EFFECTIVE MULTIMEDIA DOCUMENT REPRESENTATIONS FOR KNOWLEDGE DISCOVERY

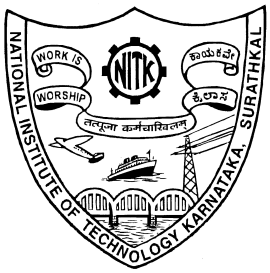
Thesis

Submitted in partial fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

by

PUSHPALATHA K



DEPARTMENT OF INFORMATION TECHNOLOGY
NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA,
SURATHKAL, MANGALORE - 575025

JULY, 2017

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA, SURATHKAL

DECLARATION

By the Ph.D. Research Scholar

I hereby declare that the Research Thesis entitled **EFFECTIVE MULTIMEDIA DOCUMENT REPRESENTATIONS FOR KNOWLEDGE DISCOVERY** which is being submitted to the **National Institute of Technology Karnataka, Surathkal** in partial fulfilment of the requirements for the award of the Degree of **Doctor of Philosophy** in **Information Technology** is a *bonafide report of the research work carried out by me*. The material contained in this Research Thesis has not been submitted to any University or Institution for the award of any degree.

IT11F01, Pushpalatha K

(Register Number, Name & Signature of the Research Scholar)

Department of Information Technology

Place: NITK, Surathkal.

Date:

CERTIFICATE

This is to *certify* that the Research Thesis entitled **EFFECTIVE MULTIMEDIA DOCUMENT REPRESENTATIONS FOR KNOWLEDGE DISCOVERY** submitted by **PUSHPALATHA K**, (Register Number: IT11F01) as the record of the research work carried out by her, is *accepted as the Research Thesis submission* in partial fulfillment of the requirements for the award of degree of **Doctor of Philosophy**.

Prof. Ananthanarayana V.S.
Research Guide

Prof. G. Ram Mohana Reddy
Chairman - DRPC

ACKNOWLEDGEMENTS

It is a genuine pleasure to express my deep felt gratitude to all the people who have made this research work possible.

I would like to express my deep gratitude towards my research guide **Prof. Ananthanarayana V S**, Professor of Information Technology Department. Professor Ananthanarayana V S has been a great motivator for me to achieve the goal with dedication and clear vision. I thank him for his guidance, suggestions and kind hearted support throughout my research work.

I express heartfelt thanks to my research progress assessment committee members **Dr. P. Santhi Thilagam**, Department of Computer Science and Engineering, National Institute of Technology Karnataka Surathkal and **Dr. B. R. Shankar**, Department of MACS, National Institute of Technology Karnataka Surathkal, for their valuable suggestions and constant encouragement to improve the research work.

I am thankful to the current head of the Department of Information Technology, **Prof. G. Ram Mohana Reddy** and the former head of the department, **Prof. Ananthanarayana V S**, who provided an independent working environment with all computing facilities required to carry out the research work.

I am indebted to all the teaching staff of Information Technology Department and my special thanks to **Dr. Sowmya Kamath S** and **Dr. Geetha V** for their support and encouragement.

I would like to express my sincere thanks to **Mrs. Saumya Hegde** and **Mrs. Sumith Nireshwalya** for their academic and personal guidance and support. I am thankful to all my research colleagues for creating a cordial working environment.

I am also very grateful to all my former teachers especially **Mr. Harisha Shetty** who have initiated my journey of knowledge and learning.

I would like to express my sincere thanks to all technical and administrative staff of the Information Technology Department for their help during my research work. I am grateful to all those people, who supported me in any kind during the completion of my research work.

I humbly thank my family for their constant support, understanding and patience throughout my research work. They have stood by me at all times and encouraged me to pursue my dreams.

As I complete my research, I know in my heart that my father would have been proud of my achievement and has left behind his blessings, which will stay with me forever.

Above all, I owe it all to Almighty God for granting me the wisdom, health and strength to undertake this research work and enabling me to its completion.

Place: NITK, Surathkal Pushpalatha K

Date:

ABSTRACT

In recent years, the rapid advances in multimedia technology have led to grow the multimedia documents explosively. In order to utilize the multimodal information of multimedia documents, sophisticated knowledge discovery systems are required. The knowledge discovery systems require efficient multimedia mining methods to extract the meaningful and useful information from the huge volume of multimedia documents. The success of multimedia mining relies on the representation of multimedia documents and its multimodal contents. The appropriate representation of multimedia documents discovers the useful patterns that can be used to assist the multimedia mining methods in discovering the useful knowledge. The multimodal nature of multimedia objects is the challenging problem for the multimedia document representation, as the features of multimodal objects are in different space with different characteristics and dimensionalities. Representation of multimodal multimedia objects in a unified feature space helps the multimedia document representation and multimedia mining methods. The research work in this thesis proposes the multimedia data representation methods, multimedia document representations, and multimedia mining methods for the effective knowledge discovery in multimedia documents.

In the first methodology, this thesis aims at the representation of multimodal multimedia objects in a unified feature space. We propose two multimedia data representation methods, Multimedia To Signal Conversion (MSC) and Multimedia to Image Conversion (MIC) to represent the multimedia objects in a unified domain. The MSC represents the multimedia objects in frequency domain by converting the multimedia objects as signal objects. The MIC converts the multimedia objects as image objects to represent them in spatial domain. The multimedia objects in unified domain are represented in the unified feature space using the features with similar dimensions and characteristics. Hence, both the multimedia data representation methods convert the

multimodal multimedia documents as unified multimedia documents. The unified multimedia documents ease the representation of multimedia documents and improve the efficiency of multimedia mining methods. The proposed multimedia data representation methods are effectively used for knowledge extraction from multimedia documents.

In the second methodology, this thesis presents the two multimedia document representations, Multimedia Suffix Tree Document (MSTD) and Multimedia Feature Pattern Tree (MFPT) to represent the unified multimedia documents. The MSTD represents the unified multimedia documents based on shared similar multimedia objects among the documents. The similarity between the multimedia objects depends on the similarity of the features. The MFPT represents the documents based on shared similar feature patterns of the multimedia objects. Both the representations are compact and provide the complete information of the documents. They function as the platform for the multimedia knowledge extraction methods.

In the third methodology, this thesis explores the multimedia mining methods based on the MSTD and MFPT representations. The MSTD and MFPT based classification algorithms effectively classifies the multimedia documents. The multimedia documents are partitioned into clusters of same multimedia concepts using the MSTD and MFPT based clustering algorithms. The MSTD representation extracts the frequent multimedia patterns to generate the multimedia class association rules for classifying the multimedia documents. The MFPT representation extracts the sequential multimedia feature patterns to derive the multimedia class sequential rules that support the classification of multimedia documents based on the object characteristics.

The efficacy of the proposed methods is evaluated by conducting the experiments with four datasets of multimodal multimedia documents. Experimental results demonstrate that the proposed multimedia data representation methods benefit the multimedia document representation and multimedia mining methods by representing the multimodal multimedia objects

in a unified feature space. The proposed multimedia document representations are effectively used to enhance the performance of multimedia mining methods in discovering the knowledge from multimedia documents.

Keywords: Knowledge discovery, Multimedia documents, Multimedia document representation, Multimedia data representation, Multimedia mining, Classification, Clustering, Frequent multimedia patterns, Multimedia class association rules, Sequential multimedia feature patterns, Multimedia class sequential rules

Table of Contents

List of Figures

List of Tables

Abbreviations

AFF	Attention Fusion Function
ARM	Association Rule Mining
ASM	Acoustic Segment Mode
ARM	Adaptive Resonance Theory
AVA	Audio-Visual Atoms
AVG	Audio-Visual Grouplet
CAR	Class Association Rule
CBR	Content Based Retrieval
CCA	Canonical Correlation Analysis
CFA	Cross-Modal Factor Analysis
CRF	Conditional Random Field
CSR	Class Sequential Rule
CWT	Continuous Wavelet Transform
DCCA	Deep Canonical Correlation Analysis
DF	Document Frequency
DFS	Depth First Search
DWT	Discrete Wavelet Transform
FIG	Feature Interaction Graph
FMP	Frequent Multimedia Pattern
GA	Glowworm Agent
GLF	Gradient-Descent-Optimization Linear Fusion

GSO	Glowworm Swarm Optimization
GSOMDC	Glowworm Swarm Optimization Multimedia Document Clustering
HMM	Hidden Markov Models
HSV	Hue, Saturation Value
IDF	Inverse Document Frequency
ISMD	Information Theory Based Similarity Measure for Multimedia Documents
KDMD	Knowledge Discovery In Multimedia Documents
LDA	Latent Dirichlet Analysis
LMOS	Laplacian Media Object Space
LPC	Linear Predictive Coefficients
LSA	Latent Semantic Analysis
LSI	Latent Semantic Indexing
MCAR	Multimedia Class Association Rule
MFCC	Mel Frequency Cepstral Coefficients
MFPT	Multimedia Feature Pattern Tree
MIC	Multimedia to Image Conversion
MIL	Multiple Instance Learning
MIR	Multimedia Information Retrieval
MMCS	Multimedia Correlation Space
MMD	Multimedia Document
MMDSG	MMD Semantic Graph
MMDSSG	MMD Semi Semantic Graph
MMG	Mixed Media Graph
MMSAR	Multi-Modal Semantic Association Rule
MSC	Multimedia to Signal Conversion
MST	Multimedia Suffix Tree
MSTD	Multimedia Suffix Tree Document

NLF	Nonlinear Fusion
NMF	Non-Negative Matrix Factorization
OMDSL	Online Multimodal Deep Similarity Learning
PD	Percentage Difference
PLSA	Probabilistic Latent Semantic Analysis
RMS	Root Mean Square
SCAE	Stacked Contractive Auto Encoders
SFFS	Sequential Forward Floating Search
SMFP	Sequential Multimedia Feature Pattern
SODA	Shrinkage Optimized Directed Information Assessment
SPM	Sequential Pattern Mining
STC	Suffix Tree Clustering
STD	Suffix Tree Document
SVM	Support Vector Machines
SVR	Support Vector Regression
TDNN	Time Delay Neural Network
TF	Term Frequency
TSP	Time-Constrained Sequential Pattern
TTS	Text To Speech
UCCG	Uniform Cross-Media Correlation Graph
UMD	Unified Multimedia Document
VSD	Vector Space Document

Chapter 1

Introduction

In recent years, due to the rapid advances in multimedia technology, multimedia objects such as texts, audios, images, videos, graphics and animations have been growing at an unprecedented speed. The vast amount of multimedia objects contains the rich information that can be used in various applications such as education, entertainment, information retrieval, security and medicine. This knowledge enriched multimedia objects cannot be represented in the traditional database structures as they are multimodal, unstructured and unordered. Multimedia objects are defined in terms of the media types they belong. Multimedia objects are diverse with different characteristics which are perceived, stored and processed in various modes. In general, the information about a multimedia concept is embedded in multiple forms of multimedia objects. Multimedia document (MMD) encloses the information of a semantically related multimedia concept in multimodal forms (Yoshitaka and Ichikawa, 1999). It is multimodal in nature due to its multimodal contents associated with unique multimodal features. Each MMD is labeled by the concept of its multimedia contents and identified by the unique identifier. For example, the MMD that has the information of tiger in the form of images, audios and text documents is labeled as 'Tiger'. Similarly the MMD 'Guitar' has multimodal information about the concept 'Guitar'. MMD can be in many forms such as a single image, a single video, a single audio, an image with caption or the collection of images, audios, texts and videos. Our research work focuses on the MMDs that are the collection of media objects images, audios and texts related to same multimedia

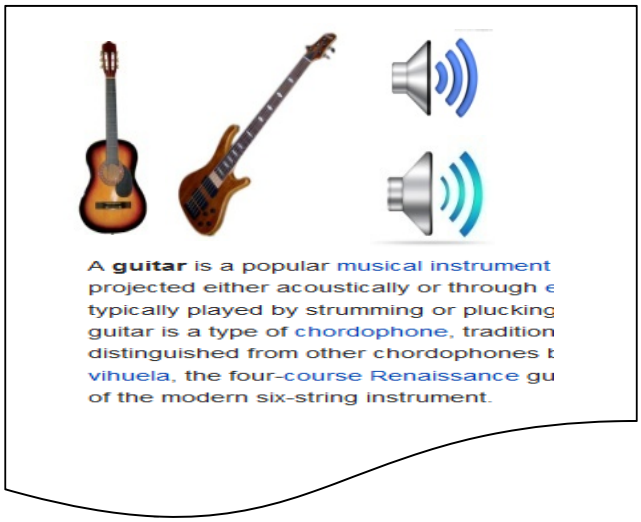


Figure 1.1: MMD 'Guitar', enclosing the media objects of guitar

concept. Figure 1.1 shows the example of MMD of 'Guitar' which has guitar images, guitar sounds and textual information of guitar.

With the vast amount of MMDs, manual analyzing will be inefficient, time consuming and tedious. The existing knowledge extraction methods are basically applied to unimodal documents such as texts, images and audios. They require additional processing to manage the MMDs. In order to organize, analyze and discover hidden knowledge of the multimedia documents, the sophisticated multimedia knowledge discovery methods are required.

1.1 Knowledge Discovery in Multimedia Documents

Knowledge discovery in multimedia documents (KDMD) discovers the hidden, understandable and useful knowledge in a massive amount of multimedia documents. The main components of KDMD process include multimedia document selection, multimedia document processing, multimedia document mining, multimedia pattern evaluation and multimedia knowledge interpretation.

Multimedia document selection process selects the MMDs relevant to the analysis task using the prior knowledge of the concept. For example, the knowledge about a sport is necessary while mining any multimedia documents related to sports. The multimedia document processing prepares the MMDs for knowledge extraction. It processes the MMDs and represents them in a form to assist the multimedia mining process. Multimedia mining also known as multimedia pattern discovery is the most essential step of KDMD process, discovers the interesting multimedia patterns using the intelligent techniques. The multimedia patterns are evaluated using the pattern evaluation process. Finally, the extracted patterns are presented in a more convenient and understandable form using the knowledge representation methods. Although the useful knowledge are discovered by multimedia mining methods, the success of multimedia mining relies on the processing of MMDs. Hence, the multimedia document processing and the multimedia mining are the two essential components of KDMD process.

1.2 Multimedia Document Processing

The main task of multimedia document processing is to process and represent the MMDs for the discovery of knowledge. Multimedia document representation portrays the multimedia documents in an appropriate representation based on the representation of multimodal multimedia objects. Multimedia data representation represents the multimedia objects in an abstract feature space by discovering the important features of multimedia objects. So, the multimedia data representation and the multimedia document representation are the important processes of multimedia document processing.

1.2.1 Multimedia Data Representation

Multimedia data representation is a very important component of knowledge discovery process, as to a great degree the success of KDMD process relies on the representations of data features. It represents the multimedia objects as useful information that would facilitate the multimedia mining process. It comprises of cleaning, transformation and feature extraction methods. The cleaning process reduces the unwanted noise from the multimedia data. The transformation process transforms the multimedia data into a form suitable for feature extraction. The transformed data are represented in an abstract feature space using the feature extraction process.

The representation of multimedia object depends on its modality. The multimedia objects of different modality require different cleaning, transformation and feature extraction methods for each modality of data. The multimedia documents with multimodal objects is the big challenge for KDMD process as it has different feature spaces with different characteristics and dimensionalities. During the last few years, in order to analyze the MMDs, some approaches employed one modality object, neglecting the other modalities. The main issue in these approaches is the selection of a particular modality of data. Although the multimedia objects of an MMD are from different information sources, they semantically complement each other. With the availability of universally understandable multimodal multimedia data, their effective use will be beneficial for many applications. Moreover, combining the outputs of several information sources may reduce the risk of an adverse selection of a poorly performing information source. Also, the information from multimodal data provides more information than a unimodal data. Even though the multimedia objects such as texts, images and audios are individually ambiguous, their integration will reduce the ambiguity. For example, the word *Tiger* may be the person's name or an animal or airways name. However, the integration of word *Tiger* and tiger image confirms that the concept is about tiger. Similarly, the visual and audio information provides more information to detect and

distinguish tiger from cat. In many applications one modality of data may coexist with other modalities. Therefore, the use of multimodal multimedia data will be advantageous in increasing the efficiency of KDMD process.

The multimedia objects in different feature space complicate the process of multimedia document representation. The improper representation of MMDs affects the efficiency of multimedia mining tasks. In recent years, the multimodal multimedia objects have been represented by integrating them using various multimedia fusion approaches. The multimedia fusion approaches are broadly classified as feature level fusion and decision level fusion (Atrey et al., 2010). The feature level fusion also known as early fusion, integrates the features of different modality of objects before the main processing is performed. The features are combined either by concatenation or mapping in a sub space. Early fusion approaches preserve the multimodal correlations and require only one classifier. The decision or late fusion achieves the final result by integrating the results obtained by processing the each modality of objects independently. It has the advantage of selecting the suitable learning approach for each modality. Some of the works employed the combination of these two fusion approaches to integrate the multimedia objects to make use of the benefits of both the methods.

1.2.2 Multimedia Document Representation

The multimedia document representation describes the approach of extracting meaningful patterns from the MMDs. It represents the MMDs in a more suitable representation to assist discovering the useful patterns in the mining process. The proper representation reduces the search time and memory space requirements, thereby benefiting the overall knowledge discovery process. Also, it discovers the useful patterns embedded in the documents to assist the multimedia mining tasks. Generally, the document information retrieval methods rely on the Vector Space Document (VSD) model (Salton et al., 1975) that represents the documents as a feature vector of words. The main drawback of VSD model is the VSD based knowledge extraction methods make use of single words which gives the vague

information. In contrast to VSD model, an n-gram based tree model known as Suffix Tree Document (STD) model has been used (Zamir and Etzioni, 1998) to represent the documents. Suffix tree is a rooted tree having internal nodes that represent the common n-grams shared by the documents. The documents that share more internal nodes are considered as more similar. Both the VSD and STD models are used to represent the text documents based on the number of similar words between them. They have limitation in managing the MMDs due to their multimodal nature.

1.3 Multimedia Mining

Multimedia mining can be defined as the analysis of large amounts of multimodal multimedia documents to discover the useful patterns that cannot be accessed by normal queries. It incorporates the strength of knowledge enriched multimedia data and intelligent mining techniques to extract the implicit meaningful patterns from multimedia documents to provide the useful knowledge for various applications. The unstructured and multimodal nature of multimedia data makes the multimedia mining process more complex compare to traditional business data. However, the objectives of the many multimedia applications are more successfully achieved with multiple modalities than the single modalities alone. The most common multimedia mining techniques that utilize the multimodal multimedia data are classification, clustering, frequent pattern mining, and sequential patterns mining.

Multimedia Classification

Classification is the process of discovering the predictive learning function to distinguish the concepts of data classes in order to predict the class of unknown data (Han et al., 2011). It is a supervised technique of mapping the target data to the predefined groups or classes. It has two phases. The first phase is the

training phase that builds a classifier from the existing data to describe their associated class labels. The testing phase classifies the unknown data based on the learned model. The most widely used classification approaches are support vector machines, k-nearest-neighbors, decision tree and Bayesian classifier. The other popular classification methods include neural networks, Bayesian belief networks, fuzzy logic and genetic algorithms etc. For a good classification, reduction of the noisy data and irrelevant attributes is necessary. The efficiency of the classification relies on the representation of data and the selection of distance measure. Some of the multimedia applications soccer goal detection, commercial detection, spam images detection, auto annotation of multimedia objects, news story segmentation, etc. are well classified with multiple modalities compared to the single modalities. The success of multimedia classification depends on the selection of appropriate classifier for multimedia objects.

Multimedia Clustering

Clustering is the process of dividing a group of data into a number of clusters of similar data. It is an unsupervised approach that does not use any predefined classes and forms the clusters based on self similarity. Clusters are formed based on the rule such that the data within a cluster must be highly similar and very dissimilar to data of other clusters. Some of the popular applications of multimedia clustering include social event detection, image clustering, video clustering, multimodal news story clustering, web document clustering, clustering of speakers etc. Multimedia clustering is used to learn the correlations among the multimedia data.

Mining Frequent Multimedia Patterns and Multimedia Associations

Frequent pattern mining (Agrawal et al., 1993) is a method of finding relationships among the items in a database. The extraction of frequent patterns helps in mining the interesting relations among the data. Association rule mining (Agrawal et al., 1993) generates the descriptive rules using the frequent patterns that discover the relationships among the set of items in a dataset. The usefulness of rules can be extended to develop the association rule based classification model (Ma, 1998). Association rule mining is used to find the multimodal correlations in many multimedia applications such as event detection in videos, image retrieval, feature detection, concept mining and medical image diagnosis.

Mining Multimedia Sequential Patterns and Sequential Rules

Sequential pattern mining (SPM) is a method of extracting the sequential patterns, itemsets or sub-sequences (Agrawal and Srikant, 1995). In SPM, the sequences maintain the order of itemsets. Sequential patterns are used to generate the sequential rules (Liu, 2007). With the set of data sequences, the sequence rules are termed as class sequential rule (CSR) as each data sequence is labeled by the class label. The CSRs are used to classify the unknown data sequences by learning the labeled data sequences. The SPM is used for the classification of images and audios by discovering the sequential patterns of the media objects.

1.4 Multimedia Information Retrieval

Multimedia information retrieval (MIR) brings the multimodal objects together by matching the multimedia content and thus by satisfying the user's needs. A

multimedia search and retrieval engine should allow the users to express their query in any suitable form and retrieve the information in any suitable forms. Also, the users should be able to get the complete view of the retrieved information.

Prior to the introduction of content based media retrieval, the text based searching has been used to retrieve the annotated media. With the evolution of vast multimodal multimedia data, the use of traditional text based search has been considered inefficient. The manual process of annotating media is complicated, timely and costlier. Some annotation failed to describe the media where as some multimedia data cannot be easily described in words. Hence, the multimodal nature of multimedia data has raised a demand for a powerful content based retrieval system. The content based retrieval system performs searching based on the content of media by avoiding the manual annotation. The content based retrieval methods allow the user to submit media object as query and gets the similar media objects as result. However, in content based retrieval methods, the modality of the query object and returned results are same. In order to exploit the multimodal correlations of multimedia objects, a powerful content based cross media retrieval approach has been emerged. In cross media retrieval, the modality of query object and retrieved results need not be same. Cross media retrieval is helpful as an auxiliary tool when the traditional content based multimedia retrieval is insufficient. For example, with the image of a bird it is possible to retrieve the text description or the chirping sound of that bird without knowing the name of the bird. However, the cross media retrieval methods are literally unimodal as they manage query and results of single modality. The necessity of utilizing multimodal objects both in queries and results had driven a way to the emergence of multimodal retrieval systems. Multimodal retrieval methods allow the multimodal queries and retrieve the multiple media objects simultaneously. The multimodal retrieval accepts the multimodal MMD as query and retrieves the similar multimodal MMDs.

1.5 Motivation

With the availability of multimodal information, many multimedia areas such as education, medicine, entertainment, security and information retrieval require the effective utilization of multimedia information for the successful achievement of their applications. Multimodal information necessitates sophisticated multimedia mining methods. The efficiency of the multimedia mining methods relies on the effective representation of multimedia documents which in turn relies on the representation of multimedia objects. The existing document representations are not suitable for MMDs due to the multimedia objects. Moreover, the divergent features of multimedia objects are the main hurdle for multimedia document representation and multimedia mining. Therefore, the multimodal data representation and multimedia documents representation are the two main challenges for knowledge discovery in multimedia documents. Our research work focuses on the development of effective multimedia data representation methods and multimedia document representations for the success of KDMD process. The objectives of the research work are motivated by following factors.

Multimedia Data Representation

Generally, the multimodal multimedia objects are represented in multimodal feature space. The multimodal features complicate the process of multimedia document representation and multimedia mining. The unified representation of multimedia objects use the same feature extraction methods resulting in the features with same dimensionalities and similar characteristics. The features in a unified feature space facilitate the multimedia document representation and multimedia mining methods thereby improving the efficiency of the multimedia knowledge extraction.

Multimedia Document Representation

The appropriate representation of MMDs discovers the useful patterns of the documents and preserves the multimodal associations which are useful for multimedia mining methods. Representing the set of MMDs in a single structure helps in reducing the search time and memory requirement of multimedia mining methods thereby improving the efficiency of overall decision making process. As per our knowledge there are no efforts made for the unified representation of multimedia documents with respect to multimodal objects and their features.

Multimedia Mining

With the vast volume of MMDs, the needs and expectations regarding the multimedia knowledge extraction has imposed a high demand for the efficient multimedia mining methods. The traditional data mining methods has been used for well structured, well defined and non-ambiguous data. With MMDs, the mining methods have to deal with the unstructured and heterogeneous multimedia objects and their correlations. Hence, there is a requirement of sophisticated multimedia mining methods to extract the explicit hidden knowledge from MMDs.

1.6 Thesis Contributions

The main contributions of this dissertation are as follows:

- Developing the multimedia data representation methods to represent the multimodal multimedia objects in a unified feature space in order to aid the multimedia document representation and multimedia document mining.
- Developing multimedia document representations based on object similarity to discover the multimedia patterns in order to benefit the multimedia document mining tasks thereby improving the efficiency of the KDMD process.

- Developing the multimedia document mining methods to discover the precise patterns from the dataset of MMDs that provide the useful knowledge for many applications.

1.7 Outline of the Thesis

Figure 1.2 shows the organization of the thesis. The remainder of this dissertation is organized as follows:

Chapter 2 presents the review of existing multimedia data representation methods, multimedia document representations, and multimedia mining methods. Followed by literature review, we highlight the open research issues in the existing multimedia data representation methods, multimedia document representations and multimedia mining. Then, the problem statement, objectives, scope and the research contributions based on the identified research issues are presented. Further, we briefly describe the multimedia document datasets used for the experiments of our research work.

In chapter 3, we discuss the multimedia data representation methods Multimedia to Signal Conversion (MSC) and Multimedia to Image Conversion (MIC) for the representation of multimodal multimedia objects in a unified feature space. The proposed methods are evaluated for the classification and retrieval of multimedia documents. Also, we describe the bio inspired glowworm swarm optimization based clustering algorithm for clustering the MMDs and the information theory based similarity measure to find the similarity between the MMDs. The validations of the proposed methods are performed by the experimental analysis with the MMD datasets.

Chapter 4 presents the Multimedia Suffix Tree Document Representation (MSTD) for representing the MMDs. The MSTD representation is evaluated by the MSTD based multimedia mining methods for the knowledge discovery from MMDs. The performance evaluation of the MSTD representation and MSTD based multimedia mining methods are presented using the experimental results.

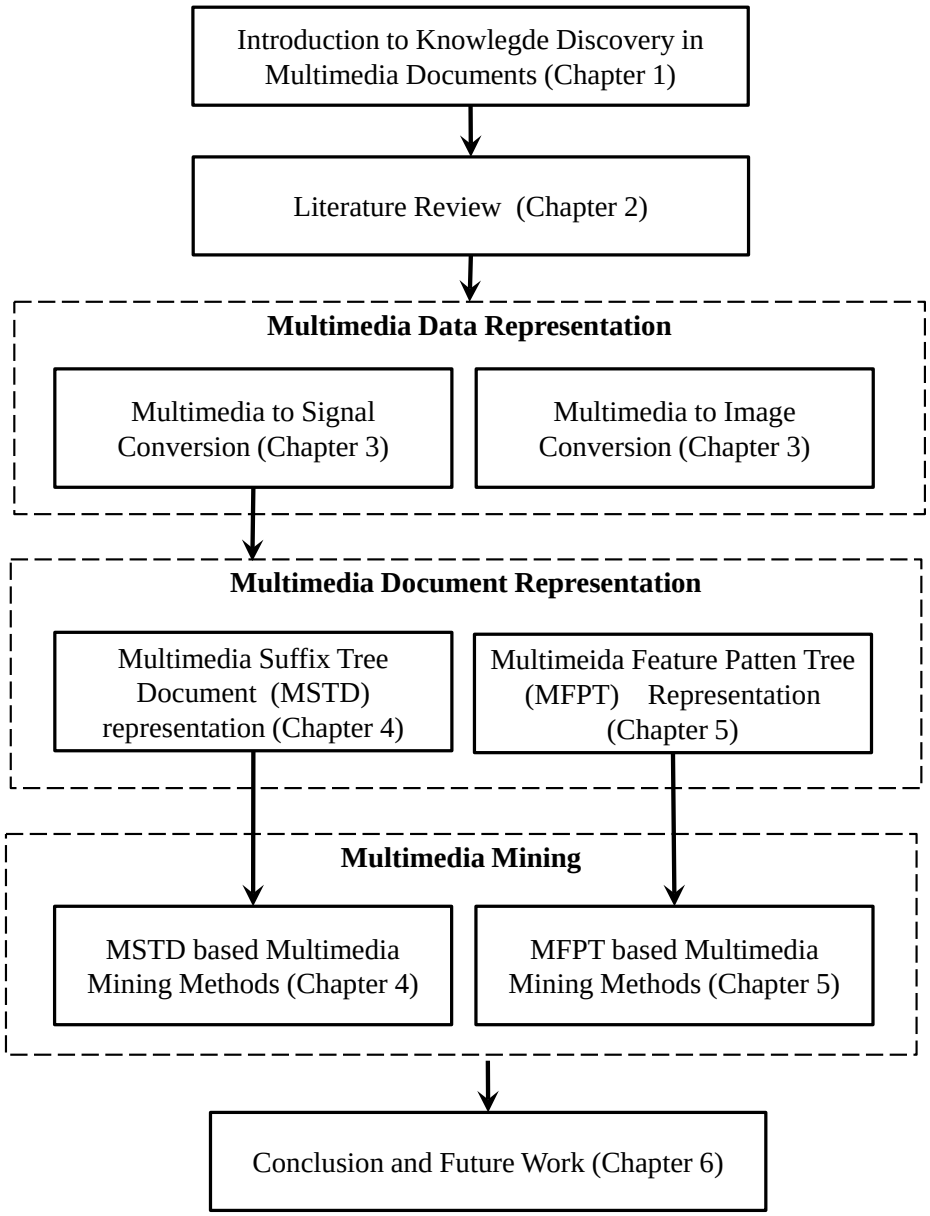


Figure 1.2: Organization of the thesis

In chapter 5, we describe the Multimedia Feature Pattern Tree (MFPT) representation for the representation of the MMDs. To evaluate the MFPT representation, MFPT based multimedia mining methods are proposed. Experimental analysis demonstrates the efficacy of the MFPT representation and MFPT based multimedia mining methods.

Finally, Chapter 6 presents the summary of the research contributions and highlights the possible future directions.

1.8 Summary

This chapter introduced the two main steps of the KDMD process: multimedia document processing and multimedia document mining. The various methods of multimedia data representation, multimedia document representation, and multimedia mining are briefly explained. The challenges and motivations for the development of multimedia data representation, multimedia document representation and multimedia mining methods are discussed. This chapter also gives the detailed organization of the thesis.

Chapter 2

Literature Review

The multimedia knowledge discovery system uses the multimedia mining methods to discover the meaningful patterns from MMDs. The success of the multimedia mining methods depends on the representation of MMDs. The appropriate representation of multimedia objects helps the representation of multimedia documents. This chapter presents the review of existing multimedia data representation, multimedia document representation and multimedia mining methods. Further, this chapter gives the problem statement, objectives, scope, and contributions of our research work. Also, the brief description of the multimedia datasets used for the experiments of our research work is presented in this chapter.

2.1 Multimedia Data Representation

Multimedia data representation processes and represents the multimodal multimedia objects for the effective utilization of multimodal information. The data representation involves two processes; multimedia data processing and multimedia fusion. Using the multimedia data processing procedure, the multimedia objects are processed, features are extracted and represented in a multimodal feature space. Multimedia fusion integrates the multimodal features of multimedia objects to utilize the multimodal information effectively.

2.1.1 Multimedia Data Processing

The representation of multimedia objects depends on their modality. The most common and basic types of multimedia data are text, audio and image. The multimedia objects require different cleaning, transformation and feature extraction methods for different modality. Hence, the representation of multimedia objects is multimodal due to the different characteristics and dimensionalities of the features. Following section briefly describes the processing of multimedia objects.

Texts

The text document is prepared for mining by splitting into words using the tokenization process. The words are refined by filtering the stop words, lemmatization and stemming (Porter, 1980). The obtained words describe the concept of given text document.

The feature extraction process for texts is extracting the important words that describe the concept of the text document. The importance of words is computed based on the characteristics of words such as term frequency(TF), document frequency(DF), inverse document frequency(IDF), tf-idf, entropy based weighting, term contribution etc. The term frequency is the number of occurrences of a word in a document. The number of documents that contain a word is termed as document frequency. IDF measures the importance of the word in a document. The tf-idf measures the importance of a word for a document in the document corpus. Entropy based weighting (Dumais, 1991) provides the quality of a word which is measured by the reduction of entropy value when the word is removed.

Audios

Audio is a mixture of silences, music, speech and noise. Audios are processed by segmenting the data of fixed window size or at word level or at the phoneme.

The phoneme level segmentation best suits for sounds. In some applications, initially the noise, silences and music are removed before further processing (Pfeiffer et al., 1997; Saraceno and Leonardi, 1997; Saunders, 1996)

Audio features describe various aspects and properties of sound. Audio features can be categorized based on their domain as temporary domain, frequency domain, cepstral domain, modulation frequency domain, eigen domain and phase space(Lu, 2001) (Mitrović et al., 2010). Many researchers have found the cepstral based features, mel frequency cepstral coefficients (MFCC) (Davis and Mermelstein, 1980) and linear predictive coefficients (LPC) (Mitrović et al., 2010) (Chu et al., 2009) (Xu et al., 2005) very useful in audio mining tasks such as environmental sound detection, music summarization and speech recognition.

Images

Images are subjected to various processes in order to represent in feature space. Initially, the noise reduction and enhancement methods have been applied to the images to suppress unwanted part and to keep the interesting part of images (Stockman and Shapiro, 2001). The images are subjected to various levels of spatial segmentation such as threshold, region and edge level (Stockman and Shapiro, 2001) depending on the applications requirement.

The images features are extracted using the image attributes color, edges, shape and texture. Generally the images are converted to any suitable color spaces before subjecting to feature extraction. The common color spaces of image are RGB, HSV, XYZ, YUV, CIE etc. Color based features include color histogram (Huang et al., 1997), color autocorrelogram (Huang et al., 1997), color moments (Stricker and Orengo, 1995) and color coherence vector (Pass et al., 1997). Edge features are extracted using edge detection kernels such as Canny, Sobel, Prewitt, Roberts, threshold based, Laplacian operator, Laplacian of Gaussian operator (Davis, 1975) etc. Texture features include co-occurrence matrices, Tamura feature, Markov random field, fractal model, Gabor and wavelet transform (Howarth and Rüger, 2004) (Tuceryan et al., 1993). Shape

based features (Yang et al., 2008a) has been extracted using the popular approaches like convex hull, chain code, boundary and region moments, wavelet transform etc.

2.1.2 Multimedia Fusion

Multimedia fusion integrates the features of multimodal objects in a unified method to utilize the multimodal information efficiently. The survey on multimodal fusion for multimedia analysis (Atrey et al., 2010) provides an overview of different strategies of multimodal fusion. Multimedia fusion methods are broadly categorized as early fusion and late fusion depending on the level of fusion. The hybrid fusion approaches employed the combination of these two methods (Lan et al., 2014; Sargin et al., 2007).

2.1.2.1 Early Fusion

Early fusion approaches integrates the multimodal information at feature level either by concatenating multimodal features or creating a new subspace of features using the multimodal correlations. The main advantage of early fusion methods is the capability to preserve the correlations between the multimodal objects. Moreover, they require only one phase of training phase (Snoek et al., 2005).

Concatenation Of Multimodal Features

Some of the early fusion approaches utilized the multimodal information by concatenating the multimodal features into a single high dimensional feature vector. In an early fusion work, the latent semantic indexing based method has utilized the concatenation of textual and visual features to extract the semantic structures of web documents (Zhao and Grosky, 2002). The method demonstrated the superior performance of textual-visual features over textual features for web document retrieval. In an audio-visual analysis framework (Jiang et al., 2009), the short-term audio-visual atoms (AVA) were extracted using the short term Region tracking. Initially visual regions were tracked within

the short-term video slices to generate the visual atoms and the corresponding audio signal was decomposed into audio atoms. Regional visual features extracted from the visual atoms and the audio features were concatenated to form the AVA feature representation. Joint audio-visual codebook was constructed based on AVAs using multiple instance learning. Finally the codebook based features were generated and used for semantic concept detection. A neutral network based multimodal feature learning approach has been proposed for the sentimental analysis of social media data(Baecchi et al., 2016). The approach concatenates the textual and visual features that are obtained by analyzing the text and images using the Continuous bag-of-words model and denoising autoencoder respectively.

Due to high dimensional feature vectors, these early fusion methods generally face the issues of curse of dimensionality and missing modality. These issues are handled by the correlation based early fusion approaches.

Multimodal Correlations based Approaches

The correlation based early fusion approaches use various methods to integrate the multimodal correlations such as Bayesian networks, graphical models, Canonical Correlation Analysis (CCA), Latent Factor Analysis (LSA), Matrix Factorization and Deep Learning.

In Bayesian network based methods, Hidden Markov Models (HMM) are used to integrate the multimodal features. The special Bayesian networks, factorial HMM and coupled HMM were adopted an early fusion approach for speech recognition (Nefian et al., 2002). In this approach, the correlations among the audio and video observation sequences preserved by upsampling the sequence of visual features to match the audio observation vector frequency. In (Wang et al., 2000), the HMM based fusion methods utilized the correlations between audio and visual features for multimedia content analysis. The speaker localization applications (Nock et al., 2003) employed the HMM based approach to analyze the effect of mutual information between the audio-visual features. The correlations between audio and visual features have been learned using the

Time Delay Neural Network for the applications of speaker detection (Cutler and Davis, 2000) and human motion detection (Zou and Bhanu, 2005).

In speech analysis applications, the Probabilistic Generative models (Beal et al., 2003; Hershey et al., 2004) were employed to fuse the audio-visual observations based on their mutual dependencies in a cluttered, noisy scene. The noise uncertainty measurement of each feature stream was used to combine the audio-visual features using the probabilistic framework (Katsamanis et al., 2006; Papandreou et al., 2007, 2009). The noise was approximated using a Gaussian model and the feature probability was modeled using Gaussian Mixture Models. The audio-visual features were combined by the Bayesian inference method to compute the joint probability of a speech segment. A shrinkage optimized directed information assessment (SODA) framework (Chen et al., 2012) has been proposed for multimodal video indexing and retrieval. The audio and visual features are fused based on the joint probability density functions. The directed information is estimated using the shrinkage estimator from the probability distributions of audio-visual features.

The CCA methods learn the feature representations for two different media types by exploiting the pairwise correlations between them. Based on the learned correlation, these approaches map the original representation of multimodal data into a shared subspace using which the data of different modalities match each other. Wu et al. designed a subspace mapping algorithm based on CCA to learn the multimodal correlations among the audio and visual features (Wu et al., 2006). A general distance function was defined in the CCA subspace using polar coordinates. The approach uses relevance feedback to improve the results. Similar CCA based approach has been used for cross media retrieval by utilizing the correlations between visual and textual features (Rasiwasia et al., 2010) (Costa Pereira et al., 2014). In (Zhen et al., 2016), Zhen et al. proposed a CCA based spectral multimodal hashing method for the retrieval of multimedia data. The method is based on learning the canonical correlations between the textual and visual features.

The LSA approaches such as LSI (Latent Semantic Indexing), PLSA (Probabilistic Latent Semantic Analysis), LDA (Latent Dirichlet Analysis) and matrix factorization algorithms has been used to construct a unified representation for multimodal data based on the correlations between them. Barnard et al. used the multimodal LDA model to predict the missing captions of unlabeled images (Barnard et al., 2003) using the visual and textual features. The PLSA model was used to index and annotate the images in (Monay and Gatica-Perez, 2007). The method modeled the images as a mixture of latent factors to generate the text and image feature and PLSA principle has been applied to learn the mixture of factors. In a music information retrieval approach (Levy and Sandler, 2009), the words from social tags and audio muswords has been learned by the PLSA model. In (Wu et al., 2016), Wu et al. proposed a multimodal random walk neural network model by utilizing the click data collected from user searching behavior. The model learns the latent representation of multimodal data in order to support the cross modal retrieval.

The interactions between textual and visual data has been modeled and projected into a latent semantic space using Parallel Field Alignment Retrieval (Mao et al., 2013). This method determines the semantic correlations between the data which was used for cross modal retrieval. A cross modal association approach called cross-modal factor analysis (CFA) (Li et al., 2003) was introduced to discover the semantic patterns between audio and video subsets. The approach has been tested for cross media retrieval and proved better performance than CCA and LSI. Based on matrix factorization algorithms, a framework (Caicedo and González, 2012) was proposed to learn the relationships between the textual and visual modalities using latent factors. The approach was experimented for retrieval of images. Similarly a Non-negative Matrix Factorization (NMF) (Caicedo et al., 2012) approach was used to generate a joint visual-textual representation for multimodal image analysis. It extracted the relationships between textual and visual features to construct a unified representation based on LSA principles using an asymmetric algorithm.

Deep Learning Approaches

Recently, the deep learning technology has attracted the attention towards the utilization of multimodal correlations. In (Ngiam et al., 2011), the multimodal auto encoders have been employed to learn the features of multiple modalities and validated the methodology for audio-visual speech classification. The multimodal Deep Boltzmann Machine (Srivastava and Salakhutdinov, 2012) has been adopted to learn a generative model of bimodal data consisting of images and texts. These two architectures require identical hidden states of the multiple modalities by neglecting the distinct qualities of the multimodal data. The deep canonical correlation analysis (DCCA) has been introduced (Andrew et al., 2013) to learn maximally correlated two deep nonlinear mappings of two views simultaneously. To identify visual objects from both the labeled image and semantic information extracted from annotated text, a deep visual-semantic embedding model was presented (Frome et al., 2013).

In (Wang et al., 2016b), a supervised multi-modal deep neural network framework has been proposed for the retrieval of multimodal data. The framework includes learning algorithms for the textual and visual data using the deep convolutional neural network model and a neural language model respectively. In a similar approach (He et al., 2016), a deep and bidirectional representation learning model has been employed for cross modal retrieval that employed convolutional neural network and word-based convolutional neural network to learn the images and texts representations respectively. Since these methods basically use bimodal objects, they cannot be extended to multimodal analysis in a straightforward manner.

An online multimodal deep similarity learning (OMDSL) (Wu et al., 2013) has been proposed by optimizing the integration of multiple deep neural networks for the retrieval of images. Here the input for the system relies on the human-crafted features. A deep learning based mapping function (Wang et al., 2014) has been proposed to explore the intra-modality and inter-modality semantic correlations for multimodal retrieval. The method uses stacked auto

encoders to map the high dimensional multimodal features into a common low dimensional latent space. Pang et al. (Pang et al., 2015) proposed a deep Boltzmann machine based joint density model to learn the multimodal low level features for the emotion analysis and retrieval of multimedia data.

Manifold Learning Approaches

In recent years, several manifold learning approaches have been proposed for the integration of multimodal multimedia data. In (Zhuang et al., 2007), Zhuang et. al. proposed a manifold learning cross modal approach to represent the semantically similar multimedia objects. They defined multimedia document (MMD) as the collection of multimodal media objects of similar semantics. The semantically similar MMDs are mapped into MMD semantic space for indexing and retrieval of MMDs. This approach has been extended by constructing a uniform cross-media correlation graph (UCCG) (Zhuang et al., 2008) to evaluate the correlations among multimodal media objects. The graph represents the multimodal media objects as vertices and the correlations as edges. The cross media retrieval has been achieved by assigning a positive score for the query example that spread over UCCG and MMDs with higher score are retrieved as result.

Yang et al. (Yang et al., 2008b) presented a two level manifold learning method for cross media retrieval. Initially the Laplacian media object space (LMOS) was constructed using the three independent Laplacian spaces for images, audios and texts. Then, the MMD semantic graph (MMDSG) was constructed to learn the MMD semantic correlations. In a similar approach (Yang et al., 2010), a MMD semi semantic graph (MMDSSG) has been constructed by learning the semantic correlations between multimedia objects. Then, the cross media indexing space has been constructed to perform cross media retrieval. In order to explore the semantic correlations among the multimodal media objects, a Multimedia Correlation Space (MMCS) (Yang et al., 2009) (Yang et al., 2012) has been constructed using the MMD distance measure for the cross media retrieval. MMCS represents each MMD as a data

point. A ranking algorithm known as ranking with local regression and global alignment (LRGA) has been applied to this space to learn the Laplacian matrix for data ranking. The main drawback of these manifold learning approaches is the use of relevance feedback for the performance improvement.

Recently, the Laplacian Eigenmaps based manifold learning methods have been proposed for the retrieval and annotation of multimedia documents (Daras et al., 2012; Lazaridis et al., 2013; Rafailidis et al., 2013). These approaches construct a low-dimensional feature space utilizing the low-level descriptors of each modality. In this feature space, semantically similar MMDs were described by multimodal descriptor vectors using the Laplacian Eigenmaps based manifold learning method. In (Daras et al., 2012), the radial basis function network is applied for query expansion. The multimodal descriptors have been indexed using as an indexing scheme to accelerate the retrieval process (Lazaridis et al., 2013). In (Rafailidis et al., 2013), Laplacian Eigenmaps are used with heat kernel to improve the performance of external queries. Although these methods achieved good performance with internal queries, they lack in case of external queries. Also they have flaws in matching similar data of same modality.

Other Correlation based Approaches

The parameter free, graph based approach, Mixed Media Graph (MMG)(Pan et al., 2004) has been constructed to discover cross-modal correlations by constructing a graph with three types of nodes namely images, annotations and regions. Then, a random walk was used to estimate the affinity of one node respect to another. A feature Interaction Graph (FIG) (Cui et al., 2010) has been built by using the textual and visual features as nodes and their correlations as edges. The similarity between the multimedia objects has been obtained using a probabilistic model based on Markov Random Field. The approach has been employed for social media data retrieval and social media recommendation. An Audio-Visual Grouplet (AVG) based framework (Jiang and Loui, 2011) was explored for general video concept classification. AVG encloses the temporally correlated set of audio and visual codewords. The AVG based

audio-visual dictionaries are constructed and used for video concept detection. Wang et. al proposed a Semantic Boosting Cross-Modal Hashing approach (Wang et al., 2016a) to project the high dimensional multimodal features into a common hamming space in order to benefit the cross modal retrieval.

