Arenberg Doctoral School of Science, Engineering & Technology
Faculty of Engineering
Department of Electrical Engineering

Learning from multi-view data: clustering algorithm and text mining application

Xinhai LIU

Dissertation presented in partial
fulfillment of the requirements for
the degree of Doctor
in Engineering

September 2011

Learning from multi-view data: clustering algorithm and text mining application

Xinhai LIU

Jury:
Prof. dr. ir. C. Vandecasteele, chair
Prof. dr. ir. B. De Moor, promotor
Prof. dr. ir. J. Vandewalle
Prof. dr. ir. Y. Moreau
Prof. dr. ir. J. Suykens
Prof. dr. M. Moens
Prof. dr. W. Daelemans (UA)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering

September 2011

D/2011/7515/111
ISBN 978-94-6018-410-9

Acknowledgement

The research work presented in this thesis was carried our at the lab of Signals, Identification, System Theory and Automation (SCD/SISTA) in the Department of Electrical Engineering (ESAT) of Katholieke University Leuven in the presence of complete and enthusiastic professors and colleagues. This PhD would have been impossible without their professional and moral support.

At first, I am very grateful to my pormotor Prof. Bart De Moor for this opportunity to start my doctoral studies in this research group. I would like to show appreciation for his continuous support with respet to my funding, academic guidance and all the university paperwork required throughout these years.

Many thanks to Prof. Carlo Vandecasteele, Prof. Joos Vandewalle, Prof. Johan Suykens, Prof. Yves Moreau, Prof. Marie-Francine Moens and Prof. Walter Daelemans for being part of the jury of this thesis and the related academic suggestions. I also want to thank Prof. Vicent Blondel, a member of my supervision committee, for introducing me to the world of graphs and networks.

I would like to express my sincere appreciation to Prof. Lieven De Lathauwer for his introduction of tensor models. When preparing the related papers, I was impressed by his concrete and rigorous attitude in research. The mathematical representations of several chapters in this thesis were significantly improved due to his suggestions.

Next, I would like to convery my gratitude to my collaborators. Dr. Frizo Janssens supervised me to start the research on text mining and clustering analysis as well as proofread several chapters of this thesis. Prof. Wolfgang Glänzel offered me some research instructions and helped me to analyze the related clustering results. Dr. Shi Yu gave me many useful suggestions and together with him, we applied and refined our methods in the scientometric problems. With the help of Dr. Olivier Gevaert and Dr. Léon-Charles Tranchevent, I began to carry out the research on biomedical applications.

This thesis would not have been finished without the collaboration with some smart and diligent colleagues in SCD/SISTA. Dr. Carlos Alzate and Dr. Mariya Ishteva offered me many insightful discussions. As a Java expert, Arnaud Installe supported me to develop a Java based information retrieval system from a zero starting point. As native speakers, Tunde Adefioye and Ernesto Iacucci helped me a lot in correcting the mistakes in my papers and thesis. I am also thankful for Mauricio Agudelo, Jiqiu Cheng, Philippe Dreesen, Marco Signoretto, Quoc Tran Dinh, Tillmann Falck, Attila Kozma, Kim Batselier, Pieter Schuddinck, Dries Geebelen, Siamak Mehrkanoon and Rocco Langone for some interesting discussions and their kindly help all along.

I would like to acknowledge the help that I received from the administrative and professional staffs, namely, Ida Tassens, Maarten Truyens, Ilse Pardon, John Vos, Mimi Deprez and Liesbeth Van Meerbeek.

I would also like to take this opportunity to express my thanks for the kind support of those outside our research group. Dr. Lei Tang (Research Scientist, Yahoo! Labs, USA) provided some useful data and professional explanation to my questions. Dr. Renaud Lambiotte (Imperial College, London) and Prof. Jean-Charles Delvenne (UCL, Belgium) have introduced lots of state of art knowledge about graph and network models to me. Meanwhile, I would like to thank my three English teachers in CLT, Ms. Hilde De Clercq, Ms. Jackie Clare and Ms. Marleen Vanderheiden for their teaching and encouragement all these years in Leuven.

Furthermore, I want to express my thanks to some friends that in one way or another have made my stay in Belgium pleasant, namely, Lianming Li, Yantian Chen, Tao Xiang, Shubin Wu, Rui Duan, Shiqiong Yang, Lihong Wang, Yingmei Feng, Ning Ma, Lin Zhang, Hang Gao, Kai Liu, Yao Yue, Qiong Yang, Tingyao Wu, Zhenggang Wang, Roberto López-Sastre, David Tingdahl, Stefaan De Roeck and so on. Also many thanks for partially financial support of China Scholarship Council (CSC), the Education Section of Chinese embassy in Belgium (in particular, the education counselor Mr. Luxin Wang and the former education counselor Ms. Huanbai Xue), and International Office of K.U.Leuven (in particular Ms. Anouk De Weerdt).

Last but not least, I want to express my deep appreciation to my family for their support. Specially I owe a debt of gratitude to my parents, brothers, who in the distance, have supported me over the years, encouraged me to continue through my most difficult times. I look forward to seeing you all soon.

Xinhai Liu

Leuven, September 2011

Abstract

The rapid development of information and computer technology (ICT) in the last two decades has fundamentally changed almost every discipline in science and engineering, transforming many fields from data-poor to increasingly data-rich, and calling for innovative data mining methods to conduct the related research. Meanwhile, as data collection sources and channels continuously evolve, data can be extracted from multiple information sources and observed by various models. Therefore, learning from multi-view data has become a crucial step in machine intelligence and knowledge discovery.

For the purpose of integrating and leveraging the mass amount of multi-view data to obtain significant and complementary high-level knowledge, this dissertation investigates learning from multi-view data from two sides: clustering algorithm and text mining application.

The dissertation is organized into three parts.

In the first part, we analyze multi-view clustering from a multilinear perspective and create several novel multi-view clustering algorithms. At first, modeling multi-view data as a tensor, we present a novel tensor based multi-view partitioning framework for integrating multi-view data in the context of spectral clustering. Within this framework, a joint optimal subspace shared by multi-view data as well as the multilinear relationships among multi-view data are revealed by the relevant tensor methods. Second, taking multi-view data as multiple graphs, we put forward a multi-view clustering strategy based on simultaneous trace maximization (STM), which analyzes multi-view data through a multilinear perspective as well. Third, a joint dimension reduction scheme based on tensor decomposition is presented, particularly for multi-view data. The dimension reduction scheme is embedded into the STM based multi-view clustering strategy, which enables us to handle large-scale multi-view data.

In the second part, we investigate text mining to extract multi-view heterogeneous data from a large-scale publication database of Web of Science (WoS).

In order to facilitate the scientific mapping that is useful for monitoring and detecting new trends in different scientific fields, hybrid clustering, either in vector spaces or in graph spaces, is carried out to integrate these multi-view data. Regarding hybrid clustering in vector spaces, various methodologies are included in a unified framework, which consists of two general approaches: clustering ensemble and kernel fusion. A mutual information based weighting scheme is proposed to leverage the effect of multiple data sources in hybrid clustering. Concerning hybrid clustering in graph spaces, various graphs are generated from multi-view data. Utilizing the complementary properties of both text graph and citation graph, we present a hybrid strategy named graph coupling. Meanwhile, based on the modularity optimization, our graph coupling strategy detects the number of clusters automatically and provides a top-down hierarchical analysis, which fits in with the practical applications. In addition, the computation of this modularity based hybrid clustering method is so efficient that it does well in partitioning large-scale data.

In the third part, we propose a novel strategy to derive knowledge from textual information from a multi-view perspective. The multiple views can be different controlled vocabularies, term weighting schemes, publishing time periods and biomedical subjects. Our strategy has been applied to the MEDLINE corpus and analyzed using a disease based data set. In particular, we investigate the effect of combining multiple views for clustering and assessed whether vertical searches can be more accurate for specific biological questions. Moreover, a Web application of our multi-view text mining strategy is developed for gene retrieval.

To conclude, the theory, algorithm, applications and software presented in this dissertation provide an interesting perspective for clustering algorithms and text mining applications. In addition, the obtained results are promising to be applied and extended to many other relevant fields besides scientific mapping and bioinformatics.

List of symbols

Variables and symbols (Page No.)

$a, b, \ldots, \alpha, \beta, \ldots$	scalars
$A, B, \ldots,$	vectors
$\mathbf{A}, \mathbf{B}, \ldots$	matrices
$\mathcal{A}, \mathcal{B}, \ldots$	tensors (pp.36)
$\tilde{\mathcal{A}}$	low multilinear rank approximation of a tensor $\mathcal{A}$ (pp.68)
$\mathcal{A}_{i=\alpha}$	subtensor of $\mathcal{A}$ obtained by fixing the index i as α (pp.66)
$\mathbf{A}_{(n)}$	n-mode matrix unfolding of a tensor $\mathcal{A}$ (pp.36)
$\mathbf{I}$	identity matrix (pp.33)
$\mathbf{U}$	relaxed assignment matrix for spectral clustering (pp.33)
$\mathbf{S}$	adjacency or similarity matrix of a graph (pp.32)
$\mathbf{D}$	degree matrix (pp.33)
$\mathbf{S}_N$	normalized similarity matrix (pp.33)
$\mathbf{L}$	Laplacian matrix (pp.33)
$\mathbf{L}_{NCut}$	NCut based normalized Laplacian matrix (pp.33)
$\tilde{\mathbf{S}}$	normalized matrix integration (pp.35)
$\mathbf{B}$	modularity matrix (pp.119)
W	weighting vector of multi-view data (pp.35)
w_v	weighting factor of the vth view data (pp.35)
V	number of views (pp.34)
K	number of clusters (pp.33)
K'	number of reduced dimension (pp.67)
$\mathbf{K}$	kernel (pp.96)
N	number of instances (pp.32)
$\mathbb{R}$	set of real numbers (pp.32)
$\mathbb{R}^{m \times n}$	real value $(m \times n)-$ matrix (pp.33)
$\mathbb{R}^{I_1 \times I_2 \times I_3}$	real value $(I_1 \times I_2 \times I_3)-$ tensor (pp.36)

Basic operations

$\mathbf{A}^T$	transpose of matrix $\mathbf{A}$
vec($\mathbf{A}$)	vector representation of matrix $\mathbf{A}$ (all columns of $\mathbf{A}$ stacked each other)
$\|\mathbf{A}\|_F$	Frobenius norm, $\sqrt{trace(\mathbf{A}^T\mathbf{A})}$
trace($\mathbf{A}$)	trace of $\mathbf{A}$, $\sum \mathbf{A}_{ii}$
$\|A\|_2$	2-norm of a vector or a matrix

Abbreviations

Acronym	**Notion** (Page No.)
AdacVote	The cumulative vote weighting method by Ayad et al. (pp.44)
ARI	Adjusted Rand Index (pp.97)
ANMI	Average Normalized Mutual Information (pp.92)
AKFCM	Average Kernel-Fusion Clustering Method (pp.99)
BGC	Bibliographic Coupling (pp.91)
BV-CRC	Binary Cross-Citation (pp.91)
BV-Text	Binary score of TFIDF (pp.91)
CCA	Canonical Correlation Analysis (pp.29)
COC	Co-Citation (pp.91)
CRC	Cross-Citation (pp.91)
EAC-AL	Evidence Accumulation Clustering by applying Average Linkage (pp.98)
ESI	Essential Science Indicators (pp.89)
EVD	Eigenvalue Decomposition (pp.42)
FI	Feature Integration (pp.44)
HOOI	Higher Order Orthogonal Iteration (pp.40)
IDF	Inverse Document Frequency (pp.48)
KFCM	Kernel-Fusion Clustering Method (pp.96)
KNN	k-Nearest Neighbour (pp.115)
LSI	Latent Semantic Indexing (pp.90)
LSI-CRC	Latent Semantic Indexing of Cross-Citation (pp.91)
LSI-TFIDF	Latent Semantic Indexing of TFIDF (pp.90)
LMF	Linked Matrix Factorization (pp.44)
MEDLINE	Medical Literature Analysis and Retrieval System Online (pp.20)
MKF	Multiple Kernel Fusion (pp.44)
MLSVD	Multilinear Singular Value Decomposition (pp.13)
MSV	Mean Silhouette Value (pp.97)
NMI	Normalized Mutual Information (pp.93)
NCut	Normalized Cut (pp.33)
PCA	Principal Component Analysis (pp.24)
PARAFAC	Parallel Factor Analysis (pp.32)
QMI	The clustering ensemble method by Topchy et al. (pp.98)
SA	Strehl's clustering ensemble algorithm (pp.44)
STM	Simultaneous Trace Maximization (pp.62)
SVD	Singular Value Decomposition (pp.37)
SVM	Support Vector Machine (pp.110)
TF	Term Frequency (pp.48)
TF-IDF	Term Frequency-Inverse Document Frequency (pp.48)
WKFCM	Weighted Kernel-Fusion Clustering Method (pp.96)
WLCDM	Weighted Linear Combination of Distance Matrices Method (pp.99)
WoS	Web of Science journal database (pp.16)

Contents

Chapter 1

Introduction

1.1 General background

1.1.1 Multi-view data and multi-view learning

The rapid development of computer and information technology in the last two decades has fundamentally changed almost every discipline in science and engineering, transforming many fields from data-poor to increasingly data-rich, and calling for innovative data mining methods to conduct the related research.

Meanwhile, as data collection sources and channels continuously evolve, data can be extracted from multiple information sources and observed by various models, which forms **multi-view data**, that is, the same instance with different representations.

In general, each view may have different formulations or statistical properties. In the common machine learning problem setting, we often assume that the data is represented in a single vector space or in a single graph space. In many real-life problems, however, the same instances of multi-view data may be represented in several different vector spaces, or in several different graph spaces, or even a mixture of vector spaces and graph spaces [150]. For example, in scientific publication analysis, documents can be represented in the text vector space as well as in the citation based graph space.

Learning from multi-view data has become a crucial step in machine intelligence and knowledge discovery. On the one hand, many machine learning tasks such as classification, regression and clustering, can significantly improve

their performance if information from multi-view data can be properly integrated and leveraged. On the other hand, regarding several emerging fields and applications, such as, healthcare informatics, computer vision and music retrieval in particular, comparing patterns from multi-view data and understanding their relationships can be extremely beneficial for these applications.

Meanwhile, according to the relevant definition [2], machine learning tasks heavily rely on the empirical data where the latent patterns are hidden. As the information contained within the data is incomplete or the data is corrupted by noise, the goals of machine learning become challenging or even impossible to achieve. As a result, multi-view learning is proposed to tackle these challenges.

Multi-view learning: It is assumed that multi-view data gives a broader understanding of the task and thus yields better performance. The goal of multi-view learning is to effectively explore and exploit the information from multi-view data for the purpose of improving the learning performance [118].

1.1.2 Benefits of multi-view learning

In the following, the three apparent benefits from multi-view learning and the relevant examples are illustrated.

Benefit one: recovering a full pattern by learning from multi-view data with an example of 3D image reconstruction

As known, actual scenes can only be directly captured in the form of 2D plane images, either by human eyes or by a camera. Because of the 3D structure of the real world, the 2D image only contains limited information which is inefficient to gain a complete understanding of the scene. Fortunately, thanks to the excellent multi-view learning mechanism of the visual perception system, human beings are able to perceive the 3D structure of the scene by seamlessly integrating images about the surrounding scene from two perspectives (two eyes). The basic visual perception mechanism of a human being is shown in Figure 1.1.

Analogous to the visual system of human beings, computer vision was put forward and its aim is to enable computers to imitate the functionality of human vision through automatically learning from multiple views. As illustrated in Figure 1.2, given multiple 2D images of a building which seem incomplete but complementary, a 3D model of the building which seems more complete is learned by collecting the essential evidence from various views.

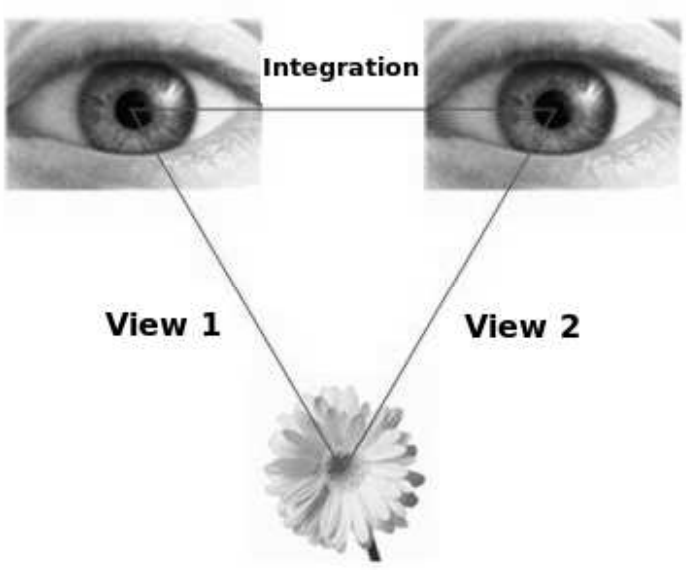

Figure 1.1: Human vision from multi-view data

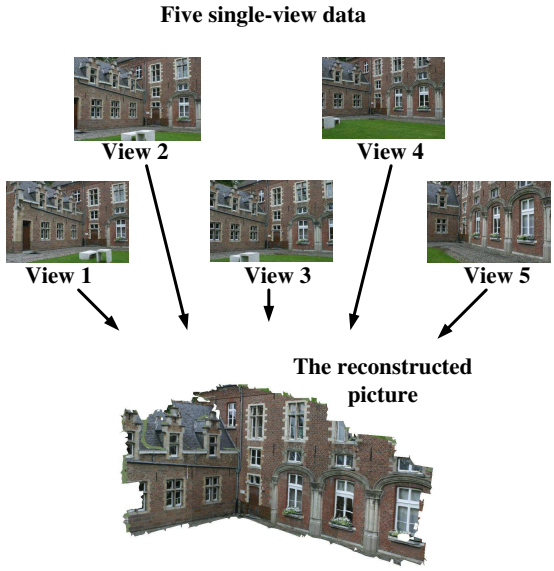

Figure 1.2: Computer vision from multi-view data

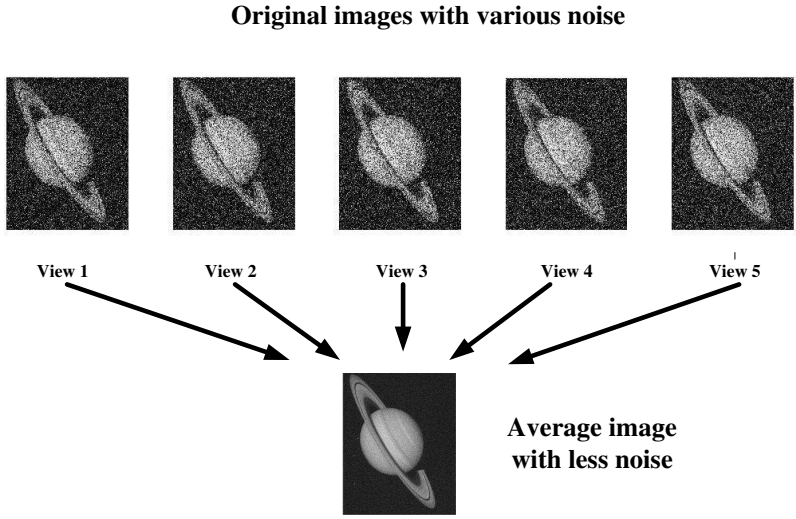

Figure 1.3: Image denoising by averaging multi-view data

This example demonstrates learning from multi-view data tends to find a full "picture". Single-view data sometimes includes incomplete information while multi-view data usually contains complementary information. As a result, a relative complete pattern could be obtained by collecting the complementary information from multi-view data, especially when the weaknesses of one view are complemented by the strengths of other views.

Benefit two: reducing noise by learning from multi-view data with an example of image denoising

Figure 1.3 illustrates a denoising example in image processing by learning from multi-view data. As shown in the above, a planet in the space is observed by five cameras at different nearby locations. All these pictures appear to be full of noise due to the limitation of imaging technology, which may bring some trouble to further observation and analysis. However, the noise of these five pictures are different from each other because they are randomly generated within each view (each camera).

On the other hand, the different random noise within each view can be reduced through simply averaging the five pictures. At the same time, the common pattern (the original image of the planet) shared by all these views is emphasized by such an average operation. Thus an integrated picture with less noise is obtained as can be seen in Figure 1.3.

This image processing example demonstrates learning from multi-view data leads to robust results by reducing the noise from each single-view data. In general, the presence of noise in each single-view data (or the corruptness of data) sometimes makes the detection of patterns (clusters) more difficult, leading to the unsatisfied analysis of singel-view data. On the other hand, multi-view learning is able to circumvent the side-effect of noise or corrupted data in each view and emphasize the common pattern shared by multi-view data.

Benefit three: facilitating the learning tasks which can not be implemented only by single-view data with an example of Webpage retrieval Another exciting application of learning from multi-view data is Web information retrieval as illustrated in Figure 1.4. Webpages contain rich multi-view data, such as textual content and hyperlinks. Each data has diverse physical property. However, appropriate integration can accomplish some tasks which is hard to implement in one single-view data. For instance, the big success of Google lies in the elaborate integration of text and hyperlink data for Webpage retrieval [113]. Given a query, a huge number of retrieved Webpages are output. According to traditional information retrieval, the retrieval results are ordered by calculating the similarity between the textual query and the content of the Webpage. Due to the huge number of Webpages, it is impossible to obtain the retrieval results with meaningful order online by this traditional way.

On the other hand, the sparse hyperlinks can be employed to efficiently calculate the relative importance (the order) of each Webpage, that is the basic idea of PageRank. Then PageRank provides a meaning ranking of the huge number of Webpages by offline computation. With such PageRank, the online retrieval can be implemented by online matching of simply textual pattern, rather than by intensive computation of textual similarity. As can be seen in Figure 1.4, the integration of online textual pattern matching and offline PageRank leads to the immediate retrieval.

This example shows that, learning from multi-view data is able to accomplish the learning tasks that are impossible to implement by single-view data due to its limitation. The complementary property among multi-view data is able to overcome the limitation of single-view data and expand their application areas.

1.1.3 Challenges of multi-view learning

Traditionally, machine learning or data mining algorithms are conceived for learning from single-view data. The need to develop general theories,

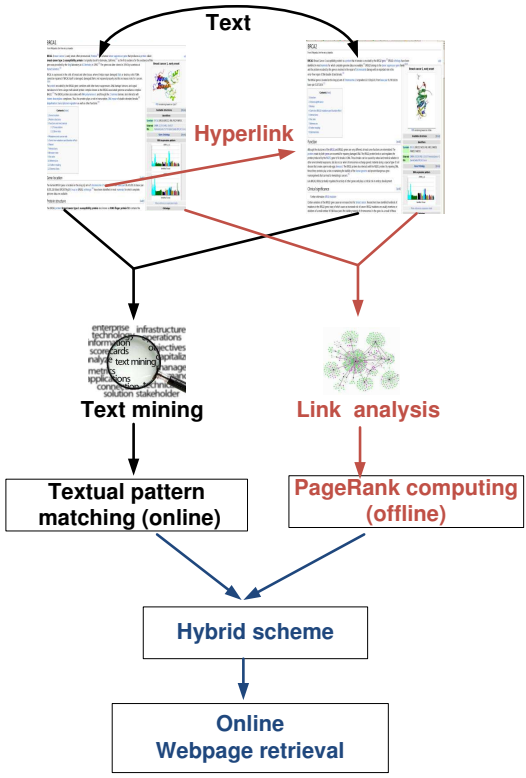

Figure 1.4: Web information retrieval from multi-view data

frameworks, data structures, and heuristics, for multi-view learning has become increasingly crucial.

Although some relevant work utilizing multi-view data has been proposed, these methods are usually rather ad-hoc and do not adequately address some of the most fundamental research issues in this field [118]. Unleashing the full power of multi-view data is, however, a very challenging task.

The model of multi-view data for joint analysis. Multi-view data may come from the same feature space or different feature spaces. The inherent properties of multi-view data may vary remarkably, for instance, in scientific publication analysis, text data denotes the attributes of each document while citation data depicts the link relationships among various documents. For the convenience of joint analysis, modeling multi-view data in a unified form is required. How to model multi-view data in a proper way is a basic issue in

multi-view learning.

Leveraging the effect of multi-view data. In many applications of multi-view learning, it is of great interest to develop a strategy that is able to investigate the underlying relationships amongst views. Then such a strategy can potentially identify the interactions between multi-view data, and also evaluate their learning capabilities. The latter can prove particularly useful in problems where collecting the necessary data from a view may be resource demanding and thus expensive [30]. Although it is known that various single-view data play different roles in the joint learning, how to leverage their effect to facilitate the learning tasks is still a challenging problem.

Dimension reduction of multi-view data. Computer power is growing by Moore's law while data volume is growing even faster. In practical applications, both the number of objects and the number of features are becoming huge. Moreover, multi-view observation could expand the data volume in a rapid rate as well. Preprocessing by dimension reduction on multi-view data seems an essential step for further analysis. Because multi-view data is more complicated than single-view data, how to implement the joint dimension reduction of multi-view data becomes a critical issue in multi-view learning.

In this Thesis, the above questions will be handled respectively. Meanwhile, under the umbrella of multi-view learning, we focus on two basic tasks: multi-view clustering and multi-view text mining. These two tasks are not separated completely. In fact, text mining data can be directly employed to clustering. On the other hand, clustering sometimes can be applied to facilitate text mining tasks.

1.2 Clustering of multi-view data

1.2.1 Clustering analysis

Data clustering is a fundamental problem in many fields, such as machine learning, data mining and computer vision. Unfortunately, there is no universally accepted definition of a cluster, probably because of the diverse forms of clusters in real applications [67]. For instance, in distance based clustering analysis, two or more objects belong to the same cluster if they are "close" according to a given distance (in this case geometrical distance) while in concept based clustering, two or more objects belong to the same cluster if this one defines a concept common to all that objects, in other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures [148]. But it is generally agreed that the objects belonging

to a cluster satisfy certain internal coherence condition, while the objects not belonging to a cluster usually do not satisfy this condition [67].

Although thousands of clustering algorithms have been published and continue to appear, some fundamental challenges associated with clustering still remain: for example, how to determine the unknown number of clusters in the given data, how to evaluate the validity of discovered clusters and partition, which clustering method is proper for current data, and how to define the pair-wise similarity [67]. Furthermore, although an ideal cluster can be defined as a set of points that is compact and isolated, the noise in the data usually makes cluster analysis more difficult.

All in all, clustering analysis is virtually an exploratory tool, and the output of clustering algorithms only suggests hypotheses [67].

1.2.2 Multi-view clustering

The aim of clustering is exploratory in nature to find the structural pattern hidden in data. As to clustering on single-view data, the data structure sometimes seems to be corrupted by noise or incomplete due to the limited information it contains. Hence, we will explore how multiple views make the clustering problem significantly more tractable. Concerning clustering of multi-view data, based on the above benefits of multi-view learning, first, a robust cluster structure is expected to be obtained because the random noise and isolated outliers are deleted by learning from multiple views; second, a cluster structure is expected to be found through fusing the complementary information of multi-view data.

Bickel and Scheffer put forward the multi-view clustering concept in 2004 and empirically find that multi-view clustering strategy greatly improves on its single-view counterparts [12]. In this original work, multi-view clustering refers to clustering instances that are represented by multiple independent sets of features. A multi-view clustering strategy via canonical correlation analysis (CCA) is presented in [27]. This method assumes that the views are uncorrelated given the cluster label. These two methods are based on an assumption that multiple views are independent of each other. In fact, the real multi-view data is not entirely uncorrelated and they usually share certain inner relationship instead. Furthermore, with the number of views increasing, it is hard to preserve the independence of these views. Therefore, such an assumption limits the application of these methods.

A strategy named multiple view semi-supervised dimensionality reduction is devised and employed to multi-view clustering [61]. A consensus pattern is

learned from multiple embeddings of multi-view data. However, both this strategy and CCA based multi-view clustering need part of labeled data which is usually unavailable in practical applications.

Consequently, we will investigate the multi-view clustering under the free assumptions of both unsupervised learning and existing certain dependent relationship amid them, which would be more challenging but more close to the reality.

1.2.3 Our multi-view clustering strategies

As to the integration of multi-view data, a natural idea is to concatenate different types of data into a single vector. In fact, inter-feature dependencies within one data set are more likely to be relevant than dependencies between two different types of data [109]. Furthermore, since the structure of each view is disparate, it is unwise to regard multiple representations as one view by simply connecting all features. Thus concatenating multi-view data as a single vector may treat the representation of each view in the same way and ignore their diversities [61]. Consequently, we take multi-view data either as different similarity matrices (kernels) or as a tensor, rather than as a concatenated vector.

By treating various views differently, we can construct a kernel (or a similarity matrix) on each data and obtain the sum of kernels for multi-view clustering. Although only kernel summation seems simple and efficient to implement and even can achieve good clustering performance, an automatic mechanism assigning different weights to each kernel is still preferred. Such a weighting mechanism can bring some apparent benefits, for instance, it can delete or reduce the noisy kernel (view) and provide a boundary error guarantee [109].

With the aim of both integrating multi-view data in a proper way and utilizing their inherent relationship to facilitate clustering, we carry out multi-view clustering in the following perspectives.

Multi-view clustering by tensor decomposition

Multi-view data can be naturally modeled as a tensor. First, a common space shared by multi-view data can be obtained by tensor decomposition. The partitioning is carried out in this common space to get the final clustering results. Second, the inherent relationship of multi-view data can be regarded as a kind of multilinear relationship, which can be captured through tensor

decomposition as well. Such a relationship actually corresponds to the weights assigned to multi-view data.

In fact, our tensor based multi-view clustering strategy provides a general framework to integrate multi-view data for joint partition. Hence, our framework can be easily extended to other multi-view learning tasks, such as classification and spectral embedding.

Our tensor based multi-view analysis is closely related to a concept named **multi-way analysis**. Multi-way analysis is the natural extension of multivariate analysis, when data are arranged in three- or higher way arrays [19]. In analogy to a matrix, each direction in a high-way array is called a way or a mode, and the number of levels in the mode is called the dimension of that mode. Tensor decomposition has been successfully employed to multi-way data analysis, such as, chemistry, food industries, social network analysis, chemometrics, signal processing, Web search, data mining, scientific computing and bioinformatics [3, 8, 19, 38, 76, 77, 89, 122, 128].

Multi-view analysis can be regarded as a special case of multi-way data analysis. First, multi-view data is a kind of three-way data (a third-order tensor), in other words, multi-view analysis is actually a type of three-way analysis. Second, the three modes in multi-view data (the three directions of such a tensor) are fix, that is, the three models refer to objects, features and views respectively.

Multi-view partitioning by simultaneous trace maximization

Multi-view data can also be modeled as multiple graphs. Each graph is usually presented by its similarity (adjacency) matrix. The multilinear relationship (weights) among multi-view data can be analyzed by simultaneous trace maximization of the corresponding similarity matrices. At the same time, the joint dimension reduction of multi-view data by tensor decomposition is taken into account, which enables our strategy to handle large-scale multi-view data.

Multi-view clustering based mutual information analysis of multi-view data

Based on the measure of average normalized mutual information (ANMI) [127], an automatic weighting strategy of multi-view data is proposed and applied to multi-view clustering methods kernel fusion and clustering ensemble.

In addition, we also investigate the complementary properties of multi-view data to facilitate the joint clustering. For example, in the scientific publication analysis, we integrate the sparse links of citation data with the rich semantic

meaning of text data, thereby leading to efficient partitioning of scalable data. Nevertheless, such a strategy is only applicable to certain ad-hoc applications, like Web mining and document analysis.

1.2.4 An example about the comparison between single-view clustering and multi-view clustering

Here is an example of our multi-view clustering on a multiplex network. Multiplex network refers to a group of networks which share the same nodes (vertices) but multiple types of links [99]. Each type of links can be regarded as a single view of the multiplex network. This synthetic multiplex network has 3 clusters, with each having 50, 100, 200 members respectively. We can generate various views of interactions among these 350 nodes and we add some noise to the network by randomly connecting any two vertices with low probability [130]. From the three-view multiplex network, we can form three interaction matrices, each of whose elements is the interaction strength of a pair of vertices. The visualization of the three adjacency matrices is shown as Figure 1.5.

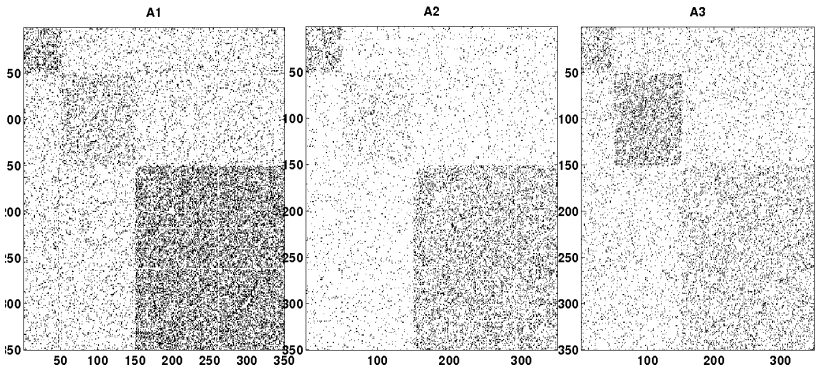

Figure 1.5: The adjacency matrices of three different views in a synthetic multiplex network

In order to demonstrate the power of multi-view clustering, we compare single-view clustering methods with multi-view clustering methods w.r.t the partitioning of this multiplex network. First, we implement spectral clustering on each of the three single-view data. From the partitioning of each single-view data, there is intensive overlap among the three clusters as shown in the upper right part of Figure 1.6. Second, we average the multi-view data

by spectral partitioning on the sum of multiple adjacency matrices and its spectral projection is illustrated in the middle part of Figure 1.6. As compared to single-view clustering, the cluster structure obtained by this multi-view clustering strategy is more clear but the overlap still remains. Third, we partition the multiplex network by a multi-view clustering strategy based on a tensor method named multilinear singular value decomposition (MLSVD). From the lower right part of Figure 1.6, it can be seen that the three clusters are separated clearly. As compared to single-view clustering, the improvement by our MLSVD based multi-view clustering strategy is apparent in terms of partitioning results. This partitioning example suggests that with an appropriate multi-view clustering mechanism, multi-view clustering can recover the latent cluster structure hidden among multi-view data, which can not be achieved by only single-view clustering.

1.3 Text mining from multiple views

Text mining comprises the intelligent automated analysis of textual data and aims for extraction of interesting facts and relationships and discovery of knowledge from large amounts of text [69]. For this purpose, text mining employs techniques and algorithms from disciplines such as data mining, information retrieval, statistics, mathematics, machine learning and natural language processing [41]. Today, text mining is even used for emerging trend detection, policy-making processes, intelligence services, press monitoring to automatically detect breaking news, marketing, data protection, law enforcement and personalized advertising [69]. As most information is currently stored as text, text mining is believed to have a high commercial potential value.

Although the successful applications of text mining have been achieved in many areas, the challenges still remain. For instance, text mining usually lacks the deep and fully understanding of the literature and the information one needs is often not recorded in textual form [36].

To tackle these challenges, first, we carry out multi-view text mining by adopting multiple models and multiple data sources. Information fusion by integrating multi-view information is expected to boost the understanding about the knowledge from literature. Second, on the basis of text mining data, we employ some machine learning or data mining schemes, such as clustering and ranking, to deeply understand the pattern or information hidden in text data.

In this research, some general text mining tasks are involved, such as text clustering and information extraction [41]. Whereas some deep text

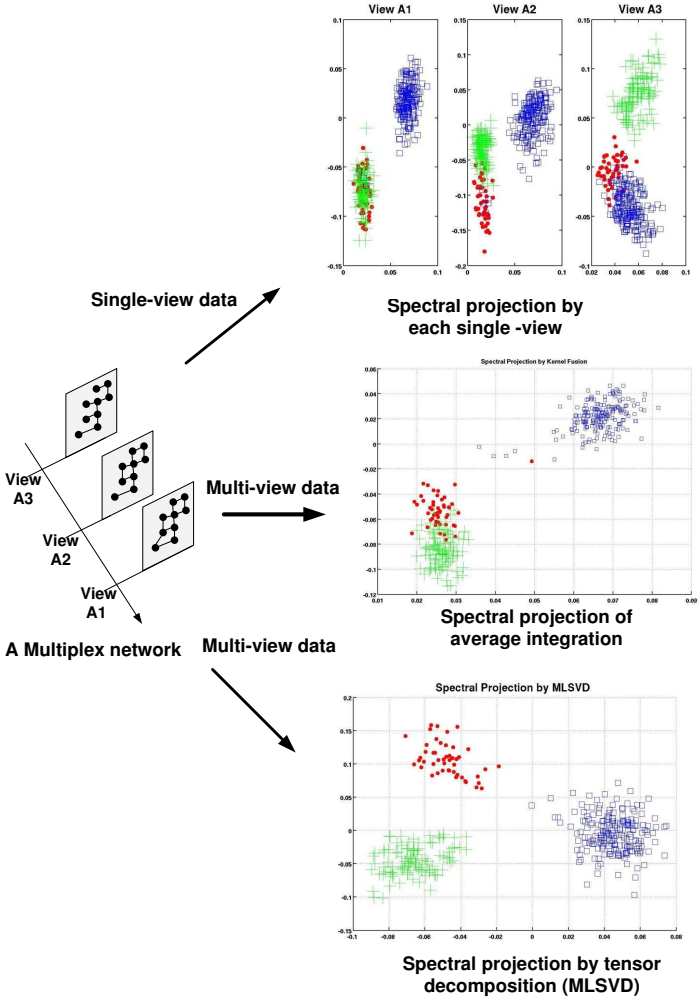

Figure 1.6: Comparison of single-view clustering and multi-view clustering in a multiplex network

analysis tasks have not been handled yet, for instance, production of granular taxonomies, sentiment analysis, document summarization, entity relation modeling and event detection.

In the following, at first, we introduce the multi-view data based on text mining analysis. Next, we briefly present the two applications of our strategy: scientific

mapping and biomedical analysis.

1.3.1 Multi-view data based on text mining

Multi-view data from text content: Multi-view text mining data can be directly extracted from various literature databases. Multi-view text mining data can also be generated by using various text mining models. Such text mining models can be (but not limited to) ontologies, weighting schemes, subjects and publication time periods of literatures.

Multi-view data by information extraction: Information extraction [35] represents a starting point for computers analyzing unstructured text and identifying key phrases and relationships within text. For instance, in scientific publication analysis, bibliometric data is extracted as well and then the related multi-view data is built up, such as cross-citation, co-citation and bibliographic coupling [90].

In addition, the multi-view data generated by text mining implicitly or explicitly can be integrated with other data to advance certain learning tasks. For example, genes can be represented in the expression vector space (corresponding to the genetic activity) and also in the term vector space (corresponding to the text information) [52]. These two heterogeneous data can be integrated together to facilitate the gene clustering [48].

1.3.2 Scientific mapping by multi-view text mining

The aim of scientific mapping is to understand the structure and evolution of various research areas and of their relationships with other fields, based on scientific publications [69]. Text mining is a powerful tool for automatic retrieval of information and for mapping of knowledge embedded in text. These documents contain textual information that can be directly mined for knowledge by using text mining techniques. Besides, other information in these documents can be extracted by text mining as well, such as the citation links, co-occurrence of certain entities, relationships and even some events.

Glenisson [50, 51] and Janssens [69, 73] have conducted research on mapping scientific disciplines by integrating multi-view data (textual content and citation links). Glenssion started a pilot study of combining full text analysis and bibliometric indicators, clustering papers within one journal [51]. Based on text mining in a large-scale bibliographic database, Janssens devises the hybrid clustering strategy by combining text model based on publication content and graph model based on bibliometric data [69]. In this Thesis, we carry on the

research of scientific mapping on a large-scale academic publication database while we extract more multi-view data by text mining. For instance, based on textual content, we generate multi-view text data of TF, IDF,TFIDF and Binary-TFIDF while multi-view citation data is generated based on co-citation, bibliographic coupling and binary-cross-citation.

In addition, we exploit some unified models either in vector spaces or in graph spaces to discover the rich knowledge embedded in this giant amount of publications. On the one hand, in vector spaces, we propose a mutual information based weighted hybrid clustering strategy to integrate text mining based multi-view data for joint mapping and such a strategy is applied to clustering ensemble and kernel fusion. On the other hand, in graph spaces, utilizing the complementary properties of text mining based multi-view data, we formulate an optimal and efficient partitioning strategy named graph coupling, which immediately provides a hierarchical cluster structure without parameter setting for scalable databases.

Figure 1.7 illustrates an example of typical scientific mapping of the Web of Science (WoS) journal database by integrating textual content and citation links. For each cluster, the three most important terms are shown. The network is visualized by Pajek [9]. The edges represent cross-citation links and darker color represents more links between the paired clusters. The circle size represents the number of journals within each cluster.

In scientific mapping, hybrid clustering is traditionally employed to refer to multi-view clustering [69, 73]. Hence, we will alternatively use the name of hybrid clustering and multi-view clustering in the following Chapters.

1.3.3 Biomedical analysis by multi-view text mining

Due to the increasing number of electronically available publications stored in databases such as PubMed, there is an increasing interest in text mining and information extraction strategies applied to the biomedical and molecular biology literature. Figure 1.8 illustrates the number of publications related to human genes each year from 1950 to 2010 and it appears that there is a rapid rise since 2000, which also reflects the fast progress of biomedical research during this period.

As known, literature can provide human beings with the best knowledge. For instance, from the published biomedical literature, by text mining, we can extract the existing knowledge (such as, the relationships and the patterns) among biology entities (such as genes). As an example, we seek the relationships between genes to aid the cancer diagnosis. By mining PubMed

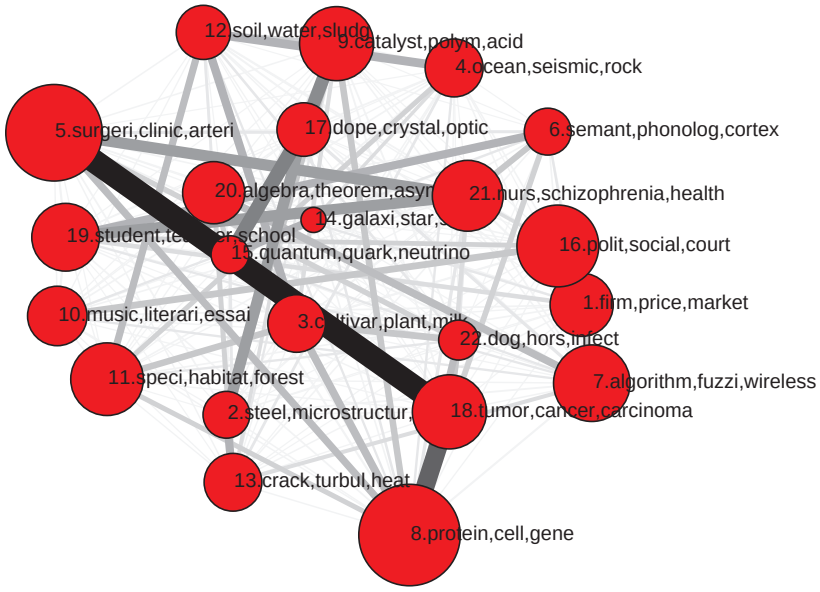

Figure 1.7: A scientific mapping example of 22 clusters on the WoS journal database by MLSVD based multi-view clustering (Data source: Thomson Reuters, Web of Science).

articles, genes are represented as term vectors in vector space model (VSM). Each argument of the gene vector corresponds to a term of a fixed vocabulary (ontology). Then a gene-by-gene similarity matrix is created by calculating the gene-to-gene distances. This matrix is used as prior information to build the basic structure of Bayesian decision network. The use of network seeds can greatly improve the ability of Bayesian network analysis. Moreover, some data mining methods, for instance, clustering, classification and ranking, can also be applied on this similarity matrix to recognize other hidden patterns among genes.

