text{TC}} M^{\text{CV}})^T \| \| \mathbf{v}^{\text{I}} \|}. \quad (\text{Eq. 3.8})$$

Finally, we can define a cross-media similarity measure based on the *dual-direction transformation* which is the arithmetic mean of  $\text{sim}^{\text{VT}}(\mathbf{v}^{\text{I}}, \mathbf{t}^{\text{TS}})$  and  $\text{sim}^{\text{TV}}(\mathbf{t}^{\text{TS}}, \mathbf{v}^{\text{I}})$  given by

$$\text{sim}_d^{\text{vt}}(\mathbf{v}^{\text{I}}, \mathbf{t}^{\text{TS}}) = \frac{\text{sim}^{\text{VT}}(\mathbf{v}^{\text{I}}, \mathbf{t}^{\text{TS}}) + \text{sim}^{\text{TV}}(\mathbf{t}^{\text{TS}}, \mathbf{v}^{\text{I}})}{2}. \quad (\text{Eq. 3.9})$$

Note that as we consider that visual-to-text and text-to-visual transformations are two equivalent and independent processes, we use arithmetic mean for calculating the final similarity between the image and the text segment. In future, other methods for combination of the results of the two transformations can be explored. For example, we can use geometric mean or involve different weight parameters for the visual-to-text and text-to-visual transformations and employ expectation-maximization (EM) algorithm [DLR77] for parameter estimation.

### 3.2.3.3 Vague Transformation with Visual Space Projection

A problem in the reversed cross-media (text-to-visual) transformation in dual-direction transformation is that the intermediate layer, i.e., information categories, may be *embedded* differently in the textual feature space and the visterm space. For example, in Figure 3.4, two information categories, “Terrorist Suspects” and “Victims”, may contain quite different text descriptions but somewhat similar images, e.g. human faces. Suppose we translate a term vector of a text segment into the visual feature space using

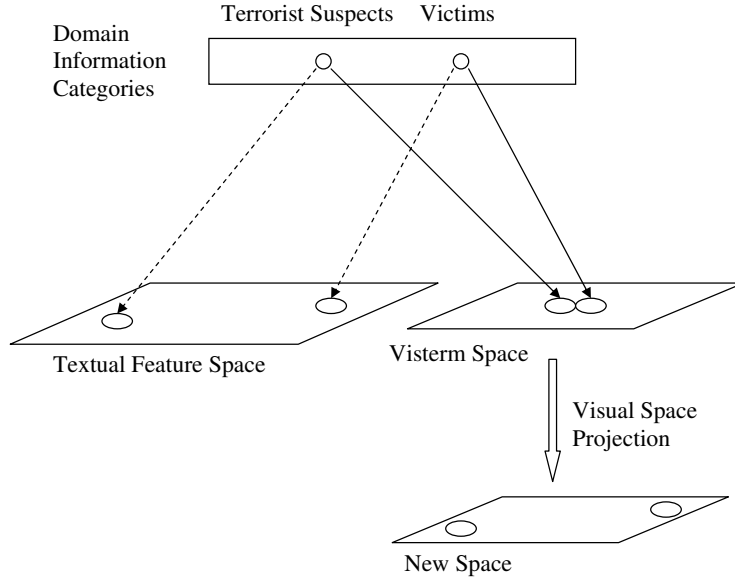


Figure 3.4: Visual space projection.

a cross-media transformation. Transforming a term vector in “Victims” category or a term vector in “Terrorist Suspects” category may result in a similar visual feature vector as these two information categories have similar representation in the visual feature space. In such a case when there are text segments belonging to the two categories in the same web page, we may not be able to select a proper text segment for an image about “Terrorist Suspects” or “Victims” based on the text-to-visual vague transformation.

For solving this problem, we need to consolidate the differences in the similarities between the information categories in the textual feature space and the visual feature space. We assume that text can more precisely represent the semantic similarities of the information categories. Therefore, we project the visual feature space into a new space in which the information category similarities defined in the textual feature space can be preserved.

We treat each row in the transformation matrix  $M^{CV}$  in Eq. 3.8 as a visual feature representation of an information category. We use a matrix  $X$  to transform the visterm space into a new space, wherein the similarity matrix of information categories can be rep-

resented as  $(M^{cv}X^T)(XM^{cvT}) = \{s_v(c_i, c_j)\}^{l \times l}$ , where  $s_v(c_i, c_j)$  represents the similarity between the information categories  $c_i$  and  $c_j$ . In addition, we suppose  $D = \{s_t(c_i, c_j)\}^{l \times l}$  is the similarity matrix of the information categories in the textual feature space, where  $s_t(c_i, c_j)$  is the similarity between the information categories  $c_i$  and  $c_j$ . Our objective is to minimize the differences between the information category similarities in the new space and the textual feature space. This can be formulated as an optimization problem of

$$\min_X \| D - M^{cv}X^T XM^{cvT} \|^2. \quad (\text{Eq. 3.10})$$

The similarity matrix  $D$  in the textual feature space is calculated based on our training data set, in which texts and images are manually classified into the domain specific information categories. Currently, two different methods have been explored for this purpose as follows.

- **Using Bipartite Graph of the Classified Text Segments** For constructing the similarity matrix of the information categories in the textual feature space, we utilize the bipartite graph of the classified text segments and the information categories as shown in Figure 3.5.

The underlying idea is that the more text segments that two information categories share, the more similar they are. We borrow the similarity measure used in [SB96] for calculating the similarity between information categories which is originally used for calculating term similarity based on bipartite graphs of terms and text documents. Therefore, any  $s_t(c_i, c_j)$  in  $D$  can be calculated as

$$s_t(c_i, c_j) = \sum_{TS_k \in c_i \cap c_j} wt(c_i, TS_k) \cdot wt(c_j, TS_k), \quad (\text{Eq. 3.11})$$

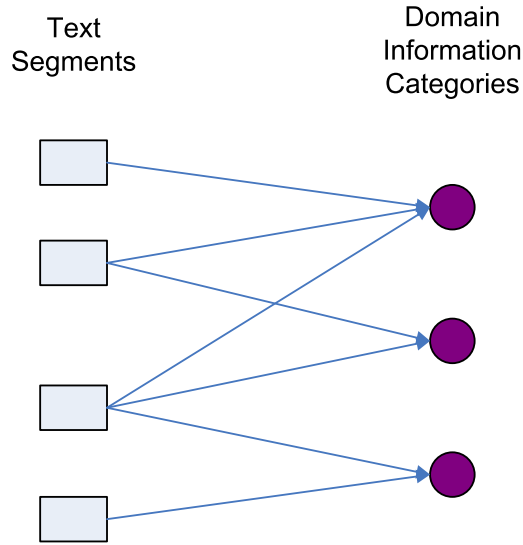


Figure 3.5: Bipartite graph of classified text segments and information categories.

where

$$wt(c_i, TS_k) = \frac{1/|TS_k|}{\sqrt{\sum_{TS_l \in c_i} (1/|TS_l|)^2}}, \quad (\text{Eq. 3.12})$$

where  $|TS_k|$  and  $|TS_l|$  represent the sizes of the text segments  $TS_k$  and  $TS_l$  respectively.

- **Using Category-To-Text Transformation Matrix** We also attempt another approach that utilizes the category-to-text vague transformation matrix  $M^{cT}$  in Eq. 3.4 for calculating the similarity matrix of the information categories. We treat each row in  $M^{cT}$  as a textual feature space representation of an information category. Then, we calculate the similarity matrix  $D$  in the textual feature space by

$$D = M^{cT} \cdot M^{cT^T}. \quad (\text{Eq. 3.13})$$

With the similarity matrix  $D$  of the information categories calculated above, the visual space projection matrix  $X$  can be solved based on Eq. 3.10. By incorporating  $X$ , Eq. 3.8 can be redefined as

$$\text{sim}^{TV}(\mathbf{t}^{\text{TS}}, \mathbf{v}^{\text{I}}) = \frac{(\mathbf{t}^{\text{TS}})^T M^{TC} M^{CV} X^T X \mathbf{v}^{\text{I}}}{\| ((\mathbf{t}^{\text{TS}})^T M^{TC} M^{CV} X^T X)^T \| \| \mathbf{v}^{\text{I}} \|}. \quad (\text{Eq. 3.14})$$

Using this refined equation in the dual-direction transformation, we expect that the performance of discovering the image-text associations can be improved. However, solving Eq. 3.10 is a non-linear optimization problem of a very large scale because  $X$  is a  $m \times m$  matrix, i.e., there are  $m^2$  variables to tune. Fortunately, from Eq. 3.14 we can see that we do not need to get the exact matrix  $X$ . Instead, we only need to solve a simple linear equation  $D = M^{CV} X^T X M^{CVT}$  to obtain a matrix

$$A = X^T X = M^{CV-1} D M^{CV-1T}, \quad (\text{Eq. 3.15})$$

where  $M^{CV-1}$  is the pseudo-inverse of the transformation matrix  $M^{CV}$ .

Then, we can substitute  $X^T X$  in Eq. 3.14 with  $A$  for calculating  $\text{sim}^{TV}(ts, i_v)$ , i.e., the similarity between a text segment represented by  $\mathbf{t}^{\text{TS}}$  and the visual content of an image represented by  $\mathbf{v}^{\text{I}}$ .

### 3.2.4 Fusion ART Based Cross-Media Similarity Measure

In the previous subsection, we presented the method for measuring the similarity between visual and textual information using a vague transformation technique. For the vague transformation technique to work on small data sets, we employ an intermediate layer of information categories to map the visual and textual information in an *indirect* manner. A deficiency of this method is that for training the transformation matrix, additional work on manual categorization of images and text segments is required. In addition, matching visual and textual information based on a small number of information categories may cause a loss of detailed information. Therefore, it would be appealing if we can find a method that learns *direct* associations between the visual and textual information. In this subsection, we present a method based on the fusion ART network, a generalization



of Adaptive Resonance Theory (ART) model [CG87b], for discovering direct mappings between visual and textual features.

#### 3.2.4.1 A similarity measure based on the resonance theory

As discussed, small data set does not have enough data samples, and thus many useful association patterns may appear in the data set implicitly. Those implicit associations may not be reflected in individual data samples but can be extracted by summarizing a group of data samples.

For learning implicit associations, we employ a method based on the fusion ART network. Fusion ART can be seen as multiple overlapping Adaptive Resonance Theory (ART) models [CG91] each of which corresponds to an individual information channel. Figure 3.6 shows a two-channel fusion ART (also known as Adaptive Resonance Associative Map [Tan95]) for learning associations between images and texts. The model consists of a  $F_2^c$  field, and two input pattern fields, namely  $F_1^{c1}$  for representing visual information channel of the images and  $F_1^{c2}$  for representing textual information channel of text segments. Such a fusion ART network can be seen as a simulation of a *physical resonance phenomenon* where each associated image-text pair can be seen as an information “*object*” that has a “natural frequency” in either visual or textual information channel represented by the visual feature vector  $\mathbf{v} = (v_1, v_2, \dots, v_m)$  or the textual feature vector  $\mathbf{t} = (t_1, t_2, \dots, t_n)$ . If two information objects have similar “natural frequencies”, strong resonance occurs. The strength of the resonance can be computed by a resonance function.

Given a set of multimedia information objects (associated image-text pairs) for training, the fusion ART learns a set of *multimedia information object templates*, or *object templates* in short. Each object template, recorded by a category node in the  $F_2^c$  field, represents a group of information objects that have similar “natural frequencies” and can

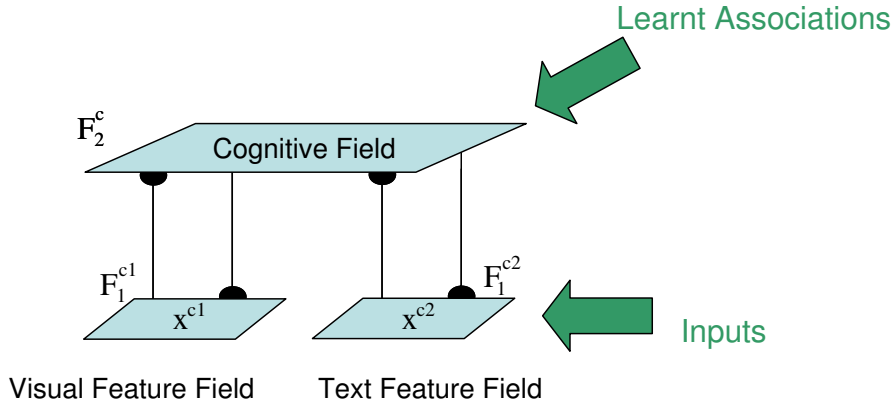


Figure 3.6: Fusion ART for learning image-text associations.

strongly resonate with each other. Initially, no object template (category node) exists in the  $F_2^c$  field. When information objects (associated image-text pairs) are presented one at a time to the  $F_1^{c1}$  and  $F_1^{c2}$  fields, the object templates are incrementally captured and encoded in the  $F_2^c$  field. The process of learning object templates using fusion ART can be summarized in the following stages:

- (i) Code Activation: A bottom-up propagation process first takes place when an information object is presented to the  $F_1^{c1}$  and  $F_1^{c2}$  fields. For each category node (multimedia information object template) in the  $F_2^c$  field, a *resonance score* is computed using an *ART choice function*. The ART choice function varies with respect to different ART models, including ART 1 [CG87b], ART 2 [CG87a], ART 2-A [CGR91a] and fuzzy ART [CGR91b]. We adopt the ART 2 choice function based on the cosine similarity which has been proven to be effective for measuring vector similarities and insensitive to the vector lengths.

Given a pair of visual and textual information feature vectors  $\mathbf{v}$  and  $\mathbf{t}$ , for each  $F_2^c$  category node  $j$  with a visual information template  $\mathbf{v}^{c_j}$  and a textual information template  $\mathbf{t}^{c_j}$ , the resonance score  $T_j$  is calculated by:

$$T_j = \gamma \frac{\mathbf{v} \cdot \mathbf{v}^{c_j}}{\|\mathbf{v}\| \|\mathbf{v}^{c_j}\|} + (1 - \gamma) \frac{\mathbf{t} \cdot \mathbf{t}^{c_j}}{\|\mathbf{t}\| \|\mathbf{t}^{c_j}\|}, \quad (\text{Eq. 3.16})$$

where  $\gamma$  is the factor for weighing the visual and textual information channels. For giving the equal weight to the visual and textual information channels, we set the  $\gamma$  value to 0.5.  $\frac{\mathbf{v} \cdot \mathbf{v}^{c_j}}{\|\mathbf{v}\| \|\mathbf{v}^{c_j}\|}$  and  $\frac{\mathbf{t} \cdot \mathbf{t}^{c_j}}{\|\mathbf{t}\| \|\mathbf{t}^{c_j}\|}$  are actually the ART 2 choice function for visual and textual information channels.

- (ii) Code Competition: A code competition process follows under which the  $F_2^c$  node  $j$  with the highest resonance score is identified.
- (iii) Template Matching: Before the node  $j$  can be used for learning, a template matching process checks that for each (visual and text) information channel, the object template of node  $j$  is sufficiently similar to the input object with respect to the norm of the input object. The similarity value is computed by using the ART 2 match function [CG87a]. Resonance can occur if for each (visual and textual) channel, the match function value meets a vigilance criterion.

The vigilance criterion defines the minimum fraction of the input object that must be similar with the matched object template. Low vigilance criterion means that the input object and the matched template can have small similarity and thus cause broad generalization and coarse template learning. On the contrary, high vigilance criterion means that the input object and the matched template should have large similarity and therefore lead to narrow generalization and fine template learning. Choosing a proper vigilance criterion usually depends on the characteristics of the data set. When a data set contains noisy data, a relevantly low vigilance criterion should be used to allow pattern generalization for noise tolerant during the pattern learning. When a data set does not contain much noise, a relevantly high vigilance criterion should be chosen to avoid too broad generalization. For learning implicit

association between visual and textual features, a median vigilance criterion of 0.6 is selected for both visual and textual information channels to encourage certain degree of template generalization.

- (iv) **Template Learning:** Once a node  $j$  is selected, its object template will be updated by linearly combining the object template with the input information object according to a predefined learning rate [CG87a].
- (v) **New Code Creation:** When no category node is sufficiently similar to the new input information object, a new category node is added to the  $F_2^c$  field. Fusion ART thus expands its network architecture dynamically in response to incoming information objects.

The advantage of using fusion ART is that its object templates are learnt by incrementally combining and merging new information objects with previously learnt object templates. Therefore, a learnt object template is a “summarization” of characteristics of the training objects. For example, the three training I-T pairs as shown in Figure 3.7 can be summarized by fusion ART into one object template and thereby the implicit associations across the I-T pairs can be captured. In particular, the implicit associations between frequently occurred visual and textual contents (such as the visual content “damaged car” and the term “kill” in Figure 3.7) can be learnt for predicting new image-text associations.

Suppose a new pair of image and text segment is presented to the  $F_1^{c1}$  and  $F_1^{c2}$  fields of a trained fusion ART. If the image and the text segment are semantically relevant, their information object should be able to strongly resonate with an learnt object template in the trained fusion ART; otherwise, a weak resonance may occur. In other words, we can measure the similarity or relevance between an image and a text segment according to their resonance score (see Eq. 3.16) in a trained fusion ART. Such a resonance based similarity measure has been used in existing work for data terrain analysis [LON05].

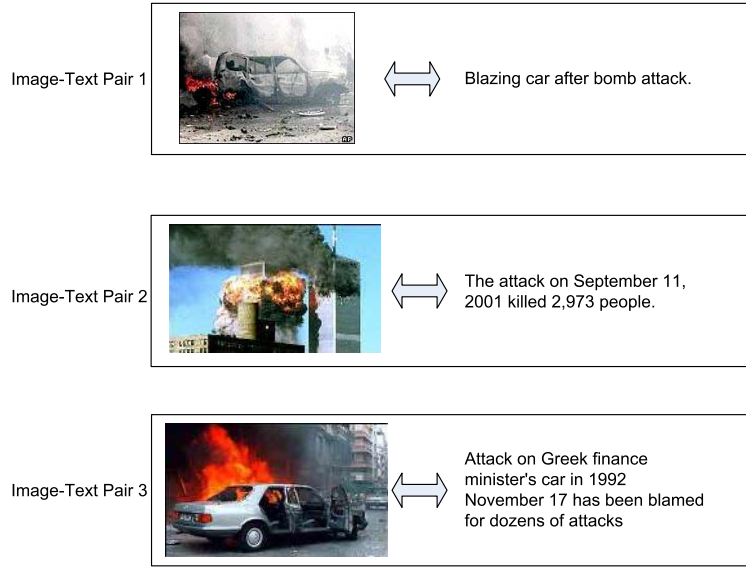


Figure 3.7: Training samples for fusion ART.

### 3.2.4.2 Image Annotation Using fusion ART

In the above discussion, we define a cross-media similarity measure based on the resonance function of the fusion ART. Based on this similarity measure, we can identify an associated text segment represented by textual feature vector  $\mathbf{t}$  that is considered most “similar” to an image represented by the visual feature vector  $\mathbf{v}$ . At the same time, we will identify an  $F_2^c$  object template  $j$  with a textual information template  $\mathbf{t}^{c_j}$  having the strongest resonance with the image-text pair represented by  $\langle \mathbf{v}, \mathbf{t} \rangle$ . Based on the  $\mathbf{t}$  and  $\mathbf{t}^{c_j}$ , we can extract a set of keywords for annotating the image. The process is described as follows:

- (i) For the  $k$ th dimension of the textual information vectors  $\mathbf{t}$  and  $\mathbf{t}^{c_j}$ , if  $\min\{t_k, t_k^{c_j}\} > 0$ , extract the term  $tm_k$  corresponding to the  $k$ th dimension of the textual information vectors for annotating the image.
- (ii) When  $tm_k$  is extracted, a confidence value of  $\min\{t_k, t_k^{c_j}\}$  is assigned to  $tm_k$  based

on which all extracted keywords can be ranked for annotating the image<sup>1</sup>. We can see that the confidence value of a term is calculated by  $\min\{t_k, t_k^{c_j}\} > 0$ , which is a combination of the term weight for a specific text segment and the weight for a learnt textual information template. Therefore, intuitively a term that is highly weighted in both the text segment associated with the image and the domain information template will be used for the image annotation.

We should note that the image annotation task is possible because the fusion ART can capture the direct associations between the visual and textual information. The performance evaluation of using fusion ART for image annotation is beyond the scope of this chapter. However, the fusion ART based image annotation method has a unique advantage over the existing image annotation methods, namely it does not require a predefined set of keywords. A set of examples of using the fusion ART based method for image annotation will be presented in the next section.

## 3.3 Experimental Results

### 3.3.1 Data Set

The experiments are conducted on a web page collection, containing 300 images related to terrorist attacks, downloaded from the CNN and BBC news web sites. For each image, we extract its caption (if available) and long text paragraphs (with more than 15 words) from the web page containing the image.

For applying vague transformation based methods that utilize an intermediate layer of information categories, we manually categorize about 1500 text segments and 300 images into 15 predefined domain information categories, i.e., *Anti-Terror*, *Attack Detail*, *After*

---

<sup>1</sup>As many other methods, a difficulty of using fusion ART for image annotation is to decide the number of annotation keywords that should be used. The discussion on this problem is out of the scope of this thesis.

*Attack, Ceremony, Government Emergency Preparation, Government Response, Impact, Investigation, Rescue, Simulation, Attack Target, Terrorist Claim, Terrorist Suspect, Victim, and Others.* The detailed data preprocessing methods are described as follows.

### 3.3.1.1 Textual Feature Extraction

We treat each text paragraph as a text segment. Preprocessing the text segments includes tokenizing the text segments, part-of-speech tagging, stop word filtering, stemming, removing unwanted terms (retaining only nouns, verbs and adjectives), and generating the textual feature vectors where each dimension corresponds to a remaining term after the preprocessing.

For calculating the term weights of the textual feature vectors, we use a model, named TF-ITSF (term frequency and inverted text segment frequency), similar to the traditional TF-IDF model. For a text segment  $TS$  in a web document  $d$ , we use the following equation to weight the  $k$ th term  $tm_k$  in  $TS$ 's textual feature vector  $\mathbf{t}^{\text{TS}} = (t_1^{\text{TS}}, t_2^{\text{TS}}, \dots, t_n^{\text{TS}})$ :

$$t_k^{\text{TS}} = tf(TS, tm_k) \cdot \log \frac{N(d)}{tsf(d, tm_k)}, \quad (\text{Eq. 3.17})$$

where  $tf(TS, tm_k)$  denotes the frequency of  $tm_k$  in the text segment  $TS$ ,  $N(d)$  is the total number of text segments in the web document  $d$ , and  $tsf(d, tm_k)$  is the text segment frequency of term  $tm_k$  in  $d$ . Here, we use the text segment frequency for measuring the importance of a term for a text segment in a web document.

After a textual feature vector  $\mathbf{t}^{\text{TS}} = (t_1^{\text{TS}}, t_2^{\text{TS}}, \dots, t_n^{\text{TS}})$  is extracted, L1-normalization is applied for normalizing the term weights into a range of  $[0, 1]$ :

$$\mathbf{t}^{\text{TS}} = \frac{(t_1^{\text{TS}}, t_2^{\text{TS}}, \dots, t_n^{\text{TS}})}{\max\{t_{i=1..n}^{\text{TS}}\}}, \quad (\text{Eq. 3.18})$$

where  $n$  is the number of textual features (i.e., terms).

### 3.3.1.2 Visual Feature Extraction from Images

Our visual feature extraction method is inspired by those used in the existing image annotation approaches [CMC05] [BF01] [JLM03] [DBdFF02] [FML04]. During image preprocessing, each image is first segmented into  $10 \times 10$  rectangular regions. There are two reasons why we use rectangular regions instead of employing more complex image segmentation techniques to obtain image regions. The first reason is that existing work on the image annotation shows that using rectangular regions can provide performance gain compared with using regions obtained by automatic image segmentation techniques [FML04]. The second reason is that using rectangular regions is much less time-consuming and therefore suitable for the online multimedia web information fusion task.

For each image region, we extract a visual feature vector, consisting of six color features and 60 gabor texture features. The color features are the means and variances of the RGB color spaces. The texture features are extracted by calculating the means and variations of the Gabor filtered image regions in six orientations at five different scales. Such a set of color and textual features have been proven to be useful for image classification and annotation tasks [Sha98] [DOP03].

After the feature vectors are extracted, all image regions are clustered using the  $k$ -means algorithm. The purpose of the clustering is to discretize the continuous color and texture features [BF01] [DBdFF02] [FML04]. The generated clusters, called *visterms* represented by  $\{vt_1, vt_2, \dots, vt_k\}$ , are treated as a vocabulary for describing the visual content of the images. An image is described by a visterm  $vt_j$  if it contains a region belonging to the  $j$ th cluster. For the terrorist domain data set, the visterm vocabulary is enriched with a high-level semantic feature, extracted by a face detection model provided by the OpenCV. In total, a visterm vector of  $k+1$  features is extracted for each image.



The weight of each feature is the corresponding visterm frequency normalized with the use of L1-normalization.

A problem of using visterms is how we can determine a proper number of visterms  $k$  (i.e., the number of clusters for k-means clustering). Note that the images have been manually categorized into the fifteen information categories which reflect the images' semantic meanings. Therefore, images belonging to different information categories should have different patterns in their visterm vectors. If we cluster images into different clusters based on their visterm vectors, in the most ideal case, images belonging to different categories should be assigned into different clusters. Based on the above consideration, we can measure the meaningfulness of the visterm sets with different  $k$  values by calculating the *information gains* of the image clustering results. The definition of the information gain given below is similar to the one used in the Decision Tree for measuring partitioning results with respect to different data attributes.

Given a set  $S$  of images belonging to  $m$  information categories, the *information need* for classifying the images in  $S$  is measured by

$$I(S) = - \sum_{i=1}^m \frac{s_i}{\|S\|} \log\left(\frac{s_i}{\|S\|}\right), \quad (\text{Eq. 3.19})$$

where  $\|S\|$  is the total number of images and  $s_i$  is the number of images belonging to the  $i$ th information category.

Suppose we cluster an image collection  $S$  into  $n$  clusters, i.e.,  $S_1, S_2, \dots, S_n$ . The information gain can be calculated as follows:

$$\text{Gain} = I(S) - \sum_{j=1}^n \frac{\|S_j\|}{\|D\|} I(S_j), \quad (\text{Eq. 3.20})$$

where  $\|S_j\|$  is the number of images in the  $j$ th cluster.

Figure 3.8 shows the information gains obtained by clustering our image collection based on visterm sets with a varying number of visterms. We can see no matter how

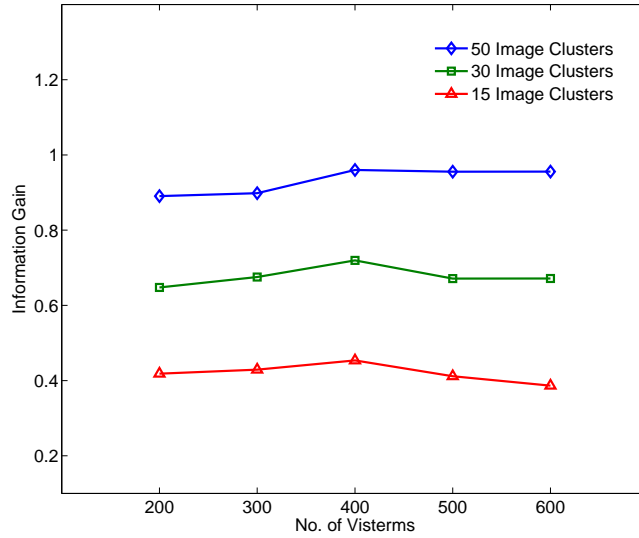


Figure 3.8: Information gains of clustering the images based on a varying number of visterms.

many clusters of images we generate, the largest information gain is always achieved when  $k$  is around 400. Based on this observation, we generate 400 visterms for the image visterm vectors.