Most of above discussed multimodal correlation based approaches utilized the correlations between the bimodal data. They cannot be easily extended for multimodal approaches. Some of these methods perform the correlation analysis and semantic analysis independently.

Table 2.1 provides the summary of all above discussed early fusion approaches. It is noticed that many of the early fusion methods are bimodal approaches. They are used either for cross modal applications or for single modality applications. The applications related to video, image or audio has used the bimodal features such as audio-visual or audio-textual or visual-textual. For instance, image has been retrieved based on visual-textual features, speech has been identified using audio-visual features or audio-textual features. Very few approaches have been developed for the retrieval of multiple multimodal objects such as MMD. Moreover, the correlations among the multimodal multimedia objects are not well studied.

Table 2.1: Summary of Early fusion Approaches

Approach	Modalities Used	Multimedia Applications
Concatenation Of Multimodal Features		
Latent Semantic Indexing	Visual and Textual	Web document retrieval (Zhao and Grosky, 2002)
Multiple Instance Learning	Audio and Visual	Video concept classification (Jiang et al., 2009)
Neutral Network	Visual and Textual	Sentimental analysis of social media (Baecchi et al., 2016)

Table 2.1 Continued: Summary of Early fusion Approaches

Approach	Modalities Used	Multimedia Applications
Multimodal Correlations based Approaches		
Bayesian Networks	Audio and Visual	Speech recognition (Nefian et al., 2002), Multimedia content analysis in videos (Wang et al., 2000), Speaker localization (Nock et al., 2003)
Time Delay Neural Network	Audio and Visual	Speaker detection (Cutler and Davis, 2000), Human motion detection (Zou and Bhanu, 2005)
Probabilistic Generative model	Audio and Visual	Object tracking (Beal et al., 2003), Speech detection (Hershey et al., 2004)
Feature Uncertainty Measure using Gaussian model	Audio and Visual	Speech recognition (Katsamanis et al., 2006; Papandreou et al., 2007, 2009)
SODA	Audio and Visual	Video indexing and retrieval (Chen et al., 2012)
CCA	Audio and Visual	Cross media retrieval (Wu et al., 2006)
	Audio and Textual	Cross media retrieval (Costa Pereira et al., 2014; Rasiwasia et al., 2010)
Multimodal LDA	Visual and Textual	Image annotation (Barnard et al., 2003)
Spectral Multimodal Hashing	Visual and Textual	Multimodal retrieval (Zhen et al., 2016)

Table 2.1 Continued: Summary of Early fusion Approaches

Approach	Modalities Used	Multimedia Applications
PLSA	Visual and Textual	Image annotation (Monay and Gatica-Perez, 2007)
	Audio and Textual	Music Information Retrieval (Levy and Sandler, 2009)
Parallel Field Alignment	Visual and Textual	Cross media retrieval (Mao et al., 2013)
Cross modal Factor Analysis	Audio and Visual	Cross media retrieval (Li et al., 2003)
Matrix Factorization	Visual and Textual	Image retrieval (Caicedo et al., 2012; Caicedo and González, 2012)
Multimodal random walk neural network	Visual and Textual	Cross modal retrieval (Wu et al., 2016)
Deep Learning Approaches		
Autoencoders	Audio and Visual	Speech classification (Ngiam et al., 2011)
Deep Boltzmann Machine	Visual and Textual	Multimodal image retrieval (Srivastava and Salakhutdinov, 2012), Affective analysis and multimedia retrieval (Pang et al., 2015)
DCCA	Audio and Visual	Multiview learning (Andrew et al., 2013)
Deep Visual Semantic Embedding	Visual and Textual	Image object identification (Frome et al., 2013)

Table 2.1 Continued: Summary of Early fusion Approaches

Approach	Modalities Used	Multimedia Applications
OMDSL	Visual	Image retrieval (Wu et al., 2013)
Stacked Autoencoder	Visual and Textual	Cross media retrieval (Wang et al., 2014)
Deep neural network	Visual and Textual	Multimedia retrieval (Wang et al., 2016b)
Deep and bidirectional representation learning model	Visual and Textual	Cross media retrieval (He et al., 2016)
Manifold Learning Approaches		
MMD Semantic Space and Relevance Feedback	Audio, Visual and Textual	Cross media retrieval (Zhuang et al., 2007)
UCCG and Relevance Feedback	Audio, Visual and Textual	Cross media retrieval (Zhuang et al., 2008)
LMOS based MMDSG and Relevance Feedback	Audio, Visual and Textual	Cross media retrieval (Yang et al., 2008b)
MMDSSG and Relevance Feedback	Audio, Visual and Textual	Cross media retrieval (Yang et al., 2010)
MMCS, ranking with LRGA and Relevance Feedback	Audio, Visual and Textual	Cross media retrieval (Yang et al., 2012, 2009)
Laplacian Eigenmaps	Audio, Visual and Textual	Multimedia retrieval (Daras et al., 2012; Lazaridis et al., 2013; Rafailidis et al., 2013)

Table 2.1 Continued: Summary of Early fusion Approaches

Approach	Modalities Used	Multimedia Applications
Other Correlation Based Approaches		
MMG and Random Walk	Visual and Textual	Cross media retrieval (Pan et al., 2004)
FIG	Visual and Textual	Social media recommendation and retrieval (Cui et al., 2010)
AVG	Audio and Visual	Video concept classification (Jiang and Loui, 2011)
Semantic Boosting Cross-Modal Hashing approach	Visual and Textual	Cross media retrieval (Wang et al., 2016a)

2.1.2.2 Late Fusion

In late fusion method, the final result is achieved by combining and analyzing the individual results of each modality of objects. The individual results are obtained by analyzing the each modality of objects based on their individual features. Late fusion methods have the advantage of selecting the best learning approach for each modality of data. They can be categorized based on the type of learning approaches used. A late fusion approach (Adams et al., 2003) was used for the retrieval of semantic concepts in videos using visual, audio and textual modalities. The concept representations are modeled using the probabilistic modeling approaches. A discriminate learning approach has been used for fusing different modalities at the semantic level. An online video recommendation system (Yang et al., 2007a) was employed multimodal fusion by combining the relevance from three modalities (text, audio and video) using

attention fusion function (AFF). Relevance of single modalities was first computed by weighted linear combinations of relevance between features and then fused using AFF. The relevance feedback dynamically adjusts the intra and inter weights between the different modalities. Anguera et al. developed a parameter free decision level multimodal fusion algorithm for video copy detection by fusing the audio and visual features (Anguera et al., 2011). The algorithm generates the list of possible matches based on the weighted sum of the normalized scores for each modality.

In a decision level fusion method (Kanluan et al., 2008), audio and visual features are used to recognize the emotions. Support Vector Machine (SVM) was used to estimate the three emotion primitives valence, activation, and dominance using the audio and visual features. The useful features are selected using the Sequential Forward Floating Search (SFFS) technique for audio features and SVR (Support Vector Regression)-SFFS for visual features. The method demonstrated the better performance by using both the modalities. An ensemble based system (Wagner et al., 2011) has been employed the audio, face and gesture feature for emotion recognition with missing data. The method demonstrated the superior performance of ensemble techniques with multimodal data over the single modality data.

For categorizing the web video, Yang et al. proposed a multimedia fusion method (Yang et al., 2007b). The rich information of web videos including visual features, latent semantic features, audio features, and surrounding text features are classified using SVM classifiers for each individual modality. The outputs of the classifiers were fused using max fusion, average fusion, and linear weighted fusion to predict the class of an unknown video. The authors state that the multimodality representation outperforms the individual representation. Wu et al. developed two late fusion approaches, gradient-descent-optimization linear fusion (GLF) and the super-kernel nonlinear fusion (NLF) for the detection of video concept (Wu et al., 2004b). The weighted linear combinations of individual modalities is learned by GLF using the kernel matrices with gradient descent

techniques. The super-kernel nonlinear fusion learns an optimal non-linear combination of individual models by fusing single-modality classifiers. The authors used SVM classifiers to classify the target concept based on the color, texture, motion and audio features.

In a late fusion based approach called MultiFusion (Wang and Kankanhalli, 2010a), the multimodal data has been segmented into atomic event to use it as a fusion unit. The performance of the system has been improved by combining the multimodal classifiers in each iteration using the multimodal correlation based Adaboost-like structure. A late fusion method based on portfolio theory (Wang and Kankanhalli, 2010b) was adapted for multimedia fusion to derive the optimal fusion weights for different information sources by maximizing expected return and minimizing the risk to achieve an overall good performance. An adaptive decision fusion method (Lee and Park, 2008) has been used for audio-visual speech recognition. Adaptive weighting scheme was implemented for audio visual features by varying the weights according to the noise level in speech. The method neglected the correlation among the information sources. A two step framework of optimal multimodal fusion (Wu et al., 2004a) has been introduced for multimedia data analysis. The first step finds the statistically independent modalities from raw features and the second step uses the super-kernel fusion to determine the optimal combination of individual modalities.

A multimodal fusion framework (Zhu et al., 2006) has been proposed to classify the text embedded images. Initially images were classified using bag-of-words. Then, the text line is detected and the text features text color, size, location, edge density, brightness, contrast are extracted. The text concept has been learned using multiple instance learning. Finally, the visual and textual features were fused using the pairwise SVM classifier. For automatic laughter detection (Reuderink et al., 2008), the authors combined the audio and visual features at decision level using audio-HMM classifier and video-SVM classifier. Final result is obtained by combining the audio and video classifiers using a weighted sum rule. A non-linear audio visual fusion scheme has been proposed in

(Muneesawang et al., 2010) that employs SVM classifier to learn semantic concepts from a given video database. Visual features were extracted using adaptive video indexing technique and audio features were extracted using Laplacian Mixtures.

In a multimodal and multilevel ranking framework (Hoi and Lyu, 2008), the multimodal resources were combined by the graphs for video retrieval task. The semi-supervised ranking method was used for graph learning by applying semi supervised learning techniques. The multilevel ranking framework has been proposed by integrating several learning methods in a cascade fashion to reduce the computational cost and improve the retrieval performance for large-scale video retrieval. In (Mourão et al., 2015), a data fusion method has been proposed that combines the textual and visual features to support the multimodal medical information retrieval. In (Zhai et al., 2014), the joint representation learning method has been proposed to learn the correlations and semantic information of different media types in different graphs. Modeling different media objects in different graphs neglects the correlation between the media objects.

The summary of all above discussed late fusion methods is given in Table 2.2. Most of these approaches used the multimodal information for the analysis of single modality applications. The late fusion approaches have the privilege to select the suitable learning and knowledge extraction methods for each modality of data. However, the late fusion methods have to deal with the selection of learning and knowledge extraction methods for each modality. Also, they need to consider the methods of analyzing and integrating the results of each modality to get the final output. The main drawback of late fusion methods is neglecting the correlations between the multimodal data that are very necessary to gain the complete knowledge about the concept of a multimedia document.

Table 2.2: Summary of Late fusion Approaches

Approach	Modalities Used	Multimedia Applications
SVM Classifier	Audio, Visual and Textual	Semantic concept detection (Adams et al., 2003), Web video categorization (Yang et al., 2007b)
	Audio and Visual	Emotion recognition (Kanluan et al., 2008), Movie clip classification and retrieval (Muneesawang et al., 2010)
	Audio, Visual and Motion	Video concept detection(Wu et al., 2004b)
	Visual and Textual	Image classification(Zhu et al., 2006)
AFF and relevance feedback	Audio, Visual and Textual	Online video recommendation (Yang et al., 2007a)
Weighted Sum Rule	Audio and Visual	Video copy detection (Anguera et al., 2011), automatic laughter detection (Reuderink et al., 2008)
Ensemble based system	Audio and Visual	Emotion recognition with missing data (Wagner et al., 2011)
Fusion of Weighted multimodal classifiers	Audio, Visual and Textual	Image concept detection and human detection (Wang and Kankanhalli, 2010a)
Portfolio based Multimedia Fusion	Audio, Visual and Textual	Human detection (Wang and Kankanhalli, 2010b)

Table 2.2 Continued: Summary of Late fusion Approaches

Approach	Modalities Used	Multimedia Applications
Independent modality analysis and super kernel fusion	Audio ,Visual and Motion	Multimedia data analysis (Wu et al., 2004a)
Multimodal and multilevel ranking framework	Visual and Textual	Video retrieval(Hoi and Lyu, 2008)
Joint representation learning	Audio, Visual, Textual, Video and 3D	Cross media retrieval(Zhai et al., 2014)
Data fusion	Visual and Textual	Medical information retrieval(Mourão et al., 2015)

2.1.2.3 Hybrid Fusion

The hybrid fusion approaches incorporated the advantages of both the early fusion and late fusion methods. A two stage hybrid fusion approach (Natarajan et al., 2012) has been proposed for video event detection. The first stage employs an early fusion method, multiple kernel learning to combine the audio, visual and motion features in different combinations. The second stage combines these subsystems using the late score level fusion. In (Chang et al., 2007), three multimodal fusion frameworks have been proposed. The first method is ensemble fusion method that combines the scores of independent detection using weighted sum. The other two methods were based on late fusion and early fusion method. The late fusion based Audio-Visual Boosted Conditional Random Field method used the global features and combined the prediction results from separately

trained models using Conditional Random Field (CRF). The early fusion based approach Audio-Visual Joint Boosting used the local visual features and utilized the VSPM kernels derived from individual concepts to learn the joint models for detecting multiple concepts simultaneously. A multimodal information fusion framework (Mansoorizadeh and Charkari, 2010) has been proposed to generate a hybrid feature space by combining the multimodal features. The system has been used feature level fusion to capture the correlations between the modalities and decision fusion to improve the performance. The system has been evaluated for audio visual speech recognition and the results proved that the use of multimodal data achieved higher accuracy than the single modality.

The combination of CCA based early fusion and late fusion (Sargin et al., 2007) has been proposed for open-set speaker identification. Initially the audio features and video has been used for each canonical pair. Then, the late fusion used the Bayesian decision fusion to integrate weak HMM classifiers to get the output. In an approach called double fusion, early fusion and late fusion were combined to incorporate the advantages (Lan et al., 2014). Initially early fusion was performed to generate various combinations of features from the single modality features. For each combination of features and single modality features, classifiers have been trained and late fusion was performed to get the output. The results demonstrated superiority of double fusion over early and late fusion. In a multimodal fusion approach (Zeppelzauer and Schopfhauser, 2016), both the early fusion and late fusion approaches has been applied for fusing the textual and visual data to classify the social events.

The summary of all above discussed hybrid fusion approaches is given in Table 2.3. These approaches are computationally expensive as they use both the early and late fusion methods.

Table 2.3: Summary of Hybrid fusion Approaches

Approach	Modalities Used	Multimedia Applications
Multiple kernel learning and Score level fusion	Audio ,Visual and Motion	Video event detection (Natarajan et al., 2012)
Audio-Visual Boosted CRF	Audio and Visual	Multimodal semantic concept detection (Chang et al., 2007)
feature level fusion and decision level fusion	Audio and Visual	Speech recognition (Mansoorizadeh and Charkari, 2010)
CCA and HMM based Classifier	Audio and Visual	Speaker identification (Sargin et al., 2007)
Multiple kernel learning and SVM classifier	Audio, Visual and Textual	Multimedia event detection (Lan et al., 2014)
Multimodal fusion approach	Visual and Textual	Social event classification (Zeppelzauer and Schopfhauser, 2016)

2.2 Multimedia Document Representation

The multimedia document Representation organizes the MMDs in an appropriate representation that assist the multimedia mining methods in discovering the useful patterns. Vector Space Document (VSD) and Suffix Tree Document (STD) are the two most widely used document representation models for knowledge extraction from the documents.

2.2.1 Vector Space Document Model

Vector space document (VSD) model represents the collection of documents in a high dimensional vector space (Salton et al., 1975). The dimensions of the vector space are related to the words of the documents. The documents are represented as a feature vector of words based on the significance of the word. In the model, the importance of a word for a document in the collection is evaluated by a statistical measure known as tf-idf (Term Frequency-Inverse Document Frequency). The VSD model extracts the words from the collection of documents and employs the tf-idf weighting function to represent each document as, $\overrightarrow{d} = \{w_{t,d} \mid_{t=1}^{M}\}$. The tf-idf weighting function $w_{t,d}$ is calculated as,

$$w_{t,d} = tf_{t,d} * idf_t \tag{2.2.1}$$

where $tf_{t,d}$ is the term frequency of term t in document d and idf_t is the inverse document frequency which measures the importance of a term t. The idf_t is calculated as,

$$idf_t = 1 + log\frac{N}{df_t} \tag{2.2.2}$$

where N is the number of documents and df_t is the document frequency that provides the number of documents containing a term t. The VSD model computes the pairwise similarity between the two documents d_A and d_B using the cosine similarity as follows,

$$Sim(\overrightarrow{d_A}, \overrightarrow{d_B}) = \frac{\overrightarrow{d_A}.\overrightarrow{d_B}}{\left\|\overrightarrow{d_A}\right\| \left\|\overrightarrow{d_B}\right\|} \tag{2.2.3}$$

Most of the text document knowledge extraction approaches use the VSD model for representing the documents. For the text retrieval, a generalized Vector Space Model (VSM) (Tsatsaronis and Panagiotopoulou, 2009) has been proposed in which the term to term relatedness was measured using WordNet. The similarity between the patent documents has been found by representing them using the International Patent Classification code based VSM vectors

(Chen and Chiu, 2011). The phrase-based VSM was proposed for the retrieval of the documents in which the similarity between the phrases has been found using the knowledge base (Mao and Chu, 2007). The topic based vector space model has been proposed for spam filtering in emails by using a semantic ontology (Santos et al., 2012).

The VSD model has been basically used for knowledge extraction of text documents. However, few of the approaches extended the VSD model for knowledge extraction from audios and images. To identify the spoken language, the VSM was employed for acoustic segment models of speech (Li et al., 2007). For the indexing and retrieval of music content (Maddage et al., 2006), a music structure based VSM approach was used to represent the music segment. The n-gram statistics of the phonemes was extracted using a VSM based approach to identify the language of a speech segment (Li et al., 2013). The importance degree of image features was identified using VSM method for the retrieval of images (Suzuki et al., 2008). In (Martinet et al., 2011), the VSM has been merged with conceptual graph formalism for the retrieval of photographs. However, to the best of our knowledge there are no efforts made to represent the multimedia documents using VSD model.

Although the VSD model is simple, it is suitable for small datasets. With very long documents, measuring similarity is difficult due to the poor representation of document with smaller similarity values and higher dimension. The main drawback of VSD model is the use of single word term analysis of documents by the VSD based knowledge extraction methods. The main problem with single word is ambiguity in knowledge representation. For example, the word “Tiger” may be related to “animal” or “person name” or “airways”. To achieve more efficiency, instead of a single word, the group of words shared by many documents known as phrases conveys more information about the documents. For example “Tiger Airways”, “Tiger lives in forest” and so on. Moreover, the VSD model neglects the associations between the words of the documents by considering the single words.

2.2.2 Suffix Tree Document Model

A string can be represented in compressed form by the suffix tree using all the suffixes of the string. In order to achieve more efficient decision making, instead of a single word, the group of words shared by many documents known as suffixes are considered for the documents. Suffix tree document (STD) model is the another model used for the representation of documents. The STD model represents the documents using the suffix tree by identifying and extracting all the overlap phrases of the documents. It considers the text document as a string and the words as the characters. It is a complete representation of set of documents containing all suffix phrases of the documents. It was first introduced by Zamir et al. (Zamir and Etzioni, 1998; Zamir et al., 1997) for clustering the web documents. The STD model has been constructed in a linear time using Ukkonen's algorithm (Ukkonen, 1995).

The STD model comprised of root node and suffix nodes. The tree edges are labeled by the suffix phrase and suffix nodes are labeled by the integration of edge labels. The suffix nodes represent the set of documents and a suffix phrase that is common to represented documents. Each suffix node provides the information about the document identifier and position of the suffix in a string. For knowledge extraction, the pairwise document similarity is computed based on the number of similar suffix nodes between the documents.

Unlike the VSD model, the STD model considers the documents as the set of ordered words. STD model extracts the word sequences from documents to represent them in a tree structure for the various data mining tasks such as document classification, clustering and retrieval. In (Zu Eissen et al., 2005), the characteristics of STD model and suffix tree clustering (STC) algorithm have been analyzed to state the advantages and drawbacks of both the methods. Due to the drawbacks of STC algorithms, several similarity methods have been proposed to improve the performance of the data mining tasks. In (Li et al., 2008), a clustering algorithm has been proposed to cluster the documents using the frequent word sequences extracted by the generalized STD model. The STD

model has been used with hierarchical clustering algorithm to cluster the text documents using the term frequency based similarity measure (Worawitphinyo et al., 2011). A phrase based document similarity (Chim and Deng, 2008) was proposed using the combination of the suffix tree model and vector space model to cluster the documents using hierarchical agglomerative clustering algorithm. Basically, the STD model has been implemented to represent the set of homogeneous text documents (Zamir and Etzioni, 1998). However, some attempts were made to use the suffix tree model for images (Ruocco and Ramampiaro, 2010) and audios (Lo et al., 2008). The suffix tree based approach was used to extract the important repeating patterns of music objects (Lo et al., 2008). To retrieve the Flickr images, the suffix tree clustering algorithm was employed by utilizing the image annotations (Ruocco and Ramampiaro, 2010). To the best of our knowledge, STD model has been not attempted to represent the multimedia documents. Table 2.4 summarizes the VSD and STD approaches discussed in this section.

Table 2.4: Summary of Multimedia Document Representation Approaches

Approach	Modalities Used	Multimedia Applications
Vector Space Document Model		
Vector Space Model	Textual	Text retrieval (Tsatsaronis and Panagiotopoulou, 2009)
International Patent Classification code based VSM vector	Textual	Retrieval of patents (Chen and Chiu, 2011)
Phrase based VSM	Textual	Document retrieval (Mao and Chu, 2007)

Table 2.4 Continued: Summary of Multimedia Document Representation Approaches

Approach	Modalities Used	Multimedia Applications
Topic based VSM	Textual	Spam filtering in emails (Santos et al., 2012)
VSM based acoustic segment models of speech	Audio	Spoken language identification(Li et al., 2007)
Music structure based VSM for representation of music segment	Audio	Music indexing and retrieval (Maddage et al., 2006)
Extraction of phoneme statistics using VSM	Audio	Speech language identification (Li et al., 2013)
VSM based importance degree retrieval	Visual	Image retrieval (Suzuki et al., 2008)
VSM with conceptual graph formalism	Visual	Photographs retrieval (Martinet et al., 2011)
Suffix Tree Document Model		
STC	Textual	Web document clustering (Zamir and Etzioni, 1998; Zamir et al., 1997)
	Visual	Image retrieval (Ruocco and Ramampiaro, 2010)
STD and Hierarchical clustering algorithm	Textual	Text document clustering (Worawitphinyo et al., 2011)

Table 2.4 Continued: Summary of Multimedia Document Representation Approaches

Approach	Modalities Used	Multimedia Applications
STD, VSD and Hierarchical clustering algorithm	Textual	Text document clustering (Chim and Deng, 2008)
STD	Textual	Document clustering (Li et al., 2008)
	Audio	Discovery of non-trivial repeating patterns in a music object (Lo et al., 2008)

2.3 Multimedia Mining

Multimedia mining methods are used to extract the knowledge from multimodal multimedia data. They make use of the multimodal information and intelligent data mining methods to extract the useful knowledge from multimedia documents. Many multimodal applications achieved better performance by utilizing the multimodal data than the unimodal data. The most widely used multimedia mining methods are classification, clustering, frequent pattern mining and sequential pattern mining.

Multimedia Classification

Multimedia classification predicts the class of unknown data based on the learning of labeled data. The most popular classifiers used for multimedia classification are decision tree, support vector machines, k-nearest neighbors, neural networks, Bayesian classifier and rule based classifier. The decision tree

based multimodal classifiers have been used to detect the events in the videos (Chaisorn et al., 2003; Chen et al., 2006, 2004; Shyu et al., 2008). In (Chaisorn et al., 2003), the classifier utilized the combination of visual, audio and textual features to segment the story in large news video corpus. In (Chen et al., 2006, 2004), the audio and visual features of the video frames have been used for the detection of soccer goals in soccer videos. In (Shyu et al., 2008), the video is syntactically segmented to extract the visual and audio features. Distance base mining techniques applied to reduce the negative instances and decision tree classifier has been used for rare event detection.

A SVM based multimodal classifier utilized the textual and visual features for news video classification (Lin and Hauptmann, 2002). A video classification algorithm called VideoMule (Ramachandran et al., 2009) has been proposed based on the consensus learning approach. It combines the classification and clustering algorithm to train textual metadata, audio and video to build a multilabel tree. In an image classification framework (Xie et al., 2014), the image tags were effectively used with the visual features. Initially, the tags were refined by removing the inaccurate tags. Later, graph based learning has been adopted by combining the visual and tag graphs. Using support vector regression, the class label of unknown image was predicted. A two stage video classification framework (Liu et al., 2016) has been proposed using the textual, audio and visual features of video. In first stage, the semantic relations between modalities have been preserved by using the stacked contractive auto encoders (SCAE) for each modality. The second stage learned the inter modality semantic interactions using multimodal SCAE by joining the output of first stage.

In (Zeppelzauer and Schopfhauser, 2016), the author employed early and late fusion approaches to investigate the capabilities of textual, visual and multimodal classification. They demonstrated with social media event classification and concluded that the multimodal classification achieved better performance than the single modality classification. In another hybrid fusion approach (Laurier et al., 2008), both the early and late fusion methods have

been used to combine the audio and textual lyrics information for music mood classification. The experiments demonstrated that the combination of audio and textual information performed better than the single modalities alone.

The major shortcomings of above discussed multimedia classification approaches are, the selection of classifier for each modality of objects and classifier fusion. Most of the approaches used distinct classifier for different modality of data. The classifier outputs are combined either using another classifier or rule based methods. Here, the correlations between multimodal data are neglected. Very few early fusion approaches employ the associations between the multimodal data for classification. Moreover, most of these methods utilized the multimodal information for the classification of single modality of data. Therefore, an effective multimedia classification is required for the classification of MMDs.

Multimedia Clustering

Multimedia Clustering is an unsupervised approach that partitions the collection of multimodal objects into a number of clusters of multimedia objects of similar concept. Due to the heterogeneous nature, various clustering algorithms have been proposed for clustering the multimedia objects. A k-means clustering algorithm (Goh et al., 2003) with audio and visual features has been employed for the detection of commercials from TV program. In (Bekkerman and Jeon, 2007), a lightweight version of Combinatorial Markov Random Fields has been proposed for clustering multimedia collections using visual and textual features. For all the modalities, partitions have been constructed simultaneously to obtain the dense distribution of the modalities. An adaptive resonance theory (ART) based heterogeneous fusion approach (Meng et al., 2014) has been proposed for clustering the web documents. The method used an adaptive weighting algorithm to combine the image and meta information features. In an image clustering application, a tripartite graph has been used to fuse the low level visual features and surrounding texts (Gao et al., 2005). The CCA based early

fusion approach was used for clustering the audio-visual speaker and Wikipedia articles (Chaudhuri et al., 2009). The authors claimed that the CCA based approach performed better than the PCA based approach.

Cai et al. (Cai et al., 2004) proposed a two level spectral clustering technique to cluster web image search results using visual, textual and link analysis. In another similar approach (Sevillano and Alías, 2014), the spectral clustering has been applied on multimodal features for the discrimination of image senses. Both the methods (Cai et al., 2004; Sevillano and Alías, 2014) need the calculation of affinity scores for every pair of images and every modality which results in unrealistically heavy representation. A multimodal spectral clustering algorithm (Petkos et al., 2012, 2014) has been proposed for the detection of social events. The method employed the supervisory signal to achieve the better performance.

In (Zhang et al., 2009) a multimodal framework has been proposed to cluster the spam images using two-level clustering method. In the first level, the image similarities were calculated with respect to the visual features and images were grouped. In the second level, clustering results were further refined using a string matching method by comparing the closeness of texts in two images. Wei et al. proposed a cross reference reranking approach (Wei et al., 2010) to improve the retrieval performance. The reranking approach combines the multimodal features to utilize the semantic understanding of video. Initially the individual results of different modalities have been clustered and the clusters were ranked based on their relevance to the query. Finally, all the ranked clusters have been combined using a cross reference approach.

From the above discussion it is observed that many of the multimedia clustering algorithms employed individual clustering for each modality and then the results are combined. Some approaches employed early fusion approach to combine the bimodal features for clustering the unimodal objects such as images, audios, and videos. Only a very few approaches developed for clustering the multimodal web documents. Thus, an effective multimedia clustering is needed for clustering the MMDs.

Mining Frequent Multimedia patterns and Multimedia Associations

Frequent patterns are the itemsets or sub-sequences or substructures that appear frequently in a data set (Han et al., 2011). Association rule mining (Agrawal et al., 1993) generates the descriptive rules using the frequent patterns that discover the relationships among the set of items in a dataset. The learned relationships are helpful in analyzing the domain.

Let I be the set of items and T is the set of transactions in dataset D. The association rules extracted from D are in the form of $X \rightarrow Y$ where $X, Y \subseteq I$, and $X \bigcap Y = \phi$. The support of the rule is the percentage of transactions in D that contain both X and Y. The confidence is the proportion of the transactions that contain X which also contains Y. The confidence of a rule can be computed as $support(X \bigcup Y)/support(X)$.

The usefulness of rules is extended to develop the association rule based classification model(Ma, 1998). Association rule based classification model is designed using two steps: 1. Association rule generation 2. Build the classifier using the subset of strong class association rules (CAR). The strong CARs are generated by pruning the generated association rules based on user defined criteria. In CAR, the antecedent is restricted to only the class. A class association rule is in the form of $itemset \rightarrow c$, where $itemset$ is the set of items such that $itemset \subseteq I$ and $c \in C$ is the class label.

The popular association rule mining algorithms, Apriori and FP-Tree are used to mine the frequent patterns and associations from the well structured alphanumeric text data. However in recent years, some efforts made to mine the multimodal frequent patterns and associations from the multimedia data. In (Chen et al., 2007), Chen et al. proposed a hierarchical temporal association mining approach to automatically capture the optimal temporal patterns for characterizing the interesting events in soccer video. The method is the modified version of Association Rule Mining (ARM) that supports the automatic selection

of threshold values. To mine the temporal patterns, an extended Apriori algorithm was proposed. Two learning methods were proposed (Jiang and Tan, 2009) for discovering the underlying associations between images and texts based on small training data sets. In first method, vague transformation extracts the associations between image and text using the information categories. The second method uses fusion Adaptive Resonance Theory to extract direct image-text associations by employing automatic summarization of associated features. A multilevel sequential association mining (Zhu et al., 2005) has been extracted using the video processing techniques. Then, the associations among the audio and visual cues were explored using multilevel sequential association mining to construct video indexing.

He et. al. developed a Multi-Modal Semantic Association Rule (MMSAR) approach to explore the associations among the visual features and keywords of images (He et al., 2011). The MMSARs were mined using a customized frequent itemsets mining algorithm. The mined MMSARs fused the visual -textual features to improve the retrieval performance. An image retrieval approach employed the association rule mining (ARM) algorithm (Alghamdi et al., 2014) to learn the semantic interactions between visual and textual information of the images. Initially image clusters has been formed based on visual features and textual features. Then, the ARM algorithm was applied to visual and textual clusters to extract the association rules. To support the medical image diagnosis, an association rule based approach (Ribeiro et al., 2008) has been proposed by correlating the medical images and diagnosis reports. The useful features were selected and the associations between image features and keywords were mined. The classification of the image was performed by the classifier using the mined correlations.

It is noticed that, the above discussed multimedia association mining approaches exploit the associations among the bimodal data. In order to detect the associations among the multimodal contents of MMDs, sophisticated multimedia association mining methods are required.

Mining Multimedia Sequential Patterns

In association rule mining, the order of the transactions is not considered. The ordered list of itemsets is known as sequence. In sequence pattern mining (Agrawal and Srikant, 1995), the itemsets maintain the order of items. Let $I = \{i_1, i_2, ...i_m\}$ be the set of items. In an itemset, one item can occur only once. A sequence s is denoted by $(x_1 x_2 ... x_r)$ where x_1 is an itemset. The itemset x_1 is denoted as $\{j_1, j_2, .., j_k\}$ where $j_k \in I$. The number of items occur in sequence is considered as the length of the sequence. A sequence of length k is called k-sequence. Sequential pattern mining extracts the sequences with user defined minimum support. Sequential patterns are used to generate the sequential rules (Liu, 2007). A sequential rule can be defined as $Y \rightarrow X$ where X is the sequence and Y is the proper subsequence of X. The sequential rules can be extended to generate the class sequential rules. A class sequential rule is defined in the form $sequence \rightarrow c$, where *sequence* is the sequence containing a set of items such that $sequence \subseteq I$ and $c \in C$ is the class label. The class sequential rules are mined using the Apriori-Sequential pattern mining algorithm (Liu, 2007).

Basically, sequential patterns are mined from the well structured text documents (Liu, 2007). However, there were some approaches that used to mine the sequential patterns from the multimedia objects. In (Ren and Jang, 2012), the time-constrained sequential pattern (TSP) mining has been introduced to classify the music genre. The music was tokenized into a sequence of ASM (acoustic segment model) indices and the TSP mining algorithm was applied to generate to genre-specific TSPs. Then, the music was represented using the weighted occurrence frequencies of all TSPs and classified using the SVM classifier. A data structure known as pattern count tree (Ananthanarayana et al., 2003) has been used to represent the database of hand-written digits. The pattern count tree builds the database based on the sequential patterns of features in a single scan of database. An image classification approach has been proposed by employing the sequential patterns of low level features of the labeled images (Lin et al., 2009). It is observed that very few attempts were made to

discover the sequential patterns from multimedia objects images and audios. However, the discovery of sequential feature patterns of multimedia objects of MMDs is not properly studied.

Table 2.5 provides the summary of all the multimedia mining methods discussed above. It is observed that most of the multimedia mining methods employed the multimodal information for the mining of single modality objects. Many methods utilized the information of only two modality of objects. The effective utilization of multimodal correlations is not properly explored.

Table 2.5: Summary of Multimedia Mining Approaches

Approach	Modalities Used	Multimedia Applications
Multimedia Classification		
Decision Tree	Audio and Visual	Goal detection in soccer videos (Chen et al., 2006, 2004), Video rare event detection (Shyu et al., 2008)
	Audio, Visual and Textual	Story segmentation (Chaisorn et al., 2003)
SVM Classifier	Visual and Textual	News video classification (Lin and Hauptmann, 2002)
VideoMule Algorithm	Audio, Visual and Textual	Video classification (Ramachandran et al., 2009)
Support Vector Regression	Visual and Textual	Image classification (Xie et al., 2014)
Hybrid Fusion	Audio and Textual	Music mood classification (Laurier et al., 2008)
	Visual and Textual	Social media event classification (Zeppelzauer and Schopfhauser, 2016)

Table 2.5 Continued: Summary of Multimedia Mining Approaches

Approach	Modalities Used	Multimedia Applications
Stacked Contractive Autoencoders	Audio, Visual and Textual	Video classification (Liu et al., 2016)
Multimedia Clustering		
K-means clustering algorithm	Audio and Visual	Commercial detection (Goh et al., 2003)
Combinatorial Markov Random Fields	Visual and Textual	Clustering multimedia collections (Bekkerman and Jeon, 2007)
Spectral clustering	Link, Visual and Textual	Image clustering (Cai et al., 2004)
	Visual and Textual	Discrimination of image senses (Sevillano and Alías, 2014)
	Visual, Text, and time	Social event detection (Petkos et al., 2012, 2014)
Two-level clustering	Visual and Textual	Spam image clustering (Zhang et al., 2009)
Cross Reference Reranking approach	Visual and Textual	Video search reranking (Wei et al., 2010)
ART based heterogeneous fusion	Visual and Textual	Web documents clustering (Meng et al., 2014)
Tripartite graph	Visual and Textual	Image clustering (Gao et al., 2005)

Table 2.5 Continued: Summary of Multimedia Mining Approaches

Approach	Modalities Used	Multimedia Applications
CCA	Audio, Visual and Textual	Clustering the audio-visual speaker and Wikipedia articles (Chaudhuri et al., 2009)
Frequent Multimedia Pattern Mining and Association Rules Generation		
Hierarchical temporal association mining	Audio and Visual	Video event detection (Chen et al., 2007)
Vague transformation and Fusion ART	Visual and Textual	Learning the associations between text and images (Jiang and Tan, 2009)
Multilevel Sequential association mining	Audio and Visual	Semantic indexing and event detection (Zhu et al., 2005)
Multi-Modal Semantic Association Rule	Visual and Textual	Web image retrieval (He et al., 2011)
Multimodal Fusion method based ARM	Visual and Textual	Image retrieval (Alghamdi et al., 2014)
ARM based suggestions generation	Visual and Textual	Medical image diagnosis (Ribeiro et al., 2008)
Multimedia Sequential Pattern Mining and Sequence Rules Generation		
TSP mining	Audio	Music genre classification (Ren and Jang, 2012)
Pattern Count Tree	Visual	Hand-written digit recognition (Ananthanarayana et al., 2003)
Sequential patterns of low level features	Visual	Image classification (Lin et al., 2009)

2.4 Outcome of Literature Review

After extensive review of multimedia data representation, multimedia document representation and multimedia mining methods, we identified the following issues and research directions for the knowledge discovery in multimedia documents.

The growing amount of knowledge enriched multimedia documents demanded for the sophisticated knowledge discovery systems. The success of the knowledge discovery in multimedia documents depends on the efficacy of the multimedia mining methods. Most of the existing multimedia mining methods were proposed for mining the unimodal objects such as image mining, audio mining and video mining. These methods have utilized the multimedia information of two different modality. For example, the image retrieval or classification applications utilized image features and features from their caption. Similarly, the video applications utilized the audio and visual information. The MMD knowledge extraction methods required to utilize the information of multiple multimodal objects to analyze the multimodal contents of a MMD. For example, the query MMD containing chirping sound with image of a bird retrieves the MMDs having different images, sounds and text information of that bird. However, very few approaches have been developed for mining the knowledge from MMDs. Moreover, these methods performed mining based on the individual analyzing of each modality neglecting the interactions between the multimedia objects of MMDs. In order to make use of rich multimedia information available in the MMDs, the effective and efficient multimedia mining methods are needed.

The achievement of the multimedia mining methods is determined by the representation of MMDs. The appropriate representation of MMDs generates the useful patterns to benefit the multimedia mining methods and also reduces the search time and memory space requirements. The classic VSD and STD models are used to represent the texts. They are not suitable for representing the MMDs due to the multimodal nature of their content. The MMDs may contain similar texts, audios and images. It is very important to consider the similarity between multimedia objects while representing the MMDs. The similarity

between words are analyzed based on the synonyms. The similarity between media objects are computed based on their features. Hence, an effective multimedia document representation is needed to depict the MMDs based on the multimedia objects to improve the performance of multimedia mining methods.

The representation of multimedia documents depends on the effective use and representation of multimodal multimedia objects. Although, several approaches were proposed for utilizing the multimedia data, some challenging problems still exist and not well studied. Most of the multimedia fusion approaches were bimodal and face difficulty in extending the approach for multimodal objects. These methods managed each modality of data independently by distinct processing with distinct features. Concatenation of multimodal features suffers from the issues of dimensionality curse and missing modality. Correlation based approaches learn the correlations based on the unimodal features of each modality. Integrating the multidimensional, multimodal features is computationally expensive for the huge volume of MMDs containing multiple multimedia objects. Moreover, addition of a MMD for the dataset requires reintegration of all the multimedia objects. Independent processing of each type of data will neglect the interactions between them, which are essential to gain the full understanding of the concept of the MMD. Managing the multimodal features in multimodal space is the major challenge for MMD knowledge extraction methods. To overcome these drawbacks, the multimedia objects can be represented in a unified form. The unified representation allows the use of same feature extraction method resulting the features in unified feature space. The multimodal multimedia objects in a unified feature space not only ease the representation of MMDs, also benefits the multimedia mining tasks. Hence, an appropriate representation of multimedia objects in unified feature space is needed to assist the multimedia document representation and the multimedia mining methods. Overall, for the success of KDMD, the effective multimedia data representation, multimedia document representation and efficient multimedia mining methods are required.

2.5 Problem Statement

In order to solve the open issues discussed in the above section, effective multimedia document representations are needed for discovering the useful knowledge from the multimodal MMDs. Accordingly, the research problem is stated as follows:

"To develop effective multimedia document representations for efficient knowledge discovery in multimedia documents."

2.6 Research Objectives

The objectives of the research work are as follows:

- To develop effective multimedia data representation methods for unified representation of multimedia objects
- To develop effective multimedia document representations for unified representation of multimedia documents
- To develop efficient multimedia mining methods for knowledge discovery in multimedia documents

2.7 Scope

The ubiquity of MMDs has raised the demand for the effective use of embedded knowledge in various diverse applications. The KDMD is the process of discovering valid, novel, potentially useful, and ultimately understandable knowledge in the large amounts of MMDs. The existing knowledge discovery process is intended to discover the knowledge from text documents. The KDMD process requires the sophisticated procedures to manage, analyze and discover the multimodal knowledge from the MMDs. In this thesis the main emphasis is given to following aspects:

1. Representing the multimodal multimedia objects in a unified representation. The unified representation avoids the use of distinct processing, feature extraction and mining methods for each modality of data. It represents the multimodal multimedia objects in a unified feature space to benefit the process of multimedia document representation. Also, it facilitates the multimedia mining tasks in achieving the better result.

2. Representing the dataset of multimodal multimedia documents in a unified representation to benefit the mining tasks of KDMD process. The appropriate representation of MMDs extracts the useful patterns that help for multimedia mining tasks in discovering the required knowledge.

3. Developing the sophisticated multimedia mining methods to discover the knowledge from MMDs for the various applications.

Our main contributions reported in this thesis are:

- Developing multimedia data representation methods for representing the multimodal multimedia objects in a unified feature space.

- Demonstrating the efficacy of multimedia data representation methods with the classification and retrieval of MMDs.

- Developing a bio inspired clustering algorithm for clustering the MMDs to validate the effectiveness of the multimedia data representation methods.

- Proposing an information theory based similarity measure to find the pairwise similarity between MMDs.

- Developing a suffix tree based abstraction for the representation of MMDs.

- Developing a feature pattern tree based abstraction for representing the MMDs.

- Developing proposed abstraction based multimedia knowledge extraction methods.

2.8 General Methodology

In order to accomplish the objectives of the research work, the framework for the KDMD process comprising the multimedia data representation methods, multimedia document representations and multimedia mining methods is shown in Fig. 2.1. The research contributions described in the thesis are applicable to the dataset of MMDs with texts, images and audios. The brief description of research contributions to aid the KDMD process is given as follows:

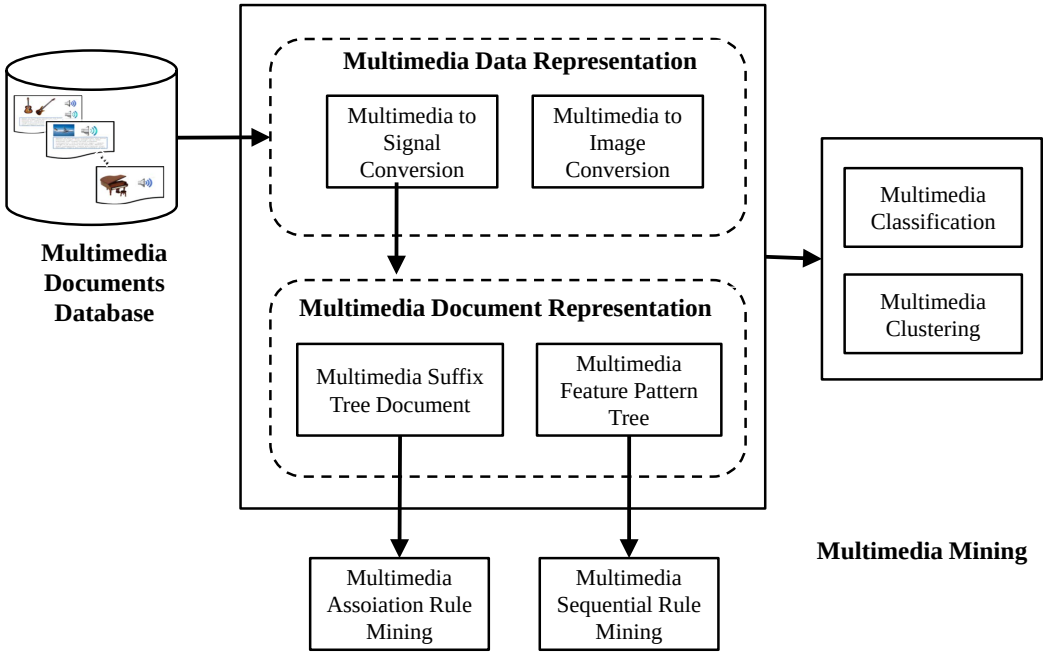


Figure 2.1: Framework for Knowledge Extraction in Multimedia Documents

2.8.1 Multimedia Data Representation

The multimodal multimedia objects of a MMD need to be represented in a unified feature space to help the multimedia mining methods. To represent the multimodal multimedia objects in a unified space, two multimedia data representation methods are proposed. The first method is Multimedia to Signal conversion that represents the multimedia objects in frequency domain by

converting the multimedia objects as signal objects. The second method is Multimedia to Image conversion that converts the multimedia objects as image objects in order to represent them in spatial domain. These two methods convert the MMD as unified multimedia document (UMD) by representing the multimedia objects texts, images and audios in unified space. The proposed multimedia data representation methods are evaluated for the classification and retrieval of MMDs. The effectiveness of the multimedia data representation methods has been evaluated for clustering the MMDs using the proposed Glowworm Swarm Optimization based multimedia documents clustering algorithm. In order to find the similarity between UMDs, an information theory based similarity measure has been proposed.