However, in the biomedical field, the analysis of such text data poses much greater challenges than traditional data analysis methods (like manual analysis). For example, genes and proteins are gigantic in size, very sophisticated in function, and the patterns of their interactions are largely unknown [56]. Thus it is a fertile field to develop sophisticated text or data mining methods for in-depth biological literature analysis. From this point of view, text mining is still very young with respect to biology and bioinformatics application. Figure 1.9 plots the number of text mining related papers in PubMed in recent ten

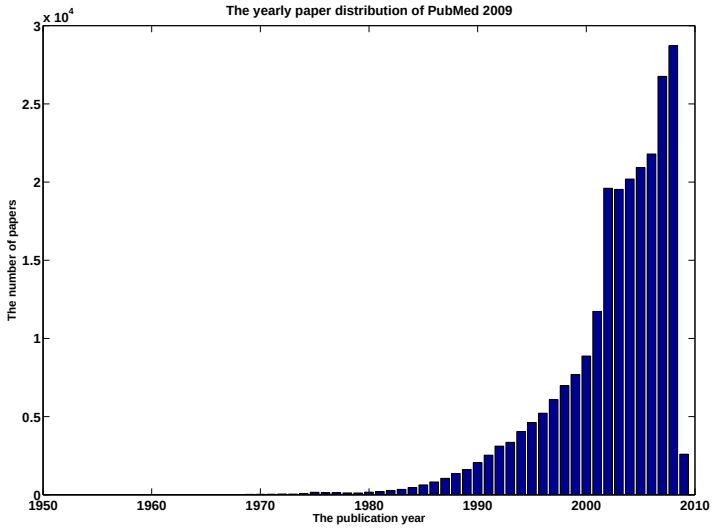

Figure 1.8: The yearly paper distribution (only the human gene related papers are included) in PubMed

years, it is obvious that text mining has attracted increasing attention in the biomedical field.

In text mining, the selection of models plays a big role in the mining stage, for example, various ontologies adopted result in diverse results. Therefore, in this Thesis, we employ a strategy named multi-view text mining for the clinical application. Multi-view text mining refers to the adoption of multiple text mining models, instead of traditionally relying on one model [147]. The multiple views can be different publication time periods, different weighting schemes, different ontologies, different biomedical disciplines and even different citations of these publications. This multi-view text mining provides a flexible and robust framework: on the one hand, one can get a "full picture" through integrating several views for information fusion; on the other hand, one can obtain the vertical observation through one specific view. In particular, we develop a search engine for gene retrieval via such a multi-view text mining scheme, which owns the functionality of both information fusion and vertical search. Glenssion [49] as well as Yu [147] have carried out some similar work, however, their research is limited to using various ontologies.

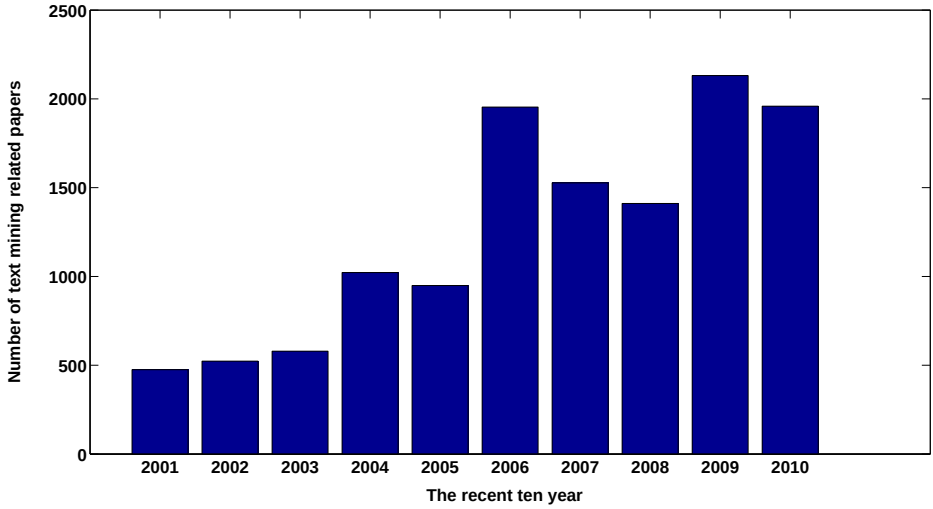

Figure 1.9: The trend of text mining research in biomedical field in recent ten years, estimated from PubMed

1.4 Chapter by Chapter Overview

The overview of each Chapter and its relationship with the author's paper are presented in the following.

Chapter 2 presents a novel tensor-based framework for integrating heterogeneous multi-view data in the context of spectral clustering. Our framework includes two novel formulations: that is, multi-view clustering based on optimization integration (MC-OI) and that based on matrix integration (MC-MI). We show that the solutions for both formulations can be computed by tensor decompositions. We evaluate our methods on synthetic data and two real-world data sets in comparison with baseline methods. Experimental results demonstrate that the proposed formulations are effective in integrating multi-view data in heterogeneous environments.

Chapter 3 puts forward a multi-view clustering strategy based on simultaneous trace maximization, which can be regarded as a multi-view extension of spectral clustering. Our strategy is able to leverage the effect of various views for simultaneous analysis. The pre-processing of dimension reduction is embedded into our strategy by tensor decomposition so that our algorithms are well suited

to the application of large-scale data processing. Our strategy can also be expanded to other clustering schemes to form their multi-view variants, for instance, k-means clustering and modularity based clustering.

Chapter 4 investigates text mining to extract multi-view heterogeneous data from a large-scale WoS database. Hybrid clustering is carried out to integrate those multi-view data to facilitate the scientific mapping. Various methodologies are included in a unified framework, which consists of two general approaches: clustering ensemble and kernel fusion. A mutual information based weighting scheme is proposed to leverage the effect of multiple data sources in hybrid clustering. Three different algorithms are extended by the proposed weighting scheme and they are employed on a large journal set retrieved from the Web of Science (WoS) database.

Chapter 5 tackles the multi-view clustering from a network analysis point of view. At first, multi-view data are modeled in graph spaces, instead of vector spaces. Then we presented a hybrid clustering strategy named graph coupling, by using the complementary properties of both text data and citation data. Based on the modularity optimization, our strategy detects the number of clusters automatically and provides a top-down hierarchical analysis, which fits in with the practical applications. In addition, the method is so efficient that it does well in partitioning large-scale data. We apply our method to cluster the journals of the WoS database.

chapter 6 proposes a novel strategy to provide text prior information from a multi-view perspective. The strategy is implemented by text mining on the Medical Literature Analysis and Retrieval System Online (MEDLINE) database. Our strategy can be applied to do information fusion by integrating multi-view data or provide certain domain knowledge from a small vertical perspective. A Web application of our strategy is developed for gene retrieval. The multiple views can be different controlled vocabularies, weighting schemes, publishing time periods and biomedical subjects. In addition, we employ a set of genes which belong to different diseases to test our multi-view gene retrieval system.

Chapter 7 summarizes the Thesis and introduces several issues that are worth further investigation.

The overview of the relationship between the different Chapters in this dissertation is illustrated in Figure 1.10. As can be seen, both Chapter 4 and Chapter 5 not only belong to multi-view clustering but also belong to multi-view text mining. In addition, the connection of different Chapters is not limited to the above illustration, for instance, in Chapter 6 that focuses on multi-view text mining, tensor based multi-view clustering strategy is still

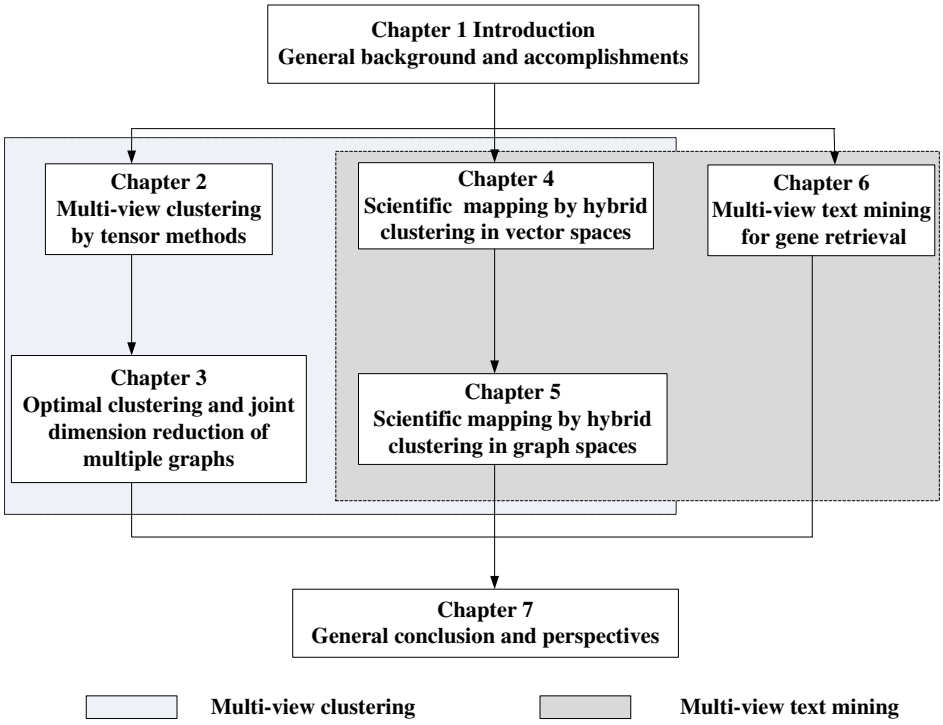

Figure 1.10: Structure of the dissertation

applied for relationship analysis.

1.5 Related research topics in ESAT-SCD, K.U.Leuven

Based on the research conducted in SISTA, an ideal mix is present of methodological and practical expertise in this Thesis. The relevant research topics are presented by one or several recently-finished PhD Thesis. The overview of the connection with SISTA research is illustrated in Figure 1.11.

- *Clustering by data fusion and multi-view text mining*
  Yu S., Kernel-based data fusion for machine learning: methods and applications in bioinformatics and text mining, PhD thesis, Faculty of Engineering, K.U.Leuven (Leuven, Belgium), Nov. 2009.

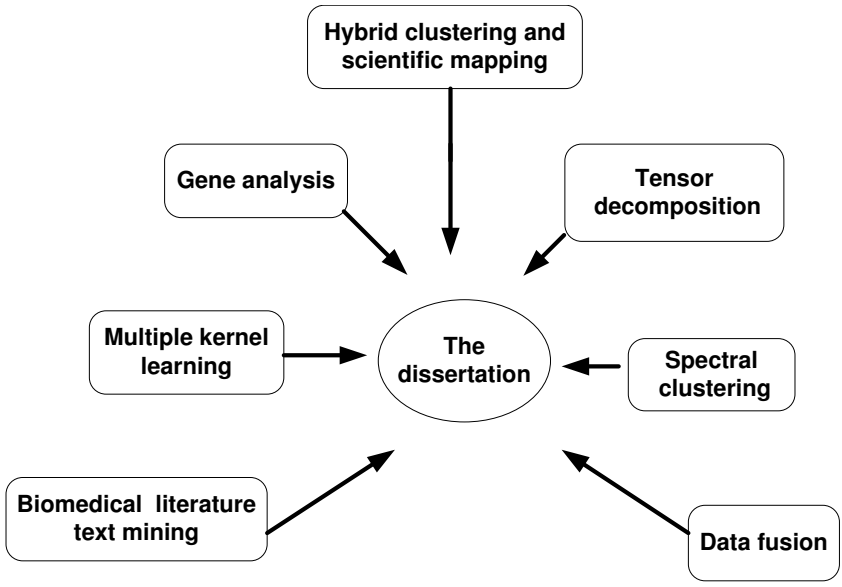

Figure 1.11: Overview of the connection between this dissertation and the relevant research in SISTA

- *Multilinear algebra and tensor decomposition*
  Ishteva M., Numerical methods for the best low multilinear rank approximation of higher-order tensors, PhD thesis, Faculty of Engineering, K.U.Leuven (Leuven, Belgium), Dec. 2009.

- *Unsupervised learning and spectral clustering*
  Alzate C., Support vector methods for unsupervised learning, PhD thesis, Faculty of Engineering, K.U.Leuven (Leuven, Belgium), May 2009.

- *Hybrid clustering and Scientific mapping*
  Janssens F., Clustering of scientific fields by integrating text mining and bibliometrics, PhD thesis, Faculty of Engineering, K.U.Leuven (Leuven, Belgium), May 2007.

- *Biomedical literature text mining*
  Glenisson P., Integrating scientific literature with large scale gene expression analysis, PhD thesis, Faculty of Engineering, K.U.Leuven (Leuven, Belgium), Jun. 2004.

Van Vooren S., Data mining for molecular karyotyping: linked analysis of array-CGH data and biomedical text, PhD thesis, Faculty of Engineering, K.U.Leuven (Leuven, Belgium), Sep. 2009

- *Data fusion of Genetics, Molecular Biology, and Biomedical sources*
  Coessens B., Data integration techniques for molecular biology research, PhD thesis, Faculty of Engineering, K.U.Leuven (Leuven, Belgium), Jun. 2006.

  Gevaert O., A Bayesian network integration framework for modeling biomedical data, PhD thesis, Faculty of Engineering, K.U.Leuven (Leuven, Belgium), Dec. 2008.

1.6 Contributions of this dissertation

1.6.1 Personal contributions

This Thesis is mainly composed of original and independent works of the author in several aspects. The content presented in Chapter 2, 3, 4, 5 and 6 represent the author's personal contribution in multi-view clustering theory, algorithmic innovation and they are all based on publications with first authorship. In Chapter 6, the author collected the corpus data, investigated the bio-ontologies, performed text mining, designed and programed the Web-interface, evaluated the performance and drafted the manuscript. The set of benchmark disease genes and the biomedical validation are based on the collaboration with the co-authors. In both Chapter 4 and Chapter 5, the author designed the algorithms and developed the software. The experimental data set adapted was collected and partially processed by the co-author. The author programmed and applied the proposed algorithms on the experimental data set, evaluated the performance and drafted the manuscript. In conclusion, 90% of the work presented in this Thesis is based on the author's independent research and contribution.

1.6.2 Main contributions

- **Tensor model based multi-view clustering**. We address multi-view clustering from a multilinear perspective. Tensor is a natural model for multi-view data. Thus, we propose to model multi-view data as a tensor and develop a new framework of multi-view spectral partitioning by tensor methods. Within this framework, two novel clustering schemes with tensor based solutions: multi-view clustering by

optimization integration and multi-view clustering by matrix integration, which can integrate data from multiple heterogeneous sources. This contribution will be discussed in Chapter 2.

The related papers:

Liu X., De Lathauwer L., Janssens F., De Moor B., Hybrid Clustering on Multiple Information Sources via HOSVD, in Proc. of the 7th International Symposium on Neural Networks (ISNN 2010), 2010, pp. 337-345.

Liu X., De Lathauwer L., De Moor B., Multi-view Partitioning via Tensor Methods, submitted to IEEE Transactions on Knowledge and Data Engineering.

- **Multi-view clustering by simultaneous trace optimization with joint dimension reduction**. Through simultaneous trace maximization of multiple similarity matrices that describe multi-view data, we obtain the multilinear relationship of multi-view data to facilitate the clustering. This efficient strategy is well suited to dealing with large-scale data. It is easily extended to other common clustering methods to formulate their multi-view variants, for example, k-means and modularity maximization based clustering. Based on a tensor decomposition method of MLSVD, we develop a joint dimension reduction strategy for multi-view data. As compared with dimension reduction by principal component analysis (PCA) on a matrix applicable to single-view data, our strategy is powerful because it is able to reduce a set of matrices simultaneously. This strategy can be utilized as a pre-processing scheme for multi-view clustering as well as other multi-view learning tasks. This contribution will be discussed in Chapter 3.

  The related paper:

  Liu X., De Lathauwer L., Glänzel W., De Moor B., Optimal Clustering and Joint Dimension Reduction of Multiple Graphs, in preparation.

- **Hybrid clustering of multi-view text mining based on mutual information**. Based on bibliometric analysis, we generate five different features in citation spaces and five other features in text spaces by text mining. Each feature provides an independent but complementary observation of the journal instances and thus we implement hybrid clustering on their combination to seek a joint scientific mapping, which is useful for monitoring and detecting new trends in different scientific fields. According to our empirical results and the observations in other related research, there exists certain relationship between average normalized mutual information (ANMI) of one data and its clustering performance. Therefore, we utilize ANMI to assign the weight to each single-view data

for hybrid clustering. This weighted hybrid clustering is carried out on two levels of information fusion: kernel fusion and partition integration (clustering ensemble). The proposed approach is able to provide a more refined structural mapping of journal sets. This contribution will be discussed in Chapter 4.

The related papers:

Liu X., Yu S., Janssens F., Glänzel W., Moreau Y., De Moor B., Weighted Hybrid Clustering by Combining Text Mining and Bibliometrics on Large-Scale Journal Database, Journal of the American Society for Information Science and Technology (JASIST), vol. 61, no. 6, 2010, pp. 1105-1119.

Yu S., Liu X., Tranchevent L., Glänzel W., Suykens J., De Moor B., Moreau Y., Optimized Data Fusion for K-means Laplacian Clustering, Bioinformatics, vol. 27, no. 21, Jan. 2011, pp. 118-126.

Liu X., Yu S., Moreau Y., De Moor B., Glänzel W., Janssens F., Hybrid Clustering of Text Mining and Bibliometrics Applied to Journal Sets, in Proc. of the SIAM Data Mining Conference 09 (SIAM DM 09), Sparks, Nevada USA, May 2009, pp. 46-60.

- **Scalable hybrid clustering based on network analysis**. By modelling our data as a graph or network with sparse links, we investigate the (textual) attributes of each node besides the (citation) links amid them and even combine these two kinds of information to facilitate the community detection task. We focus on the partitioning of the multiplex network. In addition, we tackle the practical issues, for instance, the determination of cluster number and the hierarchical partition structure. This contribution will be discussed in Chapter 5.

  The related papers:

  Zhang L., Liu X., Janssens F., Liang L., Glänzel W., Subject Clustering Analysis Based on ISI Category Classification, Journal of Informetrics, vol. 4, no. 2, Apr. 2010, pp. 185-193.

  Liu X., Yu S., Moreau Y., De Moor B., Glänzel W., Janssens F., Hybrid Clustering by Integrating Text and Citation based Graphs in Journal Database Analysis, in Proc. of the 2009 IEEE International Conference on Data Mining Workshops (ICDMW2009), Miami, Florida, Dec. 2009, pp. 521-526.

  Liu X., De Moor B., Glänzel W., A Hierarchical and Optimal Clustering of the WoS Journal Database by Hybrid Information, in Proceedings of 13th International Conference on Scientometrics and Informetrics (ISSI2011), Durban, South Africa, July 2011, pp. 485-496.

- **Information fusion and vertical search by multi-view text mining**. We extend the multi-view text mining concept from multiple controlled vocabularies to multiple publishing time periods, weighting schemes and even subjects. These multiple perspectives allow us to obtain knowledge by a specific point of view or integrate multiple views for joint analysis. The software of Text Prior is developed to implement the information fusion and vertical search for gene retrieval. This contribution will be discussed in Chapter 6.

  The related paper:

  Liu X., Gevaert O., Tranchevent L., Moreau Y., De Moor B., A Web Portal for Multi-view Text Mining and Vertical Searches, submitted to BMC Bioinformatics.

Chapter 2

Multi-view clustering by tensor methods

2.1 Introduction

In many real-world problems, objects can be described by multiple sets of features. For example, in scientific literature mining, both the textual content and the citation link between articles are often used in the knowledge discovery processes [91]. In multiplex network analysis, we are given a set of multiple networks that share the same nodes but possess network-specific links representing different types of relationships between nodes [99]. A particular instance of this scenario is the social network of university students, which may include symmetrized connections from (i) Facebook friendship, (ii) picture friendship, (iii) roommate relations, and (iv) student housing-group preference. These diverse individual activities result in multiple relationship networks among students. Such a learning scenario is called multi-view learning, since each feature set describes a view of the same set of underlying objects. A simple approach to learn from these multi-view data is to learning from each view separately. However, such approaches fail to account for the complementary information encoded into different views.

Multi-view clustering refers to the clustering of the same class of entities with multi-view representations, either from various information sources or from different feature generators. Compared with the clustering that is implemented on single-view data, multi-view clustering is expected to obtain robust and novel partitioning results by exploiting the complementary information in

different views. One of the recent developments in clustering is the spectral clustering technique, which has seen an explosive proliferation over the past few years [135]. Among many other factors, such as easy implementation and efficiency, one of the key advantages of spectral clustering is that it is based on the relaxation of a global clustering criterion (i.e., normalized cuts). Spectral clustering has been widely employed in many real applications, from image segmentation to community detection. Although spectral clustering [94] works well on single-view data, it is not well suited for the presentation of multi-view data, since it is inherently based on matrix decompositions.

Recently, several multi-view clustering algorithms have been proposed [5, 12, 27, 34, 91, 92, 127, 131, 150]. These multi-view clustering techniques have been shown to yield better performance in comparison to single-view techniques. However, as we will discuss in the related work, the limitations of some algorithms are apparent. For instance, some techniques assume that the dimensions of the features in multiple views are the same, limiting their applicability to the homogeneous settings. Some other techniques only concentrate on the clustering of two-view data so that it might be hard to extend them to more than a two-view situation [12]. In addition, an appropriate weighting scheme is lacking for these multiple views although coordinating various information from them is also one crucial step in gaining good clustering results [127, 132]. A unified framework that can integrate various types of multi-view data is lacking [92, 131].

Traditionally, tensor-based methods have been used to model multi-view data [76]. Tensors are higher-order generalizations of matrices, and some tensor methods are very powerful to analyze the latent pattern hidden in the multi-view data. Tensor decompositions [31, 77] capture multilinear structures in higher-order data-sets, where the data have more than two modes. Tensor decompositions and multi-way analysis allow naturally to extract hidden (latent) components (cluster structure) and investigate complex relationship among them. Tensors have been successfully applied to several domains, such as chemometrics, signal processing, Web search, data mining, scientific computing and bioinformatics [3, 8, 38, 76, 77, 101, 110, 122, 128].

In this Chapter, we propose a multi-view clustering framework based on tensor methods. Our formulations model the multi-view data as a tensor and seek a joint latent optimal subspace by tensor analysis. Our framework can leverage the inherent consistency among multi-view data and integrate their information seamlessly. Apart from other multi-view clustering strategies, which are usually devised for ad hoc application, our method provides a general framework in which some limitations of prior methods are overcome systematically. In particular, our framework can be extended to various types of multi-view data. Almost any multiple similarity matrices of the same

entities are allowed to be embedded into our framework. In addition, since our framework can obtain a joint optimal subspace, it can be easily extended to other related machine learning tasks, such as classification, spectral embedding and collaborative filtering. Our framework consists of two novel algorithms: multi-view clustering based on optimization integration (MC-OI) and that based on matrix integration (MC-MI). In particular, MC-MI can assign each view a suitable weight to boost the clustering. For each strategy, we provide two tensor based solutions. In fact, just as the other variants of PCA in machine learning applications [144], our strategy can be taken as a multi-view PCA analysis.

As an illustrative example of synthetic data shown in Figure 2.1, we intend to use it to compare three partitioning strategies to show the power of our tensor based multi-view partitioning. There are two groups of data points in a 3-D space, suppose based on our limited measurements (such as 2-D cameras in the real world), only the 2-D projection information of these data points could be observed (such as, X-Y projection, Y-Z projection and X-Z projection in a 3-D X-Y-Z coordinate system). We call each of the three 2-D projection data, single-view data. We can find the group information by adopting a spectral partitioning on each of these three single-view data. As shown in the upper right part of Figure 2.1, from each spectral projection of single-view data, there is significant overlap among these two groups of data points. Obviously, we could not recover the cluster structure only by each single-view data.

On the other hand, a natural idea is to integrate these three-view observations for joint partitioning, that is multi-view clustering. Thus, we investigate the spectral projection of two multi-view clustering strategies: multiple kernel fusion (MKF) and MC-OI based on MLSVD (MC-OI-MLSVD). As shown in the middle right part of Figure 2.1, the overlap among these two groups still remains by the spectral projection of MKF so that it is hard to get the correct group structure as well. Whereas, from the lower right part of Figure 2.1, the two groups are separated clearly by MC-OI-MLSVD and consequently a good group structure can be recovered from this three-view data. In this example, it shows that our tensor based multi-view clustering strategy is able to obtain the latent cluster structure hidden amid multi-view data.

To the best of our knowledge, our work is the first unified attempt to address multi-view clustering within the framework of tensor methods. The key contributions of our work can be summarized as follows:

- We propose to model multi-view data as a tensor and develop a new framework of multi-view clustering by tensor methods.

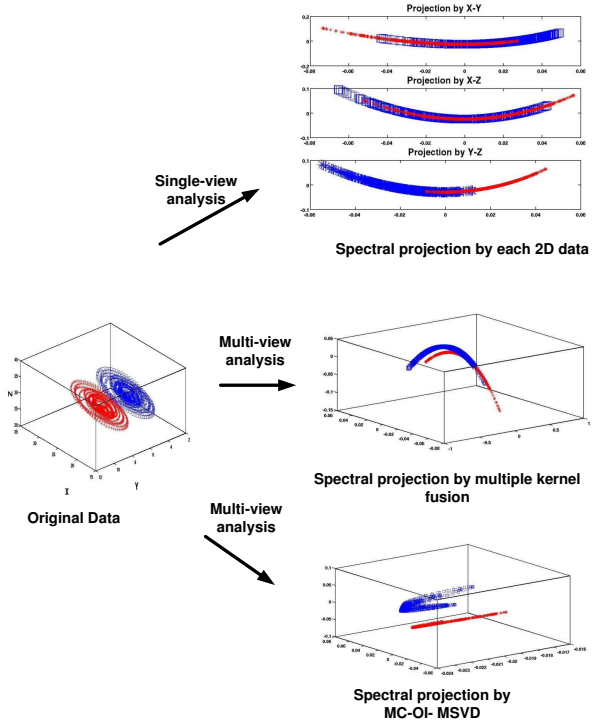

Figure 2.1: Comparison of single-view projection versus multi-view projection.

- We present two novel multi-view clustering strategies with their tensor solutions.
- We systematically evaluate our methods on both a synthetic data set and two real applications.

The rest of the Chapter is organized as follows. To start, Section 2.2 reviews the related work. Then, Section 2.3 introduces the concepts of spectral clustering. Next, Section 2.4 presents our tensor based multi-view clustering algorithms. After that, Section 2.5 demonstrates the experimental results on synthetic data and practical applications. The related research issues are discussed in Section 2.6. Finally, we conclude in Section 2.7.

2.2 Related work

2.2.1 Multi-view clustering

Bickel and Scheffere [12] propose a multi-view clustering method that extends k-means and hierarchical clustering to deal with data with two conditionally independent views. A multi-view clustering strategy via canonical correlation analysis (CCA) is presented in [27]. This method assumes that the views are uncorrelated given the cluster label. The above algorithms only concentrate on the clustering of two-view data thus it might be hard to extend them to more than two-view situations. Meanwhile our strategy is applicable to any multi-view situation. Long *et al.* [92] formulate a multi-view spectral clustering method while investigating multiple spectral dimension reduction. A clustering method based on linked matrix factorization is introduced to fuse information from multiple graphs in [132]. Zhou *et al.* [150] develop a multi-view clustering strategy via generalizing the normalized cut from a single view to multiple views and subsequently they build a multi-view transductive inference. In the above algorithms, a common problem is that the analysis of inherent relationship among multi-view data might be neglected. While in our tensor based strategy, the multilinear relationship among multi-view data is taken into account.

2.2.2 Community detection of multi-view networks

Tang *et al.* propose the concept of feature integration to implement the clustering of multi-view social networks [131]. Based on modularity optimization, Mucha *et al.* [99] develop a generalized framework of network quality functions that allow studies of community structure in a general setting encompassing networks that evolve over time, have multiple types of links (multiplexity), and have multiple scales. These methods are applicable to specific type of data with sparse links while our strategy is devised for general data.

2.2.3 Kernel fusion and clustering ensemble

Multiple kernel learning aims at finding a combination of kernels to optimize for classification or clustering [74, 91]. Such a solution might sound natural, but its underlying principal is not clear [150]. In addition, the heavy computation of their convex optimization makes them only applicable to small databases [91]. Meanwhile, with the recent research progress in tensor decomposition [121], our strategy has the potential to tackle large-scale databases. Clustering

ensemble is also known as clustering aggregation or consensus clustering, which integrates different partitions into a consolidated partition with a consensus function [5, 127]. However, clustering ensemble methods usually concentrate on single-view data to overcome the drawback of k-means. In fact, clustering ensemble is embedded into our strategy to facilitate the final partitioning.

2.2.4 Tensor based clustering

Sun *et al.* [128] introduce a dynamic tensor analysis (DTA) algorithm and its variants, and apply them to anomaly detection and multi-way latent semantic indexing. It seems their clustering method is designed for dynamic stream data. Dunlavy *et al.* [38] apply Parallel Factor Analysis (PARAFAC) decomposition for analyzing scientific publication data with multiple linkage. Selee *et al.* create a new tensor decomposition called Implicit Slice Canonical Decomposition (IMSCAND) to group information when multiple similarities are known [122]. The last two ideas that integrate multi-view data as a tensor are similar to ours. But our methods rely on a Tucker-type tensor decomposition. Furthermore, in these methods, all single-view data is considered equally important, while we will present a technique that compute weights for the different views.

2.3 Spectral clustering

Spectral clustering was originally derived based on relaxation of the normalized cut formulation for clustering [123]. In particular, spectral clustering involves a matrix trace optimization problem [94, 108]. We show in this Chapter that the spectral clustering formalism can be extended to deal with multi-view problems based on tensor computations.

Given a set of N data points $\{x_i\}_{i=1}^N$ where $x_i \in \mathbb{R}^p$ is the ith data point (p is the number of feature dimensions), a similarity $s_{ij} \geq 0$ can be defined for each pair of data points x_i and x_j based on some similarity measure. An intuitive way to represent this data set is using a graph $\mathtt{G} = (\mathtt{V}, \mathtt{E})$ in which the vertices $\mathtt{V}$ represent the data points and the edges $e_{ij} \in \mathtt{E}$ characterize the similarity between data points quantified by s_{ij}. Usually, the similarity measure is symmetric, and the graph is undirected. The affinity matrix of the graph $\mathtt{G}$ is the matrix $\mathbf{S}$ with the ijth entry $\mathbf{S}_{ij} = s_{ij}$. The degree of the vertex

d_i, defined as

$$d_i = \sum_{j=1}^{N} s_{ij}, \tag{2.1}$$

is the sum of all the weights of edges connected to d_i. The degree matrix $\mathbf{D}$ is a diagonal matrix containing the vertex degrees $d_1, ..., d_N$ on the diagonal. It follows from the spectral embedding formalism [94, 108, 123] that the Laplacian matrix is defined as $\mathbf{L} = \mathbf{D} - \mathbf{S}$, and the normalized Laplacian matrix, corresponding to the normalized cuts ($NCut$), is defined as

$$\mathbf{L}_{NCut} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2} = \mathbf{I} - \mathbf{S}_N, \tag{2.2}$$

where $\mathbf{S}_N$ is the normalized similarity matrix and defined as

$$\mathbf{S}_N = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}. \tag{2.3}$$

The matrices $\mathbf{S}_N$ and $\mathbf{L}_{NCut}$ have the same eigenvectors, and their eigenvalues are related as $\lambda^{(\mathbf{S}_N)} = 1 - \lambda^{(\mathbf{L}_{NCut})}$, where $\lambda^{(\mathbf{S}_N)}$ and $\lambda^{(\mathbf{L}_{NCut})}$ are the eigenvalues for $\mathbf{S}_N$ and $\mathbf{L}_{NCut}$, respectively.

2.3.1 Single-view spectral clustering

We first consider spectral clustering in the single-view setting [94]. Suppose $\mathbf{U} \in \mathbb{R}^{N \times K}$ is the relaxed assignment matrix, where N is the number of data points and K is the number of clusters. The spectral clustering problem can be expressed as

$$\begin{aligned} &\min_{\mathbf{U}} \text{trace}(\mathbf{U}^T\mathbf{L}_{NCut}\mathbf{U}), \\ &\text{s.t. } \mathbf{U}^T\mathbf{U} = \mathbf{I}. \end{aligned} \tag{2.4}$$

It follows from the Ky Fan theorem [112] that the optimal solution to the problem in (2.4) is the top K eigenvectors of $\mathbf{L}_{NCut}$. Considering the relationship between $\mathbf{S}_N$ and $\mathbf{L}_{NCut}$, the spectral clustering formulation can also be expressed as

$$\begin{aligned} &\max_{\mathbf{U}} \text{trace}(\mathbf{U}^T\mathbf{S}_N\mathbf{U}), \\ &\text{s.t. } \mathbf{U}^T\mathbf{U} = \mathbf{I}. \end{aligned} \tag{2.5}$$

Since $\mathbf{S}_N$ is positive semi-definite, the spectral clustering can be re-formulated as a Frobenius norm optimization problem as follows:

$$\begin{aligned} &\max_{\mathbf{U}} \|\mathbf{U}^T\mathbf{S}_N\mathbf{U}\|_F^2, \\ &\quad \text{s.t. } \mathbf{U}^T\mathbf{U} = \mathbf{I}. \end{aligned} \tag{2.6}$$

The objective functions in (2.5) and (2.6) are different, but they happen to have the same optimal solution, namely, the columns of a matrix $\mathbf{U}$ span the dominant eigenspace of $\mathbf{S}_N$.

2.3.2 Multi-view spectral clustering

Given multi-view data, the clustering result could be improved if the multiple views are integrated in an appropriate way. We have the following two strategies to integrate the multi-view data in the context of spectral clustering. Our multi-view partitioning strategies are expected to capture the complementary information conveyed in different views so that they are able to achieve better or robust clustering results.

Multi-view clustering by optimization integration (MC-OI)

Based on the spectral partitioning of each single-view data, the first strategy is to integrate the objective functions of individual partitions from each single-view data. In particular, we consider the optimization of multi-view clustering by simply adding individual objective functions as in

$$\begin{aligned} &\max_{\mathbf{U}} \sum_{v=1}^{V} \|\mathbf{U}^T\mathbf{S}_N^{(v)}\mathbf{U}\|_F^2, \\ &\quad \text{s.t. } \mathbf{U}^T\mathbf{U} = \mathbf{I}, \end{aligned} \tag{2.7}$$

where $\mathbf{S}_N^{(v)}$ is the normalized similarity matrix for the vth view and $\mathbf{U}$ is the common factor shared by multiple views.

Multi-view clustering by matrix integration (MC-MI)

The second multi-view clustering strategy is to combine the normalized similarity matrices from different views, leading to the following integrated

similarity matrix as

$$\tilde{\mathbf{S}} = w_1\mathbf{S}_N^{(1)} + w_2\mathbf{S}_N^{(2)} + ... + w_V\mathbf{S}_N^{(V)}, \tag{2.8}$$

where w_v are the weights of each view and $W = [w_1, w_2, \cdots, w_V]^T$. The multi-view clustering based on $\tilde{\mathbf{S}}$ can be formulated as follows:

$$\begin{aligned} &\max_{\mathbf{U}, w_v} \|\mathbf{U}^T\tilde{\mathbf{S}}\mathbf{U}\|_F^2, \\ &\text{s.t. } \mathbf{U}^T\mathbf{U} = \mathbf{I},\ w_v > 0 \text{ and } \sum_{v=1}^{V} w_v^2 = 1, \end{aligned} \tag{2.9}$$

where the unknown weighting factors w_v play a crucial role in the above optimization. Once w_v are determined, MC-MI can be handled as a common spectral clustering problem defined in (2.6). In addition, weighting factors w_v can also be considered as the weights (contribution) of different views during joint partitioning.

2.4 Multi-view spectral clustering via tensor methods

Following the two multi-view clustering strategies discussed above, we present the tensor-based solutions in this section. Compared to the single-view spectral clustering, which is solved by matrix decomposition, we formulate our multi-view clustering by tensor decomposition. The overview of the tensor-based method is depicted in Figure 2.2. As shown in the left part of Figure 2.2, the goal of single-view spectral clustering is to find an optimal latent subspace from single-view data. In contrast, with multi-view data, we want to obtain a joint optimal subspace with the aid of tensor methods.

2.4.1 Background on tensors

We provide some basic background on tensors and their decompositions in the following. We refer the readers to [31, 32, 77] for more detailed treatment on this topic. A tensor is a multidimensional array [77]. The order of a tensor is the number of modes (or ways). A first-order tensor is a vector, a second order tensor is a matrix and a tensor of order three or higher is called a higher-order tensor. We only investigate third-order tensor methods that are relevant to our problem.

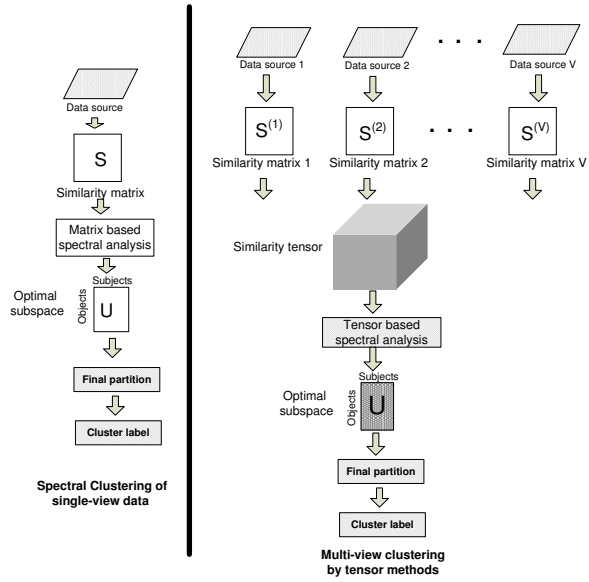

Figure 2.2: Comparison between single view (left) and multi-view (right) spectral clustering.

Matrix unfolding is the process of re-ordering the elements of a 3-way array into a matrix. The n-mode ($n = 1, 2, 3$) matrix unfoldings of a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ are denoted by $\mathbf{A}_{(1)}, \mathbf{A}_{(3)}$ and $\mathbf{A}_{(3)}$ separately. For example, the matrix unfolding $\mathbf{A}_{(1)}$ is a matrix with the number of rows I_1 and the number of its columns is the product of dimensionalities of all other modes, that is, $I_2 \times I_3$. The matrix unfolding of a third-order tensor is illustrated in Figure 2.3.

A tensor can be multiplied by a matrix. Consider a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ and a matrix $\mathbf{B} \in \mathbb{R}^{J_1 \times I_1}$, $\mathbf{C} \in \mathbb{R}^{J_2 \times I_2}$, $\mathbf{D} \in \mathbb{R}^{J_3 \times I_3}$, then the 1-mode product $(\mathcal{A} \times_1 \mathbf{B})$, 2-mode product $(\mathcal{A} \times_2 \mathbf{C})$ and 3-mode product $(\mathcal{A} \times_3 \mathbf{D})$ are defined by

$$(\mathcal{A} \times_1 \mathbf{B})_{j_1 i_2 i_3} = \sum_{i_1=1}^{I_1} a_{i_1 i_2 i_3} b_{j_1 i_1}, \ \forall j_1, i_2, i_3, \tag{2.10}$$

$$(\mathcal{A} \times_2 \mathbf{C})_{i_1 j_2 i_3} = \sum_{i_2=1}^{I_2} a_{i_1 i_2 i_3} c_{j_2 i_2}, \ \forall i_1, j_2, i_3, \tag{2.11}$$

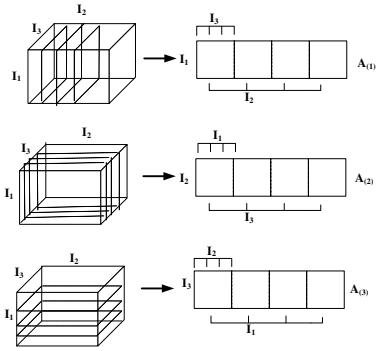

Figure 2.3: Matrix unfolding of a third-order tensor

$$(\mathcal{A} \times_3 \mathbf{D})_{i_1 i_2 j_3} = \sum_{i_3=1}^{I_3} a_{i_1 i_2 i_3} d_{j_3 i_3}, \ \forall i_1, i_2, j_3, \tag{2.12}$$

respectively.

Multilinear singular value decomposition (MLSVD) [31, 134] is a form of higher-order extension of matrix Singular Value Decomposition (SVD). It decomposes a tensor into a core tensor multiplied by a matrix along each mode. In the three-way case where $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, we have

$$\mathcal{A} = \mathcal{B} \times_1 \mathbf{U} \times_2 \mathbf{V} \times_3 \mathbf{W}, \tag{2.13}$$

where $\mathbf{U} \in \mathbb{R}^{I_1 \times I_1}$, $\mathbf{V} \in \mathbb{R}^{I_2 \times I_2}$ and $\mathbf{W} \in \mathbb{R}^{I_3 \times I_3}$ are called factor matrices or factors and can be thought of as the principal components of the original tensor along each mode. The factor matrices $\mathbf{U}, \mathbf{V}$ and $\mathbf{W}$ are assumed to be column-wise orthonormal. The tensor $\mathcal{B} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ is called the core tensor. In MLSVD, $\mathcal{B}$ has a very specific structure, namely, it satisfies "all-orthogonal" and "ordering" constraints, see [31]. The elements of $\mathcal{B}$ show the level of interaction between different components. According to [31], given a tensor $\mathcal{A}$, its matrix factors $\mathbf{U}, \mathbf{V}$ and $\mathbf{W}$ as defined in (2.13) can be computed as the left singular vectors of its matrix unfoldings $\mathbf{A}_{(1)}, \mathbf{A}_{(2)}$ and $\mathbf{A}_{(3)}$ respectively. The tensor approximation by truncating the decomposition is named truncated MLSVD. A decomposition as (2.13), with or without constraint, is also known as a Tucker decomposition [134].

2.4.2 Tensor construction

As aforementioned in the introduction section, multi-view data can be naturally modelled as a tensor. The construction of a tensor is a key step to devise our multi-view clustering algorithm. There are several options for constructing a tensor with multi-view data. In [62], a tensor is constructed by stacking the object-by-feature matrices derived from multiple views in a tensor as shown in the left part of Figure 2.4. Omberg *et al.* adopt an analogous scheme to formulate an integrative framework for joint analysis of DNA microarray data from different studies [110]. This kind of tensor construction is only applicable to the scenario of homogeneous data sources, where the dimensions of different feature spaces are the same. In fact, many multi-view applications deal with heterogeneous data sources in which the dimensions of various feature spaces are different. For instance, in our later application to scientific publication analysis, the citation feature space has the dimension of 8,305 while the dimension of the text feature space is more than 600,000.