### 3.3.2 Performance Evaluation Method

We adopt a five-fold cross-validation to test the performance of our methods. In each experiment, we use four folds of the data (240 images) for training and one fold (60 images) for testing. The performance is measured in terms of precision defined by

$$precision = \frac{N_c}{N}, \quad (\text{Eq. 3.21})$$

where  $N_c$  is the number of correctly identified image-text associations and  $N$  is the total number of images. The correctness of the extracted image-text associations are judged by human inspecting the web pages wherein the images appear. We experimented with different  $\lambda$  values for our cross-media retrieval models (see Eq. 3.1) to find the best *balance*

*point* of weighting the impact of textual and visual features. However, we should note that in principle, the best  $\lambda$  could be obtained by using an algorithm such as expectation-maximization (EM) [DLR77].

Note that whereas the most information retrieval tasks use both precision and recall to evaluate the performance, we only use precision in our experiment. This is simply due to the fact that for each image, there is only one text segment considered to be semantically relevant. In addition, we also extract only one associated text segment for each image using the various models. Therefore, in our experiments, the precision and recall values are actually the same.

### 3.3.3 Evaluation of Cross-Media Similarity Measure Based on Visual Features Only

We first evaluate the performance of the cross-media similarity measures, defined in subsection 3.2.3 and subsection 3.2.4, by setting  $\lambda = 0$  in our linear mixture similarity model (see Eq. 3.1), i.e., using only visual contents of images (without image captions) for measuring image-text associations. As there has been no prior work on image-text association learning, we implement a baseline method based on the cross-media relevance model (CMRM) proposed by Jeon et al [JLM03]. The CMRM model is designed for image annotation by estimating the conditional probability of observing a term  $w$  given the observed visual content of an image. As our objective is to associate an entire text segment to an image, we extend Jeon et al's model to calculate the average conditional probability of observing terms in a text segment given the visual content of an image. The reason of using the average conditional probability, instead of the joint conditional probability, is that we need to minimize the influence of the length of the text segments. Note that the longer a text segment is, the smaller the joint conditional probability tends to be. Table 3.1 summarizes the six methods that we experimented for discovering image-text associations based on pure visual contents of images. The first four methods are vague

## CHAPTER 3. LEARNING IMAGE AND TEXT ASSOCIATIONS FOR MULTIMEDIA INFORMATION FUSION

transformation based cross-media similarity measures that we define in subsection 3.2.3. The last two methods are the fusion ART (object resonance) based similarity measure and the CMRM model. Figure 3.9 shows the performance of using the various models for extracting image-text associations based on a five-fold cross-validation.

Table 3.1: Cross-media models.

Model	Descriptions
SDT	Single-direction vague transformation
DDT	Dual-direction vague transformation
DDT_VP_BG	DDT with visual space projection using bipartite graph based similarity matrix
DDT_VP_CT	DDT with visual space projection using category-to-text transformation based similarity matrix
fusion ART	The fusion ART based cross-media resonance model
CMRM	Cross-media relevance model

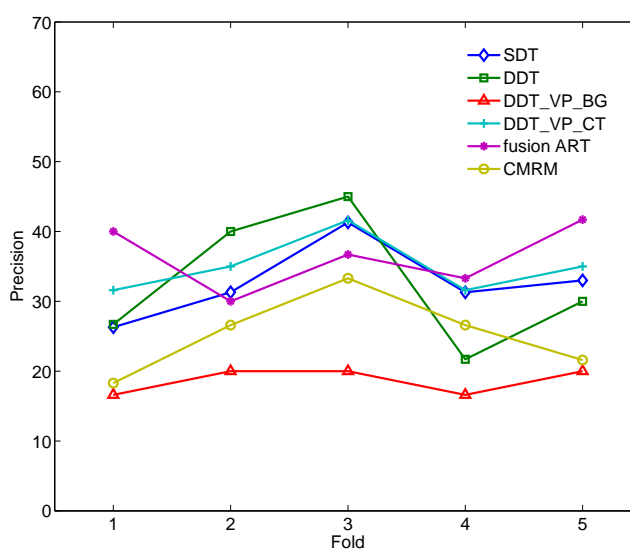


Figure 3.9: Comparison of cross-media models for discovering image-text associations.

We see that among the six methods, DDT\_VP\_CT and fusion ART provide the best performance. They outperform SDT and DDT which have a similar performance. All of these four models perform much better than CMRM model and DDT\_VP\_BG. It is

## CHAPTER 3. LEARNING IMAGE AND TEXT ASSOCIATIONS FOR MULTIMEDIA INFORMATION FUSION

surprising DDT\_VP\_BG is the worst method, hinting that the similarity matrix calculated based on bipartite graphs cannot really reflect the semantic similarity between the domain information categories. We shall revisit this issue in the next subsection.

For evaluating the impact of the size of training data on the learning performance, we also experiment with different data sizes for training and testing. Figure 3.10 shows the performance of the six cross-media similarity models with respect to training data of various sizes. We can see that when the size of the training data decreases, the precision of the CMRM model drops dramatically. In contrast, the performance of vague transformation and fusion ART drop less than 10% in terms of average precision. It shows that our methods also provide better performance stability on small data sets comparing with the statistical based cross-media relevance model.

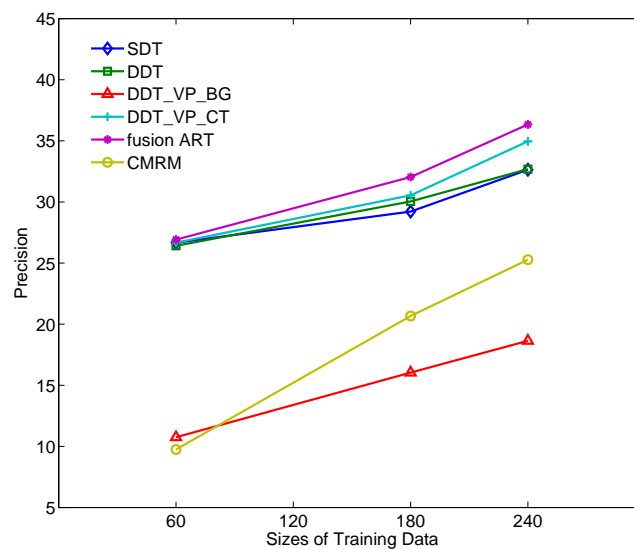


Figure 3.10: Performance comparison of cross-media models with respect to different training data sizes.

### 3.3.4 Evaluation of Linear Mixture Similarity Model

In this section, we study the effect of using both textual and visual features in the linear mixture similarity model for discovering image-text associations. Referring to the experimental results in Table 3.2, we see that textual information is fairly reliable in identifying image-text associations. In fact, the pure text similarity measure ( $\lambda = 1.0$ ) outperforms the pure cross-media similarity measure ( $\lambda = 0.0$ ) by 20.7% to 24.4% in terms of average precision.

Table 3.2: The average precision scores (%) for image-text association extraction.

Methods	$\lambda$										
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
SDT	<u>32.6</u>	41.6	48.6	56.0	59.3	60.6	61.3	<b>61.6</b>	60.6	57.6	57.0
DDT	<u>32.7</u>	51.7	59.0	59.7	<b>61.7</b>	61.3	60.7	60.0	59.3	58.3	57.0
DDT_VP_BG	<u>18.6</u>	22.3	28.0	35	40.0	45.6	51.6	54.3	56.6	<b>57.3</b>	57.0
DDT_VP_CT	<u>35.0</u>	40.3	45.3	50.6	54.6	59.3	61.3	<b>62.6</b>	62.3	60.3	57.0
fusion ART	<u>36.3</u>	42.3	49.0	55.3	57.7	60.3	<b>62.0</b>	61.7	58.3	58.0	57.0

However, the best result is achieved by the linear mixture model using both the text-based and the cross-media similarity measures. DDT\_VP\_CT with  $\lambda = 0.7$  can achieve an average precision of 62.6%, whilst the fusion ART with  $\lambda = 0.6$  can achieve an average precision of 62.0%. On the average, the mixture similarity models can outperform the pure text similarity measure by about 5%. This shows that visual features are also useful in the identification of image-text associations. In addition, we observe that combining cross-media and text-based similarity measures improves the performance of pure text similarity measure on each fold of the experiment. Therefore, such improvement is stable. In fact, keywords extracted from the captions of the images sometimes may be inconsistent with the contents of the images. For example, an image on the 911 attack scene may have a caption on the ceremony of 911 attack, such as “Victims’ families will

## CHAPTER 3. LEARNING IMAGE AND TEXT ASSOCIATIONS FOR MULTIMEDIA INFORMATION FUSION

tread Ground Zero for the first time”. In such a case, visual features can compensate the imprecision in the textual features.

			
Caption In Web Pages	Police photograph the body of the gunman.	Wreckage of the base of the World Trade Center. The CIA searched the wreckage.	Injured man being helped away.
Text-Based Measure ( $\lambda = 1.0$ )	<b>At least five people have died, and several others have been injured, in several incidents, including a shooting by a Palestinian gunman in the Israeli town of Kfar Saba, and a suicide bomb attack in north Jerusalem. (SC=0.089)</b>	<b>The CIA sent a special team to scour the wreckage for vital intelligence reports after the attack, the paper says. (SC=0.268)</b>	Others were not even able to do that. One witness said he saw several people lying on the floor of the bus, including one man whose legs had been blown off. (SC=0.110)
Cross-Media Measure (DDT_VP_CT, $\lambda = 0.0$ )	<b>At least five people have died, and several others have been injured, in several incidents, including a shooting by a Palestinian gunman in the Israeli town of Kfar Saba, and a suicide bomb attack in north Jerusalem. (SC=0.157)</b>	A secret CIA office was destroyed in the 11 September attack on the World Trade Center, the New York Times reports. (SC=0.157)	<b>It was here on Thursday that a Palestinian suicide bomber blew himself up on board a crowded bus, killing five people and injuring about 50 others. (SC=0.087)</b>
Cross-Media Measure (Fusion ART, $\lambda = 0.0$ )	<b>At least five people have died, and several others have been injured, in several incidents, including a shooting by a Palestinian gunman in the Israeli town of Kfar Saba, and a suicide bomb attack in north Jerusalem. (SC=0.430)</b>	<b>The CIA sent a special team to scour the wreckage for vital intelligence reports after the attack, the paper says. (SC=0.97)</b>	<b>It was here on Thursday that a Palestinian suicide bomber blew himself up on board a crowded bus, killing five people and injuring about 50 others. (SC=0.265)</b>
Mixture Measure (DDT_VP_CT, $\lambda = 0.7$ )	<b>At least five people have died, and several others have been injured, in several incidents, including a shooting by a Palestinian gunman in the Israeli town of Kfar Saba, and a suicide bomb attack in north Jerusalem. (SC=0.109)</b>	<b>The CIA sent a special team to scour the wreckage for vital intelligence reports after the attack, the paper says. (SC=0.181)</b>	Others were not even able to do that. One witness said he saw several people lying on the floor of the bus, including one man whose legs had been blown off. (SC=0.093)
Mixture Measure (Fusion ART, $\lambda = 0.6$ )	<b>At least five people have died, and several others have been injured, in several incidents, including a shooting by a Palestinian gunman in the Israeli town of Kfar Saba, and a suicide bomb attack in north Jerusalem. (SC=0.268)</b>	<b>The CIA sent a special team to scour the wreckage for vital intelligence reports after the attack, the paper says. (SC=0.69)</b>	<b>It was here on Thursday that a Palestinian suicide bomber blew himself up on board a crowded bus, killing five people and injuring about 50 others. (SC=0.203)</b>

Figure 3.11: A sample set of image-text associations extracted with similarity scores (SC). The correctly identified associated texts are bolded.

Among the vague transformation methods, dual-direction transformation achieves almost the same performance as single-direction transformation. However, visual space projection with dual-direction transformation can slightly improve the average precision. We can also see that the bipartite graph based similarity matrix  $D$  for visual space pro-

jection does not improve the image-text association results. By examining the classified text segments, we notice that only a small number of text segments belong to more than one categories and contribute to category similarities. This may have resulted in an inaccurate similarity matrix and a biased visual space projection.

On the other hand, the performance of fusion ART is comparable with that of vague transformation with visual space projection. Nevertheless, when using the pure cross-media model ( $\lambda = 0.0$ ), fusion ART can actually outperform vague transformation based methods by about 1% to 3%. Looking into each fold of the experiment, we see that the fusion ART based method is much more stable than the vague transformation based methods in the sense that the best results are almost always achieved with  $\lambda = 0.6$  or  $0.7$ . For the vague transformation based methods, the best result of each experiment fold is obtained with rather different  $\lambda$  values. This suggests that the vague transformation based methods are more sensitive to the training data.

A sample set of the extracted image-text associations is shown in Figure 3.11. We find that cross-media models are usually good in associating general domain keywords with images. Referring to the second image in Figure 3.11, cross-media models can associate an image depicting the attack scene of 911 attack with a text segment containing the word “attack” which is commonly used for describing terrorist attack scenes. However, for the word “wreckage” that is a more specific word, cross-media models usually cannot identify it correctly. For such cases, using image captions may be helpful. On the other hand, as discussed before, image captions may not always reflect the image content accurately. For example, the caption of the third image contains the word “man”, which is a very general term, not quite relevant to the terrorist event. For such cases, cross-media models can be useful to find the proper domain-specific textual information based on the visual features of the images.

Figure 3.12 shows a sample set of the results by using fusion ART for image annotation. We can see that such annotations can reflect the direct associations between the






Images	Keyword Annotation Using fusion ART
	complex (0.5), house (0.33), attack (0.25), suicide (0.2), people (0.17), bomb (0.16), children (0.14)
	train (0.2), bomb (0.17), thursday (0.1)
	yorker (0.5), trade (0.5), terror (0.33), america (0.33), world (0.25), center (0.2), plane (0.2)

Figure 3.12: Samples of image annotations using fusion ART.

visual and textual features in the images and texts. For example, the visual cue of “debris” in the images may be associated with words, such as “bomb” and “terror” in the text segments. Discovering such direct associations is an advantage of the fusion ART based method.

### 3.3.5 Discussions and Comparisons

In table 3.3, we provide a summary of the key characteristics of the two proposed methods. First of all, we note that the underlying ideas of the two approaches are quite different. Given a pair of image and text segment, the vague transformation based method translates features from one information space into another information space so that features of different spaces can be compared. The fusion ART based method, on the other hand, learns a set of prototypical image-text associations and then predicts the degree of association between an incoming pair of image and text segment by comparing it with the learned associations. During the prediction process, the visual and textual information are first compared in their respective spaces and the results consolidated based on a multimedia object resonance function (ART choice function).

## CHAPTER 3. LEARNING IMAGE AND TEXT ASSOCIATIONS FOR MULTIMEDIA INFORMATION FUSION

Table 3.3: Comparison of the vague transformation and the fusion ART based methods.

	Vague Transformation	fusion ART
Approach	Information are translated from one space to another space, so that information from different spaces can be compared	Visual and textual information are compared in their respective spaces and the results consolidated based on a multimedia object resonance function
Learning methodology	Statistic based batch learning	Incremental competitive learning
Information encoding	Using transformation matrices to summarize the information based on predefined information categories	Using self-organizing networks to learn typical categories of multimedia information objects
Network size	Number of predefined information categories is small. The information summarized in the transformation matrices is relatively more compact	More category nodes are created
Speed	Around 20 seconds for training; 20 seconds for testing	Around 120 seconds for training; 30 seconds for testing
Performance Stability	Unstable	Stable

The vague transformation is a statistic based method which calculates the conditional probabilities in one information space given some observations in the other information space. To calculate such conditional probabilities, we need to perform batch learning on a fixed set of training data. Once the transformation matrices are trained, they cannot be updated without building from scratch. In contrast, the fusion ART based method adopts an incremental competitive learning paradigm. The trained fusion ART can always be updated when new training data are available.

The vague transformation based method encodes the learnt conditional probabilities in transformation matrices. A fixed number of domain-specific information categories is used to reduce the information complexity. Instead of using predefined information categories, the fusion ART based method can automatically organize multimedia information objects into typical categories. The characteristics of an object category are encoded by a multimedia information object template. There are usually more category nodes learnt by the fusion ART. Therefore, the information in the fusion ART is less compact than that in the transformation matrices. In our experiments, around 70 to 80 categories are learnt by the fusion ART on a data set containing 300 images.

In terms of efficiency, the vague transformation based method runs much faster than the fusion ART based method during both training and testing. However, the fusion ART based method produces a more stable performance than that of the vague transformation based method (see discussions in Section 3.3.4).

## 3.4 Summary

The contributions of this chapter are summarized as follows:

- Two distinct methods are presented for learning and extracting associations between images and texts from multimedia web documents. The vague transformation

CHAPTER 3. LEARNING IMAGE AND TEXT ASSOCIATIONS FOR MULTIMEDIA INFORMATION FUSION

---

based method utilizes an intermediate layer of information categories for capturing indirect image-text associations. The fusion ART based method learns direct associations between image and text features by employing a resonance environment of the multimedia objects.

- Extensive experiments are conducted to show the efficacy of the proposed methods. The experimental results suggest that both methods are able to efficiently learn image-text associations from a small training data set. Most notably, they both perform significantly better than the baseline performance provided by a typical image annotation model. In addition, while the text-matching based method is still more reliable than the cross-media similarity measures, combining visual and textual features provides the best overall performance in discovering cross-media associations between images and texts in multimedia web documents.

By linking an image a web text segment, a semantic representation can be extracted from the web text for describing the image content. In the next chapter, a multimedia content representation schema is presented for describing contents of various multimedia data in an unified form of machine-processible and human-understandable semantic metadata. In addition, a method for automatically extracting semantic metadata based on textual web contents is proposed.

## Chapter 4

# Semantic Multimedia Content Representation and Metadata Extraction

This chapter endeavors to present our approach of multimedia content modelling for supporting multimedia information fusion and analysis. Different from most multimedia content modelling methods which link multimedia objects with discrete keywords or semantic concepts, our method aims to describe multimedia content with *conceptual graph*, a specific type of semantic networks. Using conceptual graphs, semantic concepts are interconnected with relations to express the detailed semantic contexts. This is inspired by the knowledge representation research in the field of artificial intelligence for simulating thinking processes of human brains. In addition, we present an approach for semantic metadata extraction, which automatically extracts conceptual graphs from web texts and encodes them in machine-processible and human-understandable RDF metadata.

## 4.1 Semantic Multimedia Content Representation

### 4.1.1 Knowledge Representation and Human Brain

Research in the field of knowledge representation shows that human thinks in the form of associations, i.e., by relating certain semantic concepts to others. Therefore, knowledge in human brains can be modelled as *semantic networks* composed of concepts and relations [Sow91]. Semantic network forms the foundation of knowledge representation and semantic modelling in the field of artificial intelligence and Semantic Web research [Sow00, bie00, LS99]. Using concepts and relations for semantic modelling is also consistent with the human's way of organizing the semantics in natural languages using terms and relations to convey semantic meanings [Sow84, Mal06].

Naturally, different domains concern different sets of concepts and relations. In the semantic based information systems, such domain-specific concepts and relations are usually defined in the form of ontologies. An *ontology* is an explicit specification of a conceptualization, which defines a representational vocabulary for a domain [Gru93, Gru95]. Based on the vocabulary defined in the ontology, domain knowledge, concerning domain specific instances (instantiated concepts) and facts (instantiated relations) about the instances, can be formally described and represented for sharing and reusing. Therefore, semantic modelling of domain knowledge involves two layers, i.e., an *ontology layer* and an *instance/fact layer*. Combining the ontology layer and the instance/fact layer forms an integrated domain-specific knowledge base (see Figure 4.1). In our framework, we follow this two-layer model for representing multimedia semantic contents. Note that our purpose is to model facts for information analysis and knowledge discovery, but not for supporting reasoning and inference [Sow00]. Therefore, logical rules or axioms, which are widely used in knowledge based systems for the reasoning and inference purpose, are not included in our framework.

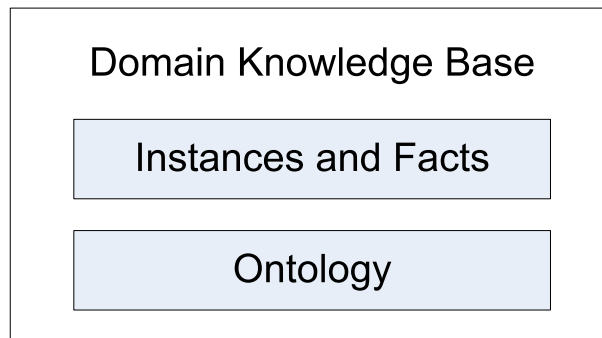


Figure 4.1: Semantic modelling of domain knowledge.

### 4.1.2 Semantic Modelling of Multimedia Contents Based on Web Texts

Considering the fact that texts, reflecting semantic model in human brains, are widely available on the Web, borrowing semantic information in texts to describe the related multimedia data can be a practical way towards automatic multimedia content modelling. To model multimedia content for the task of information analysis, a prerequisite problem we need to consider is that what form of semantic representations of texts should be extracted to allow useful knowledge to be discovered. We believe that if a piece of knowledge is useful and interesting, it should be precise and detailed without ambiguities. In addition, the extracted knowledge should also be easy to understand for human users. To achieve these requirements, the representations of multimedia contents should be in a precise, detailed and human-understandable form.

However, existing techniques used in textual information analysis, i.e., text mining [DGS99, Tan99], mainly transform text documents into simplistic intermediate forms, e.g. term vectors and bags of keywords. As terms are treated as individual items in such simplistic representations, terms lose their semantic relations and texts lose their original meanings. For example, in Figure 4.2, two short text documents with different meanings can be represented in a similar bag of keywords, e.g. {France, Defeat, Italy, World Cup, Quarter Final}. In the original documents, “France” and “Italy” have different roles in

- |  |
|--|
| <ol style="list-style-type: none"><li>1. France defeated Italy in the World Cup quarter final.</li><li>2. France was defeated by Italy in the World Cup quarter final.</li></ol> |
|--|

Figure 4.2: Two short text documents with similar keywords but of distinct semantic meanings.

the events of “Defeat”. However, the semantic relations depicting the conceptual roles are lost in the bag-of-keyword representations. Therefore, the original meanings of the documents in Figure 4.2 cannot be discriminated any more. More recently, there are some efforts conducted to design new features to capture the semantic meanings to a certain extent. For example, Xue and Zhou [XZ06] have designed a distributional feature for capturing compactness and positions of the first appearance of the words. However, such features still cannot capture the precise document semantics, such as the relations between the concepts. Without precisely representing the detailed document meanings, text mining techniques can only discover shallow patterns, such as term associations, deviations, and document clusters, which are statistical patterns of terms, not knowledge about text semantics. In this thesis, we aim to overcome the limitation of existing technologies to extract and represent the detailed meanings of the text for modelling multimedia semantics and supporting information analysis at a precise semantic level. For this purpose, we need a representation method that expresses the semantic relations between the concepts in texts.

*Resources Description Framework (RDF)*, proposed by the World Wide Web Consortium (W3C), is a language specification for modelling machine-processable and human-readable semantic metadata to describe web resources on the Semantic Web [BLHL01]. The basic element of RDF is RDF statements which are triplets in the form of <subject, predicate, object>. An RDF statement can express that there is a relation (represented by the predicate) between the subject and the object. In [BL01], Tim Berners-Lee further illustrated that RDF can be interworkable with *conceptual graphs* [Sow84] [Sow99]



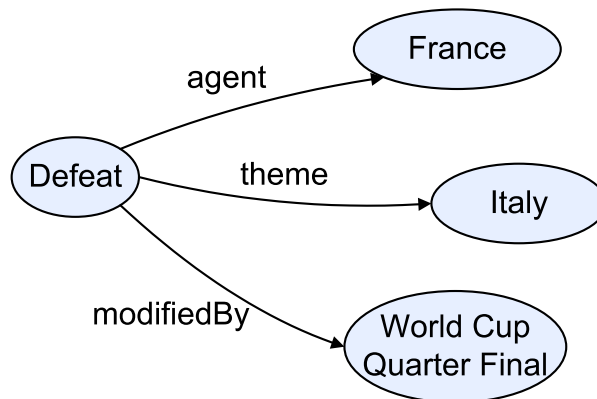


Figure 4.3: A simplified conceptual graph translated from the first sentence in Figure 4.2.

which serve as an intermediate language for translating natural languages into computer-oriented formalisms. As the full conceptual graph standard is complex for large scale applications, simplified conceptual graphs are used in many existing practices, such as [GMV99]. In our work, we also use a set of simplified conceptual graphs containing only three kinds of predicates, i.e., “agent”, “theme”, and “modifiedBy”, for representing the key semantics of text. (The detailed information of the three predicates will be introduced in Section 4.2.) As an example, Figure 4.3 shows such a simplified conceptual graph translated from the first sentence in Figure 4.2. We treat each directed arc in the conceptual graph as a semantic relation consisting of a subject (the start node of the arc), a predicate (the label or type of the arc), and an object (the end node of the arc). Each of these relations can be encoded using an RDF statement. For example, the three semantic relations in the conceptual graph in Figure 4.3 can be represented using the RDF statements  $\langle \text{Defeat}, \text{agent}, \text{France} \rangle$ ,  $\langle \text{Defeat}, \text{theme}, \text{Italy} \rangle$  and  $\langle \text{Defeat}, \text{modifiedBy}, \text{World Cup Quarter Final} \rangle$ . We can see that by using RDF and conceptual graphs, detailed semantics of textual information can be precisely captured and represented in a machine-processible and human-understandable form.