2.8.2 Multimedia Document Representation

The appropriate representation of MMDs extracts the useful patterns exist in the MMDs to benefit the multimedia mining methods in achieving the good performance. We propose two tree based representations, Multimedia Suffix Tree Document (MSTD) and Multimedia Feature pattern tree (MFPT) for the representation of UMDs. The MSTD representation represents the UMDs based on the shared similar media objects among the documents. The MFPT representation presents the UMDs based on the shared feature patterns of the media objects.

2.8.3 Multimedia Mining

Based on proposed multimedia document representations, multimedia mining methods are developed for efficient knowledge discovery. We developed MSTD and MFPT based classification for classifying the query multimedia documents. The developed MSTD and MFPT based clustering algorithms divides the set multimedia documents into clusters of same multimedia concepts. The MSTD based frequent multimedia pattern mining extracts the frequent multimedia

patterns that are used to derive the multimedia class association rules. Based on MFPT, sequential multimedia feature patterns are extracted that derives the multimedia class sequential rules for the classification of multimedia documents.

2.9 Multimedia Documents Datasets

The experimental analysis of our research work requires multimodal multimedia documents datasets. The performance analysis of proposed methods is evaluated using four multimodal multimedia documents datasets MDS1, MDS2, Flickr and MDDS. The datasets MDS1 and MDS2 were created by the I-SEARCH project for multimodal content indexing and retrieval (Rafailidis et al., 2013).

The dataset MDS1 contains 495 content objects of 10 categories with 3D objects, 2D images and sounds. The subset of 3D objects has been collected from the SHREC 2011 Generic Shape Benchmark (Li et al.) and the Princeton Shape Benchmark (Shilane et al., 2004) whereas the 2D images are the snapshots of corresponding 3D objects. The sounds were collected from publicly available websites and were manually attached to specific MMDs. The categories include various concepts birds, dogs, horse, cars, motorcycle, missile, guitar, airplane etc.

The dataset MDS2 comprises of 2779 multimedia documents belonging to 50 categories, consisting of 3D objects, 2D real images and text. The 3D objects have been collected from both the SHREC 2011 Generic Shape Benchmark (Li et al.) and the Princeton Shape Benchmark (Shilane et al., 2004). The 2D real images and text labels were collected from publicly available websites and manually attached to specific MMDs. The categories include various concepts airplanes, furniture, bird, airplane, helicopter, car, motorcycle, gun, ship, cellphone, bug, lamp, laptop etc.

The dataset Flickr is a subset of the public benchmark corpora MIRFlickr-25K dataset (Huang et al., 1997), contains 1200 documents of 10 categories. The dataset include real images and text labels of different concepts such as animal, bird, bridge, building, car, flower, river, sky and tree.

As these three datasets basically contain the bimodal data, we compiled a multimedia document dataset (MDDS) of 1000 MMDs for 20 concepts including animals, birds, airliners, motorcycles, cars and musical instruments using the real images, audio sounds and text documents. The images are collected from two benchmark image datasets (Fei-Fei et al., 2007)(Wang et al., 2013). The audio sounds and the text information related to the selected concept are collected from the Wikipedia and other publicly available websites and are manually attached to the specific documents. The MDDS dataset can be downloaded from the web link `http://dseg.nitk.ac.in/Project.html`. The maximum number of multimedia objects enclosed by a MMD for MDS1 is 2, for MDS2 is 5, for Flickr is 70 and for MDDS is 198.

2.10 Summary

This chapter provided the review of existing multimedia data representation, multimedia document representation and multimedia mining methods. The problem statement and objectives were framed based on the outcome of literature review. The scope and contributions of the present research work is presented. This chapter also provided the brief description of the MMD datasets used for the experiments of the present research work.

Chapter 3

Multimedia Data Representation

The multimedia objects need to be processed and represented in an abstract feature space for the efficient knowledge extraction from MMDs. In this chapter, we discuss two multimedia data representation methods for representing the multimedia objects in a unified feature space. We evaluate the effectiveness of multimedia data representation methods by the classification and retrieval of MMDs. Based on the proposed multimedia data representation methods, we propose a bio inspired clustering algorithm for clustering the MMDs and a similarity measure to find the pairwise similarity between the MMDs. Our contributions are:

- A Multimedia to Signal Conversion (MSC) for the unified representation of multimedia objects in frequency domain.
- A Multimedia to Image Conversion (MIC) for the unified representation of multimedia objects in spatial domain.
- A bio inspired Glowworm Swarm Optimization based clustering algorithm for the effective clustering the MMDs.

The rest of the chapter is organized as follows. In Section 3.1, we review the related work. In Section 3.2.1 and 3.2.2, we discuss the MSC and MIC method. Section 3.3 describes the characteristics of multimedia data representation methods. Section 3.4 presents the multimedia data representation based knowledge extraction methods for MMDs. The experimental results are

discussed in Section 3.5. In Section 3.6, we discuss the bio inspired clustering algorithm and its experimental evaluation. In Section 3.7, we discuss the similarity measure for MMDs and its experimental evaluation.

3.1 Related Work

Basically, all the multimedia objects are essentially digital signals. By applying the mathematical transformations, the signals can be mapped to frequency domain from their original domain (Zhang and Zhang, 2008). In recent years, some efforts have been made to convert the domain of the media objects. The text to speech (TTS) conversion methods has been discovered to convert the text to speech signal (Taylor, 2009). However, the TTS conversion is a complicated process as it requires the sequence of complex processes to process and convert text to speech. The image has been represented in frequency domain based on the changes in the image position to the corresponding changes in image intensity changing rate (Fisher et al., 1996). In (Cazan et al., 2007), the image has been converted into sound by mapping the pixel vertical position into frequency, horizontal position into time, and brightness into amplitude. The audio to image transform has been achieved using the wavelet transform for the application of audio steganography (Santosa and Bao, 2005). It is observed that these approaches have been used to convert the unimodal objects. Hence, there is a need of effective data representation method for representing the multimodal contents of MMDs in a unified feature space.

3.2 Multimedia Data Representation

We propose two multimedia data representation methods, Multimedia to Signal Conversion (MSC) and Multimedia to Image Conversion (MIC) to represent the multimedia objects in a unified domain. Our approaches are motivated by the above discussed unimodal domain conversion methods. The MSC method converts the multimedia objects as signal objects by representing in a frequency

domain. The MIC method represents the multimedia objects in a spatial domain by converting as image objects. Multimedia data representation methods represent the multimodal MMD as unified multimedia document (UMD) by converting the multimodal multimedia objects as unimodal multimedia objects. The unimodal contents of UMDs are subjected to feature extraction to extract the features and represent them in a unified feature space.

3.2.1 Multimedia to Signal Conversion

The framework for the proposed MSC method is depicted in Fig. 3.1. The framework contains three sections to process the multimedia objects texts, images and audios.

- **Texts**

  The text document is subjected to tokenization in order to extract the token words. The stop words are filtered from the extracted words and further refined by stemming (Porter, 1980). Initially, character sinusoidal signals are created for each character of a word. The signal is generated for each character of the word using the following equation:

$$s_c = sin2\pi f_c t_c \tag{3.2.1}$$

  where f_c and t_c are the frequency and time vector of the signal for each character c. The ASCII value of the character is the frequency (f_c) for the signal. The signal for a word is generated by concatenating the sinusoidal signals of its characters. In order to get unique signal for each word, it is necessary to generate a unique signal for each character of the word. The generated character signal needs to indicate a particular character of a word as well as its position in the word. For example, for a word "popular", the character signal of 'p' is different from character signal of 'o'. As the word "popular" contains two 'p's at two different locations the character signal

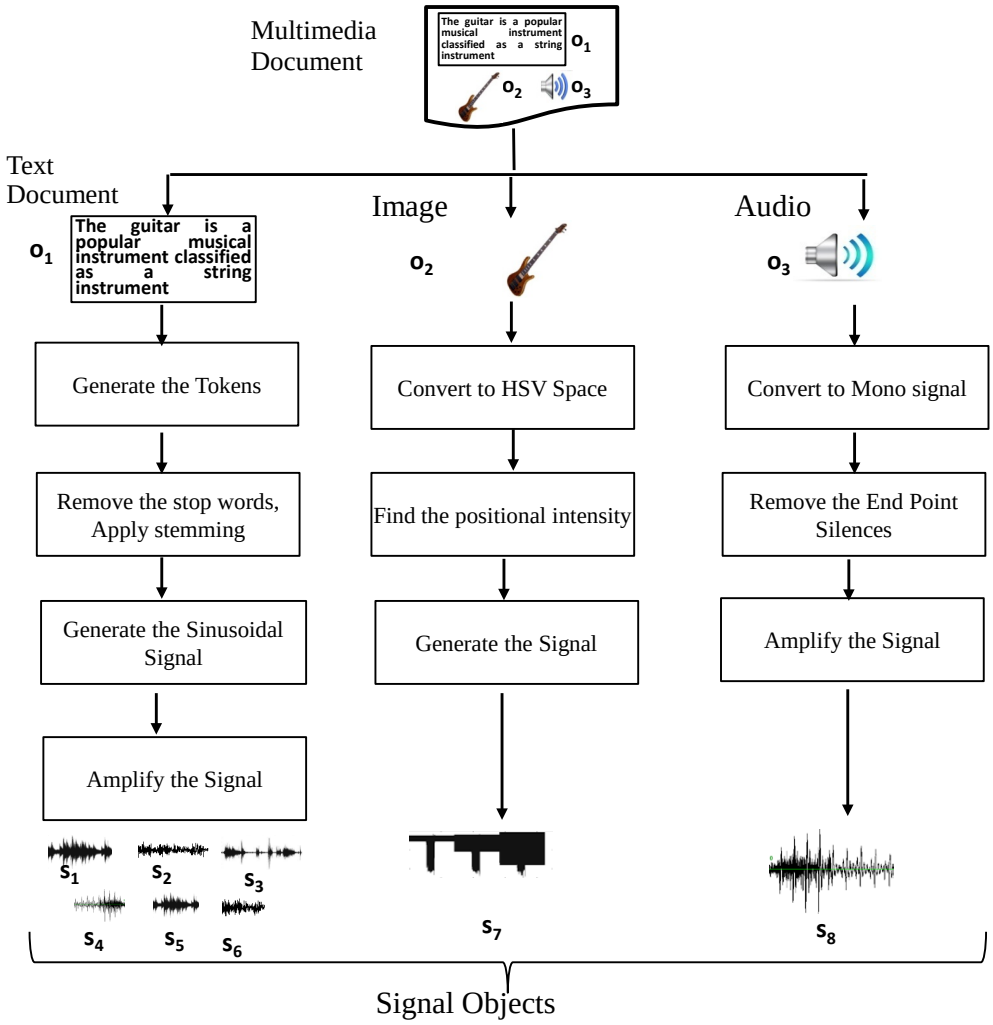


Figure 3.1: Multimedia to Signal Conversion

for 'p' at first position need to be different from the character signal 'p' at third position. In order to achieve this, based on the experimental analysis, we generated a time vector (t_c) for the character signal based on following formula:

$$t_c = (1 : \frac{ascii_{val}}{2 * char_{posn}} : ascii_{val}) \tag{3.2.2}$$

where $ascii_{val}$ is the ASCII value of the character and $char_{posn}$ is the position of the character in a word. The generated time vector indicates a particular character and its position in the word. The time vector generates $2 * char_{posn}$ time intervals starting from 1 to $ascii_{val}$.

In Fig. 3.1, the MSC tokenizes the text document o_1 to generate the words, “guitar”, “popular”, “musical”, “instrument”, “classified” and “string”. In order to convert a word “guitar” to signal object, initially the character signals are generated using the equation (3.2.1) for each character of “guitar” based on their position and ASCII value. For example, the ASCII of ’t’ is 116 and its position is 4. The time vector formed for ’t’ is [1 15.5 30 44.5 59 73.5 88 102.5] Then, the signal is generated for ’t’ using equation (3.2.1) with $f_c = 116$. Figure 3.2 shows the signal generated for character ’t’. Similarly all character signals are generated and concatenated to form a single sinusoidal signal of variable frequency over time for the word “guitar”.

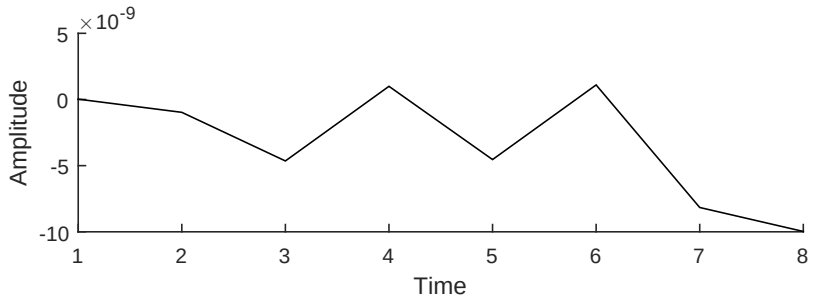


Figure 3.2: Signal form of character 't'

- **Images**

  Initially the RGB image is transformed into HSV (Hue, Saturation and Value) color space due to the potential of detaching the chromatic and achromatic components of the image. Moreover, the HSV color space is similar to the human conceptual understanding of colors. The signal form of HSV image is obtained by mapping the pixel position as time and intensity (Value) as amplitude for the signal.

  For example, let {100,120,250,245,23,30,60,255,25} are the intensity values of a 3*3 image. The signal form of image is formed by mapping pixel positions i.e. 1,2,3,..,9 in time axis with the corresponding intensity value as amplitude. The generated signal form is shown in Fig. 3.3. In Fig. 3.1, the image o_2 is converted as signal object s_7.

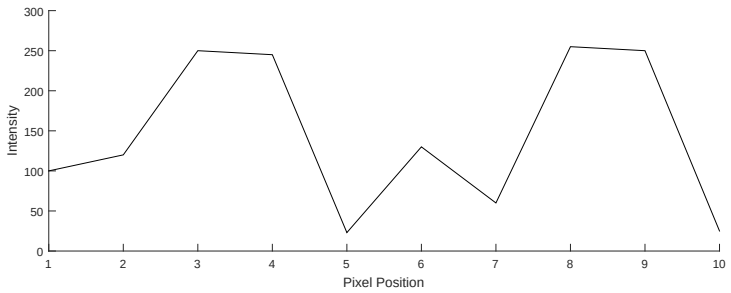


Figure 3.3: Signal form of an Image

- **Audios**

  As audio is signal by nature, it does not require any process to convert it into signal object. Initially audio signal is converted as a mono signal. Then, it is subjected to filtering to remove the noise and amplified to get the strong signal. The initial and end silences of the signal are filtered using the RMS (root mean square) value of the signal as a threshold. In Fig. 3.1, the audio signal o_3 is represented as signal object s_8.

Feature Extraction from Signal Objects

The MSC method converts each MMD as UMD and represents as a collection of m signal objects i.e. $UMD = \{s_1, s_2, ...s_m\}$. The UMD is considered as an unimodal document and the same feature extraction and mining techniques are applied. The signal objects can be subjected to any feature extraction methods that are used to extract the features from signals such as the Fourier Transform, Short Time Fourier Transform and Wavelet Transforms. In order to represent the signal objects in a unified feature space, we employed the wavelet transform (Mallat, 1989) to retrieve the features from signal objects. The wavelets are widely used in various image and audio processing applications (Bultheel and Huybrechs, 2011).

Wavelet is a waveform of effectively limited duration that has an average value of zero. Wavelet transform converts a signal into series of wavelets by providing a way for analyzing waveforms bounded in both frequency and

duration. The Continuous Wavelet Transform (CWT) analyzes the signal using wavelets by simple scaling and translation. CWT is a redundant transform, because it computes large number of redundant values both for the scale and for the translation. The Discrete Wavelet Transform (DWT) is used to decompose discrete time signals obtained by the discretization of the CWT in the time-frequency plane. The signal to be analyzed is passed through filters with different cut off frequencies at different scales. After each decomposition level, the signal is decomposed into approximation coefficients and three detail coefficients. The approximation coefficients are obtained by low pass filtering and the detail coefficients are obtained by high pass filtering the input sequence, followed by down-sampling. The frequency response for high-pass and low-pass filters determines the detail and approximation decomposition for different orders of wavelet transform. The sequence of approximation coefficients constitutes the input for the next iteration. Each decomposition level corresponds to a specified resolution. The resolution decreases with the increasing of the number of decomposition levels. Inverse DWT is used for the reconstruction of the approximation signal using the approximation and the detail from the previous resolution level. The most commonly used wavelets are Haar, Daubechies, Symlets, Shannon, Mexican Hat, Coiflet etc.

The Daubechies wavelet is the most popular wavelet used for signal and image applications (Bultheel and Huybrechs, 2011). It has the maximum number of vanishing moments for a given minimal compact support (Mallat, 2008). More vanishing moments implies higher compression, and smaller support implies less computation. The smaller the support of the wavelet is, the less of the signal it picks up in a certain wavelet coefficient. The Daubechies wavelet uses the overlapping windows, so the results reflect all changes between pixel intensities. It is orthogonal and supports the multiresolution analysis. It has balanced frequency response for both the analysis and reconstruction filters.

The signal objects from the UMD are subjected to three level Daubechies wavelet transform and decomposed into approximation coefficients and detail

coefficients. The dimensionality of wavelet coefficients varies depending on the length of signal object. In order to reduce the dimensionality of the wavelet coefficients we derived the statistics mean, standard deviation, variance, maximum and minimum from the wavelet coefficients and used as the features for the signal objects. Each of the signal object is represented as p-dimensional features $s_m = \{fs_1, fs_2, ..., fs_p\}$. The dimension p of the features depends on the decomposition level of the wavelet transform. The wavelet transform of level 3 generates 20 features i.e. five statistics are derived from each of one approximation coefficient and three detail coefficients. Figure 3.4 shows the representation of a MMD with multimodal objects as UMD with signal objects in unified feature space.

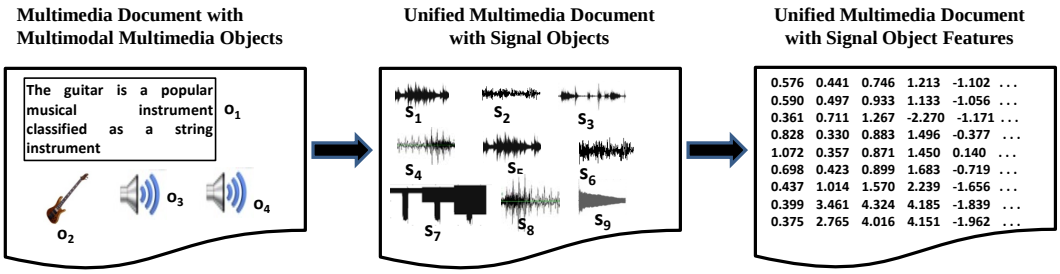


Figure 3.4: Representing MMD in Unified Feature Space using MSC

3.2.2 Multimedia to Image Conversion

The framework for the proposed MSC method is depicted in a Fig. 3.6. The framework contains three sections to process the multimedia objects texts, images and audios.

- **Texts**

  The text document is processed into words using tokenization. The words are reduced by filtering the stop words, further refined by stemming. Each extracted word is converted as a bitmap by combining the character bitmaps of all the characters of the word. For each character, a color bitmap of 20*18 pixel size is created using the predefined bitmaps of all alphanumeric

characters. The created character bitmap displays the character in colors. The background color of the character and the character color have been selected randomly. Finally the word bitmap is resized to a standard size of 256*256. Figure 3.5 shows the image form for word "guitar". In Fig. 3.6 the words ("guitar", "popular", "musical", "instrument", "classified" and "string") are extracted from text document o_1 converted into word bitmaps i_1, i_2, i_3, i_4, i_5 and i_6.

Figure 3.5: Image for a word "guitar"

- **Images**

  The images are itself represented by a spatial domain, so there is no additional processing is required to convert them as image objects. The RGB image is resized into a standard size of 384*256 to reduce the dimensionality of the image and then converted into HSV color space. In Fig. 3.6, the image o_2 is considered as i_7 after processing by the MIC method.

- **Audios**

  The audio signal is converted as a mono signal and subjected to filtering to remove the noise. The signal is amplified to get the strong signal. The initial and end silences of the audio signal are filtered using the RMS value of the signal as a threshold. The audio signal is converted as image by using the waveform of signal. The audio data samples are extracted from the audio waveform which undergoes normalization to get a positive waveform. The amplitude values of waveform are used as the positional intensity to plot the waveform image. The waveform plot has been converted into an image of standard size 640*480 pixels using the MATLAB software. In Fig. 3.6, the audio signal o_3 is converted as image i_7.

Feature Extraction from Image Objects

In general, the high dimensional images require the dimensionality reduction methods such as principal component analysis (PCA), Singular Value Decomposition (SVD), Local Linear Embedding (LLE), Isomap and Laplacian Eigenmaps to reduce the dimensionality before the extraction of features. However, the MIC method resizes the high dimensional images into standard size to reduce the dimensionality of the images. The MIC method converts the MMD as UMD and represents as a collection of m image objects i.e. $UMD = \{i_1, i_2, ...i_m\}$. In order to preserve the characteristics of color image objects, a composite feature vector is created for image objects by extracting

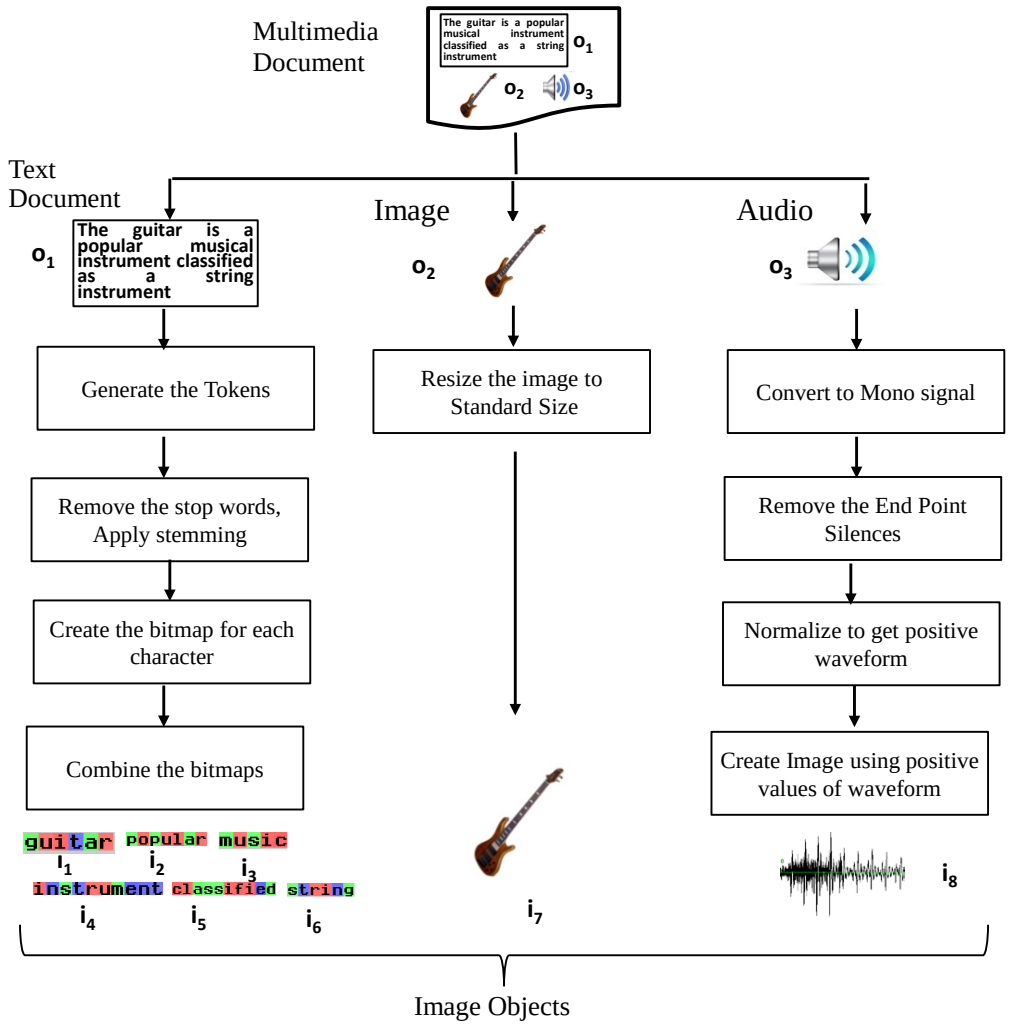


Figure 3.6: Multimedia to Image Conversion

color features and texture features of the image. The color features include HSV color histogram (Huang et al., 1997), color autocorrelogram (Huang et al., 1997), color moments (Stricker and Orengo, 1995) and texture features include Gabor filters (Grigorescu et al., 2002) and wavelet transform (Mallat, 1989).

- **HSV color histogram**

  From the HSV image H, S and V color components are extracted and quantized. For an $m * n$ image I, let $C = \{c_1, c_2, ..., c_k\}$ are the m quantized colors. The histogram h of I for color c_i can be defined as:

$$h_{c_i}(I) = \Pr\left[p \in I_{c_i}\right] \tag{3.2.3}$$

  where h_{c_i} represents the number of pixels in color c_i such that the color histogram for image I is $H(I) = \{h_1, h_2, .., h_k\}$.

- **Color Correlogram**

  A color correlogram provides the spatial correlation of pairs of colors changes with distance. It is used to describe the global distribution of local spatial correlation of colors. The correlogram of image I for color pair (c_i, c_j) is defined as:

$$\gamma^d_{c_i,c_j}(I) = \Pr_{p_1 \in I_{c_i}, p_2 \in I}\left[p_2 \in I_{c_j} \mid |p_1 - p_2| = d\right] \tag{3.2.4}$$

  where $\gamma^d_{c_i,c_j}(I)$ is the probability that a pixel p_2 of color c_j at distance d away from the given pixel p_1 is of color c_i.

  The autocorrelogram of an image captures spatial correlation between identical colors only and is defined by:

$$\alpha^c_d(I) = \gamma^d_{c,c}(I) \tag{3.2.5}$$

- **Color moments**

  Color moments are used to differentiate the images based on the color features. The three color moments provide the color distribution of the

images. The color moments are calculated from HSV color space. Any color distribution can be characterized by its moments. As the most information is concentrated on the low-order moments, the first moment (mean), the second moment (variance) and the third moment (skewness) are taken as the features for images.

Mean (E_i) is the average color value in the image and defined as:

$$E_i = \sum_{j=1}^{N} \frac{1}{N} p_{ij} \tag{3.2.6}$$

where p_{ij} is the j^{th} pixel at i^{th} color channel of image I.

The square root of the variance of the distribution is the standard deviation (σ_i) that is defined as

$$\sigma_i = \sqrt{\left[\frac{1}{N}\sum_{j=1}^{N}(p_{ij} - E_i)^2\right]} \tag{3.2.7}$$

Skewness (s_i) is a measure of the degree of asymmetry in the distribution,

$$s_i = \sqrt[3]{\left[\frac{1}{N}\sum_{j=1}^{N}(p_{ij} - E_i)^3\right]} \tag{3.2.8}$$

- **Gabor Filters**

  The Gabor Filters are well suited for texture segmentation because they possess the optimal localization properties in both spatial and frequency domain and in any number of dimensions. They are basically Gaussian functions modulated by complex sinusoidal of frequency and orientation. The Gabor features, texture energy and mean magnitude of the transform coefficients are calculated as given in (Grigorescu et al., 2002).

- **Wavelet Transform**

  Wavelet transform helps in detecting dominant points and analysis of local periodic pattern, so they are used to extract the texture features. They decompose the signal into approximation and detail coefficients. Approximation coefficients contain the original image details whereas the detail coefficients contain horizontal, vertical and diagonal details of the image. The horizontal and vertical details provide the information about the horizontal and vertical edges of the image. The mean and standard deviation of wavelet coefficients are computed and used as texture features for the image.

The above discussed features are concatenated and a high dimensional single feature vector is formed for each image object. Each feature discussed is are of different dimensions. For example, the HSV histogram generates 32 features for 32 quantized colors, the color moments generates 9 features (3 features for each color) and so on. These features are concatenated and a high dimensional single feature vector is formed for each image object. Each image object is represented as q-dimensional features $i_m = \{fi_1, fi_2, ..., fi_q\}$. Figure 3.7 shows the representation of a MMD with multimodal objects as UMD with image objects in unified feature space.

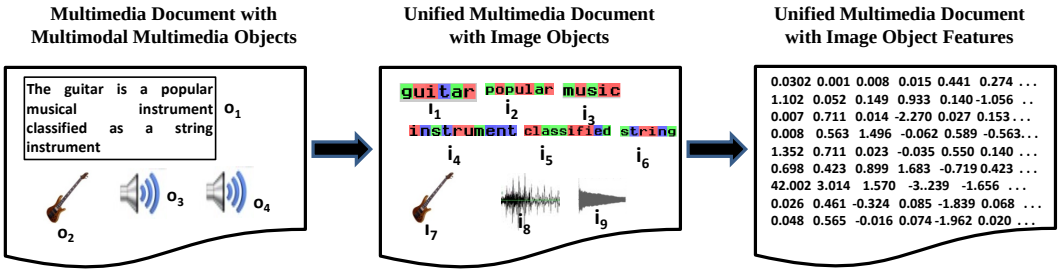


Figure 3.7: Representing MMD in Unified Feature Space using MIC

3.3 Characteristics of Multimedia Data Representation Methods

The proposed multimedia data representation methods retain the original characteristics of the image and audio objects. The similarity between the similar images or similar audios will be retained even after converting them into signal objects or image objects. The signal objects of images will never be similar to signal objects of words or signal objects of audios. Similarly, the image objects of audios will never be similar to image objects of words or image objects of images. Moreover the signal or image objects formed from different words will never be similar to each other. Although it is possible to create the similar signal or image objects for semantically similar words, it requires additional processing for the discovery of semantic meaning.

With text documents, the similarity between the words is considered based on the synonyms. The MMD is originally comprised of images, audios and text documents, there is a possibility of having similar images and similar audios. Instead of handling similar objects independently, it will be useful if the similar image and audio objects are considered as same object. To find the similar image and audio objects, the similarity between the objects is computed. The objects are considered similar when the similarity between the features of objects satisfies a user defined object similarity threshold ($thresh_{ob}$). The object similarity threshold value specifies the percentage similarity between the features of two objects.

The similarity between the objects is calculated using the percentage difference (PD) measure. The main advantage of using PD measure over the other distance measures such as city block, Euclidean, Cosine and Canberra is considering the similarity of individual features. It also considers the magnitude of feature value. Moreover, the value of PD measure is normalized by default to lie within [0,1].

The PD measure between two objects o_i and o_j is calculated as follows:

$$pd_{ij} = 100 * \frac{1}{p} \sum_{k=1}^{p} \frac{2 * |(o_{i_k} - o_{j_k})|}{(|o_{i_k}| + |o_{j_k}|)} \tag{3.3.1}$$

where p is the number of features of the objects o_i and o_j.

3.4 MSC and MIC based Knowledge Extraction from Multimedia Documents

To evaluate the effectiveness of the proposed multimedia data representation methods, experiments are conducted for the classification and retrieval of MMDs. The framework of knowledge extraction from MMDs using the proposed multimedia data representation methods is shown in Fig. 3.8. In Section 3.4.1, we discuss the classification MMDs using the MSC and MIC methods. In Section 3.4.2, we discuss the retrieval of MMDs using the MSC and MIC methods.

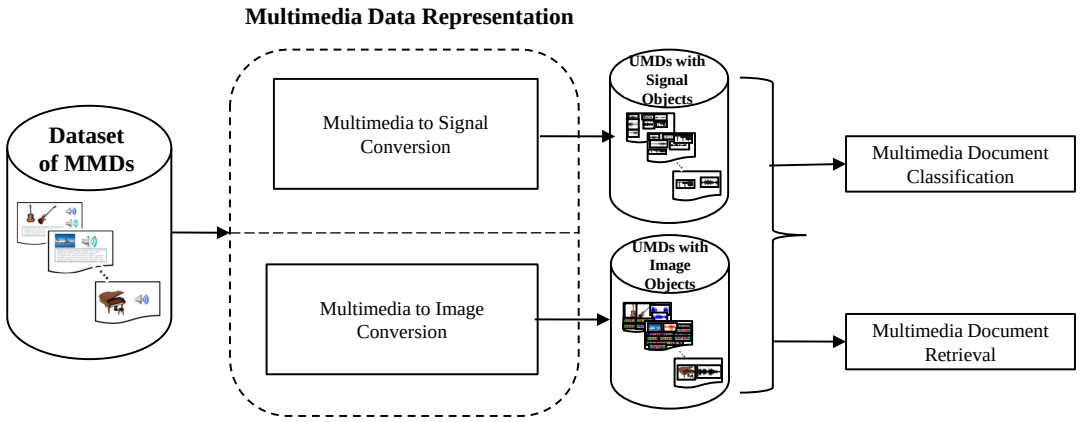


Figure 3.8: Framework for Knowledge Extraction from MMDs using the MSC and MIC methods

3.4.1 Multimedia Document Classification

In this section, we discuss how the MSC and MIC methods support the classification of MMDs. The classification of MMDs is performed by assigning the class labels for the set of test MMDs. The classification task retrieves the MMD that has maximum document similarity with the test MMD. The test MMD is correctly classified when the concept of test MMD and retrieved MMD is same. The MMDs are converted as UMDs using the proposed multimedia data representation methods and classification is performed.

Let $UMD_T = \{umd_1, umd_2, ..., umd_N\}$ be the training UMDs that have known class labels. For each test UMD umd_q, the most similar UMD umd_r is found by computing the similarity between the umd_q and $umd_t \in UMD_T$ using the equation as follows:

$$umd_r = \{umd_k : k = \underset{1 \leq t \leq N}{\operatorname{argmax}} [sim(umd_q, umd_t)]\} \tag{3.4.1}$$

where $sim(umd_q, umd_t)$ is the similarity between the umd_q and umd_t. The umd_t that has maximum similarity with umd_q is considered as the most similar UMD to umd_q. The similarity between umd_q and umd_t is computed using the Dice Coefficient as follows:

$$sim(umd_q, umd_t) = 2 * \frac{|umd_q \bigcap umd_t|}{|umd_q| + |umd_t|} \tag{3.4.2}$$

where $umd_q \bigcap umd_t$ denotes the similar objects that exist in both the umd_q and umd_t. The similar objects has been identified based on the similarity of the features calculated using the percentage difference function given as (3.3.1).

The efficiency of multimedia classification is evaluated by computing the accuracy of the classifier. The accuracy for M query UMDs is computed using the formula:

$$Accuracy = \frac{\sum_{M}^{i=1} Class\_Result}{M} \tag{3.4.3}$$

where $Class\_Result$= 1 when umd_q and retrieved similar UMD umd_r have the same concept, otherwise 0. Higher accuracy proves the efficiency of a classifier. The accuracy value for a perfect classifier is 1, i.e. the classifier correctly classifies all the test MMDs.

3.4.2 Multimedia Document Retrieval

The proposed MSC and MIC methods support the retrieval of MMDs. The MMD retrieval process retrieves the MMDs that are similar to query document. The retrieval process has been experimented for internal and external queries. Internal queries are the MMDs that belong to the training dataset whereas the external queries do not belong to the training dataset.

Let $UMD_T = \{umd_1, umd_2, ..., umd_N\}$ be the training UMDs. For each query umd_q, the UMD $umd_t \in UMD_T$ is considered relevant when their similarity satisfies the user defined document similarity threshold. The similarity, $sim(umd_q, umd_t)$ is computed using the Dice Coefficient according to (3.4.2). The effectiveness of the multimedia information retrieval is evaluated using the two standard measures, precision and recall (Manning et al., 2008). Precision measures the fraction of the returned documents that are relevant to the query and the recall measures the fraction of relevant documents in the collection was returned by the system. The precision and recall are calculated as follows:

$$Precision = \frac{No.\,of\,relevant\,MMDs\,retrieved}{No.\,of\,MMDs\,retrieved} \tag{3.4.4}$$

$$Recall = \frac{No.\,of\,relevant\,MMDs\,retrieved}{No.\,of\,relevant\,MMDs\,in\,the\,dataset} \tag{3.4.5}$$

The perfect precision value of 1 indicates that every retrieved MMD is relevant whereas the perfect recall value of 1 indicates that all relevant MMDs are retrieved.

3.5 Experimental Results and Discussion

This section discusses the results of multimedia knowledge extraction using the proposed multimedia data representation methods. The proposed multimedia data representation methods are evaluated by performing the experiments for the classification and retrieval of multimedia documents. The experiments are conducted with four multimodal multimedia documents datasets MDS1, MDS2, Flickr and MMDS. The detailed description of the datasets are given in Section 2.9. Section 3.5.1 presents the time taken by the MSC and MIC methods for representing the multimedia objects of four multimodal datasets in a unified feature space. Section 3.5.2 presents the MMD classification results. In Section 3.5.2, we discuss the results of MMD retrieval.

3.5.1 Results of Multimedia Data Representation Methods

This section compares the time taken by the MSC and MIC methods to represent the multimodal multimedia objects in a unified feature space. Table 3.1 shows the details of time taken for the multimedia data representation for each dataset. It is observed that the conversion time taken by MSC method is very small compared to MIC method. The time taken to convert image to audio by MSC method is very less compared to audio to image by MIC method. Moreover, word to audio conversion time is less compared to word to image conversion time. The total conversion time taken by MSC method is 6.96 sec for MDS1, 86.28 sec for MDS2, 30.97 sec for Flickr and 29.89 sec for MDDS. The total conversion time taken by MIC method is 85.25 sec for MDS1, 12.45 sec for MDS2, 13.55 sec for Flickr and 120.96 sec for MDDS. It is noticed that the conversion time required for the datasets depends on the modality and the number of the multimedia objects. The datasets that are dominated by the images such as MDS2 and Flickr take less conversion time with MIC. The datasets MDS1 and MDDS that have more audios and words have less conversion time with MSC method.

Table 3.1: Time taken by the MSC and MIC methods in sec

Datasets	Objects	MSC		MIC	
		Conversion Time	Feature Extraction Time	Conversion Time	Feature Extraction Time
MDS1	Images	5.89	3.39	2.49	85.97
	Audios	1.07	3.47	82.76	94.86
MDS2	Words	0.74	8.40	3.05	531.65
	Images	85.54	46.04	9.40	371.77
Flickr	Words	2.58	33.06	10.41	954.61
	Images	28.39	13.78	3.14	127.20
MDDS	Words	2.27	38.20	9.69	904.31
	Images	25.17	10.77	1.59	89.54
	Audios	2.45	6.64	109.68	121.88

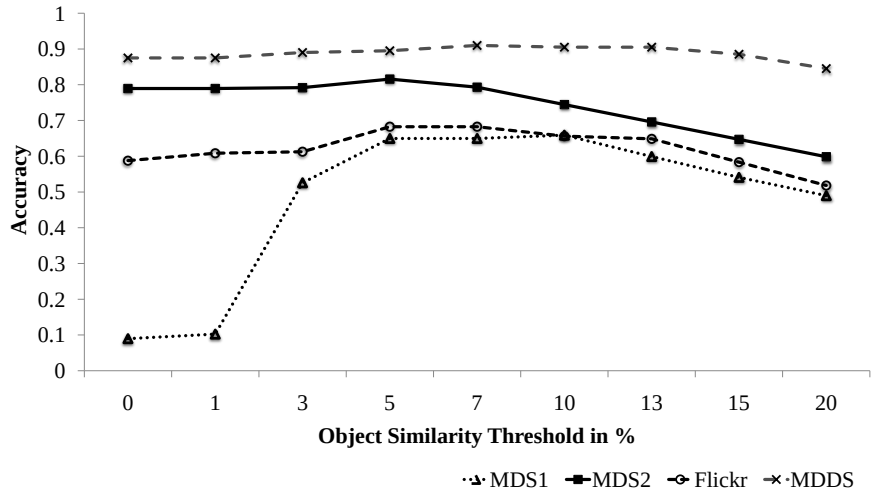
The signal objects are subjected to three level Daubechies wavelet transform, statistics are computed to extract the features and a low dimensional feature vector is formed for signal objects. A high dimensional feature vector is formed for image objects by extracting the color features and texture features of the image. The results show that the feature extraction time of image objects is more compared to the feature extraction time of signal objects due to its high dimensional features. The MSC method takes the total feature extraction time of 6.86 for MDS1, 54.44 for MDS2, 46.84 for Flickr and 55.61 for MDDS. The total feature extraction time by MIC method is 180.83 for MDS1, 903.42 for MDS2, 1081.81 for Flickr and 1115.73 for MDDS. Hence, sophisticated methods are needed for the conversion of multimedia objects to image objects. Also effective feature extraction and dimensionality reduction methods are required for image objects.

3.5.2 Multimedia Documents Classification Results

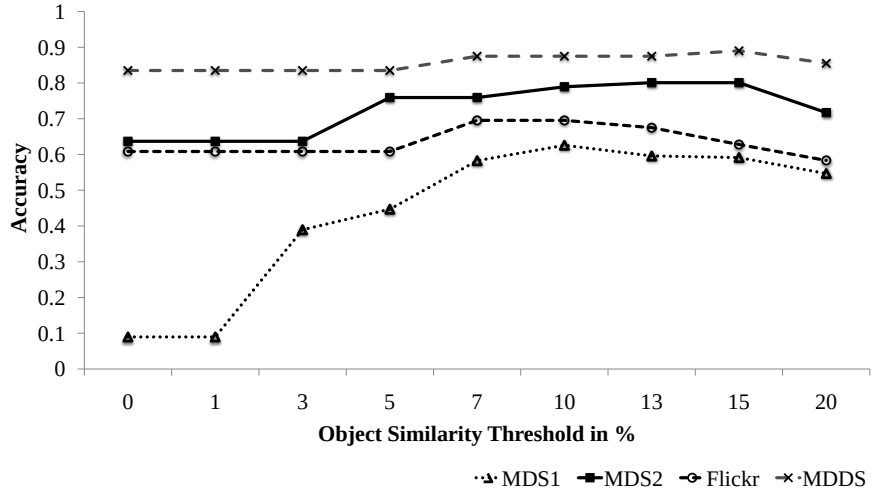
The classification of MMDs is performed by assigning the class labels for the set of test MMDs. The four multimodal MMD datasets are split into training and test datasets in the ratio of 80%-20%. All the experiments are conducted for various object similarity threshold values in %, $thresh_{ob} = \{0, 1, 3, 5, 7, 10, 13, 15, 20\}$. The

similarity between signal objects and image objects is computed using PD measure according to (3.3.1).

Figure 3.9 demonstrate the performance of the classification of UMDs using MSC and MIC methods for various object similarity values. The best accuracy values obtained using the MSC method for datasets MDS1, MDS2, Flickr and MDDS are 0.66, 0.82, 0.68, and 0.91 respectively. The best accuracy obtained using MIC method are 0.63 for MDS1, 0.80 for MDS1, 0.70 for Flickr, and 0.89 for MDDS. It is observed that the lower values of accuracy are obtained for lower and higher values of $thresh_{ob}$. The lower value of $thresh_{ob}$ may not retrieve the



(a) Classification using MSC method



(b) Classification using MIC method

Figure 3.9: Classification of Multimedia Documents

relevant MMDs for all the test MMDs due to the nonavailability of similar objects, so obtains the low accuracy. The higher values of $thresh_{ob}$ may retrieve more MMDs, but all of them may not be relevant to test MMDs resulting in lower accuracy. The optimum value of $thresh_{ob}$ retrieves more relevant MMDs resulting in a higher accuracy value. It is observed that optimum value of $thresh_{ob}$ is different for different dataset as it depends on the modality of the multimedia objects of the datasets. The values of $thresh_{ob}$ at which best accuracy achieved are 10% for MDS1, 5% for MDS2, 5% for Flickr, and 7% for MDDS. The values of $thresh_{ob}$ for MIC method that achieved best accuracy are 10% for MDS1, 13% for MDS2, 7% for Flickr, and 15% for MDDS.

Figure 3.10 shows the performance comparison of classification of MMDs for the MSC and MIC methods. The results demonstrate the achievement of MSC method over MIC method with the improvement of 3% for MDS1, 1% for MDS2 and 2% for MDDS. With Flickr, the MIC method shows the improvement of 2% as it has dominated by real color images. It is observed that the MDDS dataset performed significantly better compared to other datasets. The MDDS dataset has all the three modality of data that gives more information for the concept. Hence the classification performance of MDDS is better compared to other datasets.

3.5.3 Multimedia Document Retrieval Results

The performance of the multimedia data representation methods has been evaluated by conducting the experiments for multimedia documents retrieval in terms of precision-recall. The results of multimedia document retrieval are compared with the results of manifold learning SMMD method (Daras et al., 2012). The SMMD method has been experimented with datasets MDS1 and MDS2 for the retrieval of MMDs using internal and external queries. For internal queries, each MMD of the dataset is considered as a query and for external queries 10% of MMDs of the dataset are posed as query. The precision and recall values are computed according to (3.4.4) and (3.4.5) respectively.

We conducted experiments for the retrieval of MMDs with datasets MDS1

and MDS2 using various object similarity threshold values in %, $thresh_{ob} = \{0, 1, 3, 5, 7, 10, 13, 15, 20\}$. Based on the experimental analysis, we set a lower bound of 0.5 as the document similarity threshold. The individual precision-recall is computed for each query and then the average precision-recall is extracted for each object similarity threshold value. The precision-recall curves are drawn for each dataset by interpolating the average precision values for 11 standard recall values.