Consequently, we prefer a construction that is independent of data dimension, thereby enabling the integration of heterogeneous data sources. Based on this motivation, we propose to build a tensor $\mathcal{A}$ from the multiple similarity matrices $\{\mathbf{S}_N^{(1)}, \mathbf{S}_N^{(2)}, \cdots, \mathbf{S}_N^{(V)}\}$ derived from multiple views as the frontal slices. In this research, we call this type of tensor similarity tensor. Different from the former data tensor, this similarity tensor is partially symmetric. The first and the second dimensions I_1 and I_2 of the tensor $\mathcal{A}$ are equal to the corresponding dimension of the similarity matrices $\mathbf{S}_N^{(v)}, (v = 1, \ldots, V)$, and its third dimension V equals the number of multiple views (different similarity matrices). The construction of a similarity tensor is illustrated in the right part of Figure 2.4. Since the similarity of each view is computed in different spaces, the normalization of each similarity matrix is required. Indeed, our definition of similarity matrix in (2.3) could be regarded as a normalization step.

2.4.3 MC-OI by tensor methods

We first discuss the optimization integration approach for multi-view clustering. Suppose a similarity tensor $\mathcal{A}$ is built from similarity matrices $\mathbf{S}_N^{(v)} \in \mathbb{R}^{N \times N}$ $(v = 1, \ldots, V)$, the integration of spectral optimization can be written as

$$\sum_{v=1}^{V} \|\mathbf{U}^T \mathbf{S}_N^{(v)} \mathbf{U}\|_F^2 = \|\mathcal{A} \times_1 \mathbf{U}^T \times_2 \mathbf{U}^T \times_3 \mathbf{I}\|_F^2, \tag{2.14}$$

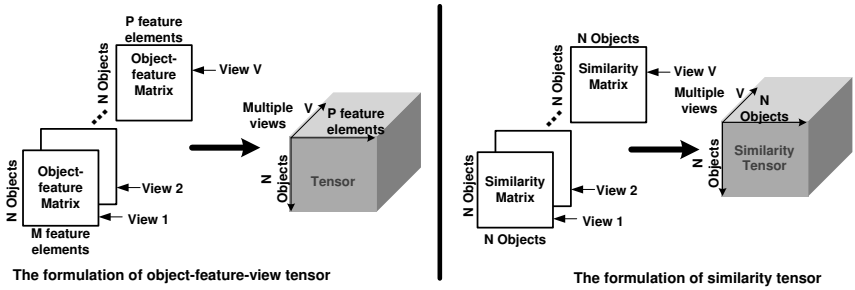

Figure 2.4: Comparison of different formulations of multi-view learning using tensor methods.

where the column space of $\mathbf{U} \in \mathbb{R}^{N \times K}$ is the joint optimal subspace and $\mathbf{I} \in \mathbb{R}^{V \times V}$ is an identity matrix. The spectral decomposition of the similarity tensor $\mathcal{A}$ in this case can be illustrated as in Figure 2.5.

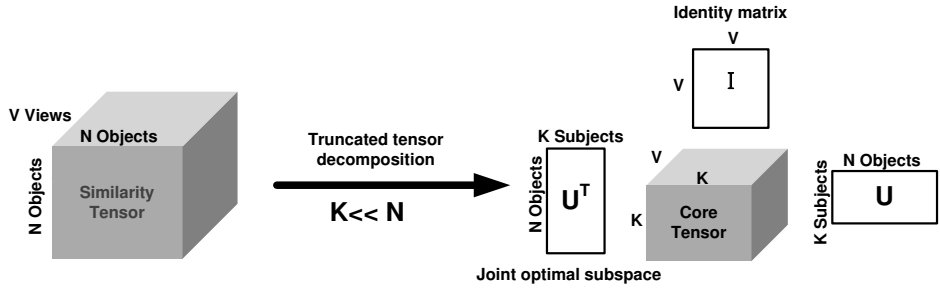

Figure 2.5: Illustration of multi-view clustering by optimization integration using tensor decomposition.

The optimization of multi-view clustering in (2.7) can be re-formulated based on tensor computation as

$$\max_{\mathbf{U}} \|\mathcal{A} \times_1 \mathbf{U}^T \times_2 \mathbf{U}^T \times_3 \mathbf{I}\|_F^2, \quad \text{s.t. } \mathbf{U}^T\mathbf{U} = \mathbf{I}. \tag{2.15}$$

This optimization can be approximated by MLSVD and we call this method multi-view clustering by optimization integration based on MLSVD (MC-OI-MLSVD). As explained in [31], projection by MLSVD on the dominant higher-

order singular vectors usually gives a good approximation of the given tensor. Consequently, we propose to take the columns of $\mathbf{U}$ to be the dominant 1-mode singular vectors. Our experimental results show that this solution usually leads to satisfactory performance. The dominant 1-mode singular vectors of $\mathbf{U}$ are equal to the dominant left singular vectors of $\mathbf{A}_{(1)}$. The truncated MLSVD obtained this way does not maximize (2.7) in general. However, the result is usually satisfactory and the algorithm is efficient and easy to implement. Moreover, there exists an upper bound on the approximation error as shown in [31]. The pseudo code of MC-OI-MLSVD is presented as follows:

Algorithm 2.4.1: MC-OI-MLSVD($\mathbf{S}^{(1)}, \mathbf{S}^{(2)}, ..., \mathbf{S}^{(V)}, K$)

comment: K is the number of clusters

1. *Build a similarity tensor* $\mathcal{A}$
2. *Obtain the unfolding matrix* $\mathbf{A}_{(1)}$
3. *Compute* $\mathbf{U}$ *from the subspace spanned by the dominant left* K *singular vectors of* $\mathbf{A}_{(1)}$
4. *Normalize the rows of* $\mathbf{U}$ *to unit length*
5. *Calculate the cluster idx with* $k-$*means on* $\mathbf{U}$

return (*idx* : *the clustering label*)

Meanwhile, there exist other tensor based solutions. For example, this optimization can be solved by a tensor approximation method called higher-order orthogonal iteration (HOOI), which is an alternating least-squares (ALS) algorithm [32, 82].

The basic idea of HOOI is to solve the following maximization problem:

$$\begin{aligned} &\max_{\mathbf{U},\mathbf{V},\mathbf{W}} \|\mathcal{A} \times_1 \mathbf{U}^T \times_2 \mathbf{V}^T \times_3 \mathbf{W}^T\|_F^2, \\ &\text{s.t. } \mathbf{U}^T\mathbf{U} = \mathbf{I},\ \mathbf{V}^T\mathbf{V} = \mathbf{I} \text{ and } \mathbf{W}^T\mathbf{W} = \mathbf{I}. \end{aligned} \tag{2.16}$$

At each step, the estimate of one of the matrices $\mathbf{U}, \mathbf{V}, \mathbf{W}$ is optimized, while the other two are fixed. In order to maximize with respect to the unknown matrix $\mathbf{U}$, the objective function of (2.16) is treated as a quadratic expression in $\mathbf{U}$. It follows from

$$\|\mathcal{A} \times_1 \mathbf{U}^T \times_2 \mathbf{V}^T \times_3 \mathbf{W}^T\|_F^2 = \|\mathbf{U}^T(\mathbf{A}_{(1)}(\mathbf{V} \otimes \mathbf{W}))\|_F^2, \tag{2.17}$$

where the columns of $\mathbf{U} \in \mathbb{R}^{I \times P}$ build an orthonormal basis for the P-dimensional left dominant subspace of the column space of $\mathbf{A}_{(1)}(\mathbf{V} \otimes \mathbf{W})$, and the solution can be obtained from the SVD of $\mathbf{A}_{(1)}(\mathbf{V} \otimes \mathbf{W})$ [31]. The optimization with respect to $\mathbf{V}$ and $\mathbf{W}$ is performed in similar ways. Usually,

the optimization of HOOI is initialized by MLSVD. Thus, the resulting algorithm called MC-OI-HOOI, is presented as follows:

Algorithm 2.4.2: MC-OI-HOOI($\mathbf{S}^{(1)}, \mathbf{S}^{(2)}, ..., \mathbf{S}^{(V)}, K$)

Build a similarity tensor $\mathcal{A}$
Obtain the unfolding matrices $\mathbf{A}_{(1)}$, $\mathbf{A}_{(2)}$ *and* $\mathbf{A}_{(3)}$
comment: obtain an initial $\mathbf{U}_0$ and $\mathbf{V}_0$ by MLSVD
while $<$!*convergence* $>$
do $\begin{cases} step1 : \mathbf{U}_{i+1} \text{ in dominant subspace of} \\ \mathbf{A}_{(1)}(\mathbf{V}_i \otimes \mathbf{I}) \\ step2 : \mathbf{V}_{i+1} \text{ in dominant subspace of} \\ \mathbf{A}_{(2)}(\mathbf{U}_i \otimes \mathbf{I}) \end{cases}$
comment: i is the counter of iteration
Normalize the rows of $\mathbf{U}$ *to unit length*
Calculate the cluster idx with $k-means$ *on* $\mathbf{U}$
return (idx : *the clustering label*)

The tensor decomposition in both MC-OI-MLSVD and MC-OI-HOOI is in fact a kind of joint matrix compression as shown in Figure 2.6, where the truncated tensor decomposition in the upper part can be understood as the joint compression of a set of matrices in the lower part. Matrix $\mathbf{U}$ in both parts is the common factor shared among multi-view data, and a set of similarity matrices $\Lambda^{(1)}, \ldots, \Lambda^{(V)}$, which form the front slices of the core tensor in the upper part, capture the characteristics of each view.

2.4.4 MC-MI by tensor methods

Since the effect of multi-view data differs from each other, each of them can be assigned an appropriate weight for joint analysis. In this case, the spectral decomposition of the similarity tensor $\mathcal{A}$ can be illustrated in Figure 2.7, in which $\mathbf{U}$ denotes the joint optimal subspace that we want to compute, and $W = [w_1, w_2, \cdots, w_V]^T$ denotes the weights of each view. The objective function of multi-view clustering of matrix integration (MC-MI) can be written as

$$\|\mathbf{U}^T\tilde{\mathbf{S}}\mathbf{U}\|_F^2 = \|\mathbf{U}^T(\sum_{v=1}^{V} w_v \mathbf{S}_N^{(v)})\mathbf{U}\|_F^2 = \|\mathcal{A} \times_1 \mathbf{U}^T \times_2 \mathbf{U}^T \times_3 W^T\|_F^2. \quad (2.18)$$

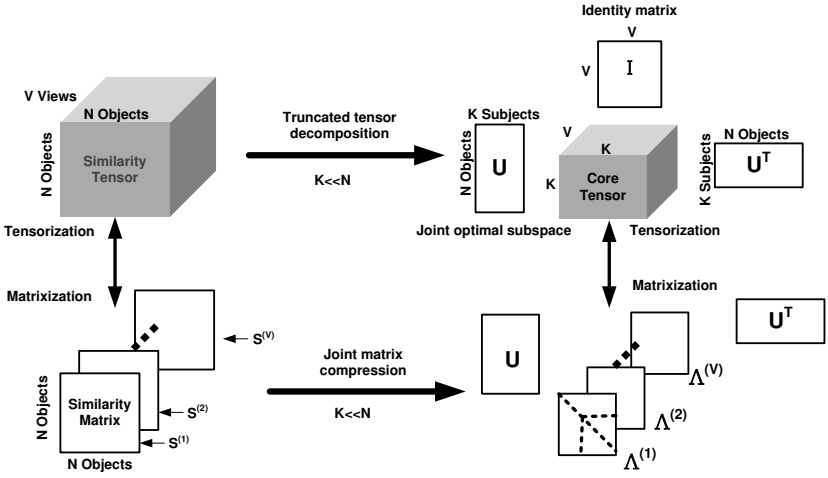

Figure 2.6: Illustration of the joint matrix decomposition for multi-view data.

Thus the optimization of multi-view clustering in (2.9) can be re-written as

$$\max_{\mathbf{U},W} \|\mathcal{A} \times_1 \mathbf{U}^T \times_2 \mathbf{U}^T \times_3 W^T\|_F^2, \; W = \begin{pmatrix} w_1 \\ \vdots \\ w_V \end{pmatrix} \tag{2.19}$$

$$\text{s.t. } \mathbf{U}^T\mathbf{U} = \mathbf{I}, \; \|W\|_F^2 = 1.$$

We can obtain the solution of this weighted multi-view clustering by HOOI. In addition, other algorithms can also be employed to solve this tensor based optimization problem, see [65] and reference herein.

Since the optimal HOOI solution usually leads to expensive computation, we simplify the main steps of the HOOI solution by an equivalent but efficient implementation. Taking the fact that W is not a matrix but a vector into account, we replace the optimization of the decomposed factors in each mode with the sum of matrices and its eigenvalue decomposition (EVD). The pseudo

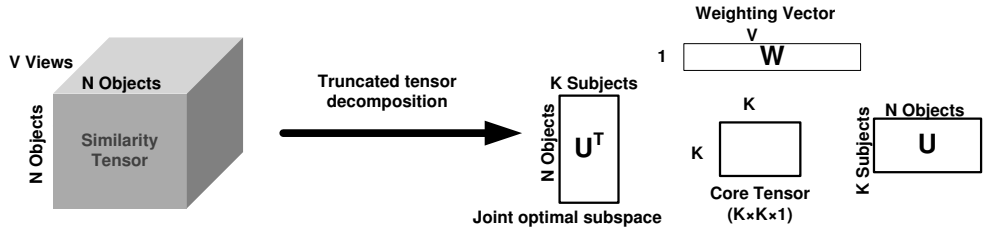

Figure 2.7: Illustration of multi-view clustering by matrix integration using tensor decomposition

code of MC-MI-HOOI is shown in the following,

Algorithm 2.4.3: MC-MI-HOOI($\mathbf{S}^{(1)}, \mathbf{S}^{(2)}, ..., \mathbf{S}^{(V)}, K$)

Build a similarity tensor $\mathcal{A}$
Obtain the unfolding matrices $\mathbf{A}_{(1)}$, $\mathbf{A}_{(2)}$ *and* $\mathbf{A}_{(3)}$
comment: obtain an initial $\mathbf{U}_0$ by MLSVD

while $<!convergence>$
do $\begin{cases} step1 : Calculating\ W_i\ as\ the\ dominant\ left\ singular\ vector\ of\ \mathbf{A}_{(3)}(\mathbf{U}_i \otimes \mathbf{U}_i) \\ step2 : Computing\ a\ new\ integration\ matrix\ \tilde{\mathbf{S}}\ by\ \sum_v^V w_v \mathbf{S}^{(i)} \\ step3 : Obtaining\ \mathbf{U}_{i+1}\ by\ eigenvalue\ decomposition\ of\ \tilde{\mathbf{S}} \end{cases}$
comment: i is the counter of iteration

Normalize the rows of $\mathbf{U}$ *to unit length*
Calculate the cluster idx with $k-means$ *on* $\mathbf{U}$
return (idx : *the clustering label*)

In the MC-OI framework, we discussed two alternative approaches, MC-OI-MLSVD and MC-OI-HOOI. In the present section, we only discussed MC-MI-HOOI. The reason is that tests indicated that in the MC-MI framework, mere truncation of the MLSVD, retaining only one vector in the third model, often yields unsatisfactory results.

2.5 Experimental Evaluation

In this section, we report experimental results of the proposed multi-view partitioning strategies in comparison with baseline multi-view clustering methods.

2.5.1 Baseline methods

The following six baseline multi-view clustering methods are employed for later comparison.

- Multiple kernel fusion (MKF): Joachims *et al.* [74] integrate different kernels by linear combination for hybrid clustering. The similarity matrix defined in (2.3) can be regarded as a linear kernel as well. In this research we adopt the average integration of multiple kernels, which actually is equal to the concatenation of the different normalized feature vectors from various single views [69].
- Feature integration (FI) [131]: With spectral analysis on each view, their structure features are extracted and then integrated together, on which SVD is implemented to obtain the cross-view principal components for clustering.
- Strehl's clustering ensemble algorithm (SA) [127]: Strehl & Ghosh formulate the optimal consensus as the partition that shares the most information with the partitions to combine. Three heuristic consensus algorithms (cluster-based similarity partition, hyper-graph partition and meta-clustering) based on graph partitioning are employed to obtain the combined partition.
- AdacVote [5]: Ayad & Kamel propose a cumulative vote weighting method to compute an empirical probability distribution summarizing the ensemble.
- CP-ALS [25, 57]: The CANDECOMP/PARAFAC (CP) decomposition is usually solved by an alternating least squares (ALS) algorithm, which we implement with a Matlab based tensor toolbox [6].
- Linked matrix factorization (LMF): In Tang's work [132], a quasi-Newton method named Limited memory BFGS (L-BFGS) is adopted for the optimization of LMF. We implement this algorithm with the aid of an optimization matlab toolbox named Poblano [37].

In our experiments, MC-OI-HOOI, MC-MI-HOOI, CP-ALS and LMF require initialization and parameter setting. The code of MLSVD and HOOI can be referred to the Matlab based tensor toolbox [6]. We develop our multi-view clustering algorithms by Matlab. Regarding CP-ALS, we adopt the default initialization and parameter setting as defined in the toolbox itself. All the other three algorithms are sensitive to the initialization, for example, LMF does not work under random initialization. So we initialize them by MLSVD

which usually provides a good initialization. On the other hand, all of these algorithms are not sensitive to the parameter setting and thus we choose their parameters as the default setting.

2.5.2 Performance measures

Regarding clustering evaluation, the data sets used in our experiments are provided with labels. Therefore the clustering performance is evaluated as compared to the automatic partitions with the labels using Adaptive Rand Index (ARI) [63] and Normalized Mutual Information (NMI) [127]. To evaluate the ARI and NMI performance, we set cluster number $M = 7$ on journal data and $M = 14$ on disease data.

In order to overcome the drawback of the k-means algorithm which is sensitive to various initializations, we adopt the combination of clustering ensemble of SA method and k-means as the final partitioning scheme for both spectral clustering and multi-view clustering. In particular, we first repeat each clustering method 50 times and deal with the 50 times clustering ensemble by SA method to obtain the final consensus partition. Consequently, the final partition by each clustering algorithm is unique.

2.5.3 Experiment on a synthetic multiplex network

We first evaluate and compare different clustering strategies applied to synthetic multi-view data. The synthetic data has 3 communties (clusters), which have 50, 100, 200 members respectively [130]. We can generate various views of interactions among these 350 vertices, that is, each view forms a network that shares the same vertices but has a different interaction pattern. For each view, group members connect with each other following a randomly generated within-group interaction probability. The interaction probability differs with respect to groups at distinct views. After that, we add some noise to the network by randomly connecting any two vertices with low probability. The different views demonstrate different interaction patterns. In this multi-view network which can also be called multiplex network according to [99], we construct four interaction matrices, each of whose elements is the interaction strength of a pair of vertices. The visualization of the four adjacency matrices is shown as Figure 2.8. They can then be used to construct a tensor. We apply spectral clustering to each single-view network.

In Table 2.1, we list the clustering evaluation on each single view data as well as those of multi-view clustering methods. First, it is clear that most multi-view

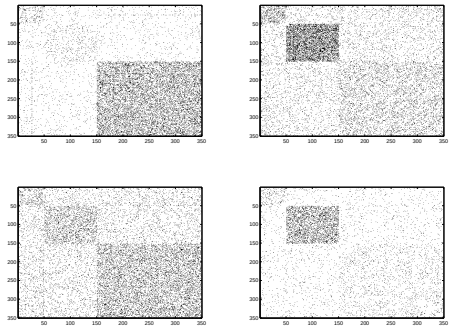

Figure 2.8: Visualization of the adjacency matrices of a synthetic multiplex network.

.

Table 2.1: Evaluation of clustering methods on a four view synthetic multiplex network

	Methods	NMI	ARI
Single View	S-A1	0.7605	0.7995
	S-A2	0.8928	0.9192
	S-A3	0.7198	0.8196
	S-A4	0.6318	0.5599
Multi View	MC-OI-MLSVD	*0.9321*	0.9508
	MC-OI-HOOI	0.9241	0.9509
	MC-MI-HOOI	**0.9431**	**0.9670**
	MKF	0.9156	0.9429
	FI	0.8893	0.9243
	SA	0.9251	*0.9540*
	AdacVote	0.8951	0.9400
	CP-ALS	0.5491	0.1274

clustering results are better than single-view clustering results. This could be easily explained by the patterns shown in Figure 2.8. The first view of the network (left above) only shows one group, and the fourth view (right below) involves another group with the other two groups hidden behind the noise. Thus, using single view is unlikely to recover the inherent cluster structure. This phenomenon is also validated by the low NMI as well as ARI of these

Table 2.2: The weighting coefficients of multi-view data by MC-MI-HOOI on synthetic data

Sources	Ranking of w_v	w_v	Performance ranking
A1	3	0.4725	3
A2	2	0.5288	1
A3	1	0.5643	2
A4	4	0.4433	4

two views. Applying multiple views helps reduce the noise and uncover the shared cluster structure. Second, compared with the other five baseline multi-view clustering strategies (since LMF works very bad on this data, we omit its comparison), our tensor based clustering methods perform better. In particular, MC-OI-MLSVD and MC-MI-HOOI are obviously superior to others based on both NMI and ARI evaluations.

To evaluate whether the optimized weights assigned on single-view data are correlated with their clustering performance, we compare the ranking of the weighting coefficients obtained by MC-MI-HOOI with the ranking of the corresponding clustering performance in Table 2.2. The ranking of the optimal weights of these multiple views is generally consistent with the ranking of their corresponding clustering performance. As shown, the top two largest coefficients correctly indicate the top two best individual data sources (A_2 and A_3).

In addition, as shown in Figure 2.8, the single-view data A3 with the largest weight contains the most information (three obvious clusters) while the single-view data A4 with the lest weight contains the lest information (one obvious cluster). The reason why our multi-view clustering algorithm of MC-MI-HOOI performs best might be due to the fact that it can leverages the effect of different single-view data in a reasonable way: during the joint clustering, the most informative view A3 plays the most part (the largest weight) while the lest informative view A4 plays lest part (the lest weight).

2.5.4 Application on scientific documents analysis

In this section, we apply our algorithms to the scientific analysis of the Web of Science (WoS) journal set. Our objective is to map these journals into different subjects by clustering algorithms.

Data description

Historically, bibliometric researchers have focused solely on citation analysis or text analysis, but not on both simultaneously. Recently, many researchers have applied text mining and citation analysis to the journal set analysis. The integration of lexical and citation information is a promising strategy towards better mappings [91]. We adopt a data set obtained from the WoS database by Thomson Scientific which contains articles, letters, notes and reviews from the years 2002 till 2006. To create a balanced benchmark data for evaluation, we select 7 categories consisting of 1424 journals. The titles, abstracts and keywords of the journal publications are indexed by a Jakarta Lucene based text mining program using no controlled vocabulary. The weights are calculated by four weighting schemes: Term Frequency (TF), Inverse Document Frequency (IDF), Term Frequency-Inverse Document Frequency (TF-IDF) and binary. Therefore, we have obtained four data sources as the lexical information of journals. These four kinds of text data are directly represented as similarity matrices. At the same time, four kinds of citation data represent link-based relationships among journals and consequently, from them, we construct corresponding affinity matrices, denoted as cross-citation, co-citation, bibliographic coupling and binary cross-citation. The detail of journal data are presented in Chapter 4.

We implement the proposed tensor based multi-view clustering methods to integrate multi-view data on journal data. To evaluate the performance, we also apply six popular multi-view clustering methods mentioned in Section V to integrate multi-view data. To verify whether the integration of multi-view data by tensor methods indeed improves the clustering performance, we first systematically compare the performance of all the individual data sources using spectral clustering. As shown in the left part of Table 2.3, the best spectral clustering is obtained on TFIDF data (NMI 0.7280, ARI 0.6601).

Afterwards, we also investigate the performance of integrating all single-view data using all compared multi-view clustering algorithms presented in the right part of Table 2.3. In particular, of all the algorithms we compared, the best performance is obtained by the MC-OI-HOOI method (NMI 0.7605, ARI 0.7262).

The comparison between the ranking of weighting coefficients by MC-MI-HOOI with the ranking of clustering performance is shown in Table 2.4. Because text and citation data are heterogeneous data sources, we compare each integration of them in their own feature spaces separately. In general, the ranking of the optimal weights obtained in this experiment is consistent with the ranking of their individual performance. For instance, within the citation feature space,

Table 2.3: Clustering performance on WoS journal database. S-BGC: Single-view clustering on Bibliographic coupling data.

SC-Algorithm	NMI	ARI	MC-Algorithm	NMI	ARI
S-TFIDF	**0.7280**	**0.6601**	MC-OI-MLSVD	0.7331	0.6615
S-IDF	*0.7020*	0.6422	MC-OI-HOOI	**0.7605**	**0.7262**
S-TF	0.6742	0.6305	MC-MI-HOOI	0.7287	0.6756
s-Binary-Text	0.6432	0.6022	MKF	0.7327	0.6787
S-cross-citation	0.6833	0.6057	FI	0.6944	0.6031
S-co-citation	0.6815	*0.6565*	SA	0.7226	0.6952
S-BGC	0.4398	0.3348	AdacVote	*0.7454*	*0.7176*
S-Binary-citation	0.5831	0.5238	CP-ALS	0.7042	0.6377
			LMF	0.5935	0.5058

Table 2.4: The weighting coefficients of multi-view data obtained by MC-MI-HOOI on journal data.

Text data	Ranking of w_v	w_v	Performance ranking
TFIDF	0.5890	1	1
IDF	0.4519	3	2
TF	0.5580	2	3
Binary-Text	0.3708	4	4
Citation	Ranking of w_v	w_v	Performance ranking
cross-citation	0.5372	2	2
co-citation	0.5771	1	1
Bibliographic coupling	0.5095	3	4
Binary-citation	0.3446	4	3

the top 2 largest coefficients correctly indicate the top 2 best individual data source (co-citation and cross-citation).

2.5.5 Experiment on disease gene clustering

Text mining helps biologists automatically collect disease-gene associations from large volumes of biological literature. Given a list of genes, we can generate a gene-by-term matrix by the retrieval from the MEDLINE database. We can also obtain multi-view gene-by-term matrices. The view represents a text

mining result retrieved by specific controlled vocabularies, hence multi-view text mining is featured as applying multiple controlled vocabularies to retrieve the gene-centric perspectives from free text publications. The clustering methods can be implemented on these genes to get the group information, which can be utilized for further disease analysis.

The data sets contain ten different gene-by-term text profiles indexed by ten controlled vocabularies. The original disease-related gene data set contains 620 genes which are known to be relevant to 29 diseases. To avoid the effect of imbalanced clusters which may affect the evaluation, we only keep the diseases that have 11 to 40 relevant genes. This data processing results in 14 genetic diseases and 278 genes. Because the present paper focuses on non-overlapping ("hard") clustering, we further remove 16 genes which are relevant to multiple diseases and 17 genes whose term vectors are empty for one of those ten vocabularies. The remaining 245 disease relevant genes are clustered into 14 clusters and biologically evaluated by their disease labels. For each vocabulary based gene-by-term data source, we create a similarity matrix using the value of the cosine similarity for two vectors. The details of disease gene data can be referred to Chapter 6.

At first, as shown in the left part of Table 2.5, the best clustering performance of individual data sources is obtained on LDDB text mining profile (NMI 0.7088, ARI 0.5942).

Afterwards, we also investigate the performance of integrating all single-view data using all compared multi-view clustering methods presented in the right part of Table 2.5. In particular, of all the methods we compared, the best performance is also obtained by the MC-OI-HOOI method (NMI 0.7732, ARI 0.6473) as the former experiment on journal data. The strategies with the next two best performances are still our tensor methods, MC-MI-HOOI (NMI 0.7494, ARI 0.6015) and MC-OI-MLSVD (NMI 0.7429, ARI 0.6030). All of our tensor methods are not only beyond spectral clustering results of any single-view data but also superior to the six baseline multi-view clustering methods, which demonstrates the power of our strategy.

In Table 2.6, we present the comparison between the ranking of weighting coefficients of multi-view data with the ranking of the corresponding clustering performance. As shown, the largest coefficient correctly indicates the best individual data source (LDDB), and also the smallest coefficient correctly indicates the worst individual data source (KO). As a whole, the optimal weights obtained in our experiments are consistent with the ranking of the corresponding performance.

In spectral clustering, by checking the "elbow" of the plot of eigenvalues of

Table 2.5: Clustering performance on disease data set.

SC-Algorithm	**NMI**	**ARI**	**MC-Algorithm**	**NMI**	**ARI**
S-GO	0.5367	0.3657	MC-OI-MLSVD	0.7429	*0.6030*
S-MeSH	*0.7072*	0.5134	MC-OI-HOOI	**0.7732**	**0.6473**
S-OMIM	0.6971	0.4901	MC-MI-HOOI	*0.7494*	0.6015
S-NCI	0.5153	0.3063	MKF	0.7002	0.5445
S-eVO	0.6048	0.3845	FI	0.6743	0.4830
S-KO	0.3187	0.1194	SA	0.7016	0.5495
S-LDDB	**0.7088**	*0.5942*	AdacVote	0.6093	0.5349
S-MP	0.6582	0.4962	CP-ALS	0.7241	0.5154
S-SNOMED	0.6819	*0.5205*	LMF	0.6058	0.4402
S-Uniprot	0.5692	0.3303			

Table 2.6: The weighting coefficients of multi-view data obtained by MC-MI-HOOI on disease data.

Sources	Ranking of w_v	w_v	Performance ranking
GO	9	0.2544	8
MeSH	7	0.2842	2
OMIM	4	0.2973	3
NCI	6	0.2931	9
eVO	3	0.3021	6
KO	10	0.2216	10
LDDB	1	**0.5303**	1
MP	2	*0.3113*	5
SNOMED	8	0.2713	4
Uniprot	5	0.2970	7

single-view data [94], it may provide a heuristic guess on the optimal cluster number. Analogously, regarding tensor based multi-view clustering, we also intend to explore the relationship between the optimal cluster number and 1-mode singular values of similarity tensors. In Figure 2.9, from each tensor of our three data sets, we plot the top 20 1-mode singular values. As shown in Figure 2.9, the "elbow" in synthetic data is quite obvious at the number of 2 or 3 (the real cluster number is 3). In journal data, the "elbow" is more likely to range from 2 to 10 (the real cluster number is 7). In disease data, the "elbow" is from 2 to 15 (the real cluster number is 14). It seems that checking the "elbow" of the plot of 1-mode singular values might also offer a heuristic estimation of the optimal cluster number for the tensor based multi-view clustering strategies. Moreover, as can be observed in Figure 2.10 and Figure 2.11, we also compare the 1-mode singular value curves using different tensors of journal data and gene-disease data. Where the tensors are generated from different number of views, for instance, in journal data, we generate different tensors by using various combinations from 2 to 7 views randomly. As shown, for each data, the 1-mode singular value plot is quite stable w.r.t. the different combination of multiple views.

To investigate the computational time, we benchmark our tensor based multi-view clustering algorithms with 6 different multi-view clustering methods on the two application data sets. As shown in Table 2.7, our three strategies (MC-OI-MLSVD, MC-OI-HOOI and MC-MI-HOOI) are efficient. For instance, they are faster than four multi-view clustering methods (SA, AdacVote, CP-ALS and LMF). Obviously, MC-OI-MLSVD is more efficient. On the other hand, compared with MKF and FI, which our three algorithms are behind, our proposed methods yield much better performance or more enriched information (the weighting factor of the individual sources). Meanwhile, the two clustering ensemble methods SA and AdacVote require more computation time since they involve the partitioning of each single-view data. Consequently, with number of views increasing, the computation of the clustering ensemble methods will become more and more intensive.

2.6 Discussion

Based on the clustering performance of the multi-view clustering strategies, MKF is efficient when compared with tensor based strategies. However, MKF only combines multiple kernels (similarity matrices) in a simple way using the average sum of multiple similarities. Such a simple combination neglects the discriminating capability of each kernel. Clustering ensemble methods (SA and AdacVote) rely on discrete hard clustering. Using only the final

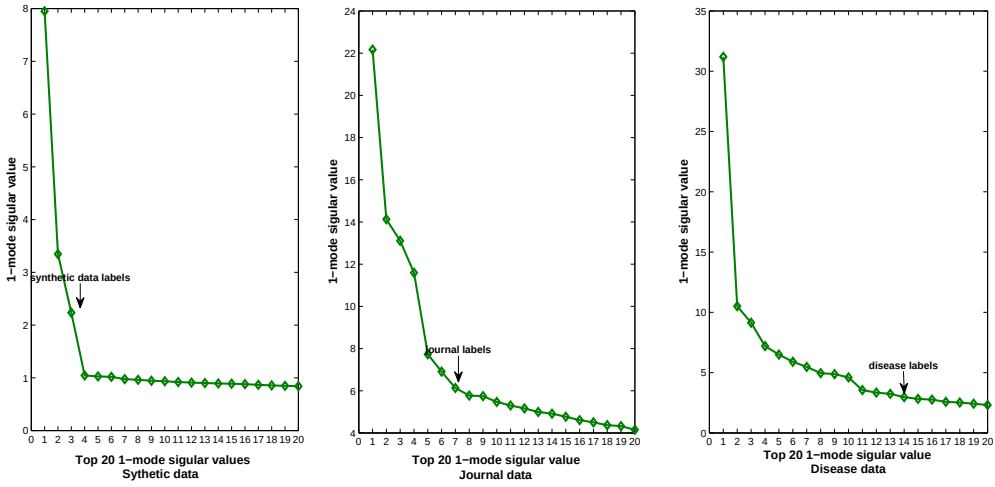

Figure 2.9: Plot of the top 20 1-mode singular values of tensors constructed from different multi-view data (synthetic data on the left, journal data on the middle and disease data on the right). All multi-view data within each data are employed to construct the relevant tensor.

Table 2.7: Comparison of CPU time of all multi-view clustering algorithms on real applications

Algorithm	disease data (seconds)	journal data (seconds)
MC-OI-MLSVD	4.82	33.34
MC-OI-HOOI	9.32	64.98
MC-MI-HOOI	2.79	41.75
MKF	1.23	3.97
FI	3.94	20.06
SA	37.29	60.94
AdacVote	37.31	44.67
CP-ALS	7.82	127.55
LMF	9.40	203.41

partition information seems too fragile to integrate. In addition, because the partitioning of every single-view data is required, the implementation of clustering ensemble methods is not efficient as shown in Table 2.7. Considering

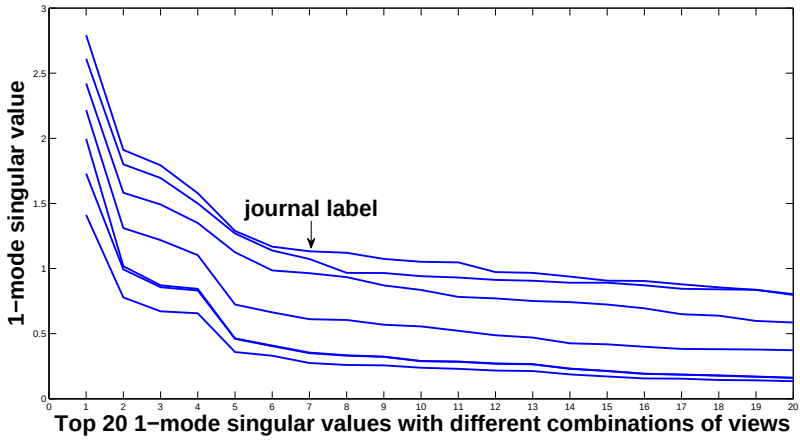

Figure 2.10: The plot of top 20 1-mode singular values by various tensor from journal data. These tensors are composed of various number of multi-view data from 2 views to 7 views. As shown, the 1-mode singular value curves are quite insensitive to which number of views are integrated.

LMF, we found that the clustering performance relies on the initialization, and hence the partitioning results are quite unstable. Moreover, its optimization mechanism consumes much time. For CP-ALS, the failure might be due to the un-orthogonal property of the relaxed assignment matrix $\mathbf{U}$ after tensor decomposition. The reason is that the similarity matrix in (2.3) we adopted to construct the tensor corresponds to the *Ncut* based Laplacian matrix which requires the orthogonal partition in spectral clustering.

Meanwhile, our tensor based multi-view spectral clustering can be thought of as a "Multi-view PCA" analysis, which integrates multi-view information seamlessly and forms a joint optimal subspace. Therefore our strategy can extract the latent pattern shared by all views and filter out irrelevant information or noise. The tensor based multi-view clustering by optimization integration strategy (MC-OI-MLSVD and MC-OI-HOOI) leverages the effect of each single-view data in an appropriate way while the tensor based multi-view clustering by matrix integration strategy (MC-MI-HOOI) is able to utilize the linear relationship of multi-view data for joint analysis.

One thing we need to emphasize is that, to some degree, our multi-view clustering relies on the complementary property of multi-view data. As shown in Figure 2.1 and Figure 2.8, as long as multi-view data owns enough

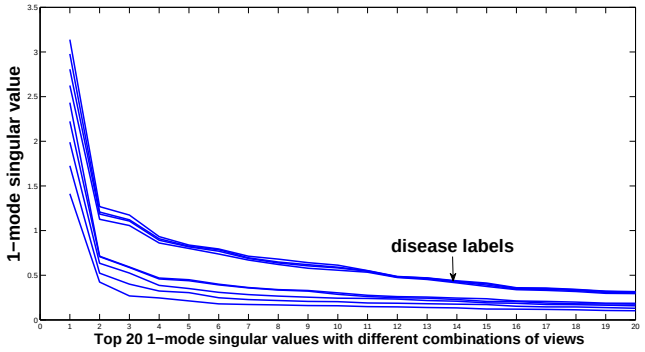

Figure 2.11: The plot of top 20 1-mode singular values by various tensor from disease gene data. These tensors are composed of various number of multi-view data from 2 views to 9 views. As shown, the 1-mode singular value curves are quite insensitive to which number of views are integrated.

complementary information and there is a clear cluster structure hidden amid them, our tensor based multi-view clustering methods are able to facilitate the detection of the latent cluster structure as expected. Although each single-view data may contain incomplete structure information or much noise.

In our former experiments, MC-OI-HOOI seems to work well to integrate multi-view heterogeneous data (data from different feature spaces) for joint clustering. For example, it works best to integrate ten-view data generated from different content (controlled vocabularies). Moreover, MC-OI-HOOI works best to integrate four views of text data with four views of citation data. The reason might be that MC-OI-HOOI provides an optimal common subspace shared by multi-view heterogeneous data by simultaneously analyzing them together. Meanwhile, MC-MI-HOOI appears to work well to integrate multi-view homogeneous data (data from the same feature space) for joint clustering because it utilizes the linear relationship of multi-view homogeneous data to boost the joint clustering. For example, MC-MI-HOOI works best to integrate our multi-view synthetic data as well. In fact, our synthetic data can be regarded as a kind of homogeneous data. Thanks to its fast computation and good clustering performance, our MC-OI-MLSVD is an efficient multi-view clustering strategy for integrating multi-view heterogeneous data.

2.7 Summary

We proposed a multi-view clustering framework based on high-order analogues of the matrix SVD and PCA. Our framework can be regarded as a multi-view extension of spectral clustering. By our tensor formulation, both heterogeneous information and homogeneous information can be integrated to facilitate the clustering task.

We presented two new multi-view clustering strategies: multi-view clustering by optimization integration (MC-OI) as well as by matrix integration (MC-MI). The relevant tensor based solutions are proposed, which are either iterative optimization or efficient approximation. All of them are capable of utilizing global information of multi-view data while taking the effect of single-view data into consideration. Furthermore, these different methods can be applied to various practical scenarios.

We employed our algorithms to both the synthetic data and two real applications. The clustering performance demonstrated that our algorithms are not only superior to single-view spectral clustering methods, but also superior to other baseline multi-view clustering methods.

In later research, we will carry out the following directions:

(1) We will investigate other alternative tensor solutions, such as Individual Difference in Scaling (INDSCAL) [25], as well as efficient tensor decomposition for scalable applications.

(2) We will extend our multi-view clustering algorithm to higher-order data (we only use three-order data in this research), such as, adding another temporal order that allows data to vary at different time points.

(3) Our framework is not limited to the clustering analysis. Since its core is to seek a joint optimal latent subspace, it can be extended to other multi-view learning tasks: for instance, classification, spectral embedding, collaborative filtering and even information retrieval.

Chapter 3

Optimal clustering and joint dimension reduction of multiple graphs

3.1 Introduction

The fast development of information technology allows us to observe the objects from different views as well as to collect these multi-view data. In this research, multi-view data refers to the same class of entities with multi-view representations. In many cases, the information provided by single-view data is insufficient to recover the inherent patterns due to the limited perspective of observation while multi-view data contains rich and complementary information, which allows us to obtain robust and better patterns.

Multi-view data is universal in a wide variety of applications. For example, two types of data are often used in journal database analysis: textual content and citation links, both of which describe the same journal entities but contain heterogeneous information. These two types of data are not entirely independent; they are actually closely correlated and supplement each other [73, 91]. Multi-view clustering of both data might provide a nice mapping of journal sets.

Another example is gene categorization in the biomedical applications. From the clinical experiments, we can obtain a description of genes while another

description can be provided by text mining from literatures. The integration of these two types of heterogeneous information for joint clustering is expected to get better understanding of gene groups.

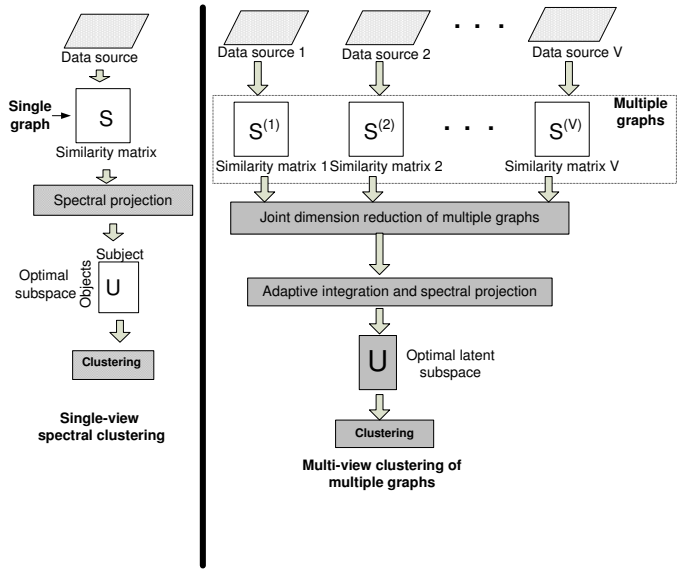

Figure 3.1: Conceptual overview of our multi-view clustering strategies based on simultaneous trace maximization

In practical spectral clustering, a graph is modeled from single-view data, while in our multi-view clustering setting, multiple graphs are extracted from multi-view data. In each graph, the nodes denote the instances that we are interested in, such as the documents and the genes in above examples.

Although spectral clustering works well in single-view data, it is not well suited to the presentation of multi-view data that might be better treated using nonlinear or multilinear methods. Therefore, by modeling multi-view data as multiple graphs, we extend spectral clustering from a multilinear algebra perspective.

In fact, many researchers turn to multi-view clustering that refers to the joint partitioning of multi-view data [12, 92, 127, 130, 150]. These multi-view clustering solutions might sound natural and even can achieve better performance. However, the limits of some algorithms are apparent. For instance, a simple and appropriate weighting scheme is required to fully utilize the inherent relationship of multi-view data (multiple graphs) [130, 132]. Some

multi-view clustering algorithms only work well in small-scale data sets [91]. Nevertheless, under practical scenarios, the number of instances is constantly huge, for example, there are nearly six million documents in the Web of Science (WoS) database from years 2002-2006. At the same time, with multiple views, the data volume becomes increasingly scalable, thus leading to the intensive computation of multi-view clustering. Hence, it is imperative to carry out the joint dimension reduction of multiple graphs before clustering, which has not yet been tackled by former research to our best knowledge.