For modelling multimedia semantics in our framework, we adopt a two-layer model, composed of domain vocabularies at the ontology layer and conceptual graphs at the

CHAPTER 4. SEMANTIC MULTIMEDIA CONTENT REPRESENTATION AND METADATA EXTRACTION

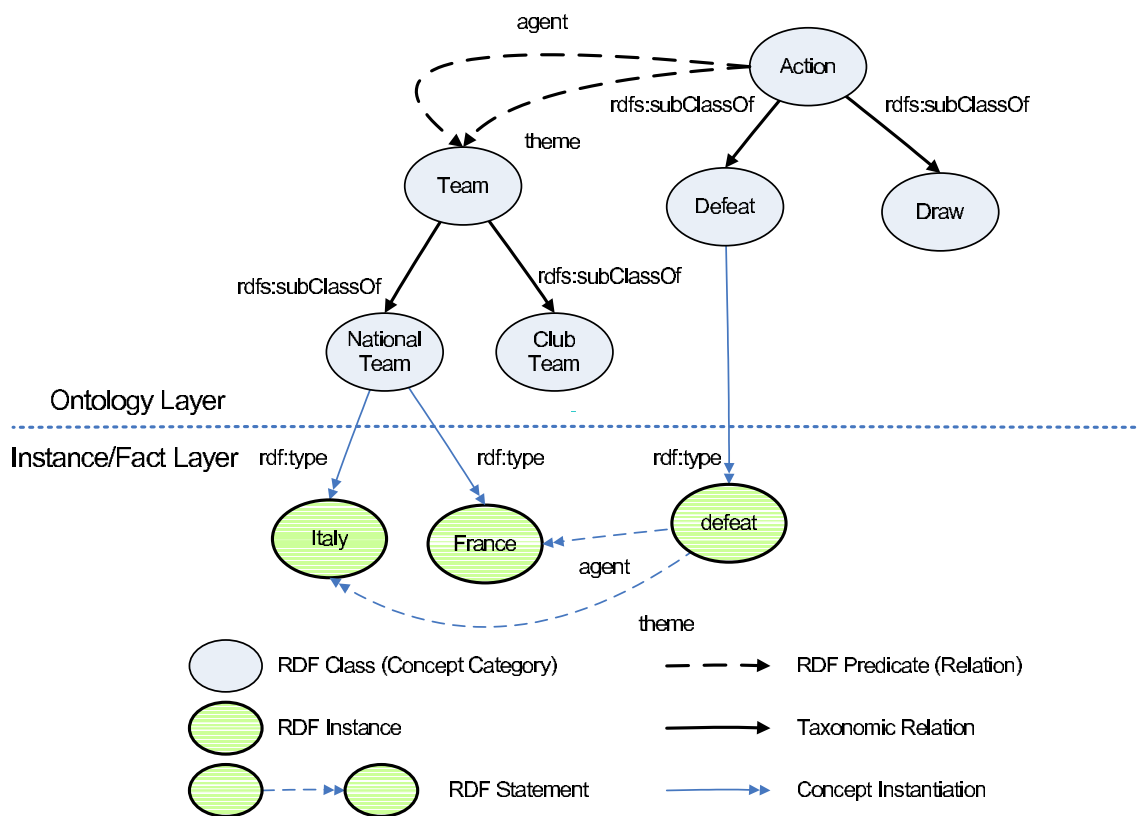


Figure 4.4: Knowledge modelling in the sports domain using RDF.

## CHAPTER 4. SEMANTIC MULTIMEDIA CONTENT REPRESENTATION AND METADATA EXTRACTION

```

<?xml version='1.0'?>
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:sports="http://www.sports.org/sports.rdf"/>

<!-- The ontology layer starts here... -->

<!-- Defining RDF Classes (Concept Categories) -->
<rdfs:Class rdf:ID="Team">
  <rdfs:label>Team</rdfs:label>
</rdfs:Class>

<rdfs:Class rdf:ID="NationalTeam">
  <rdfs:label>National Team</rdfs:label>
  <rdfs:subClassOf rdf:resource="#Team"/>
</rdfs:Class>

<rdfs:Class rdf:ID="Action">
  <rdfs:label>Action</rdfs:label>
</rdfs:Class>

<rdfs:Class rdf:ID="Defeat">
  <rdfs:label>defeat</rdfs:label>
  <rdfs:label>beat</rdfs:label>
  <rdfs:subClassOf rdf:resource="#Action"/>
</rdfs:Class>
.
.
.
<!-- Defining RDF Predicates (Relations) -->
<rdf:Property rdf:ID="agent">
  <rdfs:label>agent</rdfs:label>
  <rdfs:domain rdf:resource="#Action" />
  <rdfs:range rdf:resource="#Team" />
</rdf:Property>

<rdf:Property rdf:ID="theme">
  <rdfs:label>theme</rdfs:label>
  <rdfs:domain rdf:resource="#Action" />
  <rdfs:range rdf:resource="#Team" />
</rdf:Property>
<!-- The ontology layer ends here. -->

<!-- The instance/fact layer starts here... -->

<!-- Instances -->
<rdf:Description rdf:ID="instance1">
  <rdfs:label>defeat</rdfs:label>
  <rdf:type rdf:resource="http://www.sports.org/sports.rdf#Defeat" />
</rdf:Description>

<rdf:Description rdf:ID="instance2">
  <rdfs:label>France</rdfs:label>
  <rdf:type rdf:resource="http://www.sports.org/sports.rdf#NationalTeam" />
</rdf:Description>

<!-- Facts: relations between instances -->
<rdf:Description rdf:ID="instance1">
  <sports:agent rdf:resource="http://www.sports.org/sports.rdf#instance2" />
</rdf:Description>

<!-- The instance/fact layer ends here. -->
</rdf:RDF>

```

Figure 4.5: XML representation of the RDF metadata in the sports domain.

instance/fact layer, for representing multimedia semantic contents. For automatic construction of ontologies, we develop a set of methods for identifying the semantic-equivalent terms in web texts as concepts and organized them into a *term taxonomy* for reflecting the semantic relevance between different concepts. The details about term semantic modelling, similarity measuring, and term taxonomy construction will be introduced in Section 4.2.

Both domain ontologies and conceptual graph representations are encoded using RDF and RDF Schema. Besides the semantic modelling portion, an XML based syntax is also proposed by W3C in the RDF specification [LS99, CLS01] to facilitate RDF's interchangeability. For example, Figure 4.4 shows a sports domain RDF model. Its XML representation is presented in Figure 4.5. As RDF is online-interchangeable, human-readable, and machine-processable, it has the potential to significantly improve the efficiency of the information exchange between human and computers. For example, information users can use RDF to formulate an information query which can be well understood by RDF based search engines. Therefore, interesting information can be more precisely retrieved based on the users' information needs. Furthermore, based on the semantic based RDF data collections, data mining and machine learning techniques can also be applied to discover useful knowledge, which is easy to explain for the human users.

### 4.1.3 Integrating RDF Semantic Description and MPEG-7 standard

Although RDF and conceptual graphs provide us a fundamental framework for representing the semantic contents of multimedia objects, other techniques are still needed to allow us to relate such semantic contents with the media-specific information, such as author information, creation information, and media types. MPEG-7, a multimedia annotation standard [Mar04, Per01, Day02], developed by Moving Picture Experts Group (MPEG) is explored for this task.

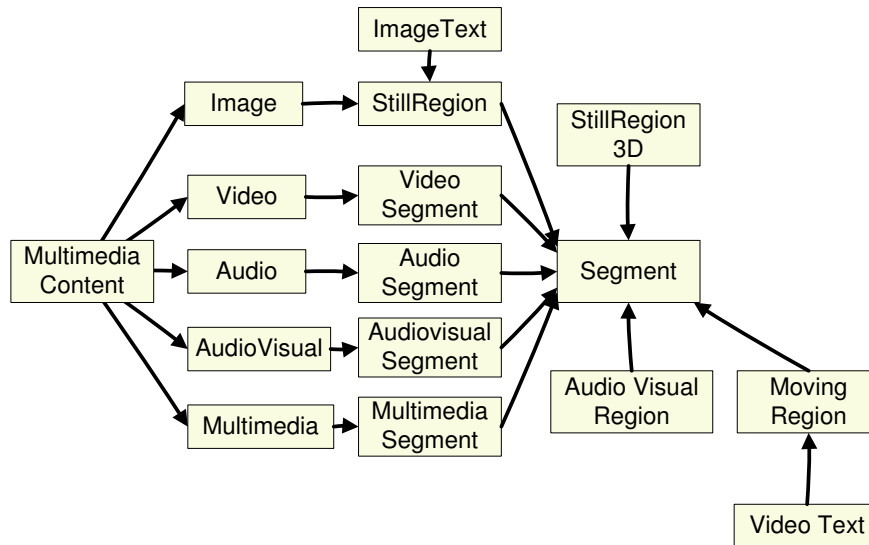


Figure 4.6: A hierarchy of media objects supported by MPEG-7.

MPEG-7, formally named "Multimedia Content Description Interface", is an ISO/IEC standard developed by MPEG (Moving Picture Experts Group) for describing the multimedia content data that supports some degree of interpretation of the information, which can be processed by a machine or a computer program. It mainly provides users with a rich set of standardized tools to enable the generation of multimedia descriptions for various media formats. Figure 4.6 shows a hierarchy of media objects that are incorporated in the MPEG-7 standard. The main elements of the MPEG-7 standard are described as follows.

- **Descriptors (D)**, representations of features, that define the syntax and the semantics of each feature representation.
- **Description Schemes (DS)**, that specify the structure and semantics of the relationships between their components. These components may be both Descriptors and Description Schemes.
- **A Description Definition Language (DDL)**, to allow the creation of new De-

scription Schemes and, possibly, Descriptors and to allow the extension and modification of existing Description Schemes.

- **System Tools**, to support multiplexing of descriptions, synchronization of descriptions with content, transmission mechanisms, coded representations (both textual and binary formats) for efficient storage and transmission, management and protection of intellectual property in MPEG-7 descriptions, etc.

Initially, MPEG-7 standard uses the XML Schema language to specify the descriptors and description schemes. However it is increasingly known that XML schema based MPEG-7 specifications have some deficiencies in reusing and interoperating with other domains. With this consideration, Hunter [Hun01] proposed to build an MPEG-7 ontology based on an RDF-based language. The RDF MPEG-7 ontology in [Hun01] mainly focuses on providing users a set of descriptors to represent media-specific information, such as author information, creation information, as well as media format and low-level perceptual features.

For conceptual and semantic aspects, though conceptual representations of multimedia contents are considered in the original MPEG-7 standard by defining XML descriptors of events, objects, and concepts, these general descriptors are too abstract to model domain-specific concepts and relations. In addition, the semantics of those XML descriptors are not clear. An XML descriptor can either have values like a relation/property or have properties/relations with other descriptors like a concept. For example, Figure 4.7 shows a semantic description of an image using the MPEG-7 descriptors. We can see that the “Event DS” descriptor has a value of “play” and a “location of” relation to the “SemanticPlace DS” descriptor. Moreover, the values of these descriptors are usually represented as keywords. For example, in Figure 4.7, the value of the “Event DS” descriptor is a text keyword “play”. As discussed in Chapter 1, the meaning of the text keywords may be vague and therefore may lower the performance of the multimedia systems.

## CHAPTER 4. SEMANTIC MULTIMEDIA CONTENT REPRESENTATION AND METADATA EXTRACTION

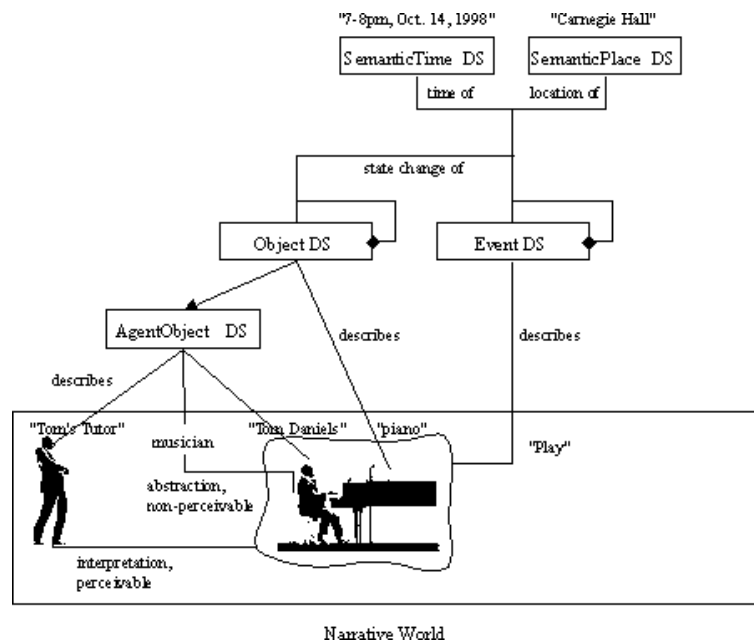


Figure 4.7: An example of conceptual aspects description using MPEG-7 descriptors, adopted from [Mar04].

To solve the semantic modelling problems in MPEG-7, we extend the MPEG-7 ontology by adding in three RDF concept definitions and one relation (property) definition as shown in Figure 4.8. Our strategy is to simplify the conceptual modelling in MPEG-7 by linking media objects through “semantic description” relation to two types of semantic concepts, i.e., “Semantic Object” and “Semantic Event”. Semantic Object is for representing a media object with an individual concept, whilst Semantic Event corresponds to multimedia content using conceptual graphs. The detailed domain-specific information is modelled by incorporating the domain ontologies. An example of combining the sports domain ontology and the extended MPEG-7 ontology is shown in Figure 4.9, wherein an image is described by a “Semantic Event” instance represented by a conceptual graph modelled based on the sports domain ontology.

CHAPTER 4. SEMANTIC MULTIMEDIA CONTENT REPRESENTATION AND METADATA EXTRACTION

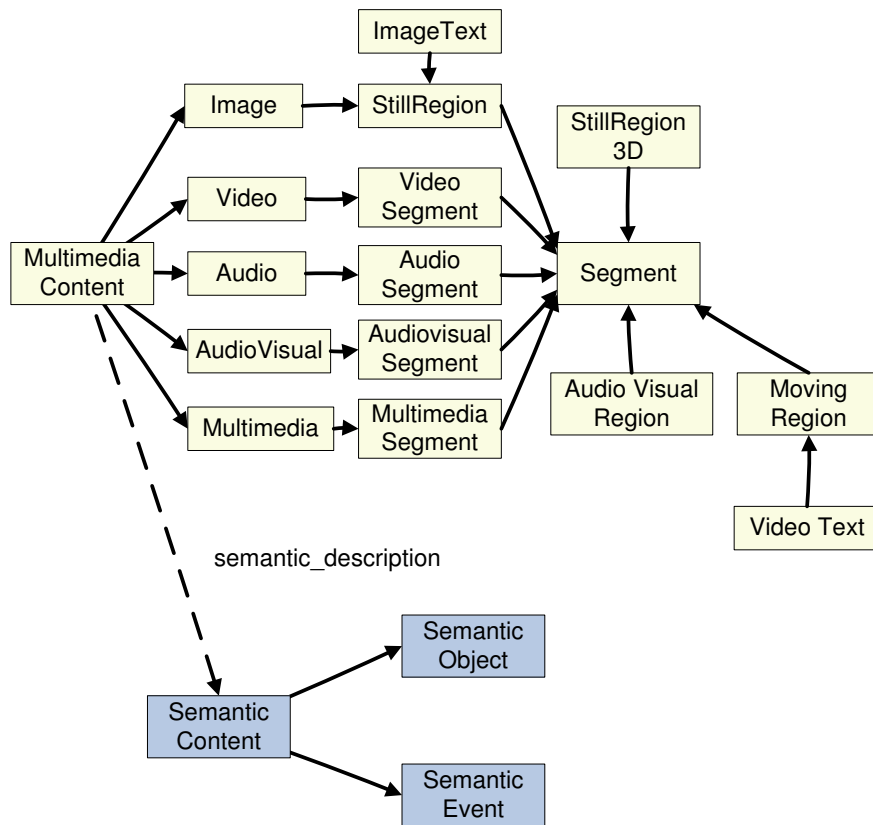


Figure 4.8: An extended MPEG-7 ontology.



## CHAPTER 4. SEMANTIC MULTIMEDIA CONTENT REPRESENTATION AND METADATA EXTRACTION

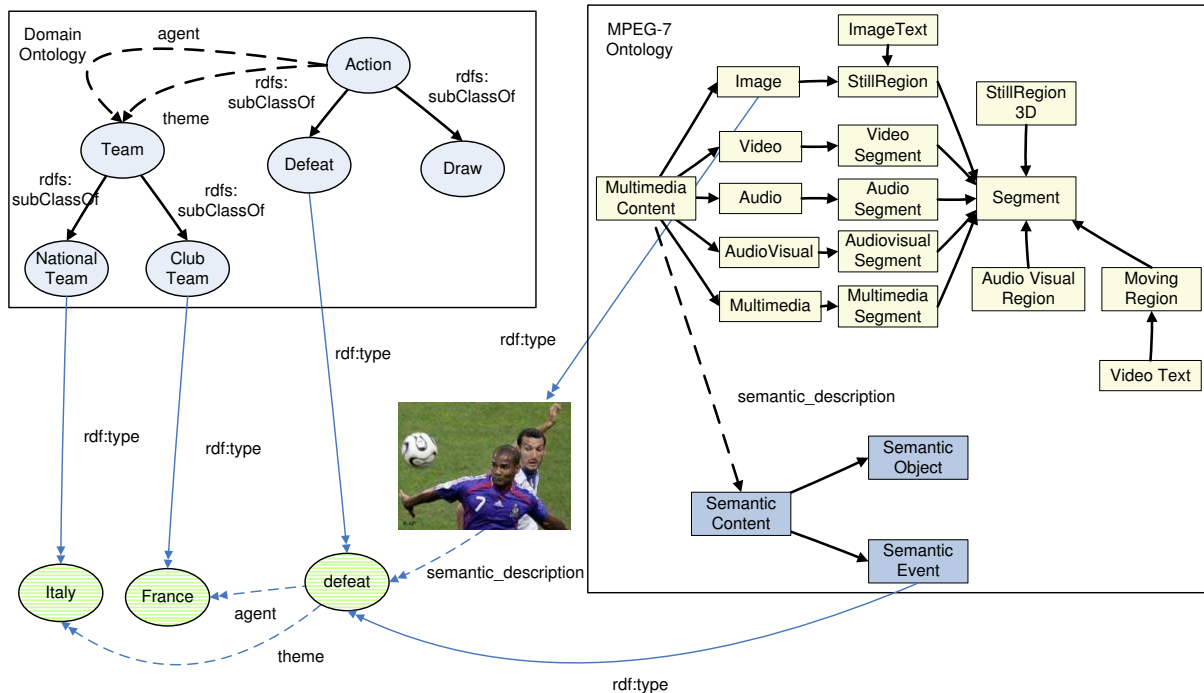


Figure 4.9: Combining domain ontology and MPEG-7 ontology for media object description.

#### 4.1.4 A Four-Layer Data Model

Based on the semantic modelling techniques introduced in the previous sections, we adopt a four-layer data model in our multimedia web information fusion and analysis framework as shown in Figure 4.10.

- (i) The **web document layer** represents original web documents without preprocessing.
- (ii) The **multimedia object layer** stores the media objects extracted from decomposed web documents.
- (iii) The **RDF metadata layer** expresses the semantic contents of the media objects as well as the content irrelevant media properties, such as image formats (e.g. JPEG, BMP, or TIFF) and video length.

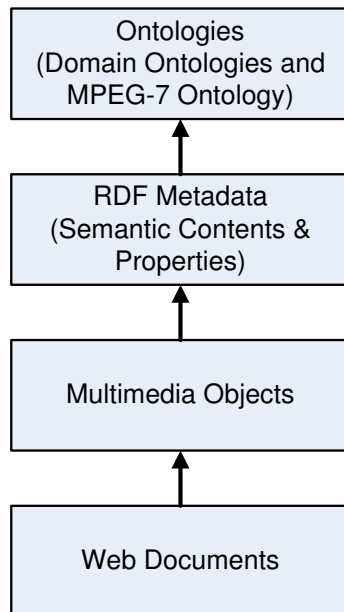


Figure 4.10: A four-layer data model in the META system.

- (iv) The **ontology layer** contains the definitions of the domain ontologies and MPEG-7 ontology, which provide vocabularies for describing semantic contents and content irrelevant media properties.

## 4.2 Semantic Metadata Extraction

A procedure for RDF metadata extraction is shown in Figure 4.11. The raw textual web content is sequentially preprocessed by a pronominal coreference resolution module and a POS (part-of-speech) tagging and syntax parsing module. For eliminating the ambiguities of the pronouns in text, we employ the pronominal coreference resolution function of Gate [CMBT02], a natural language processing (NLP) toolkit developed by the University of Sheffield. In addition, domain-specific name lexicons are embedded in Gate for identifying name entities (NE) in text. After coreference resolution, we replace each resolved pronoun with the origin term that it refers to. The text documents are then tagged and parsed by two NLP tools, namely the Eric Brill's rule-based part-of-

speech (POS) tagger [Bri92] and Collins parser [Col96]. After preprocessing, each parsed document contains a set of sentence grammar trees. Based on the sentence grammar trees, simplified conceptual graphs containing semantic relations are extracted and encoded by the following three modules.

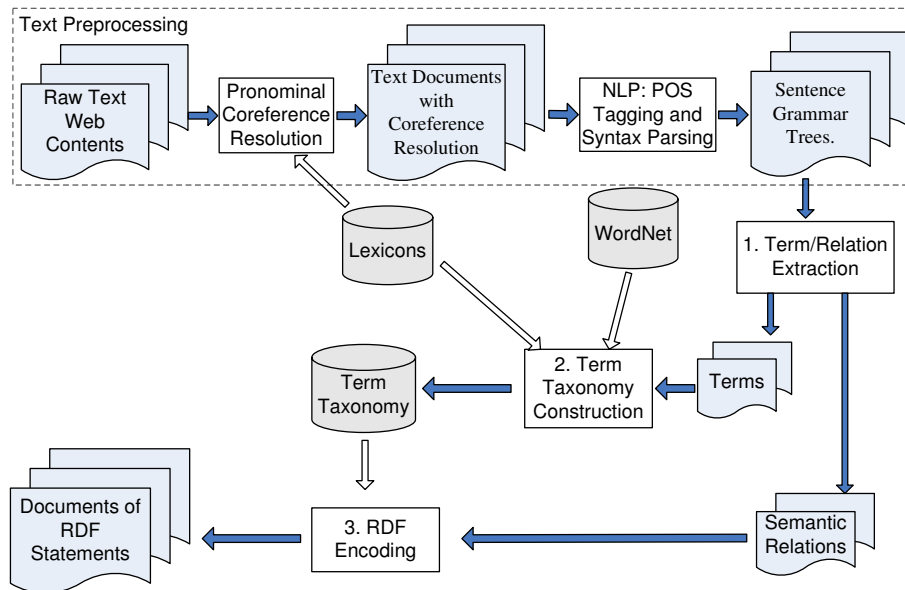


Figure 4.11: An overview of the procedure for semantic metadata extraction.

- (i) **Term and Relation Extraction:** Based on the preprocessing results, a set of predefined rules adopted from [Bar97] is used for extracting semantic relations from the sentence grammar trees. When extracting semantic relations, we first identify the important terms describing the major concepts, i.e., noun phrases (NP) and verb phrases (VP), in the grammar trees, followed by three major types of relations between these terms. The three relation types are introduced in Section 4.2.1.
- (ii) **Term Taxonomy Construction:** The terms (NP/VP) extracted from the sentence grammar trees are incrementally clustered into a term taxonomy with the assistance of WordNet [Mil95]. The atomic clusters in the term taxonomy are groups of synonyms. When a new term is inserted into the term taxonomy, we first

CHAPTER 4. SEMANTIC MULTIMEDIA CONTENT REPRESENTATION AND METADATA EXTRACTION

try to find whether there is a synonym group it can join. If there is such a synonym group, we simply insert the new term into the synonym group and the structure of term taxonomy will not change; otherwise, it will be inserted as a new synonym group and the structure of the term taxonomy will change.

- (iii) **RDF Encoding:** The term taxonomy and the semantic relations extracted from the sentence grammar trees are encoded as an RDF vocabulary and RDF statements respectively.

In the following sections, we describe the three modules in detail.

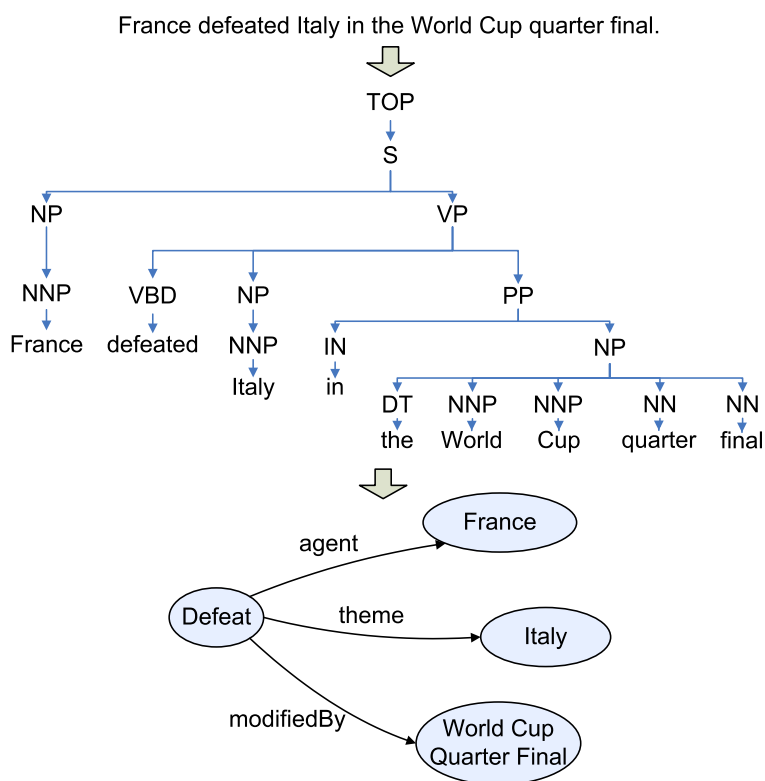


Figure 4.12: Conversion of text to semantic relations shown in the form of conceptual graph.

### 4.2.1 Extraction of Terms and Relations from Sentence Grammar Trees

As noun phrases (NP) and verb phrases (VP) convey the main meaning of text, we first extract those terms of NP or VP from the sentence grammar trees. Then, three types of relations between those terms (NP/VP) are identified based on their syntactic dependencies using a set of rules borrowed from [Bar97]. Similar to Sowa's conceptual graphs [Sow84] [Sow99] as used in many knowledge based systems [GMV99] [yGGLL02], the definitions of the three types of relations are given below.

- $\langle A, \text{agent}, B \rangle$ , where  $A$  can be a VP and  $B$  can be an NP/VP. The relation indicates that  $B$  is the *agent* that performs the *action*  $A$ .
- $\langle A, \text{theme}, B \rangle$ , where  $A$  can be a VP and  $B$  can be an NP/VP. The relation indicates that  $B$  is the *theme* (i.e., *recipient*, *object*, or *target*) of the action  $A$ .
- $\langle A, \text{modifiedBy}, B \rangle$ , where  $A$  can be an NP/VP and  $B$  can be an NP/VP. This relation indicates that  $A$  is modified by  $B$  through a proposition.

Figure 4.12 shows an example. A sentence is first parsed into a grammar tree structure. We then extract three relations, i.e.,  $\langle \text{Defeat}, \text{agent}, \text{France} \rangle$ ,  $\langle \text{Defeat}, \text{theme}, \text{Italy} \rangle$ , and  $\langle \text{Defeat}, \text{modifiedBy}, \text{World Cup Quarter Final} \rangle$ , based on the obtained grammar tree. These three relations form a simplified conceptual graph as shown in Figure 4.12.

We note that other forms of expressions can be used for modelling relations. For example, from the sample grammar tree in Figure 4.12, we can also extract a relation  $\langle \text{France}, \text{defeat}, \text{Italy} \rangle$ , where the action, its agent, and its theme are combined into one relation. However, we observe that in many sentences (in particular those in passive voice), the agents or themes of an action may be missing. For example, the sentence

“France was defeated” does not contain the agent of the action. In this case, we cannot extract a full relation containing action, agent, and theme. However, using the three relation types that we adopt, useful relations can still be extracted, even though the agents or themes of the actions are missing in the sentences.

## 4.2.2 Term Taxonomy Construction

Our purpose of building a term taxonomy is to hierarchically group similar terms into meaningful clusters. Based on such clusters, semantic relations that consist of similar terms can be generalized for deducing statistically significant patterns during the knowledge mining stage.