The retrieval performance of MSC and MIC methods are compared with the SMMD method for MDS1 and MDS2. The performance comparison of the methods is shown in Fig. 3.11. In case of internal queries, the MSC method achieved the maximum precision improvement of 5% for MDS1 and 4% for MDS2 whereas MIC method achieved the maximum precision improvement of 2% for MDS1 and 4% for MDS2 compared to SMMD method. For external queries, the MSC method attained the maximum precision improvement of 1% for both datasets and MIC method attained the maximum precision improvement of 1% for MDS2 over the SMMD method. However, the SMMD

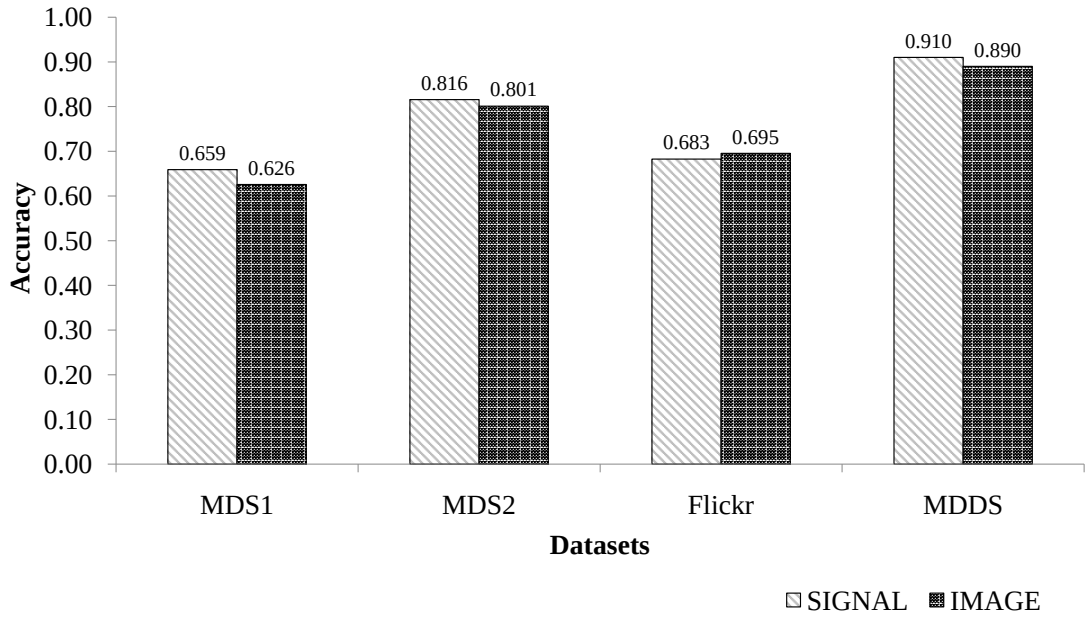
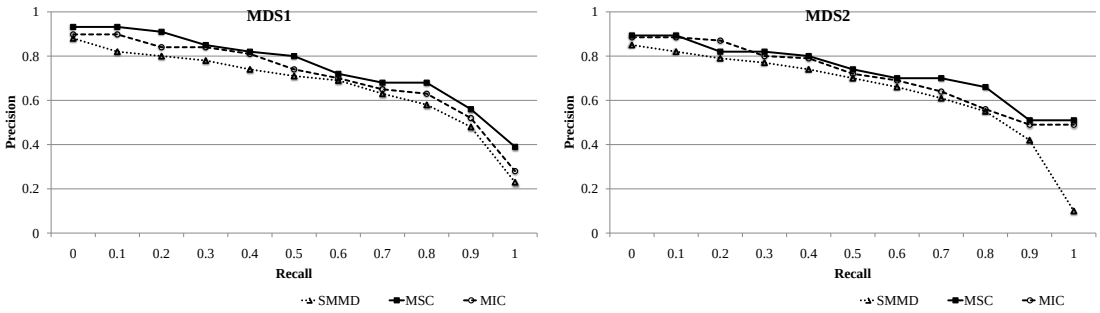
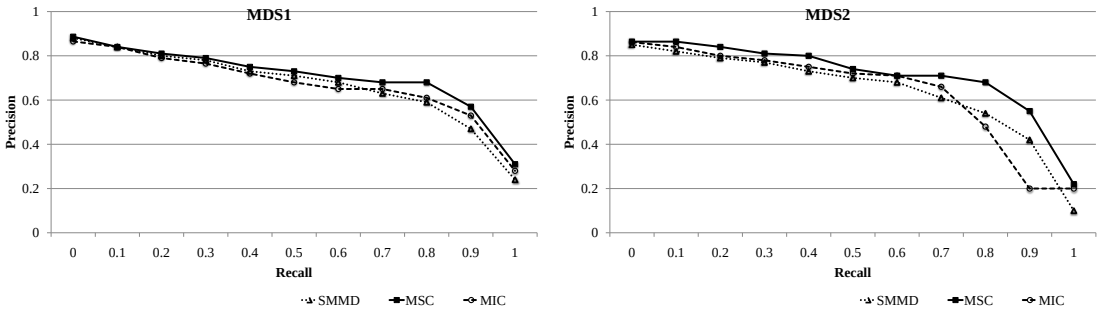


Figure 3.10: Performance Comparison of Classification of MMDs using MSC and MIC methods



(a) Precision-Recall for Internal Queries



(b) Precision-Recall for External Queries

Figure 3.11: Comparison of MSC and MIC methods with SMMD method for MMD Retrieval

method performs 1% better than MIC method for MDS1. The SMMD method searches the MMDs using a low dimensional feature space which was built based on the assumption that each modality of object of each MMD has same number of neighbors. However, this assumption may degrade the performance as the multimedia objects of different modality may have different number of neighbors. Our method retrieves the similar documents based on the similarity of multimedia objects. Hence, our method achieved better performance compared to SMMD method. The improved performance indicates that the multimedia data representation methods help to improve the efficiency of the MMD retrieval methods.

3.6 The Glowworm Swarm Optimization Based Multimedia Documents Clustering

Nowadays nature has inspired the researchers to develop the nature inspired algorithms and use them successfully for the extraction of knowledge. Swarm intelligence algorithms are the inspiration from behavior of social natural communities such as birds flocks, ant colonies, fish schools, glow worms, fireflies etc. These algorithms accomplish their task depending on the receptor of the individual's interactions while communicating with each other. Among the various Swarm intelligence algorithms, Glowworm Swarm Optimization(GSO) (Krishnanand and Ghose, 2006) algorithm has been inspired by the intelligence in lighting behavior of the worms. The glowworms have the capability to control their light emission and use the emitted light for various purposes like attracting the prey, attracting mates etc. GSO has been used in several applications such as mobile sensor networks and robotics (Krishnanand and Ghose, 2009), clustering the documents (Aljarah and Ludwig, 2013), for sensor deployment scheme in wireless sensor networks (Liao et al., 2011) and to find the optimal solution for multiple objective environmental economic dispatch (MOEED) problem (Jayakumar and Venkatesh, 2014). The majority of the swarm intelligence algorithms motivates to locate the global solution for the given optimization problem. Besides that, the GSO algorithm was the first swarm intelligence algorithm used for optimizing multimodal functions with equal or distinct objective function values.

3.6.1 Glowworm Swarm Optimization Algorithm

The GSO algorithm has the ability to simultaneously capture the multiple optima of multimodal functions with distinct objective function values. In GSO algorithm, the swarm of glowworms are scattered in search space. Each glowworm carries its own bio luminescence substance known as luciferin. The

glowworm releases luciferin to make itself visible for other glowworms within its vision range. The intensity of luciferin is associated with the objective function of glowworm's location. The vision range for each glowworm varies depending on the amount of luciferin released. All the glowworms search for their neighbors within their vision range and then move towards the brighter glowworm within their neighbor set. In each iteration, the position of the glowworm changes and then the luciferin value also gets updated. Hence at the end, most of the glowworms group together to make the compact clusters in the search space at multiple optimal solutions.

Initially, all the glowworms have an equal amount of luciferin value. In each iteration the glowworm undergoes through luciferin update phase and movement phase. In luciferin update phase, the objective function is evaluated at current glowworm position and glowworm's luciferin value updates depending on new objective function values. During movement phase, each glowworm explores its vision region to attract the neighbors that have higher luciferin value. The glowworm moves towards the best neighbor glowworm from the neighborhood set using the probability mechanism. The position of glowworm is updated depending on selected neighbor glowworm. At the end of the iteration, the vision range of glowworm is updated. One of the important characteristics of GSO algorithm is the use of variable vision range that decides the number of neighbors. The vision range increases when the neighbor density is lower and decreases when the neighbor has higher density.

Motivated by the strength of optimization and dynamic nature, we employed the GSO algorithm for finding the optimal solution for clustering the MMDs.

3.6.2 GSO based Multimedia Documents Clustering Algorithm

The proposed GSO Based Multimedia Documents Clustering Algorithm (GSOMDC) uses the GSO technology to partition the UMDs into a set of

clusters. Most of the clustering algorithms partition the dataset by extracting centroids based on the features of multimedia objects. Multimedia document is a collection of multimodal objects. The signal objects retain the original characteristics of the multimodal objects. Since the features of signal objects of different modality have different characteristics, the feature based centroids extraction is not practical for MMDs. Hence, centroid based clustering is not applicable for MMDs. Clustering the MMDs mainly relies on the similar objects between the MMDs. The proposed algorithm groups the UMDs into clusters such that the UMDs within the cluster have similar information.

The GSOMDC algorithm partitions the given UMDs, $UMD = \left\{ umd_i \mid_{i=1}^{N} \right\}$ into a set of K clusters $C = \{C_1, C_2..C_K\}$. The algorithm tries to maximize the similarity of the UMDs within the cluster and minimize the similarity between the UMDs of different clusters. In addition, each cluster must have at least one UMD within it. As our research work is dealing with MMDs of single concept, the algorithm doesn't generate the overlapped clusters. The generated clusters should not be empty and should be disjoint such that $\bigcap_{i..k} C_i = \{\}$ and $\bigcup_{i..k} C_i = UMD$.

In the initialization phase, the collection of UMDs are considered as the glowworm swam $S = \left\{ g_i \mid_{i=1}^{N} \right\}$ and each glowworm g_i acts as an agent of its own cluster in the virtual search space. The glowworm agent is responsible for the movement of glowworms and luciferin update. The glowworm agents are initialized with the number of objects found in each document as luciferin value. In each iteration, the luciferin value and vision range of each glowworm agent gets updated. The vision range of glowworm agent depends on its luciferin value. In the classical GSO algorithm, at the beginning all the glowworm agents carry same value of luciferin value. After each iteration, the luciferin value and vision range gets updated. However in the proposed algorithm, as per the real life scenario the glowworms carry the variable luciferin value depending on the number of objects belongs to them in each cluster. After initializing the swarms with the luciferin value, the glowworm agent g_i with the lower luciferin value is

selected as a candidate to search for a glowworm within its vision range and with same or higher value of luciferin. The selected glowworm agent g_i extracts the set of neighbor glowworms N_i within its vision range based on the following rule:

$$N_i = \{j : sim_{ij} < vr_i; L_i \leq L_j \ \& L_j <= 3 * L_i\} \tag{3.6.1}$$

where L_i, L_j are the luciferin values of glowworm agents g_i and g_j, $vr_i = \{max(th_{sim}); 0.5 \geq th_{sim} \leq max(sim_{ij})\}$ is the vision range of glowworm agent g_i and sim_{ij} is the similarity value between g_i and g_j. The similarity value sim_{ij} is calculated as follows:

$$sim_{ij} = 2 * \frac{\sum_{x=1}^{m} \sum_{y=1}^{n} cnt_{xy}}{L_i + L_j} \tag{3.6.2}$$

where $cnt_{ij} = \begin{cases} 1 & : pd_{xy} \geq thresh_{ob} \\ 0 & : otherwise \end{cases}$ is the count of the similar objects present in glowworms g_i and g_j and pd_{xy} is the percentage difference between the two objects $o_x \in g_i$ and $o_y \in g_j$. The $thresh_{ob}$ is the threshold for object similarity. The percentage difference between the objects o_x and o_y is computed as per the formula (3.3.1).

The glowworm g_i considers only those glowworms that have their luciferin value satisfying the criteria $L_j <= 3 * L_i$ for the selection of neighbors. In a large dataset of documents, finding the similarity between all the neighbors is time consuming. Moreover, the glowworms with higher luciferin value will have very less similarity with glowworms with lower luciferin value. To avoid computing the similarity with all the glowworms, the criteria $L_j <= 3 * L_i$ is used to select the glowworms such that the similarity between the two glowworms will be at least 0.5. The minimum threshold for the object similarity value is restricted to 0.5 to ensure that the more similar MMDs are clustered together.

For example,

Let the luciferin value L_1 for glowworm g_1 is 4. It can consider the glowworms with luciferin value of 4 or more for neighbor selection. Let the

luciferin values for glowworms g_2, g_3, g_4, g_5 and g_6 are L_2=4, L_3=8, L_4=10, L_5=12 and L_6=13 respectively. It is assumed that the selected neighbors has the objects similar to all four objects of g_1. As per equation 3.6.2, the sim(g_1,g_2)=1, sim(g_1,g_3)=0.67, sim(g_1,g_4)=0.57, sim(g_1,g_5)=0.5 and sim(g_1,g_6)= 0.47. The analysis of the similarity values indicates that the glowworm (g_6) which has luciferin value higher than L_1*3, cannot be the neighbor for glowworm with luciferin value 4, as their similarity value is less than 0.5.

Among the selected neighbors, the glowworm agent g_i selects the best neighbor that has higher or equal to its luciferin value based on the probability calculated using the following equation:

$$prob_{ij} = \frac{L_j - L_i}{\sum_{k \in N_i} L_k - L_i} \tag{3.6.3}$$

The glowworm agent g_i moves towards the cluster of the best neighbor glowworm agent g_j and becomes the member of it. while moving, glowworm agent g_i also moves the members of its cluster towards the cluster of glowworm agent g_j. The procedure continues till all the glowworms selects the best neighbor depending on their luciferin value and vision range. After each iteration the luciferin value of the glowworm agent of each cluster is updated to the number of objects of all the documents enclosed by the cluster. The procedure repeats till the number of iterations reached the maximum iterations. The algorithm terminates when there is no movement of glowworm agents. The GSOMDC algorithm is summarized in algorithm 3.1.

3.6.3 Experimental Results and Discussion

The effectiveness of the multimedia document clustering is evaluated using the two standard measures Purity and Entropy (Manning et al., 2008). Purity measures how well the clustering algorithm will be able to reproduce the classes. Entropy indicates the homogeneity of the cluster. It measures the distribution of various classes within each cluster.

The weighted average clustering purity for a set of N multimedia documents is calculated as follows:

$$Purity = \sum_{k} \frac{|c_j|}{N} \max_{j} |c_k \cap t_j| \tag{3.6.4}$$

where $C = \{c_1, c_2, ..., c_k\}$ are the number of clusters for N multimedia documents and $T = \{t_1, t_2, ..., t_j\}$ is the set of multimedia concepts. The weighted average entropy of the generated clusters is calculated as follows:

$$Entropy = -\frac{1}{logk} \sum_{j=1}^{k} \frac{|c_j|}{N} \sum_{i=1}^{l} p_{ij} log(p_{ij}) \tag{3.6.5}$$

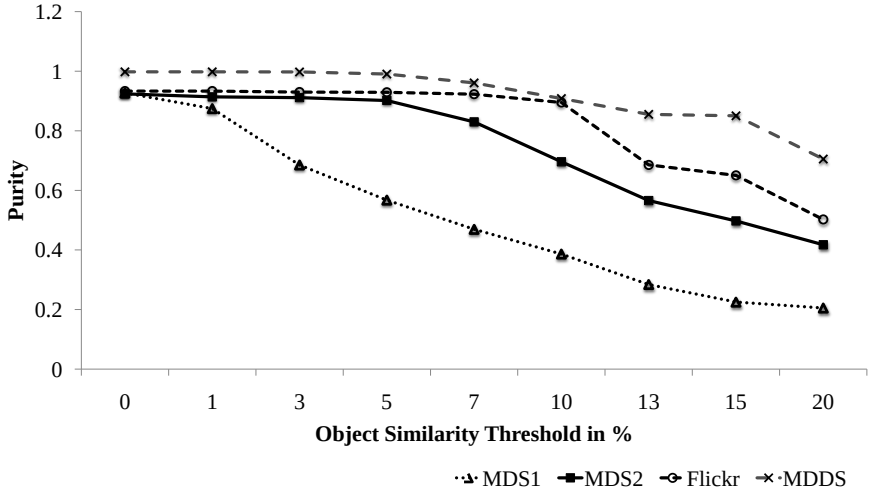
where p_{ij} is the probability that a member of cluster c_j belongs to concept t_i. Better clustering algorithms results in larger purity value and smaller entropy value. The purity and entropy value for a perfect clustering is 1 and 0 respectively.

The proposed GSOMDC algorithm is experimented with four multimodal datasets using MSC and MIC methods. The details of datasets are given in Section

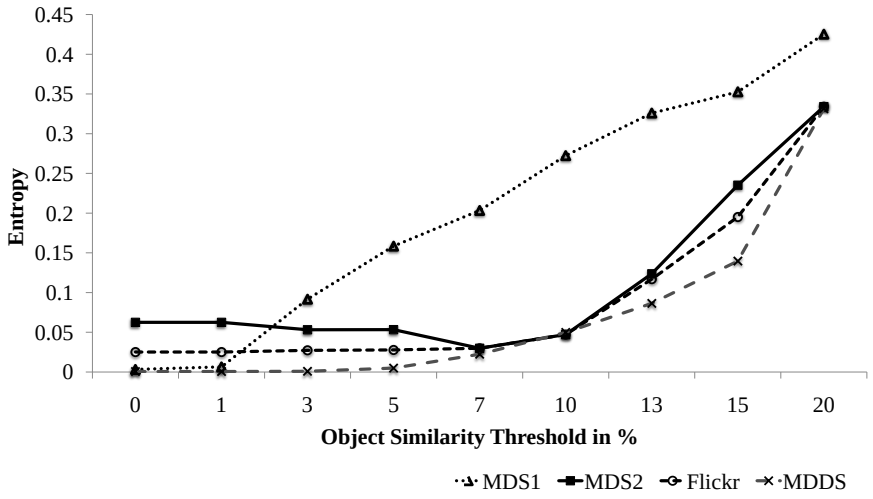
Algorithm 3.1 GSOMDC Clustering

Input: $\{UMD = umd_1, umd_2, ..., umd_N\}$, maxiterations
Output: Clusters of multimedia documents
repeat
    for each multimedia glowworm agent $g_i \in UMD$ **do**
        Initialize the number of objects present in the cluster of g_i as its luciferin value L_i
    end for
    Arrange the glowworms with respect to their luciferin (L) value in ascending order
    for each multimedia glowworm agent $g_i \in UMD$ **do**
        Find the neighbors g_j of g_i using the rule (3.6.1)
        Calculate the probability of each of the neighbor j $\in N_i$ using the formula (3.6.3),
        Select the best neighbor g_j having the minimum probability
        Move g_i towards the cluster of g_j and become its member .
    end for
    iterations=iterations+1
until no glowworms to merge or $iterations \leq maxiterations$

2.9. The efficiency of the proposed GSOMDC algorithm is explored by computing the purity and entropy values for the generated clusters. The maximum limit for the number of iterations is kept as 10 by the experimental analysis. The purity and entropy values are computed for various object similarity threshold values in % , $thresh_{ob} = \{1, 3, 5, 7, 10, 13, 15, 20\}$ for four datasets are shown in Fig. 3.12 and 3.13. The better purity and entropy values are obtained for the object similarity threshold value between 1 to 10. With the MSC method, the best purity value

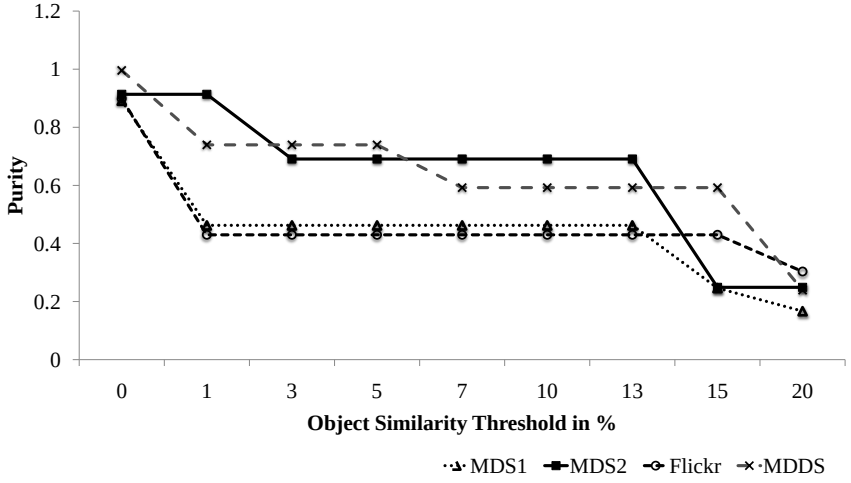


(a) Purity Values

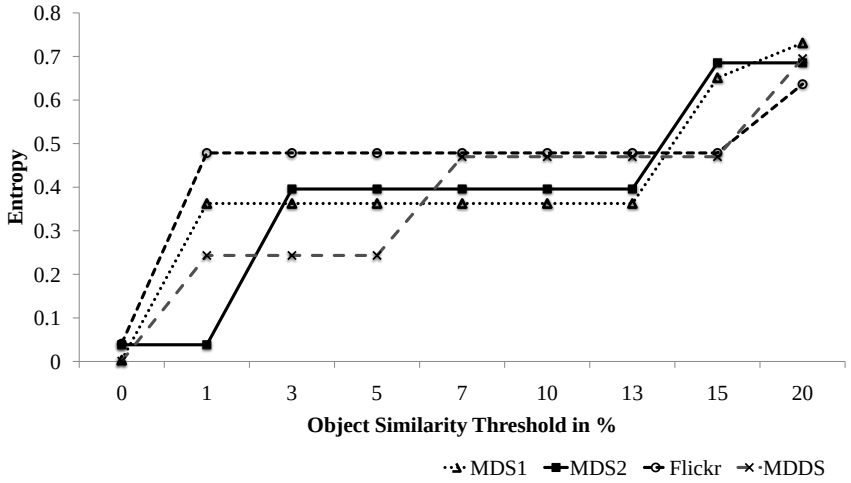


(b) Entropy Values

Figure 3.12: Purity and Entropy Scores of GSOMDC Algorithm for MSC



(a) Purity Values



(b) Entropy Values

Figure 3.13: Purity and Entropy Scores of GSOMDC Algorithm for MIC

for MDS1 is 0.9270, MDS2 is 0.9238, Flickr is 0.9332 and MDDS is 0.9986. The best entropy values are 0.0033 for MDS1, 0.0299 for MDS2, 0.0251 for Flickr and 0.0006 for MDDS. Similarly, with MIC model the best purity values are 0.8914 for MDS1, 0.9133 for MDS2, 0.0.8981 for Flickr and 0.9958 for MDDS. The best entropy values are 0.0032 for MDS1, 0.038 for MDS2, 0.040 for Flickr and 0.0007 for MDDS. The analysis of the purity and entropy values shows that the GSOMDC algorithm clusters the multimedia documents effectively.

The performance comparison of clustering of MMDs with MSC and MIC methods is presented in Table 3.2. The analysis of the results indicate that the

MSC method performs better compared to MIC method by achieving better purity and entropy values. The MSC shows the purity improvement of 3% for MDS1, 1% for MDS2, and 3% for Flickr over the MIC method. Also, it shows the improvement of 1% for MDS2, 2% for Flickr with entropy values compared to MIC method.

3.7 Similarity Measure for Multimedia Documents

The important aspect of multimedia knowledge extraction is finding the similarity between the multimedia documents. The traditional similarity measures are used to find the similarity between the unimodal objects such as text documents, images and audios. The similarity measures such as Euclidean distance, Cosine Similarity, Pearson Correlation Coefficient, Jaccard Coefficient, Averaged Kullback-Leibler Divergence, Manhattan distance, Chebyshev distance and Canberra distance are used to compute the similarity between text documents (Huang, 2008; Schoenharl and Madey, 2008). The media objects are processed and features are extracted to represent them as feature vectors. The similarity between audio features are computed using Manhattan (Seyerlehner et al., 2010), Euclidean (Helén and Virtanen, 2009), Kullback-Leibler divergence (Helén and Virtanen, 2009; Jensen et al., 2009; Kwitt and Uhl, 2008). The widely used similarity measures for image features are Euclidean (Arevalillo-Herráez et al., 2008; Grigorova et al., 2007; Guo et al., 2002),

Table 3.2: Performance Comparison of GSOMDC Algorithm for Clustering UMDs using MSC and MIC

	Purity		Entropy	
Datasets	MSC	MIC	MSC	MIC
MDS1	0.9270	0.8914	0.0033	0.0045
MDS2	0.9238	0.9133	0.0299	0.0383
Flickr	0.9332	0.8981	0.0251	0.0410
MDDS	0.9986	0.9956	0.0006	0.0007

Canberra (Singha et al., 2012), Correlation distance (Akakin and Gurcan, 2012), Hausdorff distance (Park et al., 2007) and Kullback-Leibler divergence (Goldberger et al., 2003; Schnitzer et al., 2012).

As the MMDs are the collection of multiple multimodal multimedia objects, the above discussed similarity measures are not suitable to find the similarity between the MMDs. The similarity among the MMDs depends on the presence and absence of multimedia objects. Although Dice coefficient measures the similarity between the documents based on the presence and absence of the objects, it considers only the common objects between the documents. The similarity between the documents increases when the common information between them increases. The similarity decreases with the increase in dissimilar information of the documents. In (Lin, 1998), Dekang Lin has presented a information theoretic similarity measure for text documents based on commonality and difference between the documents. Lin interpreted the similarity between two objects as the ratio between the amount of information required to state commonality of objects and the information needed to describe two objects completely.

$$Sim(D_A, D_B) = \frac{logP(common(D_A, D_B))}{logP(description(D_A, D_B))} \tag{3.7.1}$$

.

where $P(common(D_A, D_B))$ is the probability of information needed to state the commonality between the documents D_A and D_B and $P(description(D_A, D_B))$ is the probability of information that provides the complete information about the documents D_A and D_B.

We propose an **i**nformation theory based **s**imilarity measure for **m**ultimedia **d**ocuments (ISMD) that measures the document similarity based on the common information shared by the MMDs and the difference of information among the documents. The proposed similarity measure is discussed as follows:

Let umd_i, umd_j are the two UMDs. As per the information theory, the similarity between the two UMDs can be computed as follows:

$$\begin{aligned} Sim(umd_i, umd_j) &= \frac{I(umd_i \cap umd_j)}{I(umd_i, umd_j)} \\ &= \frac{I(umd_i \cap umd_j)}{I(umd_i \cap umd_j) + I(umd_i \Delta umd_j)} \end{aligned} \tag{3.7.2}$$

where $I(umd_i \cap umd_j)$ is the information related to the common objects shared by documents umd_i and umd_j and $I(umd_i \Delta umd_j)$ is the information of objects that distinguishes the documents umd_i and umd_j.

The similarity measure is defined in terms of the multimedia information that are shared between the two UMDs and the multimedia information that differentiates the two UMDs.

Let $O = \{o_k \mid_{k=1}^{m}\}$, is the set of m multimedia objects which describes the documents umd_i and umd_j. The probability of a multimedia object P_k in a UMD can be defined as the fraction of the UMD that containing the multimedia object o_k such that $\sum_{k=1}^{m} P_k = 1$. Let the umd_i can be represented by the probability vector $P_i = \{P_{i_k} \mid_{k=1}^{m}\}$, which describes the probability of object o_k in UMD umd_i.

$$I(umd_i \cap umd_j) = \sum_{k=1}^{m} min(P_{i_k}, P_{j_k})(-log\frac{P_{i_k} + P_{j_k}}{2}) \tag{3.7.3}$$

$$I(umd_i \Delta umd_j) = \sum_{k=1}^{m} diff(P_{i_k}, P_{j_k})(-log\frac{P_{i_k} + P_{j_k}}{2}) \tag{3.7.4}$$

where,

$min(P_{i_k}, P_{j_k}) = \{^{min(P_{i_k},P_{j_k});\ \ if\ \ P_{i_k}>0\ \&\ P_{j_k}>0}_{0;\ otherwise}$

$diff(P_{i_k}, P_{j_k}) = \{^{0;\ \ if\ \ P_{i_k}>0\ \&\ P_{j_k}>0}_{sum(P_{i_k},P_{j_k});\ otherwise}$

Both the documents share the $min(P_{i_k}, P_{j_k})$ of information whereas they differ by the amount of $sum(P_{i_k}, P_{j_k})$ information only when one of the object is absent in one of the document.

3.7.1 Vector Representation for Multimedia Documents

The proposed similarity measure is applied for the binary vector form of the UMDs. Binary vectors are formed for each UMD depending on the existence and the similarity of the objects. Let a document umd_i with M objects is represented by a vector $V_i = \{v_k \mid_{k=1}^{M}\}$, where $v_k \in \{0, 1\}$ indicates whether object o_k exists in umd_i or not. The similarity between the two UMDs, umd_i and umd_j is computed based on the existence of similar objects between the documents. It is assumed that there are no duplicate objects exist in the UMDs. If found, they are considered as one object by computing the similarity between the objects using any similarity measure. If the similarity between the objects satisfies the user defined threshold value, the vector value of both the objects becomes 1.

Let the umd_i has m objects and umd_j has n objects. Initially the binary vectors with $m + n$ length are formed for both the UMDs. The binary vector values assign 1 for the objects they belong and assign 0 for the objects of other UMD. Then, each object of the document umd_i is compared with all the objects document umd_j to find the similar objects. If found, then the corresponding bits of both the vectors are changed to 1. The similarity of objects is calculated using the percentage difference similarity measure which is defined in (3.3.1)

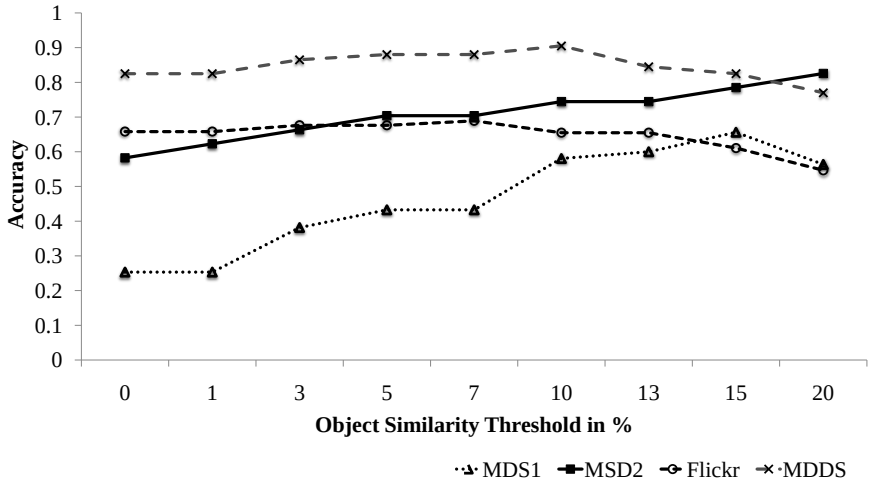
For example, the document umd_i contains five objects and the document umd_j contains three objects. So there are a total of eight objects found in the documents umd_i and umd_j. Hence, each document is represented by a vector of 8 bits, each bits representing the existence or absence of object in that document. Initially, umd_i is represented by the vector V_i= [11111000] and umd_j is represented by V_j=[00000111]. Let assume 1^{st} object of umd_i is similar to 2^{nd} object of umd_j i.e. 7^{th} object of all 8 objects. Then, the 7^{th} bit of umd_i and 1^{st} bit of umd_j becomes 1. So the vectors are changed to V_i= [11111010] and V_j= [10000111]. In this way, all the objects are checked for the similar objects in umd_j. After forming the binary vectors, the proposed similarity measure is applied between the vectors to find the similarity between the UMDs.

3.7.2 Experimental Results and Discussion

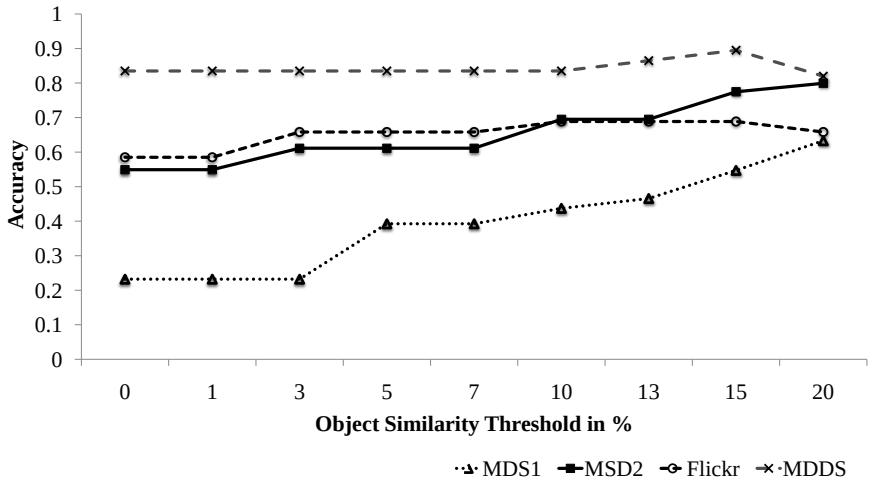
To evaluate the effectiveness of the proposed similarity measure, the experiments are conducted for the classification of MMDs with four datasets MDS1, MDS2, Flickr and MMDS. The datasets are explained in Section 2.9. The datasets are divided into training and test dataset in the ratio of 80% - 20%. The MMDs are represented using the MSC and MIC methods and ISMD measure is used for document similarity. The similarity between the media objects of UMDs is computed using the PD measure ((3.3.1)). The results of multimedia document classification is shown in Fig. 3.14. Using MSC method, the ISMD measure achieved the maximum accuracy of 0.66, 0.83, 0.69 and 0.91 for datasets MDS1, MDS2, Flickr and MDDS respectively. The maximum accuracy obtained with MIC method are 0.63, 0.80, 0.69 and 0.895 for datasets MDS1, MDS2, Flickr and MDDS respectively.

The proposed ISMD similarity measure is compared with information theory based similarity measure SMTP (Lin et al., 2014) for the classification of MMDs. The SMTP measure basically employed for measuring the similarity among the unimodal text documents. The SMTP measure was used for classification of text documents and proved better compared to other similarity measures Euclidean, Cosine, Extended Jaccard, and IT-Sim (Aslam and Frost, 2003). The proposed multimedia data representation methods proposed converts the multimodal multimedia documents into unimodal unified multimedia documents. As the use of non-textual features for SMTP is mentioned in the paper (Lin et al., 2014), we used the SMTP for unified multimedia documents that contain non-textual features. Hence we implemented the SMTP and the proposed ISMD for measuring the similarity between the UMDs. The value of the constant λ used in SMTP measure is assumed as 1 to obtain the best classification (Lin et al., 2014).

Table 3.3 shows the performance comparison of ISMD and SMTP for the classification of MMDs using MSC and MIC methods. It is observed that the proposed ISMD achieved significantly better performance compared to SMTP in



(a) Accuracy of MMDs with MSC using ISMD



(b) Accuracy of MMDs with MIC using ISMD

Figure 3.14: Performance of Classification of MMDs using ISMD measure

classifying the MMDs using the MSC and MIC methods. The ISMD shows the improvement of 14% (MDS1), 6% (MDS2), 1% (Flickr), and 6% (MDDS) with MSC method over the SMTP method. With MIC method, the ISMD attained the improvement of 13% (MDS1), 1% (MDS2), 3% (Flickr), and 6% (MDDS). The SMTP considers the count of objects in the documents and uses Gaussian function for the similarity measure. As the similarity between the documents mainly relies on the presence and absence of objects, the proposed ISMD did not consider the number of objects present in the documents. The predominance of

the classification results indicate that the proposed ISMD measure outperforms SMTP measure for the classification of multimedia documents with the MSC and MIC methods.

We compared the classification results of proposed ISMD with Dice Coefficient for four datasets with MSC and MIC methods. The results are shown in Table 3.4. The Dice Coefficient achieved improvement of 1% in accuracy over the ISMD measure for MDS2 and Flickr for the classification of MMDs using MSC method. The accuracy of Dice Coefficient is similar to accuracy of ISMD measure for MDS1 and MDDS. It is noticed that using MIC method, the classification performance of Dice Coefficient is similar to ISMD measure. Compared to Dice coefficient, the ISMD measure is computationally expensive, as it requires the computation of probability for each object in each MMDs. Therefore, we preferred the use of Dice's coefficient over ISMD measure for further experiments to find the pairwise document similarity.

Table 3.3: Comparison of MMD Classification Accuracy for ISMD and SMTP measure

	MSC		MIC	
Datasets	SMTP	ISMD	SMTP	ISMD
MDS1	0.523	0.656	0.499	0.633
MDS2	0.765	0.826	0.786	0.799
Flickr	0.679	0.689	0.658	0.689
MDDS	0.845	0.905	0.835	0.895

Table 3.4: Comparison of MMD Classification Accuracy for ISMD measure and Dice's coefficient

	MSC		MIC	
Datasets	ISMD	Dice's Coefficient	ISMD	Dice's Coefficient
MDS1	0.656	0.659	0.633	0.626
MDS2	0.826	0.816	0.799	0.801
Flickr	0.689	0.683	0.689	0.683
MDDS	0.905	0.910	0.895	0.890

3.8 Summary

In this chapter, we discussed the two multimedia data representation methods MSC and MIC which are used to represent the multimodal MMD as UMD. The MSC converts multimedia objects as signal objects whereas MIC converts multimodal signal objects as image objects. The signal objects are subjected to wavelet transform and a low dimensional feature vector is generated for the signal by computing the statistics of wavelet coefficients. In order to preserve the color, shape and texture of the image objects a high dimensional feature vector is formed based on the color, shape and texture features of the image objects. The experimental evaluation of proposed methods demonstrate that the proposed multimedia data representation methods are effectively used for classification and retrieval of MMDs. The proposed GSOMDC algorithm successfully clusters the MMDs using the MSC and MIC model by achieving better purity and entropy values. The proposed similarity measure ISMD outperforms the SMTP in classification of MMDs using both the MSC and MIC model.

The experimental validation indicate that the performance of ISMD is similar to Dice's coefficient for classification of MMDs using the MSC and MIC methods. Compared to Dice's coefficient, the computation of ISMD is computationally expensive as it requires the computation of probability for each object of each MMD. Therefore for the further experiments, we are using the Dice coefficient to find the similarity between MMDs. The performance comparison of MSC and MIC methods demonstrate that the MSC based knowledge extraction methods perform better compared to MIC based methods. The time taken by the MIC method for the conversion of multimedia objects to image objects and feature extraction is higher compared to the MSC method. Moreover, managing the image objects is computationally expensive due to its high dimensional features as discussed in Section 3.5.1. Hence, to achieve the remaining objectives of the research work, we used the MSC method for the representation of MMDs.

‘

Chapter 4

Multimedia Suffix Tree Document Representation

Multimedia document representation is the important aspect of KDMD process that favors the multimedia mining process by representing the MMDs in a suitable representation. In this chapter, we discuss an effective representation to represent the multimedia documents in a unified tree based representation. We discuss the representation based multimedia mining methods in order to extract the useful knowledge from multimedia documents. Our contributions are:

- A complete tree based representation, Multimedia Suffix Tree Document (MSTD) for the MMDs.
- Effective MSTD based multimedia mining methods for the efficient knowledge extraction from MMDs.

The rest of the chapter is organized as follows. Section 4.1 describes the MSTD representation. In Section 4.2, we discuss the knowledge extraction from MMDs using MSTD representation. In Section 4.3, we discuss the computational complexity of MSTD representation and MSTD based knowledge extraction methods. The experimental results are discussed in Section 4.4.

4.1 Multimedia Suffix Tree Document Representation

Suffix tree document model has the ability to represent the documents in a tree structure without losing any detail. Moreover, it generates the base clusters based on the common information between the documents. These attractive factors influenced us to generate a suffix tree based representation for MMDs using the shared multimedia objects. To the best of our knowledge there are no efforts made to represent the multimodal MMDs based on the shared multimodal information in one platform.

The multimodal nature is the main obstacle for STD model in representing the MMDs. The MMDs are converted as UMDs by the MSC method discussed in Section 3.2.1. An enhanced version of STD model known as multimedia suffix tree document (MSTD) representation is proposed to represent the UMDs based on the shared similar signal objects between the UMDs. The similar signal objects are found by computing the similarity between them using the PD measure according to the equation (3.3.1). Section 4.1.1 describes the construction of MSTD representation for the given UMDs and Section 4.1.2 describes the characteristics of the MSTD representation.

4.1.1 Construction of MSTD Representation

As discussed in Section 2.2.2, the standard STD model considers a document as a string consisting of words. The MSTD representation considers a unified multimedia document *umd* as a string and the signal objects $s_1, s_2, ..., s_m$ as words. For example, let $umd = \{s_1, s_2, s_3\}$. The collection of identifiers of the signal objects forms the string $s_1s_2s_3$ for *umd*. The suffix phrases of the string are $s_1s_2s_3$, s_2s_3 and s_3.

The MSTD representation is built for the set of UMDs using the similarity of the signal objects. It is comprised of two kinds of nodes: root node and phrase

node. The phrase node represents the suffix phrase shared by the documents. The phrase node consists of four attributes *Phrase*, *DocId*, *child* and *sibling*. The *Phrase* attribute stores the suffix phrase shared by the documents. The document identifiers that contain the common suffix phrase are stored in *DocId* attribute. For example, if the UMDs umd_1 and umd_2 contain the signal objects $s_1s_2s_3$, then the suffix phrase $s_1s_2s_3$ is stored in *Phrase* and the identifiers of umd_1 and umd_2 are stored in *DocId*. The *child* and *sibling* attributes point the child and sibling nodes respectively. The level of the phrase nodes are designated with respect to the root node. The phrase nodes that are directly connected to root are the first level nodes. The suffix phrases of the document are inserted in the multimedia suffix tree (MST) by comparing the first signal object of suffix phrase with the first signal object of the first level phrase nodes of the suffix tree.

Let assume that, $sfxphr$ contain the suffix phrase of the document need to be placed in the MST. Let *pnode* be the first level phrase node of MST that has first signal object similar to first signal object of $sfxphr$. The signal objects of *pnode.Phrase* and $sfxphr$ are compared and the similar signal objects between them are extracted which forms the new phrase string for *pnode*. The remaining signal objects of *pnode.Phrase* forms the child node of *pnode*. Now, the phrase of remaining signal objects of $sfxphr$ are assigned to $sfxphr$ and compared with the *pnode.child.Phrase*. If none of the signal objects found similar, they form the child node for *pnode*. If similar signal objects found, they form the new phrase for *pnode.child* and the remaining signal objects of $sfxphr$ becomes new phrase for $sfxphr$. This process continues till the all the signal objects of $sfxphr$ are covered by the child phrase nodes of *pnode* or forms a new node for signal objects of $sfxphr$.

In text mining the synonyms and antonyms play the major role in analyzing the words. However in MSTD representation, grouping the words based on synonyms and antonyms is not incorporated. Hence, the representation does not use any data structure like inverted file to store the synonyms and antonyms. With the text documents, the sequence of words plays a major role in knowledge extraction.

So, the standard STD model maintains the sequence of words of the suffix phrase in suffix tree. As the MMD is a collection of multimedia objects, considering the sequence of objects is not practical. However, in order to avoid the ambiguity, the MSTD representation maintains the position of only the first signal object of the suffix phrase in the MST tree. After obtaining the first level phrase node, the sequence of other signal objects may change while comparing them with the child nodes of selected first level phrase node.

For example, let $sfxphr$=$s_2s_1s_3s_5$ is a phrase of the document need to be inserted in the MST. Let the first level phrase nodes n_1 contain the phrase $s_1s_2s_3s_4$, n_2 contain the phrase $s_2s_3s_4$, n_3 contain the phrase s_3s_4 and n_4 contain the phrase s_4. The first signal object of $sfxphr$ and first level phrase nodes are compared. The n_2 is selected for further comparison as it has s_2 as its first signal object. The phrase node n_2 and $sfxphr$ have s_2s_3 in common, so the phrase s_2s_3 is stored in n_2. The remaining phrase (s_4) of n_2 becomes the child node of n_2. The remaining phrase (s_1s_5) of $sfxphr$ is now stored in $sfxphr$ and compared with the child nodes of n_2. As none of the child nodes of n_2 have signal objects of $sfxphr$, it becomes the child node of n_2. It is noted that, after obtaining the first level node n_2, the sequence of signal objects is changed in the phrase, i.e. phrase $s_2s_1s_3s_5$ is changed to $s_2s_3s_1s_5$. However, the position of first signal object of first level phrase nodes and the suffix phrase $sfxphr$ remains unaltered.

Let consider the UMD as a set of UMDs with signal objects.

$UMD = \{umd_1, umd_2, umd_3, umd_4, umd_5, umd_6\}$;

$umd_1 = \{s_{1_1}, s_{1_2}, s_{1_3}\}$

$umd_2 = \{s_{2_1}, s_{2_2}, s_{2_3}, s_{2_4}, s_{2_5}\}$

$umd_3 = \{s_{3_1}, s_{3_2}, s_{3_3}, s_{3_4}\}$

$umd_4 = \{s_{4_1}, s_{4_2}, s_{4_3}, s_{4_4}, s_{4_5}\}$

$umd_5 = \{s_{5_1}, s_{5_2}, s_{5_3}, s_{5_4}\}$

$umd_6 = \{s_{6_1}, s_{6_2}, s_{6_3}\}$

where $s_{1_1}...s_{1_3}$, $s_{2_1}...s_{2_5}$, $s_{3_1}...s_{3_4}$, $s_{4_1}...s_{4_5}$, $s_{5_1}...s_{5_4}$ and $s_{6_1}...s_{6_3}$ are the signal objects. Let the document identifiers of UMDs, $umd_1, ..., umd_6$ be $1, .., 6$.