Furthermore, some of these multi-view algorithms only concentrate on the clustering of using two-view data, and it is hard to extend them to a situation with more than two views [12, 27]. As a whole, even though the research of multi-view clustering has recently received considerable attention, it still seems to be at an early stage.

Therefore, to fully utilize the relationship of multi-view data (multiple graphs), we propose a clustering strategy that optimizes their multilinear relationship just by simply simultaneous trace maximization. Meanwhile, to handle the scalable clustering of multi-view data, we present a joint dimension reduction scheme for multiple graphs with the aid of multilinear singular value decomposition (MLSVD). Besides, our strategy is applicable to a situation with more than two views.

To the best of our knowledge, our work is the first unified attempt to address multi-view clustering by multilinear solution of simultaneous trace maximization together with the joint dimension reduction of multiple graphs.

Our work has three key contributions:

- By modeling multi-view data as multiple graphs, we propose a new framework of adaptive multi-view clustering by using simultaneous trace maximization (MC-STM), which is able to unravel the linear relationship among multi-view data in a simple way.
- With scalable data of multiple graphs, we put forward a joint dimension reduction scheme by MLSVD, which allows us to partition multi-view data efficiently.
- Based on the simultaneous trace maximization (STM) strategy, we formulate multi-view variants of other clustering strategies as well: multi-view clustering based on modularity optimization and multi-view k-means clustering.

3.2 Related work

The research on multi-view clustering, clustering ensemble and kernel fusion has been introduced in the Chapter 1. Other relevant work is discussed as follows.

Multi-view learning Zhu *et al.* [152] design a hybrid classification algorithm by carrying out a joint factorization on both the linkage adjacency matrix and the document-term matrix. Zhou *et al.* [151] devise a new document recommendation method by combining multiple graphs to measure document similarities; and according to the nature of different graphs, various factorization strategies are adopted. An integrated k-means-Laplacian (KL) clustering method is introduced in [136], which combines both k-means clustering on data attributes and spectral clustering on pairwise relations. This integrated KL clustering method performs well for a situation with two specific views (attribute data and relation data) and small scale data. Yu *et al.* [145] propose a clustering algorithm to combine multiple kernels and Laplacians and the coefficients of kernels and Laplacians can be optimized automatically, which formulation shares the similar flavor with our algorithms. However, that strategy works well on small databases, due to its heavy computation. In addition, based on multiple kernel learning, a strategy named heterogeneous feature machine is put forward for visual recognition [24]. Cai *et al.* propose a multi-modal spectral clustering with non-negative constrain to integrate heterogeneous image features for visual recognition [21].

Topic model To some degree, some strategies of topic model can be considered to be multiview clustering from a probabilistic perspective. For instance, PHITS-PLSA combines Probabilistic Hyperlink-Induced Topic Search (PHITS) with Probabilistic Latent Semantic Analysis (PLSA) for document clustering [20]. Erosheve *et al.* [40] combine Latent Dirichlet Allocation (LDA) with LDA-Link for network analysis. Nallapti *et al.* [102] combine the mixed membership stochastic block model with LDA, and extend the LDA-Link-Word model. A community detection method to combine the conditional link model and discriminative content model is presented in [141]. De Semt *et al.* [34] propose a unified probabilistic model to simultaneously model latent topics from bilingual corpora that discuss comparable content and use the topics as features in a cross-lingual, dictionary-less text categorization task. As known, one major problem with these topic model methods is that it is hard to implement them on data with high-dimension features. Hence, the limited scalability of topic model hinders its applications to a wider range. In addition, with the number of views increasing, it becomes difficult to apply these topic models.

Dimension reduction PCA is a well-known dimension reduction scheme;

nevertheless, it only works with vectorized representations of data. A two-dimensional PCA (2DPCA) algorithm is proposed for image processing [140], where an image is generally modeled as a matrix. Ye [143] presents data as a matrix and formulates an algorithm named generalized low rank approximations of matrices (GLRAM) by approximating a collection of matrices with a set of small-size matrices with lower rank. The formulation of that approximation algorithm is closely related to that of MLSVD. Lu *et al.* introduce a multilinear PCA (MPCA) framework for feature extraction of tensor based objects (video sequences) [93]. Wang and Ahuja [137] present a tensor based approximation approach for dimensionality reduction and apply that approach to object recognition (For example, a moving toy in video sequences). Different from these dimension reduction scenarios for vision or image data, based on MLSVD, we propose a joint dimension reduction scheme for multiple graphs, which is applicable to general data. In addition, dimension reduction by tensor methods has been used in signal processing and that scheme has been demonstrated to obtain an efficient implementation while accurate estimations [33]. Ishteva *et al* also mention a type of hierarchical Tucker format for joint dimension reduction [64].

3.3 Multi-view clustering based on spectral optimization

Regarding the formulation of spectral clustering, there are several types of Laplacian matrices to choose [94]. However, we just define our Laplacian matrix based on normalized cut (NCut). Then the optimization of spectral clustering is a maximization process, thus leading to the formulation of our multi-view clustering algorithms based on simultaneous trace maximization. Otherwise, we can not obtain our multi-view clustering formulation directly.

The single-view spectral clustering is formulated as Chapter 1, where $\mathbf{S}_N \in \mathbb{R}^{N\times N}$ is a NCut based similarity matrix; $\mathbf{U} \in \mathbb{R}^{N\times K}$ is a relaxed assignment (indicator) matrix; N is the number of instances and K is the number of clusters. From a single-view graph, we can formulate the spectral clustering by trace based optimization.

We expect that integrating multiple graphs is able to facilitate clustering tasks. By extending spectral clustering that is usually implemented on single-view data, we put forward a multi-view clustering framework by integrating the trace based clustering optimization of each graph.

From multi-view (V views) graphs, we can generate the corresponding

normalized similarity matrices $\mathbf{S}_N^{(v)}$ $(v = 1, 2, ..., V)$ accordingly. A natural idea is to link the trace based spectral optimization of each single-view graph for joint analysis,

$$\begin{aligned} &\max_{\mathbf{U}, w_v} \sum_{v=1}^{V} w_v \text{trace}(\mathbf{U}^T \mathbf{S}_N^{(v)} \mathbf{U}), \\ &\text{s.t. } \mathbf{U}^T \mathbf{U} = I, \ w_v > 0 \text{ and } \sum_{v=1}^{V} w_v^2 = 1, \end{aligned} \tag{3.1}$$

where w_v are the weighting factors of each view. It is hard to solve (3.1) directly, hence we will seek its solution by alternating optimization in next Section. Because our simultaneous trace maximization only involves norm-2 operation, we only limit our work to norm-2 based optimization. It is worthwhile pointing out that the weighting factors w_v can be interpreted in two ways from the multilinear algebra perspective. First, w_v can be considered to be the contribution of each single-view graph to the multi-view clustering. Second, w_v can be regarded as the linear coefficients of each similarity matrice to form the new integration matrix.

The comparison between single-view spectral clustering and our multi-view clustering is given in Figure 3.1. Similar to spectral clustering, the aim of our multi-view clustering is also to obtain the relaxed indicator matrix $\mathbf{U}^*$, the columns of which construct the optimal latent subspace.

3.4 Multi-view clustering via simultaneous trace maximization (MC-STM)

Since our multi-view clustering framework is based on trace optimization, we present an alternating least square (ALS) strategy named simultaneous trace maximization (STM) to solve the unknown parameters (the weighting vectors w_v and the relaxed indicator matrix $\mathbf{U}$) as defined in (3.1).

Given multiple similarity matrices $\mathbf{S}_N^{(1)}, .., \mathbf{S}_N^{(V)} \in \mathbb{R}^{N \times N}$, K is the cluster number and $K < N$, we attempt to maximize

$$f(W, \mathbf{U}) = \sum_{v=1}^{V} \sum_{k=1}^{K} w_v (\tilde{\mathbf{S}}_N^{(v)})_{kk}, \tag{3.2}$$

in which $\mathbf{U} \in \mathbb{R}^{N \times K}$ is orthonormal; $W \in \mathbb{R}^V$ is unit-norm; $\mathrm{W} = [w_1, w_2, .., w_V]^T$; $\tilde{\mathbf{S}}_N^{(v)} = \mathbf{U}^T \mathbf{S}_N^{(v)} \mathbf{U}, v = 1, ..., V$ and $(\tilde{\mathbf{S}}_N^{(v)})_{kk}$ denotes the kth

diagonal element. The matrix $\mathbf{S}_N^{(v)}$ are not necessarily positive (semi)definite. (3.2) will be written as

$$f(W, \mathbf{U}) = \sum_{v=1}^{V} w_v \text{trace}_K(\tilde{\mathbf{S}}_N^{(v)}). \tag{3.3}$$

Because both $\mathbf{U}$ and W are unknowns, (3.2) or (3.3) is not convex. Nevertheless, if we fix one of the unknowns, the optimization problem becomes quadratic and can be solved optimally. Thus, alternating least squares techniques swap between fixing $\mathbf{U}$ and fixing W. When $\mathbf{U}$ is fixed, the scheme recomputes W by solving a least-square problem, and vice verse. This optimization ensures that each step decreases (3.3) until convergence. We implement the simultaneous trace maximization as the following procedures.

3.4.1 Calculating the weighting vector W

Given $\mathbf{U}$, the objective function of $f(\bullet)$ in (3.3) can be written as

$$\begin{aligned} f(W) &= \sum_{v=1}^{V} w_v \sum_{k=1}^{K} (\tilde{\mathbf{S}}_N^{(v)})_{kk}, \\ &= W^T (\sum_{k=1}^{K} (\tilde{\mathbf{S}}_N^{(1)})_{kk}, ..., \sum_{k=1}^{K} (\tilde{\mathbf{S}}_N^{(v)})_{kk})^T. \end{aligned} \tag{3.4}$$

Suppose $Y = (\sum_{k=1}^{K} (\tilde{S}_N^{(1)})_{kk}, ..., \sum_{k=1}^{K} (\tilde{S}_N^{(V)})_{kk})^T$, we get

$$f(W) = W^T Y. \tag{3.5}$$

From (3.5), the optimal weighting vector W is given by $\dfrac{Y}{\|Y\|_2}$.

The intuition hidden in our strategy is that the objective function is to obtain the largest variance and the single-view data (or similarity matrix) with larger variance will be assigned to a larger weight for joint analysis, which utilizes the multilinear relationship among various views.

3.4.2 Obtaining the relaxed cluster indicator matrix U

Note that trace is a linear function, for instance,

$$\text{trace}(\alpha \mathbf{A} + \beta \mathbf{B}) = \alpha \text{ trace}(\mathbf{A}) + \beta \text{ trace}(\mathbf{B}). \tag{3.6}$$

Hence, given W, the objective function $f(\bullet)$ can be written as

$$f(\mathbf{U}) = \text{trace}(\mathbf{U}^T(\sum_{v=1}^{V} w_v \mathbf{S}_N^{(v)})\mathbf{U}). \tag{3.7}$$

Hence, we have to determine the optimal $\mathbf{U}$ for only one matrix, namely, $\sum_{v=1}^{V} w_v \mathbf{S}_N^{(v)}$, which we write as $\tilde{\mathbf{S}} \in \mathbb{R}^{N\times N}$. It is obvious that we can obtain the optimal $\mathbf{U}$ by the eigenvalue decomposition (EVD) of the matrix $\tilde{\mathbf{S}}$.

As we obtain the relaxed indicator matrix $\mathbf{U}$, we can carry out the final partitioning by k-means (or other partitioning strategies) to obtain the cluster labels. The pseudo code of MC-STM is presented as follows:

Algorithm 3.4.1: MC-STM($\mathbf{S}_N^{(1)}, \mathbf{S}_N^{(2)}, ..., \mathbf{S}_N^{(V)}, K$)

comment: Given an initial matrix $\mathbf{U}$ and the cluster number K

1. *Calculate the weighting vector* W
2. *Obtain the relaxed indicator matrix* $\mathbf{U}$
3. *Go back to* 1 *until convergence*
4. *Normalize the rows of* $\mathbf{U}$ *to unit length*
5. *Calculate the cluster idx with* $k-$ *means on* $\mathbf{U}$

return (*idx* : *the clustering label*)

This linear combination of multiple similarity matrices (multi-view graphs) can be understood from a PCA point of view. Taking each similarity matrix (each view) as a component in the view space, the new integration matrix generated by our STM can be understood as the principal component in the view space. Then the optimal weights are the linear coefficients of multiple views to form this principal component that keeps as much variance from multiple views as possible. Thus this analysis preserves the maximum variance from multi-view data. On the contrary, just averaging the combination of these multiple similarity matrices is unable to preserve such variance.

3.4.3 The initialization of MC-STM

To carry out our STM based algorithm, in the first place, we need to provide the initialization of the relaxed indicator matrix $\mathbf{U}$. Four initialization schemes (but not limited to these four) are listed as the following.

1. Identity matrix method: the matrix $\mathbf{U}$ is set as part of an identity matrix;

2. Random orthonormal vector method: the columns of $\mathbf{U}$ are initialized as a set of random mutually orthonormal vectors;

3. Average spectral projection method: the normalized similarity matrices can be averagely summed and the matrix $\mathbf{U}$ is taken as the top K eigenvectors of the summed matrix;

4. Truncated MLSVD method: A similarity tensor can be constructed by taking each similarity matrix as the frontal slices. The matrix $\mathbf{U}$ is obtained by MLSVD of the similarity tensor.

Various initializations might cause different clustering results, which we will investigate in the experimental part.

3.4.4 The convergence of MC-STM

W.r.t the iterative maximization scheme of our MC-STM, convergence to a local optimum is guaranteed because in each step we maximally increase the objective function $f(\bullet)$. It may be necessary to reinitialize a number of times in order to find the global optimum. In fact, as demonstrated in both Section 3.8 and Section 3.9, our strategy has a good convergence property.

3.5 Joint dimension reduction of multiple graphs for clustering

3.5.1 Dimension reduction by SVD

Working with the original, high-dimensional data may be too time-consuming or even computationally infeasible. Moreover, it is known that, with an appropriate dimension reduction scheme, the low-dimensional estimators often have a smaller variance than high-dimensional estimators, which may lead to more accurate results. Consequently, in real applications, pre-processing of dimension reduction is a vital step to handle large-scale data. For the case of matrix representation of single-view data, SVD or PCA is a powerful tool to implement the dimension reduction on high dimension data [144].

Regarding multi-view data that is in the form of multiple graphs in this research, it generally leads to scalability as compared to single-view data. Meanwhile, owing to the overlap among various views, much redundant information exists

amid them. Therefore, it is imperative to carry out joint dimension reduction of multi-view data.

Analogous to SVD on single-view data, we employ its multi-view counterpart of MLSVD to carry out the joint dimension reduction of multi-view data. MLSVD usually leads to expensive computation as well. Nevertheless, the recent progress on scalable tensor decomposition [78, 121] allows us to efficiently implement MLSVD on large-scale data .

Besides, the related tensor knowledge has been introduced in the Chapter 2, the property of the core tensor is essential to the formulation of our dimension reduction strategy, which we will introduce in the following.

3.5.2 Basic knowledge of MLSVD

Suppose $\mathcal{A}$ is a original tensor and $\mathcal{B}$ is a core tensor after MLSVD, the subtensors $\mathcal{B}_{i_n=\alpha}$, obtained by fixing the nth index to α (where $n = 1, 2, 3$ is the mode number, in this case, $\mathcal{B}_{i_n}$ is a matrix), has the following properties.

1. All-orthogonality: two subtensors $\mathcal{B}_{i_n=\alpha}$ and $\mathcal{B}_{i_n=\beta}$ are orthogonal for all possible values of n, α and β subject to $\alpha \neq \beta$.

$$\langle \mathcal{B}_{i_n=\alpha}, \mathcal{B}_{i_n=\beta} \rangle = 0, \ when \ \alpha \neq \beta, \tag{3.8}$$

2. Ordering:

$$\|\mathcal{B}_{i_n=1}\|_F \geqslant \|\mathcal{B}_{i_n=2}\|_F \geqslant \ldots, \geqslant \|\mathcal{B}_{i_n=I_n}\|_F \geqslant 0, \tag{3.9}$$

for all possible values of n, where $\| \bullet \|_F$ means Frobenius-norm.

The Frobenius-norms $\|\mathcal{B}_{i_n=i}\|_F$, symbolized by $\delta_i^{(n)}$ are n-mode singular values of $\mathcal{A}$.

The Frobenius-norm of a tensor $\mathcal{A}$ is given by

$$\|\mathcal{A}\|_F = \sqrt{\sum_{i_1} \sum_{i_2} \sum_{i_3} a_{i_1,i_2,i_3}^2}. \tag{3.10}$$

In addition, there exits a certain relationship between the core tensor and the original tensor under the full MLSVD as the following,

$$\|\mathcal{A}\|_F = \|\mathcal{B}\|_F. \tag{3.11}$$

3.5.3 MC-STM-MLSVD

During truncated MLSVD, a core tensor is able to provide a good approximation of its original tensor, preserving the maximal variance. As a result, the core tensor can be employed to replace the original tensor to carry out the relevant operation. Modeling different objects as different tensors, Phan & Cichocki [115] employ core tensors to replace the original tensors as the significant features to carry out the classification task. By formulating generalized low rank approximations of a set of matrices, Ye [143] proposes a similar strategy to use a set of low-rank decomposed matrices, implementing image compression and recognition. The set of low-rank decomposed matrices can be considered to be an alternative formulation of our core tensor. In our strategy, at first, we

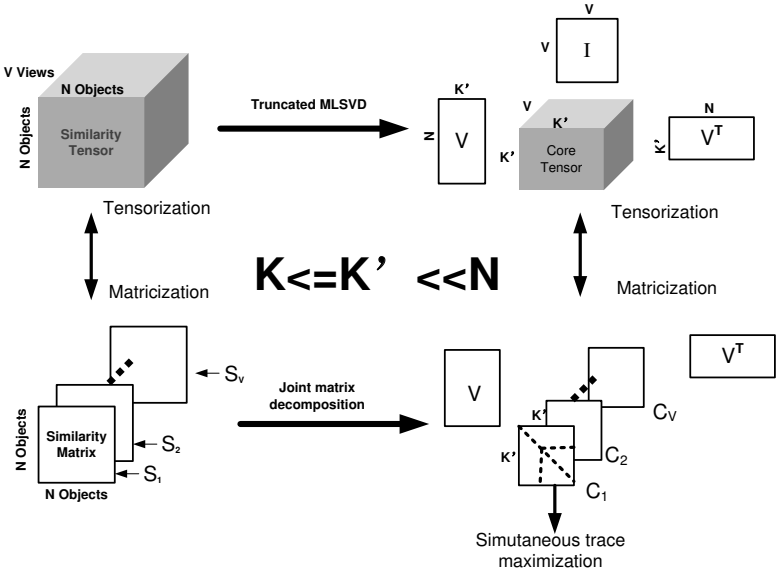

Figure 3.2: Truncated multilinear singular value decomposition (MLSVD) of a similarity tensor and its matricization. N is the number of objects (the dimensionalities of the original object space); K' is the number of the reduced dimensions; K is the number of clusters.

construct a similarity tensor from the original similarity matrices of each graph as in the Chapter 2. The frontal slice of the similarity tensor corresponds to the original similarity matrix of each view. Then we implement MLSVD to get a set of small-size matrices for each graph. The joint dimension reduction of multiple graphs is illustrated in Figure 3.2 and the truncated MLSVD of the

similarity tensor is written

$$\mathcal{A} \approx \tilde{\mathcal{A}} \times_1 \mathbf{V}^T \times_2 \mathbf{V}^T \times_3 \mathbf{I}^T, \tag{3.12}$$

where $\mathbf{V} \in \mathbb{R}^{N \times K^{'}}$ is the 1-mode factor matrix that is a common factor shared by multi-view data and $K^{'}$ is the truncated (reduced) dimension. Because of the partially symmetric property of this similarity tensor (the spaces of 1-mode and 2-mode are exactly same), the 2-mode factor matrix equals the 1-mode factor matrix V. Since we only care about the decomposition of 1 and 2-mode which are relevant to the object space for clustering, simply, we set the 3-mode factor matrix as an identity matrix $\mathbf{I}$. $\tilde{\mathcal{A}} \in \mathbb{R}^{K^{'} \times K^{'} \times V}$ is the core tensor after the truncated MLSVD, representing the interaction of each mode.

Next, we matricize the core tensor $\tilde{\mathcal{A}}$ as a set of matices: $\mathbf{C}_v = \mathcal{C}(1 : K^{'}, 1 : K^{'}, 1), v = 1, \ldots, V, \mathbf{C}_v \in \mathbb{R}^{K^{'} \times K^{'}}$. Hence, we replace the original similarity matrix $\mathbf{S}_N^{(v)}$ with $\mathbf{C}_v$ to implement the simultaneous trace maximization. Consequently, the size of the similarity matrix is reduced from $N \times N$ to $K^{'} \times K^{'}$, thus causing an efficient implementation. Subsequently, an optimal subpace of $\mathbf{U}_C \in \mathbb{R}^{K^{'} \times K}$ is obtained by simultaneous trace maximization on small-size matrices $\mathbf{C}_1, \ldots, \mathbf{C}_V$.

Finally, in order to recover the original optimal object subspace $\mathbf{U} \in \mathbb{R}^{N \times K}$, the following multiplication is required,

$$\mathbf{U} = \mathbf{V}\mathbf{U}_C. \tag{3.13}$$

Then we can partition the multiplied matrix $\mathbf{U}$ to get the final clustering labels. By embedding this joint dimension reduction scheme by MLSVD with MC-STM, we call this strategy MC-STM-MLSVD. The pseudo code of our MC-STM-MLSVD is listed as follows.

Algorithm 3.5.1: MC-STM-MLSVD($\mathbf{S}_N^{(1)}, \mathbf{S}_N^{(2)}, ..., \mathbf{S}_N^{(V)}, K$)

comment: K is the number of clusters

1. *Build a similarity tensor* $\mathcal{A}$
2. *Implement MLSVD on* $\mathcal{A}$ *and get* $\mathbf{V}$ *and core tensor* $\mathcal{C}$
3. *Matricize core tensor to a set of matrices* $\mathbf{C}_v$
4. *Carry out* $MC - STM$ *on* $\mathbf{C}_v$ *and get* $\mathbf{U}_C$
5. *Multiply* $\mathbf{U}^{'}$ *with* $\mathbf{V}$ *to recover the optimal subspace* $\mathbf{U} = \mathbf{V}\mathbf{U}_C$
6. *Normalize the rows of* $\mathbf{U}$ *to unit length*
7. *Calculate the cluster idx with* $k-means$ *clustering on* U

return (idx : *the clustering label*)

In addition, other tensor methods are also feasible to implement this joint dimension reduction, such as higher-order orthogonal iteration (HOOI) [32].

During this joint dimension reduction scheme, however, the problem remains as to how to select the proper number of the reduced dimensions $K^{'}$. There are two possible ways to handle this problem. In the first place, $K^{'}$ can be set according to test and error. Within a range of $K^{'}$, certain clustering evaluation measures are adopted to check which $K^{'}$ leads to good clustering performance. The second solution is based on 1-mode singular value analysis of the similarity tensor, which is analogous to the singular value analysis of matrix decomposition. Through the observation of the distribution of 1-mode singular values in descending order, we seek $K^{'} >= K$ so that the 1-mode singular values $\delta^{(1)}_{K^{'}+1} << \delta^{(1)}_{K^{'}}$, where $\delta^{(1)}_{K^{'}}$ denotes the $K^{'}$th 1-mode singular value and all the 1-mode singular values are sorted in a descending order. This selection is based on the assumption that the corresponding part of the core tensor $\tilde{\mathcal{A}}$ does not really contribute much to the approximation of the original tensor $\mathcal{A}$ during this tensor truncation.

3.6 Extension to other multi-view clustering

As known, several other clustering algorithms can be presented by alternative optimal formulation of trace maximization, which allows us to extend our MC-STM strategies to formulate the corresponding multi-view solutions. Meanwhile, the joint dimension reduction of multi-view data by MLSVD is applicable to these extensions as well.

3.6.1 Multi-view clustering by modularity optimization

Both modularity matrix and modularity based spectral optimization is referred to Chapter 5 and the related paper [106]. Given modularity matrices $\mathbf{B}_v$ ($v = 1, 2, ..., V$) from multiple graphs, the multi-view clustering is formulated as,

$$\begin{aligned} &\max_{\mathbf{U}, w_v} \sum_{v=1}^{V} w_v \text{trace}(\mathbf{U}^T \mathbf{B}_v \mathbf{U}), \\ &\text{s.t. } \mathbf{U}^T\mathbf{U} = I,\ w_v > 0 \text{ and } \sum_{v=1}^{V} w_v^2 = 1, \end{aligned} \tag{3.14}$$

which can be solved by simultaneous trace maximization as well.

In fact, the modularity matrix $\mathbf{B}_v$ is not guaranteed to be positive (semi) definite. If $\mathbf{B}_v$ has not less than K positive eigenvalues, $\mathbf{U}$ is taken equal

to the $K^{'}$ eigenvectors corresponding to the largest eigenvalues of $\mathbf{B}_v$. If $\mathbf{B}_v$ has only $\tilde{K} < K^{'}$ positive eigenvalues, then the first $\tilde{K}$ rows of $\mathbf{U}$ are taken equal to the $\tilde{K}$ eigenvectors corresponding to the positive eigenvalues of $\mathbf{B}_v$, complemented with the $K^{'} - \tilde{K}$ eigenvectors corresponding to the least negative eigenvalues of $\mathbf{B}_v$.

Another remedy is that we can regularize the modularity matrix to guarantee that it is positive (semi)definite [106].

3.6.2 Multi-view *k*-means clustering

According to [145], if the single-view data $\mathbf{X}$ (feature-by-object matrix) has zero sample means, the objective function of k-means is given by

$$\begin{aligned} &\max_{\mathbf{U}} \operatorname{trace}(\mathbf{U}^T\mathbf{X}^T\mathbf{X}\mathbf{U}), \\ &\text{s.t. } \mathbf{U}^T\mathbf{U} = \mathbf{I}. \end{aligned} \tag{3.15}$$

Given multi-view data $\mathbf{X}_v, v = 1, \ldots, V$, we can also formulate multi-view k-means clustering as the following,

$$\begin{aligned} &\max_{\mathbf{U}, w_v} \sum_{v=1}^{V} w_v \operatorname{trace}(\mathbf{U}^T\mathbf{X}_v^T\mathbf{X}_v\mathbf{U}), \\ &\text{s.t. } \mathbf{U}^T\mathbf{U} = \mathbf{I},\ w_v > 0 \text{ and } \sum_{v=1}^{V} w_v^2 = 1. \end{aligned} \tag{3.16}$$

3.7 Experimental setting

In this and next section, we will cross compare our multi-view spectral clustering methods and the other seven baseline multi-view clustering methods MKF, SA, FI, AdacVote, LMF, MC-OI-MLSVD and MC-MI-HOOI, which have been introduced in the last Chapter. In our experiments, both MC-MI-HOOI and LMF are initialized by MLSVD. We develop the STM algorithm by Matlab since the algorithm only involves basic matrix operations. The implementation of MLSVD can be referred to the Matlab based tensor toolbox [6].

In addition, we will exam other five aspects of our strategies, such as, the weighting analysis of STM, the initialization of STM, the convergence of STM, the joint dimension reduction scheme of STM and even the computation time.

3.7.1 Clustering evaluation

We adopt two clustering validation measures Adaptive Rand Index (ARI) [63] and Normalized Mutual Information (NMI) [127] to evaluate our clustering results. k-means clustering is employed as final partitioning method; then each clustering method is repeated for 50 times and the mean value is taken for comparison.

3.8 Experiment on disease gene clustering

Disease gene data with ten views has been presented in the Chapter 2. The clustering instances are 245 genes belonging to 14 diseases (14 clusters).

3.8.1 Clustering performance on disease gene data

At first, we implement the spectral clustering on the ten single-view data respectively and the clustering performance is presented in Table 3.1. As can be seen, the best clustering performance of individual data sources is obtained on the LDDB text mining profile (NMI 0.7065, ARI 0.5669).

Afterwards, we investigate the multi-view clustering performance of integrating all single-view data. The comparison of all related multi-view clustering algorithms (we also list the best clustering on single-view data LDDB for comparison) is presented in Table 3.2.

In addition, we also list the best clustering on single-view data LDDB for comparison with these multi-view clustering strategies. Of all the methods we compared, the top two best performance is obtained by our two STM based strategies MC-STM-MLSVD and MC-STM. Thus our multi-view clustering methods are not only beyond spectral clustering of any single-view data but also superior to the seven baseline multi-view clustering methods. For instance, the NMI values is improved from 0.7065 by LDDB (best single-view data) to 0.7451 by MC-STM-MLSVD while the ARI value is improved from 0.5474 by MC-MI-HOOI (the best alternative multi-view clustering strategy) to 0.5938 by MC-STM-MLSVD. As shown, the improvement by our STM based strategy is statistically significant, thus demonstrating the effectiveness of our multi-view weighting strategy by simultaneous trace maximization.

Table 3.1: The clustering performance of single-view data on disease gene data. The mean values and standard deviations are observed from 50 random repetitions. The best performance is indicated in bold.

Method	NMI	ARI
GO	0.5502±0.0138	0.3458±0.0367
MeSH	0.6988±0.0227	0.5200±0.05
OMIM	0.6912±0.0193	0.4965±0.0458
NCI	0.5088±0.0125	0.2833±0.0168
eVO	0.5896±0.0185	0.3633±0.0267
KO	0.3222±0.0083	0.1161±0.0081
LDDB	**0.7065±0.0147**	**0.5669±0.0283**
MP	0.6516±0.0220	0.4643±0.0418
SNOMED	0.6673±0.0231	0.4868±00453
Uniprot	0.5713±0.0183	0.3460±0.034

Table 3.2: The clustering performance of various multi-view clustering strategies on disease gene data. The mean values and standard deviations are observed from 50 random repetitions. The best performance is shown in bold. The p-values are statistically evaluated with the best performance using paired t-test.

Algorithm	NMI	p-value	ARI	p-value
MC-STM	0.7319±0.0137	1.43e-04	0.5841±0.026	0.3202
MC-STM-MLSVD	**0.7451±0.0123**	—	**0.5938±0.0287**	—
MKF	0.7028±0.0229	2.83e-06	0.5215±0.0527	2.42e-06
FI	0.6868±0.0207	1.67e-13	0.5106±0.0532	4.58e-07
AdacVote	0.6439±0.0308	2.16e-16	0.4103±0.0704	1.39e-15
SA	0.6896±0.0190	2.51e-13	0.5314±0.0316	4.87e-08
MC-OI-MLSVD	0.7282±0.0183	3.16e-05	0.5205±0.0409	5.13e-04
LDDB	0.7065±0.0147	6.22e-09	0.5669±0.0283	0.0017
LMF	0.5481 ± 0.003	6.16e-20	0.3518 ± 0.002	3.19e-18
MC-MI-HOOI	0.7221 ± 0.003	3.27e-11	0.5474 ± 0.002	2.63e-06

3.8.2 The analysis of the weighting coefficients of multiple graphs on disease gene data

To evaluate whether the optimized weights assigned on each single-view data are correlated with the clustering performance, the comparison between the ranking of weighting coefficients of each single-view data and the ranking of their clustering performance is illustrated in Table 3.3.

The comparison suggests that, in general, there exits a corresponding relationship between the weighting factors and the clustering performance. For instance, the largest weighting coefficient corresponds to the best clustering performance (LDDB) while the least weighting coefficient to the worst clustering performance (KO). The comparison results suggest that the single-view data with best clustering performance makes most contribution to the joint analysis while the single-view data with worst clustering performance makes least contribution.

However, as can be seen in Table 3.3, the ranking of these optimal weighting coefficients are not completely consistent with the ranking of the corresponding clustering performance. This inconsistency may be due to the fact that there is certain linear overlap among the different views.

In addition, we also list the weighting coefficients obtained by MC-MI-HOOI for comparison as presented in Table 3.3. Although the weighting coefficients by MC-MI-HOOI are slightly different from these by our STM based multi-view clustering strategies, the rankings by these strategies are almost the same as that by our strategies (except the exchange of 5th and 6th). This conincidence may be due to the fact that all these strategies are based on multilinear analysis.

3.8.3 The analysis of the initialization schemes of STM on disease gene data

We investigate four various initialization methods introduced in Subsection 3.4.3 to see the effect of these initialization schemes on our clustering strategies. The running time and clustering evaluation of our clustering strategies together with these four initialization schemes are provided in Table 3.4.

As can be seen, regarding these four initial schemes, all their clustering performance under our multi-view clustering strategies (either by MC-STM or by MC-STM-MLSVD), is quite similar, indicating that our multi-view clustering strategy is insensitive to initializations. Hence, we adopt the

Table 3.3: The weighting coefficients of multi-view data on disease gene data. Pranking refers to the ranking of clustering performance. $w_v^{(1)}$ obtained by MC-STM, $w_v^{(2)}$ obtained by MC-STM-MLSVD and $w_v^{(3)}$ obtained by MC-MI-HOOI.
.

Sources	Ranking	$w_v^{(1)}$	Ranking	$w_v^{(2)}$	Ranking	$w_v^{(3)}$	Pranking
GO	9	0.1821	9	0.1819	9	0.2544	8
MeSH	7	0.2328	7	0.2326	7	0.2842	2
OMIM	4	0.2540	4	0.2538	4	0.2973	4
NCI	5	0.2422	5	0.2420	6	0.2931	9
eVO	3	0.2721	3	0.2720	3	0.3021	6
KO	10	0.1311	10	0.1310	10	0.2216	10
LDDB	1	0.7218	1	0.7223	1	0.5303	1
MP	2	0.2729	2	0.2727	2	0.3113	5
Snomed	8	0.2113	8	0.2111	8	0.2713	3
uniprot	6	0.2410	6	0.2409	5	0.2970	7

Table 3.4: The clustering performance of our STM based multi-view clustering algorithms with four initialization schemes on disease gene data

Schemes	Clustering methods	NMI	ARI	Time (seconds)
Random	MC-STM	0.7289±0.01	0.5782±0.03	0.0077
	MC-STM-MLSVD	0.7460±0.01	0.5940±0.03	0.0082
Identity	MC-STM	0.7306±0.01	0.5787±0.02	4.0e-05
	MC-STM-MLSVD	0.7423±0.01	0.581±0.04	4.1e-05
MLSVD	MC-STM	0.7292±0.02	0.5786±0.03	0.0947
	MC-STM-MLSVD	0.7493±0.01	0.5972±0.04	0.0944
Average	MC-STM	0.7317±0.01	0.5828±0.03	0.0391
	MC-STM-MLSVD	0.7451±0.01	0.5938±0.03	0.0397

initialization scheme of identity matrix due to the lest computation time it requires. In addition, in light of the insensitive property of the initialization scheme, our strategies appear to be beyond some multi-view clustering strategies that rely on a proper initialization, such as LMF.

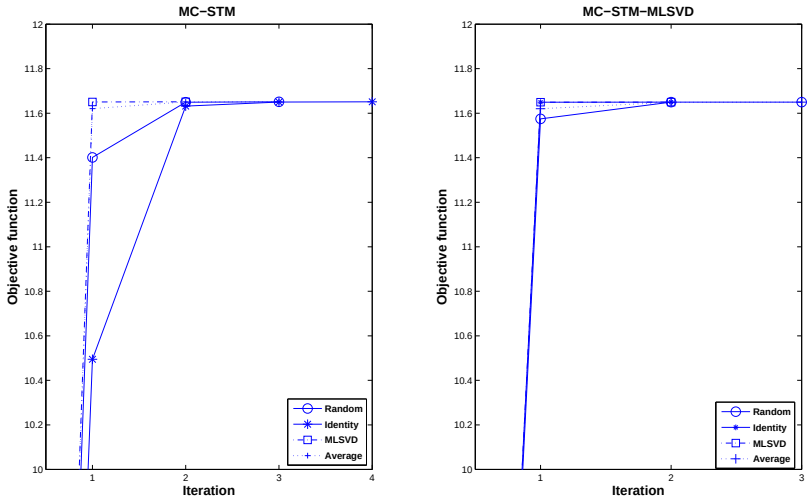

Figure 3.3: The convergence of MC-STM and MC-STM-MLSVD with various initialization schemes on disease gene data

3.8.4 The analysis of the convergence of STM on disease gene data

We investigate the convergence of both MC-STM and MC-STM-MLSVD during the optimization stage. Within each strategy, the values of their objective functions at various iterative steps are plotted in Figure 3.3. In the first place, both of our strategies with various initializations converge within less than three iterative steps, which seems very fast. Second, with various initialization methods, all the objective functions of our multi-view clustering strategies constantly converge to the same point.

3.8.5 The analysis of the joint dimension reduction of multiple graphs on disease gene data

At first, as presented in Table 3.1, our strategy MC-STM–MLSVD which contains the joint dimension reduction scheme, achieves the best clustering results and its improvement over other alternative clustering algorithms, is significant.

Next, as can be observed in Table 3.3, both the ranking of weighting factors, and even the values of these factors, by both MC-STM and MC-STM-MLSVD, are almost the same. Since these weighting coefficients denote the linear relationship of multi-view data, this result demonstrates the MLSVD based dimension reduction strategy is capable of capturing the inherent linear relationship among multi-view data as implemented in the original space.

Afterwards, according to Figure 3.3, it is apparent that MC-STM-MLSVD (around two iterations) converges faster than MC-STM (around three iterations). It seems that the faster convergence is also caused by the joint dimension reduction of multiple graphs.

Besides, we investigate the performance of dimension reduction by varying the number of the dimensions, based on the distribution of the 1-mode singular values of the similarity tensor. The 1-mode singular values of the similarity tensor on disease gene data are partially plotted in Figure 3.4. It is obvious that the 1-mode singular values become smaller when the descending order is over 50. Hence, we vary the dimension from 2 to 50, and the relevant clustering performance is shown in Figure 3.5. It can be observed that there is a peak region around 14 (in particular, in the ARI observation), which is exactly the number of clusters (diseases). This analysis implicates that the scheme of joint dimension reduction may achieve good results when the number of the reduced dimension is chosen around the number of clusters. As a result, in our MC-STM-MLSVD, the number of dimension is chosen as exactly the number of clusters in above clustering analysis.

3.9 Experiment of scientific mapping

In contrast to disease gene data, this Web of Science (WoS) database is large (8,305 journal objects and 669,860 terms) as well as the cluster distribution is biased (the number of members within each cluster varies from 25 to 1140). Consequently, the clustering task is full of challenges. The two views we adopt to obtain scientific mapping are text mining (TFIDF) and bibliometric data (cross-citation). In order to provide a scientific mapping of WoS, integrating the two heterogeneous data, we apply our algorithms to cluster the 8305 journals into 22 clusters, which is the number of Essential Science Indicator (ESI) subjects as the reference categories in WoS. The detail of journal data and ESI subjects is presented in Chapter 4.

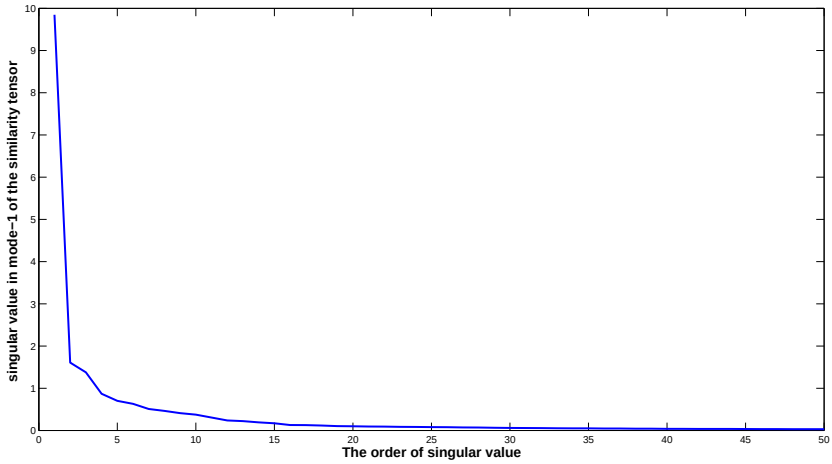

Figure 3.4: The distribution of the top 50 1-mode singular values of the similarity tensor on disease gene data.

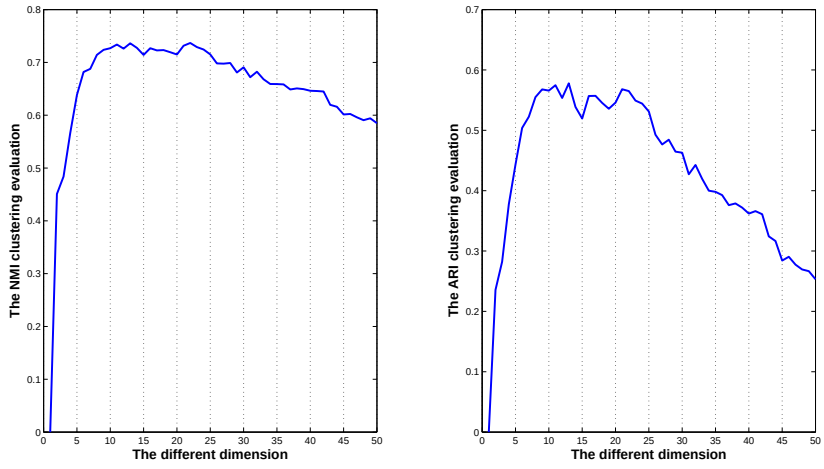

Figure 3.5: Clustering performance of MC-STM-MLSVD with varied dimension on disease gene data.

3.9.1 Clustering performance on journal data

We implement spectral clustering on two single-view data and the related multi-view clustering strategies for comparison. The clustering evaluation of the related clustering algorithms can be observed in Table 3.5. First, the best single-view clustering performance (NMI 0.5386, ARI 0.3312) is obtained by TFIDF data. Next, except MC-MI-HOOI and MC-STM, the performance of our MC-STM-MLSVD (NMI 0.5615, ARI 0.3531) is superior to other clustering strategies, including multi-view clustering as well as single-view clustering. Moreover, the improvement by MC-STM-MLSVD is statistically significant.

In addition, MC-STM has the similar performance (NMI 0.5613, ARI 0.3512) as MC-STM-MLSVD and MC-MI-HOOI and there is no difference between the three multi-view clustering algorithms in term of statistical significance (p-value between MC-STM and MC-STM-MLSVD: 0.8708; p-value between MC-STM and MC-STM-MLSVD: 0.4215).

3.9.2 The analysis of the weighting coefficients of multiple graphs on journal data

Table 3.6 gives the weighting factors optimized by three multi-view clustering strategies as well as the corresponding clustering performance (ARI evaluation). As can be seen, according to the results of these three clustering schemes, the weighting factor of CRC is larger than that of TFIDF. At the same time, the clustering performance of CRC is better than that of TFIDF as well. The comparison suggests that, in this two-view case, the ranking of weighting factors of both single-view data is consistent with the ranking of their clustering performance. This experimental result is also in line with the common sense of bibliometric analysis: the citation is more informative and reliable while text contains much noise, hence citation data should dominantly contribute to the joint analysis.

3.9.3 The analysis of initialization schemes of STM on journal data

As can be observed in Table 3.7, we analyze the clustering performance of our clustering strategies with various initializations on the journal data. Similar to the analysis on disease gene data, there is no big difference among these four initialization solutions, all of which lead to almost the same clustering

Table 3.5: Clustering evaluation of the related clustering algorithms on journal data. Two single-view clustering methods and eight multi-view clustering methods. The mean values and standard deviations are observed from 50 random repetitions. The best performance is shown in bold. The p-values are statistically evaluated with the best performance using paired t-test.