### 4.2.2.1 Existing work in term taxonomy construction

Recently, there has been an increasing amount of attention on automatic taxonomy construction in the field of ontology engineering [Hea92] [MPS02] [CHS04] [Car99]. Existing methods for taxonomy construction mainly fall into two categories: symbolic approaches and statistics-based methods, and both have their limitations. Symbolic approaches, which directly find taxonomic relations from text using lexico-syntactic patterns [Hea92], can hardly exhaustively extract all possible taxonomic relations, especially when some commonly-known domain-specific information, such as “goalkeeper is a kind of soccer player”, does not explicitly exist in the text documents. Statistics-based methods usually employ bottom-up (agglomerative) or top-down (divisive) hierarchical clustering methods to build term hierarchies based on statistical context features of terms (such as frequencies of surrounding terms) [Car99] [MPS02]. The disadvantages of these methods include the poor traceability of the taxonomy construction process and the difficulty in labelling non-leaf nodes (inner clusters) of the taxonomies. The extracted taxonomies are thus difficult for human users to understand. Furthermore, both symbolic and statistics-based approaches require a large domain-specific text corpus, which is usually unavailable, for

taxonomy construction. In addition, the costs of computation and update of taxonomies may be very large on such corpus. Besides the symbolic and statistics-based methods, a more recent work presented in [Hep06] focuses on deriving produces and services ontologies, including concept taxonomies, based on the existing industrial categorization standards. The work itself is interesting. However, as well-defined categorization standards do not widely exist, the reusability of the proposed method is limited. In our opinion, most existing techniques for taxonomy construction are more suitable to bootstrap an ontology acquisition process but have limited usages in data mining tasks such as the one presented in this paper. In subsection 4.2.2.3, we introduce a light-weight incremental clustering strategy for taxonomy construction. Instead of using a large text corpus, it utilizes the word sense hierarchies in WordNet [Mil95] as the basis for building the term taxonomy. The constructed taxonomy is thus more understandable for human users. In addition, because it constructs the taxonomy in an incremental manner, the computation and update costs are minimal.

#### 4.2.2.2 Term Representation and Similarity Measure Selection

A term (VP/NP) extracted from text is represented as a bag of senses in WordNet [Mil95] in the form of  $S = \{s_1, s_2, \dots, s_n\}$ , where each sense represents a meaning of a word and corresponds to a set of synonyms in WordNet. For each word in a VP/NP, we add all its WordNet senses into the bag representation. For each sense added in the bag, we recursively add their *hypernyms* and *derivationally related senses* into the bag. However, adding all senses of a term and their related senses (hypernyms and derivationally related senses) can generate a very large bag which will slow down the process of term taxonomy construction. Therefore, we impose a restriction on the WordNet search depth (*WNSD*) when building the bag of senses for a term. A set of sample terms represented by bags of senses is listed in Table 4.1. A WordNet sense (e.g. *10283858\_player*) is expressed using

its ID (e.g. *10283858*) in WordNet conjuncted with its representative word (e.g. *player*). If two terms have the same set of senses, they are called *synonyms*. The extracted NP/VPs thus can be classified into groups of synonyms.

Table 4.1: A set of sample terms represented by bags of WordNet senses.

Order	Terms	Bags of Senses
1	Midfield Player	$S_{Midfield\_Player} = \{08451381\_midfield, 08405214\_center, 10283858\_player, 09476765\_contestant, 10260253\_performer, \dots, 10246540\_participant\}$
2	Player	$S_{Player} = \{10283858\_player, 09476765\_contestant, 10260253\_performer, \dots, 10246540\_participant\}$
3	Attack Player	$S_{Attack\_Player} = \{00453042\_attack, 00452701\_play, 10283858\_player, 09476765\_contestant, 10260253\_performer, \dots, 10246540\_participant\}$
4	Winner	$S_{Winner} = \{09619712\_achiever, 09476765\_contestant, 10621652\_winner, 09968260\_gambler, 10621801\_victor, 00007626\_person\}$

For clustering the terms extracted from the text, we need a method to measure the semantic similarity between the terms. There have been many term similarity measures proposed in the existing literatures, such as Jiang and Conrath’s measure [JC97], Leacock’s measure [LC98], and Seco et al’s measure [SVH04]. Some measures uniquely rely on the topology information in concept taxonomies such as WordNet [LC98] [SVH04], whilst some others use both concept taxonomy and large text corpus for combining the topology information and word statistics [JC97]. Evaluations and comparisons of various term similarity measure are presented in [BH01] and [SVH04]. However, the above-mentioned similarity measures are seldom used in the existing work of term taxonomy construction. One reason can be that these similarity measures (in particular, those relying on the word statistics) are computationally intensive. Another reason may be that these measures usually do not take multi-word terms (e.g. “Attack Player” in Table 4.1) into account. For handling multi-word terms, an additional strategy must be used.

As we represent each term as a bag of senses, we calculate the term similarities based on the number of senses shared by two terms. However, our similarity measure is



intrinsically equivalent to the *cosine similarity measure* which is widely used and has achieved satisfactory results in many term taxonomy construction tasks [Car99] [MPS02].

The similarity measure is defined as

$$sim(t_1, t_2) = \frac{|S(t_1) \cap S(t_2)|}{\sqrt{|S(t_1)| \cdot |S(t_2)|}}, \quad (\text{Eq. 4.1})$$

where  $t_i$  denotes a term and  $S(t_i)$  denotes its corresponding bag of senses ( $i = 1$  or  $2$ ). This measure is equivalent to the cosine similarity measure if we convert the bags of senses into sense vectors where each sense is a feature dimension and the feature values of the senses are set to 1 or 0, depending on whether a sense is present in a bag of senses.

#### 4.2.2.3 Incremental Term Taxonomy Construction

In our work, the term taxonomy is dynamically built on the fly using an incremental hierarchical clustering strategy. Here, we treat each group of synonyms as an atomic cluster. Several clusters can be merged to form a larger cluster, which is treated as the parent (super-cluster) of the merged clusters. A cluster is also represented as a bag of WordNet senses, containing the senses shared by all terms in this cluster. We can thus use Eq. 4.1 to measure the similarity between two clusters. The root of the term taxonomy corresponds to a cluster containing all terms. The bag of senses associated with the *root cluster* is set to  $\emptyset$ .

When a new term (NP/VP) is extracted, we first try to find an existing atomic cluster (composed of its synonyms) to which it can be directly assigned. If there is no such atomic cluster, we create one for the new term and add it into the term taxonomy using our incremental hierarchical clustering strategy. The clustering process can be summarized in the following steps:

**Step 1** : We first find a cluster in the term taxonomy that is most similar to the new term<sup>1</sup>. If no similar cluster can be found (i.e., there is no existing cluster that has

<sup>1</sup>For simplicity, we here use “new term” to represent the atomic cluster created for the new term.

non-zero similarity with the new term), add the new term as a sub-cluster of the root cluster and the process is completed.

**Step 2** : The new term and its most similar cluster are merged to form a new cluster in three different ways:

- (i) If the sense bag of the new term is a superset of the sense bag of its most similar cluster (i.e., the new term is more specific), we merge the new term into its most similar cluster as a sub-cluster (Figure 4.13 (b)) and the process is completed.
- (ii) If the sense bag of the new term is a subset of the sense bag of its most similar cluster (i.e., the new term is more general), we merge its most similar cluster into the new term as a sub-cluster (Figure 4.13 (c)) and go to Step 1 to recursively insert the merged cluster (the expanded cluster of the new term);
- (iii) Else merge the new term with its most similar cluster to form a new cluster (Figure 4.13 (d)) and go to Step 1 to recursively insert the merged cluster.

In case (ii) and case (iii) of Step 2 described above, the merged cluster is to be recursively inserted into the taxonomy. According to Step 1, we need to find a most similar cluster for the merged cluster. We know that the merged cluster is generated from and thus similar to the most similar cluster  $c_{sim}$  of the new term. Intuitively, the merged cluster is very likely to be similar to a super-cluster of  $c_{sim}$ . Thus, a *local search* (i.e., searching the super-clusters of  $c_{sim}$ ) is adopted for locating the most similar cluster for the merged cluster.

In addition, we use another heuristic strategy to simplify the taxonomy structure. We observe that terms sharing little meaning may still be grouped into a cluster, e.g. terms “go” and “stop”. To avoid grouping such irrelevant terms into a cluster, we define

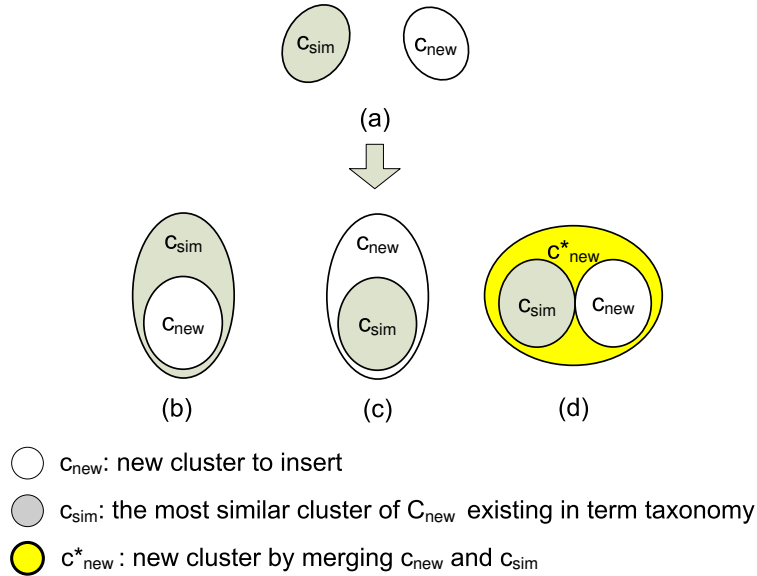


Figure 4.13: Three cases for merging the new cluster and its most similar cluster.

a *minimum similarity threshold* ( $mins_{im}$ ). If the similarity between a new cluster and its most similar cluster is below  $mins_{im}$ , we directly insert the new cluster as a sub-cluster of the root. For setting a proper  $mins_{im}$  value, we use a three-stage method in our practices. Firstly, we randomly select a small set of sample terms. For each pair of sample terms, we manually determine whether two terms are relevant and take the results as our ground truth. Secondly, the similarities for each pair of terms are calculated. Thirdly, we explore the term similarities and manually select a  $mins_{im}$  value that can best separate the irrelevant and relevant pairs of terms in the ground truth.

We illustrate the term taxonomy construction process using the sample terms listed in Table 4.1.

- As shown in Figure 4.14, when the first term “Midfield Player” (an atomic cluster) is inserted, the root cluster is the only cluster in the taxonomy. Therefore, we cannot find a most similar cluster for “Midfield Player”. Thus, “Midfield Player” is directly added as a sub-cluster of the root cluster.

## CHAPTER 4. SEMANTIC MULTIMEDIA CONTENT REPRESENTATION AND METADATA EXTRACTION

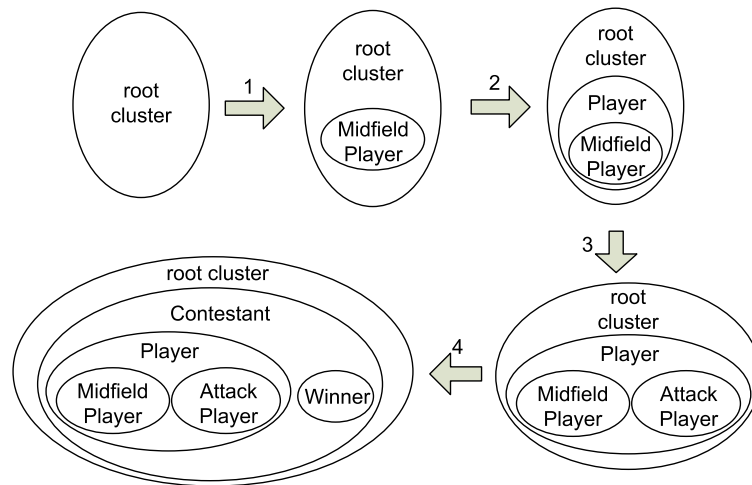


Figure 4.14: Term Taxonomy Construction Using the Sample Terms in Table 4.1.

- When the term “Player” is inserted, “Midfield Player” is found as the most similar cluster. As the sense bag of “Player” is a subset of the sense bag of “Midfield Player” (i.e., “Player” is more general), “Midfield Player” is merged into “Player” as a sub-cluster. Note that we need to recursively insert the expanded “Player” cluster into the term taxonomy. As the root cluster is the only super-cluster of “Midfield Player”, we cannot find a most similar cluster that has non-zero similarity to the expanded “Player” cluster. “Player” is thus added as a sub-cluster of the root cluster.
- When the term “Attack Player” is inserted, the “Player” cluster is identified as the most similar cluster. As “Attack Player” is more specific than “Player”, “Attack Player” is directly inserted as a sub-cluster of the cluster “Player”.
- Finally, when the term “Winner” is inserted, the most similar cluster “Player” is identified. “Player” and “Winner” are merged into a new cluster, labelled by “contestant” (as sense 09476765\_contestant appears in both sense bags of “Player” and “Winner”). We then recursively add “contestant” cluster into the term taxonomy.

As the root cluster is the only super-cluster of “Player”, we cannot find a cluster that has non-zero similarity to “contestant”. Therefore, the “contestant” cluster is inserted as a sub-cluster of the root.

### 4.2.3 RDF Encoding

The term taxonomy is encoded using RDFS as a part of our RDF vocabulary (a schema file) for describing semantic relations. Each term cluster in the taxonomy is mapped to an RDFS class. For any two clusters  $c_1$  and  $c_2$  where  $c_2$  is a sub-cluster of  $c_1$ , the RDFS class of  $c_2$  is defined as a subclass of the RDFS class of  $c_1$  using the “rdfs:subClassOf” predicate.

To encode semantic relations in RDF, the three predicates, i.e., *agent*, *theme*, and *modifiedBy*, are defined as *RDF predicates* (instances of *rdf:Property*) in our RDF vocabulary.<sup>2</sup> In addition, we treat each NP/VP extracted from text as an *RDF resource* with a Unified Resource Identifier (URI). As atomic clusters (synonym groups) are defined as RDFS classes, each NP/VP is thus defined as an instance of the RDF class corresponding to its synonym group.

## 4.3 Summary

In this chapter, we first present an approach, which combines conceptual graph based knowledge representation, RDF based semantic modelling, and MPEG-7 based multimedia information description for expressing multimedia contents in an integrated framework.

More importantly, we present a method for automatic extraction of multimedia semantic contents from Web texts. A myriad of natural language processing techniques

---

<sup>2</sup>Note that the technologies used in our work are not limited by the predefined types of relations. New types of relations can be included by expanding the set of rules for relation extraction. In addition, our generalized relation association mining algorithm can be applied on any RDF document collection with the existence of an RDF vocabulary.

CHAPTER 4. SEMANTIC MULTIMEDIA CONTENT REPRESENTATION AND METADATA EXTRACTION

---

is employed to parse the text sentences into grammar trees, based on which conceptual graphs composed of semantic concepts and relations are extracted. In addition, we propose an incremental clustering technique for organizing the extracted semantic concepts (terms) into a term taxonomy which forms a domain vocabulary. Both the conceptual graphs and the term taxonomies are encoded using RDF and RDF Schema. The extracted semi-structured RDF metadata serve as a foundation for applying data mining algorithms to discover interesting patterns. In the next chapter, a generalized association pattern mining algorithm, called GP-Close, is proposed for deducing frequently-occurred RDF relation patterns.

We should note that evaluating the overall semantic metadata extraction method is difficult. On one hand, there is a lack of a standard evaluation criteria. On the other hand, as the extraction approach is a combination of various techniques including the existing NLP methods and our proposed techniques, it is difficult to estimate the contributions of the different methods to the overall extraction results. Therefore, instead of provide an evaluation of the whole metadata extraction framework, we tend to evaluate a most important technique proposed for taxonomy construction, i.e., the proposed term similarity measure. The importance of the term similarity measure lies on that it determines the final structure of the term taxonomy which will further influence the RDF metadata mining results. The results of evaluation of the proposed term similarity measure are shown in Appendix B. We believe that though the performance of the whole semantic extraction framework is not assessable at this stage, it can be reflected by the experimental results of the RDF metadata mining, which show the quality of the knowledge discovered based on the extracted RDF metadata. In the next chapter, the experimental evaluation of the RDF metadata mining will be presented.

In stead of introducing a new knowledge representation and extraction framework, the work presented in this chapter mainly aims to employs and combines the strength of

#### CHAPTER 4. SEMANTIC MULTIMEDIA CONTENT REPRESENTATION AND METADATA EXTRACTION

---

the existing mature knowledge representation techniques for supporting data mining and knowledge discovery based on detailed representation of the multimedia data semantics. Therefore, the contribution of this chapter is more lying on the practice aspect.

## Chapter 5

# Mining RDF Semantic Metadata for Multimedia Analysis

While machine-processible RDF semantic metadata and domain ontologies (term taxonomy) can be automatically extracted for describing multimedia web contents, multimedia information analysis can be achieved by mining semi-structured RDF metadata for useful knowledge. In this chapter, we present a generalized association pattern mining technique for discovering useful knowledge based on RDF semantic metadata and concept taxonomies in domain ontologies.

### 5.1 Association Rule Mining: An Introduction

Association Rule Mining (ARM) [AIS93], since its introduction, has become one of the key data mining techniques in the field of Knowledge Discovery in Database (KDD). Given a set of items  $\mathcal{I}$  and a large database of transactions  $\mathcal{D}$ , where each transaction is a set of items  $T \subseteq \mathcal{I}$  with a unique identifier  $tid$ , an association rule is an implication of the form  $X \Rightarrow Y$ , where  $X, Y \subseteq \mathcal{I}$  (called itemsets or patterns) and  $X \cap Y = \emptyset$ . In the domain of market-basket analysis, such an association rule indicates that the customers who buy the set of items  $X$  are also likely to buy the set of items  $Y$ . An essential task of ARM is discovering frequent itemsets, i.e., itemsets whose *supports* (occurring frequencies) are greater than a user defined threshold (minimum support).



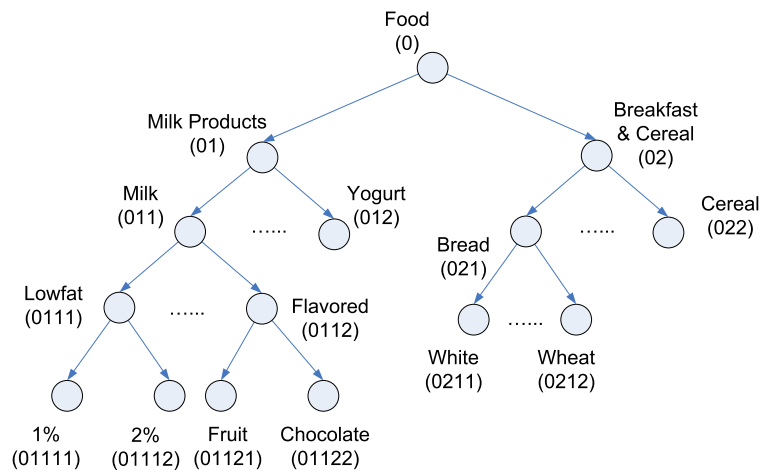


Figure 5.1: A sample market-basket taxonomy.

Motivated by the observation that ARM may generate too many rules that are not statistically significant, generalized association rule mining (GARM) has been proposed [SA95, HF95] to exploit item taxonomies for generalizing trivial rules into more statistically significant and informative rules. For example, the finding of “six percent of people buy **2% milk** if they buy **wheat bread**” may not be statistically significant. However, by utilizing the item taxonomy shown in Figure 5.1, it can be combined with another fact “four percent of people buy **1% milk** if they buy **wheat bread**” to form a statistically significant rule: “10 percent of people buy **low-fat milk** if they buy **wheat bread**”<sup>1</sup>.

Despite its initial focus on market-basket analysis, generalized association rule mining has become influential in many other application domains. In particular, with the rising popularity of ontologies and concept hierarchies, real-world data sets with predefined taxonomies (such as gene ontology, lexicon hierarchy, and web page subject hierarchy) have become available. Empowered by well-defined taxonomies, generalized association rule mining has been used in many emerging real-world applications, including text

<sup>1</sup>Here, an underlying assumption is that no person who buy wheat bread will buy 1% milk and 2% milk at the same time

mining [yGGLL02], web mining [Evv03], bio-informatics mining [KSB04], biomedical mining [BLLL05], multimedia mining [TWS05], and semantic web mining [MMST05]. In particular, on the Semantic Web, information is described using RDF metadata composed of domain-specific concepts, which are usually organized in concept taxonomies in RDF Schema based domain ontologies. Employing such domain ontologies, GARM can be applied to extract generalized association patterns of RDF statements. Similarly, in the semantic based multimedia information fusion and analysis framework, wherein RDF semantic metadata and RDF Schema based term taxonomies can be automatically extracted, GARM can be applied to extract useful patterns from the multimedia semantic contents.

Although GARM is originally designed to generalize trivial associations, it introduces another type of redundant patterns, called over-generalized patterns. A pattern is *over-generalized* if its *occurrence frequency* or *support* is the same as one of its specialized patterns. Referring back to our market-basket example, a pattern “10 percent of customers buy both **milk** and **wheat bread**” is over-generalized given its specialized pattern “the proportion of customers who buy both **low-fat milk** and **wheat bread** is also 10 percent”. With the same support, the semantic meaning conveyed by the specialized pattern is more precise than that conveyed by the over-generalized pattern. In fact, an over-generalized pattern can always be inferred from its specialized patterns, and therefore, over-generalized patterns are redundant. Moreover, combinations and extensions of over-generalized patterns are also over-generalized. This causes a serious pattern combinatorial explosion [Per98] which can dramatically increase pattern redundancy and the computation cost. Therefore, solutions for reducing over-generalized patterns are clearly needed.

In this chapter, we define the problem of *over-generalization reduction*, wherein we aim to generate all frequent generalized patterns WITHOUT generating over-generalized

patterns. By employing the notion of *generalization closures*, we translate the over-generalization reduction problem into a closed pattern mining problem [BTP<sup>+</sup>00, CCS00, PBTL99a, PHM00, ZH02] and present an algorithm, called *GP-Close* (Closed Generalized Pattern Mining), for mining frequent generalized patterns with over-generalization reduction. The contributions of this chapter are summarized as follows:

- (i) We introduce a mathematical foundation to translate the over-generalization reduction problem into a closed pattern mining problem.
- (ii) We present the GP-Close algorithm and an additional set of pattern pruning strategies for eliminating over-generalized patterns during the GARM process. In particular, the proposed pattern pruning strategies can efficiently minimize the combinatorial explosion problem of the over-generalized patterns which is not addressed by the existing algorithms.
- (iii) We provide strong experimental results to show that the GP-Close algorithm can substantially reduce pattern redundancy and perform much better than the existing generalized and closed itemset mining algorithms, namely *Cumulate* [SA95], *Prutax* [HMWG98], and *CHARM* [ZH02], in terms of time efficiency.

## 5.2 Association Rule Mining Algorithms

### 5.2.1 Generalized Association Mining Algorithms

Generalized association rule mining was first proposed in [SA95] and [HF95] to extract associations among items across various levels of taxonomies. Representative GARM algorithms include *Cumulate* [SA95] and *Prutax* [HMWG98], both of which use an Apriori-like [AS94] bottom-up and breadth-first search strategy for finding patterns. More recently, an algorithm called *SET* [ST02] is reported. However, SET adopts a

*taxonomy-based* and *join-based* strategy for itemset enumeration, which may cause a pattern duplication problem when the taxonomies are not tree-structured.

A common limitation of the above algorithms is that they generate all frequent generalized patterns including over-generalized ones. The over-generalization problem is first identified in [Ino04] when mining generalized substructure in connected graphs. Unfortunately, though [Ino04] adopts a set of pruning strategies to remove some over-generalized patterns, those pruning strategies do not guarantee that all over-generalized patterns will be eliminated.

## 5.2.2 Closed Itemsets Mining Algorithms

Mining closed itemsets, first proposed in [PBTL99b], aims to discover the itemsets that do not have proper supersets with the same support. According to [Zak00], mining closed itemsets can generate a much smaller result set without losing any information conveyed by all frequent itemsets.

Well-known algorithms for mining closed itemsets include A-Close [PBTL99b], CLOSET [PHM00], CLOSET+ [WHP03] and CHARM [ZH02]. A-Close first uses Apriori-like bottom-up traversal of itemset lattice to search the closed itemset generators, each of which is the smallest subset of a closed itemset with the same support. Then a closed itemset can be computed based on its generator by intersecting all the transactions in which the generator appears. CLOSET and CLOSET+ use a frequent pattern tree (FP-tree) structure to compress the transaction information and employ a recursive divide-and-conquer strategy to mine long frequent patterns. CHARM adopts depth-first search and tidset-based counting strategies to find patterns in an itemset-tidset tree (IT-tree). Each branch of the IT tree can be seen as an independent partition of the original pattern search space. When mining long closed itemsets, traversing IT tree shrinks the search space quickly.

In this chapter, we prove that the over-generalization reduction problem in GARM can be translated into a closed pattern mining problem. However, the existing closed itemset mining algorithms are not designed for generalized association rule mining. As they do not use taxonomic information for pattern pruning, the combinatorial explosion problem of the over-generalized patterns is not addressed. Therefore, the efficiencies of these algorithms may not be satisfactory. Based on such considerations, our GP-Close algorithm is designed, which adopts a set of additional pattern pruning strategies for pruning the pattern search space and preventing the combinatorial explosion problem.

## 5.3 Over-Generalization Reduction

### 5.3.1 Over-Generalization Problem

We illustrate the over-generalization problem using an example as follows. Given a sample market-basket taxonomy  $\mathcal{T}$  (Figure 5.1) and a transaction database  $\mathcal{D}$  (Table 5.1), the set of all frequent generalized itemsets with a minimum support of 40% is given in Table 5.2. Among those frequent itemsets, we would like to highlight two patterns, i.e., {Milk, Bread} and {Milk, Breakfast & Cereal}. As “Breakfast & Cereal” is a generalized item of “Bread” in Figure 5.1, {Milk, Breakfast & Cereal} is a generalized pattern of {Milk, Bread} and they have the same support of 60% (see also Figure 5.2). Intuitively, with the same support, a specialized pattern should be more interesting than its generalized pattern as the information conveyed by the specialized pattern is more precise than that conveyed by the generalized one. Therefore, the second pattern is considered as redundant and not useful. Based on the above observation, we define generalized patterns and the over-generalization problem formally as follows.

**Definition 5.1** *Given a set of items  $\mathcal{I}$ , a taxonomy  $\mathcal{T}$ , and two itemsets (patterns)  $X, Y \subseteq \mathcal{I}$ ,  $X$  is called a **generalized itemset (pattern)** of  $Y$  and  $Y$  is called a **specialized itemset (pattern)** of  $X$ , if and only if: (1)  $X \neq Y$ , (2)  $\forall i^* \in X, \exists i \in Y$  such*

Table 5.1: A sample transaction database.

TID	Transactions
1	Fruit Milk
2	1% Milk, White Bread
3	Wheat Bread
4	2% Milk, White Bread
5	Chocolate Milk, Wheat Bread

Table 5.2: Frequent generalized itemsets in sample database ( $minsup=40\%$ ).