The UMD umd_1 is scanned to extract the signal objects s_{1_1}, s_{1_2} and s_{1_3}. The multimedia suffix tree is constructed using the suffix phrases $s_{1_1}s_{1_2}s_{1_3}$, $s_{1_2}s_{1_3}$ and s_{1_3}. Let MST be the empty suffix tree with root node. As there are no similar signal objects found in umd_1, all the phrases form three phrase nodes n_{1_1}, n_{1_2} and n_{1_3} as shown in Fig. 4.1a

The extracted signal objects from umd_2 are s_{2_1}, s_{2_2}, s_{2_3}, s_{2_4} and s_{2_5}. It is assumed that the signal objects s_{2_2} and s_{2_5} are similar to each other as per the object similarity and both are termed as s_{2_2}. Similarly the signal objects s_{2_1} is similar to s_{1_1}, s_{2_2} is similar to s_{1_2} and s_{2_3} is similar to s_{1_3}. After removing the duplicates, the signal objects of umd_2 are designated as s_{1_1}, s_{1_2}, s_{1_3}, s_{2_4}. Let $sfxphr$ be the first suffix phrase $s_{1_1}s_{1_2}s_{1_3}s_{2_4}$. The first signal object s_{1_1} of $sfxphr$ is compared with the first signal object of the first level phrase nodes of MST. The node n_{1_1} has s_{1_1} as first element and selected for further comparison. The signal objects of $n_{1_1}.Phrase$ are compared with the signal objects of $sfxphr$ to find the common elements between them. The signal objects s_{1_2} and s_{1_3} are found common to both $n_{1_1}.Phrase$ and $sfxphr$. Thus, the common phrase $s_{1_1}s_{1_2}s_{1_3}$ is replaced as $n_{1_1}.Phrase$ and $sfxphr$ is replaced by the remaining string s_{2_4}. Now, the $sfxphr$ is compared with the child nodes of n_{1_1}. The child nodes of n_{1_1} does not contain any signal object of $sfxphr$, so the $sfxphr$ becomes the child node n_{2_1} for n_{1_1}. Similarly the remaining suffixes of umd_2 are placed in appropriate place depending on object similarity. The MSTD representation for umd_1 and umd_2 is shown in Fig. 4.1b.

The document umd_3 is scanned and the signal objects s_{3_1}, s_{3_2}, s_{3_3} and s_{3_4} are extracted. It is assumed that the signal objects s_{3_1} is similar to s_{1_2}, s_{3_2} is similar to s_{1_1} and s_{3_3} is similar to s_{1_3}. Hence, the signal objects of umd_3 are designated as s_{1_2}, s_{1_1}, s_{1_3}, s_{3_4}. Let the first suffix phrase $s_{1_2}s_{1_1}s_{1_3}s_{3_4}$ is stored in $sfxphr$. The first signal object s_{1_2} of $sfxphr$ is compared with the first level phrase nodes of MST. The node n_{1_2} is selected for further comparison as it has s_{1_2} as first element. The phrase node n_{1_2} and $sfxphr$ have $s_{1_2}s_{1_3}$ in common, so the phrase $s_{1_2}s_{1_3}$ is stored in n_{1_2}. The remaining phrase $s_{1_1}s_{3_4}$ is stored in $sfxphr$

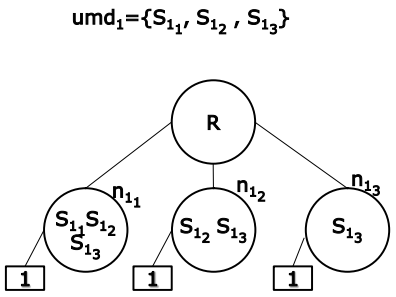
and compared with the child nodes of n_{1_2}. As none of the child nodes of n_{1_2} have signal objects of $sfxphr$, it becomes the child node of n_{1_2}.

For the second phrase $sfxphr = s_{1_1}s_{1_3}s_{3_4}$, the phrase node n_{1_1} is selected since it has s_{1_1} as first element. Both the $sfxphr$ and n_{1_1} has $s_{1_1}s_{1_3}$ in common. Hence the phrase $s_{1_1}s_{1_3}$ replaces the $n_{1_1}.Phrase$ and the remaining elements of $n_{1_1}.Phrase$ becomes the child node of n_{1_1}. The $sfxphr$ is changed to the remaining elements of $sfxphr$ i.e. s_{3_4} and compared with the child nodes of n_{1_1}. As there are no child nodes have the elements of $sfxphr$ it forms the child node for n_{1_1}. In the similar way all the suffix phrases of umd_3 are placed in the MST. The MSTD representation after inserting all the suffix phrases of umd_3 is shown in the Fig. 4.1c. The procedure is continued for all the UMDs of the dataset and the constructed MSTD representation for the UMDs is shown in Fig. 4.2.

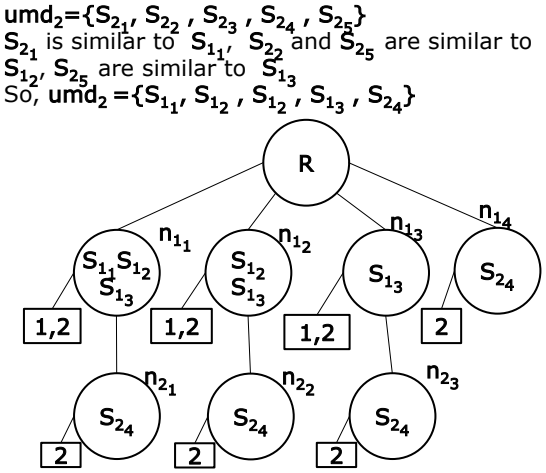
4.1.2 Characteristics of MSTD Representation

The MSTD representation provides the complete information of all the MMDs. The multimedia objects of every multimedia document is represented by the phrase nodes of the tree. Hence, the MSTD is a complete representation of MMDs that provides the complete information needed for knowledge extraction process. As the MSTD representation is constructed based on the similarity of the signal objects, similar signal objects are grouped and enclosed by the same phrase node. Hence the representation is compact by avoiding the separate representation for similar signal objects. Also, it reduces the search time for the knowledge extraction methods.

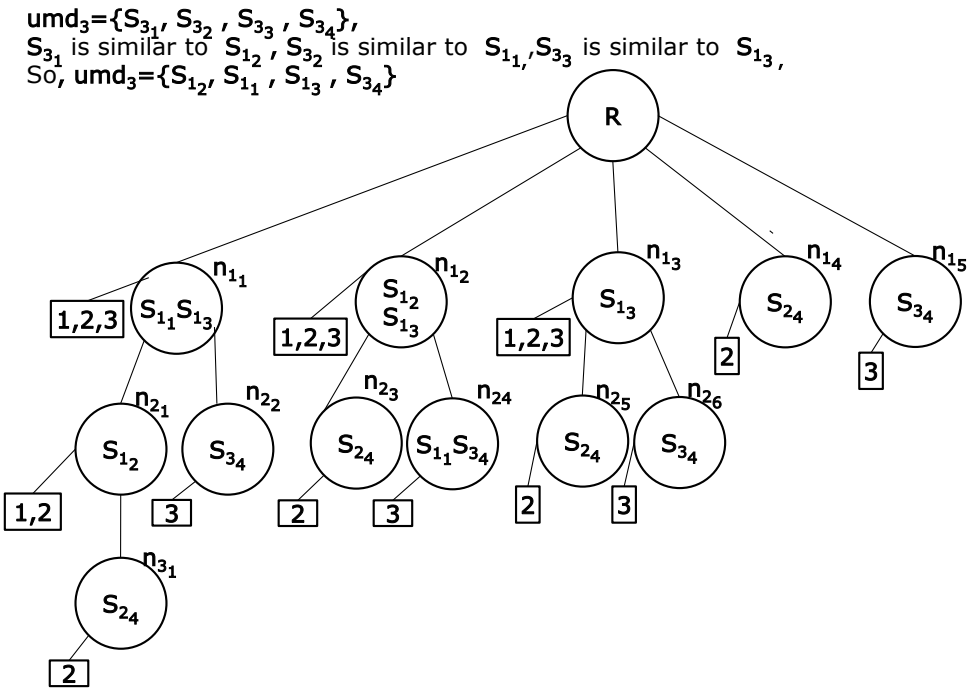
The number of first level phrase nodes equals to the number of unique signal objects present in the UMDs. Each phrase node contains the maximum common objects shared by the documents. They cluster the documents based on the common objects. Hence, the MSTD representation generates the base clusters for the MMDs based on the common information they have. The phrase nodes can be further merged to get the optimum clusters using their similarity.



(a) MSTD representation for umd_1



(b) MSTD representation for umd_1, umd_2



(c) MSTD representation for umd_1, umd_2,umd_3

Figure 4.1: Construction of MSTD representation for UMDs

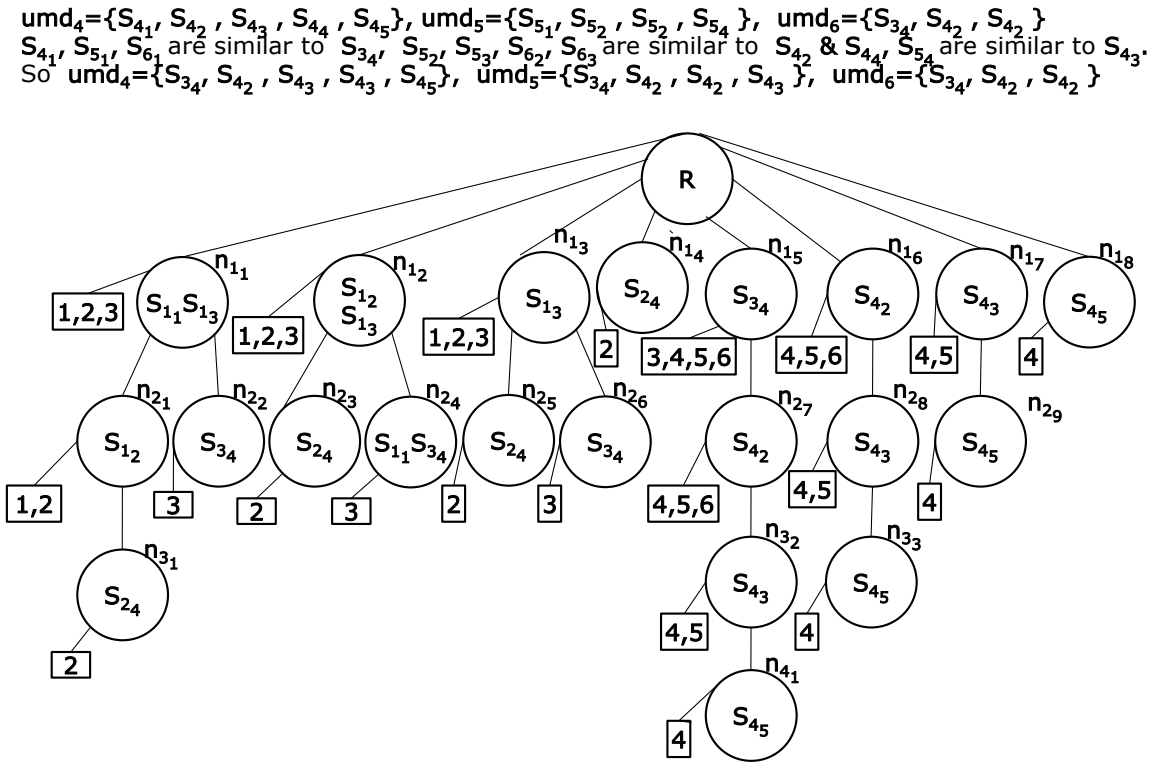


Figure 4.2: MSTD Representation For dataset $UMD = \{umd_1, umd_2, umd_3, umd_4, umd_5, umd_6\}$

4.2 Knowledge Extraction from Multimedia Documents using MSTD Representation

Figure 4.3 shows the framework for knowledge extraction from UMDs using the proposed MSTD representation. MSTD representation is used as a platform for knowledge extraction methods classification, clustering, frequent pattern mining and association rule generation. MSTD based representation assigns the class label for the test MMD. MMDs are grouped into topics using the MSTD based clustering without having any prior knowledge about the topics. MSTD representation extracts the frequent patterns exist in the MMDs using which the association rules are discovered to describe the relationships among the contents of the MMDs. In Section 4.2.1, we discuss the classification of MMDs using MSTD representation. In Section 4.2.2, we discuss the clustering of MMDs using

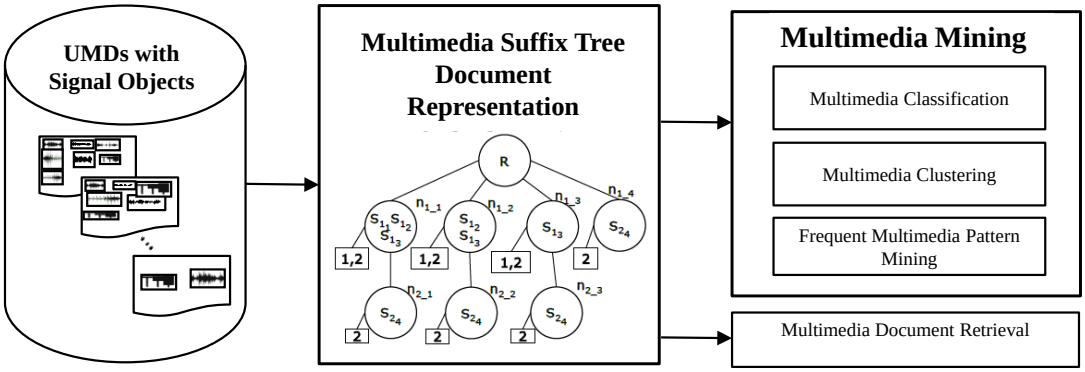


Figure 4.3: MSTD Representation for Knowledge Extraction from MMDs

MSTD representation. The MSTD based frequent pattern mining and association rule generation is discussed in Section 4.2.3.

4.2.1 MSTD based Classification of Multimedia Documents

In this section we will demonstrate how the MMDs are classified using the MSTD representation. Let $UMD_T = \{umd_1, umd_2, .., umd_N\}$ be the N training UMDs and umd_q be the test document. The MSTD representation, $MSTD_T$ is created for UMD_T as described in section 4.1.1.

Let $\{s_{q_1}, s_{q_2}, .., s_{q_Q}\}$ be the test signal objects extracted from umd_q. The test signal objects are compared with the signal objects stored as first element in first level phrase nodes of suffix tree $MSTD_T$. Let *pnode* be the first level phrase node of the $MSTD_T$ which contains the test signal object s_{q_1}. Then, depth first search (DFS) algorithm is applied starting from *pnode* to store the child phrase nodes that have signal objects similar to test signal objects in a list *nodelist*. The *nodelist* is processed to extract the phrase nodes that have maximum similar signal objects of test document. For each of these extracted phrase node, let the documents that has the phrase node are stored in *ResDocId* and the number of maximum signal objects is stored in *ComIdLen*. Similarly, for all the test signal objects, the phrase nodes that have maximum signal objects are extracted and the

details are stored in $ComIdLen$ and $ResDocId$. Each document from $ResDocId$ is assigned a weight based on the joint probability of having the common objects in both training document and test document. The weight for each document from $ResDocId$ is computed as follows:

$$w_i = \frac{|ComIdLen_i|}{|ResDocId_i|} \times \frac{|ComIdLen_i|}{|umd_q|} \tag{4.2.1}$$

where w_i is the weight of document $ResDocId_i \in ResDocId$ and $ComIdLen_i$ is the length of similar objects found in $ResDocId_i$.

The document that has higher weight is selected as the most relevant document. If the concept of the retrieved relevant document is similar to that of test document, then the test document is considered as correctly classified. The MSTD based classification algorithm is summarized in Algorithm 4.1. The MSTD based classification is illustrated in the following example:

Let the test document be $umd_q = \{s_{3_4}, s_{1_1}, s_{4_2}, s_{4_3}\}$. For the first signal object s_{3_4}, the selected first level phrase node is n_{1_5}. The child node of n_{1_5} that has maximum similar signal objects of test document is n_{3_2} i.e. $s_{3_4}s_{4_2}s_{4_3}$. So, the documents umd_4 and umd_5 which stored in node n_{3_2} are retrieved. For s_{1_1}, the selected first level phrase node is n_{1_1}. Since none of its child nodes have signal objects similar to test signal objects except s_{1_1}, the documents umd_1, umd_2 and umd_3 of node n_{1_1} are retrieved. Similarly for s_{4_2} and s_{4_3}, the documents umd_4 and umd_5 are retrieved. For all the retrieved documents, the weight is computed using the formula (4.2.1). The documents with weights are $umd_1(0.08)$, $umd_2(0.05)$, $umd_3(0.06)$, $umd_4(0.45)$ and $umd_5(0.56)$. The document umd_5 with higher weight is selected as most relevant document for umd_q.

4.2.2 MSTD based Clustering of Multimedia Documents

This section explains how the clustering is performed to group the MMDs into the clusters of multimedia topics using the MSTD representation. Let $UMD =$

Algorithm 4.1 MSTD based Classification Algorithm

1: **Input**: Training Dataset $\{UMD_T = umd_1, umd_2, ..., umd_N\}$, Test document umd_q
2: **Output**: $Class_q$=Class of umd_q
3: Construct the MSTD representation, $MSTD_T$ for the training dataset UMD_T
4: **for** each signal object $s_{q_i} \in umd_q$ **do**
5: $pnode \leftarrow$ first level phrase node with first element of $pnode.Phrase = s_{q_i}$
6: $nodelist \leftarrow$ list of child phrase nodes of $pnode$ obtained by applying DFS algorithm that has signal objects similar to umd_q
7: **if** nodelist is Empty **then**
8: $selnode \leftarrow pnode$
9: **else**
10: $selnode \leftarrow$ phrase node from $nodelist$ that has maximum similar signal objects compared to umd_q.
11: **end if**
12: $ResDocId_i \leftarrow selnode.DocId$
13: $ComIdLen_i \leftarrow$ number of similar test signal objects present in $selnode.Phrase$
14: **end for**
15: compute the weight w_i for each document $ResDocId_i \in ResDocId$ using the formula (4.2.1)
16: Let $RelDocId \leftarrow$ most relevant document that has higher weight w_i
17: $Class_q \leftarrow$ Class of $RelDocId$

$\{umd_1, umd_2, .., umd_N\}$ be the set of UMDs. Clustering the MMDs require to compute the pairwise similarity between the UMDs. In order to compute the similarity between the UMDs, the list of first level phrase nodes are required as they carry the information of each UMD. Algorithm 4.2 explains the steps to get the list of first level phrase nodes for each UMD.

After collecting the list of first level phrase nodes for all the UMDs, for each document umd_i the pairwise similarity with all other UMDs is computed based on the similar first level nodes between them. The $UMDs$ with maximum similarity are the candidates to form cluster with the document umd_i. The $UMDs$ that are having maximum similarity with umd_i are selected using the following equation,

$$UMD_R = \{umd_k : k \in \underset{1 \leq j \leq N}{\operatorname{argmax}} [sim(umd_i, umd_j)]\} \tag{4.2.2}$$

where $sim(umd_i, umd_j)$ is the similarity between two documents umd_i and umd_j. The similarity is calculated as

$$sim(umd_i, umd_j) = 2 * \frac{|PhrNdList_i \bigcap PhrNdList_j|}{|PhrNdList_i + PhrNdList_j|} \tag{4.2.3}$$

where $PhrNdList_i$ and $PhrNdList_i$ are the first level phrase node lists of umd_i and umd_j respectively. The minimum threshold for $sim(umd_i, umdj)$ is restricted to 0.5 to ensure that more similar UMDs are clustered together. The $UMDs$ with maximum similarity are clustered with umd_i by combining their phrase node list with the phrase node list of umd_i. The procedure continues for all the documents till they grouped into clusters. The clusters are then merged based on the similarity between the clusters. The algorithm repeats till there are no more clusters to merge resulting in optimum clusters. The MSTD based multimedia clustering algorithm is summarized in Algorithm 4.3 and illustrated in following example.

According to Fig. 4.2, $\{n_{1_1}, ..., n_{1_8}\}$ are the first level phrase nodes of MSTD representation. As per Algorithm 4.2 the extracted first levels phrase nodes for each $umd \in UMD$ are as follows:

$umd_1 \longrightarrow \{n_{1_1}, n_{1_2}, n_{1_3}\}$

$umd_2 \longrightarrow \{n_{1_1}, n_{1_2}, n_{1_3}, n_{1_4}\}$

$umd_3 \longrightarrow \{n_{1_1}, n_{1_2}, n_{1_3}, n_{1_5}\}$

$umd_4 \longrightarrow \{n_{1_5}, n_{1_6}, n_{1_7}, n_{1_8}\}$

$umd_5 \longrightarrow \{n_{1_5}, n_{1_6}, n_{1_7}\}$

$umd_6 \longrightarrow \{n_{1_5}, n_{1_6}\}$

The similarity between the phrase node lists for each UMD with other UMDs is calculated and most similar UMDs are selected according to (4.2.2). As per the similarity the umd_1 is merged with umd_2 and umd_3 as it has higher similarity with them. Similarly, umd_4 is merged with umd_5 and umd_6 is merged with umd_4. Finally, obtained two clusters are $c_1(umd_1, umd_2, umd_3)$ and $c_2(umd_4, umd_5, umd_6)$.

Algorithm 4.2 First level phrase nodes collection for a Multimedia Document

Input: $B = \{b_1, b_2, .., b_O\}$ are branches of MSTD representation for dataset UMD, umd_{id}: Document ID
Output: $PhrNdList$: first level phrase node list for umd_{id}
$PhrNdList_{id} \leftarrow \{\}$
for each branch $b_k \in B$ **do**
    Let $pnode_k$ be the first level node of b_k
    if $umd_{id} \in pnode_k.DocId$ **then**
        $PhrNdList_{id} = PhrNdList_{id} \bigcup pnode_k$
    end if
end for
return $PhrNdList$

Algorithm 4.3 MSTD based Clustering Algorithm

Input: $LIST = \{PhrNdList_1, PhrNdList_2, .., PhrNdList_N\}$: set of first level phrase node lists for $UMD = \{umd_1, umd_2, ..umd_N\}$
Output: $C = \{c_1, c_2, ..c_K\}$ Clusters for UMD
repeat
    for each document $umd_i \in UMD$ **do**
        $c_i \leftarrow umd_i$
        Let $PhrNdList_i$ be the leaf node list of umd_i
        Let $PhrNdList_1$ .. $PhrNdList_j$ be the leaf node list of $umd_1..umd_j$ and $i \neq j$
        $UMD_R \leftarrow UMDs$ with maximum similarity computed as per equation (4.2.2)
        $c_i \leftarrow c_i \bigcup UMD_R$
        $PhrNdList_i \leftarrow PhrNdList_j \bigcup PhrNdList_i$
        $PhrNdList_j \leftarrow \{\}$
    end for
until no clusters to merge
return $C = \{c_1, c_2, ..c_K\}$

4.2.3 MSTD based Frequent Pattern Mining and Association Rule Generation for Multimedia Documents

In this section we discuss how the MSTD representation supports the frequent pattern mining and association rule generation from UMDs. The multimedia document is considered as a transaction as it has the collection of patterns of multimedia objects. With multimedia document, the frequent pattern is termed

as frequent multimedia pattern (FMP) as the patterns are formed using multimedia objects. The main task of frequent multimedia pattern mining is to generate the frequent patterns of multimedia objects that exist in user defined fraction of multimedia documents. The MSTD representation generates the frequent multimedia patterns in the construction stage itself. The phrase nodes of the MSTD representation represent the frequent pattern of multimedia objects shared by the MMDs in the dataset. Thus, the FMPs are obtained from the phrase nodes of the MSTD representation. The FMPs are extended to generate the multimedia class association rule (MCAR) for the classification of MMDs.

Let $UMD = \{umd_1, ..., umd_N\}$ be the set of unified multimedia documents, $S = \{s_1, ..., s_m\}$ be the set of all signal objects present in UMD and $C = \{c_1, ..., c_k\}$ be the set of class labels.

Definition 4.2.1. *The MCAR is defined in the form $R_m : objset \longrightarrow c$ where $objset$ is the set of signal objects such that $objset \subseteq S$ and $c \in C$.*

Definition 4.2.2. *The support of the multimedia class association rule, $sup(R_m)$ is defined as the number of UMDs that contain $objset$ and are labeled with c.*

Definition 4.2.3. *The occurrence $occ(R_m)$ of rule R_m is the number of UMDs that contain the signal objects present in R_m's antecedent.*

Definition 4.2.4. *The confidence of rule R_m, $conf(R_m)$ is defined as,*

$$conf(R_m) = \frac{sup(R_m)}{occ(R_m)}$$

Each FMP will generate $2^n - 1$ rule combinations. All the FMPs may not help in building the effective classifier. Moreover, they may produce huge number of MCARs which reduces the performance of the classifier. Therefore, the FMPs are filtered based on the user defined minimum support. The change

in user defined minimum support changes the number of MCARs used for classification thereby affecting the classification time for the documents. To avoid the duplicate multimedia association rules, the closed FMPs are mined. The rule combinations of the FMPs form the antecedent part of the MCAR whereas the class label of MMDs that own the phrase becomes the consequent part of the MCAR. A MCAR is pruned if its antecedent is a subset of another MCAR's antecedent and both share the same class label i.e. a MCAR R_{m1} is pruned if its antecedent is a subset of another MCAR R_{m2} antecedent and the consequent of R_m1 is same as R_m2. If a MCAR associates with more than one class label, then the class label that contains more documents has been selected. The following example explains the generation of FMPs and MCARs from the MSTD representation shown in Fig. 4.2

The phrase node n_{1_1} generates the FMP $\{s_{1_1}s_{1_3}\}$ which is shared by the documents umd_1,umd_2 and umd_3 with support value of 3/6. It is represented in the following form,

$\{s_{1_1}s_{1_3}\}[sup = 3/6](1, 2, 3)$

where 1,2 and 3 are the identifiers of documents umd_1,umd_2 and umd_3.

Let the FMPs be the part of at least two documents i.e. the minimum support of a FMP is 2. The mined closed FMPs are listed as follows:

$\{s_{1_1}s_{1_3}\}[sup = 3/6](1, 2, 3)$

$\{s_{1_1}s_{1_3}s_{1_2}\}[sup = 2/6](1, 2)$

$\{s_{1_2}s_{1_3}\}[sup = 3/6](1, 2, 3)$

$\{s_{3_4}\}[sup = 4/6](3, 4, 5, 6)$

$\{s_{3_4}s_{4_2}s_{4_3}\}[sup = 2/6](4, 5)$

$\{s_{3_4}s_{4_2}\}[sup = 3/6](4, 5, 6)$

Let c_1 is the class label of UMDs $\{1, 2, 3\}$ and c_2 is the class label of UMDs $\{4, 5, 6\}$, then the generated MCAR are,

$s_{1_1}s_{1_3} \longrightarrow c_1$

$s_{1_1}s_{1_3}s_{1_2} \longrightarrow c_1$

$s_{1_3}s_{1_2} \longrightarrow c_1$

$s_{3_4} \longrightarrow c_1, c_2$

$s_{3_4}s_{4_2}s_{4_3} \longrightarrow c_2$

$s_{3_4}s_{4_2} \longrightarrow c_2$

Since the MCAR $s_{1_1}s_{1_3}s_{1_2} \longrightarrow c_1$ also includes $s_{1_1}s_{1_3} \longrightarrow c_1$ hence the latter MCAR is pruned. For the MCAR $s_{3_4} \longrightarrow c_1, c_2$, the class label c_2 has selected as it contains more documents(4,5,6) compared to class label c_1 (3). Finally, the MCARs used for classification are,

$s_{1_1}s_{1_3}s_{1_2} \longrightarrow c_1$ and $s_{3_4}s_{4_2}s_{4_3} \longrightarrow c_2$

MSTD-MCAR based Classification of Multimedia Documents

The usefulness of MCARs is extended to classify the unlabeled test UMDs. The training UMDs are represented using the MSTD representation. The training MCARs are generated as explained in section 4.2.3. The signal objects from the test UMD form the test multimedia patterns and generate the test MCARs. The training MCARs and test MCARs are sorted according to the length of antecedent part. The antecedent part of the test MCAR is compared with that of training MCAR. The training MCAR that covers more test signal objects is selected and its label is assigned to the test UMD.

For Example,

Let the test document be $umd_q = \{s_{3_4}, s_{1_1}, s_{4_2}, s_{4_3}\}$. The test MCAR is $s_{3_4}s_{1_1}s_{4_2}s_{4_3} \longrightarrow c_x$ where c_x is unknown label. The MCARs generated in example of section 4.2.3 are compared with test MCAR. The training MCAR $s_{3_4}s_{4_2}s_{4_3}$ has more items of test MCAR, so the corresponding label c_2 is assigned to umd_q.

MSTD-FMP based Clustering of Multimedia Documents

The UMDs can be clustered using the FMPs mined from the MSTD representation. Since the FMPs are extracted from the phrase nodes of MSTD representation, each FMP is connected to the set of documents. The strong FMPs are selected based on the minimum support and further refined by selecting the closed FMPs. The procedure of clustering the UMDs using FMPs is explained in following steps.

1. Collect the list of closed FMPs $fmpList$ for each UMD
2. For each $umd_i \in UMD$, compute the similarity with other UMDs based on their $fmpList$.
3. Select the UMDs that has higher similarity with umd_i according to (4.2.2), form the cluster with umd_i by merging their $fmpList$ with that of umd_i.
4. Repeat the steps 1-3 for all UMDs.
5. Repeat the steps 2-4 until no clusters are there to merge.

For Example, let the closed FMPs mined are as follows:

$fmp_1 \longrightarrow \{s_{1_1} s_{1_3}\}(1, 2, 3)$

$fmp_2 \longrightarrow \{s_{1_1} s_{1_3} s_{1_2}\}(1, 2)$

$fmp_3 \longrightarrow \{s_{1_2} s_{1_3}\}(1, 2, 3)$

$fmp_4 \longrightarrow \{s_{3_4}\}(3, 4, 5, 6)$

$fmp_5 \longrightarrow \{s_{3_4} s_{4_2} s_{4_3}\}(4, 5)$

$fmp_6 \longrightarrow \{s_{3_4} s_{4_2}\}(4, 5, 6)$

The extracted list of closed FMPs, $fmpList$ for each $umd \in UMD$ is as follows:

$umd_1 \longrightarrow \{fmp_1, fmp_2, fmp_3\}$

$umd_2 \longrightarrow \{fmp_1, fmp_2, fmp_3\}$

$umd_3 \longrightarrow \{fmp_1, fmp_3, fmp_4\}$

$umd_4 \longrightarrow \{fmp_4, fmp_5, fmp_6\}$

$umd_5 \longrightarrow \{fmp_4, fmp_5, fmp_6\}$

$umd_6 \longrightarrow \{fmp_4, fmp_6\}$

The pairwise similarity between the UMDs is computed based on their $fmpList$. The $umd_i \in UMD$ forms the cluster with other UMDs which are more similar to umd_i. In the above example, the umd_2 is merged with umd_1, as it has higher similarity with umd_1 according to (4.2.2). Similarly, umd_1 is merged with umd_3, umd_5 is merged with umd_4 and umd_5 is merged with umd_6. So, finally two clusters (umd_1, umd_2, umd_3) and (umd_4, umd_5, umd_6) are formed for the dataset.

4.3 Computational Complexity of MSTD Representation and MSTD based Knowledge Extraction Methods

MSTD representation is constructed for the data set of UMDs that are obtained using the MSC method by converting multimedia objects as signal objects. The computation complexity of domain conversion step depends on the methodologies used for the conversion of text, image and audio into signal objects. The time complexity for constructing a standard STD model for the documents having m words is $O(m^2)$. However, it can be constructed in a linear time of $O(m)$ using Ukkonen's algorithm (Ukkonen, 1995).

The MSTD representation is constructed by computing the similarity among the m signal objects with p features. The representation has k first level phrase nodes for UMDs with k unique signal objects such that $k = m - d$ where d is duplicate objects. While constructing the MSTD representation, each suffix phrase compares with k first level phrase nodes. Once match is found, the v child phrase nodes are traversed to place suffix phrase in the suffix tree. Hence, the computation complexity for the construction of MSTD representation is $O((k^2 + v)p)$ for k signal objects with p features. As p is constant and very small compared to k, it can be ignored.

In the MSTD based classification, the test MMD objects need to be compared with k first level phrase nodes of MSTD representation. The matched first level phrase node traverses through its child phrase nodes v to get the information about the MMDs that has multimedia objects similar to test MMD. Hence the time complexity of MSTD based classification is $O(k + v)$ assuming the length of test MMD is very small compared to k.

The MSTD based clustering algorithm requires the collection of first level phrase nodes for each document. The time complexity for retrieving all the first level phrase nodes that has the document identifier of an UMD is $O(k)$. Therefore, the time cost of building first level phrase nodes for N UMDs is $O(Nk)$. In order to cluster the documents, pairwise similarity in terms of leaf node list has to be computed which leads time cost of $O(N^2)$. Thus, the total time complexity for clustering the multimedia documents is $O(Nk) + O(N^2)$.

The phrase nodes of the MSTD representation provides the FMPs for the UMDs. The time complexity for retrieving FMPs from MSTD representation is $O(n)$ with n phrase nodes.

4.4 Experimental Results and Discussion

The MSTD representation is evaluated by experimenting with four datasets of MMDs for the knowledge extraction MMDs. The details about the datasets MDS1, MDS2, Flickr and MDDS are explained in Section 2.9. The classification is performed by dividing the four datasets into training and test datasets in the ratio of 80%-20%. The experiments are performed for various object similarity thresholds in %, $thresh_{ob} = \{0, 1, 3, 5, 7, 10, 13, 15, 20\}$

In Section 4.4.1, we evaluate the performance of MSC model for the representation of MMDs using VSD model. In Section 4.4.2, we discuss the results of classification of MMDs using MSTD representation. In Section 4.4.3, we discuss the results of MSTD based clustering of MMDs. The evaluation of

MSTD based frequent pattern mining and association rule generation is discussed in Section 4.4.4.

4.4.1 Results of VSD based Classification of Multimedia Documents

In this section we evaluate the performance of MSC model for the representation of MMDs by implementing the VSD model. The VSD model is applied for the training dataset to get the representations of the MMDs. Each MMD from test dataset is posed as query in order to get the similar MMD. The retrieved MMD is considered relevant only when its concept is same as the concept of test MMD. The performance of the classification is evaluated by computing the classification accuracy as given in (3.4.3).

In order to evaluate the performance, two experiments have been conducted with the VSD model for four multimodal multimedia datasets. The first experiment is conducted with the original objects of the datasets with original features. Each modality processed separately and uses distinct feature extraction techniques. The images are represented using visual features, audio sounds are represented using audio features and the text documents are represented using bag of keywords. The visual features are the color features and texture features. The color features include HSV color histogram, color autocorrelogram, color moments and the texture features include Gabor filters and wavelet transform. The audio signals are represented by a composite feature vector consists of MFCC (Mel frequency cepstral coefficients) features, RMS value of the signal and energy based features. The similarity between each modality of objects found individually with distinct similarity measure. Based on the experimental analysis we used cosine measure for images and Euclidean measure for audios. Based on the similarity, three separate vectors are constructed for each modality and finally concatenated to form a single vector for each MMD. The test MMD is also processed using the similar procedure. For the second experiment, all the

MMDs are converted as UMDs of signal objects using the MSC method and wavelet transform based signal features are extracted. As all the objects are represented in a unified space, a single vector is formed for each document.

Table 4.1 shows the representation and query time taken by the VSD model for both the experiments. It is noticed the representation time depends on the number of multimedia objects present in the MMDs of the dataset. As the number of multimedia object present in the MMDs of the dataset increases the representation time also increases. The minimum time taken by the VSD model is 0.59 sec for representing the UMDs and 0.61 sec for representing the MMDs of MDS1. The VSD model has taken the maximum of 52.43 sec to represent the MMDs and 51.87 sec to represent the UMDs of MDDS. The query time of the VSD model depends on the representation of training documents and the number of multimedia objects of the test document. The minimum time taken by the VSD model is 1.68 sec for querying the UMDs and 1.755 sec for querying the MMDs of MDS1. The VSD model has taken the maximum of 90.37 sec to query the MMDs and 86.09 sec to query the UMDs of MDS2. The analysis of the results prove that the VSD model has taken less time to represent the UMDs compared to the time taken to represent the MMDs. Also, the time taken to query the UMD is less compared to query the MMD. In order to represent and query the MMDs, the VSD model requires to compute the similarity between the features of each modality of objects individually with distinct similarity measure which consumes more time. The advantage with UMDs is a common similarity measure is used to compute the similarity between the features of signals objects. Hence, the representation and query processing time is reduced for UMDs.

Table 4.1: Comparison of time taken by VSD model for MMDs and UMDs

	Representation Time in sec		Query Time in sec	
Datasets	VSD for MMDs	VSD for UMDs	VSD for MMDs	VSD for UMDs
MDS1	0.61	0.59	1.755	1.68
MDS2	17.10	11.50	90.37	86.09
Flickr	40.51	38.89	60.65	61.10
MDDS	52.43	51.87	39.14	29.91

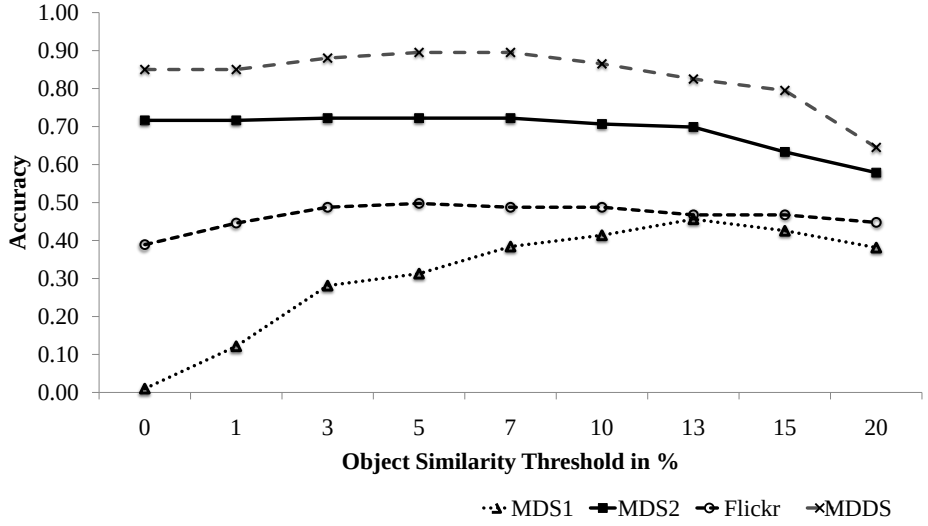


Figure 4.4: Classification using VSD model with Multimodal Objects

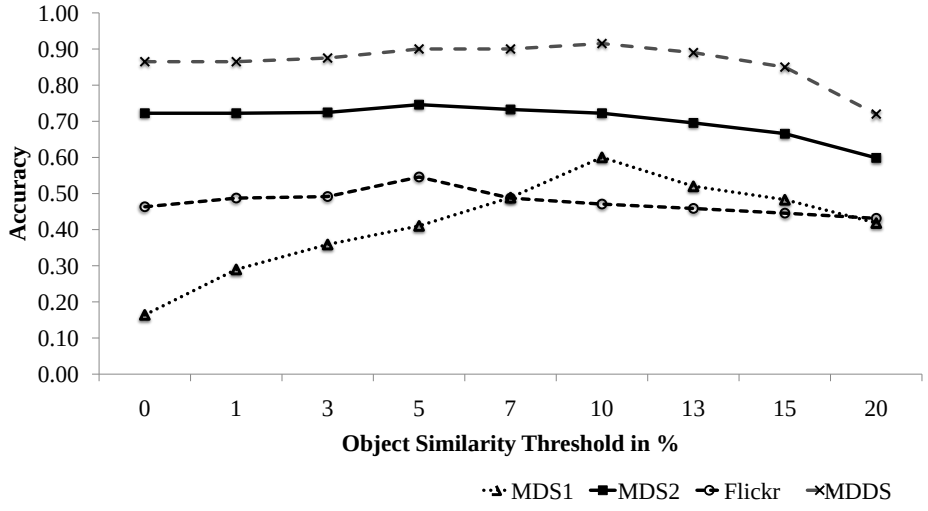


Figure 4.5: Classification using VSD model with Signal Objects

The performance of VSD model for the classification of MMDs with multimodal objects is shown in Fig. 4.4. The VSD model for MMDs achieved the best classification accuracy of 0.456 for MDS1, 0.722 for MDS2, 0.498 for Flickr, and 0.895 for MDDS. Figure 4.5 shows the VSD model based classification of UMDs with signal objects. The VSD model for UMDs achieved the best classification accuracy of 0.60 for MDS1, 0.746 for MDS2, 0.546 for Flickr and 0.920 for MDDS.

The performance comparison of VSD model based classification for MMDs

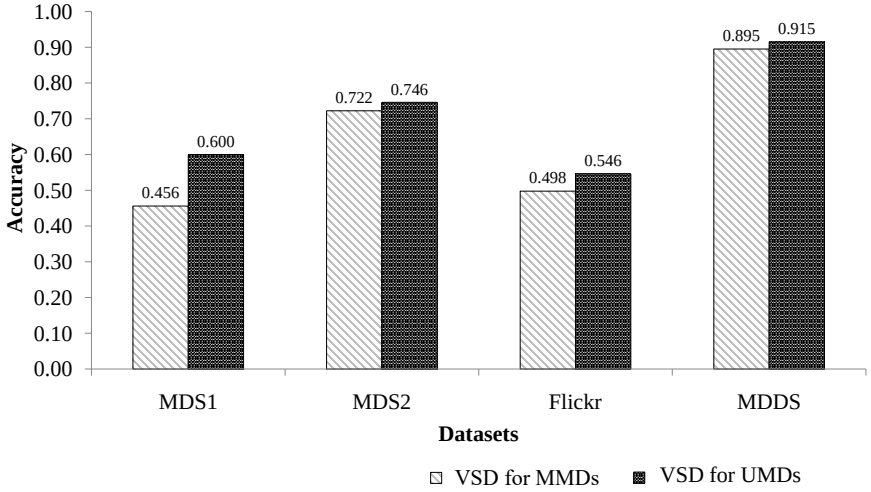


Figure 4.6: Performance Comparison of Classification using VSD model for MMDs and UMDs

and UMDs is shown in Fig. 4.6. The results demonstrate the supremacy of the VSD model with UMDs over VSD model with MMDs. The VSD model with UMDs achieved the improvement in accuracy of 14% for MDS1, 2% for MDS2, 5% for Flickr and 2% for MDDS over the VSD model with MMDs. The improved performance of VSD based classification with UMDs prove that the MSC method has been effectively used for the representation of MMDs using the VSD model.

4.4.2 Results of MSTD based Classification of Multimedia Documents

In this section, we discuss the performance of MSTD based classification for four datasets. Table 4.2 shows the comparison of the time taken by the VSD and MSTD representations to represent the UMDs. It is noticed the representation time of MSTD representation depends on the number of MMDs and the number of multimedia objects exist in each MMD of the dataset. The MSTD representation takes more time for the dataset that has more multimedia objects in the MMDs. The dataset MDDS has taken maximum time of 157.25 sec for the

Table 4.2: Comparison of time taken by VSD and MSTD representation for UMDs

	Representation Time in sec		Query Time in sec	
Datasets	VSD	MSTD	VSD	MSTD
MDS1	0.59	0.50	1.68	0.34
MDS2	11.49	11.27	86.09	3.20
Flickr	38.89	90.55	61.10	9.29
MDDS	51.87	157.25	29.91	17.47

representation as it has the MMDs with maximum number of multimedia objects. The dataset MDS1 has MMDs with two multimedia objects, so it has taken minimum time of 0.50 sec for representation. The MMD querying time depends on the number of branches and branch level of the MSTD representation. The maximum query time is 17.47 for MDDS and minimum time is 0.34 for MDS1. The analysis of the results indicate that the time taken by the MSTD representation is significantly higher for datasets Flickr and MDDS. As the MMDs of these two datasets contain higher number of multimedia objects, time taken to construct the MSTD representation with lengthy branches is more compared to VSD model. However, it is noticed that the time taken by the MSTD representation for processing the query is significantly lower compared to VSD model. The MSTD representation has shown the minimum improvement of 1.34 sec for MDS1 and maximum improvement of 82.89 sec for MDS2 compared to VSD model. The lower time taken to process the query proves that the MSTD representation reduces the search time of the query.

The results shown in Fig. 4.7 demonstrates the performance of the proposed MSTD based classification. The method achieved the best classification accuracy for MDS1 is 0.66, MDS2 is 0.83 and Flickr is 0.69 and MDDS is 0.93. The optimal value of object similarity threshold at which maximum classification accuracy achieved is 7% for MDS1, 10% for MDS2, 5% for Flickr, and 5% for MDDS.