Algorithm	NMI	p-value	ARI	p-value
MC-STM-MLSVD	**0.5615±0.0039**	—-	**0.3531±0.0168**	—–
MC-STM	0.5613±0.0036	0.8708	0.3512±0.0175	0.6069
TFIDF	0.5256±0.0062	1.31e-19	0.3033±0.0077	4.06e-23
CRC	0.5386±0.0065	9.39e-13	0.3312±0.0161	1.03e-5
MKF	0.5524±0.0061	1.47e-10	0.3028±0.009	7.937e-22
FI	0.5540±0.0064	5.84e-7	0.335±0.0218	9.5e-4
AdacVote	0.5482±0.0065	1.28e-12	0.3264±0.0224	2.0723e-6
SA	0.5105±0.0334	4.52e-14	0.2769±0.0333	5.33e-18
MC-OI-MLSVD	0.5550±0.006	1.18e-4	0.3392±0.0181	2.02e-3
MC-MI-HOOI	0.5618 ± 0.0004	0.4215	0.3534 ± 0.0032	0.7201

Table 3.6: The weighting coefficients of multi-view data on journal data. $w_v^{(1)}$ obtained by MC-STM, $w_v^{(2)}$ obtained by MC-STM-MLSVD and $w_v^{(1)}$ obtained by MC-MI-HOOI.

Source	$w_v^{(1)}$	$w_v^{(2)}$	$w_v^{(3)}$	ARI
TFIDF	0.3190	0.3226	0.3829	0.3033
CRC	0.9478	0.9465	0.9234	0.3412

performance. The comparison results also demonstrate our STM based optimization strategy is insensitive to the various initializations.

3.9.4 The analysis of convergence of STM on journal data

The convergence of our clustering strategies on journal data can be observed in Figure 3.6. In the first place, regarding this large-scale data, both of our strategies still gain good performance, converging within less than three iterative steps. Second, all objective functions of various initialization schemes converge to the same point (the maximal objective function) nicely.

Table 3.7: The clustering performance of our STM based multi-view clustering strategies with four initialization schemes on journal data

Schemes	Clustering methods	NMI	ARI	Time(s)
Random	MC-STM	0.5627+0.0064	0.3558+0.022	92.61
	MC-STM-MLSVD	0.5627+0.0048	0.3513+0.0141	0.0028
Identity	MC-STM	0.5622+0.0061	0.3534+0.017	0.015
	MC-STM-MLSVD	0.5624+ 0.0047	0.3527+0.0154	0.0125
MLSVD	MC-STM	0.5617+0.0053	0.3513+0.016	863
	MC-STM-MLSVD	0.5630+0.0045	0.3549+0.0171	0.2876
Average	MC-STM	0.5621+0.005	0.3525+0.0172	145
	MC-STM-MLSVD	0.5630+0.0050	0.3567+0.0165	0.6156

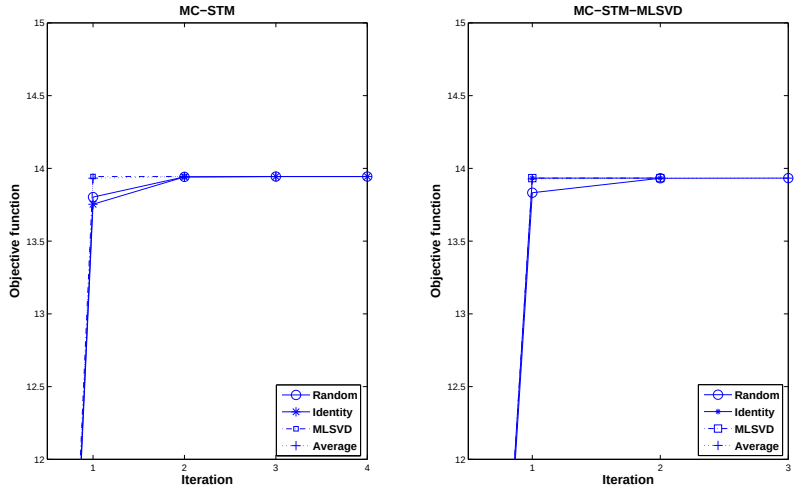

Figure 3.6: The convergence of MC-STM and MC-STM-MLSVD with various initialization schemes on journal data

3.9.5 The analysis of joint dimension reduction of multiple graphs on journal data

First, as shown in Table 3.5, on this large-scale journal data, our MC-STM-MLSVD still achieves better clustering performance.

Second, Table 3.6 reveals that MC-STM-MLSVD is still able to capture the inherent linear relationship of multi-view data as MC-STM does in the original space because the weighting coefficients by both strategies are almost the same.

Like the convergence analysis of disease gene data, MC-STM-MLSVD (around two iterations) converges faster than MC-STM (around three iterations) as illustrated in Figure 3.6.

Finally, the 1-mode singular values of the similarity tensor on journal data are partially plotted in Figure 3.7. It can be observed that the 1-mode singular values become tiny after the top 100. Thus, we exam the clustering performance of MC-STM-MLSVD by varying the dimension from 2 to 100. As can be observed in Figure 3.8, regarding the clustering performance of these dimension reduction cases, there is a peak region around 22 (in particular, in the ARI observation), which is exactly the number of standard benchmark categories (the 22 ESI categories). This analysis echoes the dimension reduction analysis of disease gene data: it seems that MC-STM-MLSVD is able to achieve good clustering performance when the dimension is set around the number of clusters.

3.9.6 Comparison of the computation time of the relevant clustering algorithms

To investigate the computational time of the multi-view clustering algorithms, we benchmark our multi-view clustering strategies with other alternative methods on the two applications. As can be seen in Table 3.8, both of our strategies, MC-STM and MC-STM-MLSVD, seem to be efficient. In particular, on the large-scale journal data, they are only behind MC-OI-MLSVD. However, in contrast to MC-OI-MLSVD, our two strategies are able to yield more enriched information (the weighting factor of each view). Furthermore, among our two strategies, MC-STM-MLSVD appears superior in terms of computation time due to the joint dimension reduction scheme. At the same time, MC-MI-HOOI which is able to provide weighting factors as well is comparable to our strategies in term of clustering performance; nevertheless, its computation time is nearly six times longer than our strategies on large-scale journal data, which appears quite inefficient in real applications.

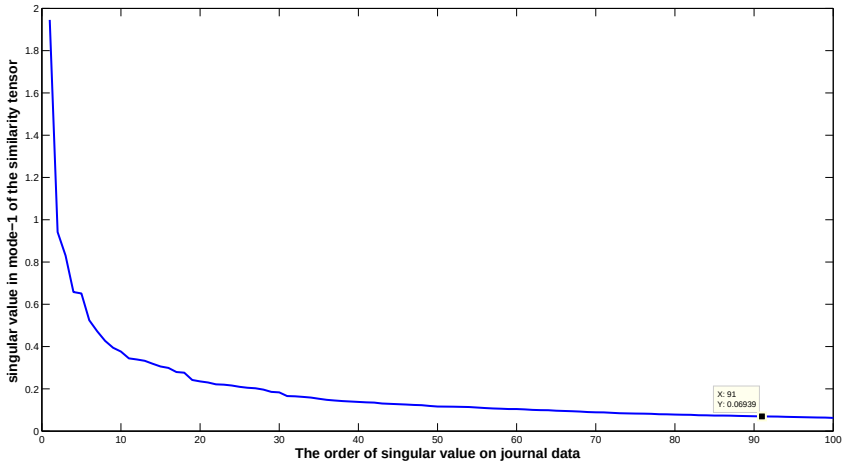

Figure 3.7: The distribution of top 100 1-mode singular values of the similar tensor on journal data.

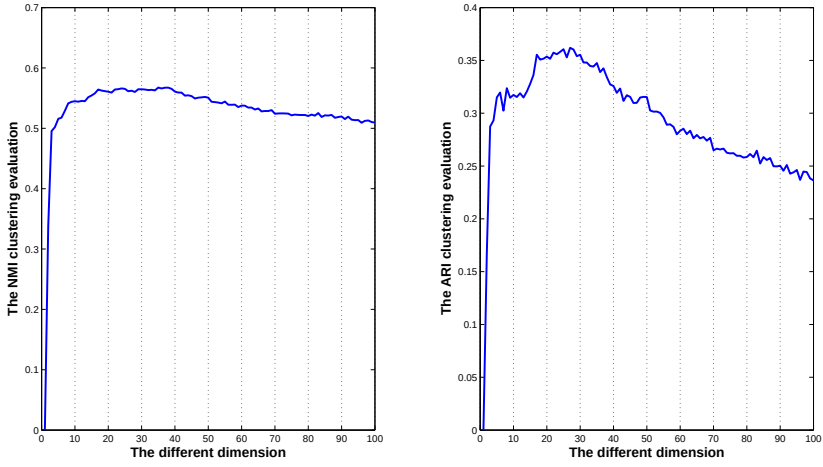

Figure 3.8: Clustering performance of MC-STM-MLSVD with varied dimension on journal data.

Table 3.8: Comparison of CPU time of all multi-view clustering algorithms on real applications

Algorithm	journal data (seconds)	disease data (seconds)
SA	1200	95
AdacVote	1200	98
MKF	1726	7
FI	2093	21
MC-OI-MLSVD	939	9
LMF	3683	207
MC-MI-HOOI	7460	8.29
MC-STM	1325	32
MC-STM-MLSVD	**1081**	**11**

3.10 Discussion

Our proposed methods provide a loose framework for the clustering of multi-view data, which is not limited to spectral formulation. For instance, it can be extended to multi-view modularity based clustering and multi-view k-means clustering. Hence, theoretically, any types of clustering with the alternative formulation of trace maximization can be extended to its multi-view clustering variants by our STM strategy.

In this Chapter, we mainly discuss the research of multi-view clustering. In fact, as illustrated in Figure 3.1, our strategy aims to seek an optimal latent subspace of objects. Therefore, it can be extended to other multi-view learning tasks: such as classification [142, 152], spectral embedding and collaborative filtering [151].

Many researchers are concerned with the convex optimization of clustering methods. Based on the experimental analysis, the alternative maximization solution of our multi-view clustering strategies seems satisfactory. In addition, our strategies are able to achieve good convergence. Consequently, we neglect the step of finding the convex solution that would be very time-consuming.

Concerning the strategy of MC-STM, it can efficiently captures the linear relationship of multiple graphs through simply simultaneous trace maximization. Through this kind of multilinear analysis, the complementary information among multi-view data, such as the text data and citation data, is fully employed, which can facilitate the joint clustering based on our experimental analysis. Meanwhile, our MC-STM-MLSVD is the joint dimension reduction version of our multi-view clustering strategy. With MLSVD, the noise or

the redundant data among multi-view data is removed to some degree before the joint clustering analysis. Therefore as compared to MC-STM, MC-STM-MLSVD has several apparent advantages based on our applications:

- More efficient implementation with short computational time (or even less memory);
- Same or even better clustering performance;
- Better convergence property;
- Recovering the same linear relationship among multi-view data within the low dimension spaces

Nevertheless, the efficiency of MC-STM-MLSVD depends on the dimension reduction scheme of MLSVD, which usually leads to heavy computation w.r.t large-scale data. At the same time, some research work about the efficient implementation of MLSVD has been proposed [78, 121]. For instance, MLSVD can be simplified as eigenvalue decomposition (EVD) and the relevant tensor operations can be simplified to the operation between vectors and matrices. We will investigate this issue in the further research.

3.11 Summary

The main points of this paper are three-fold:

- Our multi-view clustering strategies are able to efficiently utilize the linear relationship of multiple graphs for joint analysis.
- The joint dimension reduction scheme of multiple graphs allows us to handle large-scale data. Through joint dimension reduction, without destroying the linear relationship of multi-view data, our strategy MC-STM-MLSVD can achieve the same or even better clustering performance.
- The general framework of our multi-view clustering can be easily extended to other trace maximization based clustering (modularity optimization based clustering and k-means clustering).

In addition, as shown in (3.7), because our STM scheme mainly involves the EVD of a matrix during multi-view clustering stage no matter how many graphs

(views) are involved, our multi-view clustering strategies are applicable to any multi-view conditions.

We applied our strategies to two real applications: grouping genes in the disease gene data and scientific mapping of WoS data. Both applications demonstrate the effectiveness of both our STM based multi-view clustering strategy and the joint dimension reduction scheme of multiple graphs.

Chapter 4

Scientific mapping by hybrid clustering in vector spaces

4.1 Introduction

In scientometrics, information from journals can be categorized lexically or with citations. An important area of scientometric research is the clustering or mapping of scientific publications. The widely used method of co-citation clustering was introduced independently by Small [124, 125] and Marshakova [95]. Cross-citation based cluster analysis for science mapping is different; while the former is usually based on links connecting individual documents, the latter requires aggregation of documents to units like journals or subject fields among which cross-citation links are established. Some advantages of this method (for instance, the possibility to analyze directed information flows) are undermined by possible biases. For example, bias could be caused by the use of predefined units (journals, subject categories etc.) implying already certain structural classification. Journal cross-citation clustering has been used by Leydesdorff [87], Leydesdorff and Rafols [88], and Boyack, Börner, and Klavans [15], while Moya-Anegón *et al.* [98] applied subject co-citation analysis to visualize the structure of science and its dynamics.

The integration of lexical similarities and citation links has also attracted interest in other fields such as search engine design (i.e., Google combines text and links [18]). The combination of link-based clustering with a textual approach was suggested as early as 1990 to improve the efficiency and usability of co-citation and co-word analysis. One of the aims was to

improve the apparently low recall of co-citation analysis concerning current work [16, 17, 153]. The combination of link-based and textual methods also makes it possible to cluster objects whenever links are weak or missing (e.g. in the case of poorly cited or un-cited papers). The present study is based on a new combined citation/lexical-based clustering approach [70], which forms a hybrid solution in two respects. First, it combines citations and text, and second, it uses individual papers to cluster the journals in which they appear. Furthermore, the lexical component is used to label the journal clusters obtained for interpretation.

Hybrid clustering has also been applied in various document analysis applications [12, 59, 97, 138] as well as science mapping research [50, 69, 91]. Although these approaches all combine lexical and citation information, the actual algorithms applied are quite diverse. For web document analysis, Modha & Spangler [97] integrated similarity matrices from terms, out-links and in-links by a weighted linear combination, and the data partition was obtained from the combined similarity matrix using the toric k-means algorithm. He *et al.* [59] incorporated three types of information (hyperlink, textual and co-citation information) to cluster web documents using a graph-cut algorithm. Bickel & Scheffer [12] investigated web documents and combined intrinsic views (page content) with extrinsic views (anchor texts of inbound hyperlinks). Three clustering algorithms (generic Expectation-Maximization (EM), k-means and agglomerative) were applied to combine the different views as hybrid clustering. With exception of Web page analysis, Glenisson *et al.* [49] combined textual analysis and bibliometrics to improve the performance of journal publication clustering. Janssens [69] proposed an un-biased combination of textual content and citation links on the basis of Fisher's inverse chi-square for agglomerative clustering. Liu *et al.* [91] reviewed some popular hybrid clustering techniques within a unified computational framework and proposed an Adaptive Kernel k-means Clustering (AKKC) algorithm to learn the optimal combination of kernels constructed from heterogeneous data sources.

The present study advances the hybrid clustering approach in terms of using larger-scale experimental data and combining more refined data models. Large-scale journal data presents a challenge to hybrid clustering, because the journal sets are usually expressed in a high dimension vector space and a massive amount of journals usually represents a large number of scientific fields. Moreover, the present study combines the lexical and citation data into ten heterogeneous representations for hybrid clustering. Therefore, when the dimensionality, the number of samples, and the number of categorizations are all large, many existing algorithms become inefficient. To tackle this problem, we present a new hybrid clustering approach for large-scale journal data in terms of scalability and efficiency. The data used in

this research was collected from the Web of Science (WoS) journal database from the period 2002-2006, which contains over 6,000,000 publications. In our approach, the above mentioned ten data sources are combined in a weighted manner, where the weights are determined by the Average Normalized Mutual Information (ANMI) between the single source partitions and the hybrid clustering partitions based on combined data. To evaluate the reliability of the clustering obtained on journal sets, we compared the clustering results with the standard categorizations, Essential Science Indicators (ESI), provided by Thomson Scientific (Philadelphia, PA, USA). We systematically compare the automatic clustering results obtained by all methods with the standard ESI categorizations. We also apply some statistical evaluation methods to produce label-independent evaluations. In total, twelve different hybrid clustering algorithms are investigated and benchmarked using two external and two internal validation measures. The experimental results show that the proposed algorithms can achieve both improved clustering result and high efficiency.

In sum, our contributions are three-fold:

- We propose an ANMI-based weighted hybrid clustering scheme and formulate the related variants of clustering ensemble methods and multiple kernel fusion.
- We generate ten multi-view text mining data from both textual perspective and citation perspective for joint analysis.
- The strategy is applied to the scientific mapping of WoS journal database and several clustering evaluation measurements are adopted.

This Chapter is organized as follows. The adopted data set and the standard ESI categorizations are described in next Section. We then present the proposed hybrid clustering methodologies and the ANMI weighting scheme. Next, the experimental results are analyzed. Afterward, we illustrate and investigate the mapping of journal sets obtained from hybrid clustering.

4.2 Journal database analysis

4.2.1 Data sources and data processing

The original journal data contains more than six million published papers from 2002 to 2006 (i.e., articles, letters, notes, reviews, etc.) indexed in the WoS database provided by Thomson Scientific. Citations received by these

papers have been determined for a variable citation window beginning with the publication year, up to 2006. An item-by-item procedure was used with special identification-keys made up of bibliographic data elements, which were extracted from the first-author names, journal title, publication year, volume and the first page. To resolve ambiguities, journals were checked for the name changes and the papers were checked for name changes and merged accordingly. Journals not covered in the entire period (from 2002 to 2006) have been omitted. Two criteria were applied to select journals for clustering: at first, only the journals with at least 50 publications from 2002 to 2006 were investigated, and others were removed from the data set; then only those journals with more than 30 citations from 2002 to 2006 were kept. With this kind of selection criteria, we obtained 8305 journals (in paper level, there are more than six million papers) as the data set adopted in this Chapter.

4.2.2 Text mining analysis

The titles, abstracts and keywords of the journal publications were indexed with a Jakarta Lucene [55] based text mining program using no controlled vocabulary. The index contains 9,473,061 terms but we cut the Zipf curve of the indexed terms at the head and the tail to remove rare terms, stopwords and common words [73]. These words are known to be usually irrelevant and noisy for clustering purposes. After the Zipf cut, 669,860 meaningful terms were used to represent the journals in a vector space model where the terms are attributes and the weights are calculated using four weighting schemes: TF-IDF, IDF, TF and binary. The paper-by-term vectors are then aggregated to journal-by-term vectors as the representations of the lexical data. Therefore, we have obtained four sub-models as the textual data sources varied with the term weighting scheme. We applied Latent Semantic Indexing (LSI) [11] on the TF-IDF data to reduce the dimensionality to 200 LSI factors. LSI is implemented on the basis of the SVD algorithm. The number of LSI factors was selected empirically in a similar way as the preliminary work of Janssens [69]. For the 8305 journals, on a dual Opteron 250 with 16 GB RAM, time taken for LSI computation was around 105 minutes. Then this new textual feature is named LSI-TFIDF.

4.2.3 Citation analysis

We investigated the citations among the selected publications in five aspects. These citation data can be generated by information extraction from WoS database.

- **CRC**: Cross-citation between two papers is defined as the frequency of citations between each other. We ignored the direction of citations by symmetrizing the cross-citation matrix.
- **BV-CRC**: To neglect the side effect of the large amount of citations appearing in the journals, we used binary value 1 (0) to represent whether there is (no) citation between two journals, termed binary cross-citation.
- **COC**: Co-citation refers to the number of times two papers are cited together in subsequent literature. The co-citation frequency of two papers is equal to the number of papers that cite them simultaneously.
- **BGC**: Bibliographic coupling occurs when two papers reference a common third paper in their bibliographies. The coupling frequency is equal to the number of papers they simultaneously cite.
- **LSI-CRC**: We also applied LSI on the sparse matrix with cross-citations to reduce the dimensionality. The selection of the number of the LSI factors was also based on the previous work [69] and was set to 150.

The citations among papers were all aggregated to the journal level. The citation All the textual data sources and citation data sources were converted into kernels using a linear kernel function. In particular, for the textual data, the kernel matrices were normalized and their elements correspond to the cosine value of pair-wise journal-by-term vectors.

4.2.4 Reference labels of journals

As mentioned in last Section, to evaluate the science mapping results, we refer to the twenty-two categorizations of ESI, which are curated by various professional experts. Our main objective is thus to compare the automatic mapping obtained by the proposed hybrid methods against the ESI categorizations. As shown in Table 4.1, the number of journals contained in the different ESI fields is quite imbalanced. For instance, the largest field (Clinical Medicine) contains 1410 journals, whereas the smallest (Multidisciplinary) only has 25 journals.

4.3 Weighted hybrid clustering for large-scale data

The hybrid clustering algorithms considered in our experiments can be divided into two approaches: clustering ensemble and kernel-fusion clustering.

Table 4.1: The 22-field Essential Science Indicator (ESI) labels of the WoS journal database

Field #	ESI field	Num	Field #	ESI field	Num
1	Agricultural Science	183	12	Mathematics	312
2	Biology & Biochemistry	342	13	Microbiology	87
3	Chemistry	441	14	Molecular Biology & Genetics	195
4	Clinical Medicine	1410	15	Multidisciplinary	25
5	Computer Science	242	16	NeroScience & Behavior	194
6	Economics & Business	299	17	Pharmacology & Toxicology	135
7	Engineering	704	18	Physics	264
8	Environment/ Ecology	217	19	Plant & Animal Science	608
9	Geoscience	277	20	Psychology/Psychiatry	448
10	Immunology	73	21	Social Science	968
11	Materials Sciences	258	22	Space Science	47

Clustering ensemble is also known as clustering aggregation or consensus clustering, which integrates different partitions into a consolidated partition with a consensus function. Kernel-fusion clustering maps the data sets into a high dimensional feature space and combines them as kernel matrices. Then a kernel based clustering algorithm is applied to the combined kernel matrix. The details about these two approaches are mentioned in our earlier work [91]. The present study proposes a novel weighting scheme on the basis of ANMI to leverage the effect of multiple sources in hybrid clustering. For all sub-models, the one with the largest ANMI value is expected to have the most relevant information and therefore it should contribute dominantly to the hybrid clustering.

4.3.1 Definition of ANMI

ANMI has been employed in clustering ensemble algorithms [127], where the optimal cluster ensemble is obtained by maximizing the ANMI value. Given a set of cluster labels $P = \{P_1, \ldots, P_v, \ldots, P_V\}$, where V is the number of views, P_v represents the labels obtained from a single sub-model and N is the number of sub-models. ANMI measures the average normalized mutual information between P_v and P, given by

$$ANMI(P_v, P) = \frac{1}{V-1} \sum_{j=1, j\neq v}^{V} NMI(P_v, P_j), \tag{4.1}$$

where NMI is the normalized mutual information indicating the common information shared by two partitions, given by

$$NMI(P_v, P_j) = \frac{\sum_{k=1}^{C}\sum_{m=1}^{C} c_{km} log(\frac{nc_{km}}{a_k b_m})}{\sqrt{(\sum_{k=1}^{C} e_k log(\frac{e_k}{n}))(\sum_{m=1}^{C} f_m log(\frac{f_m}{n}))}} \tag{4.2}$$

In the formulation above, C is the cluster number; e_k is the number of data points contained in the k-th cluster in the partition P_v; f_m is the number of samples contained in the m-th cluster in the partition P_j; c_{km} is the number of intersection samples between the k-th cluster from P_v and the m-th cluster from P_j. In particular, if P_j is the standard reference labels, $NMI(P_v, P_j)$ evaluates the performance of P_v with the standard labels.

4.3.2 Comparison of ANMI with other evaluation measures

In data fusion applications, the use of external validation indicators is an appropriate way to provide data-independent evaluations about the clustering quality, however, they rely on the known reference labels. In contrast, the statistical validation indicators (internal validation indicators) depend on the scales, the structures and the dimensionalities of data, thus they are not suitable to be compared among multiple data sources. In this case, the reliability of the internal and the external validation indicators can be judged by cross-comparing with each other. The ANMI adopted in our approach belongs to the internal validation case because it does not require any reference labels. To prove its reliability, we compare the ANMI with external validation indicators (NMI and Adjusted Rand Index [ARI]) using the individual sub-models of journal sets. Besides the ANMI, we also compare the other two internal validation indicators (Mean Silhouette Value [MSV] and modularity). As illustrated in Figure 4.1, the ANMI shows almost the same trend as the NMI and the ARI when predicting the model performance. In contrast, the MSV and the modularity show some similar trends but are not very consistent with the curve of the NMI and the ARI. The merit of ANMI is that the performance is evaluated on the basis of information criterion, which avoids the data dependency on scales, structures and dimensionalities. In our problem, the ANMI shows similar evaluation on sub-models as the NMI and the ARI, which both need the extra reference labels for evaluation. Therefore ANMI is reliable to apply in explorative data analysis. Furthermore, the validity of ANMI as an evaluation measure has also been introduced by Strehl & Ghosh [127].

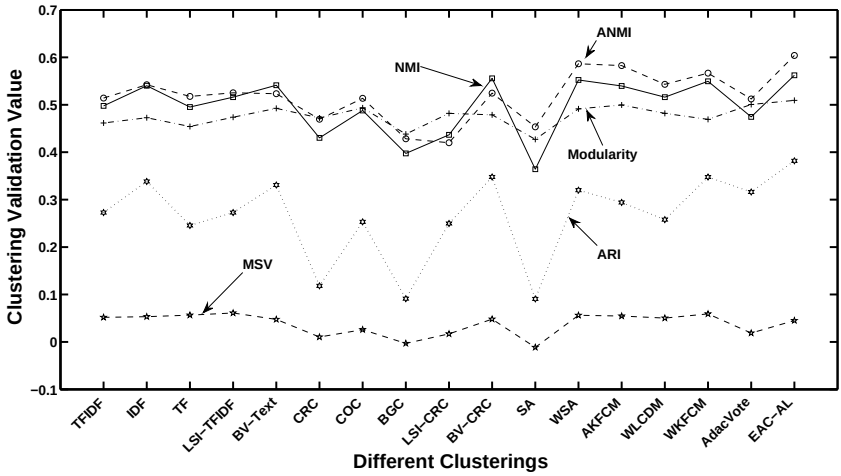

Figure 4.1: Comparison of ANMI with the external-validation indicators (NMI and ARI) and the internal-validation indicators (MSV and Modularity).

4.3.3 Weighting scheme

As explained, our approach assumes that when different sub-models are applied for the hybrid clustering, the more relevant sub-models should contribute more to the hybrid clustering. A straightforward way to leverage the sub-models is to weigh them according to the values of their indicators (i.e., the ANMI values, the MSV values, the modularity values, etc.). Based on this assumption, we propose an ANMI-based weighting scheme to combine the kernel matrices (similarity matrices) of multiple sub-models as a weighted convex linear combination. The conceptual scheme of our proposed weighting scheme is depicted in Figure 4.2.

As illustrated in Figure 4.2, the weighted hybrid clustering consists of several steps which may be summarized as follows:

- Step 1: The kernels of all sub-models are constructed and clustered individually by Ward's linkage based hierarchical clustering [68]. The obtained partition of each sub-model is denoted as P_v. For all the submodels, the set of partitions is denoted as $P = \{P_1, P_2, \ldots, P_V\}$. As introduced, ten sub-models are involved so V is equal to 10.
- Step 2: Based on P, the clustering result of each sub-model is evaluated

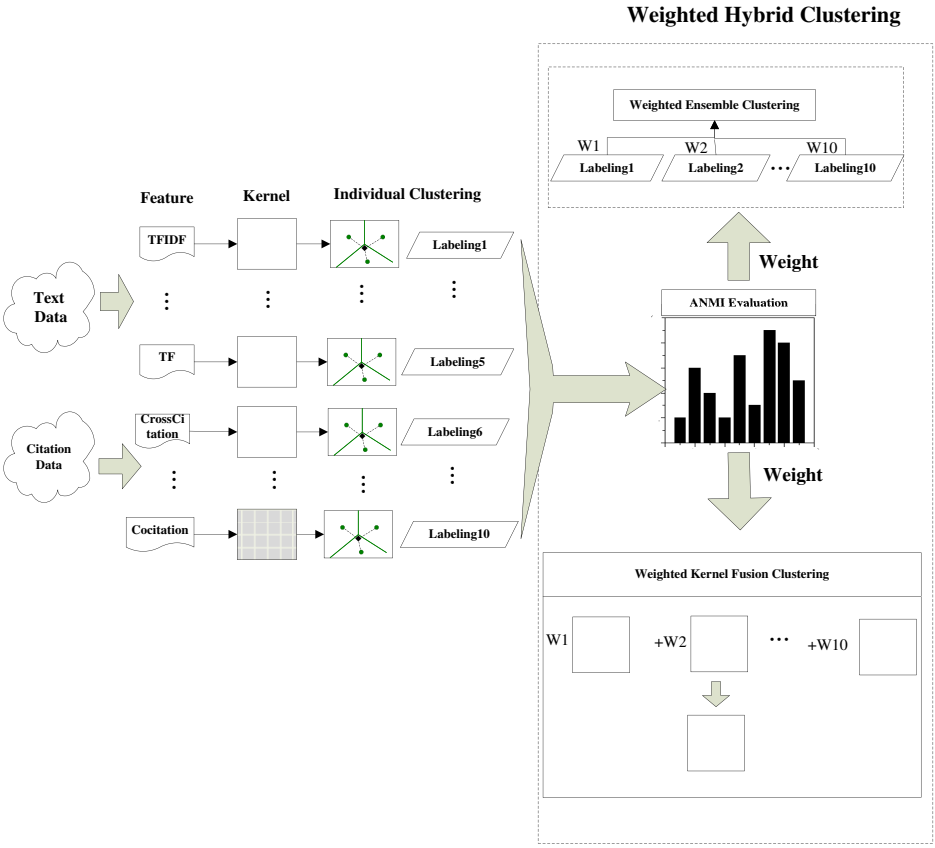

Figure 4.2: Conceptual framework of the ANMI-based weighted hybrid clustering.

using the ANMI as defined in (4.1). The ANMI index is denoted as a_v, given by

$$a_v = ANMI(P_v, P),\ v \in \{1, 2, \ldots, V\}. \tag{4.3}$$

- Step 3: We compute the weights w_v of sub-models as proportional to their ANMI values, given by

$$w_v = \frac{a_v}{a_1 + \ldots + a_v + \ldots + a_V},\ v \in \{1, 2, \ldots, V\}. \tag{4.4}$$

- Step 4: Using the weights obtained in step 3, we combine the kernels in a weighted manner, and alternatively, we integrate the labels of sub-models as weighted clustering ensemble. The algorithms are briefly described as follows:

  – **Weighted kernel-fusion clustering method (WKFCM)**. In kernel-fusion clustering method (KFCM), given a set of kernels $\mathbf{K}_v, v = 1, \ldots, V$, constructed from V sub-models, to leverage their effects in hybrid clustering, we integrate their kernels as a weighted combination, given by

$$\mathbf{K} = \sum_{v=1}^{V} w_v \mathbf{K}_v. \tag{4.5}$$

  The combined kernel $\mathbf{K}$ is further applied by single kernel based clustering algorithms (i.e., kernel k-means, hierarchical clustering based on kernel spaces, spectral clustering, etc.).

  – **Weighted clustering ensemble (WSA and WEAC-AL)**. In clustering ensemble, the partitions of all sub-models are $\{P_1, \ldots, P_V\}$ usually considered as equally important. To incorporate the weights, we extend the algorithm of SA proposed by Strehl & Ghosh [127], as the Weighted Strehl's Clustering Ensemble Algorithm (WSA). Moreover, we also analogously extend the Evidence Accumulation Clustering with Average Link (EAC-AL) algorithm proposed by Fred & Jain [44] as the weighted EACA-AL algorithm (WEAC-AL). Both extensions are straightforward: in the original versions, the partitions of multiple sub-models are considered as the input; in the weighted versions, the input is formulated as $\{w_1 P_1, \ldots, w_V P_V\}$.

Collectively, we have proposed three weighted hybrid clustering methods on the basis of ANMI (WKFCM, WSA, WEAC-AL).

4.3.4 Clustering evaluation

Mean silhouette value (MSV) The silhouette value of a clustered object (e.g., journal) measures its similarities with the objects within the cluster versus the objects outside of the cluster [117], given by:

$$S(i) = \frac{min(B(i, C_j)) - W(i)}{max(min(B(i, C_j), W(i)))}, \tag{4.6}$$

where $W(i)$ is the average distance from object to all other objects within its cluster, and $B(i, C_j)$ is the average distance from object i to all objects in another cluster C_j. The MSV for all objects is an intrinsic measure on the overall quality of a clustering solution. MSV may vary with the number of clusters, which is also useful to find the appropriate cluster number statistically. In the journal database, the dimensionality of lexical data is extremely high so the distance based calculation of MSV is computationally expensive. As an alternative solution, we pre-compute the paired distances of all samples and store it as a kernel, in this way, the average distance required in the MSV value is directly computable in the kernel of paired distances.

Modularity. Newman [107] introduced modularity as a graph-based evaluation of the clustering quality. Up to a multiplicative constant, modularity calculates the number of intra-cluster links minus the expected number in an equivalent network with the same clusters, but with links given at random. It means good clustering may have more links within (and fewer links between) the clusters than could be expected from the random links. Modularity is defined as follows: a $k \times k$ symmetric matrix e is defined as the element, e_{ij} is the fraction of all the edges in the network that link vertices in community or cluster i to vertices in cluster j. The trace of this matrix $\text{trace}(e) = \sum_i e_{ii}$ represents the fraction of edges in the network that connect vertices in the same cluster. The sum of rows (or columns) $a_i = \sum_j e_{ij}$ represents the fraction of edges that connect to vertices in cluster i. The modularity Q is then defined as:

$$Q = \sum_i (e_{ii} - a_i^2) = \text{trace}(e) - \|e^2\|, \tag{4.7}$$

where $\|x\|$ is the sum of the elements in matrix x and $\|e^2\|$ refers to the expected fraction of edges that connect vertices in the same cluster with edges given at random in the network.

ARI. The Adjusted Rand Index (ARI) is the corrected-for-chance version of the rand index [63]. The adjusted rand index measures the similarity between two partitions. Let us assume that two partitions X and Y are obtained from a given set of n elements $S = \{O_1, \ldots, O_n\}$, given by $X = \{x_1, \ldots, x_r\}$ and

$Y = \{y_1, \ldots, y_s\}$, we define the following: a, as the number of pairs of elements in S that are in the same set in X and in the same set in Y; b, as the number of pairs of elements in S that are in different sets in X and in different sets in Y; c, as the number of pairs of elements in S that are in the same set in X and in different sets in Y; d, as the number of pairs of elements in S that are in different sets in X and in the same set in Y.

The ARI R is defined as:

$$R = \frac{2(ab - cd)}{((a+b)(b+d) + (a+c)(c+b))}. \tag{4.8}$$

The ARI implementation can be referred to the Matlab code [1].

Normalized mutual information (NMI). NMI is another external clustering validation measure, which relies on the reference labels. NMI is defined in (4.2).

Both MSV and modularity are internal evaluation measures that do not rely on the benchmark categories but the data structure itself. These internal validation measures sometimes are used to select the number of clusters [69]. As a result, in this Chapter, we adopt them to evaluate the clustering results with varied cluster numbers.

4.3.5 Other hybrid clustering algorithms

In addition to the three proposed hybrid clustering algorithms, we also apply six hybrid clustering algorithms for comparison.

- Strehl's clustering ensemble algorithm (SA) [127]: as introduced in Chapter 2. The code is provided by the author [2].
- EAC-AL: Fred & Jain [44] introduce evidence accumulation clustering (EAC) that maps the individual data partitions as a clustering ensemble by constructing a co-association matrix. The final data partition is obtained by applying average linkage (AL) based hierarchical clustering algorithm on the co-association matrix.
- AdacVote [5]: as introduced in Chapter 2.
- QMI: Topchy, Jain & Punch [133] phrase the combination of partitions as a categorical clustering problem. Their method adopts a category utility

[1] http://www.kernel-methods.net/matlab_tools.html
[2] http://www.lans.ece.utexas.edu/ strehl/soft.html

function proposed by Mirkin [96] that evaluates the quality of a "median partition" as a summary of the ensemble.

- AKFCM: The averagely combined kernel is treated as a new individual data source and the partitions are obtained by standard clustering algorithms in the kernel spaces.
- WLCDM: The weighted linear combination of distance matrices method proposed by Janssens *et al.* [70] is actually a simplified version of AKFCM: it is achieved by equally-weighted linear combination of a text based kernel and a citation based kernel. In this way, WLCDM is equal to the concatenation of the different normalized feature vectors from various single views [69].

The first four algorithms belong to the category of clustering ensemble, whereas the next two algorithms are kernel-fusion clustering methods. Regarding our weighted hybrid clustering, the developing of ANMI based weighting scheme is based on ClusterPack Matlab Toolbox [3].

4.4 Experimental result

In this part, at first, we analyze our clustering result on WoS journal database. Then we discuss the clustering under various number of clusters and the computational complexity of different clustering schemes.

4.4.1 Evaluation of clustering results

We applied all algorithms to combine the ten sub-models to cluster the journal data into 22 partitions. The ten sub-models were also clustered individually as single sources and the performance was compared with the hybrid clustering. To determine statistical significance, we used the bootstrap t-test [39]. The bootstrap sampling was repeated 30 times and for each repetition, approximately 80% of the journals were sampled. After bootstrapping, the duplicated samples were normalized as one sample for clustering. To evaluate the performance, we applied both ARI and NMI using the standard ESI categorizations. The MSV and the standard deviations (STD) of the 30 bootstrapped samples are shown in Table 4.2.

[3] http://www.lans.ece.utexas.edu/ strehl/soft.html

Table 4.2: Comparison of different clustering methods by NMI and ARI

Method	NMI	ARI	Method	NMI	ARI
TFIDF	0.5080±0.008	0.2676±0.017	WLCDM	0.5161±0.008	0.2885±0.012
IDF	0.5478±0.009	0.3071±0.019	AKFCM	0.5175±0.006	0.2841±0.012
TF	0.5124±0.009	0.2816±0.022	WKFCM	0.5495±0.006	0.3246±0.023
LSI-TFIDF	0.5242±0.006	0.2925±0.02	QMI	0.5477±0.012	0.3069±0.024
BV-Text	0.5399±0.009	0.3213±0.023	AdacVote	0.4851±0.027	0.2824±0.056
CRC	0.4532±0.016	0.1604±0.032	SA	0.4722±0.025	0.1697±0.07
COC	0.4672±0.016	0.1786±0.032	WSA	0.5532±0.016	0.3057±0.026
BGC	0.4191±0.012	0.1256±0.025	EAC-AL	0.5562±0.006	0.3387±0.019
LSI-CRC	0.4378±0.009	0.2221±0.018	WEAC-AL	0.5757±0.008	0.3710±0.014
BV-CRC	0.5544±0.008	0.3350±0.02			

Table 4.3: Comparison of different weighted clustering performance by t-test.

Compared clustering methods	p-value
WSA vs. SA	2.22e-12
WKFCM vs. AKFCM	1.84e-8
WEAC-AL vs. EAC-AL	5.8e-03
WEAC-AL vs. BV-CRC	3.5e-03

Weighted hybrid clustering performs better than its non-weighted counterpart

As shown in Table 4.2, all the weighted methods outperformed their non-weighted counterparts. For the EAC-AL algorithm, the weighted version improved the ARI value by 9.54% and the NMI value by 3.51%. For the kernel-fusion clustering, the weighted algorithm increased the ARI index by 14.23 % and the NMI index by 5.99%. The weighted combination in WSA also improved the ARI value of SA method by more than 50 % and the NMI index by 18.32 %. The improvement of the weighted methods was shown to be statistically significant and the p-values obtained from the bootstrapped t-test are presented in Table 4.3. The reason why our weighted hybrid clustering algorithms perform better might be due to the fact that they emphasize the consensus pattern among multi-view data through mutual information based weighting schemes, which causes the removal of individual noise to some degree.

Weighted hybrid clustering performs better than the best individual sub-model

We also compared the performance of individual sub-models with the hybrid results. As shown in Table 4.2, WEAC-AL gained improvement by heterogeneous data fusion and led to better performance than the best individual sub-model (BV-CRC). Compared to other hybrid clustering algorithms listed in previous section, WEAC-AL outperformed them as well.

Comparison of the lexical data and the citation data

When using the base algorithm on a single sub-model, the lexical data generally performed better than the citation data. This was probably because the sparse structures in the citation data could be more thoroughly analyzed using the graph cut algorithms than using the kernel clustering methods (we will handle this issue in Chapter 5). However, the main objective of this study is to show the validity of the weighted hybrid clustering scheme. To keep the problem simple and concise, we do not distinguish the heterogeneities of data structure. Combining different structures with different clustering algorithms is an interesting and novel problem, and it will be presented in our forthcoming publication.

The investigation of individual sub-models also substantiated the validity of our proposed weighting scheme: the sub-models with higher clustering performance were assigned with larger weights. For example, the sub-model IDF with the largest weight performed the second best individually; the sub-model (BV-CRC) with the second largest weight performed the best individually.

Comparison of kernel-fusion clustering with clustering ensemble

Our experiment compared 6 clustering ensemble and 4 kernel-fusion clustering methods on the same large-scale journal database. As shown in Table 4.2, the clustering ensemble methods generally showed better clustering performance. This was probably because the clustering ensemble relies more on the "agreement" among various partitions to find the optimal consensus partition. In our experiment, ten sub-models were combined and most of them were highly relevant, so the combination of sufficient and correlated partitions was helpful in finding the optimal consensus partition. In our related work [91], the notion of "sufficient number" was also shown to be important for clustering ensemble. In contrast, kernel-fusion clustering algorithms were less affected by the number of sub-models.

Table 4.4: Clustering performance of different weighted clustering schemes.

Weighted hybrid clustering method	NMI	ARI
MSV-based SA	0.5309	0.2866
ANMI-based SA (WSA)	0.5532	0.3057
MSV-based KFCM	0.5447	0.3067
ANMI-based KFCM (WKFCM)	0.5495	0.3246
MSV-based EAC-AL	0.5491	0.3414
ANMI-based EAC-AL (WEAC-AL)	0.5757	0.3710

Comparison of ANMI-based and MSV-based weighting schemes

Alternatively, we could also base our weighting scheme on the MSV criterion to leverage different sub-models in hybrid clustering. To compare the effects of MSV and ANMI in weight calculation, we applied the MSV-based weighting scheme to create three analogous hybrid clustering methods. The comparison of the two weighting schemes is shown in Table 4.4. As illustrated, the weighting scheme by ANMI works better than that based on MSV.

The failure of MSV based weighting schemes may be due to the fact that MSV value relies on the data structure property of each view alone so that the different MSV values from various views might not be comparable. Whereas ANMI is based on the mutual information of various views, so the different ANMI values from various views are comparable, thus leading to a better weighting scheme.

4.4.2 Clustering by various number of clusters

So far, the presented results were all obtained for the number of clusters equal to the number of standard ESI categorizations. How to determine the appropriate cluster number from multiple data sources still remains an open issue. As known, in single data clustering, the optimal cluster number can be explored by comparing indices for various cluster numbers. In our approach, we compared the MSV and modularity indices from 2 clusters to 30 clusters. As depicted in Figure 4.3, the indices of the proposed algorithm are almost all higher than those of the non-weighted methods. Moreover, they are also generally better than the best individual data (BV-CRC).