Support	Frequent Generalized Itemsets ( $minsup=40\%$ )
40%	{Lowfat Milk} {Flavored Milk} {White Bread} {Wheat Bread} {Lowfat Milk, White Bread}
60%	{Milk, Bread} {Milk, Breakfast and Cereal} {Milk Products, Bread} {Milk Products, Breakfast and Cereal}
80%	{Milk} {Bread} {Milk Products} {Breakfast and Cereal}
100%	{Food}

that  $i^* = i$  or  $i^*$  is an ancestor of  $i$  in  $\mathcal{T}$ , and (3)  $\forall i \in Y, \exists i^* \in X$  such that  $i^* = i$  or  $i^*$  is an ancestor of  $i$  in  $\mathcal{T}$ .

A generalized itemset can be seen as a result of replacing some items in an itemset with their ancestors. Likewise, a specialized itemset can be created by replacing some items in an itemset with their descendants.

**Definition 5.2** An itemset  $X$  is **over-generalized** if there exists a specialized itemset  $Y$  of  $X$  with  $supp(X) = supp(Y)$ .  $X$  is also called an **over-generalization** of  $Y$ . If an itemset  $X$  is not over-generalized, we say  $X$  is **generalization-complete**.

Based on the above definitions, the task of over-generation reduction can be stated as follows.

**Problem Statement:** *Over-generalization reduction* is to avoid generating over-generalized itemsets during the generalized association rule mining process.

For example, in Table 5.2, five frequent patterns highlighted by underline (i.e., {Milk, Breakfast and Cereal}, {Milk Products, Bread}, {Milk Products, Breakfast and Cereal}, {Milk Products}, and {Breakfast and Cereal}) are *over-generalized* as they have specialized itemsets with the same supports. It means that almost 36% (5 out of 14) of the frequent patterns are not useful. In some real world data sets, such as RDF data sets (see Section 5.6), the proportion of over-generalized patterns may be much higher. This motivates us to develop a new method for over-generalization reduction.

Itemset X (supp=60%) : {Milk, Breakfast & Cereal}
Itemset Y (supp=60%) : {Milk, Bread}

Figure 5.2: An illustration of over-generalization. The itemset  $X$  is an over-generalization of the itemset  $Y$ .

Generalization Closure of X (supp=60%) : {Milk, Milk Product, Breakfast & Cereal, Food}
Generalization Closure of Y (supp=60%) : {Milk, Bread, Milk Product, Breakfast & Cereal, Food}

Figure 5.3: The generalization closures of the itemsets  $X$  and  $Y$  in Figure 5.2.

### 5.3.2 Over-Generalization Reduction

In this section, we introduce a method for over-generalization reduction based on the notion of generalization closures. Informally, the generalization closure of an itemset  $X$ , denoted as  $gc(X)$ , is a set of items containing all items in  $X$  and all their ancestors. In fact,  $gc(X)$  can be seen as the union of  $X$  and all generalized patterns of  $X$ . As an example, Figure 5.3 shows the generalization closures of the itemsets  $X$  and  $Y$  in Figure 5.2. We know that a transaction which supports  $X$  must support its generalized patterns [SA95] and thus support  $gc(X)$ . On the other hand, as  $X \subseteq gc(X)$ , a transaction

which supports  $gc(X)$  must also support  $X$ . Therefore,  $X$  and  $gc(X)$  have the same support, i.e.,  $supp(X) = supp(gc(X))$ . The formal definition of generalization closure is given below.

**Definition 5.3** Given a set of items  $\mathcal{I}$  and a taxonomy  $\mathcal{T}$  (a directed acyclic graph), we define a function  $gc$  on  $2^{\mathcal{I}}$ :  $gc(X) = \{i \mid i \in X \text{ or } \exists i' \in X \text{ where } i \text{ is an ancestor of } i' \text{ in } \mathcal{T}\}$ , where  $X$  is an itemset and  $X \subset \mathcal{I}$ .  $gc(X)$  is called the **generalization closure**<sup>2</sup> of  $X$ .

Based on the above definition and referring to Figure 5.2 and Figure 5.3, we can see that if an itemset  $X$  is a generalized itemset of  $Y$ ,  $gc(X)$  must be a proper subset of  $gc(Y)$ , i.e.,  $gc(X) \subset gc(Y)$ . Therefore, if an itemset  $X$  is an over-generalization of another itemset  $Y$ ,  $gc(X) \subset gc(Y)$  and  $supp(gc(X)) = supp(gc(Y))$  holds, i.e.,  $gc(X)$  is not *closed* [PBTL99a, ZH02]. The definition of closed itemsets is given below.

**Definition 5.4** An itemset (pattern) is **closed** if it does not have a proper superset having the same support. Analogically, we say a generalization closure  $gc(X)$  of an itemset  $X$  is **closed** if there does not exist an itemset  $Y$  such that  $gc(X) \subset gc(Y)$  and  $supp(gc(X)) = supp(gc(Y))$ .

Based on the above observation, we see that if we extract only the itemsets whose generalization closures are closed, all over-generalized itemsets can be eliminated. This motivates us to mine closed generalization closures for over-generalization reduction.

In doing so, we also need to ensure that mining closed generalization closures will not remove potentially useful patterns besides the over-generalized patterns. This is guaranteed by the following theorem which shows that mining closed generalization closures will remove only the non-closed or over-generalized (i.e., not generalization-complete) itemsets. In other words, all removed patterns are not useful.

---

<sup>2</sup> $gc$  is in fact a special form of *closure operator*. We call it **generalization closure operator**. All generalization closures form a *closure system*. More information about closure system and closure operator can be found in [GW97].



**Theorem 5.1** *Given a frequent itemset  $X$ ,  $X$  is closed and generalization-complete, if and only if  $gc(X)$  is a closed generalization closure.*

**Proof:** By contradiction:

- (i) Given  $X$  is closed and generalization-complete, we suppose  $gc(X)$  is not a closed generalization closure. It follows that there exists an itemset  $Y$ , where  $X \subseteq gc(X) \subset gc(Y)$  and  $supp(gc(Y)) = supp(gc(X))$ . As  $gc(X) \subset gc(Y)$ , one of the following statements holds: (1)  $Y$  is a specialized itemset of  $X$  or (2)  $\exists i \in Y, \nexists i^* \in X$  such that  $i^*$  is an ancestor of  $i$  (i.e.,  $Y$  is not a specialized itemset of  $X$ ). If (1) holds, as  $supp(Y) = supp(X) = supp(gc(Y)) = supp(gc(X))$ ,  $X$  is over-generalized. If (2) holds, let  $Z = X \cup \{i\}$ . As  $X \subset Z \subseteq gc(Y)$  and  $supp(X) = supp(gc(Y))$ , we can see  $supp(Z) = supp(X) = supp(gc(Y))$ . Therefore,  $X$  is not closed. This contradicts with the statement that  $X$  is closed and generalization-complete.
- (ii) Given that  $gc(X)$  is a closed closure, we suppose  $X$  is not closed or  $X$  is over-generalized. It follows that  $\exists Y$ , such that (1)  $supp(Y) = supp(X)$  and (2)  $X \subset Y$  (if  $X$  is not closed) or  $Y$  is a specialization of  $X$  (if  $X$  is over-generalized). It follows that  $gc(X) \subset gc(Y)$  and  $supp(gc(X)) = supp(gc(Y))$ . This contradicts with the statement that  $gc(X)$  is a closed closure.

[End of Proof]

## 5.4 GP-Close Algorithm

Based on the discussions in the last section, we see that extracting the closed generalization closures will remove all over-generalized patterns and thus achieve over-generalization reduction. In this section, we present an algorithm, called GP-Close (Closed Generalized Pattern Mining), that discovers the set of closed generalization closures.

Referring to Definition 5.4, we can see that closed generalization closures can be seen as a special kind of closed itemsets. Therefore, the existing algorithms for closed pattern mining [PBTL99a, ZH02] could be used for mining closed generalization closures based on an expanded database in which a transaction is enriched by adding in the ancestors of the original items. However, as the existing algorithms are not designed for generalized pattern mining, they do not utilize the information in a taxonomy for pattern pruning during the mining process. In particular, they do not provide efficient strategies for reducing combinatorial explosion of the over-generalized patterns. For example, referring to the example shown in Figure 5.2 and Figure 5.3, once we know that “{Milk, Breakfast & Cereal}” is over-generalized, combinations of this pattern and other patterns must also be over-generalized and can be removed. This can large prevent combinatorial explosion of the patterns. However, traditional closed pattern mining algorithms do not consider such taxonomy based strategy for pattern pruning. In view of the limitations of the existing methods, our GP-Close algorithm adopts an additional set of pattern pruning strategies. Specifically, we utilize taxonomic information for reducing the pattern search space and minimizing the combinatorial explosion problem of the over-generalized patterns. Experimental results in Section 5.6 show that our algorithm can significantly improve the mining efficiency and outperform the state of the art closed pattern mining algorithm, CHARM [ZH02], in terms of the time efficiency.

In the rest of this section, we first present an overview of the GP-Close algorithm that follows a closed pattern mining paradigm based on *depth-first search* (DFS). Then, we introduce our strategies for pattern pruning and support counting.

### 5.4.1 Algorithm Overview

Seeing that given two generalization closures  $gc(X)$  and  $gc(Y)$ ,  $gc(X) \cup gc(Y)$  is also a generalization closure, i.e.,  $gc(X \cup Y)$  (denoted as  $gc(XY)$  for simplicity), our GP-Close

CHAPTER 5. MINING RDF SEMANTIC METADATA FOR MULTIMEDIA ANALYSIS

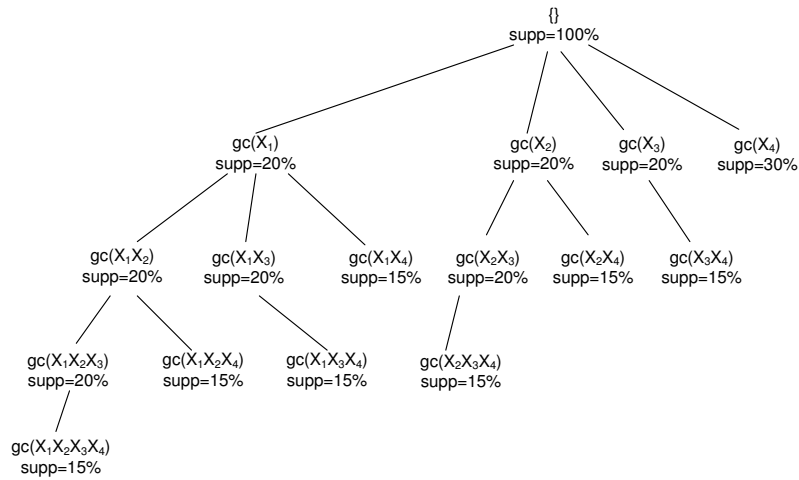


Figure 5.4: A full closure enumeration tree.

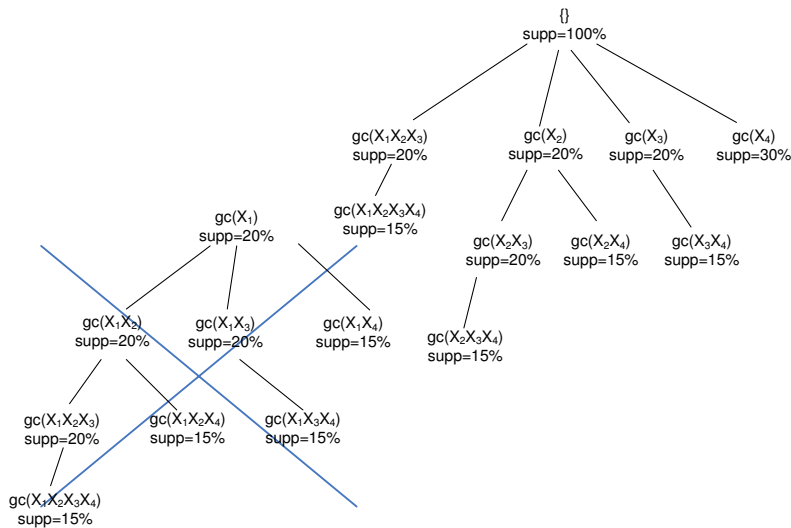


Figure 5.5: The closure enumeration tree after pruning the non-closed closure  $gc(X_1)$ .

algorithm employs a closure enumeration method to traverse the generalization closure space in a depth-first search (DFS) manner. Beginning with the generalization closures of 1-frequent itemsets (frequent items), the algorithm gradually enumerates larger closures by merging smaller ones. Figure 5.4 shows a full closure enumeration (search) tree based on four 1-frequent itemsets ( $X_1$ ,  $X_2$ ,  $X_3$ , and  $X_4$ ).

Intuitively, mining closed generalization closures is to find the largest generalization

**Algorithm 1** GP-Close*Input:*Transaction database:  $\mathcal{D}$ A taxonomy:  $\mathcal{T}$  over the set of items  $\mathcal{I}$ Support Threshold:  $minsup$ *Output:*The set of all closed frequent generalization closures:  $\mathcal{C}$ 

- 1:  $root = \emptyset$ ;  $root.support = 1$  //Initialize closure enumeration tree
- 2:  $root.children = \{gc(\{i\}) | i \in \mathcal{I} \wedge support(i) \geq minsup\}$  //Construct closures of 1-frequent itemsets (by looking up  $\mathcal{T}$ )
- 3:  $Sort(root.children)$  //Sort child-closures in a specialization-first (length-decreasing/support-increasing) manner
- 4: Closure-Enumeration( $root, \mathcal{C} = \emptyset$ )
- 5: return  $\mathcal{C}$

closure with the same support, i.e., any expansion of this pattern will have a lower support. For example, in Figure 5.4,  $gc(X_1X_2X_3)$  with the support of 20% is a closed generalization closure because any expansion of  $gc(X_1X_2X_3)$  will have a support lower than 20%. On the other hand,  $gc(X_1)$  with the support of 20% is not a closed generalization closure and should be pruned because its expansion  $gc(X_1X_2)$  has the same support of 20%.

In the closure enumeration tree, each node is a unique generalization closure. The children or descendants of the node are its expansions, i.e., proper supersets. We can see that if a closure and its children have the same support, this closure is not closed and thus can be pruned. We prune such a non-closed closure by replacing it with the union of its *equal-support children* (i.e., its child-closures that have the same support). For example, in Figure 5.4,  $gc(X_1)$  and its two child-closures  $gc(X_1X_2)$  and  $gc(X_1X_3)$  have the same support of 20%, so  $gc(X_1)$  is replaced by  $gc(X_1X_2X_3)$ , the union of  $gc(X_1X_2)$  and  $gc(X_1X_3)$  (see Figure 5.5).

For a non-closed generalization closure  $gc(X)$ , the union of its equal-support children is the largest expansion of  $gc(X)$  that can preserve  $gc(X)$ 's support in the subtree rooted

**Algorithm 2** Closure-Enumeration*Input:*A node in the closure enumeration tree :  $n$ A set of discovered frequent closed generalization closures:  $\mathcal{C}$ *Output:*The expanded set of frequent closed generalization closures:  $\mathcal{C}$ 

- 1: If  $\exists c^* \in \mathcal{C}$  such that  $c^*$  subsumes  $n$ , return  $\mathcal{C}$ . //Subtree Pruning
- 2: Children-Prune( $n.children$ ). //Child-Closure Pruning
- 3:  $c = \text{Closed-Closure}(n)$  //Derive the closed generalization closure based on  $n$ .
- 4:  $\mathcal{C} = \mathcal{C} \cup \{c\}$  //Insert  $c$  into the frequent closed generalization closure set.
- 5: **for** each child-closure  $child_i$  of  $n$ ,  $child_i \in n.children$  **do**
- 6:   Initialize the subtree of  $child_i$ : generate child-closures, perform support counting, and remove infrequent child-closures.
- 7:   Closure-Enumeration( $child_i, \mathcal{C}$ )
- 8: **end for**
- 9: return  $\mathcal{C}$

at  $gc(X)$ . Therefore, this union is *locally closed* in the subtree. For example, in Figure 5.4,  $gc(X_1X_2X_3)$  is the largest expansion of  $gc(X_1)$  that maintains the support of 20% in the subtree with the root  $gc(X_1)$ . It is thus locally closed. If a node in the closure enumeration tree does not have any equal-support child, the node itself is locally closed (see  $gc(X_2X_3)$  or  $gc(X_3)$  in Figure 5.4 for an example).

The locally closed closures are candidates to be *globally closed*. To determine whether a locally closed closure is globally closed, we need to examine whether there is a discovered (globally) closed closure containing it and having the same support (i.e., *subsuming*<sup>3</sup> it). If there is no such discovered closed closure, it is globally closed. Therefore, by recursively traversing the closure enumeration tree, the entire set of the closed generalization closures can be discovered.

The key steps for closure enumeration tree construction and closed closure discovery are listed in Algorithm 1 and 2. In Algorithm 1, the closure enumeration tree is first

<sup>3</sup>For simplicity, in rest of this chapter, we say a generalization closure  $gc(X_1)$  can *subsume* another generalization closure  $gc(X_2)$  if  $gc(X_2) \subset gc(X_1)$  and  $supp(gc(X_1)) = supp(gc(X_2))$ .

initialized (line 1-2). The tree initialization includes constructing the tree root (an empty closure with the support of 100%) and the child-closures of the root (closures of 1-frequent itemsets). Using the initialized closure enumeration tree as the input, Algorithm 2 is called in Algorithm 1 (line 4) to recursively access and build the enumeration tree nodes and discover closed generalization closures.

In Algorithm 2, for each tree node  $n$  visited, we first perform *subtree pruning* (line 1) to see whether the current sub enumeration tree rooted at  $n$  can be pruned. It is followed by *child-closure pruning* (line 2) to remove redundant child-closures (pruning tree branches). These two pruning strategies are efficient at prevention of the pattern combinatory explorations (see next section for a detailed discussion). Then based on  $n$  and its child-closures, we generate a closed generalization closure (line 3-4). Finally, for each child-closure  $child_i$  of  $n$ , we initialize a sub-enumeration tree rooted at  $child_i$  and recursively access the generated subtree (line 5-8). Initializing the subtrees includes (1) generating the child-closures of  $child_i$  by merging  $child_i$  with other child-closures of  $n$ , (2) counting the support of the new generated closures and (3) removing infrequent child-closures of  $child_i$ . For example, in Figure 5.4, suppose that the tree root ( $\{\}$ ) is the currently visited tree node  $n$ . If we want to initialize the subtree of  $gc(X_1)$ , we will generate the child-closures of  $gc(X_1)$  by merging  $gc(X_1)$  with  $gc(X_2)$ ,  $gc(X_3)$ , and  $gc(X_4)$ . As a result, the nodes  $gc(X_1X_2)$ ,  $gc(X_1X_3)$ , and  $gc(X_1X_4)$  will be created.

In the following subsections, we introduce our pattern pruning and support-counting strategies in detail.

### 5.4.2 Pruning Child-Closures and Subtrees

Besides the closed pattern pruning strategy, we further employ two pruning techniques to reduce the pattern search space.

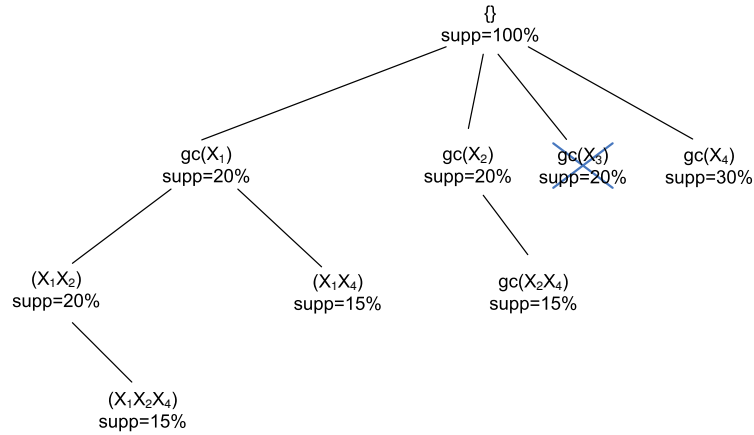


Figure 5.6: Child-closure pruning.

#### 5.4.2.1 Child-Closure Pruning

We note that a full closure enumeration tree, whose root has  $n$  child-closures, has a total of  $2^n$  nodes. Thus, pruning one child-closure of the tree root will reduce *half of the pattern search space*. For example, as the tree in Figure 5.4 has four child-closures under the root, the total number of nodes in the tree is 16 ( $2^4$ ). Suppose  $X_3$  is a generalized itemset of  $X_2$ . This implies  $gc(X_3) \subset gc(X_2)$ . As  $supp(X_3) = supp(X_2) = 20\%$ , we conclude that  $X_3$  is an over-generalization of  $X_2$ . Therefore,  $gc(X_3)$  is subsumed by  $gc(X_2)$ , i.e., any pattern containing  $gc(X_3)$  must also contain  $gc(X_2)$ . Therefore,  $gc(X_3)$  can be pruned as shown in Figure 5.6. Upon pruning of  $gc(X_3)$ , half of the tree nodes are gone. Therefore, early removal of redundant child-closures of a (sub) closure enumeration tree is very efficient for minimizing the pattern combinatorial explosions and thus highly desirable. In Algorithm 2 line 2, the function *Children-Prune* prunes the redundant child-closures of the current tree node visited. We call this pruning technique *child-closure pruning*. This is a taxonomy based pruning technique which is very efficient for preventing combinatorial explosion problem and is not yet used in other closed pattern mining approaches.

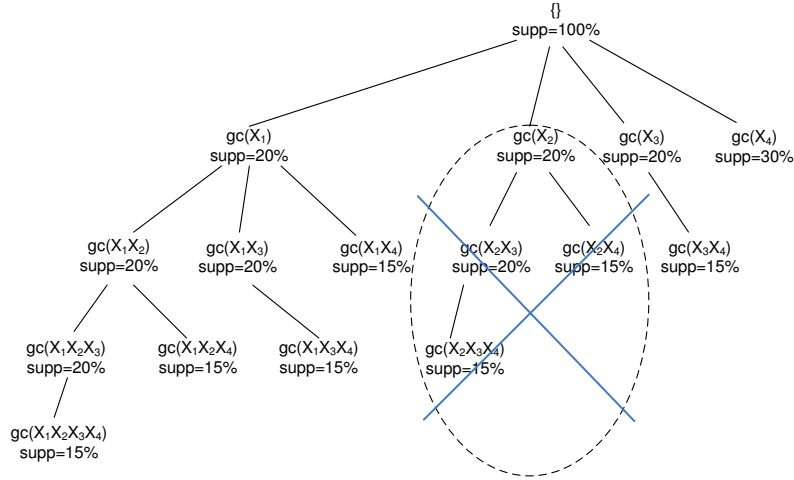


Figure 5.7: Subtree pruning.

#### 5.4.2.2 Subtree Pruning

In addition, we note that all descendants of an enumeration tree node are its expansions (proper supersets). If the tree node can be subsumed by an identified closed closure (i.e., the tree node is not closed), traversing the (sub) enumeration tree rooted at this node cannot generate new closed generalization closures. Therefore, the entire (sub) tree can be pruned (Algorithm 2 line 1). For example, in Figure 5.7, the closure  $gc(X_2)$  is subsumed by the closed closure  $gc(X_1X_2X_3)$  and all its descendants can be subsumed by the descendant of  $gc(X_1X_2X_3)$ . Thus, the subtree with the root  $gc(X_2)$  can be pruned. Pruning sub enumeration tree from the root is important to avoid a small redundant pattern to be expanded to form larger redundant patterns and involve further redundancies. In addition, we can see that if a tree node  $n$  cannot be subsumed by any identified closed closure, the locally closed closure generated based on  $n$  also cannot be subsumed by any identified closed closure, i.e., it is globally closed. This is because if there is an identified closed closure that can subsume the locally closed closure which is a superset of  $n$ , it must also subsume  $n$ . It contradicts with the statement that  $n$  cannot be subsumed.

In general, there are three cases in which one closure can subsume another, as listed



below.

- (i) A closure can be subsumed by one of its specialized closures if they have the same support.
- (ii) A closure can be subsumed by one of its *super closures* if they have the same support.
- (iii) A closure can be subsumed by a *superset of its specialized closures* if they have the same support.

Based on the above observations, a support-increasing and length-decreasing strategy is adopted for dynamic closure sorting (Algorithm 1 line 3). This sorting method implies that specialized closures will be enumerated first. This increases the occurrences of subsumption case 1 and case 3 so that there is a higher probability that a subtree can be pruned. In this way, the tree traversing space will be further reduced.

### 5.4.3 Support Counting

We calculate the support of generalization closures based on the *transaction ID sets* (*tidsets*). A tidset of an itemset is a set of IDs of transactions that contain (support) that itemset. For example, given the sample database in Table 5.1, the tidset of 1-itemset {0211} is {2, 4} as item 0211 appears in transactions 2 and 4. This type of support counting is widely used by many association rule mining algorithms, as in [SON95] and [ZH02].

Using tidsets for support counting is straightforward. The support of an itemset is given by the cardinality of its tidset divided by  $|\mathcal{D}|$  (the number of transactions in the database  $\mathcal{D}$ ). Given the tidsets of two itemsets  $X$  and  $Y$ ,  $supp(X \cup Y)$  can be calculated by the cardinality of the intersection of the tidsets divided by  $|\mathcal{D}|$ . Usually, the tidsets are built during the first database scan (DB scan) for counting the support of 1-itemsets (single items). The tidsets built can be subsequently used for counting the support of larger itemsets.

### 5.4.4 Complexity Analysis

The time complexity of GP-Close is as follows.

**Theorem 5.2** *The running time of GP-Close is  $O(|C| \cdot (l_{gc} \cdot \log|C| + l_{tidset}))$ , where  $l_{gc}$  is the average length of the generalization closures,  $l_{tidset}$  is the average length of tidsets, and  $|C|$  is the number of frequent closed generalization closures.*

**Proof:** Note that when traversing the generalization closure enumeration tree, we only visit those nodes based on which a closed generalization closure will be generated. This is guaranteed by our subtree pruning strategy. Therefore, GP-Close performs  $O(|C|)$  full tree node accesses. Each full access of a tree node  $n$  consists of three major operations, i.e., subtree pruning (Algorithm 2 line 1), child-closure pruning (Algorithm 2 line 2), and generating the child-closures for  $n$ 's children (involving support counting) (Algorithm 2 line 6).

Subtree pruning is to check whether  $n$  can be subsumed by an identified closed closure, i.e., whether there is an identified closed closure that has the same support with  $n$  and also contains  $n$ . The cost of finding the (hashed) closed closures having the same support is  $O(\log|C|)$ . The cost of checking whether a closed closure contains  $n$  is  $O(l_{gc})$ . Therefore, the total cost of subtree pruning is  $O(l_{gc} \cdot \log|C|)$ .

Note that in a closure enumeration tree, each non-leaf tree node on average has 2 child-closures. Child-closure pruning on a tree node  $n$  involves checking whether a child-closure can be subsumed by another child-closure. Referring to the analysis of subsumption checking for subtree pruning, we can see that the total cost of child-closure pruning is  $O(l_{gc} \cdot \log 2 \cdot 2)$  or  $O(l_{gc})$ .

Finally, generating the child-closures for  $n$ 's children means to merge any pair of  $n$ 's children to create larger closures. As  $n$  typically has two children, the algorithm on average generates only one child-closure for  $n$ 's children. The main cost here is counting

the support of the new closure by intersecting their tidsets. The cost of intersecting two tidsets is  $O(l_{tidset})$ .