The MSTD based classification is compared with VSD based classification for UMDs and the performance comparison is shown in Fig. 4.8. In comparison with VSD model, the MSTD representation achieved improvement of 6% for MDS1, 9% for MDS2, 14% for Flickr, and 1% for MDDS. The significant improvement

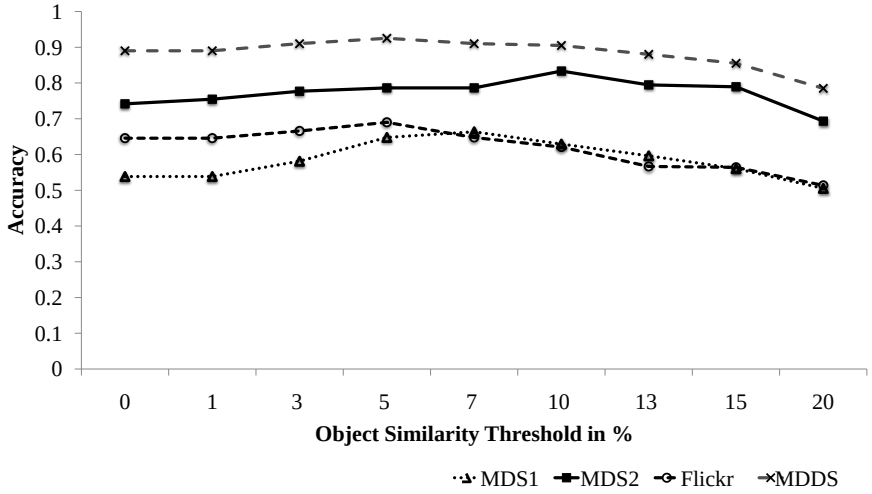


Figure 4.7: Classification using MSTD representation

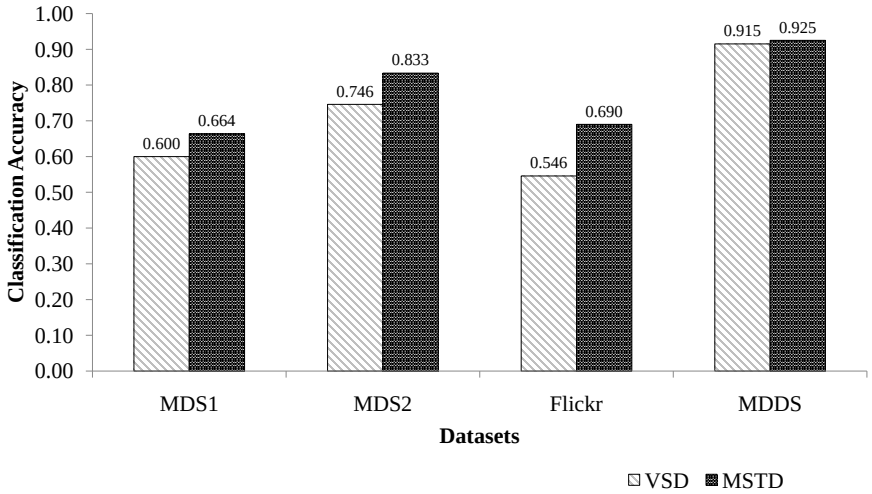


Figure 4.8: Performance Comparison of Classification of MMDs using VSD model and MSTD Representation

of MSTD based classification over VSD based classification prove that the MSTD representation is efficiently used for the classification of multimedia documents.

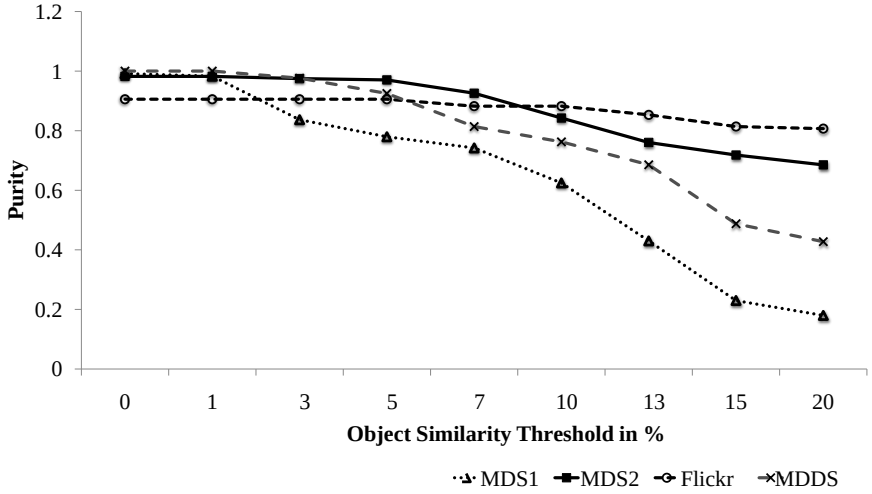
4.4.3 Results of MSTD based Clustering for Multimedia Documents

In this section, the effectiveness of the MSTD based clustering is evaluated by clustering the MMDs. The analysis of clustering quality of clusters formed using

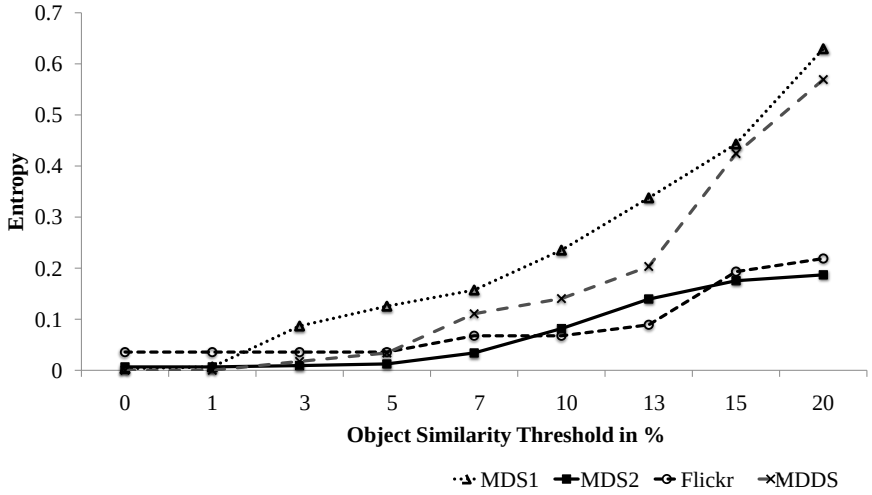
the MSTD based clustering algorithm is expressed in terms of purity and entropy. The purity and entropy values are obtained according to equations (3.6.4) and (3.6.5). Figure 4.9 demonstrates the performance of the proposed MSTD based clustering. The higher purity values obtained using the proposed algorithm are 0.99 for MDS1, 0.98 for MDS2, 0.91 for Flickr, and for MDDS is 1. Similarly, the lower entropy value is 0.003 for MDS1, 0.006 for MDS2, 0.03 for Flick, and 0 for MDDS. The results indicate that the purity and entropy values remain stable for the object similarity threshold between 0 to 7 for all the four datasets. The higher values of purity and lower values of entropy indicates the goodness of the MSTD based clustering algorithm. The experimental results prove that the MSTD based representation is effectively used to cluster the MMDs.

4.4.4 Results of MSTD based Frequent Multimedia Pattern Mining and Multimedia Association Rule Generation

The usage of MSTD representation for frequent multimedia pattern mining and multimedia association rule generation is presented in this section. Table 4.3 shows the details of the number of FMPs and MCARs generated for four datasets. It is observed that the number of FMPs and MCARs are decreasing with the increase in object similarity threshold. The increase in object similarity threshold results in grouping of more similar signal objects. This also causes to decrease in the length of FMPs. It has been noticed that the maximum length of FMPs generated from MDS1 and MDS2 are 2 and 5 respectively. These values remain constant with increase in object similarity threshold due to the limited number of multimedia data present in these two datasets. These two datasets are restricted to have only one multimedia object of one type in a multimedia document. The dataset Flickr contains more text labels and the dataset MDDS contains more number of multimedia objects of different modality. Hence, the



(a) Purity values



(b) Entropy values

Figure 4.9: Performance Analysis of MSTD based Clustering of MMDs

number of generated FMPs for these two datasets are high compare to other two datasets. The maximum FMP length of Flickr is 70 and MDDS is 161. It can be seen that the number of MCARs generated for dataset Flickr and MDDS are very huge because of longer length of FMP.

The advantages of MCARs has been experimented with MCAR based classification for UMDs. The performance evaluation of MCAR based classification is depicted in Fig. 4.10. The MCAR based classification achieves the maximum classification accuracy of 0.69 for MDS1, 0.84 for MDS2, 0.73 for Flickr, and 0.93 for MDDS. Figure 4.11 shows the performance comparison of

Table 4.3: Number of FMPs and MCARs generated using MSTD representation for four datasets

$thresh_{ob}$ in %	FMPs	Closed FMPs	Max length of FMP	MCARs
0	450	346	2	568
1	446	343	2	565
2	452	313	2	535
3	390	288	2	510
5	363	264	2	486
7	327	240	2	456
10	261	203	2	393
13	199	159	2	323
15	165	134	2	272
20	113	94	2	209

(a) MDS1

$thresh_{ob}$ in %	FMPs	Closed FMPs	Max length of FMP	MCARs
0	2280	1953	5	10035
1	2280	1953	5	10035
2	2280	1953	5	10029
3	2280	1953	5	10013
5	2256	1933	5	9981
7	2193	1871	5	9871
10	2154	1725	5	9575
13	1910	1577	5	9263
15	1834	1498	5	9078
20	1718	1353	5	8763

(b) MDS2

$thresh_{ob}$ in %	FMPs	Closed FMPs	Max length of FMP	MCARs
0	13985	2171	70	1.61E+20
1	13985	2169	70	8.68E+19
2	13982	2169	70	8.68E+19
3	13979	2163	70	8.68E+19
5	13965	2160	70	7.27E+19
7	13938	2125	70	7.15E+19
10	13804	2106	68	7.06E+19
13	13328	2047	60	1.79E+18
15	12819	2029	55	6.93E+16
20	12342	1987	49	4.33E+16

(c) Flickr

$thresh_{ob}$ in %	FMPs	Closed FMPs	Max length of FMP	MCARs
0	15739	1687	161	7.31E+47
1	15715	1687	161	7.31E+47
2	15660	1682	160	3.66E+47
3	15607	1655	160	7.34E+46
5	15359	1514	152	5.94E+45
7	14322	1409	136	1.08E+39
10	11650	1302	92	6.29E+26
13	9167	966	57	4.06E+16
15	7402	955	41	3.29E+12
20	5243	738	27	7.23E+11

(d) MDDS

MSTD and MSTD-MCAR based classification of multimedia documents. In comparison with results of MSTD based classification, the MCAR based classification attained performance improvement of 1% for MDS1, 1% for MDS2, and 4% for Flickr.

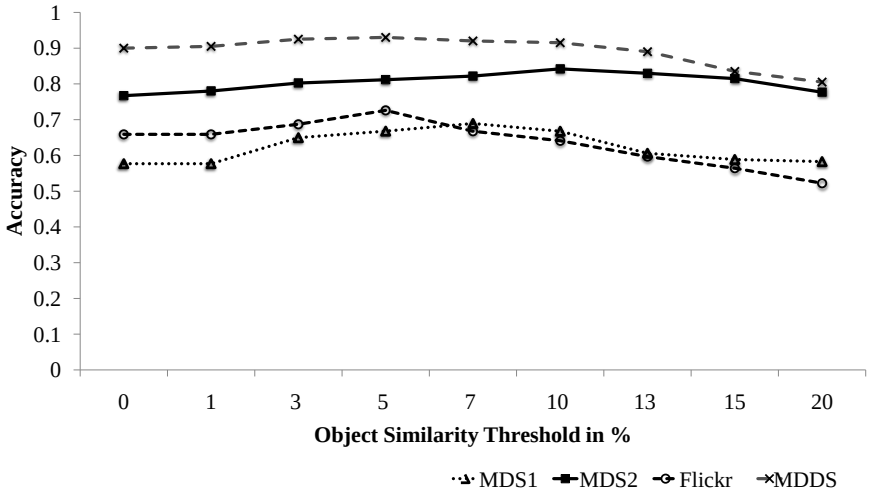


Figure 4.10: MSTD-MCAR based Classification of MMDs

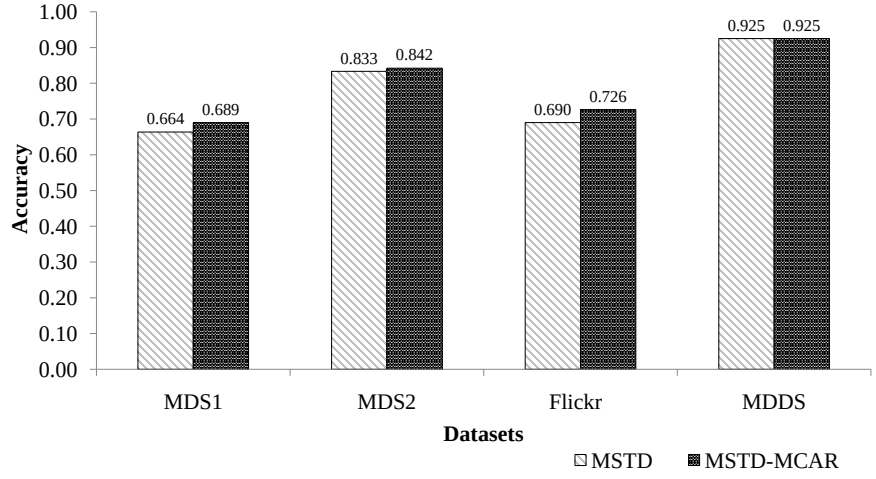
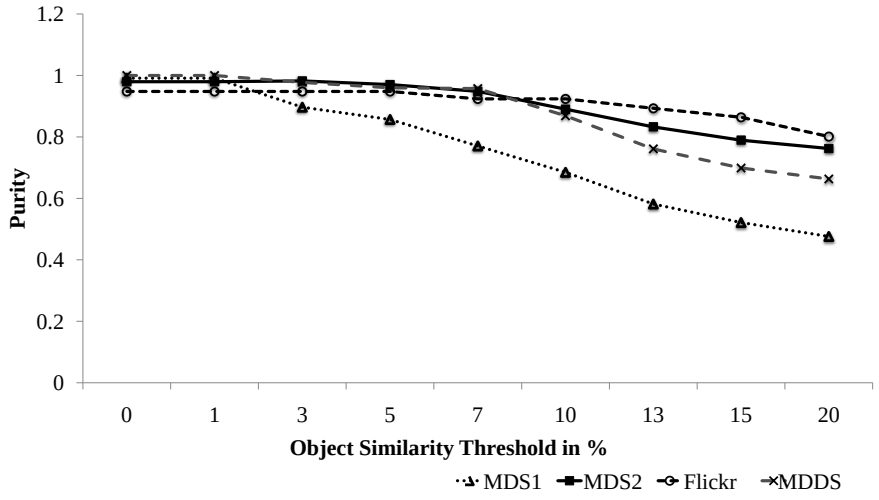
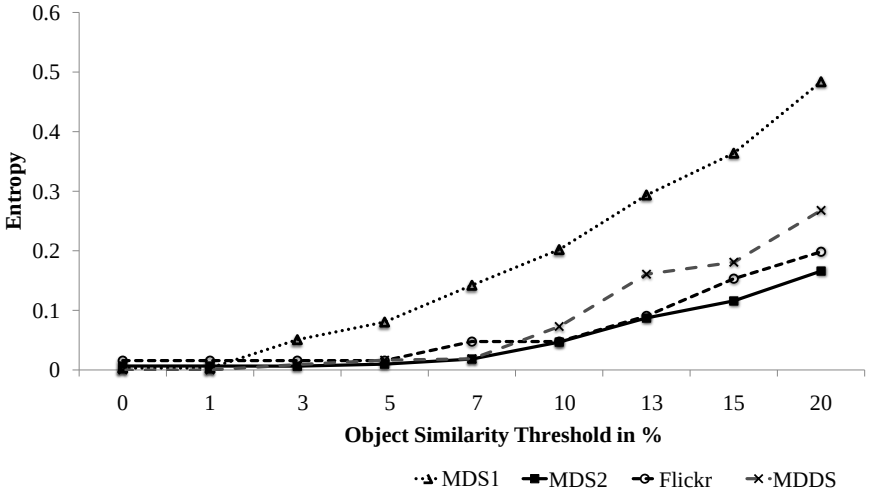


Figure 4.11: Comparison of MSTD-MCAR and MSTD based Classification



(a) Purity Values



(b) Entropy Values

Figure 4.12: Purity and Entropy values of MSTD-FMP based clustering

The closed FMPs mined from MSTD representation are used to cluster the MMDs. Figure 4.12 illustrates the purity and entropy values computed from the FMP based clustering method for four datasets MDS1, MDS2, Flickr, and MDDS. The purity values show the steady value for the object similarity

Table 4.4: Comparison of MSTD and MSTD-FMP based Clustering

Datasets	Purity		Entropy	
	MSTD	MSTD-FMP	MSTD	MSTD-FMP
MDS1	0.991	0.991	0.003	0.003
MDS2	0.982	0.982	0.007	0.006
Flickr	0.906	0.948	0.036	0.016
MDDS	1.000	1.000	0.000	0.000

threshold value between 0 to 7. The best purity values obtained by the MSTD-MFP based clustering are 0.99 for MDS1, 0.98 for MDS2, 0.95 for Flickr, and 1 for MDDS. The best entropy values are 0.003 for MDS1, 0.006 for MDS2, 0.015 for Flickr, and 0 MDDS.

The best purity and entropy values of MSTD-FMP based clustering are compared with MSTD based clustering and the performance comparison is presented in Table 4.4. The MSTD-FMP based clustering shows the steady performance compared to MSTD based clustering. It is observed that, the best values of purity and entropy are almost same for both the algorithms. The experimental results of MSTD-MCAR based classification and MSTD-FMP clustering prove that the MSTD based representation effectively used for the generation of FMPs and MCARs.

4.5 Summary

In this chapter, we discussed the MSTD representation that represents the MMDs in a unified representations using the MSC method. MSTD representation represents the UMDs based on the similar objects between the documents without losing any information. It is used as a platform for multimedia knowledge extraction methods. To validate the efficacy of the MSTD representation, we developed MSTD based classification, clustering, frequent multimedia pattern mining and multimedia class association rules generation approaches. The experimental results prove that the MSTD based classification

achieved superiority over the VSD based classification. Moreover, the MSTD representation takes significantly less time to query the MMDs compared to VSD model. The higher purity values and lower entropy values prove the efficacy of the MSTD based clustering algorithm. The FMPs has been discovered from the phrase nodes of the MSTD representation. The FMPs has been employed for clustering the UMDs. The MSTD-FMP based clustering achieved the best purity and entropy values close to MSTD based clustering. The FMPs are extended to generate the MCARs that has been used to classify the UMDs. The MSTD-MCAR based classification attained the better classification accuracy compared to MSTD based classification. The significant performance of the MSTD based multimedia mining methods prove that the MSTD representation helps in improving the performance of knowledge extraction methods for MMDs.

Chapter 5

Multimedia Feature Pattern Tree Representation

In the previous chapter, we discussed the representation of multimedia documents based on the shared multimedia objects using the MSTD representation. In this chapter, we propose a multimedia document representation based on the similarity of the features of multimedia objects. The proposed representation discovers the meaningful feature patterns hidden in the MMDs. In order to discover the useful knowledge of MMDs, we propose the multimedia mining techniques based on the proposed multimedia document representation. Our contributions are:

- A compact and complete representation, Multimedia Feature Pattern Tree (MFPT) for the MMDs.
- Effective MFPT based multimedia mining methods for the efficient knowledge extraction from MMDs.

The rest of the chapter is organized as follows. In Section 5.1, we discuss the MFPT representation. In Section 5.2, we discuss the knowledge extraction from MMDs using the MFPT representation. In Section 5.3, we discuss the computational complexity of MFPT representation and MFPT based multimedia mining methods. The experimental results are discussed in Section 5.4.

5.1 Multimedia Feature Pattern Tree Representation

The generation of abstraction for the document representation will benefit the overall decision of knowledge discovery systems by the reduction of search time and memory requirements. In knowledge discovery, patterns are the main representation of any transaction. With multimedia documents, the transactions are viewed as the collection of patterns of multimedia objects whereas each multimedia object is represented by the collection of pattern of features. The completeness and compactness of the PC tree (Ananthanarayana et al., 2003) motivated us to propose an enhanced version of PC tree known as Multimedia Feature Pattern tree (MFPT) to represent the MMDs in a compact representation. To the best of our knowledge this is the first effort to represent the multimodal MMDs in a compact representation with respect to the features of the multimedia objects.

With multimodal multimedia objects, it is difficult to get useful feature patterns, as multimodal features are in different feature space with different dimensions and characteristics. To deal with multimodal multimedia objects, we employed the MSC method discussed in section 3.2.1 to convert the MMDs as UMDs. Hence, the proposed MFPT representation represents the UMDs in a compact representation based on the patterns of the features of the signal objects.

MFPT is a rooted tree with the branches representing the signal objects of the UMDs. The branches contain the feature nodes to store the feature values of the signal object. The feature node of the MFPT consists of three attributes *val*, *child* and *sibling*. The *val* attribute stores the feature value of the signal object. The *child* attribute points the child node and the *sibling* attribute points the sibling node. MFPT is built by evaluating the similarity between the feature patterns of the signal objects of UMDs. The feature patterns are evaluated based on the similarity of the sequential features. The branches of the MFPT

represents the feature patterns of signal objects. The signal objects of an UMD are inserted as the branches of the MFPT, depending on the similarity between the feature values and the corresponding feature node values sequentially. Section 5.1.1 describes the construction of MFPT representation for the given dataset of UMDs and Section 5.1.2 describes the characteristics of the MFPT representation.

5.1.1 Multimedia Feature Pattern Tree Construction

Let a UMD umd_i be the collection of m signal objects, $umd_i = \{s_1, s_2, ..., s_m\}$. Let $s_1 = \{fs_{1_1}, fs_{1_2}, ..., fs_{1_p}\}$ be the signal object with p feature values, $b_k = \{fnd_{k_1}, fnd_{k_2}, ..., fnd_{k_p}\}$ be the branch of the MFPT with p feature nodes. The signal objects are placed in the MFPT as branches by storing the feature values in feature nodes. The pairwise similarity between a feature value of the signal object and the corresponding feature node value is computed in order to place the signal object in the MFPT. The similarity is computed based on whether the feature values of signal objects are within the user defined threshold range. Let $thresh_{ob}$ be the user defined object similarity threshold. The feature value fs_{1_p} of s_1 and feature node value $fnd_{k_p}.val$ of b_k are considered similar if $|fs_{1_p} - fnd_{k_p}.val| * 100 <= fnd_{k_p}.val * thresh_{ob}$.

According to object similarity threshold, when all the feature values of the signal object and the feature node values of a branch are similar to each other, then the signal object uses the same feature nodes of that branch. The feature nodes store the minimum value among the feature value and feature node value. If the similarity is not within the object similarity threshold, the sequence of features form a new sub branch from the feature node where the dissimilarity is found. Every branch of the MFPT is terminated by the leaf node which stores the document identifiers of the objects that are traversed through it. Due to the MSC's ability of retaining the original characteristics of the media object, MFPT forms the separate branches for image, audio and text signal objects. The algorithm for the MFPT creation is given in Algorithm 5.1.

Algorithm 5.1 Construction of Multimedia Feature Pattern Tree

Input: Dataset $UMD = \{umd_1, umd_2, ..., umd_N\}$, object similarity threshold $thresh_{ob}$
Output: MFPT Representation
Let root of the MFPT be R
for each unified multimedia document $umd_i \in UMD$ **do**
    for each signal object $s_j \in umd_i$ **do**
        Let $s_j = \{fs_{j_1}, fs_{j_2}, ..fs_{j_p}\}$ are the features of signal object s_j
        Let B be the branches of the tree
        for each branch $b_k \in B$ **do**
            Let $b_k = \{fnd_{k_1}, fnd_{k_2}, ..fnd_{k_p}\}$ are the feature nodes of branch b_k
            $m \leftarrow 1$
            repeat
                if $|fnd_{k_m}.val - fs_{j_m}| * 100 <= fnd_{k_m}.val * thresh_{ob}$ **then**
                    $fnd_{k_m}.val \leftarrow min(fnd_{k_m}.val, fs_{j_m})$
                    $m \leftarrow m + 1$
                end if
            until ($m > 1$ and $m <= p$)
            if $m > 1$ **then**
                Create sub branch $b_{k'} \in B$: $b_{k'} = \{fnd_{k'_x} |_{x=m}^{p}\}$; $fnd_{k'_x}.val = fs_{j_x}$
            else
                Create a new branch $b_n \in B$: $b_n = \{fnd_{n_x} |_{x=1}^{p}\}$; $fnd_{n_x}.val = fs_{j_x}$
            end if
        end for
    end for
end for

An Example

Let us consider the UMD is a dataset of UMDs comprised of signal objects such that,

$UMD = \{umd_1, umd_2, umd_3, umd_4, umd_5, umd_6\}$;

$umd_1 = \{s_{1_1}, s_{1_2}, s_{1_3}\}$

$umd_2 = \{s_{2_1}, s_{2_2}, s_{2_3}, s_{2_4}, s_{2_5}\}$

$umd_3 = \{s_{3_1}, s_{3_2}, s_{3_3}, s_{3_4}\}$

$umd_4 = \{s_{4_1}, s_{4_2}, s_{4_3}, s_{4_4}, s_{4_5}\}$

$umd_5 = \{s_{5_1}, s_{5_2}, s_{5_3}, s_{5_4}\}$

$umd_6 = \{s_{6_1}, s_{6_2}, s_{6_3}\}$

where $s_{1_1}...s_{1_3}$, $s_{2_1}...s_{2_5}$, $s_{3_1}...s_{3_4}$, $s_{4_1}...s_{4_5}$, $s_{5_1}...s_{5_4}$ and $s_{6_1}...s_{6_3}$ are the signal objects. Each signal object is subjected to feature extraction to extract the features and represented by the feature vector. Let us consider following are the feature values of the signal objects.

$s_{1_1} = \{f_{1_{1_1}}..f_{1_{1_9}}\} \Rightarrow \{0.37, 1.14, 1.63, 2.32, -1.97, -6.74, -5.72, -4.86, 2.65\}$

$s_{1_2} = \{f_{1_{2_1}}..f_{1_{2_9}}\} \Rightarrow \{0.97, 0.40, 0.81, 1.34, -0.05, -2.52, -1.00, 0.06, 2.71\}$

$s_{1_3} = \{f_{1_{3_1}}..f_{1_{3_9}}\} \Rightarrow \{0.98, 0.40, 0.84, -1.82, -1.02, -0.68, -0.92, 0.94, 2.40\}$

$s_{2_1} = \{f_{2_{1_1}}..f_{2_{1_9}}\} \Rightarrow \{0.40, 1.14, 1.65, 2.35, -1.97, -6.74, -5.72, -4.86, 2.18\}$

$s_{2_2} = \{f_{2_{2_1}}..f_{2_{2_9}}\} \Rightarrow \{0.97, 0.40, 0.81, 1.34, -0.05, -2.52, -1.00, 0.06, 2.71\}$

$s_{2_3} = \{f_{2_{3_1}}..f_{2_{3_9}}\} \Rightarrow \{0.97, 0.40, 0.81, 1.34, -0.05, -2.52, -1.00, 0.06, 2.35\}$

$s_{2_4} = \{f_{2_{4_1}}..f_{2_{4_9}}\} \Rightarrow \{0.98, 0.40, 0.84, -1.82, 1.27, 0.35, -0.58, -1.90, 1.56\}$

$s_{2_5} = \{f_{2_{5_1}}..f_{2_{5_9}}\} \Rightarrow \{0.98, 0.40, 0.84, -1.82, -1.02, -0.68, -0.92, 0.12, 1.93\}$

$s_{3_1} = \{f_{3_{1_1}}..f_{3_{1_9}}\} \Rightarrow \{0.97, 0.40, 0.81, 1.34, -0.05, -2.52, -1.00, 0.06, 2.36\}$

$s_{3_2} = \{f_{3_{2_1}}..f_{3_{2_9}}\} \Rightarrow \{0.39, 1.14, 1.63, 2.32, -1.97, -6.70, -5.68, -4.84, 2.88\}$

$s_{3_3} = \{f_{3_{3_1}}..f_{3_{3_9}}\} \Rightarrow \{0.98, 0.40, 0.84, -1.82, -1.02, -0.68, -0.92, 0.94, 2.38\}$

$s_{3_4} = \{f_{3_{4_1}}..f_{3_{4_9}}\} \Rightarrow \{0.15, 0.62, 0.45, 0.93, 1.17, -0.95, -1.15, -1.90, -1.55\}$

$s_{4_1} = \{f_{4_{1_1}}..f_{4_{1_9}}\} \Rightarrow \{0.15, 0.62, 0.45, 0.93, 1.17, -0.95, -1.15, -0.54, -0.94\}$

$s_{4_2} = \{f_{4_{2_1}}..f_{4_{2_9}}\} \Rightarrow \{0.52, 0.42, 0.85, 1.00, -1.31, -1.17, -1.08, -2.42, 0.96\}$

$s_{4_3} = \{f_{4_{3_1}}..f_{4_{3_9}}\} \Rightarrow \{1.07, 0.33, 0.60, 1.10, 0.14, -1.57, -0.24, -0.36, 1.63\}$

$s_{4_4} = \{f_{4_{4_1}}..f_{4_{4_9}}\} \Rightarrow \{1.12, 0.36, 0.62, 1.10, 0.14, -1.57, -0.20, -0.36, 1.86\}$

$s_{4_5} = \{f_{4_{1_1}}..f_{4_{1_9}}\} \Rightarrow \{0.15, 0.62, 0.45, 0.93, 1.17, -0.95, -1.15, -1.90, -1.25\}$

$s_{5_1} = \{f_{5_{1_1}}..f_{5_{1_9}}\} \Rightarrow \{0.75, 2.62, 2.45, -0.73, -0.17, 0.35, -1.15, -1.90, -1.55\}$

$s_{5_2} = \{f_{5_{2_1}}..f_{5_{2_9}}\} \Rightarrow \{0.52, 0.42, 0.85, 1.00, -1.31, -1.17, -1.08, -2.42, 0.97\}$

$s_{5_3} = \{f_{5_{3_1}}..f_{5_{3_9}}\} \Rightarrow \{0.52, 0.42, 0.85, 1.00, -1.31, -1.17, -1.08, -2.42, 0.76\}$

$s_{5_4} = \{f_{5_{4_1}}..f_{5_{4_9}}\} \Rightarrow \{1.07, 0.33, 0.60, 1.10, 0.14, -1.57, -0.24, -0.36, 1.73\}$

$s_{6_1} = \{f_{6_{1_1}}..f_{6_{1_9}}\} \Rightarrow \{0.15, 0.62, 0.45, 0.93, 1.17, -0.95, -1.15, -1.90, -1.25\}$

$s_{6_2} = \{f_{6_{2_1}}..f_{6_{2_9}}\} \Rightarrow \{0.52, 0.42, 0.85, 1.00, -1.31, -1.17, -1.08, -2.42, 0.99\}$

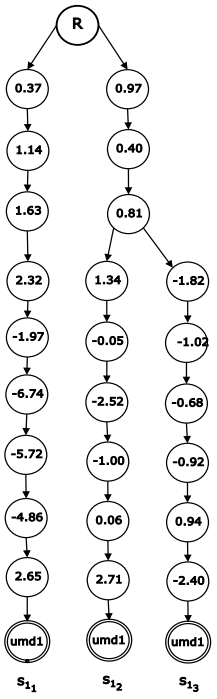
$s_{6_3} = \{f_{6_{3_1}}..f_{6_{3_9}}\} \Rightarrow \{0.55, 0.45, 0.86, 1.03, -1.31, -1.17, -1.08, -2.42, 1.26\}$

The MFPT representation is constructed for the given dataset of UMDs by scanning the UMDs one by one. To begin with, the document umd_1 is scanned

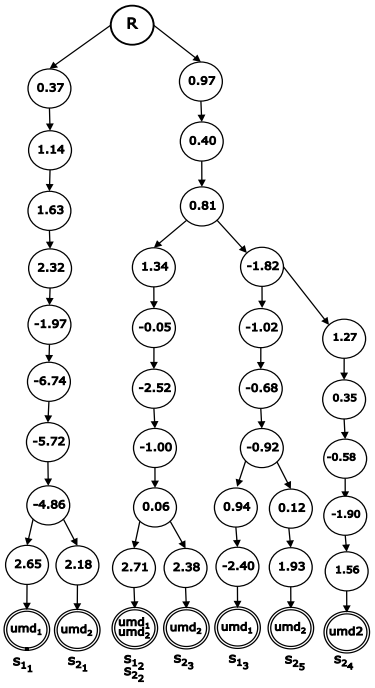
and the signal objects s_{1_1}, s_{1_2} and s_{1_3} are extracted. The value of $thresh_{ob}$ is assumed as 5. Initially MFPT will be empty, so the signal object s_{1_1} forms the first branch b_1 using all feature values $f_{1_{1_1}}..f_{1_{1_9}}$ as feature nodes $fnd_{1_1}..fnd_{1_9}$. The branch is terminated by the leaf node that stores the document identifier of umd_1.

With s_{1_2}, the feature value $f_{1_{2_1}}$ (0.97) is compared with the value of feature node fnd_{1_1} (0.37) of branch b_1. The value of $|f_{1_{2_1}} - fnd_{1_1}.val| * 100$ is more than $fnd_{1_1}.val * thresh_{ob}$ i.e. $|0.97 - 0.37| * 100 > 0.37 * 5$. Hence, a new branch b_2 is formed for s_{1_2} using the feature values $f_{1_{2_1}}..f_{1_{2_9}}$ as feature nodes $fnd_{2_1}..fnd_{2_9}$.

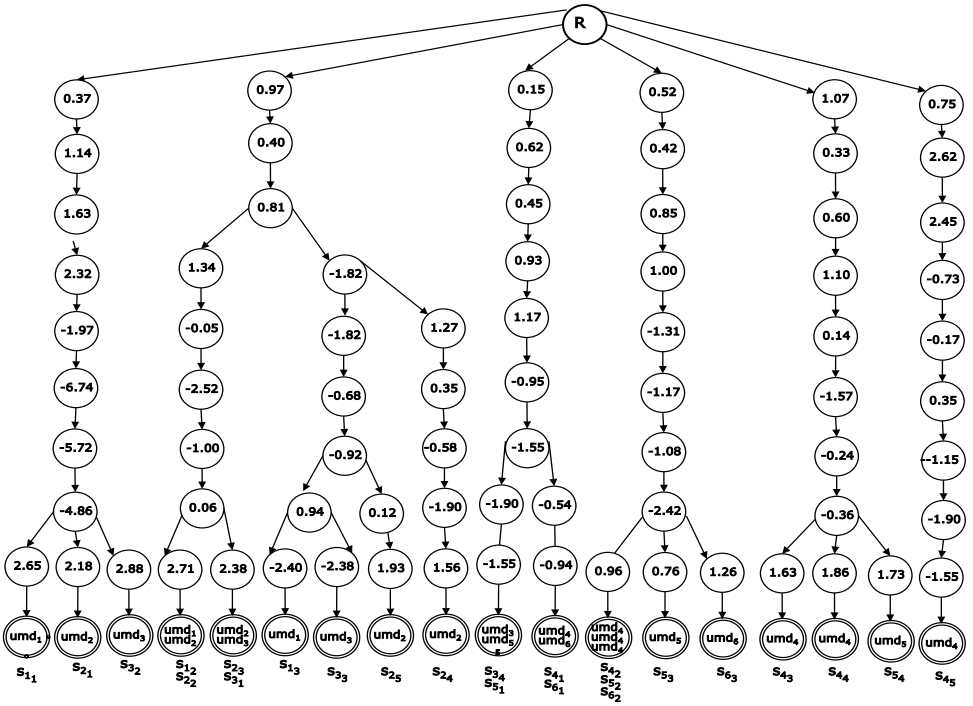
With s_{1_3}, the feature values $f_{1_{3_1}}..f_{1_{3_3}}$ are similar to values of feature nodes $fnd_{2_1}..fnd_{2_3}$ of branch b_2. i.e. $|0.97 - 0.98| * 100 <= 0.97 * 5$, $|0.40 - 0.40| * 100 <= 0.40 * 5$ and $|0.81 - 0.84| * 100 <= 0.81 * 5$. So, s_{1_3} uses the feature nodes $fnd_{2_1}..fnd_{2_3}$ for features $f_{1_{3_1}}..f_{1_{3_3}}$. The feature nodes $fnd_{2_1}..fnd_{2_3}$ store the minimum of similar feature values. For example, $fnd_{2_1}.val$ stores the minimum of (0.97, 0.98) i.e. 0.97. The remaining features $f_{1_{3_4}}..f_{1_{3_9}}$ of s_{1_3} forms the sub branch from feature node fnd_{2_3} of branch b_2. The construction of MFPT for umd_1 is shown in Fig. 5.1a. Similarly the umd_2 is scanned and the signal objects s_{2_1}, s_{2_2}, s_{2_3}, s_{2_4} and s_{2_5} are extracted. The features of the signal objects are compared with the values of the feature nodes. The signal objects use the features nodes of the branches for similar feature values. As shown in Fig. 5.1b, the signal object s_{2_1} follows the branch of s_{1_1} for features $f_{2_{1_1}}..f_{2_{1_8}}$ and forms the sub branch for feature $f_{2_{1_9}}$. The signal object s_{2_2} has all the features similar to that of s_{1_2}, so both use the feature nodes of same branch. The leaf node of the branch stores the identifiers of both the documents umd_1 and umd_2. Similarly the signal object s_{2_3}, s_{2_4}, and s_{2_5} uses the branches of the tree for the similar features and forms the sub branches for the remaining features. If none of the feature values of the signal object are similar to the values of the feature nodes of the branches, a new branch is created. This procedure continues for all the signal objects of all the UMDs and a MFPT representation is constructed. Fig. 5.1c shows the MFPT for the example dataset.



(a) MFPT for document umd_1



(b) MFPT for documents umd_1, umd_2



(c) MFPT for dataset of UMDs, $UMD = \{umd_1, ..., umd_6\}$

Figure 5.1: Construction of MFPT Representation

5.1.2 Characteristics of MFPT Representation

MFPT is a compact representation of UMD dataset in terms of features. It is constructed by evaluating the similarity between the features of the signal objects enclosed by the UMDs. When the sequence of features of two signal objects found similar based on object similarity threshold $thresh_{ob}$, both the signal objects use the same feature nodes of a branch of the MFPT. Hence the representation is compact by avoiding the separate representation for similar signal objects. MFPT is a complete representation of UMDs in terms of features. The completeness of the MFPT depends on the user defined threshold $thresh_{ob}$. The representation is complete when the threshold $thresh_{ob}$ is zero. In this case, the feature nodes store the exact value of the features, as the similarity between the feature values is zero. When the value of $thresh_{ob}$ is more than zero, the feature nodes store the minimum of similar values which results in an almost complete representation of dataset. As the MFPT is constructed for UMDs, the height of all branches of MFPT is same and equal to the dimension of signal object features.

5.2 Knowledge Extraction from Multimedia Documents using MFPT Representation

The MFPT representation supports the multimedia knowledge extraction methods classification, clustering and sequential multimedia pattern mining. Figure 5.2 shows the framework for knowledge extraction from MMDs using the proposed MFPT representation. In Section 5.2.1, we discuss the MFPT based classification of MMDs. In Section 5.2.2, we discuss the clustering of MMDs using MFPT representation. The MFPT based sequential multimedia pattern mining is discussed in Section 5.2.3.

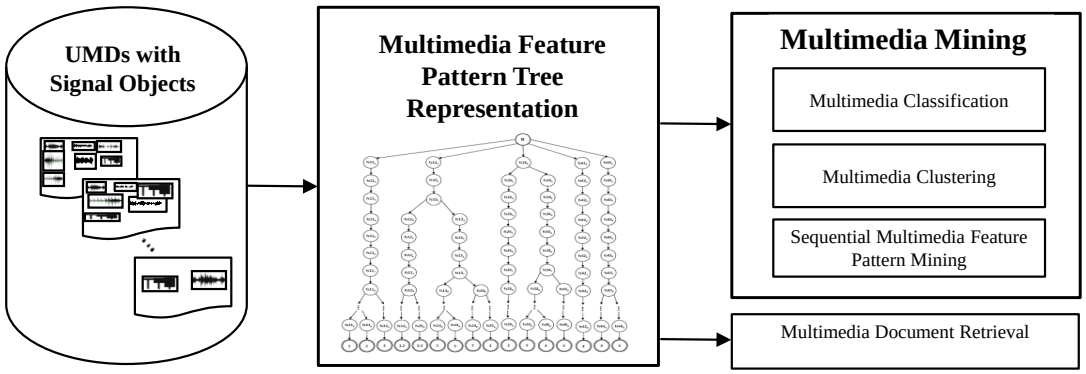


Figure 5.2: Knowledge Extraction from MMDs using MFPT Representation

5.2.1 MFPT based Classification for Multimedia Documents

This section demonstrates how the MMDs are classified using the MFPT representation. Let $UMD_T = \{umd_1, umd_2, .., umd_N\}$ be the training dataset of UMDs and umd_q be the test UMD. The MFPT representation, $MFPT_T$ is constructed for the training dataset UMD_T as explained in algorithm 5.1. Similarly the $MFPT_q$ is constructed for each test document umd_q.

The branches of the $MFPT_q$ are compared with all the branches of the $MFPT_T$ and the relevant documents are retrieved based on the similarity between them. Each branch of the $MFPT_q$ is compared with each branch of the $MFPT_T$ to find the similar branch by computing the similarity between the values of the feature nodes of the branches. The branches of the $MFPT_T$ with maximum similar feature nodes and with minimum difference of feature values are selected as the similar branches.

Let $MFPT_T$ has K branches such that $MFPT_T = \{bt_1, bt_2, .., bt_K\}$. The branches of the $MFPT_T$ which are similar to branches of the test $MFPT_q$ are retrieved as per the following rule,

$$B = \left\{ bt_k : k \in \left(\underset{1 \le j \le K}{\operatorname{argmax}} \left[\sum_{m=1}^{p} cnt_{j_m} \right] \bigcap \underset{1 \le j \le K}{\operatorname{argmin}} \left[\sum_{m=1}^{p} |ft_{j_m} - fq_m| \right] \right) \right\} \quad (5.2.1)$$

In the above equation, bt_k is the k^{th} branch of the $MFPT_T$, $cnt_{j_m} = \begin{cases} 1 & : |ft_{j_m} - fq_m| \leq thresh_{ob} * ft_{j_m} \\ 0 & : otherwise \end{cases}$, ft_{j_m} is the m^{th} feature node of branch bt_j, and fq_m is the m^{th} feature node of query branch. Similarly all the branches of the $MFPT_q$ are compared with $MFPT_T$ and the similar branches having maximum similar feature nodes with minimum dissimilarity of features are selected. The UMDs belong to all the selected branches are retrieved using the leaf nodes of the branches.

The rule (5.2.1) retrieves the documents based on the similarity of the objects of the test document. But all of them may not be relevant to test document. For example, the test document has the details about cat in the form of cat image, cat sound and text document. Let the test text document of cat contains the words *"black cat drinks milk "*. Let the training documents contains the documents about *cat* and *bench*. The *bench* document contain the text information *"black bench "*. As a result the *bench* document also selected as the relevant document to test document *cat* because it contains the word *"black "* which is actually not relevant. Thus, further filtering is needed to select the most relevant documents. Initially, the UMDs that exists in maximum number of branches are selected. The retrieved UMDs are further refined by filtering the documents with minimum average difference for those objects that exist in the test document.

Let $UMD_R = \{umd_1, umd_2, .., umd_o\}$ be the set of selected similar UMDs based on the similarity of the test objects. Among UMD_R, the set of documents UMD'_R that contain maximum objects similar to test document are chosen based on the following rule,

$$UMD'_R = \left\{ umd_k : k \in \underset{1 \leq j \leq o}{\operatorname{argmax}} \left[\sum_{x=1}^{d} cnt_{j_x} \right] \right\} \tag{5.2.2}$$

where umd_k is the k^{th} document of UMD_R, $cnt_{j_x} = \begin{cases} 1 & : s_{j_x} \in umd_q \\ 0 & : otherwise \end{cases}$, s_{j_x} is the x^{th} signal object of $umd_j \in UMD_R$, $s_{j_x} \in umd_q$ implies that umd_q has one of the signal object similar to s_{j_x} and d is the number of similar signal objects exist between umd_j and umd_q. The documents UMD'_R are further refined by filtering the documents with minimum average difference for those data that exist in test document.

Let $UMD'_R = \{umd_1, umd_2, .., umd_z\}$ be the documents obtained as per rule (5.2.2). The number of similar signal objects exist between the each of the document of UMD'_R and test document umd_q is d. The document UMD_{RS}, that is more similar to test document, selected using the following rule,

$$UMD_{RS} = \left\{ umd_k : k = \underset{1 \le j \le z}{\operatorname{argmin}} \left[\frac{1}{d} \sum_{x=1}^{d} \sum_{m=1}^{p} |ft_{j_{x_m}} - fq_{x_m}| \right] \right\} \tag{5.2.3}$$

where $ft_{j_{x_m}}$ is the m^{th} feature of x^{th} similar signal objects of document $umd_j \in UMD'_R$, fq_{x_m} is the m^{th} feature of x^{th} similar signal objects of umd_q. When the concept of the retrieved document UMD_{RS} is same as that of test document umd_q, then it is considered that the retrieved document is most relevant to test document and the test document is correctly classified. The summary of the MFP based classification algorithm is given in Algorithm 5.2.

For Example,

Let the test multimedia document be $umd_q = \{s_{3_4}, s_{4_2}, s_{4_3}\}$. According to MFPT of example dataset shown in Fig. 5.1c, the MMDs that contain signal objects similar to s_{3_4} are umd_3 and umd_5. Both of these MMDs are considered as similar documents as they have the signal objects similar to s_{3_4} with same feature difference value. The MMDs that contain signal objects similar to s_{4_2} are umd_4, umd_5 and umd_6. Although these three documents have maximum similar feature nodes, the difference between the feature values (0.03) of s_{4_2} and s_{6_2} is

Algorithm 5.2 MFPT based Classification Algorithm

1: **Input**: Training Dataset UMD_T of N UMDs, Test Dataset UMD_Q of M UMDs
2: **Output**: $Class_q$=Class of umd_q
3: Construct the MFPT, $MFPT_T$ for the training dataset UMD_T .
4: Construct the MFPT, $MFPT_q$ for umd_q
5: $UMD_R \leftarrow \{\}$
6: **for** each branch $bq_{i_j} \in MFPT_{q_i}$ **do**
7: Find the branch $bt_k \in MFPT_T$ with maximum number of similar features and with minimum difference as compared with bq_{i_j} as per rule (5.2.1)
8: $UMD_R \leftarrow UMD_R \bigcup DocIds$ stored in leaf node of branch bt_k
9: **end for**
10: Filter the relevant documents $UMD'_R \in UMD_R$ using the rule (5.2.2)
11: Select the most relevant document $UMD_{RS} \in UMD'_R$ using the rule (5.2.3)
12: $Class_q \leftarrow$ Class of UMD_{RS}

more compared to the difference between the feature values (0.01) of s_{4_2} and s_{5_2}. Hence, according to rule (5.2.1), umd_6 is ignored. Similarly for s_{4_3}, umd_4 is considered as similar MMD. Hence the UMDs similar to umd_q are retrieved as follows:

$s_{3_4} \longrightarrow \{umd_3, umd_5\}$

$s_{4_2} \longrightarrow \{umd_4, umd_5\}$

$s_{4_3} \longrightarrow \{umd_4\}$

The similar documents retrieved based on the object similarity for umd_q are umd_3 ,umd_4 and umd_5. All these documents are not relevant because umd_3 has only one matching object of test document. Hence, the document with maximum number of matching objects and maximum object similarity is selected as most similar document. Among the retrieved documents, the documents umd_4 and umd_5 are having the maximum number of matching signal objects of test document. The document umd_4 is selected as the most relevant document, as its matching signal objects are more similar to umd_q.