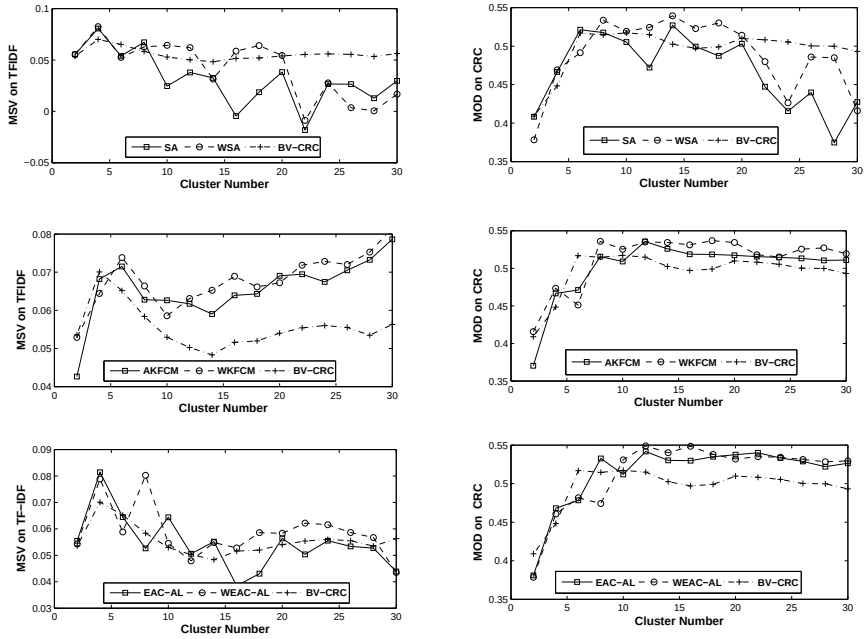

Figure 4.3: Internal validations of weighted hybrid clustering methods on different cluster numbers. The two figures on the top compare the weighted clustering ensemble methods. The figures in the middle evaluate the weighted kernel fusion clustering method of WSA. The figures on the bottom investigate the WEAC-AL clustering methods. The figures on the left represent the MSV indices. The figures on the right side represent the modularity (MOD) indices. The MSV is calculated on the TF-IDF sub-model and the MOD is verified on the CRC sub-model.

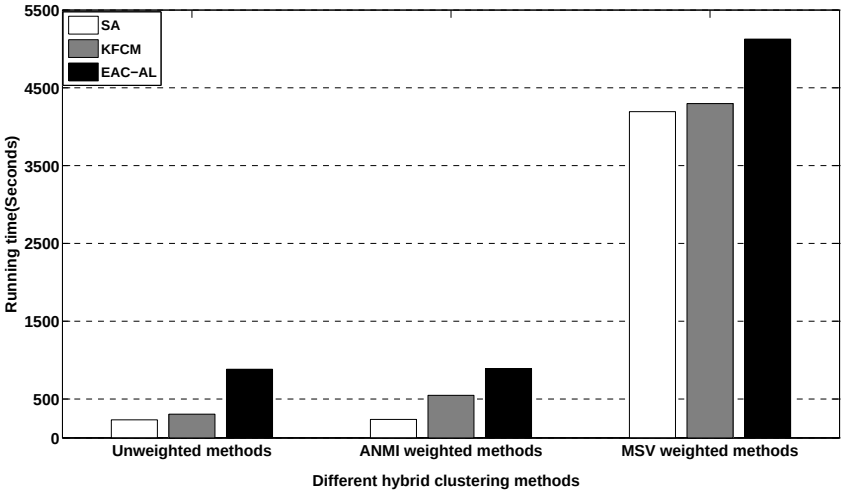

Figure 4.4: Comparison of the running time of different hybrid clustering methods. The running time is measured when clustering all the journals to 22 partitions.

4.4.3 Computational complexity on different weighting schemes

We also compared the computational time of the ANMI-based hybrid clustering algorithms with the un-weighted and the MSV-based weighted algorithms. The experiment was carried out on a CentOS 5.2 Linux system with a 2.4G Hz CPU and 16 G Bytes memory. As illustrated in Figure 4.4, the ANMI-based weighting scheme is more efficient than the MSV-based weighting scheme. Moreover, the ANMI-based weighting method performs as efficiently as the un-weighted version.

4.5 Mapping of the journal sets

To visualize the clustering result of journal sets, the structural mapping of the 22 categorizations obtained using the WEAC-AL method is presented in Figure 4.5.

Network Structure of the 22 Journal Clusters. For each cluster, the three most important terms are shown. The network is visualized by Pajek [9]. The edges represent cross-citation links and darker color represents more links between

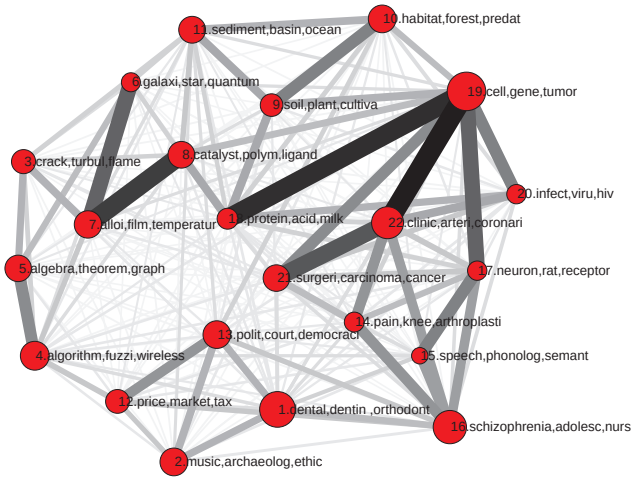

Figure 4.5: Network structure of the 22 journal clusters.

the paired clusters. The circle size represents the number of journals in each cluster.

To better understand the structure of clustering, we applied a modified Google PageRank algorithm [73] to analyze the journals within each cluster. The algorithm is also applied to rank a journal within each cluster according to the number of papers it published and the number of cross-citations it received. The algorithm is defined as,

$$PR_i = \frac{1-\alpha}{n} + \alpha \sum_j PR_j \frac{a_{ji}/P_i}{\sum_k \frac{a_{jk}}{P_k}}, \tag{4.9}$$

where PR_i is the PageRank of the journal i, α is a scalar between 0 and 1 (we set $\alpha = 0.9$ in our implementation), n is the number of journals in the cluster, a_{ji} is the number of citations from journal j to journal i, and P_i is the number of papers published by the journal i. The self-citations among all the journals were removed before the algorithm was applied. Using the algorithm, as (4.9), we investigated the five most highly ranked journals in each cluster and presented them in Table 4.5. Moreover, for the journals presented in Table 4.5, we re-investigated the titles, abstracts and keywords that have been indexed in the text mining process, the indexed terms were sorted by their frequencies and for each cluster, the thirty most frequent terms were used to label the obtained clusters. The textual labels of each journal cluster are shown in Table 4.6.

Table 4.5: The five most important journals of each cluster ranked by the modified PageRank algorithm.

Cluster 1	Cluster 2	Cluster 3	Cluster 4
1. Tech High Educ	1. Pub Histo	1. Acous R L	1. Austra Compu J
2. Strojarstvo	2. Histo Euro Idea	2. J App Mech Asme	2. J Res Prac Infor T.
3. Verter Econ	3. Pub Culture	3. Zamm Ange Math	3. Technome
4. Urban Educ	4. R. Du Lou Rev	4. App Energy	4. IEEE Multimedia
5. Thero Lingu	5. Antiquity	5. AIAA J	5. J Quaity Tech
Cluster 5	Cluster 6	Cluster 7	Cluster 8
1. P London Math Soc	1. Physc Rev A	1. Plat Surf finishing	1. Poly Inter
2. Grap Combin	2. Astrono Astrophy	2. J App Physics	2. Indi J Chem Sec
3. P Japan Ac S A-math	3. A Rev Nucl Parti Sci	3. Plastic Rub Comp	3. Poly Eng & Sci
4. Algeb Geom Topo	4. Astroph J	4. App Phys L	4. Afinidad
5. Stat Meth Med Res	5. Jetp L	5. J Phase Equili	5.Studies Surf SCI & C.
Cluster 9	Cluster 10	Cluster 11	Cluster 12
1. J Plant Grow Regul	1. Neotro Entomo	1. Physic Earth Plant	1. J corpo Finance
2. A J Enol & Viticul	2. Environ Entomo	2. IEEE T Geosci & Rem	2. Finance a Uver
3. Agronomie	3. Nauti	3. Phys & Chem Earth	3. A J Agricu & Reso E
4. J Range Manage	4. Ameghi	4. Aquatic Geochem	4. A Occupa Hygiene
5. A Rev Phytop	5. Wilson J Ornith	5. Spe Drilling & Compl	5. Manage Learn
Cluster 13	Cluster 14	cluster 15	Clsuter 16
1. Popul & Environ	1. J A Board Fami Med	1. Brain & Langua	1. Work & Stress
2. Geogra Zeitschrift	2. Arthroscopy	2. Behav Res Methods	2. Telemedi J & E-health
3. Politis Viertelj	3. Archi Environ Health	3. Clinic Linguis & Phon	3. Med Hygiene
4. A Rev Pub Admin	4. Birth Perina Care	4. J Nerolinguis	4. Fami Soc J Contem H
5. Washing Quarte	5. I J Geri Psychi	5. Behavior & Brain Sci	5. Zeits Entwich Padag P
Cluster 17	Cluster 18	Cluster 19	Cluster 20
1. Neuromole Med	1. J Food Sci & Tech	1. Math Biosci	1. R Med Microbio
2. Behaviou Brain Res	2. Archiv Fur Gefluge	2. Lab Animal	2. A Virology
3. Archives Itali Biolo	3. App Environ M	3. Methods Enzymology	3. A Agricul & Environ M
4. Brain	4. Worlds Poultry Sci J	4. Meth Compan Meth E	4. Avian Pathology
5. I J Neuroscience	5. Arch Latin Oamer N	5. Maydica	
Cluster 21	Cluster 22		
1. Pathology	1. J Aero Med Clea E L.		
2. Grae A Clin & Exper O	2. Obster & Gynecology		
3. Pathologe	3. Clin J A S Nephrology		
4. A j Neuroradiology	4. J D Maladies Vascul		
5. Skull Surgery	5. Nutr Metab & Cardi D		

According to Tables 4.5 and Table 4.6, we obtained the following structure. In the natural and applied sciences, we have found nine clusters, particularly, cluster #3 through #11. On the basis of the most important journals and terms, we have labeled them engineering (ENGN), computer science (COMP), mathematics (MATH), astronomy, astrophysics, physics of particles and fields (ASTR), physics (PHYS), chemistry (CHEM), agriculture, environmental science (AGRI), biology (BIOL) and geosciences (GEOS). The interpretation of the most characteristic terms of the nine life-science and medical clusters is somewhat more complicated. In particular, we have a biomedical, a clinical and psychological group. The latter one has some overlap with the third group, the social sciences and humanities clusters. Although the overlap of the most important terms within the life-science and medical clusters is considerable, the terms provide an excellent description for at least some of the medical clusters. Thus cluster #16 (PSYC) stands for psychology, #17 (NEUR) for the neuroscience and #15 (COGN) for cognitive science. While NEUR represents the medical and clinical of neuro- and behavioural sciences, COGN comprises cognitive psychology and neuroscience and PSYC rather psychology and psychiatry which is traditionally considered part of the social

Table 4.6: The textual labels of the journal clusters.

Cluster	30 best terms	Subject
1	teacher, detal, student, dentin, teeth,school,patient, educ, cari, orhodont implant, resin,dentur,enamel,tooth,mandibular,classroom,maxillari,polit,children social bond teach dentist discours cement librari incisor endodont learner	SCO1
2	music archaeolog polit ethic moral religi literari christian essai god philosoph religion church philosophi artist war centuri poetri historian hi roman text narr poem aesthet social theologi fiction argu kant spiritu	HUMA
3	crack turbul finit flame heat shear concret combust vibrat beam reynold temperatur veloc elast steel thermal vortex wilei fuel acoust convect coal load plate flow equat lamin fatigu jet buckl	ENGN
4	algorithm fuzzi wireless robot queri semant ltd qo packet traffic xml user graph network multicast fault wilei machin cdma web server bit servic cach bandwidth scheme architectur watermark sensor simul circuit	CSCI
5	algebra theorem finit graph asymptot polynomi infin equat inc manifold let algorithm semigroup ltd singular cohomolog inequ conjectur convex omega lambda infinit ellipt eigenvalu abelian automorph hilbert bound hyperbol epsilon sigma	MATH
6	galaxi star quantum optic neutrino quark stellar brane luminos magnet laser redshift galact beam solar cosmolog photon superconduct qcd spin ngc atom meson neutron nucleon rai boson temperatur ion hadron	ASTR
7	alloi film temperatur dope crystal magnet si anneal dielectr diffract microstructur gan quantum silicon epitaxi steel metal ceram sinter atom nanotub fabric oxid nm layer spin thermal ion electron coat	PHYS
8	catalyst polym ligand acid crystal bond ion atom nmr hydrogen solvent adsorpt wilei angstrom copolym oxid ltd poli temperatur molecul polymer electrochem metal chiral film spectroscopi aqueou electrod anion	CHEM
9	soil plant cultivar leaf crop seedl seed arabidopsi shoot wheat gene speci flower rice weed biomass ha tillag germin fruit irrig maiz forest protein acid fertil manur water pollen root	AGRI
10	speci habitat forest predat fish larva prei nov egg lake genu femal taxa bird plant forag male larval biomass season river breed parasitoid nest phylogenet abund mate fisheri soil beetl	BIOC
11	sediment basin soil ocean ltd seismic rock fault water sea magma tecton earthquak mantl isotop river crustal aerosol volcan subduct groundwat lake magmat atmospher climat wind cloud crust metamorph temperatur ozon	GEOS
12	firm price market tax wage busi polici capit organiz economi trade worker employe invest monetari earn investor financi auction asset brand inc corpor compani stock welfar incom job employ retail bank	ECON
13	polit polici social ltd court parti democraci democrat urban reform forest elector women vote discours war sociolog land tourism geographi market welfar crime voter labour elect poverti econom economi govern citi	SOC2
14	patient pain knee arthroplasti hip injuri fractur tendon athlet clinic muscl ligament femor women ankl bone exercis cruciat arthroscop rehabilit surgeri flexion tibial hospit shoulder score dementia radiograph cancer nurs	CLI1
15	speech phonolog semant lexic word task children sentenc auditori memori cognit perceptu verb cue languag stimuli stimulu ltd speaker patient vowel neuropsycholog erp aphasia verbal noun hear distractor syllabl stutter listen	COGN
16	patient schizophrenia adolesc children nurs women health disord depress symptom psychiatr clinic anxieti mental student suicid social smoke abus ptsd emot hospit interview cognit psycholog child physician ltd questionnair sexual	PSYC
17	neuron rat patient receptor brain cortex mice seizur epilepsi hippocamp synapt cell axon gaba hippocampu cortic protein ltd cerebr stroke dopamin nmda sleep astrocyt spinal inc motor nerv diseas gene glutam eeg	NEUR
18	protein acid milk diet gene ferment cell cow chees intak enzym meat starch fat dietari coli ltd strain broiler ph dna food carcass fed bacteria fatti rat antioxid dairi mutant yeast	BIOC
19	cell protein gene receptor mice rat tumor kinas patient bind transcript mrna cancer apoptosi dna mutat il phosphoryl mutant inhibitor inhibit ca2 peptid insulin acid enzym mous tissu beta vitro	BIOS
20	infect viru hiv vaccin patient dog protein cell antibodi viral gene pcr clinic hors mice strain antigen immun hcv parasit diseas rna malaria cd4 tuberculosi assai serotyp influenza virus pneumonia	MBIO
21	patient tumor surgeri carcinoma cancer postop lesion surgic clinic resect liver cell laparoscop diseas hepat endoscop arteri therapi ct gastric pancreat flap tissu preoper biopsi histolog mri malign tumour bone corneal	CLI2
22	patient cancer clinic arteri coronari renal diseas therapi transplant tumor diabet blood cell ventricular hypertens surgeri cardiac asthma hospit myocardi pulmonari lung children stent dose women prostat serum aortic graft	CLI3

sciences. Cluster #14, #21 and #22 represent different subfields of clinical and experimental medicine, and are therefore labeled (CLI1 through CLI3). CLI1 represents issues like health care, physiotherapy, sport science and pain therapy while CLI2 and CLI3 share many terms (cf. Table 4.6), but have somewhat different focus as can be seen on the basis of the most important journals (cf. Table 4.5). Finally, clusters #18 (BIOC), #19 (BIOS) and #20 (MBIO) stand for biochemistry, biosciences and microbiology, respectively (see [47]). It should be noted that links and overlaps among the life-science clusters are rather strong. The last group is formed by the social sciences and humanities (four clusters in total). Cluster #12 (ECON) is labeled as economics and business, cluster #2 (HUMA) represents the humanities and clusters #1 (SOC1) and #13 (SOC2) two different subfields on the social sciences. While SOC1 stands for educational sciences, cultural sciences and linguistics, SOC2 represents sociology, geography, urban studies, political science and law.

The 22 clusters are more or less strongly interlinked (cf. Figure 4.5). The strong links between clusters #6 and #7, #7 and #8 or the "chain" leading from #18 to #21 via #19 and #22 might just serve as an example. Therefore we have combined those clusters which are strongly interlinked to larger structures. These "mega-clusters" are presented in Figure 4.6. The first mega-cluster is formed by the social sciences clusters (SOC1, SOC2, ECON and HUMA). The second one comprises MATH and COMP and the third one is formed by the natural and engineering sciences (without mathematics and computer science). Biology, agricultural, environmental and geosciences (BIOL, AGRI, GEOS) form the fourth mega-structure. The fifth and sixth one are formed by the biomedical clusters and the neuroscience clusters, respectively. The large neuroscience cluster (#15 - #17) acts as a bridge connecting the life-science mega-cluster with the social sciences and humanities, whereas the agricultural-environmental mega-cluster connects the life sciences with the natural and applied sciences (cf. Figure 4.6).

4.6 Discussion

4.6.1 The analysis of mutual information based weighted hybrid clustering

In this scheme, the view that shares larger mutual information with other views is assigned a larger weight. To certain extent, mutual information can measure the consensus pattern shared by different views. Accordingly, in our weighting scheme, the view containing the most consensus pattern plays the most important roles during the joint clustering. Hence, the hybrid strategy

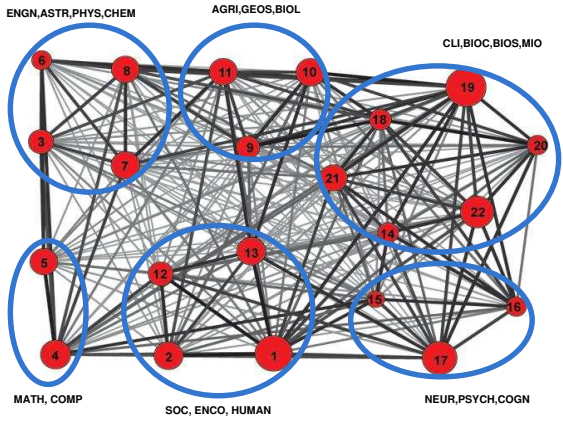

Figure 4.6: Subgroups of the WoS journal network by weighted hybrid clustering

emphasizes the consensus pattern (not the complementary information) among multiple views, which can reduce the noise from each view respectively, thus leading to a robust clustering result. (The effect of this multi-view clustering analysis is similar to that of the multi-view learning demo as presented in the Figure 1.3).

4.6.2 comparison of various weighting schemes

W.r.t weighting schemes for multi-view data, the only difficulty, of course, is how best to select the weighting factor λ. The value of the weighting factor can be set using cross-validation over several choices for its value [109]. Janssens employs Silhouette Value per Clustering (SVC) for each data type to estimate the relative quality of each data source; and he applies this estimation as an educated guess for weights of text mining data and bibliometric data [69],

$$\lambda = \frac{SVC^t}{SVC^t + SVC^{bc}}, \tag{4.10}$$

where λ denotes the weighting factor of text mining data, SVC^t refers to the SVC value of text mining data and SVC^{bc} refers to the SVC value of bibliometric data. When using a large number of kernels, however, this weighting strategy is no longer practical because of the intensive computation of the SVC on each view. Hence, more appropriate approach for weighting the different kernels is required.

Rather than requiring that weights be assigned a prior, Lanckriet and coworkers train a Support Vector Machine (SVM) and learn the kernel weights simultaneously, using a technique known as Semidefinite Programming (SDP) [84, 85]. Yu et al. have also carried out similar work to seek an automatic weighting scheme [145]. But the heavy computation of these multiple kernel learning based strategies seems not suitable for large-scale databases. Thus it triggers our ANMI based weighting scheme whose implementation is more efficient than above weighting strategies when handling large-scale data. However, the scalable issue still remains: our ANMI based weighting scheme will become inefficient as the number of views is increased, because the partition of each view data is involved in our ANMI based weighting scheme.

4.6.3 Comparison of various multi-view clustering schemes

In this Subsection, we mainly provide a comparison between multilinear based multi-view clustering methods and clustering ensemble methods. Multilinear based multi-view clustering methods refer to the clustering strategies proposed in Chapter 2 and Chapter 3.

First, clustering ensemble strategy focuses on the integration of multi-view data on the partition level. Therefore the computation of the clustering ensemble (only the clustering labels, instead of the original data, are involved) mainly depends on the partitioning step of each single-view data. Furthermore, its clustering performance is closely related to the number of views (each partitioning corresponds to one view) [91]. The more the number of views, the better its clustering performance.

While multilinear based multi-view clustering strategy focuses on the integration of multi-view data on the similarity matrix level. Its computation relies on the relevant multilinear operations, such as tensor decomposition. The clustering performance does not depend on the number of views.

Next, compared with the partition level of multi-view data, the similarity matrix level contains much more information. Thus more information is utilized. This difference may lead to better clustering performance of multilinear based multi-view clustering strategy. (see Table 2.1, 2.3, 2.5, 3.2, 3.5).

Third, the original clustering ensemble strategy takes each view equally (in this research, we extend some clustering ensemble methods to their weighted version) while the multilinear based multi-view clustering strategy can leverage the effect of multiple views automatically.

4.7 Summary

We proposed an ANMI-based weighting scheme for hybrid clustering and applied this scheme to a real application to obtain the structural mapping of a large-scale journal database. The main contributions are concluded as follows.

We presented an open framework of hybrid clustering to combine heterogeneous lexical and citation data for journal sets analysis from the scientometric point of view. We exploited two main approaches in this framework as clustering ensemble and kernel-fusion clustering. The performance of all approaches has been cross compared and evaluated using multiple statistical and information based indices.

The analysis of lexical and citation information in this research was carried out at more refined granularities. The lexical information was represented in five independent data sources by the different weighting schemes of text mining. The citation information was also investigated with five different views, resulting in five independent citation data sources. These lexical and citation data sources were combined in hybrid clustering as refined representations of journals. On the basis of the ANMI, we proposed an efficient weighting scheme for hybrid clustering. Three clustering algorithms were extended using the weighting scheme and they were systematically compared with the concerned algorithms using multiple evaluations.

To thoroughly investigate the journal clustering result, we visualized the structural network of journals on the basis of citation information. We also ranked the journals of each partition using a modified PageRank algorithm. Furthermore, we provided multiple textual labels for each cluster on the basis of text mining results. The obtained journal network integrates lexical and citation information and it can be employed as a good reference for journal categorization. The proposed method is also efficient to be applied in large-scale data to detect new trends in different scientific fields. The proposed weighted hybrid clustering framework can also be applied to retrieve multi-aspect information, which is useful to a wide range of applications pertaining to heterogeneous data fusion (i.e., bioinformatics research and Web mining).

In this Chapter, we focus on hybrid clustering in vector spaces, that is, both text and citation data are modeled in vector spaces. According to our experimental analysis, it seems the computation of this hybrid clustering strategy is still intensive as the data set becomes larger. Consequently, in Chapter 5, we will investigate hybrid clustering in graph spaces, with the aim of achieving efficient implementation against scalable data.

Chapter 5

Scientific mapping by hybrid clustering in graph spaces

5.1 Introduction

The objective of this research is an accurate unsupervised clustering of science or technology fields, towards the detection of new emerging fields. The idea of combining citation information with textual content is not new for it has already been pursued to obtain improved performance in information retrieval [23], bibliometric mapping of science [16, 17, 50, 72, 100, 126], clustering [97, 138], and classification issues [22, 74]. Sometimes textual information can indeed indicate similarities invisible through citation links, and vice verse. On the other hand, based on text alone, true document similarity might be obscured by differences in vocabulary use, or spurious similarities might be introduced as a result of textual pre-processing like stemming, or because of polysemous words or words with little semantic value. For instance, documents about music information retrieval might erroneously be linked to patent-related research based on common terms used in both contexts, such as title, record, creative and business. Consequently, the combination of textual data and citation data is thought as a promising method to deal with scientific publication. Some hybrid clustering has been carried out, such as Janssens *et al.* [71, 72] put forward two hybrid clustering strategies based on a convex combination of distance matrices method (WLCDM) and Fisher's inverse χ^2 method respectively. Most of the applied hybrid methods are based on vector space model.

Due to the growth of information and the availability of huge databases during the last decades, handling the amount of data has become a real challenge to information science. Hybrid clustering in vector spaces is usually limited to tackle scalable data. Furthermore, the number of clusters is often considered to be known or is estimated based on a variety of measures. Therefore a non-parameter or no-hypothesis clustering is needed. In addition, clustering methods often return a one level cluster structure which does not always reflect the nature of data structure correctly because real data is pretty complicated. Hence, multi-level of cluster structure or hierarchical cluster structure should be preferred.

With widely-available large-scale networks in various fields, community detection is gaining increasing attention from a variety of disciplines including physics, economics, epidemiology, business marketing and bioinformatics. The extracted communities can be utilized for further analysis such as visualization, viral marketing, determining the causal factors of group formation, detecting group evolution or stable clusters, relational learning and building ontology for semantic web [131]. However, in the last few years, there has been a concerted interdisciplinary effort to develop mathematical tools and computer algorithms to detect community structure in large networks. Such a problem is often computationally intractable and therefore requires approximation methods in order to find reasonably good partitions in a reasonably fast way. The rapidity of the algorithm has become a crucial factor due to the increasing size of the networks to be analyzed. A large variety of methods have been developed in order to address this problem [42].

In particular, the recent method called "Louvain method", which is based on approximate modularity optimization, outperforms the alternative methods in terms of computation time, while having an excellent accuracy [14]. The Louvain method has been employed in the analysis of scientific knowledge. Lambiotte and Panzarasa [83] discussed the community detection of a scientific collaboration network by Louvain method. Rafols and Leydesdorff [88] investigated the clustering of Louvain method of the 2006 edition of the Journal Citation Report (JCR) and compared the results with those of other three classification schemes. In former research, we have used this method to cluster the subjects structure of the WoS based on ISI classification based on citation link data [149]. However, graph partitioning methods usually focus on link structure and ignore attribute similarities. Therefore, we put forward a hybrid strategy based on network (graph) model to deal with the clustering problems mentioned before. Our strategy is able to facilitate clustering tasks by several ways: combining citation links and textual information, being the self optimizing and providing a hierarchical analysis.

In a related approach, He *et al.* [59] implemented the combination of hyperlink

structure and textural similarity to cluster the Webpages. Here we use the cross-citation link and the text based k-Nearest Neighbor (KNN) relationship and modularity optimization based on Louvain method.

The rest of this Chapter is organized as follows. At first, the database of WoS is introduced in Section 5.2. Then, community detection based on modularity optimization is presented in Section 5.3. Next, we study the hybrid strategy of community detection in Section 5.4. The experimental results are demonstrated in Section 5.5. Conclusion can be found in Section 5.6.

Scientists studying community detection and those studying data clustering are apparently looking at the same coin [116], thus, we employ the term of both community detection and clustering in this research alternatively.

5.2 Data sources and methodology

The raw dataset contains more than 6,000,000 publications (articles, letters, notes, and reviews) indexed by the WoS database of Thomson Reuters for the period 2002-2006. At first, these publications are aggregate to journal level in order to easily handle the scientific mapping of this large scale data. In pre-processing, the ambiguities of journal names, author names and bibliographic data are resolved. We only keep the journals with both at least 50 papers and more than 30 citations. After pre-processing, we obtain 8,305 journals as the data set adopted in this research (the detail can be referred to Chapter 4).

5.2.1 Text mining analysis

In the first step, we analyzed text mining data according to Chapter 4. All textual content was encoded in vector space model using the TF-IDF weighting scheme [7]. The paper-by-term vectors are then aggregated to journal-by-term vectors as the representations of the lexical data. Text-based similarities were calculated as the cosine of the angle between the vector representations of two journals [120].

According to [94], there are two ways to transform a given set of data points with pairwise similarities into a graph: (1) $\varepsilon-$ neighborhood method which reduces the pairwise similarities smaller than a threshold ε; (2) $k-$ nearest neighbor method which lets the pairwise similarity between node i and j exist only if i is among the KNN of j. Because the threshold ε is hard to determine and there is no guarantee that the graph is still connected as some similarities

below the threshold are deleted. Consequently, we adopt the strategy of KNN to build the text based graph.

5.2.2 Citation analysis

In the second step, we analyzed the cross-citations links among the selected journals. The citation data is generated by information extraction from the WoS database. Citations among individual papers were aggregated to the journal level. We ignored the direction of citation links by symmetrizing the cross-citation matrix. When we analyzed the cross-citation links among the selected journals, a problem arises: some journals own huge cross-citations while some journals share tiny cross-citations. For instance, the largest number of cross-citations is 37162; the smallest number of cross-citations is 1 and the mean number is 10. This heavily uneven distribution is caused by the various number of papers within each journal. That uneven paper distribution leads to the detection of wrong communities by modularity based methods. Therefore we normalize the cross-citation matrix before using it as an adjacency matrix for community detection in the following way [149],

$$\mathbf{A}_{ij} = \frac{\mathbf{C}_{ij}}{\sqrt{(\sum_u \mathbf{C}_{iu})}\sqrt{(\sum_u \mathbf{C}_{uj})}}, \tag{5.1}$$

where $\mathbf{C}$ is the raw cross-citation matrix; $\mathbf{C}_{ij}$ is the raw cross-citations between journal i and journal j and $\mathbf{A}_{ij}$ is the normalized cross-citations between journal i and journal j. Though the representation of citations actually forms a sparse graph, we can regard it as journal-by-citation vectors as well, where the similarities of journals are measured by the cosine value of journal-by-citation vectors.

5.3 Community detection by modularity optimization

5.3.1 Modularity

Modularity is a benefit function used in the analysis of networks or graphs. Its most common use is as a basis for optimization methods for detecting community structure in networks (graphs) [107]. In this research, all graphs are regarded as weighted graphs and modularity is defined as [105]

$$Q = \frac{1}{2m} \sum_{ij} \left[\mathbf{A}_{ij} - \frac{d_i d_j}{2m} \right] \delta(c_i, c_j), \tag{5.2}$$

where $\mathbf{A}_{ij}$ represents the weight of the edge between vertex i and vertex j; $d_i = \sum_j \mathbf{A}_{ij}$ is the sum of the weights of the edges attached to vertex i; c_i is the community to which vertex i belongs to; the δ function $\delta(u, v)$ is 1 if $u = v$ and 0 otherwise and $m = \frac{1}{2} \sum_{ij} \mathbf{A}_{ij}$. The value of the modularity lies in the range [-1,1]. It is positive if the number of edges within groups exceeds the number expected on the basis of chance. The fast approximation algorithm for optimizing modularity on large graphs was proposed by Clauset *et al.* [29]. That method consists in recurrently merging communities that optimize the production of modularity as denoted in the below,

$$\triangle Q = \begin{cases} \dfrac{\mathbf{A}_{ij}}{2m} - \dfrac{d_i d_j}{(2m)^2} & \text{if } i, j \text{ are connected,} \\ 0 & \text{otherwise.} \end{cases} \tag{5.3}$$

5.3.2 Louvain method [14]

Based on modularity optimization, the Louvain method incorporates a multi-level organization and consists of two phases repeated iteratively. First, the algorithm looks for "small" communities by optimizing modularity in a greedy, local way. Second, the algorithm aggregates nodes of the same community and builds a new network whose nodes are the communities. These phases are repeated iteratively until a maximum of modularity is attained and an optimal partitioning of the network into communities is found. The choice of this method for community detection is motivated by its excellent accuracy and its rapidity which allows us to study networks of unprecedented size (for instance, the analysis of a typical network of 2 million nodes only takes 2 minutes). The Louvain method has also been shown to be very accurate by focusing on ad-hoc networks with known community structure.

Moreover, due to its hierarchical structure, which is reminiscent of renormalization methods, it allows to look at communities at varied partition levels. The output of the program therefore gives several partitions. The partition found after the first step typically consists of many communities of small sizes. At subsequent steps, larger and larger communities are found due to the aggregation mechanism. This process naturally leads to hierarchical decomposition of the network [42]. Then the Louvain method can be regarded as a hierarchical partitioning method from the perspective of graph spaces.

Part of the algorithm's efficiency results from the fact that the gain in modularity $\triangle Q$ obtained by moving an isolated node i into a community C can easily be computed by

$$\begin{aligned}\triangle Q = \frac{1}{2}[\frac{\sum_{in}+2d_{i,in}}{2m} - (\frac{\sum_{tot}+d_i}{2m})^2] \\ -[\frac{\sum_{in}}{2m} - (\frac{\sum_{tot}}{2m})^2 - (\frac{d_i}{2m})^2],\end{aligned} \tag{5.4}$$

where $\sum_{in}$ is the sum of the weights of the links inside C; $\sum_{tot}$ is the sum of the weights of the links incident to nodes in C; k_i is the sum of the weight of the links incident to node i; $k_{i,in}$ is the sum of the weights of the links from i to nodes in C and m is the sum of the weights of all the links in the network. A similar expression is used in order to evaluate the change of modularity when i is removed from its community. In practice, one therefore evaluates the change of modularity by removing i from its community and then by moving it into a neighboring community. Louvain method has been applied to identify language communities in a Belgian mobile phone network of 2 million customers by analyzing a web graph of 118 million nodes and more than one billion links [81].

5.3.3 Finding communities at different resolutions

In order to uncover communites of different characteristic sizes, a tuning resolution parameter t is introduced to define a new quality function about the optimal partitions of a network [83]:

$$Q_t = (1-t) + \frac{1}{2m}\sum_{i,j}[t\mathbf{A}_{ij} - \frac{d_i d_j}{2m}]\delta(c_i, c_j), \tag{5.5}$$

where t is a resolution parameter. When $t = 1$, the above function is equivalent to modularity as defined in (5.2).

The introduction of the resolution parameter t is able to circumvent the problem of resolution limit which modularity suffers from in its original formulation as in (5.2). Resolution limit by modularity optimization means that some communities smaller than certain threshold size could not be detected due to the size of network and the extent of interconnectedness of its communities [43]. Communities smaller than the threshold tend to be merged into larger communities, thereby missing important structures [116]. When t is decreased, some smaller communities could be detected and in the limit cases $t = 0$, the optimal partition is made of N single nodes. On the other hand, when t is

increased, the optimal partitions are made of larger and larger communities. In particular, the optimal partitions of the limit cases $t = \infty$ are made of one community containing the whole network respectively.

5.3.4 Matrix formulation of modularity maximization

An alternative formulation of the modularity, useful particularly in spectral optimization algorithms, is as follows [106]. Define $\mathbf{U}_{ir}$ to be 1 if vertex i belongs to group r and zero otherwise, then

$$\delta(c_i, c_j) = \sum_r \mathbf{U}_{ir}\mathbf{U}_{jr}, \tag{5.6}$$

and hence

$$Q = \frac{1}{2m}\sum_{ij}\sum_r [\mathbf{A}_{ij} - \frac{d_i d_j}{2m}]\mathbf{U}_{ir}\mathbf{U}_{jr} = \frac{1}{2m}\text{trace}(\mathbf{U}^T\mathbf{B}\mathbf{U}), \tag{5.7}$$

where $\mathbf{U}$ is the (non-square) matrix having elements $\mathbf{U}_{ir}$ and $\mathbf{B}$ is the so-called modularity matrix, which has elements

$$\mathbf{B}_{ij} = \mathbf{A}_{ij} - \frac{d_i d_j}{2m}. \tag{5.8}$$

Relaxing $\mathbf{U}$ to be continuous, it can be inferred that the optimal $\mathbf{U}$ is composed of the top k eigenvectors of the modularity matrix [106].

In fact, unlike Laplacian matrix in spectral clustering, the modularity matrix $\mathbf{B}$ is not guaranteed to be positive semi-definite. However, the modularity matrix can be regularized to guarantee that it is positive semi-definite [106]. Once $\mathbf{U}$ is obtained, the final partitioning could be obtained by k-means, as implemented in standard spectral partitioning.

Although this spectral analysis of modularity optimization is well formulated in term of linear algebra, this approach is neither appropriate nor realistic for community detection in most application contexts because one typically does not know the number of communities in advacne. Furthermore, this strategy needs extra regularization work and is incapable of providing the hierarchical partition structure as Louvain method.

5.4 Hybrid clustering by modularity optimization

Thanks to the merits of Louvain method, taking each journal as a vertex, we can directly carry out the clustering analysis of the database by cross-citation

data. Since the text data of WoS is available, we attempt to utilize the textual information which is supposed to complement the citation data. By combining these two types of information, it is expected to obtain a robust cluster structure. However, the question remains of how to deal with these two heterogeneous data in graph spaces.

5.4.1 Hybrid clustering by graph integration

A simple strategy for integrating the two heterogeneous graphs is to average their adjacency matrices as presented in (5.9). Because of the different measurements of these adjacency matrices, they should be normalized during the combination. Here we just normalize each adjacency matrix as the following,

$$\bar{\mathbf{A}} = \frac{1}{2}\left(\frac{\mathbf{A}^{(T)}}{\| \mathbf{A}^{(T)} \|_2} + \frac{\mathbf{A}^{(C)}}{\| \mathbf{A}^{(C)} \|_2}\right). \tag{5.9}$$

Correspondingly,

$$\bar{m} = \frac{1}{2}(m^{(T)} + m^{(C)}), \quad \bar{d}_j = \frac{1}{2}(d_j^{(T)} + d_j^{(C)}). \tag{5.10}$$

With $\bar{A}$, this hybrid strategy boils down to classic community detection in a single graph. Based on the integrated graph, we can obtain the modularity gain as the following,

$$\triangle \bar{Q} = \frac{\bar{\mathbf{A}}_{ij}}{2\bar{m}} - \frac{\bar{d}_i \bar{d}_j}{(2\bar{m})^2}. \tag{5.11}$$

In fact, how to normalize the multiple graphs in a proper way poses a challenge for this hybrid strategy because it is of difficulty to make different graph spaces with various statistic property compare with one another. Consequently, we continue seeking the hybrid partitioning strategy of multiple graphs ahead.

5.4.2 Hybrid clustering by graph coupling

Inspired by the research work in Webpage clustering [59], we are able to integrate the cross-citation with text in the following way. The link structure is determined by cross-citation, that is, if there is a cross-citation relationship between two journals, there will be a link, while the edge strength is determined by the textual similarity. Consequently, a coupled graph is generated and the Louvain method can be implemented on it to obtain the final partitioning result.

Given $\mathbf{A}^T$, the adjacency matrix of the text network and $\mathbf{A}^C$, the adjacency matrix of the citation network, the adjacency matrix of the coupled graph $\tilde{\mathbf{A}}$ is obtained in the following way,

$$\tilde{\mathbf{A}} = \frac{\mathbf{A}^{(T)} \otimes \mathbf{A}^{(C)}}{\| \mathbf{A}^{(T)} \otimes \mathbf{A}^{(C)} \|_2}, \tag{5.12}$$

where,

$$(\mathbf{A}^{(T)} \otimes \mathbf{A}^{(C)})_{ij} = \begin{cases} \mathbf{A}_{ij}^{(T)} & \text{if } \mathbf{A}_{ij}^{(C)} \neq 0 \\ 0 & \text{if } \mathbf{A}_{ij}^{(C)} = 0 \end{cases}. \tag{5.13}$$

In the empirical test on the coupled graph, we found that some edges with weak strength (weak textual similarity) have negative impact on the final partitioning. This kind of edge can be understood as although two journals are cross-cited (relevant) but they share less textual terms (unsimiliar). In a common sense, if two journals are cross-cited each other but share less textual terms, they should not be clustered into the same category. Therefore, to neglect their negative impact in the partitioning, we even go further to use the KNN constraint to filter out these edges. By this means, we strengthen the effect of those journal vertices which are cross-cited and share more textual similarity.

As compared to hybrid clustering in vector spaces [71, 72], our hybrid strategy is completely distinct, in particular:

- Out strategy does not require any previous setting, such as the number of clusters;
- Integration schemes are different; we use the cross-citation link structure to couple the textual similarity, plus the KNN constraint to neglect the un-useful edges;
- Partition schemes are diverse; in the vector space model, we can use k-means or Ward's linkage method while in the graph space model, the modularity optimization based on the Louvain method is employed;
- Final cluster structures are varied; in the vector space model, usually one level clustering result is provided. Whereas, in our strategy, the optimal hierarchical structure is offered which would more fit in with the practical tasks.

In addition, we grasp the core information of the textual and citation data while neglecting large data without sufficient information, thus our strategy is applicable to large-scale applications.

5.5 Experimental results

5.5.1 Fixing the community resolution (the number of clusters as 22) for comparison with standard ESI category

Experimental setting

In order to compare our hybrid clustering strategy (graph coupling) with other alternative clustering strategies in this research, we fix the cluster number the same as the number of standard ESI fields (used by Thomson Reuters). Otherwise, it is impossible to compare various clustering results with different cluster numbers.

In this case, with Louvain method, we need adjust the resolution parameter t as in (5.5) to seek the expected cluster structure with cluster number as 22. As we set the resolution parameter t as 0.4, such a cluster structure is obtained. In addition, unlike the partitioning by k-means which is sensitive to the initialization, the clustering result of our strategy is unique because of the partitioning property of Louvain method. The implementation of Louvain method is available [1].

Three vector space model based clustering solutions are compared: clustering on TF-IDF data, clustering on cross-citation data and hybrid strategy of WLCDM [70]. The final partitioning of these three clustering strategies is implemented by Ward's linkage method [68].

In graph spaces, clustering strategies of both spectral modularity optimization and Louvain method are implemented on TF-IDF data (we apply the top 100 nearest neighbour constraint), cross-citation data, graph integration and graph coupling respectively.

Two external evaluation measures are adopted to gauge the clustering performance against ESI fields as benchmark category. One is ARI [63], and the other one is NMI [127]. Both are trying to measure the overlap between clustering results and benchmark category. The larger the evaluation values, the better clustering performance.

Comparison of clustering performance among the clustering strategies

The clustering performance of the relevant clustering strategies based on NMI and ARI evaluation is shown in Table 5.1. The clustering results based on

[1] http://sites.google.com/site/findcommunities/

Table 5.1: The clustering evaluation with fixed number of clusters (22). Spectral: spectral modularity optimization; LM: Louvain method.