Based on the above analysis, we can see that the overall computation cost of GP-Close is  $O(|C| \cdot (l_{gc} \cdot \log|C| + l_{gc} + l_{tidset})) \approx O(|C| \cdot (l_{gc} \cdot \log|C| + l_{tidset}))$ . When the document set is large (i.e.,  $|D|$  and  $l_{tidset}$  is large), the running time of GP-Close is approximately  $O(|C| \cdot l_{tidset})$ . [End of Proof]

## 5.5 Experiments

Our experiments are performed on a desktop PC with a P4-2.6G CPU and 1G RAM. The GP-Close algorithm is implemented using Java (JDK 1.4.2). We also implement two GARM algorithms, namely Cumulate [SA95] and Prutax [HMWG98], and a closed pattern mining algorithm, CHARM [ZH02], as the reference for performance evaluation and comparison. To apply CHARM for discovering closed generalization closures, we expand the transactions in the original data sets by inserting the ancestors of the transaction items.

Two types of data sets are used for the experiments. We first apply the algorithms on a suit of synthetic data sets which have been used for simulating market-basket transactional databases and evaluating GARM algorithms for the market-basket analysis. In addition, we further adopt two Semantic Web data sets to estimate and compare the efficiencies of the algorithms on the real world Semantic Web mining applications.

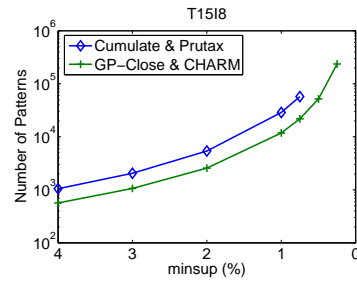
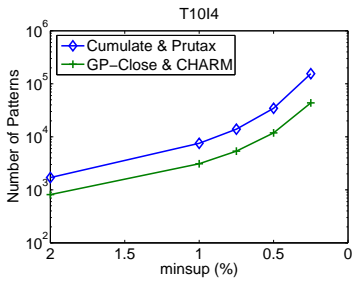
### 5.5.1 Synthetic Data Sets

The synthetic data sets are generated by a synthetic data generator provided by IBM Almaden Research Center (<http://www.almaden.ibm.com/software/projects/hdb/resources.shtml>), which is widely used in association rule mining and generalized association rule mining research [AS94, HMWG98, SA95, ST02]. The benefit of using synthetic data sets is the

CHAPTER 5. MINING RDF SEMANTIC METADATA FOR MULTIMEDIA ANALYSIS

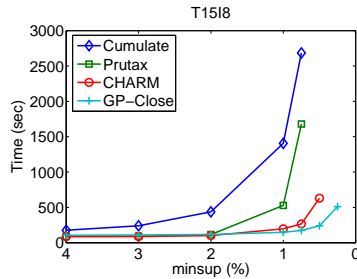
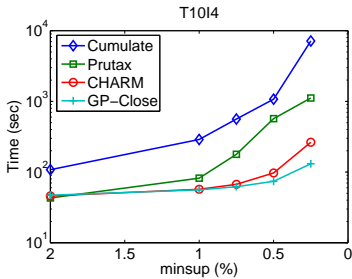
Table 5.3: The characteristics of the synthetic data sets.

Settings	Comments	T10I4	T15I8
$ \mathcal{D} $	Number of transactions	1,000,000	1,000,000
$ \mathcal{T} $	Average size of transactions	10	15
$I$	Average size of patterns	4	8
$ \mathcal{I} $	Number of different items	100,000	100,000
$ \mathcal{P} $	Number of seed patterns	10,000	10,000
$ \mathcal{R} $	Number of taxonomy roots	250	250
$F$	Fanout	5	5
$L$	Number of taxonomy levels	5-6	5-6



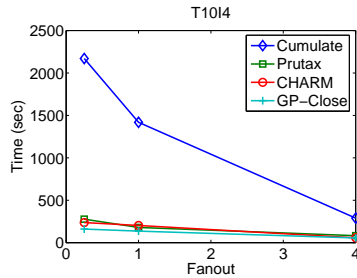
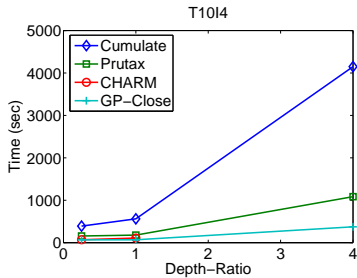
a. Number of patterns discovered in T10I4

b. Number of patterns discovered in T15I8



c. Execution time on T10I4

d. Execution time on T15I8



e. Execution time with respect to depth-ratio

f. Execution time with respect to fanout

Figure 5.8: The performance of GP-Close compared with Cumulate and Prutax on the synthetic data sets.

convenience in changing the settings of the data sets and studying the performance of our algorithm on various kinds of data. The parameters and their default values used by the synthetic data generator are listed in Table 5.3. In our experiments, we generate two synthetic data sets, namely T10I4 and T15I8. T10I4 is generated using the default settings of the IBM data generator. T15I8 is generated by changing the default values of  $|\mathcal{T}|$  and  $I$  to 15 and 8. T10I4 and T15I8 correspond to short transaction/pattern case and long transaction/pattern case respectively.

Figure 5.8.a and Figure 5.8.b show the number of patterns discovered by GP-Close, Cumulate, Prutax, and CHARM. We can see that the number of frequent generalized itemsets discovered by Cumulate and Prutax is usually two to three time larger than the number of closed generalization closures discovered by GP-Close and CHARM on the synthetic datasets. The execution times of the four algorithms with respect to the minimum support values are shown in Figure 5.8.c and Figure 5.8.d. We can see that GP-Close can perform almost an order of magnitude faster than Prutax and more than two order of magnitude faster than Cumulate on both T10I4 and T15I8. In particular, GP-Close can also perform two to three times faster than CHARM on both synthetic data sets. This can be seen as the benefit of our algorithm owing to its taxonomy based pattern pruning strategy.

We evaluate the influence of the taxonomy on the mining performance by varying the depth-ratio (see Figure 5.8.e) and fanout (see Figure 5.8.f) of the synthetic data sets. A low depth-ratio means that the items in the frequent generalized itemsets tend to appear at the low level of the taxonomy. This implies that there potentially exist more over-generalized itemsets. On the other hand, decreasing fanout means increasing the number of levels in the taxonomy. A lower fanout will cause more over-generalized itemsets to be generated in the data sets as well. Therefore, from Figure 5.8.e and Figure 5.8.f, we can see that the GP-Close algorithm demonstrates significant benefits

of over-generalization reduction at high depth-ratio and low fanout values. Figure 5.8.e also shows that CHARM cannot finish the mining task when the depth-ratio is set to 4 because of a memory-full failure. This may be explained as follows: CHARM does not use taxonomy based pattern pruning strategy; it potentially enumerates more candidate itemsets than GP-Close which may cause large memory occupation during the mining process.

### 5.5.2 Real World Semantic Web Data Sets

Besides the synthetic data sets for simulating the traditional application of market-basket analysis, a real world Semantic Web data set is adopted for evaluating the performance of the proposed GP-Close algorithm for mining RDF semantic metadata. We should note that the basic elements of RDF are not atomic items as in the market-basket analysis, but *RDF statements* (or *RDF relations*), each consisting of a subject, a predicate, and an object. For example, Figure 5.9 shows a sports domain RDF ontology containing two RDF concept taxonomies (with “Action” and “Player” as roots) and one predicate type “agent”. With this ontology, an event “Kaka scores” in a match can be represented by an RDF statement as  $\langle \text{Score, agent, Kaka} \rangle$ . When applying GARM to RDF metadata, an RDF statement can be generalized in different ways, such as generalizing the subject, generalizing the object, and generalizing both the subject and the object. Figure 5.10 shows a taxonomy of generalized statements which can be inferred from the original RDF taxonomy in Figure 5.9. We can see that the RDF statement taxonomy is much more complicated than the original RDF concept taxonomy. In fact, mining generalized patterns from RDF metadata based on such a taxonomy usually imposes high computational cost due to the high possibility of over-generalization and the pattern combinatorial explosion.

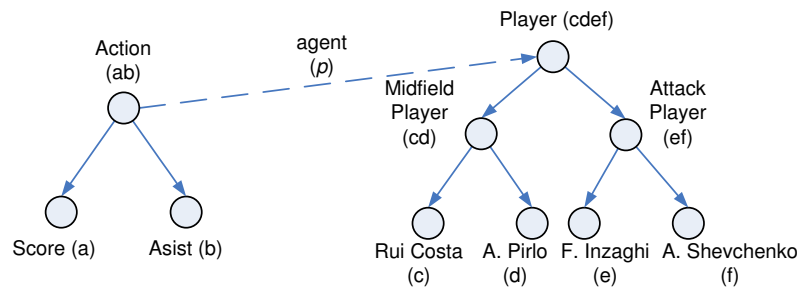


Figure 5.9: A sample RDF vocabulary.

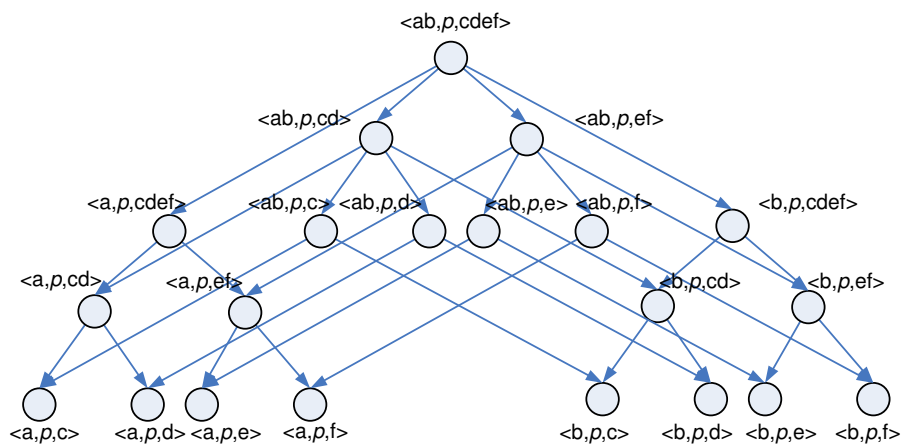


Figure 5.10: The RDF statement taxonomy inferred from the RDF vocabulary defined in Figure 5.9.

The real world RDF data set adopted in our experiments is the foafPub data set (<http://ebiquity.umbc.edu/resource/html/id/82/>) provided by the UMBC eBiquity Research Group. foafPub is a set of RDF files that describe people and their relationships with the use of the FOAF vocabulary (<http://xmlns.com/foaf/0.1/>).

Table 5.4 lists the key statistics of the RDF vocabularies used by the foafPub data set. The features of the RDF documents are summarized in Table 5.5. We can see that most of the RDF statements (semantic relations) in the Semantic Web data sets are distinct. This heightens the necessity of mining generalized patterns from the RDF documents. In addition, Table 5.5 shows that the set of the generalized RDF statements is much larger than the original set of the instance-level statements in the RDF documents. This implies

that there is a very large generalized pattern searching space in which over-generalized patterns may widely exist.

For mining frequent generalized patterns from the RDF data sets, we convert the RDF documents into binary transaction files. We assign a unique identifier (*rid*) for each RDF statement (including generalized statement) in the RDF data sets. Each RDF document is thus encoded as a transaction containing a set of *rids*. The RDF statement taxonomies are pre-computed and stored in a generalized statement lookup table, which is also stored in a binary file. All algorithms, namely Cumulate, Prutax, CHARM, and GP-Close, refer to this table for fast looking up of generalized RDF statements.

Table 5.4: The statistics of the RDF vocabularies in the foafPub and ICT-SB data sets.

RDF vocabulary	RDF predicates	RDF classes	Average Fanout
foafPub	36	6801	13

Table 5.5: The characteristics of the foafPub and ICT-SB data sets.

RDF Data Set	$ \mathcal{D} $	$N_r$	$N_r^*$	$N_{gr}$	Average length
foafPub	6170	85778	83759	207119	14

$|\mathcal{D}|$  denotes the number of RDF documents,  $N_r$  denotes the total number of RDF statements,  $N_r^*$  denotes the total number of distinct RDF statements, and  $N_{gr}$  denotes the total number of distinct generalized RDF statements.

The performance comparison of the four algorithms on the foafPub and ICT-SB data sets is shown in Figure 5.11. Figure 5.11.a shows that the number of closed generalization closures discovered by GP-Close and CHARM can be one to two orders of magnitude smaller than the number of all frequent generalized patterns discovered by Cumulate and Prutax. This is because the RDF statement taxonomies (see Figure 5.10) have a complex structure wherein one semantic relation can be generalized via either its subject



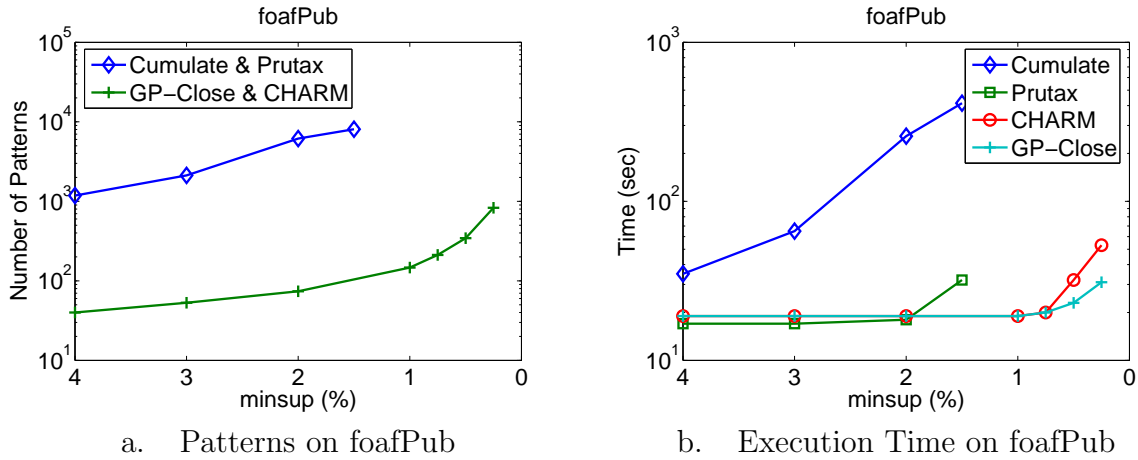


Figure 5.11: GP-Close compared with Cumulate and Prutax on the foafPub and ICT-SB data sets.

or object. This leads to potentially more over-generalized patterns in the Semantic Web data sets. Figure 5.11.b shows the execution time of the four algorithms on the foafPub and ICT-SB data sets. We can see that the GP-Close algorithm runs one to two orders of magnitude faster than the Cumulate algorithm. For Prutax, it can only run with a high *minsup* value. This is due to the fact that Prutax needs to store all discovered generalized itemsets in the main memory for candidate pruning. When *minsup* is below a certain value (1.5% for foafPub), the number of all frequent generalized itemsets can be too large to store in the main memory. In such cases, Prutax cannot work properly. In addition, GP-Close can still perform two to three times faster than CHARM when *minsup* is low.

To analyze the usefulness of the patterns extracted from the RDF data sets, we examine the closed closures extracted by GP-Close from the foafPub data with a *minsup* of 0.5%. For each closed closure  $gc(X)$ , we convert it into a simplified pattern (itemset) by removing those RDF statements that have specialized statements in  $gc(X)$ . For illustration, three sample patterns are listed below:

- (1)  $\{ \langle \text{foaf:Person, foaf:depiction, foaf:Document} \rangle, \langle \text{foaf:Person, foaf:weblog, foaf:Document} \rangle, \langle \text{foaf:Person, foaf:homepage, foaf:Document} \rangle \}$  ( $supp=4.4\%$ ).

(2) {<foaf:Person, foaf:depiction, http://holonet.thebhg.org/>, <foaf:Person, foaf:img, http://holonet.thebhg.org/>, <foaf:Person, foaf:interest, http://www.thebhg.org/index.htm>} (*supp*=1.0%).

(3) {<foaf:Person, foaf:interest, http://techedbloggers.net/>, <foaf:Person, foaf:homepage, http://blogs.msdn.com/>, <foaf:Person, foaf:weblog, http://blogs.msdn.com/>} (*supp* = 0.98%).

The first pattern indicates that three kinds of predicates, i.e., “foaf:depiction”, “foaf:weblog”, and “foaf:homepage”, frequently co-occur in the documents. Such knowledge can be used to improve the storage efficiency of RDF databases [WSKR03]. The second pattern looks more interesting as it reflects the correlation between the predicate values, i.e., for people interested in “http://www.thebhg.org/index.htm”, their depiction and image often can be found on the same website “http://holonet.thebhg.org/”. This can be a typical situation in the online interest communities that people who are interested in a community may have personal community accounts which contain their depictions and images. The third pattern represents an association between the predicate values related to two different websites which, in fact, are allied.

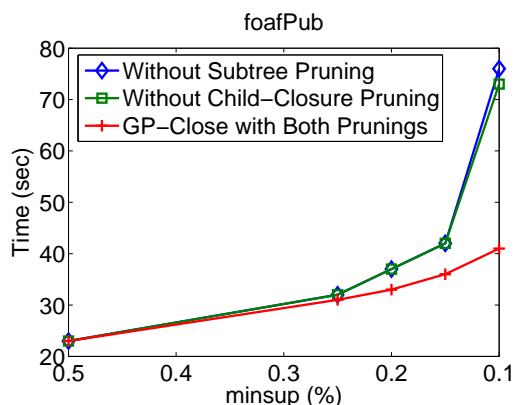


Figure 5.12: Efficiency of using child-closure and subtree based pruning strategy.

### 5.5.3 Effect of Pruning Strategies

For evaluating the effect of our proposed pruning strategies, we compare the performance of the GP-Close algorithm with and without using child-closure pruning and subtree pruning based on the foafPub data set. Three versions of GP-Close are compared, one without using the child-closure pruning strategy, one without using the subtree pruning strategy, and one using both child-closure pruning and subtree pruning. As shown in Figure 5.12, we can see that both child-closure pruning and subtree pruning can efficiently speed up our algorithm in particular when the *minsup* is low. Each pruning strategy can reduce about 30% of the execution time of GP-Close with a *minsup* of 0.1%. Similar properties are also observed on other data sets.

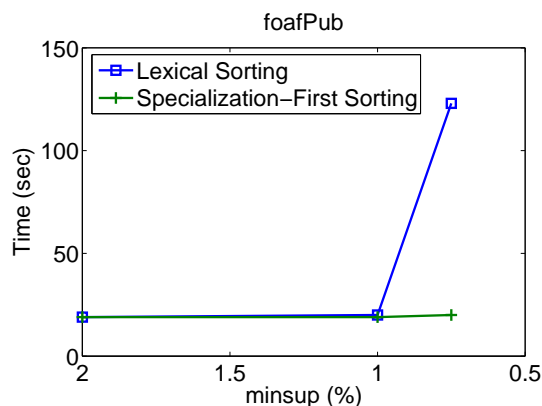


Figure 5.13: Efficiency of the specialization-first sorting compared with the lexical sorting strategy.

### 5.5.4 Effect of Specialization-First Sorting

For evaluating the effect of our proposed specialization-first sorting strategy, we compare two versions of the GP-Close algorithm. In one version, we sort the closures in the support-increasing and length-decreasing (i.e., specialization-first) manner. The other version sorts the closures based on their lexical order. The execution times of applying the two GP-Close on T15I8 data set are summarized in Figure 5.13. As expected,

the specialization-first sorting performs better than lexical sorting strategy when the *minsup* is less than 1. Furthermore, when the *minsup* decreases, the benefit of using specialization-first sorting becomes increasingly obvious.

## 5.6 Analysis of Terror Attack Documents: A Case Study

For an overall evaluation of the semantic based multimedia web information analysis method for real world information analysis tasks, we applied the proposed RDF metadata extraction and mining approaches on two online terror attack databases on the web site of the International Policy Institute for Counter-Terrorism (ICT) <sup>4</sup>. The contents of the online documents are mainly text descriptions of car bombing and suicide bombing events.

### 5.6.1 Semantic Metadata Extraction

We apply RDF metadata extraction to extract semantic metadata from the ICT suicide bombing (ICT-SB) documents and the ICT car bombing (ICT-CB) documents with a WordNet search depth (*WN<sub>SD</sub>*) of 2 and a minimum similarity threshold (*minsim*) of 0.1. The RDF extraction task is performed by using our in-house developed semantic metadata extraction tool. Figure 5.14 shows the interface of the extraction tool including a graphic illustration of an RDF conceptual graph extracted from a web document.

The statistics of the RDF metadata extraction and term taxonomy construction are summarized in Table 5.6. Figure 5.15 shows a subset of the extracted term taxonomy. We see that most of the extracted RDF statements (semantic relations) are distinct, heightening the necessity of mining generalized patterns. In addition, Table 5.6 shows

---

<sup>4</sup><http://www.ict.org.il/>

## CHAPTER 5. MINING RDF SEMANTIC METADATA FOR MULTIMEDIA ANALYSIS

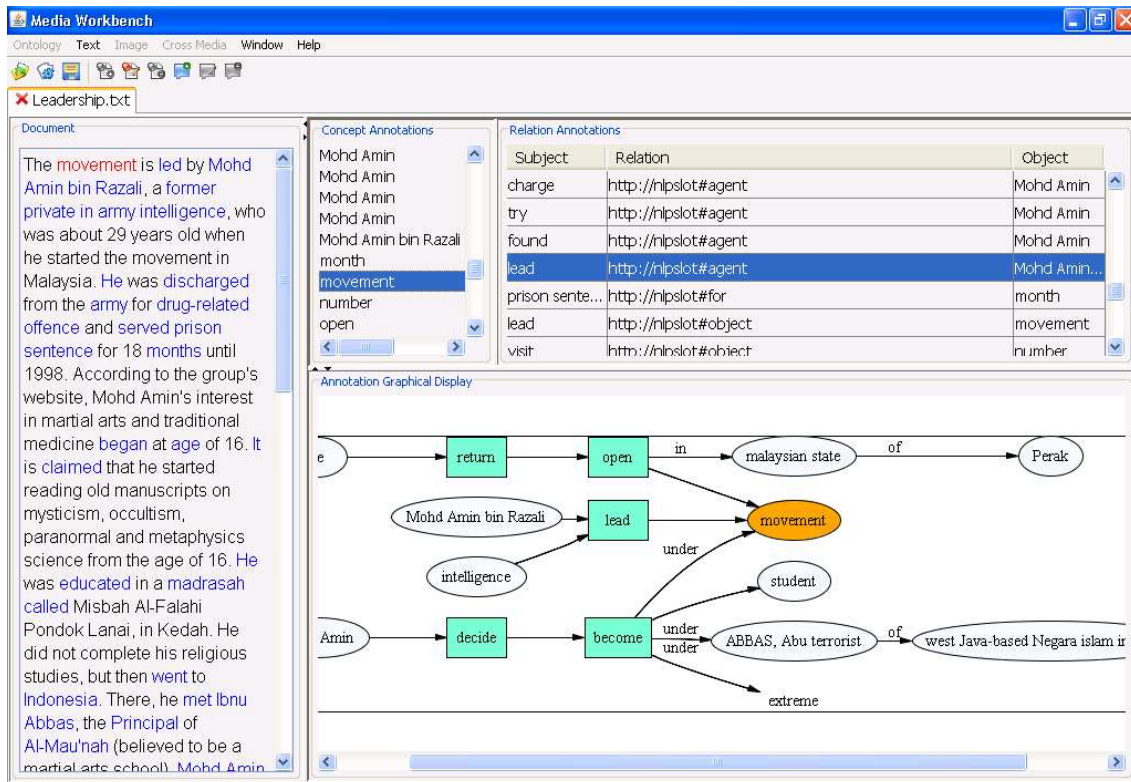


Figure 5.14: The semantic metadata extraction tool.

that the set of generalized RDF statements, which can be derived from the term taxonomy, is much larger than the original set of RDF statements stored in RDF metadata. This implies that there is a very large generalized pattern search space.

We apply both the GP-Close algorithm and the Cumulate algorithm [SA95] for performance evaluation and comparison. We choose the Cumulate algorithm as the reference algorithm because Cumulate uses hard disk for support counting and pattern storage and therefore can work properly at low support level when huge number of over-generalized pattern exist. Other memory based algorithms, such as Prutax and CHARM, store frequent patterns and candidate patterns in main memory and therefore can only work properly at high support level.

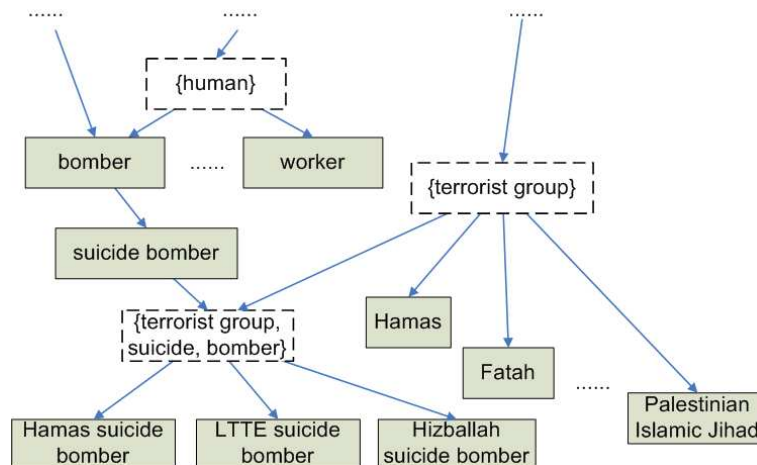


Figure 5.15: A subset of the extracted term taxonomy.

Table 5.6: Summary of RDF metadata extraction results.

RDF Data set	Documents	$N_r$	$N_r^*$	$N_{gr}$	$n_{gr}$	$NoC$	$NoR$
ICT-SB	106	1938	1665	48691	29	1578	1988
ICT-CB	128	2224	1814	52673	29	1790	2198

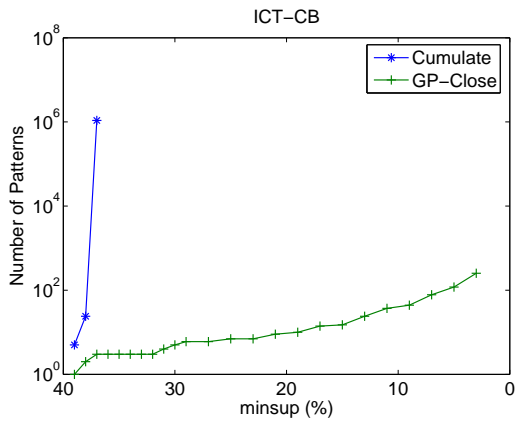
$N_r$  denotes the number of RDF statements extracted,  $N_r^*$  denotes the number of distinct RDF statements,  $N_{gr}$  denotes the number of distinct generalized RDF statements,  $n_{gr}$  denotes the average number of generalized statements for an RDF statement at instance level,  $NoC$  denotes the number of term clusters in the term taxonomy (i.e., RDF classes in the RDF vocabulary), and  $NoR$  denotes the number of *is-a* relations in the term taxonomy (i.e., “rdfs:subClassOf” relations between RDF classes in the RDF vocabulary).