5.2.2 MFPT based Clustering for Multimedia Documents

This section discusses about the clustering of MMDs using the MFPT representation. Let $UMD = \{umd_1, umd_2, .., umd_N\}$ be the dataset of N UMDs. The UMDs are clustered based on the pairwise similarity between the UMDs. The similarity between the UMDs is found based on the leaf nodes of the MFPT. The leaf nodes contain the documents that has the signal objects with sequential similar features. Hence for each UMD, the list of leaf nodes is retrieved using the Algorithm 5.3. After collecting the list of leaf nodes for all the UMDs, the pairwise similarity between any two documents $umd_i \in UMD$ and $umd_j \in UMD$ is computed based on the similarity of their leaf node lists. The $UMDs$ with maximum similarity are the candidates to form cluster with the UMD umd_i.

Let $LnodeList_i$ be the leaf node list of umd_i and $LnodeList_j$ be the leaf node list of umd_j. The $UMDs$ that are having maximum similarity with umd_i are selected using the following equation,

$$UMD_R = \{umd_k : k \in \underset{1 \leq j \leq N}{\operatorname{argmax}} [sim(LnodeList_i, LnodeList_j)]\} \tag{5.2.4}$$

where $sim(LnodeList_i, LnodeList_j)$ is the similarity between umd_i and umd_j based on their leaf node lists. The similarity is calculated as follows:

$$sim(LnodeList_i, LnodeList_j) = 2 * \frac{|LnodeList_i \bigcap LnodeList_j|}{|LnodeList_i + LnodeList_j|} \tag{5.2.5}$$

The minimum value for $sim(LnodeList_i, LnodeList_j)$ is assumed as 0.5 in order to cluster the more similar documents. The $UMDs$ with maximum similarity are clustered with umd_i by combining their leaf node lists with the leaf node list of umd_i. The procedure continues for all the documents till they group them into concepts. The formed clusters are further merged based on the similarity between the clusters. The algorithm repeats till there are no more clusters to merge resulting in optimum clusters.

Algorithm 5.3 Leaf nodes collection for a UMD

1: **Input**: $B = \{b_1, b_2, .., b_K\}$ are branches of MFPT for dataset UMD, umd_{id}: Document ID
2: **Output**: $LnodeList_{id}$: leaf node list for UMD_{id}
3: $LnodeList_{id} \leftarrow \{\}$
4: **for** each branch $b_k \in B$ **do**
5: Let $Lnode_k$ be the leaf node of b_k
6: **if** $umd_{id} \in Lnode_k.DocId$ **then**
7: $LnodeList_{id} = LnodeList_{id} \bigcup Lnode_k$
8: **end if**
9: **end for**
10: return $LnodeList_{id}$

5.2.3 MFPT based Sequential Multimedia Feature Pattern Mining and Sequential Rule Generation

Multimedia objects are characterized by the collection of the features. The patterns of the features describe the characteristics of the multimedia object. Therefore, the MMDs are described by the collection of the feature patterns. Mining sequential feature patterns is advantageous in many applications. For example, in order to retrieve black color animals living in forest, the first feature should describe the background forest and the second feature describes the color of the animal. The MFPT is constructed on the basis of sequential feature patterns in order to represent the MMDs. The sequential feature patterns of MMDs are termed as sequential multimedia feature patterns (SMFP) as the patterns are formed using the sequential features of multimedia objects. The main task of SMFP mining is to generate the sequential feature patterns of multimedia objects that exist in user defined fraction of multimedia documents. The branches of the MFPT represent the SMFPs of each object. The MFPT representation generates the base clusters of MMDs based on the SMFPs of the objects. Hence, the SMFPs are extracted from the branches of the MFPT.

The length of the SMFP is defined as the number of sequential features in the SMFP. As the features determine the characteristics of the objects, the different length of SMFP describes different objects.

Let $UMD = \{umd_1, umd_2, ..., umd_N\}$ be the set of unified multimedia documents, $F = \{f_1, ..., f_m\}$ be the set of all features present in UMD and $C = \{c_1, ..., c_k\}$ be the set of class labels.

Definition 5.2.1. *The multimedia class sequential rule (MCSR) is defined in the form* $SR_m : SeqFeat \longrightarrow c$ *where* $SeqFeat$ *is the sequence of features of given length such that* $SeqFeat \subseteq F$ *and* $c \in C$.

Definition 5.2.2. *The support of the MCSR,* $sup(SR_m)$ *is the number of UMDs that contain* $SeqFeat$ *and are labeled with* c.

Definition 5.2.3. *The occurrence* $occ(SR_m)$ *of rule* SR_m *is the number of UMDs that contain the sequential feature patterns present in* SR_m*'s antecedent.*

Definition 5.2.4. *The confidence of the rule* SR_m *is defined as,*

$$conf(SR_m) = \frac{sup(SR_m)}{occ(SR_m)}$$

The number of SMFPs depend on the user defined minimum support and the length of the SMFP. Each SMFP generates only one MCSR. The SMFPs form the antecedent part of the MCSR and the class label of UMDs that own the SMFP becomes the consequent part of the MCSR. If a MCSR associates with more than one class label, then the class label that has more documents satisfying the MCSR has been assigned. If both the class labels have equal number of documents, then any one class label is assigned to MCSR. Following example explains the generation of SMFPs and MCSRs using the MFPT representation shown in Fig. 5.1

An Example,

Let the length of SMFP is 5 and the user defined minimum support is 2. The SMFP that extracted from the first branch of the MFPT shown in Fig. 5.1 is

<0.37 1.14 1.63 2.32 1.97>. The SMFP is represented in the following form,

$< 0.37\ 1.14\ 1.63\ 2.32\ 1.97 > \{1, 2, 3\}$

where $< 0.37\ 1.14\ 1.63\ 2.32\ 1.97>$ is the SMFP formed by the values of sequential features and 1,2,3 are the identifiers of UMDs umd_1,umd_2 and umd_3 that share the SMFP. Following are the SMFPs extracted from the MFPT representation shown in Fig. 5.1;

$< 0.37\ 1.14\ 1.63\ 2.32\ 1.97> \{1, 2, 3\}$

$<0.97\ 0.40\ 0.81\ 1.34\ 0.05> \{1, 2, 3\}$

$<0.97\ 0.40\ 0.81\ 1.82\ 1.82> \{1, 2, 3\}$

$<0.15\ 0.62\ 0.45\ 0.93\ 1.17> \{3, 4, 5, 6\}$

$<0.52\ 0.42\ 0.85\ 1.00\ 1.31> \{4, 5, 6\}$

$<1.07\ 0.33\ 0.60\ 1.10\ 0.14 > \{4, 5, 6\}$

Let c_1 is the class label of UMDs $\{1, 2, 3\}$ and c_2 is the class label of UMDs $\{4, 5, 6\}$, then the generated MCSR based on the above SMFPs are as follows;

$<0.37\ 1.14\ 1.63\ 2.32\ 1.97> \longrightarrow c_1$

$<0.97\ 0.40\ 0.81\ 1.34\ 0.05> \longrightarrow c_1$

$<0.97\ 0.40\ 0.81\ 1.82\ 1.82> \longrightarrow c_1$

$<0.15\ 0.62\ 0.45\ 0.93\ 1.17> \longrightarrow c_1, c_2$

$<0.52\ 0.42\ 0.85\ 1.00\ 1.31 > \longrightarrow c_2$

$<1.07\ 0.33\ 0.60\ 1.10\ 0.14 > \longrightarrow c_2$

For the MCSR $<0.15\ 0.62\ 0.45\ 0.93\ 1.17> \longrightarrow c_1, c_2$, the class label c_2 is selected as it has more UMDs (4,5,6) compared to class label c_1 (3).

MFPT-MCSR based Classification

The MCSRs are used to classify the unknown UMDs by finding their class labels. The training UMDs are represented using the MFPT representation $MFPT_T$ based on the features of its contents. The SMFPs are extracted from $MFPT_T$ which generates the training MCSRs as discussed in section 5.2.3. The test document umd_q is represented using the features of signal objects by the MFPT representation $MFPT_q$. The SMFPs are extracted from $MFPT_q$ and the test

MCSRs are generated. The antecedent part of the test MCSR is compared with that of training MCSR. The training MCSR that covers more test signal objects is selected and its label is assigned to the test UMD.

For Example,

Let the test document be $umd_q = \{s_{3_4}, s_{1_1}, s_{4_2}, s_{4_3}\}$. The feature values of test signal objects are given in Section 5.2.3. Initially, the test UMD is represented by the MFPT representation using the feature values of test signal objects as explained in Section 5.1.1. The test SMFPs are extracted from the representation and the test MCSRs are generated as discussed above. Let the length of SFMP is 5. The generated test MCSRs are as follows;

$< 0.15\ 0.62\ 0.45\ 0.93\ 1.17 > \longrightarrow c_x$

$< 0.37\ 1.14\ 1.63\ 2.32\ 1.97 > \longrightarrow c_x$

$< 0.52\ 0.42\ 0.85\ 1.00\ 1.31 > \longrightarrow c_x$

$<1.07\ 0.33\ 0.60\ 1.10\ 0.14 > \longrightarrow c_x$

where c_x is the unknown class label.

The test MCSRs are compared with training MCSRs and the corresponding class labels are obtained. The class label that has more test objects is selected as class label for test UMD. In the example c_2 selected from the obtained labels c_1, c_2 as it has more test objects $(s_{3_4} s_{4_2} s_{4_3})$.

5.3 Computational Complexity of MFPT representation and MFPT based Multimedia Mining Methods

MFPT representation is constructed for the data set of UMDs that are obtained using the MSC method by converting multimedia objects as signal objects. The computation complexity of domain conversion depends on the methodologies used for the conversion of text, image and audio into signal objects.

The MFPT is constructed by scanning the dataset of documents only once. The branches of the MFPT are created by comparing the feature nodes of the branches. Let the dataset contains N documents with m signal objects. Each signal object has p features. MFPT is constructed by comparing the features of the signal objects. The time complexity of constructing a MFPT for m signal objects with p features is $O(m(m-1)p)$. The best time complexity is $O(mp)$ when the dataset contains all similar objects. If the dataset contains the k unique words such that $k = m - d$ where d is duplicate words, the time complexity is $O(mkp)$. As p is constant and very small compared to m, it can be ignored.

The MFPT based classification requires comparing the $MFPT_q$ branches with the branches of the $MFPT_T$. Let k be number of leaf nodes of the branches of the $MFPT_T$. Hence the time complexity of MFPT based classification is $O(k)$ assuming the length of test document is very small compared to k.

The MFPT based clustering algorithm requires the collection of leaf nodes for each document. The time complexity of obtaining a leaf node list which gathers all the leaf nodes traversed by an UMD is $O(k)$. Here, the length of branch is neglected as it is very small compared to k. Therefore, the time cost of building leaf node list for N UMDs is $O(Nk)$. In order to cluster the documents, pairwise similarity in terms of leaf node list has to be computed which leads time cost of $O(N^2)$. Thus, the total time complexity for clustering the multimedia documents is $O(Nk) + O(N^2)$.

The SMFPs are extracted by traversing the branches of the MFPT representation towards the leaf nodes. The time complexity for extracting the SMFPs is $O(k)$ with k leaf nodes.

5.4 Experimental Results and Discussion

This section discusses the results obtained by the experiments performed to validate the MFPT representation for the knowledge discovery in MMDs. The

experiments are conducted for four multimodal datasets MDS1, MDS2, Flickr and MDDS. The details of the datasets are discussed in Section 2.9. As a preliminary step, all the four MMD datasets are converted into UMD datasets using the MSC model as discussed in Section 3.2.1. Multimedia classification is evaluated for all four datasets by dividing MMD as training and test datasets in the ratio of 80%-20%. The experiments are performed for object similarity thresholds, $thresh_{ob} = \{0, 1, 3, 5, 7, 10, 13, 15, 20\}$.

Section 5.4.1 discusses the memory requirement for MFPT representation. Section 5.4.2 presents the results of classification of MMDs using MFPT representation. In Section 5.4.3, we discuss the results of MFPT based clustering of MMDs. The MFPT based sequential multimedia feature pattern mining is discussed in Section 5.4.4.

5.4.1 Results of Memory required for MFPT Representation

This section shows how the memory requirement for the representation of MMDs reduces with the MFPT representation. The compactness property of MFPT reduces the memory requirements for storing the multimedia documents with respect to features of the objects. Figure 5.3 shows the comparison of memory requirements for the datasets MDS1, MDS2, Flickr and MDDS with MFPT representation and with original representation. The analysis of the graph shows that the memory requirement of MFPT is less compared to the original representation. The memory reduction ratio for MFPT increases with the increase in the object similarity threshold value from 0% to 20%. The increase in object similarity threshold results in more multimedia objects sharing the same branch of MFPT. The reduction ratio of memory for MDS1 is 0% to 60% , for MDS2 is 29% to 75% , for MDDS is 55% to 82% and for Flickr is 48% to 82%. It is observed that the memory reduction ratio depends on the multimodal objects of the dataset. As the size of the dataset increases the memory reduction rate

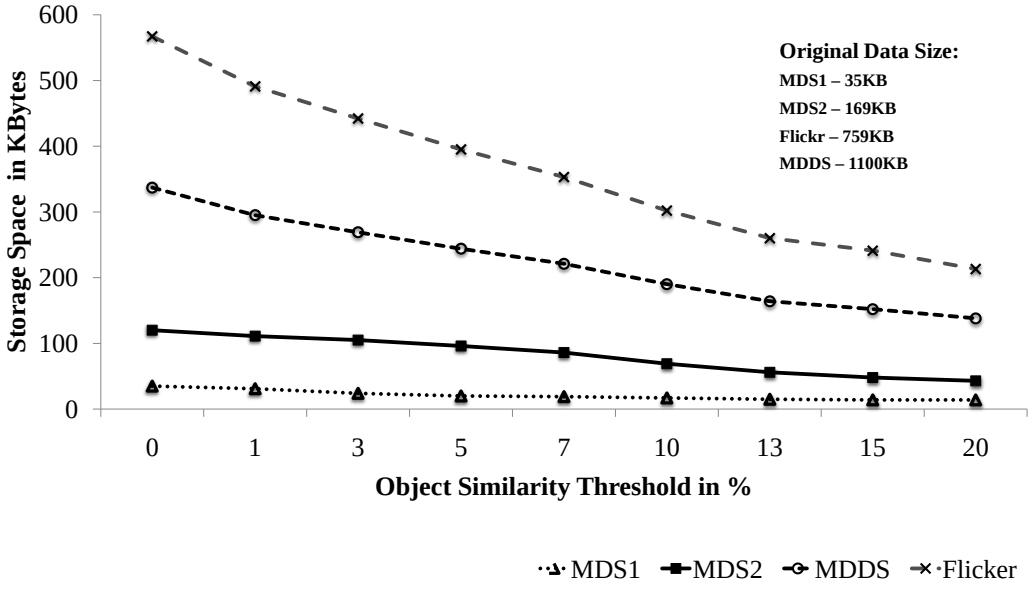


Figure 5.3: Comparison of storage space requirement for MFPT with various object similarity thresholds

also increases. For MDS1 dataset of 35KB the maximum memory reduction is 60% whereas the memory reduction is 82% for Flickr dataset of size 1100KB. This indicates more MMDs will have more similar multimedia objects. The MFPT can made still compact by merging the branches based on similar feature nodes of the branches.

5.4.2 Results of MFPT based Classification for Multimedia Documents

In this section, we discuss the results of classification of UMDs based on MFPT representation. Table 5.1 shows the comparison of the time taken by the MSTD and MFPT representations to represent the UMDs. It is noticed the representation time of MFPT depends on the number of multimedia objects and the similar features of the multimedia objects of the MMDs. The representation time increases with the increase in the number of multimedia objects of the MMDs. It reduces with the increase in similarity of sequential features of the multimedia objects. Also, it is observed that the representation time increases

with increase in modality of the multimedia objects. The dataset MDS1 has taken 0.40 sec for representation as it has less number of MMDs with two multimedia objects of two modality. The dataset MDDS has taken 53.88 sec for representation as it has more number of MMDs with maximum of 198 multimedia objects of three modality. The MMD querying time depends on the number of branches of the MFPT representation. The maximum query time is 15.95 sec for MDDS and minimum time is 0.30 sec for MDS1. The analysis of the results indicate that the time taken by the MFPT representation is significantly lower for all the four datasets compared to MSTD representation. Moreover the time taken by the MFPT for query processing is also improved compared to MSTD representation. As the MFPT representation is based on the individual feature comparison, the computation time for measuring similarity is eliminated. As a result, the representation time and query time is reduced. Compared to MSTD, the MFPT has shown minimum reduction 0.10 sec for MDS1 and maximum reduction of 103.4 sec for MDDS to represent the MMDs. The MFPT has reduced minimum of 0.004 sec for MDS1 and maximum of 1.57 sec for MDDS for query processing compared to MSTD representation. The lower time taken to represent and process the query proves that the MFPT representation reduces search time for knowledge extraction methods.

The classification performance is evaluated by computing the accuracy according to the equation (3.4.3). The accuracy of MFPT based classification for various object similarity threshold values is shown in Fig. 5.4. The maximum accuracy for MDS1 is 0.78, for MDS2 is 0.85, for Flickr is 0.796 and for MDDS is 0.945. The optimal $thresh_{ob}$ value at which maximum accuracy obtained is 10% for MDS1, 7% for MDS2, 5% for Flickr and 5% for MDDS.

The performance of the MFPT based classification is compared with the MSTD based classification and the performance comparison is shown in Fig. 5.5. It is observed that the MFPT based classification outperforms the MSTD based classification by 12% for MDS1, 2% for MDS2, 11% for Flickr, and 2% for MDDS. The MSTD based knowledge extraction methods search for the similar

MMD based on the overall feature similarity of multimedia objects. The objects that has huge difference in any one of the feature may result in lower similarity between the multimedia objects. This issue has been overcome by the MFPT representation as the MFPT based knowledge extraction methods search the similar MMDs based on the number of similar features of multimedia objects. The improvement in accuracy proves that the MFPT based representation can be successfully used for the classification of multimedia documents.

5.4.3 Results of MFPT based Clustering of MMDs

The analysis of clustering quality of clusters is expressed in terms of purity and entropy. The weighted purity and entropy values are calculated as per the equations (3.6.4) and (3.6.5). Figure 5.6 shows the purity end entropy values of the clusters formed using MFPT based clustering algorithm. It is observed that the better purity and entropy values are resulted for the object similarity threshold value between 1 to 13. As the threshold values increases the performance of clustering reduces resulting in lower purity and higher entropy values. The best purity value for MDS1 is 0.99, for MDS2 is 0.95, for Flickr is 0.95 and for MMDS is 1. Similarly the best entropy value is 0.0.003 for MDS1, 0.018 for MDS2, 0.015 for Flickr and 0 for MMDS. The higher values of purity and lower values of entropy indicates that the MFPT representation is effectively used for clustering the MMDS.

Table 5.1: Performance Comparison of time taken by MSTD and MFPT

	Representation Time in secs		Query Time in secs	
Datasets	MSTD	MFPT	MSTD	MFPT
MDS1	0.50	0.40	0.34	0.30
MDS2	11.27	4.40	3.20	2.93
Flickr	90.55	30.31	9.29	8.46
MDDS	157.25	53.88	17.47	15.95

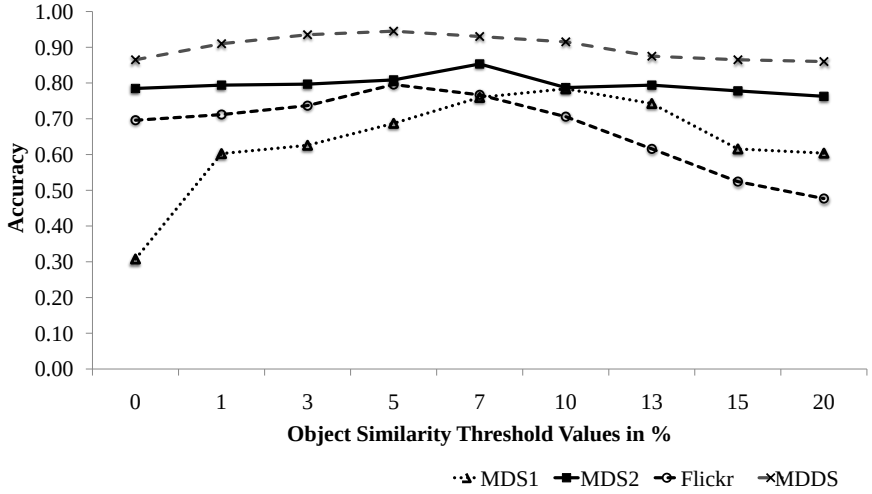


Figure 5.4: MFPT based Classification for Multimedia Documents

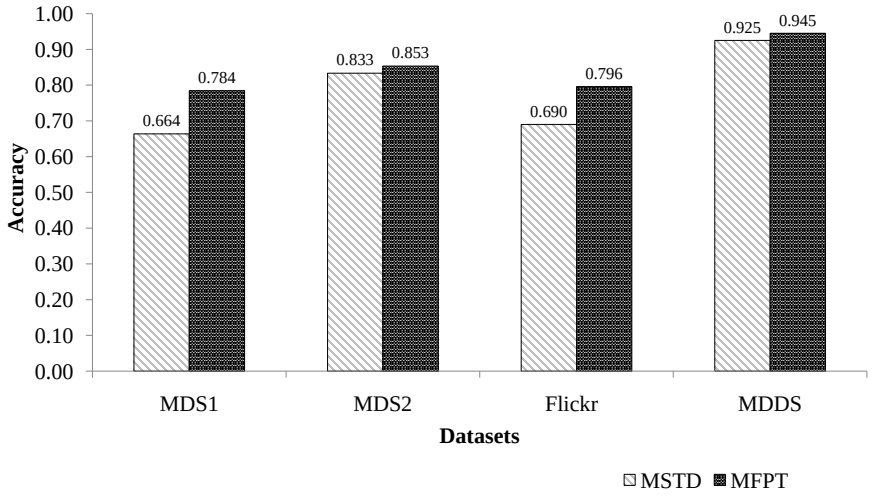
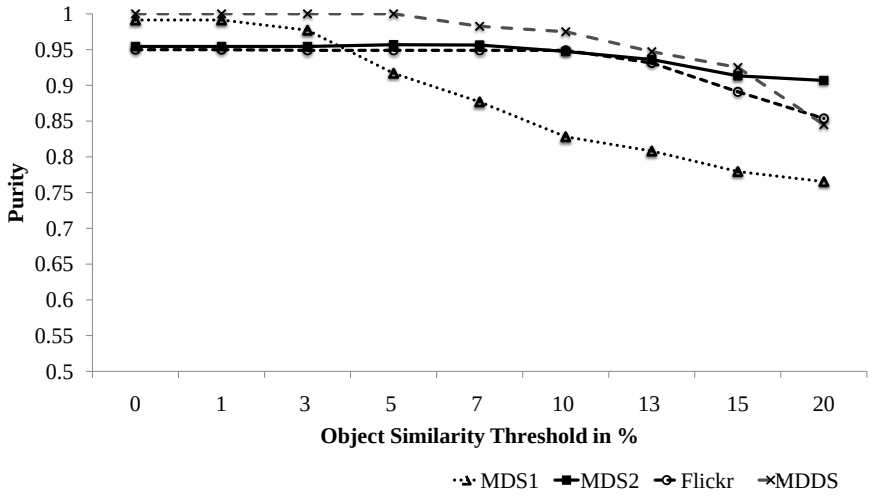


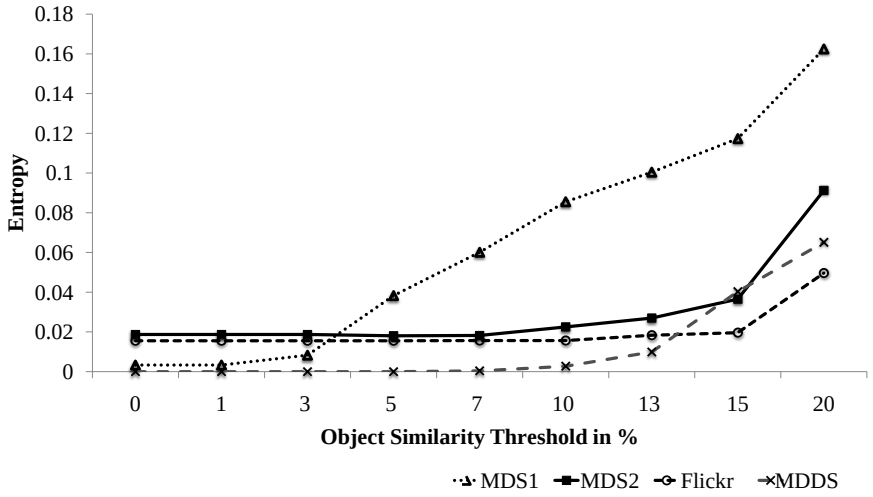
Figure 5.5: Performance Comparison of Classification of MMDs using MSTD and MFPT Representations

The clustering results of MFPT based clustering is compared with MSTD based clustering algorithm. Table 5.2 shows the comparison of MFPT and MSTD based clustering in terms purity-entropy values. It is noted that, the best values of purity and entropy of both the algorithms are almost same for MDS1 and MDDS. The MFPT based clustering achieved the improvement of 4% in purity and 2% in entropy for Flickr dataset. However, the MSTD based clustering achieved the improvement of 2% in purity and 1% in entropy for MDS2. Most of the multimedia documents of MDS2 dataset has single image

with more text tokens. The impact of sequential feature patterns is less for text tokens compared to images and audios. Hence the overall feature similarity based MSTD model performs better for MDS2 compared to sequential feature similarity based MFPT model.



(a) Purity values of the clusters



(b) Entropy values of the clusters

Figure 5.6: Purity and Entropy values for MFPT based Clustering

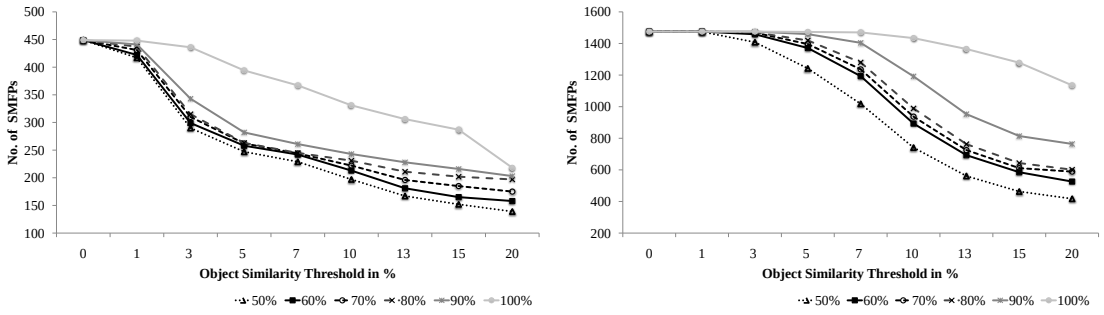
5.4.4 Results of MFPT based Sequential Multimedia Feature Pattern Mining and Multimedia Class Sequence Rules Generation

In this section we discuss the mining of SMFPs and generation of MCSRs. The MFPT represents the multimedia documents based on the sequential feature patterns of the multimedia objects. The branches of the MFPT are mined to extract the SMFPs. The number of SFMPs varies for various length of SMFP. Figure 5.7 demonstrates the number of sequential multimedia feature patterns generated for various length of SMFPs and various object similarity thresholds. For the experiments, the length of SMFPs are selected as 50%, 60%, 70%, 80%, 90% and 100% of total features of the signal object. The number of SMFPs extracted depends on the dataset. The maximum number of SMFPs extracted from MDS1 is 449, MDS2 is 1475, Flickr is 6425 and MDDS is 3869. The minimum number of SMFPs is 139 for MDS1, 417 for MDS2, 1389 for Flickr and 1047 for MDDS. It is seen that the number of SMFP increases with the increase in length of SMFP and decreases with the increase in object similarity threshold.

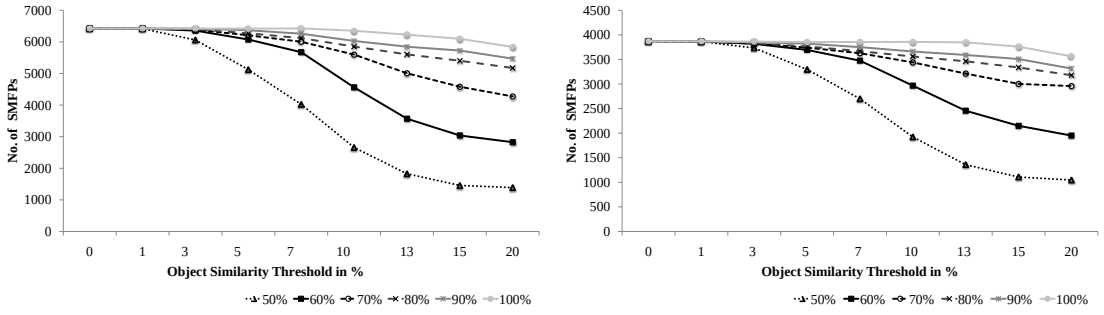
The SMFPs are used to generate the MCSRs that are used for the classification of MMDs. The results of MCSR based classification is shown in fig. 5.8. The experiments has been conducted for four datasets using various length of SMFPs i.e. 50%, 60%, 70%, 80%, 90% and 100% of total features of the signal object. The experiments are performed for object similarity thresholds in %, $thresh_{ob} = \{0, 1, 3, 5, 7, 10, 13, 15, 20\}$. The results demonstrate that the accuracy

Table 5.2: Comparison of MSTD and MFPT based clustering

	Purity		Entropy	
Datasets	MSTD	MFPT	MSTD	MFPT
MDS1	0.991	0.991	0.003	0.003
MDS2	0.982	0.957	0.007	0.018
Flickr	0.906	0.950	0.036	0.016
MDDS	1.000	1.000	0.000	0.000



(a) MDS1 **(b)** MDS2

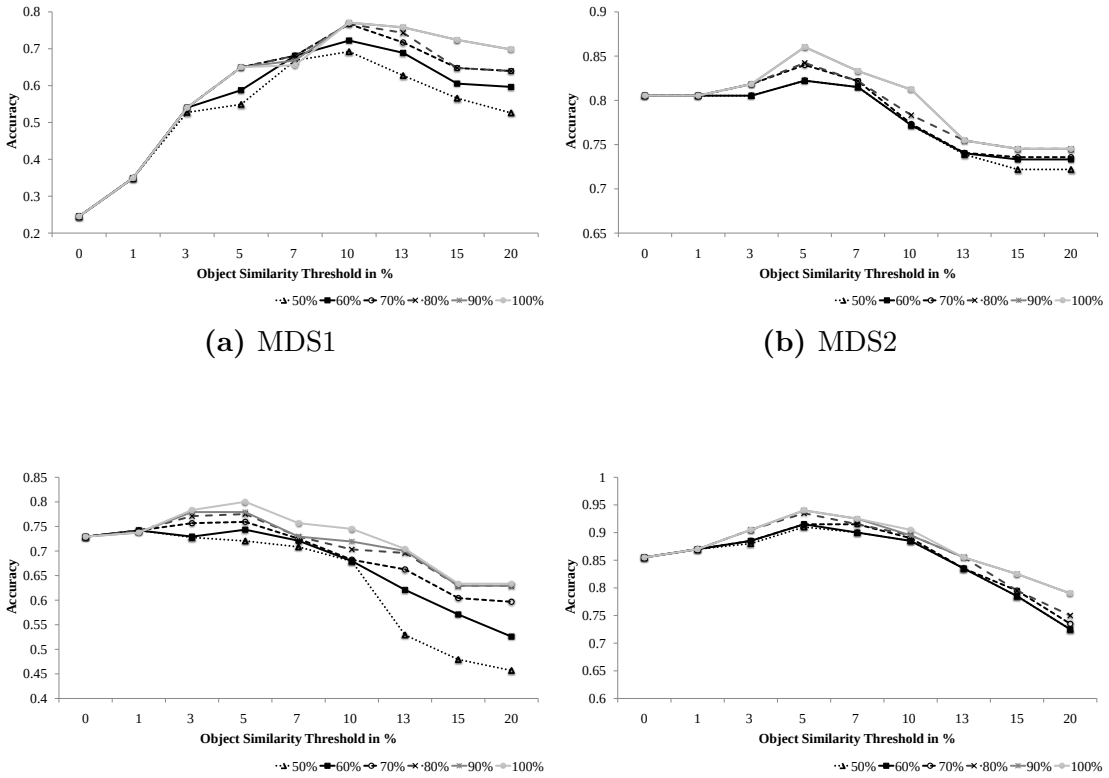


(c) Flickr **(d)** MDDS

Figure 5.7: Number of SMFPs generated for various length of SMFPs

increases with the increase in the length of SMFPs. The best accuracy for MDS1 is 0.771, MDS2 is 0.866, Flickr is 0.805 and for MDDS is 0.925.

The accuracy of MFPT-MCSR based classification is compared with that of MFPT classification and the results are shown in Fig. 5.9. The MFPT-MCSR based classification has been successful in achieving the improvement of 2% for MDS2 and 1% for Flickr. However, it lagged by 2% for MDDS compared to MFPT based classification. The variation in result is due to the use of sequential features of different length by the MFPT-MCSR classification. The MFPT based classification considers all the similar features of the signal objects.



(a) MDS1 **(b)** MDS2

(c) Flickr **(d)** MDDS

Figure 5.8: Accuracy of Multimedia Documents using MFPT-MCSR based Classification for various length of SMFPs

5.4.5 Comparison of the Proposed Multimedia Document Representations With Other Multimodal Retrieval Methods

As per our knowledge there are few representations CCA (Rasiwasia et al., 2010), LRGA (Yang et al., 2009) (Yang et al., 2012), NMF (Caicedo et al., 2012), SMMD (Daras et al., 2012) and UGMDR(Rafailidis et al., 2013) has been proposed for representing the multimedia documents. Among these approaches, the two manifold multimodal earning approaches SMMD and UGMDR have been used for the retrieval of multimedia documents with more than two modalities. The proposed MFPT representation is experimented for the retrieval of multimedia documents and compared with MSTD based multimedia

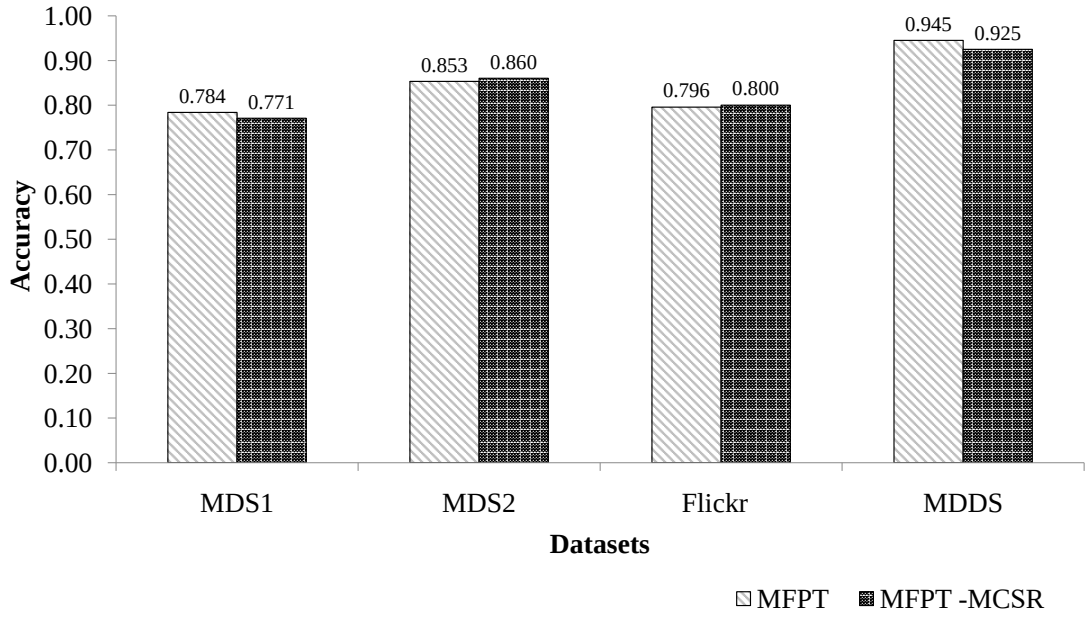


Figure 5.9: Performance Comparison of MFPT and MFPT-MCSR based Classification

document retrieval and SMMD (Daras et al., 2012) and UGMDR (Rafailidis et al., 2013). These approaches have been supported for query and retrieval of multimodal documents. The MSTD and MFPT based retrieval methods are experimented for internal and external queries with two datasets MDS1 and MDS2. For internal queries, each document of the dataset has given as a query to retrieve the most relevant documents. To evaluate the performance for external queries, 10% of MMDs of the dataset are posed as query and the remaining MMDs are used for learning the proposed MSTD and MFPT representation. The individual precision-recall is computed for each query and then the average precision-recall is extracted for each object similarity threshold value. The precision-recall curves are drawn for each dataset by interpolating the average precision values for 11 standard recall values. The performance comparison of the four methods for internal and external queries is presented in Fig. 5.10 and 5.11 respectively.

The results show that the MSTD and MFPT based retrieval outperforms the SMMD and UGMDR methods. The MSTD based approach attained the maximum precision improvement of 10% (MDS1), 14% (MDS2) for internal

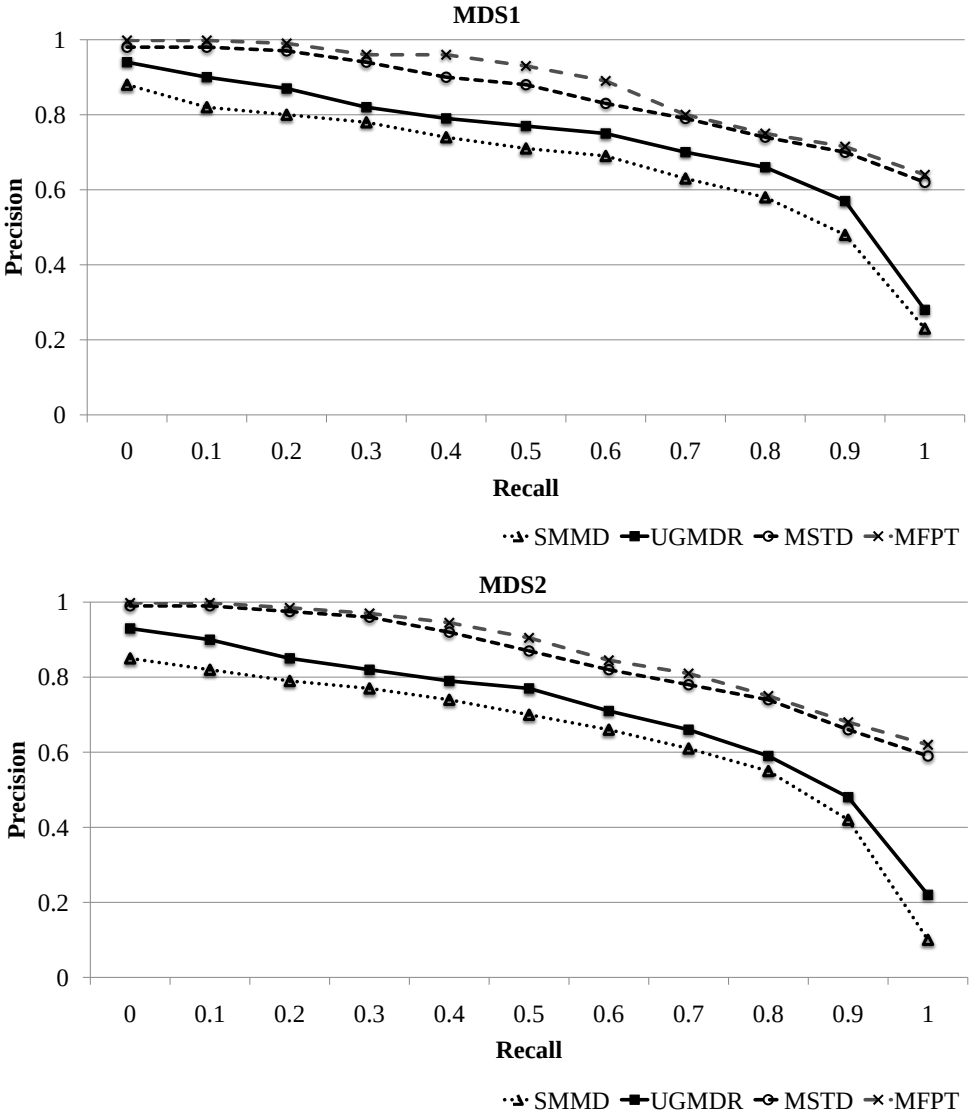


Figure 5.10: Performance Comparison of MFPT based Multimedia Document Retrieval with MSTD, SMMD, and UGMDR Retrieval for Internal Queries

queries and 5% (MDS1), 8% (MDS2) for external queries over the SMMD method. It attained the improvement of 4% (MDS1), 6% (MDS2) for internal queries and 1% (MDS2) for external queries in maximum precision value over the UGMDR method. The MFPT based retrieval achieved the maximum precision improvement of 12% (MDS1), 15% (MDS2) for internal queries and 6% (MDS1), 8% (MDS2) for external queries over the SMMD method. Compared to UGMDR method, it achieved the improvement of 6% (MDS1), 7% (MDS2) for internal

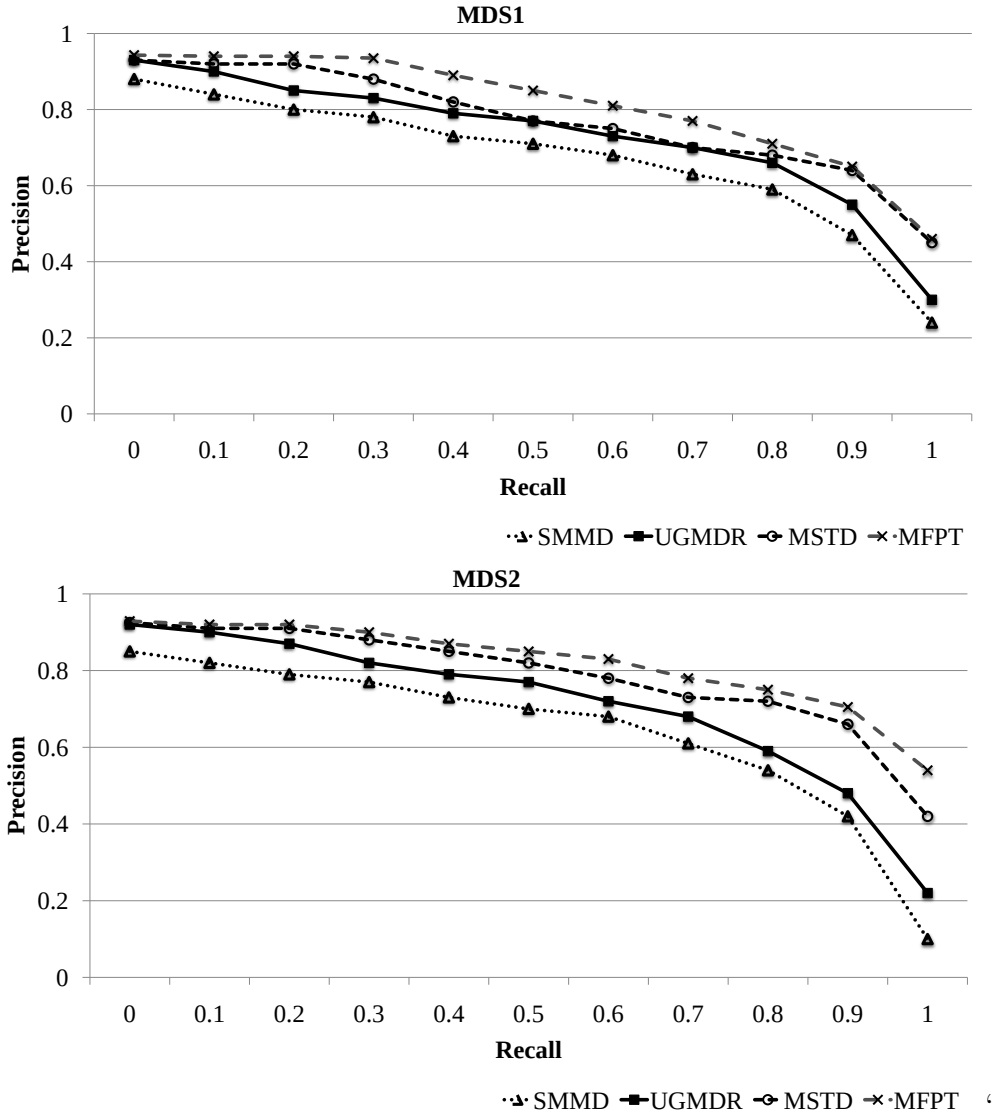


Figure 5.11: Performance Comparison of MFPT based Multimedia Document Retrieval with MSTD, SMMD, and UGMDR Retrieval for External Queries

queries and 1% (MDS1), 1% (MDS2) in maximum precision value for external queries. In order to retrieve the MMDs, the SMMD and UGMDR methods were constructed a multimodal feature space using LE maps based on the unimodal similarity matrix. The similarity matrix is created using the equal number of neighbors for each modality. However, the multimedia objects of different modality may have different number of neighbors. Our method retrieves the similar documents based on the document similarity which is computed based on

the similarity of multimedia objects. Hence, our method achieved better performance compared to SMMD and UGMDR method. It is demonstrated that, in comparison with MSTD based retrieval, the MFPT based retrieval achieved better performance. The MFPT based retrieval obtained the maximum precision improvement of 2% (MDS1), 1% (MDS2) for internal queries and 1% (MDS1) for external queries. The analysis of the result proves that the proposed MFPT representation is efficiently used for the retrieval of MMDs.