Models	Methods	NMI	ARI
Vector spaces	TFIDF	0.5080	0.2676
	CRC	0.4532	0.1604
	WLCDM	0.5161	0.2885
Graph spaces (Spectral)	TFIDF	0.5429	0.2945
	CRC	0.5645	0.3515
	Graph integration	0.5517	0.2939
	Graph coupling	0.5590	0.3100
Graph spaces (LM)	TFIDF	0.5309	0.29
	CRC	0.5640	0.3209
	Graph integration	0.5418	0.3013
	Graph coupling	0.5768	0.3407

Ward's linkage and Louvain method are unique. While w.r.t the clustering results by spectral modularity optimization, we take the average values by repeating 50 times.

Comparison of different clustering models. As can be seen in Table 5.1, clearly, the clustering performance by graph model is beyond that by vector space model, both on spectral modularity optimization and on Louvain method. For instace, the NMI value of cross-citation data is increased from 0.4532 by partitioning in vector spaces (Ward's linkage) to 0.5645 by spectral partitioning (0.5640 by Louvain method).

The reason why the vector space model does not work well on our data might be two-fold: (1) the high-dimension of TFIDF feature leads to the failure of the related clustering algorithms which are successful in low-dimension vector spaces; (2) the natural feature of cross-citation data is link structure, instead of vector structure, and in consequence, clustering in vector spaces may undermine its original structure information.

Regarding the two graph spaces based clustering models, as shown, the spectral modularity optimization achieves the almost the same clustering performance as the Louvain method. However, Louvain method is able to provide a flexible clustering analysis framework (the optimal cluster number and the hierarchical partition structure) as demonstrated in Subsection 5.5.2. Furthermore, the efficiency of Louvain method is remarkable as presented in Table 5.3, for instance, the partitioning of graph coupling with spectral

Table 5.2: The comparison of link strength of text graph and citation graph

Link strength	text graph	citation graph
Minimum	7.1e -8	2.2e-7
Mean	8.5e -7	6e -6
Maximum	2.495e -6	1.704e -4

modularity optimization consumes 949 seconds while the partitioning of graph coupling with Louvain method only needs 12 seconds.

Comparison of different hybrid clustering strategies. Table 5.1 also provides the comparison between the two hybrid strategies of graph integration and graph coupling. It is obvious that the strategy of graph coupling is winner, in particular, based on Louvain method. For instance, ARI value is improved from 0.3013 by graph integration to 0.3407 by graph coupling.

The failure of graph integration may be due to the lack of proper normalization scheme before integrating the two heterogeneous graphs (text and citation). Table 5.2 gives the comparison of the link strength of these two graphs after normalization. Although the different graphs are normalized by (5.9) before integration, the link strengths of them still seem incomparable. As shown, it is clear that the link strength in citation graph is much higher than that of text graph. For instance, the mean link strength of text graph is 8.5e -7 while the mean link strength of citation graph is 6e -6. Consequently, when these two graphs are integrated, the citation graph dominates the structure information of the integrated graph and to some degree, the link strength of text graph can only be regarded as noise.

On the other hand, graph coupling is able to circumvent this comparable integration problem by coupling link structure of one graph with the link strength of the other graph. Consequently, this hybrid strategy proved to be the best clustering method according to both NMI and ARI evaluation. For example, as presented in Table 5.1, regarding graph coupling with Louvain method, the NMI value is increased from 0.5418 by graph integration to 0.5768 by graph coupling.

Comparison of computation time among the clustering strategies

Table 5.3 provides the comparison of the computational time of the related algorithms. The experiment was carried out on a CentOS 5.2 Linux system with a 2.4G Hz CPU and 16 G Bytes memory. As shown, the clustering

Table 5.3: The comparison of running time by different clustering schemes

Models	Methods	Time (seconds)
Vector spaces	TFIDF	624
	CRC	631
	WLCDM	760
Graph spaces (spectral)	TFIDF	591
	CRC	770
	Graph integration	712
	Graph coupling	949
Graph spaces (LM)	TFIDF	8
	CRC	9
	Graph integration	12
	Graph coupling	12

solutions by Louvain method in the graph model are distinctly (by almost two orders of magnitude) faster than their counterparts in the vector spaces. The heavy computation of spectral modularity optimization is caused both by the EVD of modularity matrix, the elements of which are full connected. Meanwhile, Louvain method is a kind of hierarchical partitioning, thus leading to a much efficient partitioning. In fact, we attempted to implement the above partitioning directly by the k-means clustering method (the typical clustering strategy in vector space model) for comparison with above clustering strategies mentioned. However, we found the partitioning of k-means clustering is incomparable to the relevant schemes (k-means clustering requires more than 432,000 seconds to partition the TFIDF text data). The high dimension of the feature vector (669,860 terms in text feature) is probably a major cause of the clustering failure of vector space model based clustering strategies on large-scale data. Although some dimension reduction strategies, like LSI, can boost the computation of the vector space model based clustering, they still require some extra heavy computation, such as, the SVD of the original large data. Therefore, as the volume of the data increases, the computation problem of these strategies still remains.

As a whole, the excellent performance of the hybrid strategy of graph coupling by Louvain method is based on two facts: (1) the heterogeneous information from different graph spaces is fused in an appropriate way (without normalizing the multiple heterogeneous data before integration); (2) the partitioning efficiency of the Louvain method in terms of computation time as presented in Table 5.3.

Furthermore, during the implementation of these algorithms, the comparison of memory consumption also indicates the similar trend, demonstrating the efficiency of Louvain method. Due to the page limitation, we omit the detail of that comparison.

5.5.2 The hierarchical clustering structure optimized by the Louvain method

Based on the above results, we decided to implement the optimal clustering by graph coupling, that is, both the optimal cluster number and the optimal hierarchical structure are automatically found during the partitioning process. No input parameters are thus needed for this optimum partitioning, only except the adjacency matrix of this coupled graph, which is generated by combining the cross-citation link structure with textual similarity. The partitioning strategy of the Louvain method is able to find the optimal cluster number by maximizing the modularity. In this case, the resolution parameter t in (5.5) is set as 1. When a local maximum modularity is reached, the number of clusters and its related clustering structure are recorded. Because of the aggregating optimal mechanism, the different cluster structure has a hierarchical relationship. Thus we can obtain a graph structure with various partitioning levels.

We have stopped at two levels of partitions obtained by the hierarchical strategy. At the higher level, we got 9 clusters for the coupled graph while in the lower level, we obtained 45 clusters. The cognitive analysis of the higher level of 9 clusters is denoted below.

Since the optimal cluster number (9 or 45) differs from the number of ESI fields (22), we can not gauge the clustering performance against the two evaluations of the previous subsection. However, we still can take the ESI as a reference standard to determine if our clustering results are meaningful by finding any concordance between them. As shown in Figure 5.1, the concordance between our clustering solution (9 clusters) and the ESI scheme is visualized by gray-scaled cells representing the Jaccard index [66] for each cluster and field pair. The darkest cells represent the best-matching pairs of fields and clusters. It is clear that each of our 9 clusters corresponds to one ESI field or several relevant ESI fields. For instance, cluster #7 corresponds to the Clinical Medicine, # 6 corresponds to the fields of Economic & Business and Social Sciences, and #3 corresponds to Computer Science, Engineering and Mathematics.

To better understand the structure of clustering, we applied a modified Google PageRank algorithm [73] to analyze the journals within each cluster. Using

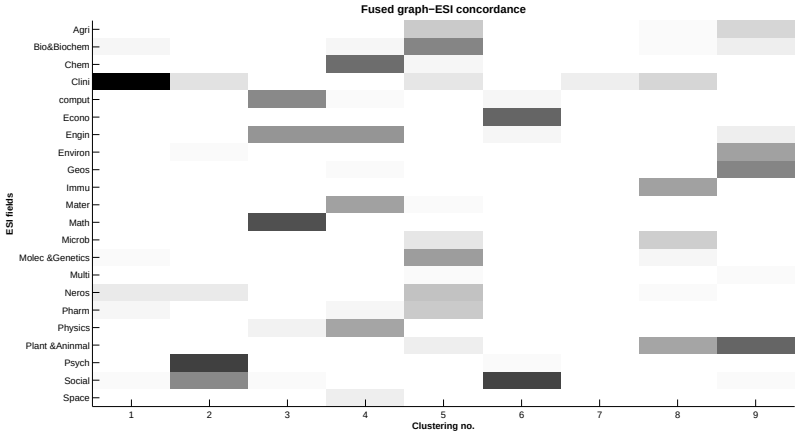

Figure 5.1: The concordance between 9 clusters obtained by our strategy and the 22 ESI categories. The first level of the hierarchical cluster structure. [Data source: Thomson Reuters, Web of Science]

Table 5.4: The five most important journals of the 9 clusters, the first level of the hierarchical clustering structure.

Cluster 1	Cluster 2	Cluster 3
1.Nat Rev cancer	1. Annu Rev Psycho	1. J Roy Stat Soc S.B
2. Caner cell	2. Psycho Meth	2. Fund comp math
3. CA-cancer J Clin	3. Psych B.	3. Biostat
4. Annu Rev Med	4. Psych Rev	4. J Amer Math Soc
5. B. B. Rev cancer	5. Behav Brai Sci	5. Anna Math
Cluster 4	Cluster 5	Cluster 6
1. Rev Mod Phys	1. Nat Rev molec cell bio	1. Quart J econ
2. Nat material	2. Nat Rev genetics	2. J econ liter
3. Chem Rev	3. Deve cell	3. J finance
4. Annu Rev Astron & Astrop	4. Nat Rev neruos	4. J finance econ
5. Mate sci & eng Rep	5. Annu Rev Bioche & 5. J poli econ	
Cluster 7	Cluster 8	Cluster 9
1. Prog retin & eye res	1. Nat Rev immu	1. Annu Rev Ecolog evo & Sys
2. Invest ophth & visua sci	2. Annu Rev Immu	2. Ocean & Marin Bio
3. Surv ophth	3. Nat immu	3. Syst Bio
4. Molec vision	4. Nat Medi	4. A Muse Novi
5. Archi ophth	5. J Exp Med	5. Annu Rev Entom

Table 5.5: The 30 best TF-IDF terms describing the 9 hybrid citation-textual clusters, the first level of the hierarchical clustering structure.

Cluster	Best 30 terms
1	patient tumor cancer clinic cell arteri diseas therapi surgeri carcinoma renal diabet coronari lesion transplant pain postop surgic gene blood dose bone breast prostat women liver resect hospit rat protein
2	patient children schizophrenia student health adolesc nurs women disord depress symptom clinic cognit school teacher mental psychiatr social educ anxieti hospit smoke emot suicid psycholog interview child questionnair abus sleep
3	algorithm algebra graph fuzzi finit wireless theorem antenna wilei queri polynomi semant nonlinear robot asymptot qo equat packet infin bandwidth xml network user scheme multicast manifold fault server nois bit
4	film temperatur alloi crystal atom ion polym quantum catalyst galaxi dope magnet metal oxid hydrogen diffract optic particl thermal wilei bond beam spin spectroscopi spectroscopi rai angstrom electron si spectra
5	protein gene cell receptor rat neuron mice kinas bind mutant transcript acid mrna dna ca2 phosphoryl mutat enzym inhibit peptid inhibitor apoptosi membran beta genom brain mous insulin muscl rna
6	polit firm polici market price social busi tax wage economi capit organiz war trade welfar reform court parti democraci labour corpor invest women discours democrat countri employe pavement econom compani
7	corneal retin ey patient glaucoma acuiti iop iol macular cataract intraocular lasik ocular surgeri cornea retina len choroid postop myopia vitrectomi astigmat refract phacoemulsif rpe ophthalmolog cnv retinopathi vitreou keratoplasti
8	infect cell patient il hiv viru vaccin mice protein gene antibodi immun antigen cd4 ifn cow diseas dog clinic cytokin receptor cd8 viral milk pcr lymphocyt calv serum hla hcv macrophag
9	soil speci plant forest sediment habitat water lake basin ocean river biomass season sea fish leaf cultivar seedl rock seismic seed predat temperatur fault climat larva veget isotop ecosystem rainfal

the algorithm, we investigated the five most highly ranked journals in each cluster and presented them in Table 5.4. Moreover, for the journals presented in Table 5.4, we re-investigated the titles, abstracts and keywords that have been indexed in the text mining process. The indexed terms were sorted by their frequencies and for each cluster, the thirty most frequent terms were used to label the obtained clusters. The best TF-IDF terms of each journal cluster are denoted in Table 5.5.

For the lower level partitioning with 45 clusters, the best 15 representative terms in each cluster are shown in Table 5.7. The top three journals of each cluster are listed in Table 5.6 (Cluster #13 only has two journals). Its cluster structure is visualized in Figure 5.3 analogously to Figure 5.2. We illustrate the hierarchical structure between the two partitions in Figure 5.4, which provides different resolutions of the scientific mapping. The clusters of different partitioning level are annotated by the number of journals within this cluster and its related subjects. For instance, cluster #1 stands for clinical medicine and neuroscience, and has the following substructure: 9 subfields including obstetrics, dental dermatology, cancer, medicine, surgery, audio, neurology and bone.

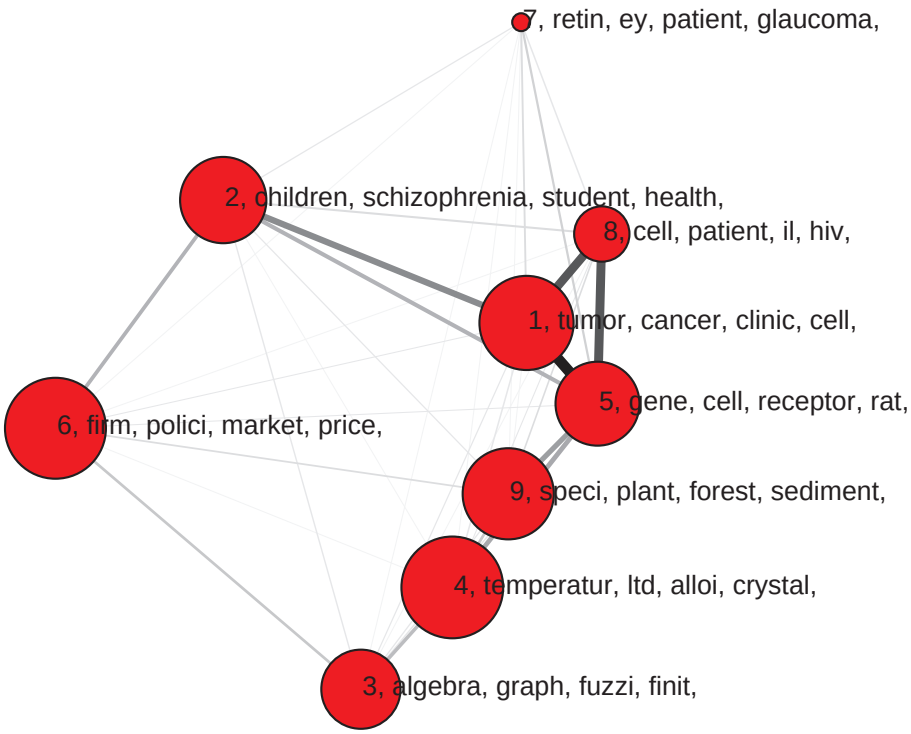

Figure 5.2: Network structure of the 9 journal clusters, the first level of the hierarchical clustering structure. [Data source: Thomson Reuters, Web of Science]

5.6 Summary

In this study we have presented a new hybrid clustering strategy based on graph model, which proved efficient and extremely fast. It is able to automatically provide optimum partitions at several hierarchical levels without any previous input. By combining textual and citation information, the strategy provided more robust cluster structures than hybrid clustering strategies based on vector space model. Even for given number of clusters, the new method outperformed analogous cluster algorithms based on the vector space model.

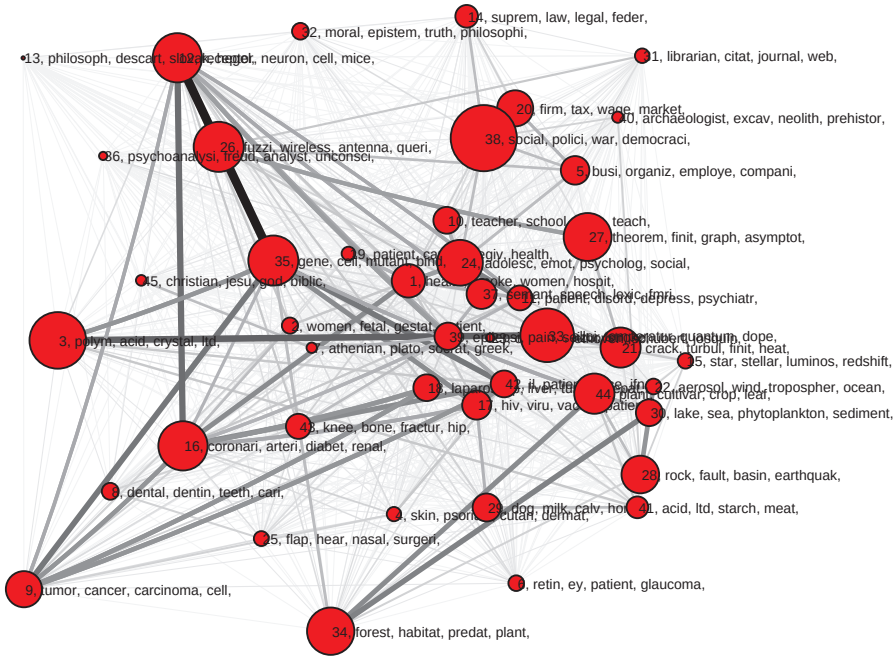

Figure 5.3: Network structure of the 45 journal clusters, the second level of the hierarchical clustering structure. [Data source: Thomson Reuters, Web of Science]

The self-optimization scheme of the Louvain method provided an optimum two-level hierarchical cluster structure. The cognitive analysis based on the textual component provided information for labelling and term annotation; the ranked journals and the visualization of the cluster structure also verified the validity of the new strategy.

The hybrid strategy is expect to provide a powerful tool to scientometrics and informetrics, as it can handle large-scale data, carry out the immediate partitioning, automatically optimize the cluster and provide a hierarchical system in practically one process and in an very short time. Because of the close relationship between bibliometric analysis and Web mining, our hybrid partitioning strategy can be directly extended to detect the communities in Webpages and Web social networks.

The key point of our strategy is to utilize the complementary property of

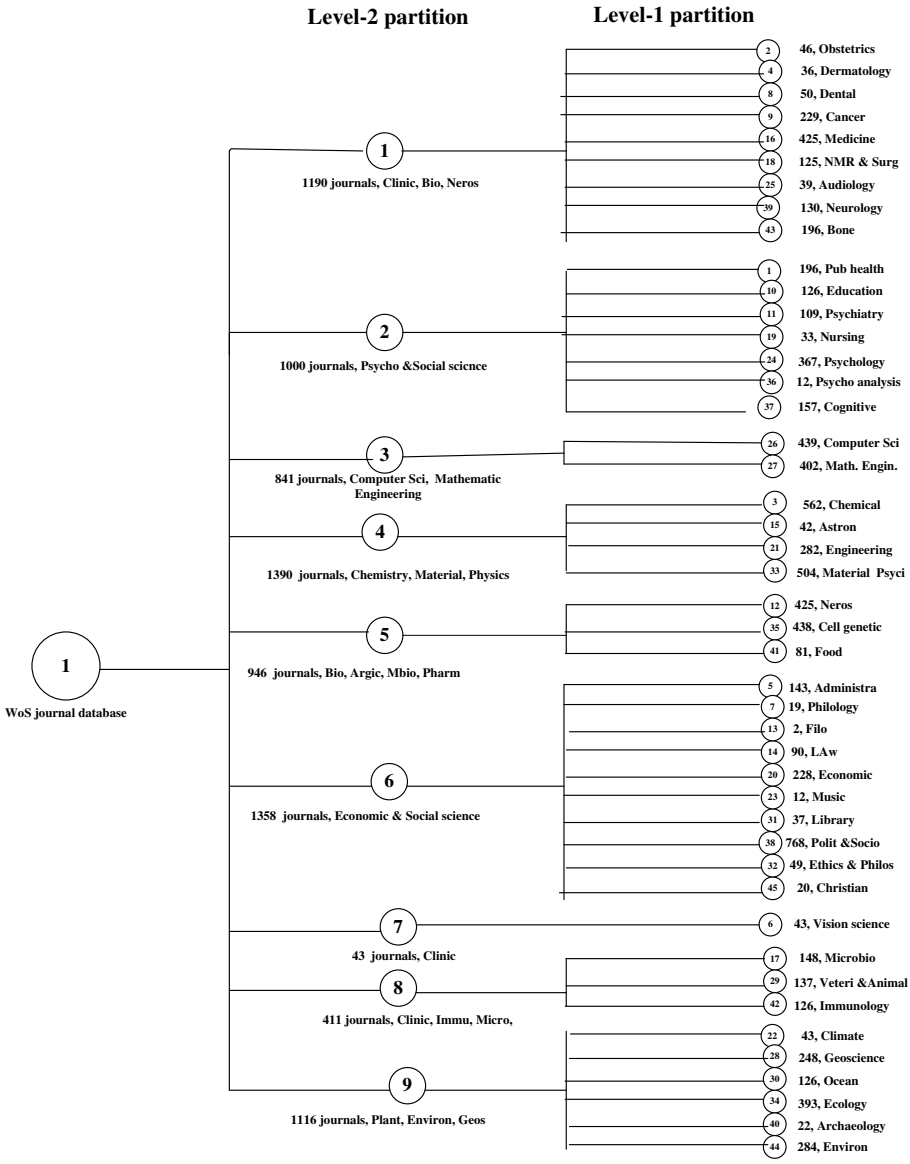

Figure 5.4: The hierarchical structure of the whole WoS journal database. Each cluster is annotated by both the number of journals it owns and the subject information. [Data source: Thomson Reuters, Web of Science]

Table 5.6: The three most important journals of the 45 clusters, the second level of our hierarchical clustering structure.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1. Milkb Quar	1. Twin Res	1. Chem R	1.J Inv Derm S P	1. Admi Sci Q
2. A R P Healt	2. I J Obst & G	2. P Ploy S	2. A J Clin Derm	2. Mis Quar
3. A J Epide	3. Hum Repr	3. Acc Chem R	3. J Inve D	3. Aca Mana J
Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
1. P. R. eye R	1. Class Antiq	1. Crit R Ora B M	1. N R cancer	1. R Educ R
2. Inv Ophth & V	2. T A Philo Asso	2. J Dent R	2. Cancer cell	2. A Educ R J
3. Sur Ophth	3. A J Philo	3. Dent Mater	3. Ca-cancer J C	3. Educa Eval P A
Cluster 11	Cluster 12	Cluster 13	Cluster 14	Cluster 15
1. A G Pysychi	1. N R Neros	1. Filoso Casopis	1. Yale Law J	1. A R Astr. & A
2. Molec Psychi	2. Physi R	2. Filozo	2. Univ Chica L	2. Astrop J S
3. Bio Psychi	3. A R Neros		3. Stanf Law R	3. Astroph J
Cluster 16	Cluster 17	Cluster 18	Cluster 19	Cluster 20
1. A R Med	1. N R Microb	1. Gastroe	1. Geronto	1. Quart J Econ
2. N E J Med	2. Clin Microb R	2. Anna Surg	2. A J Criti Care	2. J Econ L
3. Circul	3. Lanc Infec D	3. Hepato	3. Nurs Res	3. J Finace
Cluster 21	Cluster 22	Cluster 23	Cluster 24	Cluster 25
1. A R flu Mecha	1. J Hydrom	1. J A Musico S	1. A R Psych	1. Audio Neuro
2. P E Comb SCi	2. Clima Dynam	2. Musi Theo S	2. Psych Meth	2. Ear & Hear
3. J Mecha Phys S	3. J Clima	3. Music Anal	3. Psycho Bull	3. Laryngp
Cluster 26	Cluster 27	Cluster 28	Cluster 29	Cluster 30
1. A Comp Surv	1. J R Stat S S B	1. R Minerl & G	1. Veter Res	1. Oce. & M. B
2. J Acm	2. BioStat	2. Earth Sci Rev	2. J Feli Med & S	2. Fish & Fisher
3. J Machi learn R	3. J A Mat Sco S B	3. A R Earth & P S	3. J Dairy SCi	3. P Oceanog
Cluster 31	Cluster 32	Cluster 33	Cluster 34	Cluster 35
1. L & I Sci R	1. Ethics	1. R Mode Physi	1. A R Ecolo E	1. N R Molec cell
2. P Libra & Acad	2. Philos & P A	2. N Mater	2. Syst Bio	2. N R genetics
3. Colle & Res L	3. J Philos	3. Mater S & E R R	3. A Muse Novi	3. Devel cell
Cluster 36	Cluster 37	Cluster 38	Cluster 39	Cluster 40
1. Psychoa Dial	1. Psycho R	1. A Polit S R	1. Lanc Neuro	1. J Anthr Archae
2. Psychoa quart	2. Behav & B S	2. A R Soc	2. Brain	2. A Antiq
3. J A Psycho Asso	3. Tre Cogn S	3. A Soci R	3. Anna Neuro	3. J Archa M & T
Cluster 41	Cluster 42	Cluster 43	Cluster 44	Cluster 45
1. C R food S & N	1. N R Immu	1. Exerc & S S R	1. Global Chan B	1. J Ear Chris S
2. I J food Microb	2. A R Immu	2. J bone Min R	2. Criti R P S	2. J Bib L
3. A J Gra & W	3. N Immu	3. Bone	3. Adva Envir R	3. N Testa S

multi-view data: the integration of the citation links of one view with the textual similarity of the other view. Nevertheless, the drawbacks of strategy are apparent: only limited to two views and only applicable to such a multi-view scenario: text content and citation links. We will tackle the hybrid clustering issue of integrating more views in the graph spaces in later research.

Table 5.7: The 15 best TFIDF terms describing the 45 hybrid citation-textual clusters, the second level of the hierarchical clustering structure.

Cluster	Best 15 terms
1	health;smoke;women;hospit;children;physician;alcohol;cancer;care;clinic;risk;medic;adolesc
2	pregnanc;women;fetal;gestat;patient;ivf;preterm;vagin;matern;uterin;cesarean;endometriosi
3	catalyst;polym;acid;crystal;ligand;wilei;nmr;ion;angstrom;bond;adsorpt;hydrogen;solvent;atom
4	skin;psoriasi;cutan;dermat;lesion;keratinocyt;dermatolog;melanoma;hair;clinic;acn;wound;cell;atop
5	firm;busi;organiz;employe;compani;market;custom;brand;retail;supplier;corpor;advertis;strateg
6	corneal;retin;ey;patient;glaucoma;acuiti;iop;iol;macular;cataract;intraocular;lasik;ocular;surgeri;cornea
7	roman;athenian;plato;socrat;greek;ovid;cicero;homer;poem;aristotl;poet;horac;herodotu;catullu;euripid
8	periodont;dental;dentin;teeth;cari;patient;implant;mandibular;enamel;gingiv;orthodont;tooth
9	tumor;cancer;carcinoma;cell;prostat;breast;tumour;gene;chemotherapi;p53;malign;apoptosi
10	student;teacher;school;educ;teach;classroom;curriculum;learn;learner;faculti;instruct;skill
11	patient;disord;depress;psychiatr;suicid;antipsychot;symptom;sleep;bipolar;mental;antidepress
12	rat;receptor;neuron;cell;mice;protein;mrna;ca2;brain;gene;insulin;kinas;muscl;inhibit;patient
13	philosophi;philosoph;descart;slovak;hegel;ethic;moral;masaryk;kant;husserl;cogito;frege
14	court;suprem;law;legal;feder;doctrin;litig;wto;crimin;judici;justic;lawyer;statut;claus;amend
15	galaxi;star;stellar;luminos;redshift;galact;ngc;solar;telescop;dwarf;supernova;accret;quasar
16	coronari;arteri;diabet;renal;transplant;clinic;ventricular;diseas;hypertens;cardiac;therapi
17	infect;hiv;viru;vaccin;patient;protein;viral;cell;gene;hcv;antibodi;mice;strain;pcr;malaria
18	laparoscop;liver;tumor;hepat;resect;pancreat;gastric;surgeri;cancer;pylori;endoscop;postop;surgic;ct
19	nurs;patient;care;caregiv;health;student;hospit;educ;clinic;women;staff;midwiv;midwiferi
20	price;firm;tax;wage;market;pavement;polici;trade;economi;monetari;capit;earn;invest;forecast;traffic
21	crack;turbul;finit;heat;flame;shear;vibrat;concret;beam;reynold;veloc;acoust;elast;vortex;temperatur
22	cloud;aerosol;wind;troposphere;ocean;atmospher;stratospher;radar;convect;ozon;rainfal;sst
23	music;opera;beethoven;schubert;josquin;symphoni;bach;tonal;song;motet;handel;brahm;sonata
24	children;adolesc;emot;psycholog;social;child;anxieti;student;women;cognit;school;sexual;violenc
25	flap;hear;nasal;surgeri;cochlear;postop;ear;surgic;nerv;implant;cleft;laryng;endoscop;neck;sinu
26	algorithm;fuzzi;wireless;antenna;queri;semant;robot;qo;packet;graph;xml;user;bandwidth
27	algebra;theorem;finit;graph;asymptot;infin;equat;polynomi;manifold;let;nonlinear;banach
28	seismic;rock;fault;basin;earthquak;magma;sediment;tecton;mantl;crustal;volcan;subduct;magmat
29	cow;dog;milk;calv;hors;diet;broiler;cattl;herd;pig;dairi;breed;carcass;heifer;lamb
30	fish;lake;sea;phytoplankton;sediment;speci;fisheri;habitat;river;ocean;spawn;estuari;benthic;larva
31	librari;citat;journal;web;metadata;catalog;bibliometr;literaci;academ;book;user;librarianship
32	philosoph;moral;epistem;truth;philosophi;metaphys;kant;argument;semant;argu;epistemolog
33	film;alloi;temperatur;quantum;dope;magnet;crystal;optic;si;beam;atom;laser;spin;anneal;ion
34	speci;forest;habitat;predat;plant;soil;seedl;prei;bird;tree;larva;egg;genu;femal;forag
35	protein;gene;cell;mutant;bind;dna;transcript;kinas;receptor;mutat;enzym;genom;phosphoryl
36	psychoanalysi;freud;analyst;unconsci;countertransfer;psychoanalyst;psychic;dream;analysand
37	phonolog;semant;speech;lexic;fmri;word;task;verb;sentenc;languag;children;cognit;memori
38	polit;social;polici;war;democraci;democrat;parti;women;discours;religi;reform;crime;sociolog
39	epilepsi;pain;seizur;stroke;aneurysm;clinic;cerebr;migrain;headach;brain;spinal;lesion;arteri
40	archaeologist;excav;neolith;prehistor;potteri;settlement;maya;ritual;palaeolith;burial;bronz
41	chees;acid;starch;meat;milk;flour;ferment;wine;antioxid;protein;juic;cook;food;monocytogen
42	cell;il;patient;mice;ifn;cytokin;cd4;immun;receptor;cd8;antigen;gene;hla;antibodi;protein
43	knee;bone;fractur;hip;arthroplasti;tendon;femor;ligament;injuri;pain;muscl;bmd;athlet;flexion
44	soil;plant;cultivar;crop;leaf;water;wheat;sludg;shoot;seedl;irrig;seed;biomass;sediment;ha
45	gospel;christian;jesu;god;biblic;hebrew;psalm;testament;theologi;luke;paul;church;bibl;divin

Chapter 6

Multi-view text mining for gene retrieval

6.1 Introduction

6.1.1 The importance of text mining in biomedical world

Text mining helps biologist to automatically collect structured biomedical knowledge from large volumes of biological literature. During the past ten years, there was a surge of interest in automatic exploration of the biomedical literature, ranging from the modest approach of annotating and extracting keywords from text [79] to more ambitious attempts such as Natural Language Processing (NLP) [13], and text mining based network construction and inference [86]. One of the main objectives of text mining is to structure the knowledge contained in the biological literature in order to extract biological entities and relations between them. In particular, these efforts effectively help biologists to identify the most likely disease candidate genes for further experimental validation [146]. It is often the case that text mining data is combined with other biological data within an elaborated workflow. For instance, text mining can serve as prior information for typical clinical decision support algorithms such as Bayesian networks [4]. It is also possible to unify heterogeneous data sources such as clinical data with text mining based data sources [49].

6.1.2 Multi-view text mining

In general, a successful text mining approach relies much on an appropriate mining model, and the efficiency of biomedical knowledge discovery varies greatly between different models. Which text mining model is superior depends on the problem under consideration. This makes multi-view models more suited since they are more flexible to answer various biological applications. In our early work, we propose a multi-view text mining model based on the use of several controlled vocabularies [146]. We now propose to also consider the use of several term scoring (weighting) schemes, and the mining of distinct document corpus as additional views. More precisely, we define distinct document corpus by distributing the journals based on their biomedical subjects, or by grouping the papers based on their publication year. The different views are redundant but also complementary. Therefore the integration of multiple views is expected to allow for a more accurate definition of our current knowledge in genetics and medicine. Another motivation behind our work is to provide a vertical search engine, in order to get insight into specific biomedical fields. In contrast to general search engines that attempt to index large portions of the World Wide Web or whole databases, vertical search engines typically attempt to index only the documents that are relevant to a pre-defined topic [10]. In our case, this segment can be defined by selecting one or several biomedical subjects, vocabularies, or time periods.

6.1.3 Related work

The concept of multi-view document analysis was originally proposed by Bickel and Scheffer who describe a web document clustering strategy that combines intrinsic view of web pages (text based similarity) and extrinsic view (citation link based similarity) [12]. More recently, Gaulton *et al.* have adopted three different ontologies on eight text sources and built the CAESAR system that annotates human disease genes and identifies potentially novel disease genes [46]. Lately, Névéol *et al.* have combined three different models (dictionary lookup, post-processing rules and NLP rules) to identify Medical Subject Headings (MeSH) main heading/subheading pairs from medical text [104]. Much effort has been put into the automatic extraction of disease gene relations from free text [28] [103]. To improve the performance of mapping biomedical sentences into an ontology, Kim *et al.* proposed an integrated information retrieval technique that combines a simple language model with document frequencies and a distance measure, and followed by clustering [75]. In 2005, we implemented a framework called TXTGate that combines literature indices of selected public biological resources in a flexible text-mining system designed

towards the analysis of gene sets [49]. More recently, we have used multi-view text mining data for gene prioritization and clustering [146]. Our work shares the same flavor, however we extend multi-view to broad concepts and emphasize vertical search from certain specific perspective. When compared to our former multi-view text mining work, our research brings three novel items: extension of the multi-view concept to a broad and flexible framework, implementation of vertical text mining from multiple perspectives, and a tensor based data fusion method. In the current study, we extend the multi-view concept to the use of several weighting schemes in addition to the use of several vocabularies. We also propose a vertical search engine by restricting the text mining analysis to a subset of the original document corpus. The subset can be defined either by biomedical subjects or by publication time periods, and only the relevant papers are then indexed. We have implemented this scheme into a freely available computational framework that can be used to investigate genes or gene sets through similarity analysis and clustering.

6.2 Materials and methods

6.2.1 Document corpus

One of the most important resources for biological text mining applications is MEDLINE database. MEDLINE contains more than 18 millions publications that cover many aspects of biology, chemistry and medicine. There is almost no limit to the types of information that may be recovered through careful and exhaustive mining. There are more than 10,000 biomedical related journals, accumulating over 700,000 new publications each year. In the current study, we use the MEDLINE repository as of April 2010. Each publication is represented by its title and its abstract (when available). The full article is never retrieved. The mapping between genes and publications from Entrez GeneRIF was used to index the MEDLINE repository. The GeneRIF data was also collected in April 2010, and consists of 290,000 associations between 13,633 human genes and 322,639 MEDLINE publications (from 3,276 journals).

Among the 3,276 journals that are relevant to human genes, the top 30 journals with the number of papers are listed in Table 6.1.

6.2.2 Indexing

In the first step, documents are indexed and a document-by-term matrix is computed. The indexing process is performed using the Java Lucene package

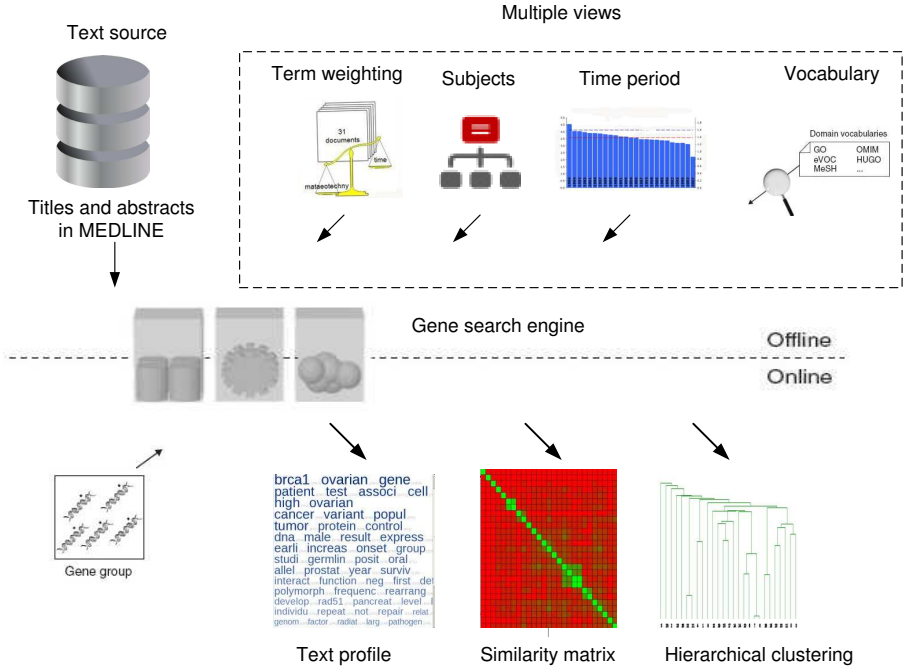

Figure 6.1: Conceptual overview of our text mining system. The whole corpus is indexed with several vocabularies, weighting schemes, biomedical subjects and publication time periods (multiple views). Sets of genes can then be investigated on-line: the text profiles of the genes are retrieved. Furthermore, similarity matrices can be computed and hierarchical clustering is performed.

[58], and more details can be found in our earlier work [146]. In the second step, we averagely combine the document-by-term vectors to obtain gene-by-term vectors according to the GeneRIF mapping. Each feature of the gene vector then corresponds to the score of a term from a fixed vocabulary (ontology). The multiple views adopted in this research refer to different weighting schemes, controlled vocabularies, and biomedical subjects.

Weighting schemes

A weight is a statistical measure used to evaluate how important a term is to a document in a corpus [119]. The importance increases proportionally to the number of times this term appears in the document but is offset by its frequency

Table 6.1: The top 30 journals related to human gene within the MEDLINE database by 2009

The No.	Paper's Number	Journal Name
1	23121	The Journal of biological chemistry
2	6735	Proceedings of the National Academy of Sciences
3	6174	Biochemical and biophysical research communications
4	3762	Molecular and cellular biology
5	3741	Genomics
6	3723	Blood
7	3624	Oncogene
8	3491	Journal of immunology
9	2848	Cancer research
10	2774	FEBS letters
11	2726	Biochemistry
12	2596	Human molecular genetics
13	2375	American journal of human genetics
14	2373	The EMBO journal
15	2261	Nature
16	2153	The Biochemical journal
17	2004	Biochimica et biophysica acta
18	1998	The Journal of clinical endocrinology and metabolism
19	1833	Nature genetics
20	1742	Cell
21	1719	Science
22	1689	Journal of virology
23	1648	Human mutation
24	1633	Human genetics
25	1602	Clinical cancer research
26	1583	Nucleic acids research
27	1552	The Journal of cell biology
28	1551	Gene
29	1493	Journal of medical genetics
30	1484	International journal of cancer

within the whole corpus. In the current study, we used three different weighting schemes: TF, IDF, and TFIDF. TFIDF is often used in information retrieval and text mining, but IDF is also found to work well in biomedical related text mining [146]. Since it is hard to estimate beforehand which scheme is universally superior, both are made available. In addition, TF is also proposed but mainly for comparative studies since it was shown to give less meaningful results [90].

Controlled vocabularies

We have selected four vocabularies from four bio-ontologies as follows.

The Gene Ontology (GO) GO [14] provides consistent descriptions of genes and gene-product attributes in the form of three structured controlled vocabularies that each provide a specific angle of view (biological processes,

cellular components and molecular functions). GO is built and maintained with the explicit goal of applications in text mining and semantic matching in mind [9]. Hence, it is an ideal source as domain-specific views in our approach. We extract all the terms in GO (due to the version released in December, 2008) as the controlled vocabulary of GO.

Medical Subject Headings (MeSH) MeSH is a controlled vocabulary produced by National Library of Medicine (NLM) for indexing, cataloging, and searching biomedical and health-related information and documents. The descriptors or subject headings of MeSH are arranged in a hierarchy. MeSH covers a broad range of topics and its current version consists of 16 top level categories. Though most of the articles in MEDLINE are already manually annotated with MeSH terms, our text mining process does not rely on these annotations but indexes the MEDLINE repository automatically with the MeSH descriptors (version 2008).

Online Mendelian Inheritance in Man's Morbid Map (OMIM) OMIM [15] is a database that catalogues all the known diseases with genetic components. It contains available links between diseases and relevant genes in the human genome and provides references for further research and tools for genomic analysis of a catalogued gene. OMIM is composed of two mappings: the OMIM Gene Map, which presents the cytogenetic locations of genes that are described in OMIM; the OMIM Morbid Map, which is an alphabetical list of diseases described in OMIM and their corresponding cytogenetic locations. Our approach retrieves the disease descriptions from the OMIM Morbid Map (version due to December, 2008) as the CV.

National Cancer Institute Dictionary (NCI) The NCI Thesaurus is a public domain description logic-based terminology produced by the National Cancer Institute, distributed as a component of the NCI Center for Bioinformatics caCORE distribution [53]. It is deep and complex compared to most broad clinical vocabularies, implementing rich semantic inter-relationships between the nodes of its taxonomies. The semantic relationships in the Thesaurus are intended to facilitate translational research and to support the bioinformatics infrastructure of the Institute. Topics described in the ontology include diseases, drugs, chemicals, diagnoses, genes, treatments, anatomy, organisms, and proteins. The NCI Thesaurus evolved from the NCI Metathesaurus, which is based on the National Library of Medicine Unified Medical Language System (UMLS) Metathesaurus. The NCI Metathesaurus has been operational since 1999. A public version is available at http://ncimeta.nci.nih.gov.