## 5.6.2 Mining Generalized Associations on RDF Metadata

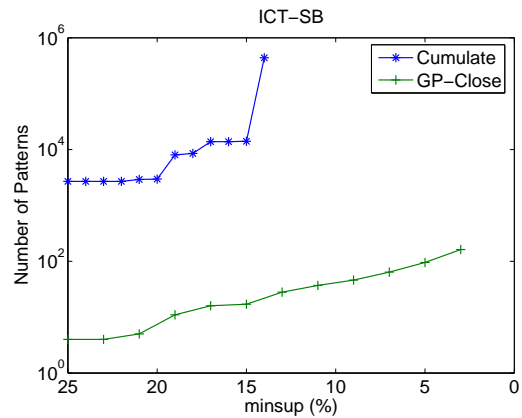
### 5.6.2.1 Number of Patterns

We first do a comparison on the number of the patterns extracted by the two algorithms, Cumulate and GP-Close. Figure 5.16 shows that the number of closed generalization closures can be one to two orders of magnitude smaller than the number of frequent relationsets discovered by Cumulate. This is especially so with a low *minsup*.

We can see that as *minsup* decreases, the number of frequent generalized patterns increases rapidly. The reason is that when *minsup* is low, larger and more specialized



5.16.a: ICT-CB



5.16.b: ICT-SB

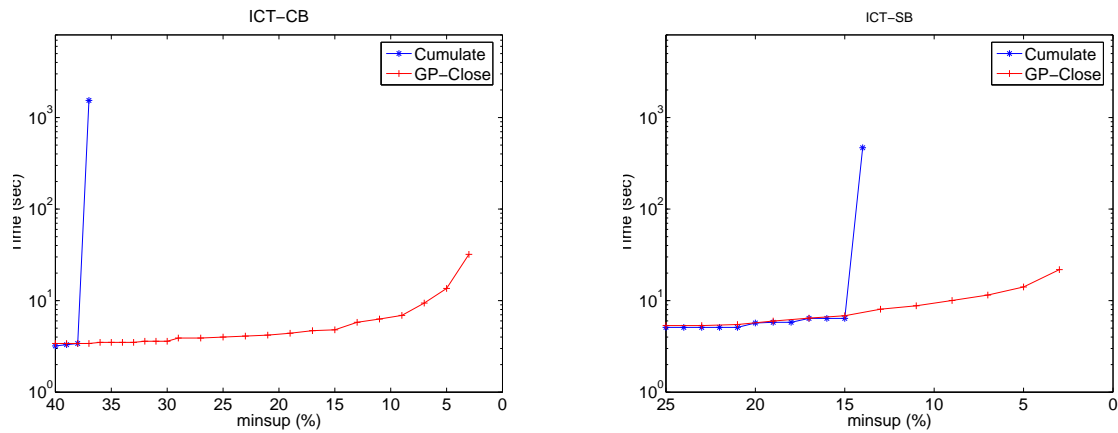
Figure 5.16: Number of Patterns

patterns will be discovered. Note that the more specialized and larger a frequent pattern is, the more generalized patterns it contains. And all these generalized patterns are also frequent. When mining associations of RDF statements, the problem becomes critical as an RDF statement is a triplet which consists of a subject, a predicate, and an object, which means a statement can be generalized in many different ways. Therefore, when patterns become larger and more specialized, the number of their generalized patterns will increase dramatically.

As the Cumulate algorithm generates all frequent patterns, we can expect that the execution time of the Cumulate algorithm also increases very fast when *minsup* decreases. For the GP-Close algorithm, as it generates only a small set of the closed generalization closures, its execution time is expected to be less sensitive to *minsup*.

### 5.6.2.2 Time Efficiency

Figure 5.17 shows the execution times of the algorithms by varying the minimum support (*minsup*). As expected, Cumulate can work properly only with high *minsup*. When *minsup* is high, the performance of the algorithms are close. However, when *minsup* is



5.17.a: ICT-CB

5.17.b: ICT-SB

Figure 5.17: Execution Time

low, the GP-Close algorithm can run more than one to two orders of magnitude faster than Cumulate. This can be explained by the number of patterns discovered by the various algorithms, i.e., Cumulate generates many more patterns than GP-Close.

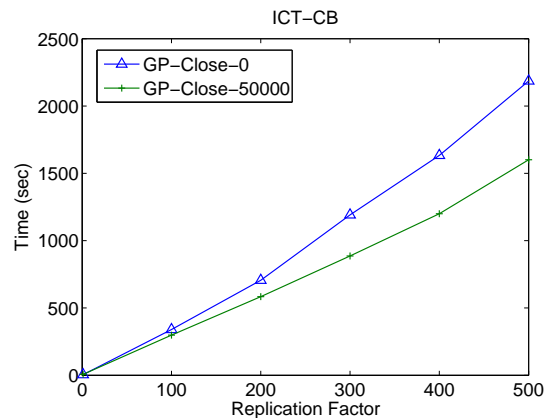


Figure 5.18: Scalability of the GP-Close algorithm.

### 5.6.2.3 Scalability

The scalability of the GP-Close algorithm is also evaluated by replicating the ICT-CB database 100 to 500 times with an incremental step of 100, i.e., the largest data set



contains more than 50,000 RDF documents. For introducing variations in the replicas, we randomly modify 1% of the RDF statements by replacing them with noisy statements. Then the GP-Close algorithm is tested on the replicated data sets with a fixed *minsup* of 10%. As shown in Figure 5.18, the execution time of the GP-Close algorithm increases linearly with the replication factor.

### 5.6.3 Analysis of Patterns

For evaluating the quality of the patterns discovered, we analyze the patterns generated using GP-Close on the ICT-CB data set with a *minsup* of 7%. We examine each of the 78 generalized patterns (relationsets), to verify (1) if it is previously known and (2) if it is significant.

We observe that 71.8% (56 out of 78) of the patterns are commonsense patterns already known by people. For example, the pattern  $\{\langle \text{explode}, \text{agent}, \text{bomb} \rangle, \langle \text{wound}, \text{theme}, \text{people} \rangle\}$  (with a support of 7.0%) describes a commonsense fact that when a bomb explodes, people get wounded. Though such patterns may not be significant knowledge for human users, they may still be useful for the tasks of clustering or classification of web documents as they reflect the underlying semantic structures of a particular domain. 12.8% (10 out of 78) of the patterns are identified as previously unknown and not useful. These patterns usually involve general terms, such as ‘group’ or ‘city’. For example,  $\{\langle \text{kill}, \text{theme}, \text{group} \rangle\}$  can be interpreted as “something related to a certain group is killed”, with a support of 12.6%. However, terms like ‘group’ or ‘city’ are too general to inform detailed semantic knowledge. The remaining 15.4% (12 out of 78) of the patterns are previously unknown and potentially useful. An example of such patterns is  $\{\langle \text{detonate}, \text{theme}, \text{truck} \rangle\}$ , with a support of 11.8%. In the ICT-CB (car bombing) data set, this pattern can be interpreted as “truck is often used by terrorists for carrying out car bombing”. Another example is  $\{\langle \text{claim}, \text{agent}, \text{Terrorist\_Group\_001} \rangle, \langle \text{claim},$

theme, responsibility>}, with a support of 7.8%. This pattern can be explained as “the terrorist group with the identifier *Terrorist\_Group\_001* is often claims responsibility for a car bombing event”.

## 5.7 Summary

The contribution of this chapter lies on that we present a systematic approach for mining frequent generalized patterns from large databases with *over-generalization reduction*. We theoretically prove that over-generalization reduction problem can be translated into a closed pattern mining problem. Specifically, the proposed GP-Close algorithm can efficiently discover a small set of closed generalization closures from which all frequent and non-over-generalized patterns can be derived. Extensive experiments have been conducted to illustrate the performance of GP-Close in mining generalized patterns from large databases. The experimental results show that our method can substantially reduce pattern redundancy and perform much faster than the original generalized association rule mining algorithms, namely *Cumulate* and *Prutax*, and the closed pattern mining algorithm, *CHARM*, in terms of time efficiency, especially when the minimum support is low. In particular, we show that the proposed GP-Close algorithm can perform especially well on RDF semantic metadata collection and therefore suitable for our semantic based multimedia information analysis framework.

# Chapter 6

## Conclusions and Future Work

### 6.1 Conclusion

This thesis has presented a framework for addressing the problem of multimedia web information fusion and analysis. In particular, we have proposed, implemented, and evaluated key techniques for image and text information fusion, semantic multimedia content representation, semantic metadata extraction, and metadata based multimedia analysis.

This chapter aims to retrospect the challenges we identified for the multimedia information fusion and analysis task and summarize the strategies and methods developed to tackle the problems.

#### 6.1.1 Media Fusion

The aim of media fusion is to integrate heterogeneous but semantically related multimedia information. As the semantics of visual and audio information are difficult to understand by computer programs, we adopt a complementary method to associate media objects with related surrounding texts for disambiguating their semantic meanings.

In particular, we study two methods for discovering the underlying associations between images and texts in web pages in chapter 3. The first method is based on a multilingual retrieval model and employs a vague transformation technique for measuring

the semantic relevance between images and texts. Another method uses a self-organizing neural network to learn direct mapping between the visual and textual features by automatically and incrementally summarizing the associated image perceptual features and textual features into a set of information templates, which are subsequently used as prototypes for identifying new associated image-text pairs.

Extensive experiments conducted show that combining textual features and visual features can achieve a reasonable performance for identifying image-text associations. The two proposed methods are comparable in term of the testing performance and all perform much better than the baseline image annotation method CMRM, in particular on small training data sets.

### **6.1.2 Multimedia Content Representation and Metadata Extraction**

As multimedia data are inherently heterogeneous in their formats, fusion and analysis of multimedia information based on raw data or low-level perceptual features is computationally impractical. In spite of the various low-level data representations, semantic meanings conveyed by the multimedia objects are constituted by the same sort of real world concepts. However, existing methods usually describe multimedia semantics with shallow lexical keywords or individual concepts inconsistent with the conceptual models in human brains, where concepts are interconnected with relations.

In view of the limitations of the existing methods, we present in chapter 4 a strategy which combines conceptual graphs, Resource Description Framework, and MPEG-7 standard for multimedia content representation. Conceptual graph is a knowledge representation approach borrowed from the field of artificial intelligence. As conceptual graphs are derived from knowledge models of natural languages, they conform to the conceptual structures used by human reasoning. In addition, as they can be automatically extracted based on the syntax in text sentences, conceptual graphs also provide a grammar

foundation for automatic multimedia semantic extraction. While conceptual graph based knowledge representation method focuses on the semantic aspect of multimedia contents, the MPEG-7 standard is used for describing media-specific information of the multimedia objects, including author descriptions, creation contexts, media formats and low-level perceptual features. Finally, Resources Description Framework (RDF) provides a well-defined data model for encoding both semantic contents and media-specific information in a machine-processible and human-understandable form. The proposed multimedia content representation method has been implemented and used in our multimedia fusion and analysis framework for indexing and describing heterogeneous multimedia information, integrating both the low-level perceptual features and high-level semantics, in a unified representation for supporting cross-media information computing.

More importantly, chapter 4 also presents a method that automatically extracts semantic metadata for describing online multimedia content based on web texts. Natural language processing techniques are employed for analyzing and parsing web text sentences into text grammar trees, based on which semantic concepts and relations are extracted and converted into conceptual graphs encoded in RDF language. The semantics of the extracted concepts are defined as bags of word senses in WordNet. Based on the WordNet sense representations, a similarity measure and an incremental clustering strategy are developed for organizing the extracted concepts into a term taxonomy which serves as a domain vocabulary encoded using RDF Schema. A real world case study presented in Chapter 5 shows that our method can extract informative semantic metadata and vocabulary for the RDF mining algorithm to discover interesting knowledge.

### **6.1.3 Multimedia Information Analysis**

Our ultimate goal is to support cross-media information analysis to extract interesting and useful knowledge. In this thesis, this goal is achieved by developing and applying

data mining techniques to RDF semantic metadata that describe the multimedia content. As the RDF semantic metadata are semi-structured data, analyzing semantic metadata may pose performance problems.

In chapter 5, a generalized association pattern mining algorithm is presented for discovering association patterns from the semi-structured metadata composed of semantic concepts and relations. Concept hierarchies are also employed for discovering knowledge at various semantic levels.

In view that the existing algorithms generate a large number of redundant *over-generalized patterns*, especially from RDF metadata, GP-Close adopts a systematic method for full over-generalization reduction based on a mathematical notion of *generalization closure*. Extensive experiments have been conducted to show that the GP-Close algorithm can significantly reduce pattern redundancy and outperform the existing generalized pattern mining algorithms in terms of time efficiency. A case study of applying GP-Close to a terrorist domain document collection is also presented.

## 6.2 Future Work

The following sections describe how the framework and techniques presented in this thesis can be improved and refined.

### 6.2.1 Media Fusion

Our current methods for identifying image-text associations are basically conducted by measuring the similarity between the image perceptual information in the visual feature space and the text semantic information in the textual feature space. The two similarity models, i.e., the vague transformation based model and the Fusion ART based resonance model, can also be extended in two aspects.

CHAPTER 6. CONCLUSIONS AND FUTURE WORK

---

Firstly, contextual information is currently not considered in the similarity calculation. For example, an image region in orange color together with an image region of a *debris* scene may represent the *fire* concept. However, an image region with a similar orange color together with an image region of a *human face* may represent *orange clothes of security people*. Therefore, visually similar information may not be semantically similar. For achieving a better similarity estimation, methods that employ contextual information are worthwhile to explore.

Secondly, visual and textual information taxonomies can also be utilized for calculating information similarity. In fact, term or word taxonomies have been proven useful for improving similarity matching in text mining applications [Res99]. We can thus borrow the idea of the taxonomy based text similarity measures and employ visterm taxonomies for improving the similarity computations in the visual feature space.

Figure 6.1 (a) shows a visterm taxonomy, where the visterm 1 and the visterm 2 are more visually similar to each other than to the visterm 3. Given such a taxonomy, we can calculate the similarity between each pair of visterms using a formula, such as  $sim(v_i, v_j) = \frac{1}{dis_{tax}^2(v_i, v_j)}$ , where  $dis_{tax}(v_i, v_j)$  is the length of the path between the visterm  $v_i$  and the visterm  $v_j$ . We can then convert the visterm taxonomy into a similarity matrix  $S$  (see Figure 6.1 (b)) and use  $S$  as a kernel matrix for visterm vector similarity calculation. For example, given the two visterm vectors  $V_i$  and  $V_j$ , we can calculate the similarity between  $V_i$  and  $V_j$  using a formula  $SIM(V_i, V_j) = V_i S^T V_j^T$  (Figure 6.1 (c)).

Using visterm taxonomy however raises two challenges. The first challenge is how to build such visterm taxonomies. A suitable hierarchical clustering method can be chosen or developed for this purpose. Moreover, we can either build only one visterm taxonomy by combining all visual features, such as color and texture features, for clustering or create multiple visterm taxonomies by performing clustering on different feature sets. The second challenge is that while visterm taxonomies reflect visual similarities of different

CHAPTER 6. CONCLUSIONS AND FUTURE WORK

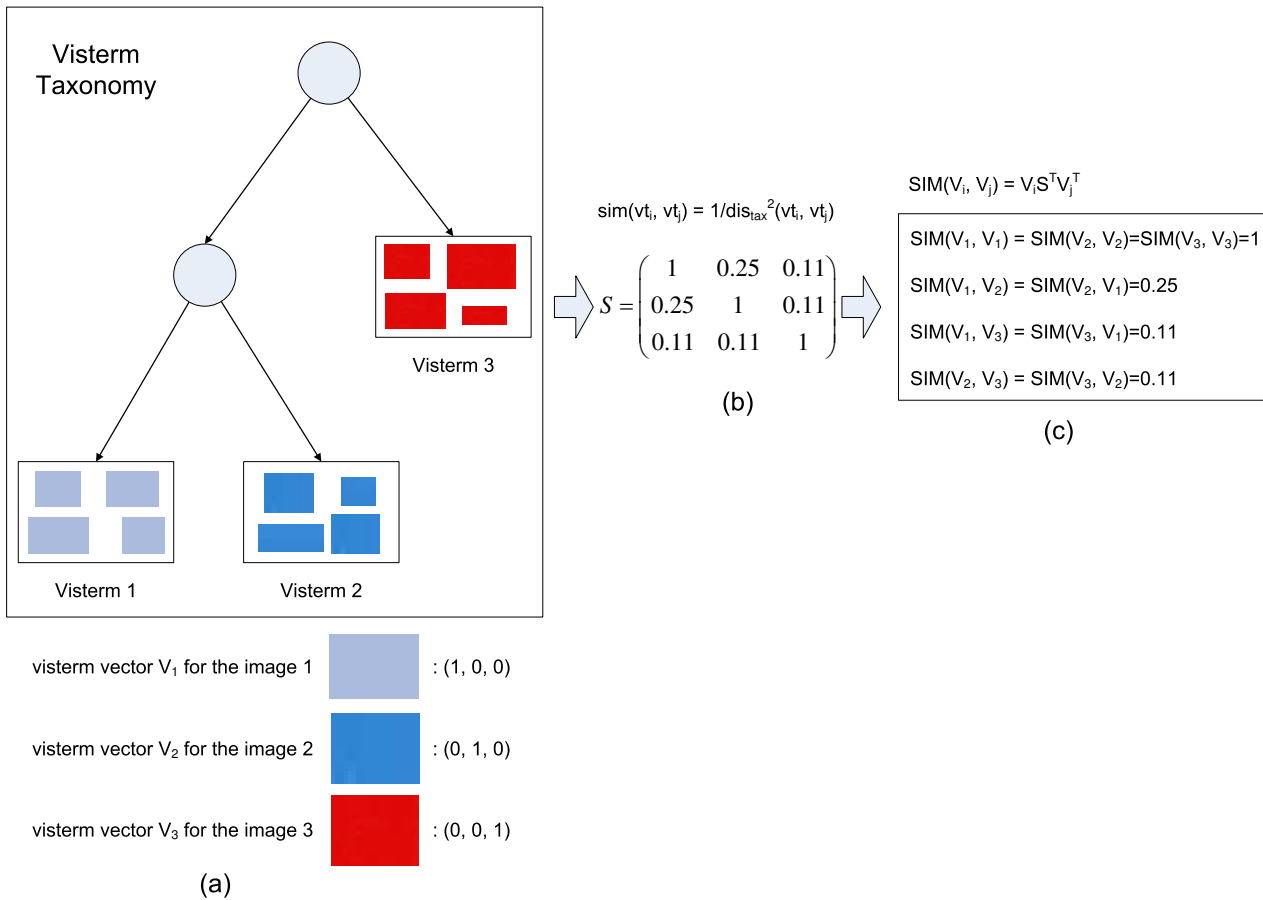


Figure 6.1: Using visterm taxonomy for similarity calculation in visual information space.

visterms, they may not be consistent with the semantic similarities of the visterms. Methods need to be proposed for combining and consolidating the visterm taxonomies with the term or concept taxonomies, which reflect the semantic similarities of the visual information.

On a broader perspective, we see that learning the associations between media objects and texts provides a compromised solution for linking multimedia contents with semantic descriptions. However, in some cases, the meanings of the web texts may not exactly match with the media contents. Therefore, methods for directly learning the semantics based on low-level perceptual information are undoubtedly useful for media fusion and worthwhile to investigate further.



### 6.2.2 Multimedia Content Representation and Metadata Extraction

Currently, the proposed multimedia fusion and analysis framework mainly models multimedia data using the information at the semantic layer. For extending the current framework, low-level perceptual information can be incorporated to provide a more complete representation of the multimedia contents. As low-level features, such as color, texture, or edge features, can be automatically extracted from the raw media data, existing media feature extraction tools, such as the MPEG-7 XM software [Mar04], can be utilized in our RDF metadata extraction module for automatically generating perceptual information descriptions.

Moreover, as low-level features are usually expressed in numeric forms and not readable for human users, a perceptual information ontology or vocabulary can be useful for bridging the gap between the numeric low-level features and human-understandable perceptual information concepts, such as “red” or “coarse texture”. This is important for information analysis tasks to represent the data contents and the analysis results in a human-understandable form. For example, a cloth design pattern “most of the male jackets sold in this winter have color within the range from #F0F0F0 to #FFFFFF” may be quite difficult to understand for a fashion journalist without the prior knowledge of the underlying color representation schemas. With a perceptual information ontology which defines a color concept, such as “black”, the cloth design pattern can be represented in a more understandable form, such as “most of the male jackets sold in this winter are in black color”. To build such a perceptual information ontology, two methodologies can be explored. First is a *top-down* approach, by which a vocabulary or taxonomy containing a set of perceptual concepts, such as color concept “black” and texture concept “coarse texture”, can be defined by human in advance or extracted from lexical ontologies, such as WordNet. Then, sample media data can be assigned to the concepts for building or

training low-level feature prototypes. Secondly, such a perceptual ontology can also be created in a *bottom-up* manner, in which sample media data can be first clustered into groups to form a set of low-level feature prototypes and then perceptual concepts can be defined on top of the generated prototypes.

### 6.2.3 Multimedia Information Analysis

For multimedia information analysis, the generalized association pattern mining methods can be extended in two directions. The first direction is to incorporate low-level perceptual information in the pattern mining process so that the knowledge combining conceptual and perceptual information can be discovered. Another direction is to extend our GP-Close algorithm and its generalization closure based pattern reduction techniques to handle graph based data structures. Currently, our algorithm treats RDF semantic metadata as a set of RDF statements or relations. We can also treat such a set of RDF relations as a whole graph and discover frequent generalized semantic graph patterns. For doing this, a graph based notion of generalization closure needs to be defined. Applying generalization closure based method for mining generalized graph patterns will be an interesting part of our future work.

Besides association pattern mining, more techniques can be applied to knowledge discovery based on RDF semantic metadata. One possible application is to perform RDF metadata clustering for identifying groups of multimedia contents with similar semantics. Such RDF metadata clustering methods can be used for topic detection, redundant information detection, and multimedia data summarization. To perform clustering on RDF metadata collection, a similarity measure needs to be proposed to evaluate the closeness of two RDF semantic graphs.

# Appendix A

## List of Publications

### A.1 Book Chapters

- (i) Tao Jiang, Ah-Hwee Tan. *Mining Association Rules from RDF Documents*. In S. Bandyopadhyay, U. Maulik, L. Holder and D. Cook (Eds.), *Advanced Methods for Knowledge Discovery from Complex Data*, Springer-Verlag, 2005.

### A.2 Journals

- (i) Tao Jiang, Ah-Hwee Tan, and Ke Wang. *Mining Generalized Associations of Semantic Relations from Textual Web Content*. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, Vol 19, No. 2 164-179.
- (ii) Tao Jiang, Ah-Hwee Tan. *Towards Multimedia Web Information Fusion: Learning Image-Text Associations*. Submitted to *IEEE Transactions on Knowledge and Data Engineering (TKDE)*.
- (iii) Tao Jiang, Ah-Hwee Tan, and Ke Wang. *Mining Generalized Itemsets with Over-Generalization Reduction*. Submitted to *IEEE Transactions on Knowledge and Data Engineering (TKDE)*.

### A.3 Conferences

- (i) Tao Jiang and Ah-Hwee Tan. *Discovering Image-Text Associations for Cross-Media Web Information Fusion*. In proceedings, 17th European Conference on Machine Learning and the 10th European Conference on Principles and Practice of

CHAPTER A. LIST OF PUBLICATIONS

---

- Knowledge Discovery in Databases (ECML/PKDD'06), LNCS 4213, pp. 461-468, Berlin, Germany, September 2006.
- (ii) Tao Jiang and Ah-Hwee Tan. *Mining RDF Metadata for Generalized Association Rules*. In proceedings, 17th International Conference on Database and Expert Systems Applications (DEXA'06), LNCS 4080, pp. 223-233, Krakow, Poland, 4-8 September 2006.
- (iii) Tao Jiang and Ah-Hwee Tan. *Mining RDF metadata for generalized association rules: knowledge discovery in the semantic web era*. In Proceedings, 15th International Conference on World Wide Web (WWW'06), Edinburgh, Scotland, pp. 951-952, May 2006.

## Appendix B

# Evaluation of the Similarity Measure for Term Taxonomy Construction

For evaluating the efficacy of our defined NNS similarity measure (Chapter 4.2.2.2), we compare it with a state of the art similarity measure developed by Seco et al [SVH04]. This measure is chosen because it is reported more closely correlated to human’s assessments of the term similarities. This is important as we hope that the final term taxonomy will conform to the human’s perspectives. In addition, Seco et al Measure is based purely on WordNet without using a large training corpus.

For comparing the two similarity measures, we randomly select 200 terms which are extracted by our proposed term/relation extraction process (see Chapter 4.2.1). For each term, we use different similarity measures to select a set of most similar terms from the other 199 terms. The correctness of the selection results are evaluated by humans. Table B.1 and B.2 show the evaluation results. The first column of the tables shows the total number of terms that have most similar terms selected out by the similarity measures. The second column shows the total number of terms that do not have any similar terms selected out by the similarity measure, i.e., the similarity scores are all zero.

We notice that the overall precision of NSS measure (63.0%, 126 out of 200) is higher than that of Seco et al measure (54.5%, 109 out of 200). Seco et al measure has a large probability to miss the most similar terms for a specific term, even some of those similar terms can be easily identified by human, e.g. “damage” and “injure”. On the other hand, both measures have certain amount of wrongly selected most similar terms. These wrong selections can fall in two categories which we call Type-I error and Type-II error. Type-I error refers to assigning similar terms to a term which should have no similar term. Type-II error refers to assigning wrong most similar terms to a term. We can see

## CHAPTER B. EVALUATION OF THE SIMILARITY MEASURE FOR TERM TAXONOMY CONSTRUCTION

Table B.1: NSS measure with  $WNSD = 3$ 

	Most similar terms selected	No similar term selected	Total
Correct	111	15	126 (63%)
Wrong	73 (Type I error: 41; Type II error: 32)	1	74 (37%)
Total	184	16	200

Table B.2: Seco et al measure

	Most similar terms selected	No similar term selected	Total
Correct	85	24	109 (54.5%)
Wrong	74 (Type I error: 32; Type II error: 42)	17	91 (45.5%)
Total	159	41	200

that Type-II error is far more serious than Type-I error. In fact, Type-II error may cause a term to be assigned to a wrong cluster, while Type-I error at most causes an outlier (a term that is not similar with any other terms) to be assigned to a cluster. Note that most wrong selections of similar terms by the NSS measure are Type-I errors. On the contrary, most of those wrong selections by the Seco et al measure are Type-II errors. Based on the above discussions, we see that using the similarity measure defined in Eq. 4.1 can achieve better performance than the Seco et al measure.

# References

- [AIS93] Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami. Mining association rules between sets of items in large databases. In Peter Buneman and Sushil Jajodia, editors, *SIGMOD'93*, pages 207–216. ACM Press, 1993.
- [AS94] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *Proceedings of VLDB'94, Santiago de Chile, Chile*, pages 487–499. Morgan Kaufmann, 1994.
- [Bar97] Caroline Barriere. *From a children's first dictionary to a lexical knowledge base of conceptual graphs*. PhD thesis, 1997. Adviser-Fred Popowich.
- [BB97] Matthias Blume and Dan R. Ballard. Image annotation based on learning vector quantization and localized haar wavelet transform features. volume 3077, pages 181–190. SPIE, 1997.
- [BBT06] Marco Bertini, Alberto Del Bimbo, and Carlo Torniai. Automatic annotation and semantic retrieval of video sequences using multimedia ontologies. In *MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia*, pages 679–682, New York, NY, USA, 2006. ACM Press.
- [BCS01] Ana B. Benitez, Shih-Fu Change, and John R. Smith. Imka: a multimedia organization system combining perceptual and semantic knowledge. In *MULTIMEDIA '01: Proceedings of the ninth ACM international conference on Multimedia*, pages 630–631, New York, NY, USA, 2001. ACM Press.
- [Ber02] Pavel Berkhin. Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA, 2002.