5.5 Summary

In this chapter, we discussed the MFPT representation and MFPT based multimedia mining methods. The UMDs are represented using the MFPT representation based on the similar features of the signal objects. MFPT representation reduces the memory space requirement for the storage of UMDs with their features. The effectiveness of the MFPT representation is validated using the MFPT based classification, clustering, sequential multimedia feature pattern mining and multimedia class sequence rules generation approaches. The improvement in time taken for the representation and query processing compared to MSTD representation proves that the MFPT representation effectively represents the MMDs for efficient query processing. The MFPT based classification performed better compared to MSTD based classification. The MFPT based clustering effectively clustered the MMDs by attaining the higher purity and lower entropy values. The branches of the MFPT representation are used to extract the SMFPs. The MCSRs are generated from the SMFPs and used for the classification of the UMDs. Also, the MFPT based retrieval outperformed the manifold learning SMMD and UGMDR methods. The significant performance of the MFPT based knowledge extraction methods prove that the MFPT representation is very effective and beneficial for knowledge extraction from MMDs.

Chapter 6

Conclusion and Future Work

The discovery of useful knowledge from the multimedia documents is beneficial for many applications. The multimodal nature of multimedia objects is the main challenge for multimedia document knowledge discovery methods. Therefore, sophisticated multimedia mining methods are required for the knowledge discovery in multimedia documents. The success of multimedia mining methods rely on the representation of multimedia documents. The multimedia documents representation relies on the representation of multimodal multimedia objects. The existing representations are applicable for text documents. To the best of our knowledge, there no representations has been discussed for the representations of multimedia documents. The appropriate representation of multimedia objects aids the multimedia document representation and multimedia mining methods. Hence, effective multimedia data representations and multimedia document representations are required for the efficient mining of multimedia documents. The research work in this thesis is directed towards the development of effective multimedia data representations, multimedia document representations, and multimedia mining methods for the efficient knowledge discovery in multimedia documents.

We developed two multimedia data representation methods to fulfill the first objective of the research work: Multimedia to Signal conversion (MSC) and Multimedia to Image Conversion (MIC). The MSC method converts the multimedia objects to signal objects to represent them in frequency domain. MIC represents the multimedia objects in spatial domain by converting them as

image objects. The representation of multimedia objects in a unified domain allows the use of same feature extraction methods resulting the features in unified space. Thus, both the multimedia data representation methods convert the multimodal multimedia documents as unified multimedia documents by representing them in a unified feature space. The representation of multimedia objects in a unified space eases the multimedia document representation and multimedia mining methods. The efficacy of the proposed MSC and MIC methods has been evaluated by conducting the experiments for retrieval, classification and clustering of multimedia documents. Experimental results demonstrate that the MSC and MIC based classification has achieved the best classification accuracy of 0.91 for the MDDS dataset which has images, audios and text documents as the contents of the multimedia document. The multimedia document retrieval using MSC and MIC methods outperformed the existing manifold learning multimedia document retrieval method by achieving the maximum precision improvement of 4% for the dataset MDS2. The multimedia documents are clustered using the proposed bio inspired Glowworm Swarm Optimization based Multimedia Documents Clustering (GSOMDC) algorithm. The GSOMDC algorithm clusters the multimedia documents by achieving the best purity value of 99.86 with MSC method and 99.58 with MIC method for the MDDS dataset. The best entropy values of the clusters are 0.0006 with MSC method and 0.0007 with MIC method for the MDDS dataset. The higher purity values and lower entropy values indicate that the GSOMDC algorithm effectively clusters the MMDs with MSC and MIC methods. We proposed an information theory based similarity measure known as ISMD to find the pairwise similarity between the multimedia documents. The ISMD based multimedia documents classification achieved the maximum improvement in accuracy of 14% for MSC and 13% for MIC with the MDS1 dataset over the existing information theory based similarity measure SMTP. The significant performance of multimedia document classification, clustering and retrieval methods prove that the proposed multimedia data representation methods are

effectively used for representing the multimodal multimedia objects in a unified feature space in order to benefit the knowledge discovery in multimedia documents. However, there is a scope for further improvement of the proposed multimedia data representation methods as given below:

1. In order to improve the efficiency of the knowledge extraction methods, the other signal feature extraction methods need to be investigated for signal objects.

2. The methods of image object creation and feature extraction need to be refined as image objects generate high dimensional features that require the computationally expensive processing.

3. The computational efficiency of the data representation methods can be further improved by parallelizing the conversion of each modality of data.

4. The proposed multimedia data representations can be extended to represent the other modality objects such as video and animation in a unified representation.

The time taken by the MIC method for the conversion of multimedia objects to image objects and feature extraction is higher compared to MSC method. Moreover, the high dimensional features of image objects made the processing computationally expensive. Hence, for the further experiments multimedia documents are represented as UMDs using the MSC method. In order to fulfill the second objective, we developed two novel multimedia document representations for the representation of UMDs based on the their similarity. The similarity between the multimedia documents depends on the presence of similar multimedia objects between them. The first representation is Multimedia Suffix Tree Document (MSTD) that represents the UMDs based on the shared similar signal objects among the UMDs. The similarity among the objects is depend on the similarity of the features. We proposed the second representation known as Multimedia Feature Pattern Tree (MFPT) that represents the UMDs based on similar patterns of features of the signal objects. As the MSTD

represents shared similar objects in same branch, it is a compact representation of documents. Compared to MSTD, the MFPT is more compact as the objects having similar sequential features share the same branch. Our experimental results showed that the MFPT representation is achieved the maximum reduction of 82% when the similarity between features of the multimedia document contents is 20% for the MDDS dataset. The main achievement of the MSTD and MFPT representations is they provide the complete information about the represented documents in one structure. Both the representations generate the base clusters of the UMDs in the construction stage itself. MSTD generates the clusters based on the similar objects whereas MFPT generates the clusters of objects based on their similar features that indirectly clusters the documents. In comparison with MSTD representation, the MFPT representation takes significantly less time for the representation of multimedia documents. The MFPT representation takes of 53.88 sec to represent a MDDS dataset whereas the MSTD representation has taken 157.25 secs to represent the same dataset.

The MSTD and MFPT representations provide the platform for the multimedia mining methods to extract the useful patterns from multimedia documents. Hence, to fulfill the third objective of the research work, we proposed MSTD and MFPT based multimedia mining methods to extract the knowledge from the multimedia documents. In comparison with Vector Space Document (VSD) model, the MSTD and MFPT representations take significantly less time to query the MMDs. For a dataset MDS2, the VSD model has taken maximum of 86.09 sec for querying the MMD whereas the MSTD has taken 3.20 sec and MFPT has taken 2.93 sec. Besides, the MSTD representation outperformed the VSD model for the classification of multimedia documents of Flickr dataset with the maximum improvement of 14% in accuracy. The MSTD based clustering efficiently clusters the multimedia documents of MDDS dataset with higher purity value of 1 and lower entropy value of 0. One of the significant characteristics of the MSTD representation is the extraction of frequent multimedia patterns (FMP) exists in the MMDs. The FMPs are used to cluster

the MMDs based on the common FMPs between the MMDs. Also, FMPs are used to generate the multimedia class association (MCAR) rules which are used to classify the MMDs. The MSTD-MCAR based classification achieved maximum of 4% improvement in accuracy over the MSTD based classification for the MDS1 dataset. The MFPT based classification achieved better performance over the MSTD based classification with the maximum improvement of 12% in accuracy for dataset MDS1. In MFPT representation, the objects are searched based on the similarity of their individual features which results in the retrieval of more UMDs. So the accuracy is higher compared to searching the objects based on the overall feature similarity. MFPT based clustering achieved the maximum improvement of 4% in purity and 2% in entropy value for Flickr dataset compared to MSTD based clustering. The higher purity and lower entropy values of MSTD based clustering and MFPT based clustering prove that both the representations are effectively used for clustering the documents. The branches of the MFPT representation represent the sequential multimedia feature patterns (SMFPs) that are used to cluster the multimedia documents based on the shared SMFPs between the MMDs. The SMFPs generate the multimedia class sequential rules(MCSR). As the MCSRs represent the partial sequential features of signal objects, they are best used for classifying the MMDs based on the partial characteristics of the objects. The performance of the MSTD and MFPT based multimedia document retrieval is compared with the existing multimedia document retrieval methods. The MSTD based retrieval achieved the maximum precision improvement of 6% for internal queries and 1% for external queries for MDS2 dataset compared to the manifold learning UGMDR method. The MFPT based retrieval achieved the maximum precision improvement of 7% for internal queries and 1% for external queries for MDS2 dataset compared to UGMDR method. The higher precision values of MSTD and MFPT representations demonstrate the significance of multimedia document representations for the retrieval of multimedia documents. The experimental analysis of MSTD and MFPT based knowledge extraction methods prove that

the proposed multimedia document representations are effectively used to represent the multimedia documents and help in efficient discovery of knowledge from MMDS. However, following issues need to investigated in the future:

1. A methodology can be developed to find the optimal object similarity threshold for each dataset based on the characteristics of the multimedia objects of the dataset for the proposed multimedia document representations.

2. The computational complexity of construction of multimedia document representations can be improved by parallelizing the representation procedure.

3. The proposed representations can be extended for the representation of multimedia documents with more multimedia objects such as video and animation.

4. An algorithm can be developed to find the optimal length of SMFPs such that SMFPs with optimal length define the complete characteristics of the multimedia object it denotes.

To summarize, the thesis proposes the effective multimedia data representations and multimedia document representations for the unified representation of multimodal multimedia documents in order to improve the performance of multimedia mining methods of KDMD process.

References

Adams, W., Iyengar, G., Lin, C.-Y., Naphade, M. R., Neti, C., Nock, H. J., and Smith, J. R. (2003). "Semantic indexing of multimedia content using visual, audio, and text cues". *EURASIP Journal on Applied Signal Processing*, 2:170–185.

Agrawal, R., Imieliński, T., and Swami, A. (1993). "Mining association rules between sets of items in large databases". *ACM SIGMOD Record*, 22(2):207–216.

Agrawal, R. and Srikant, R. (1995). "Mining sequential patterns". In *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*, pages 3–14. IEEE.

Akakin, H. C. and Gurcan, M. N. (2012). "Content-based microscopic image retrieval system for multi-image queries". *Information Technology in Biomedicine, IEEE Transactions on*, 16(4):758–769.

Alghamdi, R. A., Taileb, M., and Ameen, M. (2014). "A new multimodal fusion method based on association rules mining for image retrieval". In *Mediterranean Electrotechnical Conference (MELECON), 2014 17th IEEE*, pages 493–499. IEEE.

Aljarah, I. and Ludwig, S. (2013). "A new clustering approach based on glowworm swarm optimization". In *Evolutionary Computation (CEC), 2013 IEEE Congress on*, pages 2642–2649. IEEE.

Ananthanarayana, V., Murty, M. N., and Subramanian, D. (2003). "Tree structure for efficient data mining using rough sets". *Pattern Recognition Letters*, 24(6):851–862.

Andrew, G., Arora, R., Bilmes, J., and Livescu, K. (2013). "Deep canonical correlation analysis". In *Proceedings of the 30th International Conference on Machine Learning*, pages 1247–1255.

Anguera, X., Barrios, J. M., Adamek, T., and Oliver, N. (2011). "Multimodal fusion for video copy detection". In *Proceedings of the 19th ACM international conference on Multimedia*, pages 1221–1224. ACM.

Arevalillo-Herráez, M., Domingo, J., and Ferri, F. J. (2008). "Combining similarity measures in content-based image retrieval". *Pattern Recognition Letters*, 29(16):2174–2181.

Aslam, J. A. and Frost, M. (2003). "An information-theoretic measure for document similarity". In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 449–450. ACM.

Atrey, P. K., Hossain, M. A., El Saddik, A., and Kankanhalli, M. S. (2010). "Multimodal fusion for multimedia analysis: a survey". *Multimedia systems*, 16(6):345–379.

Baecchi, C., Uricchio, T., Bertini, M., and Del Bimbo, A. (2016). "A multimodal feature learning approach for sentiment analysis of social network multimedia". *Multimedia Tools and Applications*, 75(5):2507–2525.

Barnard, K., Duygulu, P., Forsyth, D., De Freitas, N., Blei, D. M., and Jordan, M. I. (2003). "Matching words and pictures". *The Journal of Machine Learning Research*, 3:1107–1135.

Beal, M. J., Jojic, N., and Attias, H. (2003). "A graphical model for audiovisual object tracking". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(7):828–836.

Bekkerman, R. and Jeon, J. (2007). "Multi-modal clustering for multimedia collections". In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE.

Bultheel, A. and Huybrechs, D. (2011). "Wavelets with applications in signal and image processing".

Cai, D., He, X., Li, Z., Ma, W.-Y., and Wen, J.-R. (2004). "Hierarchical clustering of WWW image search results using visual, textual and link information". In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 952–959. ACM.

Caicedo, J. C., BenAbdallah, J., González, F. A., and Nasraoui, O. (2012). "Multimodal representation, indexing, automated annotation and retrieval of image collections via non-negative matrix factorization". *Neurocomputing*, 76(1):50–60.

Caicedo, J. C. and González, F. A. (2012). "Multimodal fusion for image retrieval using matrix factorization". In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, page 56. ACM.

Cazan, A., Vârbanescu, R., and Popescu, D. (2007). "Algorithms and Techniques for Image to Sound Conversion for Helping the Visually Impaired People-Application Proposal". In *Systems, Signals and Image Processing, 2007 and 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services. 14th International Workshop on*, pages 471–474. IEEE.

Chaisorn, L., Chua, T.-S., Koh, C.-K., Zhao, Y., Xu, H., Feng, H., and Tian, Q. (2003). "A two-level multi-modal approach for story segmentation of large news video corpus". In *TRECVID Conference,(Gaithersburg, Washington DC, November 2003). Published on-line at http://wwwnlpir. nist. gov/projects/tvpubs/tv. pubs. org. html.*

Chang, S.-F., Ellis, D., Jiang, W., Lee, K., Yanagawa, A., Loui, A. C., and Luo, J. (2007). "Large-scale multimodal semantic concept detection for consumer video". In *Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 255–264. ACM.

Chaudhuri, K., Kakade, S. M., Livescu, K., and Sridharan, K. (2009). "Multi-view clustering via canonical correlation analysis". In *Proceedings of the 26th annual international conference on machine learning*, pages 129–136. ACM.

Chen, M., Chen, S.-C., and Shyu, M.-L. (2007). "Hierarchical temporal association mining for video event detection in video databases". In *Data Engineering Workshop, 2007 IEEE 23rd International Conference On*, pages 137–145. IEEE.

Chen, M., Chen, S.-C., Shyu, M.-L., and Wickramaratna, K. (2006). "Semantic event detection via multimodal data mining". *Signal Processing Magazine, IEEE*, 23(2):38–46.

Chen, S.-C., Shyu, M.-L., Chen, M., and Zhang, C. (2004). "A decision tree-based multimodal data mining framework for soccer goal detection". In *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on*, volume 1, pages 265–268. IEEE.

Chen, X., Hero III, A. O., and Savarese, S. (2012). "Multimodal video indexing and retrieval using directed information". *Multimedia, IEEE Transactions on*, 14(1):3–16.

Chen, Y.-L. and Chiu, Y.-T. (2011). "An IPC-based vector space model for patent retrieval". *Information Processing & Management*, 47(3):309–322.

Chim, H. and Deng, X. (2008). "Efficient phrase-based document similarity for clustering". *Knowledge and Data Engineering, IEEE Transactions on*, 20(9):1217–1229.

Chu, S., Narayanan, S., and Kuo, C. J. (2009). "Environmental sound recognition with time–frequency audio features". *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(6):1142–1158.

Costa Pereira, J., Coviello, E., Doyle, G., Rasiwasia, N., Lanckriet, G. R., Levy, R., and Vasconcelos, N. (2014). "On the role of correlation and abstraction in cross-modal multimedia retrieval". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(3):521–535.

Cui, B., Tung, A. K., Zhang, C., and Zhao, Z. (2010). "Multiple feature fusion for social media applications". In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 435–446. ACM.

Cutler, R. and Davis, L. (2000). "Look who's talking: Speaker detection using video and audio correlation". In *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, volume 3, pages 1589–1592. IEEE.

Daras, P., Manolopoulou, S., and Axenopoulos, A. (2012). "Search and retrieval of rich media objects supporting multiple multimodal queries". *Multimedia, IEEE Transactions on*, 14(3):734–746.

Davis, L. S. (1975). "A survey of edge detection techniques". *Computer graphics and image processing*, 4(3):248–270.

Davis, S. and Mermelstein, P. (1980). "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences". *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366.

Dumais, S. T. (1991). "Improving the retrieval of information from external sources". *Behavior Research Methods, Instruments, & Computers*, 23(2):229–236.

Fei-Fei, L., Fergus, R., and Perona, P. (2007). "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories". *Computer Vision and Image Understanding*, 106(1):59–70.

Fisher, B., Perkins, S., Walker, A., and Wolfart, E. (1996). "Hypermedia image processing reference". Wiley Chichester, UK.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., et al. (2013). "Devise: A deep visual-semantic embedding model". In *Advances in Neural Information Processing Systems*, pages 2121–2129.

Gao, B., Liu, T.-Y., Qin, T., Zheng, X., Cheng, Q.-S., and Ma, W.-Y. (2005). "Web image clustering by consistent utilization of visual features and surrounding texts". In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 112–121. ACM.

Goh, K.-S., Miyahara, K., Radhakrishnan, R., Xiong, Z., and Divakaran, A. (2003). "Audio-visual event detection based on mining of semantic audio-visual labels". In *Electronic Imaging 2004*, pages 292–299. International Society for Optics and Photonics.

Goldberger, J., Gordon, S., and Greenspan, H. (2003). "An efficient image similarity measure based on approximations of KL-divergence between two Gaussian mixtures". In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 487–493. IEEE.

Grigorescu, S. E., Petkov, N., and Kruizinga, P. (2002). "Comparison of texture features based on Gabor filters". *Image Processing, IEEE Transactions on*, 11(10):1160–1167.

Grigorova, A., De Natale, F. G., Dagli, C., and Huang, T. S. (2007). "Content-based image retrieval by feature adaptation and relevance feedback". *Multimedia, IEEE Transactions on*, 9(6):1183–1192.

Guo, G.-D., Jain, A. K., Ma, W.-Y., and Zhang, H.-J. (2002). "Learning similarity measure for natural image retrieval with relevance feedback". *IEEE Transactions on Neural Networks*, 13(4):811–820.

Han, J., Kamber, M., and Pei, J. (2011). "Data mining: concepts and techniques". Elsevier.

He, R., Xiong, N., Yang, L. T., and Park, J. H. (2011). "Using multi-modal semantic association rules to fuse keywords and visual features automatically for web image retrieval". *Information Fusion*, 12(3):223–230.

He, Y., Xiang, S., Kang, C., Wang, J., and Pan, C. (2016). "Cross-Modal Retrieval via Deep and Bidirectional Representation Learning". *IEEE Transactions on Multimedia*, 18(7):1363–1377.

Helén, M. and Virtanen, T. (2009). "Audio query by example using similarity measures between probability density functions of features". *EURASIP Journal on Audio, Speech, and Music Processing*, 2010(1):1.

Hershey, J., Attias, H., Jojic, N., and Kristjansson, T. (2004). "Audio-visual graphical models for speech processing". In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 5, pages V–649. IEEE.

Hoi, S. C. and Lyu, M. R. (2008). "A multimodal and multilevel ranking scheme for large-scale video retrieval". *Multimedia, IEEE Transactions on*, 10(4):607–619.

Howarth, P. and Rüger, S. (2004). "Evaluation of texture features for content-based image retrieval". In *Image and Video Retrieval*, pages 326–334. Springer.

Huang, A. (2008). "Similarity measures for text document clustering". In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, pages 49–56.

Huang, J., Kumar, S. R., Mitra, M., Zhu, W.-J., and Zabih, R. (1997). "Image indexing using color correlograms". In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 762–768. IEEE.

Jayakumar, D. N. and Venkatesh, P. (2014). "Glowworm swarm optimization algorithm with topsis for solving multiple objective environmental economic dispatch problem". *Applied Soft Computing*, 23:375–386.

Jensen, J. H., Christensen, M. G., Ellis, D. P., and Jensen, S. H. (2009). "Quantitative analysis of a common audio similarity measure". *IEEE Transactions on Audio, Speech, and Language Processing*, 17(4):693–703.

Jiang, T. and Tan, A.-H. (2009). "Learning image-text associations". *Knowledge and Data Engineering, IEEE Transactions on*, 21(2):161–177.

Jiang, W., Cotton, C., Chang, S.-F., Ellis, D., and Loui, A. (2009). "Short-term audio-visual atoms for generic video concept classification". In *Proceedings of the 17th ACM international conference on Multimedia*, pages 5–14. ACM.

Jiang, W. and Loui, A. C. (2011). "Audio-visual grouplet: temporal audio-visual interactions for general video concept classification". In *Proceedings of the 19th ACM international conference on Multimedia*, pages 123–132. ACM.

Kanluan, I., Grimm, M., and Kroschel, K. (2008). "Audio-visual emotion recognition using an emotion space concept". In *Signal Processing Conference, 2008 16th European*, pages 1–5. IEEE.

Katsamanis, A., Papandreou, G., Pitsikalis, V., and Maragos, P. (2006). "Multimodal fusion by adaptive compensation for feature uncertainty with application to audiovisual speech recognition". In *Signal Processing Conference, 2006 14th European*, pages 1–5. IEEE.

Krishnanand, K. and Ghose, D. (2006). "Glowworm swarm based optimization algorithm for multimodal functions with collective robotics applications". *Multiagent and Grid Systems*, 2(3):209–222.

Krishnanand, K. and Ghose, D. (2009). "Glowworm swarm optimization for simultaneous capture of multiple local optima of multimodal functions". *Swarm intelligence*, 3(2):87–124.

Kwitt, R. and Uhl, A. (2008). "Image similarity measurement by Kullback-Leibler divergences between complex wavelet subband statistics for texture retrieval". In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 933–936. IEEE.

Lan, Z.-z., Bao, L., Yu, S.-I., Liu, W., and Hauptmann, A. G. (2014). "Multimedia classification and event detection using double fusion". *Multimedia Tools and Applications*, 71(1):333–347.

Laurier, C., Grivolla, J., and Herrera, P. (2008). "Multimodal music mood classification using audio and lyrics". In *Machine Learning and Applications, 2008. ICMLA'08. Seventh International Conference on*, pages 688–693. IEEE.

Lazaridis, M., Axenopoulos, A., Rafailidis, D., and Daras, P. (2013). "Multimedia search and retrieval using multimodal annotation propagation and indexing techniques". *Signal Processing: Image Communication*, 28(4):351–367.

Lee, J.-S. and Park, C. H. (2008). "Adaptive decision fusion for audio-visual speech recognition". INTECH Open Access Publisher.

Levy, M. and Sandler, M. (2009). "Music information retrieval using social tags and audio". *Multimedia, IEEE Transactions on*, 11(3):383–395.

Li, B., Godil, A., Aono, M., Bai, X., Furuya, T., Li, L., López-Sastre, R. J., Johan, H., Ohbuchi, R., Redondo-Cabrera, C., et al. "SHREC'12 Track: Generic 3D Shape Retrieval."

Li, D., Dimitrova, N., Li, M., and Sethi, I. K. (2003). "Multimedia content processing through cross-modal association". In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 604–611. ACM.

Li, H., Ma, B., and Lee, C.-H. (2007). "A vector space modeling approach to spoken language identification". *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(1):271–284.

Li, H., Ma, B., and Lee, K. A. (2013). "Spoken language recognition: from fundamentals to practice". *Proceedings of the IEEE*, 101(5):1136–1159.

Li, Y., Chung, S. M., and Holt, J. D. (2008). "Text document clustering based on frequent word meaning sequences". *Data & Knowledge Engineering*, 64(1):381–404.

Liao, W.-H., Kao, Y., and Li, Y.-S. (2011). "A sensor deployment approach using glowworm swarm optimization algorithm in wireless sensor networks". *Expert Systems with Applications*, 38(10):12180–12188.

Lin, D. (1998). "An information-theoretic definition of similarity." In *ICML*, volume 98, pages 296–304.

Lin, M.-Y., Hsueh, S.-C., Chen, M.-H., and Hsu, H.-Y. (2009). "Mining sequential patterns for image classification in ubiquitous multimedia systems". In *Intelligent Information Hiding and Multimedia Signal Processing, 2009. IIH-MSP'09. Fifth International Conference on*, pages 303–306. IEEE.

Lin, W.-H. and Hauptmann, A. (2002). "News video classification using SVM-based multimodal classifiers and combination strategies". In *Proceedings of the tenth ACM international conference on Multimedia*, pages 323–326. ACM.

Lin, Y.-S., Jiang, J.-Y., and Lee, S.-J. (2014). "A similarity measure for text classification and clustering". *Knowledge and Data Engineering, IEEE Transactions on*, 26(7):1575–1590.

Liu, B. (2007). "Web data mining: exploring hyperlinks, contents, and usage data". Springer Science & Business Media.

Liu, Y., Feng, X., and Zhou, Z. (2016). "Multimodal video classification with stacked contractive autoencoders". *Signal Processing*, 120:761–766.

Lo, Y.-L., Lee, W.-L., and Chang, L.-h. (2008). "True suffix tree approach for discovering non-trivial repeating patterns in a music object". *Multimedia Tools and Applications*, 37(2):169–187.

Lu, G. (2001). "Indexing and retrieval of audio: A survey". *Multimedia Tools and Applications*, 15(3):269–290.

Ma, B. L. W. H. Y. (1998). "Integrating classification and association rule mining". In *Proceedings of the fourth international conference on knowledge discovery and data mining*.

Maddage, N. C., Li, H., and Kankanhalli, M. S. (2006). "Music structure based vector space retrieval". In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 67–74. ACM.

Mallat, S. (2008). "A wavelet tour of signal processing: the sparse way". Academic press.

Mallat, S. G. (1989). "A theory for multiresolution signal decomposition: the wavelet representation". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 11(7):674–693.

Manning, C. D., Raghavan, P., Schütze, H., et al. (2008). "Introduction to information retrieval", volume 1. Cambridge university press Cambridge.

Mansoorizadeh, M. and Charkari, N. M. (2010). "Multimodal information fusion application to human emotion recognition from face and speech". *Multimedia Tools and Applications*, 49(2):277–297.

Mao, W. and Chu, W. W. (2007). "The phrase-based vector space model for automatic retrieval of free-text medical documents". *Data & Knowledge Engineering*, 61(1):76–92.

Mao, X., Lin, B., Cai, D., He, X., and Pei, J. (2013). "Parallel field alignment for cross media retrieval". In *Proceedings of the 21st ACM international conference on multimedia*, pages 897–906. ACM.

Martinet, J., Chiaramella, Y., and Mulhem, P. (2011). "A relational vector space model using an advanced weighting scheme for image retrieval". *Information processing & management*, 47(3):391–414.

Meng, L., Tan, A.-H., and Xu, D. (2014). "Semi-supervised heterogeneous fusion for multimedia data co-clustering". *Knowledge and Data Engineering, IEEE Transactions on*, 26(9):2293–2306.

Mitrović, D., Zeppelzauer, M., and Breiteneder, C. (2010). "Features for content-based audio retrieval". *Advances in computers*, 78:71–150.

Monay, F. and Gatica-Perez, D. (2007). "Modeling semantic aspects for cross-media image indexing". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(10):1802–1817.

Mourão, A., Martins, F., and Magalhães, J. (2015). "Multimodal medical information retrieval with unsupervised rank fusion". *Computerized Medical Imaging and Graphics*, 39:35–45.

Muneesawang, P., Guan, L., and Amin, T. (2010). "A New Learning Algorithm for the Fusion of Adaptive Audio-Visual Features for the Retrieval and Classification of Movie Clips". *Journal of Signal Processing Systems*, 59(2):177–188.

Natarajan, P., Wu, S., Vitaladevuni, S., Zhuang, X., Tsakalidis, S., Park, U., Prasad, R., and Natarajan, P. (2012). "Multimodal feature fusion for robust event detection in web videos". In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1298–1305. IEEE.

Nefian, A. V., Liang, L., Pi, X., Liu, X., and Murphy, K. (2002). "Dynamic Bayesian networks for audio-visual speech recognition". *EURASIP Journal on Applied Signal Processing*, 2002(1):1274–1288.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). "Multimodal deep learning". In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696.

Nock, H. J., Iyengar, G., and Neti, C. (2003). "Speaker localisation using audio-visual synchrony: An empirical study". In *Image and Video Retrieval*, pages 488–499. Springer.

Pan, J.-Y., Yang, H.-J., Faloutsos, C., and Duygulu, P. (2004). "Automatic multimedia cross-modal correlation discovery". In *Proceedings of the tenth ACM*

SIGKDD international conference on Knowledge discovery and data mining, pages 653–658. ACM.

Pang, L., Zhu, S., and Ngo, C.-W. (2015). "Deep multimodal learning for affective analysis and retrieval". *IEEE Transactions on Multimedia*, 17(11):2008–2020.

Papandreou, G., Katsamanis, A., Pitsikalis, V., and Maragos, P. (2007). "Multimodal fusion and learning with uncertain features applied to audiovisual speech recognition". In *Multimedia Signal Processing, 2007. MMSP 2007. IEEE 9th Workshop on*, pages 264–267. IEEE.

Papandreou, G., Katsamanis, A., Pitsikalis, V., and Maragos, P. (2009). "Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition". *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(3):423–435.

Park, B. G., Lee, K. M., and Lee, S. U. (2007). "Color-based image retrieval using perceptually modified Hausdorff distance". *EURASIP Journal on Image and Video Processing*, 2008(1):1.

Pass, G., Zabih, R., and Miller, J. (1997). "Comparing images using color coherence vectors". In *Proceedings of the fourth ACM international conference on Multimedia*, pages 65–73. ACM.

Petkos, G., Papadopoulos, S., and Kompatsiaris, Y. (2012). "Social event detection using multimodal clustering and integrating supervisory signals". In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, page 23. ACM.

Petkos, G., Papadopoulos, S., Schinas, E., and Kompatsiaris, Y. (2014). "Graph-based multimodal clustering for social event detection in large collections of images". In *MultiMedia Modeling*, pages 146–158. Springer.

Pfeiffer, S., Fischer, S., and Effelsberg, W. (1997). "Automatic audio content analysis". In *Proceedings of the fourth ACM international conference on Multimedia*, pages 21–30. ACM.

Porter, M. F. (1980). "An algorithm for suffix stripping". *Program*, 14(3):130–137.

Rafailidis, D., Manolopoulou, S., and Daras, P. (2013). "A unified framework for multimodal retrieval". *Pattern Recognition*, 46(12):3358–3370.

Ramachandran, C., Malik, R., Jin, X., Gao, J., Nahrstedt, K., and Han, J. (2009). "Videomule: a consensus learning approach to multi-label classification from noisy user-generated videos". In *Proceedings of the 17th ACM international conference on Multimedia*, pages 721–724. ACM.

Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G. R., Levy, R., and Vasconcelos, N. (2010). "A new approach to cross-modal multimedia retrieval". In *Proceedings of the 18th ACM international conference on Multimedia*, pages 251–260. ACM.

Ren, J.-M. and Jang, J.-S. R. (2012). "Discovering time-constrained sequential patterns for music genre classification". *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1134–1144.

Reuderink, B., Poel, M., Truong, K., Poppe, R., and Pantic, M. (2008). "Decision-level fusion for audio-visual laughter detection". Springer.

Ribeiro, M. X., Traina, C., and Azevedo-Marques, P. M. (2008). "An association rule-based method to support medical image diagnosis with efficiency". *Multimedia, IEEE Transactions on*, 10(2):277–285.

Ruocco, M. and Ramampiaro, H. (2010). "Event clusters detection on flickr images using a suffix-tree structure". In *Multimedia (ISM), 2010 IEEE International Symposium on*, pages 41–48. IEEE.

Salton, G., Wong, A., and Yang, C.-S. (1975). "A vector space model for automatic indexing". *Communications of the ACM*, 18(11):613–620.

Santos, I., Laorden, C., Sanz, B., and Bringas, P. G. (2012). "Enhanced topic-based vector space model for semantics-aware spam filtering". *Expert Systems with applications*, 39(1):437–444.

Santosa, R. A. and Bao, P. (2005). "Audio-to-image wavelet transform based audio steganography". In *ELMAR, 2005. 47th International Symposium*, pages 209–212. IEEE.

Saraceno, C. and Leonardi, R. (1997). "Audio as a support to scene change detection and characterization of video sequences". In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 4, pages 2597–2600. IEEE.

Sargin, M. E., Yemez, Y., Erzin, E., and Tekalp, A. M. (2007). "Audiovisual synchronization and fusion using canonical correlation analysis". *Multimedia, IEEE Transactions on*, 9(7):1396–1403.

Saunders, J. (1996). "Real-time discrimination of broadcast speech/music". In *icassp*, pages 993–996. IEEE.

Schnitzer, D., Flexer, A., and Widmer, G. (2012). "A fast audio similarity retrieval method for millions of music tracks". *Multimedia Tools and Applications*, 58(1):23–40.

Schoenharl, T. W. and Madey, G. (2008). "Evaluation of measurement techniques for the validation of agent-based simulations against streaming data". In *Computational Science–ICCS 2008*, pages 6–15. Springer.

Sevillano, X. and Alías, F. (2014). "A one-shot domain-independent robust multimedia clustering methodology based on hybrid multimodal fusion". *Multimedia Tools and Applications*, 73(3):1507–1543.

Seyerlehner, K., Widmer, G., and Pohle, T. (2010). "Fusing block-level features for music similarity estimation". In *Proc. of the 13th Int. Conference on Digital Audio Effects (DAFx-10)*, pages 225–232.

Shilane, P., Min, P., Kazhdan, M., and Funkhouser, T. (2004). "The princeton shape benchmark". In *Shape modeling applications, 2004. Proceedings*, pages 167–178. IEEE.

Shyu, M.-L., Xie, Z., Chen, M., and Chen, S.-C. (2008). "Video semantic event/concept detection using a subspace-based multimedia data mining framework". *Multimedia, IEEE Transactions on*, 10(2):252–259.

Singha, M., Hemachandran, K., and Paul, A. (2012). "Content-based image retrieval using the combination of the fast wavelet transformation and the colour histogram". *Image Processing, IET*, 6(9):1221–1226.

Snoek, C. G., Worring, M., and Smeulders, A. W. (2005). "Early versus late fusion in semantic video analysis". In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402. ACM.

Srivastava, N. and Salakhutdinov, R. R. (2012). "Multimodal learning with deep boltzmann machines". In *Advances in neural information processing systems*, pages 2222–2230.

Stockman, G. and Shapiro, L. G. (2001). "Computer Vision". Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition.

Stricker, M. A. and Orengo, M. (1995). "Similarity of color images". In *IS&T/SPIE's Symposium on Electronic Imaging: Science & Technology*, pages 381–392. International Society for Optics and Photonics.

Suzuki, Y., Mitsukawa, M., and Kawagoe, K. (2008). "A image retrieval method using tfidf based weighting scheme". In *Database and Expert Systems Application, 2008. DEXA'08. 19th International Workshop on*, pages 112–116. IEEE.

Taylor, P. (2009). "Text-to-speech synthesis". Cambridge university press.

Tsatsaronis, G. and Panagiotopoulou, V. (2009). "A generalized vector space model for text retrieval based on semantic relatedness". In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 70–78. Association for Computational Linguistics.

Tuceryan, M., Jain, A. K., et al. (1993). "Texture analysis". *Handbook of pattern recognition and computer vision*, 2:207–248.

Ukkonen, E. (1995). "On-line construction of suffix trees". *Algorithmica*, 14(3):249–260.

Wagner, J., Andre, E., Lingenfelser, F., and Kim, J. (2011). "Exploring fusion methods for multimodal emotion recognition with missing data". *Affective Computing, IEEE Transactions on*, 2(4):206–218.

Wang, K., Tang, J., Wang, N., and Shao, L. (2016a). "Semantic boosting cross-modal hashing for efficient multimedia retrieval". *Information Sciences*, 330:199–210.

Wang, S., Joo, J., Wang, Y., and Zhu, S.-C. (2013). "Weakly supervised learning for attribute localization in outdoor scenes". In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3111–3118. IEEE.

Wang, W., Ooi, B. C., Yang, X., Zhang, D., and Zhuang, Y. (2014). "Effective multi-modal retrieval based on stacked auto-encoders". *Proceedings of the VLDB Endowment*, 7(8):649–660.

Wang, W., Yang, X., Ooi, B. C., Zhang, D., and Zhuang, Y. (2016b). "Effective deep learning-based multi-modal retrieval". *The VLDB Journal*, 25(1):79–101.

Wang, X. and Kankanhalli, M. (2010a). "MultiFusion: a boosting approach for multimedia fusion". *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 6(4):25.

Wang, X. and Kankanhalli, M. (2010b). "Portfolio theory of multimedia fusion". In *Proceedings of the 18th ACM international conference on Multimedia*, pages 723–726. ACM.

Wang, Y., Liu, Z., and Huang, J.-C. (2000). "Multimedia content analysis-using both audio and visual clues". *Signal Processing Magazine, IEEE*, 17(6):12–36.

Wei, S., Zhao, Y., Zhu, Z., and Liu, N. (2010). "Multimodal fusion for video search reranking". *Knowledge and Data Engineering, IEEE Transactions on*, 22(8):1191–1199.

Worawitphinyo, P., Gao, X., and Jabeen, S. (2011). "Improving suffix tree clustering with new ranking and similarity measures". In *Advanced Data Mining and Applications*, pages 55–68. Springer.

Wu, F., Lu, X., Song, J., Yan, S., Zhang, Z. M., Rui, Y., and Zhuang, Y. (2016). "Learning of multimodal representations with random walks on the click graph". *IEEE Transactions on Image Processing*, 25(2):630–642.

Wu, F., Zhang, H., and Zhuang, Y. (2006). "Learning semantic correlations for cross-media retrieval". In *Image Processing, 2006 IEEE International Conference on*, pages 1465–1468. IEEE.

Wu, P., Hoi, S. C., Xia, H., Zhao, P., Wang, D., and Miao, C. (2013). "Online multimodal deep similarity learning with application to image retrieval". In *Proceedings of the 21st ACM international conference on Multimedia*, pages 153–162. ACM.

Wu, Y., Chang, E. Y., Chang, K. C.-C., and Smith, J. R. (2004a). "Optimal multimodal fusion for multimedia data analysis". In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 572–579. ACM.

Wu, Y., Lin, C.-K., Chang, E. Y., and Smith, J. R. (2004b). "Multimodal information fusion for video concept detection". In *Image Processing, 2004. ICIP'04. 2004 International Conference on*, volume 4, pages 2391–2394. IEEE.

Xie, W., Lu, Z., Peng, Y., and Xiao, J. (2014). "Graph-based multimodal semi-supervised image classification". *Neurocomputing*, 138:167–179.

Xu, C., Maddage, N. C., and Shao, X. (2005). "Automatic music classification and summarization". *Speech and Audio Processing, IEEE Transactions on*, 13(3):441–450.

Yang, B., Mei, T., Hua, X.-S., Yang, L., Yang, S.-Q., and Li, M. (2007a). "Online video recommendation based on multimodal fusion and relevance feedback". In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 73–80. ACM.

Yang, L., Liu, J., Yang, X., and Hua, X.-S. (2007b). "Multi-modality web video categorization". In *Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 265–274. ACM.

Yang, M., Kpalma, K., and Ronsin, J. (2008a). "A survey of shape feature extraction techniques". *Pattern recognition*, pages 43–90.

Yang, Y., Nie, F., Xu, D., Luo, J., Zhuang, Y., and Pan, Y. (2012). "A multimedia retrieval framework based on semi-supervised ranking and relevance feedback". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(4):723–742.

Yang, Y., Wu, F., Xu, D., Zhuang, Y., and Chia, L.-T. (2010). "Cross-media retrieval using query dependent search methods". *Pattern Recognition*, 43(8):2927–2936.

Yang, Y., Xu, D., Nie, F., Luo, J., and Zhuang, Y. (2009). "Ranking with local regression and global alignment for cross media retrieval". In *Proc. of the 17th ACM international conference on Multimedia*, pages 175–184. ACM.

Yang, Y., Zhuang, Y.-T., Wu, F., and Pan, Y.-H. (2008b). "Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval". *Multimedia, IEEE Transactions on*, 10(3):437–446.

Yoshitaka, A. and Ichikawa, T. (1999). "A survey on content-based retrieval for multimedia databases". *Knowledge and Data Engineering, IEEE Transactions on*, 11(1):81–93.

Zamir, O. and Etzioni, O. (1998). "Web document clustering: A feasibility demonstration". In *Proceedings of the 21st annual international ACM SIGIR*

conference on Research and development in information retrieval, pages 46–54. ACM.

Zamir, O., Etzioni, O., Madani, O., and Karp, R. M. (1997). "Fast and Intuitive Clustering of Web Documents."

Zeppelzauer, M. and Schopfhauser, D. (2016). "Multimodal classification of events in social media". *Image and Vision Computing*.

Zhai, X., Peng, Y., and Xiao, J. (2014). "Learning Cross-Media Joint Representation With Sparse and Semisupervised Regularization". *Circuits and Systems for Video Technology, IEEE Transactions on*, 24(6):965–978.

Zhang, C., Chen, W.-B., Chen, X., Tiwari, R., Yang, L., and Warner, G. (2009). "A multimodal data mining framework for revealing common sources of spam images". *Journal of multimedia*, 4(5):313–320.

Zhang, Z. and Zhang, R. (2008). "Multimedia data mining: a systematic introduction to concepts and theory". CRC Press.

Zhao, R. and Grosky, W. I. (2002). "Narrowing the semantic gap-improved text-based web document retrieval using visual features". *Multimedia, IEEE Transactions on*, 4(2):189–200.

Zhen, Y., Gao, Y., Yeung, D.-Y., Zha, H., and Li, X. (2016). "Spectral multimodal hashing and its application to multimedia retrieval". *IEEE Transactions on cybernetics*, 46(1):27–38.

Zhu, Q., Yeh, M.-C., and Cheng, K.-T. (2006). "Multimodal fusion using learned text concepts for image categorization". In *Proceedings of the 14th annual ACM international conference on Multimedia*, pages 211–220. ACM.

Zhu, X., Wu, X., Elmagarmid, A. K., Feng, Z., and Wu, L. (2005). "Video data mining: semantic indexing and event detection from the association perspective". *Knowledge and Data Engineering, IEEE Transactions on*, 17(5):665–677.

Zhuang, Y., Yang, Y., Wu, F., and Pan, Y. (2007). "Manifold learning based cross-media retrieval: a solution to media object complementary nature". *The Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology*, 46(2-3):153–164.

Zhuang, Y.-T., Yang, Y., and Wu, F. (2008). "Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval". *Multimedia, IEEE Transactions on*, 10(2):221–229.

Zou, X. and Bhanu, B. (2005). "Tracking humans using multi-modal fusion". In *Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, pages 4–4. IEEE.

Zu Eissen, S. M., Stein, B., and Potthast, M. (2005). "The suffix tree document model revisited". In *Proceedings of the 5th International Conference on Knowledge Management*, pages 596–603.

Publications

Journal Articles

1. Pushpalatha, K., & Ananthanarayana, V.S. "Feature pattern based representation of multimedia documents for efficient knowledge discovery." Multimedia Tools and Applications, 1-27., doi: https://doi.org/10.1007/s11042-016-3434-y.
2. Pushpalatha, K., & Ananthanarayana, V. S. (2017). "A tree based representation for effective pattern discovery from multimedia documents." Pattern Recognition Letters, 93, 143-153., doi: https://doi.org/10.1016/j.patrec.2016.10.005

Conference Publications

1. Pushpalatha, K., & Ananthanarayana, V. S. (2015, December). "A New Glowworm Swarm Optimization Based Clustering Algorithm for Multimedia Documents." In 2015 IEEE International Symposium on Multimedia (ISM) (pp. 262-265). IEEE, Miami. USA
2. Pushpalatha, K., & Ananthanarayana, V. S. (2014, December). "An information theoretic similarity measure for unified multimedia document retrieval." In 7th International Conference on Information and Automation for Sustainability (pp. 1-6). IEEE. Srilanka.
3. Pushpalatha, K. & Ananthanarayana, V. S (2014, December). "An Unified Approach for Multimedia Document Representation and Document Similarity." In Computational Science and Engineering (CSE), 2014 IEEE 17th International Conference on (pp. 249-256). IEEE. China.
4. Pushpalatha K & Ananthanarayana V. S. (2014, August) "An Approach for Multimedia Information Retrieval Based on Harmonized Representation of Multimedia." In Data Mining and Warehousing (ICDMW), 2014 Elsevier 8th International Conference on (pp. 251-258), Bangalore.

Brief Bio-data

Pushpalatha K

Associate Professor

Sahyadri College of Engineering & Management

Sahyadri Campus, Adyar

Mangaluru - 575007

Permanent address

Pushpalatha K

D/o Late. K Rukmaya,

"Pruthvichaya"

Kumpala, Mithranagar

Kotekar, Mangaluru - 575022

Karnataka.

Phone: 9902338353

Email: pushpak35@gmail.com