## REFERENCES

- [BF01] Kobus Barnard and David Forsyth. Learning the semantics of words and pictures. In *ICCV 2001.*, volume 2, pages 408–415, 2001.
- [BFL<sup>+</sup>88] Peter Biebricher, Norbert Fuhr, Gerhard Lustig, Michael Schwantner Technische Hochschule Darmstadt, and Fachbereich Informatik. The automatic indexing system air/phys - from research to applications. In *SIGIR '88*, pages 333–342, New York, 1988. ACM Press.
- [BH95] Dick C. A. Bulterman and Lynda Hardman. Multimedia authoring tools: State of the art and research challenges. In *Computer Science Today*, pages 575–591. 1995.
- [BH01] Alexander Budanitsky and Graeme Hirst. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures, 2001.
- [bie00] Iso/iec 13250:2000 topic maps: Information technology–document description and markup language. Technical report, ISO/IEC, 2000.
- [BKKR01] John Bateman, Jorg Klein, Thomas Kamps, and Klaus Reichenberger. Towards constructive text, diagram, and layout generation for information presentation. *Comput. Linguist.*, 27(3):409–449, 2001.
- [BL97] Michael J. Berry and Gordon Linoff. *Data Mining Techniques: For Marketing, Sales, and Customer Support*. John Wiley & Sons, Inc., New York, NY, USA, 1997.
- [BL01] Tim Berners-Lee. Conceptual graphs and semantic web - reflections on web architecture, 2001.
- [BLHL01] Tim Berners-Lee, James Hendler, and Ora Lassila. Semantic web. *Scientific American*, 284(5):35–43, 2001.
- [BLLL05] Margherita Berardi, Michele Lapi, Pietro Leo, and Corrado Loglisci. Mining generalized association rules on biomedical literature. In *IEA/AIE'2005: Proceedings of the 18th international conference on Innovations in Applied Artificial Intelligence*, pages 500–509, London, UK, 2005. Springer-Verlag.



REFERENCES

---

- [BPS<sup>+</sup>04] Stephan Bloehdorn, Kosmas Petridis, Nikos Simou, Vassilis Tzouvaras, Yannis Avrithis, Siegfried Handschuh, Yiannis Kompatsiaris, Steffen Staab, and Michael G. Strintzis. Knowledge representation for semantic multimedia content analysis and reasoning. In *Proceedings of the European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology (EWIMT)*, NOV 2004.
- [Bri92] Eric Brill. A simple rule-based part of speech tagger. In *ANLP*, pages 152–155, 1992.
- [BSC00] Ana B. Benitez, John R. Smith, and Shih-Fu Chang. Medianet: A multimedia information network for knowledge representation. In *SPIE Conference on Internet Multimedia Management Systems (IS&T/SPIE)*, volume 4210, Boston, MA, November 2000.
- [BTP<sup>+</sup>00] Yves Bastide, Rafik Taouil, Nicolas Pasquier, Gerd Stumme, and Lotfi Lakhal. Mining frequent patterns with counting inference. *SIGKDD Explorations*, 2(2):66–75, 2000.
- [Car99] Sharon A. Caraballo. Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 120–126, Morristown, NJ, USA, 1999. Association for Computational Linguistics.
- [CC96] Ross Cutler and Kasim Selçuk Candan. Multimedia authoring systems. pages 279–296, 1996.
- [CCS00] Dana Cristofor, Laurentiu Cristofor, and Dan A. Simovici. Galois connections and data mining. *J. UCS*, 6(1):60–73, 2000.
- [CG87a] Gail A. Carpenter and Stephen Grossberg. ART 2: Self-organization of stable category recognition codes for analog input patterns. *Applied Optics*, 26:4919–4930, 1987.
- [CG87b] Gail A. Carpenter and Stephen Grossberg. A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics and Image Processing*, 37:54–115, 1987.

REFERENCES

---

- [CG91] Gail A. Carpenter and Stephen Grossberg. *Pattern Recognition by Self-Organizing Neural Networks*. Cambridge, MA: MIT Press, 1991.
- [CGR91a] Gail A. Carpenter, Stephen Grossberg, and David B. Rosen. ART 2-A: an adaptive resonance algorithm for rapid category learning and recognition. *Neural Netw.*, 4(4):493–504, 1991.
- [CGR91b] Gail A. Carpenter, Stephen Grossberg, and David B. Rosen. Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks*, 4(6):759–771, 1991.
- [CHS04] Philipp Cimiano, Andreas Hotho, and Steffen Staab. Comparing conceptual, divisive and agglomerative clustering for learning taxonomies from text, 2004.
- [CLS01] K. Selcuk Candan, Huan Liu, and Reshma Suvarna. Resource description framework: metadata and its applications. *SIGKDD Explor. Newsl.*, 3(1):6–19, 2001.
- [CMBT02] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, 2002.
- [CMC05] Shih-Fu Chang, R. Manmatha, and Tat-Seng Chua. Combining text and audio-visual features in video indexing. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005 (ICASSP '05)*., pages 1005–1008, 2005.
- [Col96] Michael Collins. A new statistical parser based on bigram lexical dependencies. In *ACL*, pages 184–191, 1996.
- [Day02] Neil Day. Mpeg-7 applications, v 11.0. 2002.
- [DBdFF02] Pinar Duygulu, Kobus Barnard, Joao F. G. de Freitas, and David A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV '02*, pages 97–112, London, 2002.

REFERENCES

---

- [DDP02] Qin Ding, Qiang Ding, and William Perrizo. Association rule mining on remotely sensed images using p-trees. In *PAKDD '02: Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pages 66–79, London, UK, 2002. Springer-Verlag.
- [DGS99] Jochen Dörre, Peter Gerstl, and Roland Seiffert. Text mining: Finding nuggets in mountains of textual data. In *KDD*, pages 398–401, 1999.
- [DH04] Pinar Duygulu and Alexander G. Hauptmann. What’s news, what’s not? associating news videos with words. In *CIVR*, volume 3115 of *Lecture Notes in Computer Science*, pages 132–140. Springer, 2004.
- [DLR77] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.
- [DMH<sup>+</sup>00] Stefan Decker, Sergey Melnik, Frank Van Harmelen, Dieter Fensel, Michel Klein, Jeen Broekstra, Michael Erdmann, and Ian Horrocks. The semantic web: The roles of xml and rdf. *IEEE Internet Computing*, 04(5):63–74, 2000.
- [DOP03] Pinar Duygulu, Ozge Can Ozcanli, and Norman Papernick. *Comparison of Feature Sets Using Multimedia Translation*. Lecture Notes in Computer Science. 2869 edition, 2003.
- [DPW03] Ng D. Duygulu, P., N. Papernick, and H Wactlar. Linking visual and textual data on video. In *Proceedings of the Workshop on Multimedia Contents in Digital Libraries*, 2003.
- [DW03] Pinar Duygulu and Howard D. Wactlar. Associating video frames with text. In *Proceedings of the ACM SIGIR Conference*, 2003.
- [EVV03] Magdalini Eirinaki, Michalis Vazirgiannis, and Iraklis Varlamis. Sewep: using site semantics and a taxonomy to enhance the web personalization process. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 99–108, New York, NY, USA, 2003. ACM Press.

REFERENCES

---

- [FML04] Shaolei Feng, Raghavan Manmatha, and Victor Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *CVPR (2)*, pages 1002–1009, 2004.
- [FN06] Kateryna Falkovych and Frank Nack. Context aware guidance for multimedia authoring: harmonizing domain and discourse knowledge. *Multimedia Syst.*, 11(3):226–235, 2006.
- [FNvOR03] Kateryna Falkovych, Frank Nack, Jacco van Ossenbruggen, and Lloyd Rutledge. Sample: Towards a framework for system-supported multimedia authoring. *CWI Report: INS-E0302, ISSN 1386-3681*, 2003.
- [FNvOR04] Kateryna Falkovych, Frank Nack, Jacco van Ossenbruggen, and Lloyd Rutledge. Sample: Towards a framework for system-supported multimedia authoring. In *MMM '04: Proceedings of the 10th International Multimedia Modelling Conference*, page 362, Washington, DC, USA, 2004. IEEE Computer Society.
- [GBvOH03] Joost Geurts, Stefano Bocconi, Jacco van Ossenbruggen, and Lynda Hardman. Towards ontology-driven discourse: From semantic graphs to multimedia presentations. In *International Semantic Web Conference*, pages 597–612, 2003.
- [GG92] Sharon R. Garber and Mitch B. Grunes. The art of search: a study of art directors. In *CHI '92: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 157–163, New York, NY, USA, 1992. ACM Press.
- [GMV99] Nicola Guarino, Claudio Masolo, and Guido Vetere. Ontoseek: Content-based access to the web, 1999.
- [Gru93] Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowl. Acquis.*, 5(2):199–220, 1993.
- [Gru95] Thomas R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum.-Comput. Stud.*, 43(5-6):907–928, 1995.

REFERENCES

---

- [GW97] Bernhard Ganter and Rudolf Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1997. Translator-C. Franzke.
- [Han05] Jiawei Han. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.
- [Har98] Lynda Hardman. Modelling and authoring hypermedia documents. *PhD thesis, University of Amsterdam*, 1998.
- [Hea92] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics*, pages 539–545, Morristown, NJ, USA, 1992. Association for Computational Linguistics.
- [Hep06] Martin Hepp. Products and services ontologies: A methodology for deriving owl ontologies from industrial categorization standards. *Int. J. Semantic Web Inf. Syst.*, 2(1):72–99, 2006.
- [HF95] Jiawei Han and Yongjian Fu. Discovery of multiple-level association rules from large databases. In *VLDB*, pages 420–431, 1995.
- [HMWG98] Jochen Hipp, Andreas Myka, Rüdiger Wirth, and Ulrich Güntzer. A new algorithm for faster mining of generalized association rules. In *PKDD*, pages 74–82, 1998.
- [Hos98] Philipp Hoschka. Synchronized multimedia integration language (smil) 1.0 specification. 1998.
- [HPYM04] Jiawei Han, Jian Pei, Yiwen Yin, and Runying Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Min. Knowl. Discov.*, 8(1):53–87, 2004.
- [HSL<sup>+</sup>06] Jonathon S. Hare, Patrick A. S. Sinclair, Paul H. Lewis, Kirk Martinez, Peter G. B. Enser, and Christine J. Sandom. Bridging the semantic gap in multimedia information retrieval: Top-down and bottom-up approaches, 2006.

REFERENCES

---

- [HTT03] Ji He, Ah-Hwee Tan, and C. L. Tan. On machine learning methods for chinese document categorization. *Appl. Intell.*, 18(3):311–322, 2003.
- [Hun01] Jane Hunter. Adding multimedia to the semantic web - building an mpeg-7 ontology. In *International Semantic Web Working Symposium (SWWS)*, 2001.
- [Hun02] Jane Hunter. Combining the cidoc crm and mpeg-7 to describe multimedia in museums. In *Museums on the Web*, 2002.
- [Inc06] NetRatings Japan Inc. Google makes gains. *Nielsen/NetRatings NetView Audience Measurement Service Report*, 2006.
- [Ino04] Akihiro Inokuchi. Mining generalized substructures from a set of labeled graphs. In *ICDM*, pages 415–418, 2004.
- [JC97] Jay J. Jiang and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *International Conference Research on Computational Linguistics*, 1997.
- [JLM03] Jiwoon Jeon, Victor Lavrenko, and Raghavan Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *SIGIR '03*, pages 119–126, New York, NY, USA, 2003. ACM Press.
- [JT06a] Tao Jiang and Ah-Hwee Tan. Discovering image-text associations for cross-media web information fusion. In *PKDD*, pages 561–568, 2006.
- [JT06b] Tao Jiang and Ah-Hwee Tan. Mining rdf metadata for generalized association rules. In *DEXA*, pages 223–233, 2006.
- [JT06c] Tao Jiang and Ah-Hwee Tan. Mining rdf metadata for generalized association rules: knowledge discovery in the semantic web era. In *WWW*, pages 951–952, 2006.
- [JT07] Tao Jiang and Ah-Hwee Tan. Towards multimedia web information fusion: Learning image-text associations. *Submitted to IEEE Trans. Knowl. Data Eng.*, 2007.

REFERENCES

---

- [JTW07a] Tao Jiang, Ah-Hwee Tan, and Ke Wang. Mining generalized associations of semantic relations from textual web content. *IEEE Trans. Knowl. Data Eng.*, 19(2):164–179, 2007.
- [JTW07b] Tao Jiang, Ah-Hwee Tan, and Ke Wang. Mining generalized itemsets with over-generalization reduction. *Submitted to IEEE Trans. Knowl. Data Eng.*, 2007.
- [KAH<sup>+</sup>02] S. Kim, H. Alani, W. Hall, P. Lewis, D. Millard, N. Shadbolt, and M. Weal. Artequakt: Generating tailored biographies with automatically annotated fragments from the web. In *Workshop Semantic Authoring, Annotation and Knowledge Markup (15th European Conf. Artificial Intelligence)*, pages 1–6, 2002.
- [KAH<sup>+</sup>03] Sanghee Kim, Harith Alani, Wendy Hall, Paul H. Lewis, David E. Millard, Nigel R. Shadbolt, and Mark J. Weal. Automatic ontology-based knowledge extraction from web documents. *IEEE Intelligent Systems*, 18(1):14–21, 2003.
- [Koh01] Teuvo Kohonen. *Self-Organizing Maps*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2001.
- [KSB04] Anand Kumar, Barry Smith, and Christian Borgelt. Dependence relationships between gene ontology terms based on TIGR gene product annotations. In *Proc. 3rd Int. Workshop on Computational Terminology (CompuTerm 2004)*, pages 31–38, Geneva, Switzerland, 2004.
- [LC98] Claudia Leacock and Martin Chodorow. Combining lexical context and wordnet similarity for word sense identification. In Christiane Fellbaum, editor, *WordNet: An electronic lexical database*. MIT Press, 1998.
- [LDK00] Jean-Charles Lamirel, Jacques Ducloy, and Hager Kammoun. A self organizing map (som) extended model for information discovery in a digital library context. In *MDM/KDD*, pages 60–66, 2000.
- [LON05] Wenyuan Li, Kok-Leong Ong, and Wee Keong Ng. Visual terrain analysis of high-dimensional datasets. In Alípio Jorge, Luís Torgo, Pavel Brazdil, Rui Camacho, and João Gama, editors, *PKDD*, volume 3721 of *Lecture Notes in Computer Science*, pages 593–600. Springer, 2005.

REFERENCES

---

- [LS99] Ora Lassila and Ralph R. Swick. Resource description framework (rdf) model and syntax specification. 1999.
- [LW03] Jia Li and James Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(9):1075–1088, 2003.
- [Mal06] Rainer Malik. Conan : Text mining in the biomedical domain. *Doctoral thesis Utrecht University*, 2006.
- [Man98] Thomas Mandl. Vague transformations in information retrieval. In *ISI 98.*, pages 312–325, 1998.
- [Mar04] José M. Martínez. Mpeg-7 overview (version 10). 2004.
- [Mil95] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.
- [MMST05] Penelope Markellou, Ioanna Mousourouli, Sirmakessis Spiros, and Athanasios Tsakalidis. Using semantic web mining technologies for personalized e-learning experiences. In *WBE 2005: Web-based Education*, pages 461–826, Grindelwald, Switzerland, 2005. ACTA Press.
- [MPS02] Alexander Maedche, Viktor Pekar, and Steffen Staab. Ontology learning part one – on discovering taxonomic relations from the web, 2002.
- [MS00] Marjo Markkula and Eero Sormunen. End-user searching challenges indexing practices in the digital newspaper photo archive. *Inf. Retr.*, 1(4):259–285, 2000.
- [MS03] Ali Mustafa and Ishwar K. Sethi. Creating agents for locating images of specific categories. volume 5304, pages 170–178. SPIE, 2003.
- [Nel65] Theodor Holm Nelson. Complex information processing: a file structure for the complex, the changing and the indeterminate. In *Proceedings of the 1965 20th national conference*, pages 84–100, New York, NY, USA, 1965. ACM Press.
- [OD96] Douglas W. Oard and Bonnie J. Dorr. A survey of multilingual text retrieval. Technical report, College Park, MD, USA, 1996.



REFERENCES

---

- [PBS<sup>+</sup>06] Kosmas Petridis, Stephan Bloehdorn, Carsten Saathoff, Nikos Simou, Stamatia Dasiopoulou, Vassilis Tzouvaras, Siegfried Handschuh, Yannis Avrithis, Yiannis Kompatsiaris, and Steffen Staab. Knowledge representation and semantic annotation of multimedia content. *IEEE Proceedings on Vision, Image and Signal Processing - Special issue on the Integration of Knowledge, Semantics and Digital Media Technology*, 153(3):255–262, JUN 2006.
- [PBTL99a] Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lotfi Lakhal. Discovering frequent closed itemsets for association rules. In *ICDT*, pages 398–416, 1999.
- [PBTL99b] Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lotfi Lakhal. Discovering frequent closed itemsets for association rules. In *ICDT '99: Proceeding of the 7th International Conference on Database Theory*, pages 398–416, London, UK, 1999. Springer-Verlag.
- [Per98] Leonid I. Perlovsky. Conundrum of combinatorial complexity. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(6):666–670, 1998.
- [Per01] Fernando Pereira. Mpeg-7 requirements document, v 15. 2001.
- [PHM00] Jian Pei, Jiawei Han, and Runying Mao. Closet: An efficient algorithm for mining frequent closed itemsets. In *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 21–30, 2000.
- [PK06] Valery A. Petrushin and Latifur Khan. *Multimedia Data Mining and Knowledge Discovery*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
- [PLSH<sup>+</sup>99] Seungyup Paek, Carl L. Sable, Vasileios Hatzivassiloglou, Alejandro Jaimes, Barry H. Schiffman, Shih-Fu Chang, and Kathleen R. McKeown. Integration of visual and text based approaches for the content labeling and classification of photographs. In *ACM SIGIR'99 Workshop on Multimedia Indexing and Retrieval*, Berkeley, CA, August 1999.
- [Res99] Philip Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.

REFERENCES

---

- [RHB99] Lloyd Rutledge, Lynda Hardman, and Dick C. A. Bulterman. Grins: a graphical interface for smil. In *ACM Multimedia (2)*, page 200, 1999.
- [SA95] Ramakrishnan Srikant and Rakesh Agrawal. Mining generalized association rules. In Umeshwar Dayal, Peter M. D. Gray, and Shojiro Nishio, editors, *VLDB'95*, pages 407–419. Morgan Kaufmann, 1995.
- [SB96] Paraic Sheridan and Jean Paul Ballerini. Experiments in multilingual information retrieval using the spider system. In *SIGIR '96*, pages 58–65, New York, 1996. ACM Press.
- [SBC<sup>+</sup>02] Guus Schreiber, Inger I. Blok, Daan Carlier, Wouter P. C. van Gent, Jair Hokstam, and Uri Roos. A mini-experiment in semantic annotation. In *ISWC '02: Proceedings of the First International Semantic Web Conference on The Semantic Web*, pages 404–408, London, UK, 2002. Springer-Verlag.
- [Sch97] Peter Schauble. *Multimedia Information Retrieval: Content-Based Information Retrieval from Large Text and Audio Databases*. Kluwer Academic Publishers, Norwell, MA, USA, 1997.
- [SCS01] Ishwar K. Sethi, Ioana L. Coman, and Daniela Stan. Mining association rules between low-level image features and high-level concepts. volume 4384, pages 279–290. SPIE, 2001.
- [SDWW01] A. Th. (Guus) Schreiber, Barbara Dubbeldam, Jan Wielemaker, and Bob Wielinga. Ontology-based photo annotation. *IEEE Intelligent Systems*, 16(3):66–74, 2001.
- [Sha98] Mona Sharma. Performance evaluation of image segmentation and texture extraction methods in scene analysis. Master's thesis, 1998.
- [SHSG02] Arnold W.M. Smeulders, Lynda Hardman, Guus Schreiber, and Jan-Mark Geusebroek. An integrated multimedia approach to cultural heritage e-documents. In *ACM workshop on multimedia information retrieval*, 2002.
- [SM86] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.

REFERENCES

---

- [SON95] Ashok Savasere, Edward Omiecinski, and Shamkant B. Navathe. An efficient algorithm for mining association rules in large databases. In *VLDB*, pages 432–444, 1995.
- [Sow84] John F. Sowa. *Conceptual structures: information processing in mind and machine*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1984.
- [Sow91] John F. Sowa. *Principles of Semantic Networks: Explorations in the Representation of Knowledge*. Representation and Reasoning. Morgan Kaufmann, 1991.
- [Sow99] John F. Sowa. Conceptual graphs: Draft proposed american national standard. In *ICCS*, pages 1–65, 1999.
- [Sow00] John F. Sowa. *Knowledge representation: logical, philosophical and computational foundations*. Brooks/Cole Publishing Co., Pacific Grove, CA, USA, 2000.
- [SS01] Daniela Stan and Ishwar K. Sethi. Mapping low-level image features to semantic concepts. In *Proceedings of SPIE: storage and retrieval for media databases*, pages 172–179, 2001.
- [SS03] Daniela Stan and Ishwar K. Sethi. eid: a system for exploration of image databases. *Inf. Process. Manage.*, 39(3):335–361, 2003.
- [ST02] Kritsada Sriphaew and Thanaruk Theeramunkong. A new method for finding generalized frequent itemsets in generalized association rule mining. In *ISCC*, pages 1040–1045, 2002.
- [SVH04] Nuno Seco, Tony Veale, and Jer Hayes. An intrinsic information content metric for semantic similarity in wordnet. In *ECAI*, pages 1089–1090, 2004.
- [SWS<sup>+</sup>00] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, 2000.

REFERENCES

---

- [TAKG06] Ankur M. Teredesai, Muhammad A. Ahmad, Juveria Kanodia, and Roger S. Gaborski. Comma: a framework for integrated multimedia mining using multi-relational associations. *Knowl. Inf. Syst.*, 10(2):135–162, 2006.
- [Tan95] Ah-Hwee Tan. Adaptive resonance associative map. *Neural Netw.*, 8(3):437–446, 1995.
- [Tan99] Ah-Hwee Tan. Text mining: The state of the art and the challenges. In *Proc of the Pacif Asia Conf on Knowledge Discovery and Data Mining PAKDD'99 workshop on Knowledge Discovery from Advanced Databases*, pages 65–70, 1999.
- [Tan00] Robert Tansley. The multimedia thesaurus: Adding a semantic layer to multimedia information. *Ph. D. Thesis, Computer Science, University of Southampton*, August 2000.
- [TCG07] Ah-Hwee Tan, Gail A. Carpenter, and Stephen Grossberg. Intelligence through interaction: Towards a unified theory for learning. In *D. Liu et al. (Eds.): International Symposium on Neural Networks (ISNN) 2007*, volume 4491 of *Lecture Notes in Computer Science*, pages 1098–1107, 2007.
- [TNM03] Jelena Tesic, Shawn Newsam, and B. S. Manjunath. Mining image datasets using perceptual association rules. In *SIAM Sixth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Third SIAM International Conference (SDM)*, May 2003.
- [TPB99] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.
- [TWS05] Vincent S. Tseng, Ming-Hsiang Wang, and Ja-Hwung Su. A new method for image classification by using multilevel association rules. In *ICDEW '05: Proceedings of the 21st International Conference on Data Engineering Workshops*, page 1180, Washington, DC, USA, 2005. IEEE Computer Society.
- [UES+01] Yusuke Uehara, Susumu Endo, Shuichi Shiitani, Daiki Masumoto, and Shigemi Nagata. A computer-aided visual exploration system for knowledge discovery from images. In *MDM/KDD*, pages 102–109, 2001.

REFERENCES

---

- [vRJMB93] Guido van Rossum, Jack Jansen, K. Sjoerd Mullender, and Dick C. A. Bulterman. Cmifed: a presentation environment for portable hypermedia documents. In *MULTIMEDIA '93: Proceedings of the first ACM international conference on Multimedia*, pages 183–188, New York, NY, USA, 1993. ACM Press.
- [W3C] W3C. W3c RDF Schema Specification.
- [WHP03] Jianyong Wang, Jiawei Han, and Jian Pei. Closet+: searching for the best strategies for mining frequent closed itemsets. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 236–245, New York, NY, USA, 2003. ACM Press.
- [WSKR03] Kevin Wilkinson, Craig Sayers, Harumi A. Kuno, and Dave Reynolds. Efficient rdf storage and retrieval in jena2. In *SWDB*, pages 131–150, 2003.
- [XZ06] Xiao-Bing Xue and Zhi-Hua Zhou. Distributional features for text categorization. In *ECML*, pages 497–508, 2006.
- [yGGLL02] Manuel Montes y Gómez, Alexander F. Gelbukh, and Aurelio López-López. Text mining at detail level using conceptual graphs. In *ICCS '02*, pages 122–136, London, UK, 2002. Springer-Verlag.
- [YW95] Hong H. Yu and Wayne H. Wolf. Scenic classification methods for image and video databases. volume 2606, pages 363–371. SPIE, 1995.
- [Zak00] Mohammed Javeed Zaki. Generating non-redundant association rules. In *Knowledge Discovery and Data Mining*, pages 34–43, 2000.
- [ZH02] Mohammed Javeed Zaki and Ching-Jiu Hsiao. Charm: An efficient algorithm for closed itemset mining. In Robert L. Grossman, Jiawei Han, Vipin Kumar, Heikki Mannila, and Rajeev Motwani, editors, *SDM*. SIAM, 2002.
- [ZHL<sup>+</sup>98] Osmar R. Zaïane, Jiawei Han, Ze-Nian Li, Sonny Han Seng Chee, and Jenny Chiang. Multimediaminer: A system prototype for multimedia data mining. In Laura M. Haas and Ashutosh Tiwary, editors, *SIGMOD 1998, Proceedings ACM SIGMOD International Conference on Management of Data, June 2-4, 1998, Seattle, Washington, USA*, pages 581–583. ACM Press, 1998.

## REFERENCES

---

- [ZHLH98] Osmar R. Zaiane, Jiawei Han, Ze-Nian Li, and Jean Hou. Mining multimedia data. In *CASCON '98: Proceedings of the 1998 conference of the Centre for Advanced Studies on Collaborative research*, page 24. IBM Press, 1998.
- [ZH00] Osmar R. Zaiane, Jiawei Han, and Hua Zhu. Mining recurrent items in multimedia with progressive resolution refinement. *icde*, 00:461, 2000.
- [ZZ06] Zhi-Hua Zhou and Min-Ling Zhang. Multi-instance multi-label learning with application to scene classification. In *NIPS*, pages 1609–1616, 2006.
- [ZZ07] Min-Ling Zhang and Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, 2007.

# Index

- Adaptive Resonance Theory (ART), 39
- Association Rule Mining (ARM), 98
  
- Closed Itemset, 106
- Conceptual Graph, 71
  
- Extensible Markup Language (XML), 21
  
- Generalization Closure, 106
- Generalized Association Rule Mining (GARM),  
99
- GP-Close, 107
  
- Image-Text Association, 37
- Implicit Association, 38
  
- MPEG-7, 22
  
- Ontology, 21
- Over-Generalization, 104
- Over-Generalization Reduction, 105
  
- RDF Schema (RDFS), 21
- Resource Description Framework (RDF),  
21
  
- Semantic Networks, 72
- Semantic Web, 20
  
- Universal Resource Identifier (URI), 20
  
- Vague Transformation, 42