Three of them (GO, MeSH, OMIM) have proved their merit in our earlier work [147]. In addition, we have also selected an ontology from the National Cancer Institute (NCI) to cover more specifically cancerous diseases. The ontological

Table 6.2: The 114 biomedical subjects associated with human gene based on the PubMed journal categories

Acquired Immunod. Syndro	Communicable Diseases	Histocytochemistry	Occupational Med
Aerospace Medicine	Complementary Therapies	Histology	Ophthalmology
Allergy and Immunology	Critical Care	History of Medicine	Optometry
Anatomy	Dentistry	Hospitals	Orthodontics
Anesthesiology	Dermatology	Internal Medicine	Orthopedics
Anthropology	Diagnostic Imaging	Jurisprudence	Otolaryngology
Anti-Bacterial Agents	Drug Therapy	Laboratory Tech. and Pro.	Parasitology
Antineoplastic Agents	Education	Medical Informatics	Pathology
Audiology	Embryology	Medicine	Pediatrics
Bacteriology	Emergency Medicine	Mental Disorders	Perinatology
Behavioral Sci.	Endocrinology	Metabolism	Pharmacology
Biochemistry	Environmental Health	Microbiology	Pharmacy
Biology	Epidemiology	Military Medicine	Physical Medicine
Biomedical Engineering	Ethics	Molecular Biology	Physiology
Biophysics	Gastroenterology	Nanotechnology	Psychiatry
Biotechnology	General Surgery	Neoplasms	Psychology
Botany	Genetics	Nephrology	Psychopharmaco
Brain	Genetics, Medical	Neurology	Psychophysiology
Cardiology	Geriatrics	Neurosurgery	Public Health
Cell Biology	Gynecology	Nuclear Medicine	Pulmonary Med
Chemistry	Health Services	Nursing	Radiology
Chemistry Tech. Anal.	Health Services Research	Nutritional Sciences	Radiotherapy
Chemistry, Clinical	Hematology	Obstetrics	Rehabilitation
Reproductive Medicine	Rheumatology	Science	
Sexually Transmitted Dise	Sexually Transmitted Dise	Social Med	
Social Sciences	Speech-Language Pathology	Sports Medicine	
Statistics as Topic	Substance-Related Disorders	Technology	
Teratology	Therapeutics	Toxicology	
Transplantation	Traumatology	Tropical Medicine	
Urology	Vascular Diseases	Veterinary Medicine	
Virology	Women's Health		

terms are first extracted, stored as bag-of-words, and then preprocessed for text mining. This pre-processing includes transformation to lower case, segmentation of long phrases, and stemming. After preprocessing, these vocabularies are fed into a Java program based on the Apache Java Lucene API to index the titles and abstracts of MEDLINE publications relevant to human genes.

Biomedical subjects

The National Library of Medicine (NLM) assigns MeSH terms to each journals to describe their main focus. Not all journals are associated to MeSH terms, and we therefore only keep the journals with at least one term (52 journals discarded over 3,276 journals in total). There are in total, 114 distinct MeSH terms used to define the journal's scope, and there are sometimes several terms per journal. The distribution of the 114 MeSH terms is heavily biased. For instance, the term with the largest number of publications is 'Molecular biology', with 33,164 publications. At the other end of the spectrum, 'Optometry' is only linked to a single paper. More details about these MeSH terms can be found in Table 6.2.

Table 6.3: The various publishing time periods with the number of papers and the number of occurring genes

The publication period	Number of papers	Number of occurring genes
1950-1990	10,026	1
1991-2000	52,508	42
2001-2005	91,973	241
2006-2010	79,880	235

Publication year

For the current MEDLINE, the publication year ranges from 1950 to 2010. Notice that 5,537 papers have been removed since their publication year is missing. The yearly paper distribution (human gene related) is shown in Figure 1.8 of Chapter 1. As expected, the number of papers that are linked to human gene is increasing since the sequencing of the human genome. We have roughly divided the papers into four categories according to the publication year: 1950-1990, 1991-2000, 2001-2005, and 2006-2010 (see also Table 6.3).

6.2.3 Web application

The Web application was developed using the Google Web Toolkit Version 2.0[1]. A conceptual overview of our system is illustrated in Figure 6.1. It can be fed with a set of genes and returns the text profiles of these genes as well as the similarity matrix and the associated clustering results. These results can be downloaded for further analysis. For each query gene of the input gene set, a text profile is retrieved. This profile contains the annotation terms and the corresponding scores. It is possible to display the top 10 terms that are annotated to the genes (terms with the highest scores). It is also possible to visualize the profile as a tag cloud (or term cloud), for which the font size of a term is proportional to its score. To compare gene profiles, we compute the cosine similarity between the two corresponding gene-by-term vectors. We offer the possibility to cluster on-line the gene set by means of hierarchical clustering. The clustering is performed in Java (own implementation) using the average link [68] and the aforementioned similarity measure. The hierarchical structure can also be visualized to allow for an exploratory clustering strategy. For convenience, the clustering can only be achieved with 100 genes or less.

[1] http://code.google.com/webtoolkit/

6.2.4 Hybrid clustering approach

Clustering is helpful to identify the functional relationship between genes [60]. In this study, we apply a clustering strategy in order to assess whether combining multi-views leads to an increased performance. Hybrid clustering refers to joint clustering that integrates multi-view data, and is expected to boost the clustering performance. The hybrid clustering strategy we adopted is tensor based MC-OI-MLSVD method as introduced in Chapter 2.

6.2.5 Biomedical validation data

We validate our approach with the human disease benchmark data set of Endeavour [145], from which we selected 14 diseases and the 264 associated genes. The 14 diseases are presented in Table 6.4. To compare different views, the cosine similarities between all gene-by-term vectors are computed for each views, which leads to the generation of one similarity matrix per view. The cosine similarity is then computed between these two matrices and used as an estimate of the global similarity of the two underlying views [129]. Given two similarity matrices $\mathbf{S}_i$ and $\mathbf{S}_j$, and their corresponding vectorizations are $\text{vec}(\mathbf{S}_i)$ and $\text{vec}(\mathbf{S}_j)$, where $\text{vec}(\mathbf{S})$ means all columns of $\mathbf{S}$ are stacked each other, the cosine similarity cross two views is computed as,

$$\cos(\theta_{i,j}) = \text{vec}(\mathbf{S}_i) \times \text{vec}(\mathbf{S}_j)/(\|\text{vec}(\mathbf{S}_i)\|_2 \times \|\text{vec}(\mathbf{S}_j)\|_2) \tag{6.1}$$

where $\cos(\theta_{i,j}) = \cos(\theta_{j,i})$ and it ranges from 0 (two views are completely different) and 1 (two views are identical).

Regarding clustering evaluation, the gene data sets used in our experiments are provided with disease labels, therefore the clustering performance is evaluated by comparing the automatic partitions with the labels using ARI [63] and NMI [127]. We set the cluster number K to 14 since there are 14 diseases.

6.3 Results

This section presents a similarity analysis performed on the individual views, a benchmark of the method based on clustering, and introduces our web tool.

Table 6.4: Genetic diseases in disease data and the number of genes relevant to each disease. The numbers in parentheses are the removed overlapping genes in each disease

Number	Disease	Number of genes
1	breast cancer	24(5)
2	cardoomuopathy	22(5)
3	cataract	20(1)
4	charcot marie tooth disease	14(4)
5	colorectal cancer	21(6)
6	diabetes	26(3)
7	emolytic anemia	13(1)
8	epilepsy	15(1)
9	lymphoma	31(4)
10	mental retardation	24(4)
11	muscular dystophy	24(6)
12	neuropathy	18(3)
13	obesity	13(1)
14	retinitis pigmentosa	30(2)

6.3.1 The similarities among multiple views

Each view provides text information from a certain perspective. In this section, before combining multiple views, we measure the similarities or the differences among these views.

Similarities among vocabularies

To compare the different vocabularies, we set the weighting scheme to IDF, and the whole corpus was indexed. The global similarities among the four vocabularies are shown in Table 6.5. The largest similarity exists between GO and NCI (0.8966) while the smallest similarity between MeSH and OMIM (0.7565). Altogether, the results indicate that although there are differences among the vocabularies, these are not huge. Similar results are obtained with TFIDF (data not shown).

Table 6.5: The cosine similarity between the four vocabularies. The largest non-self similarity is shown in bold; and the smallest non-self similarity is shown in italics.

Vocabulary	GO	MeSH	OMIM	NCI
GO	1	0.7925	0.8499	0.8966
MeSH	0.7925	1	0.7565	0.8111
OMIM	0.8499	*0.7565*	1	0.8192
NCI	**0.8966**	0.8111	0.8192	1

Table 6.6: The cosine similarity between the three weighting schemes. The largest non-self similarity is shown in bold; and the smallest non-self similarity is shown in italics.

Weighting scheme	TF	IDF	TFIDF
TF	1	0.7326	0.8715
IDF	*0.7326*	1	0.8524
TFIDF	**0.8715**	0.8524	1

Similarities among weighting schemes

To analyze multiple weighting schemes, we used MeSH as the vocabulary, and the whole corpus was indexed. The global similarities among the three weighting schemes are presented in Table 6.6. The largest similarity exists between TF and TFIDF (0.8715); and the smallest similarity between TF and IDF (0.7326). Similar results are obtained with different vocabularies (data not shown).

Similarities between biomedical subjects

In order to compare the different biomedical subjects, we used MeSH as the vocabulary; we set the weighting scheme to IDF; and we selected all publication time periods. Among the 114 biomedical subjects, we selected the six that are associated with the largest number of papers. The global similarities among these six subjects are shown in Table 6.7. The largest similarity exists between Molecular Biology and Cell biology (0.7919); and the smallest similarity between Allergy & Immunology and Genetic medical (0.2212). As can be observed from the Table 6.7, the use of different subjects gives more different results than the use of different vocabularies or weighting schemes.

Table 6.7: The cosine similarity of multi-view subjects. Except the self-similarity, the largest similarity is shown in bold while the smallest similarity is shown in italics.

Medical subject	Aller. & Immuno.	Cell bio	Genetic med	Molecular bio	Neroplasms	Sci
Aller. & Immuno.	1	0.3527	0.2212	0.3032	0.3958	0.3657
Cell biology	0.3527	1	0.5080	0.7919	0.7471	0.5574
Genetic medical	*0.2212*	0.5080	1	0.7766	0.4634	0.4967
Molecular Biology	0.3032	**0.7919**	0.7766	1	0.6900	0.5843
Neroplasms	0.3958	0.7471	0.4634	0.6900	1	0.5613
Science	0.3657	0.5574	0.4967	0.5843	0.5613	1

Table 6.8: The 26 genes associated with diabetes.

GYS1	NEUROD1	AQP2	HNF4A	IRS2	AVP	PDX1
CTLA4	IRS1	INSR	PLAGL1	PPARG	MAPK8IP1	TCF1
FOXP3	SPINK1	SLC2A2	INS	CAPN10	IAPP	
GPD2	SLC2A4	AVPR2	GCK	TCF2	RRAD	

It also motivates vertical searches that are able to provide precise and unique information.

Influence of the biological question

We have also performed the analysis on a set of diabetes related genes to investigate the differences among the multiples views. The 26 genes associated with diabetes are selected for this test as presented in Table 6.8. In this experiment, we provide a text profile for this group of genes by averaging the text profiles of each gene. The detail of multi-view text mining profiles can be observed in Table 6.9, Table 6.10, Table 6.11 and Table 6.12. We can see that, for diabetes, there is less overlap among the multiple views in general, and in particular for vocabularies and biomedical subjects. This analysis indicates that different results can be obtained with different biological questions.

6.3.2 Multi-views clustering by MC-OI-MLSVD

In this section, we compare the clustering results when applied on a single view and when applied on multiple views.

Table 6.9: The text profile of diabetes related genes by multi-view vocabularies (IDF weighting scheme, all publication time periods and all biomedical subjects).

GO	MeSH	OMIM	NCI
mbf	cod liver oil	menier	bangladesh
hexos transport	parot	antidiuresi	atcc
glucokinas activ	etodolac	tropic calcif pancreat	etanercept
glycolipid bind	menier	pck1	smad4 protein
densa	oletf	yemenit	20q12
cyclopentenon	f18	leprechaun	ibuprofen
garp	bottl	pdx1	neural network
glycerophosph	pouchiti	leiomyomata	conjunctiva
glucos bind	insulin antibodi	por	croatia
basilar	erythrocyt deform	hyperinsulinem hypoglycemia	sulindac

Table 6.10: The text profile of diabetes related genes by multi-view weighting schemes (MeSH vocabulary, all publication time periods and all biomedical subjects).

TF	IDF	TDIDF
rho	cod liver oil	parot
curv	parot	pox
activ	etodolac	adrenomedullin
adrenomedullin	menier	cyp27a1
invas	oletf	vitiligo
abl	f18	liposarcoma
insulin	bottl	altitud
vitiligo	pouchiti	lactas
multipl sclerosi	insulin antibodi	thymoma
obes	erythrocyt deform	ophthalmopathi

Table 6.11: The text profile of diabetes related genes by multi-view publication time periods (MeSH vocabulary, IDF weighting scheme and all biomedical subjects).

1980-2000	2001-2005	2006-2010
oxytocin	hyperamylasemia	cod liver oil
tempera	parot	menier
hypothyroid	etodolac	bottl
thyrotropin	menier	pouchiti
vasopressin	oletf	insulin antibodi
ppar gamma	f18	erythrocyt deform
dimens	insulin antibodi	basilar arteri
overweight	rapa	macrosomia
charact	salivari duct	multipl trauma
dosag	cholestyramin	glycyrrhetin acid

Table 6.12: The text profile of diabetes related genes by multi-view biomedical subjects (MeSH vocabulary, IDF weighting scheme and all publication time periods).

Allergy & Immun	Cell bio	Genetic med	Molecular bio	Neroplasms	Sci
filari	smad4 protein	bodi size	pox	etodolac	trachoma
hypoparathyroid	osteonecrosi	indel	amber	f18	glycogen syn
flagellin	raptor	minisatellit	smad4 protein	ibuprofen	proinsulin
uveiti	gata4	aquaporin 2	sulfid	sulfid	relaxin
cd1d	vitiligo	bcg	sulindac	sulindac	tryptas
anergi	proinsulin	dietari fat	caveolin 2	caveolin 2	vasopressin
granuloma	asc	panic disord	hydatidiform mole	biliari tract	longev
sea	anabol	arginin vasopressin	theca	bodi size	arrestin
heart transplant	ht29	tempera	beta caroten	coup	ppar gamma
th1 cell	tuber sclerosi	lipodystrophi	osteonecrosi	bcg	rna splice

Clustering using multiple vocabularies

Using the four different vocabularies, we build four different gene-by-term matrices. Therefore, four normalized similarity matrices are generated in total. We then applied our multi-view clustering method (MC-MI-MLSVD) to combine the multiple views and compare to the use of any single view. The clustering results with IDF weighting scheme are presented in Table 6.13. It can be observed that the best single-view performance is obtained by using MeSH vocabulary (NMI 0.7012, ARI 0.5157). However, the integration of multiple views (NMI 0.7290, ARI 0.5393), MeSH and OMIM vocabulary in this case, is significantly superior to the use of MeSH or OMIM vocabulary alone. These results demonstrate that the integration of multiple vocabularies is able to

Table 6.13: Clustering results for multiple vocabularies with IDF weighting scheme. The mean values and standard deviations are observed from 50 repetitions. The best values are shown in bold. "Combined" refers to the integration of MeSH and OMIM.

Vocabularies	NMI	P-value	ARI	P-value
Combined	**0.7290± 0.02**	—	**0.5393±0.05**	—
GO	0.5537± 0.01	4.4e-39	0.3575±0.03	1.34e-23
MeSH	0.7012±0.02	1.14e-7	0.5157±0.05	0.0316
OMIM	0.6893±0.02	7.01e-12	0.4868±0.05	8.42e-6
NCI	0.5109±0.01	2.97e-46	0.2844±0.02	3.10e-34

Table 6.14: Clustering results for multiple weighting schemes with MeSH vocabulary. The mean values and standard deviations are observed from 50 repetitions. The best values are shown in bold. "Combined" refers to the integration of TFIDF and IDF data.

Weighting scheme	NMI	P-value	ARI	P-value
Combined	**0.7001±0.02**	—	**0.5236±0.05**	—
TFIDF	0.6868±0.02	0.0017	0.5021±0.04	0.0466
TF	0.4963±0.02	4.52e-49	0.2882±0.02	1.14e-35
IDF	0.6872±0.01	2.85e-4	0.5039±0.04	0.0232

enhance the clustering performance. Similar results are obtained with TFIDF weighting scheme (data not shown).

Clustering using multiple weighting schemes

In this study, we have implemented three weighting schemes: TF, IDF and TFIDF, which allows us to get three gene-by-term matrices. We expect that integrating this type of multiple views will enhance the clustering performance. The clustering results with MeSH vocabulary are presented in Table 6.14. The best performance for single view is obtained by TFIDF (NMI 0.7001, ARI 0.5236), just slightly ahead of IDF (NMI 0.6872, ARI 0.5039). However, the hybrid clustering using multiple views is still significantly superior, showing that the integration of multiple weighting schemes, TFIDF and IDF in our case, can boost the clustering performance as well. Similar results are obtained with other vocabularies (data not shown).

Table 6.15: The number of overlaping terms between the four vocabularies. The total number of terms for each vocabulary is denoted between brackets.

	GO (37,069)	MesH (29,709)	OMIM (5,021)	NCI (27,247)
GO (37,069)	—	9,952	1,431	5,409
MesH (29,709)	9,952	—	3,191	6,399
OMIM (5, 021)	1,431	3,191	–	1,071
NCI (27,347)	5,409	6,399	1,071	–

6.4 Discussion

We have developed a literature based gene retrieval system that is able to provide multi-view observations as well as vertical search. The aim of our system is to aid the clinical analysis and biomedical research. We illustrate its usefulness through a clustering validation that proved the efficiency of the multi-view strategy. With respect to vertical search, our search engine is able to help the users who want very specific knowledge (corresponding to one of the several branches of the whole biomedical world). Biomedical research is a fast developing field, it is therefore divided into more and more tiny specialized fields. Hence, such a system that proposes precise searches is really more and more required.

Based on the similarity analysis of multi-view text mining, as can be seen in Table 6.5 (multi-view vocabularies), Table 6.6 (multi-view weighting schemes) and Table 6.7 (multi-view biomedical subjects), the views appear different but redundant. This redundancy was expected because the multiple vocabularies we used share common terms as denoted in Table 6.15. The largest overlap is observed between MesH and OMIM, with 3,191 common terms, which represents 64% of OMIM. In addition, the multiple weighting schemes (TF, IDF and TFIDF) we used also share part of their formulas. However, beside this redundancy, we can observe that the integration of multiple views is almost always leading to a better representation of the data.

Regarding clustering of multi-view data, the tensor based hybrid clustering method is able to make the best of the data. As long as the multi-view data has complementray information and that the noise level is kept under control, the combination with the tensor based strategy is always able to improve the clustering performance.

The idea of multi-view text mining is not restricted to the several views mentioned in this study. For example, according to the citation impact factor

of the related journals and their citations, the papers in MEDLINE could be classified into different categories, which thus would correspond to other types of views for text mining. One research avenue to explore in the future, we can use gene-by-concept vector by latent semantic analysis [60], which is expected to identify the implicit relationship among genes.

6.5 Summary

In this study, we have developed a Web based system that can be used to profile a gene set from a text-mining point of view. On the one hand, the information from multiple views can be combined to provide rich and complementary information. On the other hand, information from a specific view offers a vertical observation with a specific focus. The system can be utilized to identify the relationships between genes to aid the clinic diagnosis straightforwardly or to provide text prior information for further analysis. We have benchmarked the overall approaches with a set of disease genes. The results demonstrate the power of combining multiple views when performing clustering due to the synergic effect of the fusion. However, we also observed that better results can be obtained for a specific biological question when using a single highly relevant view. In the further research, we plan to apply our system to enhance candidate gene prioritization. Meanwhile, we are also planning an extension of the approach to other biomedical entities, such as diseases or biological pathways.

The web application of our multi-view text mining strategy is named Text Prior, which is available: http://aulne8.esat.kuleuven.be/TextPrior/.

Chapter 7

General conclusions and perspectives

7.1 Conclusions

The common sense that collecting evidence from multiple perspectives is able to facilitate discovering the latent patterns hidden in objects motivates our research to integrate multi-view data for joint learning. Two main topics associated with multi-view data are covered in this Thesis: clustering algorithm and text mining application.

7.1.1 Multi-view clustering algorithms

Clustering is a challenging task because the cluster structure inherent within the data is hard to define and consequently the cluster pattern is not easy to detect. Hence, statistic models sometimes can not reflect the nature of data in a proper way. On the other hand, evidence collected from multiple perspectives, as long as they complement each other enough, seems helpful to understand the nature of data, thus making the cluster structure more clear to observe and analyze. As a result, based on various theoretical analysis, we have proposed several multi-view clustering methods to facilitate the clustering tasks.

Multi-view partitioning via tensor methods

Tensor is a natural model for multi-view data from either vector spaces or graph spaces. Two basic strategies are presented: optimization integration (MC-OI) and matrix integration (MC-MI). A joint optimal subspace of multi-view data can be obtained by some tensor methods, for example, MLSVD and HOOI. Among our tensor based algorithms, MC-OI-MLSVD as well as MC-OI-HOOI, provide a joint matrix compression of multi-view data while MC-MI-HOOI offers a multilinear analysis of multi-view data. In particular, weights of multi-view data obtained by MC-MI-HOOI reflect the linear relationship of multiple views.

Simultaneous partitioning and joint dimension reduction of multiple graphs

Since tensor decomposition sometimes is still stuck by heavy computation, we keep on simplifying weighted multi-view clustering by simultaneous trace maximization. An algorithm named MC-STM is put forward, which analyzes the multilinear relationship of multi-view data just by trace operation and EVD while the multilinear relationship captured by MC-STM is almost the same as MC-MI-HOOI. In addition, a joint dimension reduction scheme by MLSVD is employed to reduce the abundant data which is very rich in multi-view data.

Mutual information based weighted hybrid clustering

According to the empirical observations that the ANMI value of each single-view data generally corresponds to their clustering performance, we developed a strategy to measure the contribution of each single-view by calculating the mutual information among their partitions. The weighting scheme is subsequently embedded into the multi-view clustering strategies of kernel fusion and clustering ensemble.

Network analysis in graph spaces

Multi-view data can also be modeled as a sparse multiplex network (same nodes with different links). Hence we carry out the multi-view clustering from a network analysis point of view. Taking into account the complementary properties of two heterogeneous data, that is, both the sparse link structure of citation data and the rich semantic meaning of text data, we present a modularity maximization based hybrid clustering scheme. Our hybrid clustering scheme is able to implement data fusion and hierarchical partitioning

simultaneously. This scheme is particularly devised to handle large-scale data in scientific publication analysis or Web mining.

As compared to traditional single-view clustering, the merits of multi-view clustering are obvious:

- Advanced clustering performance. If the multi-view data complements each other well, multi-view clustering is able to improve clustering performance, by discovering more complete cluster patterns.
- Robust clustering results. Multi-view clustering generally leads to robust partitioning. For instance, by collecting evidence from multiple perspectives, it is able to lower the partitioning risks which happen to single-view data, and reduce the side-effect by noise, outliers, different samplings and random initializations.
- Novel clustering patterns. Multi-view clustering enables us to discover the partterns which could hardly be discovered by any single-view data.

Thanks to the above advantages of multi-view clustering, it has a wide variety of potential applications:

- Scientific publication analysis: multi-view data refers to text data, description data (titles, authors and journals) and citation data;
- Web mining: multi-view data refers to text data, hyperlink data, image data and even click data;
- Community detection in social networks: multi-view data refers to exchange relationship by E-mail, organization relationship and collaboration relationship;
- Biomedical information processing: for example, genes can be represented in the expression vector space (corresponding to the genetic activity) and also in the term vector space (corresponding to the text information) [52];
- Multimedia information retrieval: For instance, in a video retrieval system, broadcast news videos can be represented as different models, such as text and image, which are independent but complement each other [139]. Thus multi-view clustering can be used to integrate the various information to facilitate the video retrieval.

7.1.2 Multi-view text mining applications

We applied multi-view text mining to both bibliometric and bioinformatics applications.

Scientific mapping

The complex nature of mapping various aspects of knowledge motivates the approaches that incorporate different viewpoints on the same data collection. By text mining and information extraction, we obtain multi-view text mining data as well as generate multi-view bibliometric data. Textual and bibliometric data provide different perceptions of similarity between documents or groups of documents.

We proposed various schemes to integrate textual and bibliometric methods, in particular, the mutual information based weighted hybrid clustering scheme in vector spaces as well as network analysis based graph coupling scheme in graph spaces. Our hypothesis was confirmed that such multi-view analysis leads to better comprehension of the cluster structure. Such hybrid methodologies with multi-view text mining are valuable tools to facilitate endeavors in mapping fields of science and technology and in research evaluation. The mapping of scientific fields is helpful to understand the structure and evolution of various research areas and of their relationships with other fields.

Text Prior for clinical diagnosis

We solidified our effort in multi-view text mining as Text Prior software. Text Prior provides a gene search engine in terms of data fusion by integrating multi-view text mining or vertical search by a specific biomedical perspective. Term annotation, gene relationship and clustering structure are offered. The retrieval results can be downloaded for direct analysis or for further research, such as integration with gene expression data. The software is freely accessible online and it will play a useful role for clinician and bioinformaticians in their research.

7.2 Future direction

7.2.1 Multi-view learning by tensor analysis

Dynamic tensor analysis. In this research, tensor analysis is limited to 3 dimensional arrays, and it would be easy to extend to 4 dimensions by adding the time dimension. Dynamic tensor analysis can detect evolving patterns in time series, for instance, the dynamic multi-view clustering.

Computation of large-scale tensor decomposition. Although tensor decomposition has appeared in many machine learning or data mining tasks, most work is still limited to the algorithm analysis as well as the applications on small-size databases. The heavy computation of tensor decomposition has become a bottleneck for further applications. Efficient implementation of tensor decomposition seems crucial to meet the practical requirement.

Meanwhile, real data structure (like matrices and vectors) in many data mining tasks is very sparse, which aspect we can utilize to speed up tensor decomposition. The basic idea is to transform tensor operations to operations amid sparse matrices and sparse vectors. Kolda *et al.* have carried out some similar work [78, 122].

Currently, scalable computation is also a promising topic in the field of tensor decomposition and application. Based on some successful scalable matrix decomposition applications, the Power method and Krylov method can be directly extended to implement tensor decomposition [54, 121]. Besides, the "tensor train" concept by [111] provides a powerful solution to scalable tensor decomposition.

Joint dimension reduction of multi-view data. Tensor decomposition is a powerful tool for dimension reduction and it has been applied to signal processing and computer vision [33, 137]. As introduced in Chapter 3, tensor decomposition, in particular MLSVD, is able to sharply reduce the dimensionalites of multi-view data while the inherent patterns are still preserved. Thus we will keep on working to unleash the dimension reduction potential of tensor decomposition in later work. For instance, we will apply this joint dimension reduction scheme to other multi-view learning tasks; and we will handle multi-view data by hierarchical Tucker compression [65].

7.2.2 Transfer learning on multi-view text mining

Transfer learning aims at transferring knowledge from source tasks to target tasks, where the training data from source domains and the test data from a target domain may follow different distributions or are represented by different features [114]. For example, the abilities acquired while learning to walk presumably apply when one learns to run, and knowledge gained while learning to recognize cars could apply when recognizing trucks. Researchers have applied techniques of transfer learning to problems in text classification, spam filtering, and urban combat simulation [114]. W.r.t. multi-view text mining data, we can apply transfer learning to unleash the power of multi-view text mining, transferring the knowledge or pattern learned from one view to aid the analysis of the other view. Moreover, transfer learning can be implemented between the gene patterns learned from text mining and expression data resulting from expensive experiments.

7.2.3 Incomplete data and multi-look clustering

In contrast to multi-view learning, multi-look learning also refers to learning from different representations of the same type of data. As opposed to learning from multiple views where it is assumed that the exact same instances have multiple representations, we only assume the availability of samples of the same learning task in different domains [45]. In fact, multi-look data is more common in real application.

One example is the task of medical diagnostic, in which case the outlooks are medical tests, such as blood samples and medical imagery. The different tests need not be from the same patient. Taking the multiple outlooks into account allows us to learn from the input of all tests without having all test results for each patient in all outlooks. Since not all tests are done on all patients, the outlooks perspective enables a better nonrestrictive use of data for the learning of the medical classification task [45].

The problem of incomplete data, i.e., data with missing or unknown values in multi-way arrays is ubiquitous in biomedical signal processing, network traffic analysis, bibliometrics, social network analysis, chemometrics, computer vision, communication networks, etc. [1]. In multi-view formulation, incomplete data means the representations of some instances in one view are available but in other view unavailable. Incomplete data poses a challenge to multi-view learning. However, it can be formulated as a multi-look learning problem.

7.2.4 Detection of gene outliers by collecting multi-view evidence

Outlier detection refers to the problem of finding patterns in data that do not conform to expected normal behavior [26]. These anomalous patterns are often referred to as outliers, anomalies, discordant observations, exceptions, faults, defects, noise, errors, damages, novelty or contaminants in various application domains [26]. Outlier detection is applicable in a variety of domains, such as intrusion detection, fraud detection, fault detection, system health monitoring, event detection in sensor networks, and detecting eco-system disturbances [80].

In public health data, outlier detection techniques are widely used to detect anomalous patterns in patient medical records which could be symptoms of a new disease [26]. On the other hand, some irrelevant gene outliers need to be detected when the clinicians want to analyze a set of genes associated with a certain disease.

Although some successful outliers detection methods and applications have appeared, some challenges still remain [26], for instance, the boundary between normal and outlying instances is often fuzzy. Meanwhile, more evidence collected from multiple perspectives is helpful to detect noise and the actual outliers. Consequently, instead of using only one kind of information which might contain the incomplete information, we will carry out outliers detection with multi-view data.

Appendix A

List of algorithms

No.	Name	Chapter No. (Page No.)
1	MC-OI-MLSVD	2 (40)
2	MC-OI-HOOI	2 (41)
3	MC-MI-HOOI	2 (43)
4	MC-STM	3 (64)
5	MC-STM-MLSVD	3 (68)
6	Multi-view modularity maximization clustering	3 (69)
7	Multi-view k-means clustering	3 (70)
8	WKFCM	4 (96)
9	WSA	4 (96)
10	WEAC-AL	4 (96)
11	Hybrid clustering by graph integration	5 (120)
12	Hybrid clustering by graph coupling	5 (121)

Bibliography

[1] Acar, E., Dunlavy, D. M., Kolda, T. G., and Mørup, M. Scalable tensor factorizations for incomplete data. *Chemometrics and Intelligent Laboratory Systems*. pages 158

[2] Alpaydin, E. *Introduction to Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2004. pages 2

[3] Alter, O., and Golub, G. H. Reconstructing the pathways of a cellular system from genome-scale signals by using matrix and tensor computations. *Proceedings of the National Academy of Sciences USA (PNAS) 102*, 49 (2006), 17559–17564. pages 10, 28

[4] Antal, P., Fannes, G., Timmerman, D., Moreau, Y., and De Moor, B. Using literature and data to learn bayesian networks as clinical models of ovarian tumors. *Artificial Intelligence in Medicine 30*, 3 (2004), 257–281. pages 135

[5] Ayad, H. G., and Kamel, M. S. Cumulative voting consensus method for partitions with variable number of clusters. *IEEE Transactions on Pattern Analysis and Machine Intelligence 30*, 1 (2008), 160–173. pages 28, 32, 44, 98

[6] Bader, B. W., and Kolda, T. G. Matlab tensor toolbox version 2.4. http://csmr.ca.sandia.gov/ tgkolda/TensorToolbox/, March 2010. pages 44, 70

[7] Baeza-Yates, R. A., and Ribeiro-Neto, B. *Modern Information Retrieval*. 1999. pages 115

[8] Balasubramanian, K., Kim, J., Puretskiy, A. A., Berry, M. W., and Park, H. A fast algorithm for nonnegative tensor factorization using block coordinate descent and an active-set-type method. In *Text Mining*

Workshop, Proceedings of the Tenth SIAM International Conference on Data Mining (2010), SDM'10, pp. 25–34. pages 10, 28

[9] Batagelj, V., and Mrvar, A. Pajek - analysis and visualization of large networks. In *Graph Drawing Software* (2003), vol. 2265, pp. 77–103. pages 15, 104

[10] Battelle, J. *The Search: How Google and its Rivals Rewrote the Rules of Business and Transformed Our Culture*. Portfolio, New York, 2005. pages 136

[11] Berry, M. W., Dumais, S. T., and O'Brien, G. W. Using linear algebra for intelligent information retrieval. *SIAM Review 37* (1995), 573–595. pages 90

[12] Bickel, S., and Scheffer, T. Multi-view clustering. In *Proceedings of the Fourth IEEE International Conference on Data Mining* (Washington, DC, USA, 2004), ICDM '04, IEEE Computer Society, pp. 19–26. pages 8, 28, 31, 58, 59, 88, 136

[13] Björne, J., Ginter, F., Pyysalo, S., Tsujii, J., and Salakoski, T. Complex event extraction at PubMed scale. *Bioinformatics (Oxford, England) 26*, 12 (2010), 382–390. pages 135

[14] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment 2008*, 10 (2008), P10008+. pages xiii, 114, 117

[15] Boyack, K. W., Börner, K., and Klavans, R. Mapping the structure and evolution of chemistry research. *Scientometrics 79*, 1 (2009), 45–60. pages 87

[16] Braam, R. R., Moed, H. F., and van Raan, A. F. J. Mapping of science by combined co-citation and word analysis, part I: Structural aspects. *Journal of the American Society for Information Science 42*, 4 (1991), 233–251. pages 88, 113

[17] Braam, R. R., Moed, H. F., and van Raan, A. F. J. Mapping of science by combined co-citation and word analysis, part II: Dynamical aspects. *Journal of the American Society for Information Science 42*, 4 (1991), 252–266. pages 88, 113

[18] Brin, S., and Page, L. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems 30* (1998), 107–117. pages 87

[19] Bro, R. *Multi-way analysis in the food industry-models, algorithms and applications*. PhD thesis, Department of Analytical Chemistry, University of Amsterdam, Amsterdam, 1998. pages 10

[20] Burning, D. C., Cohn, D., and Hofmann, T. The missing link-a probabilistic model of document content and hypertext connectivity. In *In Proceedings of Neural Information Processing Systems* (Vancouver,British Columbia, 2001), vol. 13, pp. 430–436. pages 60

[21] Cai, X., Nie, F., Huang, H., and Kamangar, F. Heterogeneous image features integration via multi-modal spectral clustering. In *Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)* (2011), pp. 1977–1984. pages 60

[22] Calado, P., Cristo, M., Gonçalves, M. A., de Moura, E. S., Ribeiro-Neto, B., and Ziviani, N. Link-based similarity measures for the classification of web documents. *Journal of the American Society for Information Science and Technology 57* (2006), 208–221. pages 113

[23] Calado, P., Ribeiro-Neto, B., Ziviani, N., Moura, E., and Silva, I. Local versus global link information in the web. *ACM Transactions on Information Systems 21* (2003), 42–63. pages 113

[24] Cao, L., Luo, J., Liang, F., and Huang, T. S. Heterogeneous feature machines for visual recognition. In *Proceedings of IEEE 12th International Conference on Computer Vision* (2009), pp. 1095–1102. pages 60

[25] Carroll, J. D., and Chang, J. J. Analysis of individual differences in multidimensional scaling via an n-way generalization of 'echart-young', decomposition. *Psychometricka 35* (1970), 283–319. pages 44, 56

[26] Chandola, V., Banerjee, A., and Kumar, V. Anomaly detection: A survey. *ACM Computing Surveys 41* (2009), 15:1–15:58. pages 159

[27] Chaudhuri, K., Kakade, S. M., Livescu, K., and Sridharan, K. Multi-view clustering via canonical correlation analysis. In *Proceedings of the 26th Annual International Conference on Machine Learning* (New York, NY, USA, 2009), ICML'09, ACM, pp. 129–136. pages 8, 28, 31, 59

[28] Chun, H.-w., Tsuruoka, Y., Kim, J.-d., Shiba, R., Nagata, N., and Hishiki, T. Extraction of gene-disease relations from MEDLINE using domain dictionaries and machine learning. In *In Proceedings of the 11th Pacific Symposium on Biocomputing* (2006). pages 136

[29] Clauset, A., Newman, M. E. J., and Moore, C. Finding community structure in very large networks. *Physical Review E 70* (2004), 066111. pages 117

[30] Culp, M., Michailidis, G., and Johnson, K. On multi-view learning with additive models. *The Annals of Applied Statistics 3*, 1 (2009), 292–318. pages 7

[31] De Lathauwer, L., De Moor, B., and Vandewalle, J. A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications 21*, 4 (2000), 1253–1278. pages 28, 35, 37, 39, 40

[32] De Lathauwer, L., De Moor, B., and Vandewalle, J. On the best rank-1 and rank-$(r_1, r_2, \ldots, r_n)$ approximation of higher-order tensors. *SIAM Journal on Matrix Analysis and Applications 21*, 4 (2000), 1324–1342. pages 35, 40, 68

[33] De Lathauwer, L., and Vandewalle, J. Dimensionality reduction in higher-order signal processing and rank-$(r_1, r_2, \ldots, r_n)$ reduction in multilinear algebra. *Linear Algebra and its Applications 391* (2004), 31–55. pages 61, 157

[34] De Smet, W., Tang, J., and Moens, M.-F. M. Knowledge transfer across multilingual corpora via latent topics. In *Proceedings of the 15th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD2011)* (2011), pp. 549–560. pages 28, 60

[35] Deschacht, K. *Weakly supervised methods for information extraction.* Phd thesis, Faculty of Engineering, K.U.Leuven, Leuven, Belgium, 2010. pages 14

[36] Dörre, J., Gerstl, P., and Seiffert, R. Text mining: finding nuggets in mountains of textual data. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (1999), KDD '99, pp. 398–401. pages 12

[37] Dunlavy, D. M., Kolda, T. G., and Acar, E. Poblano v1.0: A matlab toolbox for gradient-based optimization. Tech. Rep. SAND2010-1422, Sandia National Laboratories, Albuquerque, NM and Livermore, CA, March 2010. pages 44

[38] Dunlavy, D. M., Kolda, T. G., and Kegelmeyer, W. P. Multilinear algebra for analyzing data with multiple linkages. Tech. Rep. SAND2006-2079, Sandia National Laboratories, 2006. pages 10, 28, 32

[39] Efron, B., and Tibshirani, R. *An introduction to the bootstrap*. Chapman & Hall/CRC, Boca Raton, FL, 1993. pages 99

[40] Erosheva, E., Fienberg, S., and Lafferty, J. Mixed membership models of scientific publications. In *Proceedings of the National Academy of Sciences* (2004), vol. 101, pp. 5220–5227. pages 60

[41] Feldman, R., and Sanger, J. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge university press, Brook Hill Drive West Nyack, NY, 2007. pages 12

[42] Fortunato, S. Community detection in graphs. *Physics Reports 486* (2010), 75–174. pages 114, 117

[43] Fortunato, S., and Barthelemy, M. Resolution limit in community detection. *Proceedings of the National Academy of Sciences 104* (2007), 36. pages 118

[44] Fred, A. L. N., and Jain, A. K. Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 27* (2005), 835–850. pages 96, 98

[45] Gal-on, M., and Mannor, S. Learning from multiple outlooks. *CoRR abs/1005.0027* (2010). pages 158

[46] Gaulton, K. J., Mohlke, K. L., and Vision, T. J. A computational system to select candidate genes for complex human traits. *Bioinformatics 23* (2007), 1132–1140. pages 136

[47] Glänzel, W., and Schubert, A. A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics 56*, 3 (2003), 357–367. pages 108

[48] Glenisson, P. *Integrating scientific literature with large scale gene expression analysis*. PhD thesis, Faculty of Engineering, K.U.Leuven, Leuven, Belgium, June 2004. pages 14

[49] Glenisson, P., Coessens, B., Van Vooren, S., Mathys, J., Moreau, Y., and De Moor, B. TXTGate: profiling gene groups with text-based information. *Genome Biology 5(6)* (2005), R43. pages 17, 88, 135, 137

[50] Glenisson, P., Glänzel, W., Janssens, F., and De Moor, B. Combining full text and bibliometric information in mapping scientific disciplines. *Information Process Management 41* (2005), 1548–1572. pages 14, 88, 113

[51] Glenisson, P., Glänzel, W., and Olle, P. Combining full-text analysis and bibliometric indicators. a pilot study. *Scientometrics 63(1)* (2005), 163–180. pages 14

[52] Glenisson, P., Mathys, J., and De Moor, B. Meta-clustering of gene expression data and literature-based information. *ACM SIGKDD Explorations Newsletter 5(2)* (2003), 101–112. pages 14, 155

[53] Golbeck, J., Fragoso, G., Hartel, F., Hendler, J., Parsia, B., and Oberthaler, J. The national cancer institute's thesaurus and ontology. *Journal of Web Semantics 1* (2003), 75–80. pages 140

[54] Goreinov, S. A., Oseledets, I. V., and Savostyanov, D. V. Wedderburn rank reduction and krylov subspace method for tensor approximation. part 1: Tucker case. *arXiv:1004.1986v2 [math.NA]* (2010). pages 157

[55] Gospodnetic, O., and Hatcher, E. *Lucene in action.* Manning Publications, New York, 2005. pages 90

[56] Han, J., and Gao, J. *Research Challenges for Data Mining in Science and Engineering.* Chapman & Hall, June 2009. pages 16

[57] Harshman, R. A. Fundations of the parafac procedure: Model and conditions for an "explanatory" multi-modal factor analysis. *UCLA working papers in phonetics 16* (1970), 1–84. pages 44

[58] Hatche, r., and Gospodnetić, O. *Lucene in Action.* Manning Publications Co., 2004. pages 138

[59] He, X., Zha, H., Ding, C., and Simon, H. Web document clustering using hyperlink structures. *Computational Statistics and Data Analysis. 41*, 1 (2002), 19–45. pages 88, 114, 120

[60] Homayouni, R., Heinrich, K., Wei, L., and Berry, M. W. Gene clustering by latent semantic indexing of MEDLINE abstracts. *Bioinformatics 21(1)* (2005), 104–115. pages 143, 151

[61] Hou, C., Zhang, C., Wu, Y., and Nie, F. Multiple view semi-supervised dimensionality reduction. *Pattern Recogn 43(3)* (2010), 720–730. pages 8, 9

[62] Huang, H., Ding, C., Luo, D., and Li, T. Simultaneous tensor subspace selection and clustering: the equivalence of high order svd and k-means clustering. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (New York, NY, USA, 2008), ACM, pp. 327–335. pages 38

[63] Hubert, L., and Arabie, P. Comparing partitions. *Journal of Classification 2*, 1 (1985), 193–218. pages 45, 71, 97, 122, 143

[64] Ishteva, M.and De Lathauwer, L., Absil, P.-A., and Van Huffel, S. Differential-geometric Newton algorithm for the best rank-(r_1, r_2, r_3) approximation of tensors. *Numerical Algorithms 51(2)* (2009), 179–194. pages 61

[65] Ishteva, M., De Lathauwer, L., Absil, P.-A., and Van Huffel, S. Best low multilinear rank approximation of higher-order tensors, based on the Riemannian trust-region scheme. *SIAM Journal on Matrix Analysis and Applications 32(1)* (2011), 115–135. pages 42, 157

[66] Jaccard, P. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société Vaudoise des Sciences Naturelles 37* (1901), 547–579. pages 126

[67] Jain, A. K. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters 31(8)* (2010), 651–666. pages 7, 8

[68] Jain, A. K., and Dubes, R. C. *Algorithms for Clustering Data*. Prentice Hall, 1988. pages 94, 122, 142

[69] Janssens, F. *Clustering of scientific fields by integrating text mining and bibliometrics.* PhD thesis, Faculty of Engineering, K.U.Leuven, 2007. pages 12, 14, 15, 44, 88, 90, 91, 98, 99, 109

[70] Janssens, F., Glänzel, W., and De Moor, B. A hybrid mapping of information science. *Scientometrics 75*, 3 (2008), 607–631. pages 88, 99, 122

[71] Janssens, F., Leta, J., Glänzel, W., and De Moor, B. Towards mapping library and information science. *Information Processing Management 42* (2006), 1614–1642. pages 113, 121

[72] Janssens, F., Tran Quoc, V., Glänzel, W., and De Moor, B. Integration of textual content and link information for accurate clustering of science fields. In *Proceedings of the I International Conference on Multidisciplinary Information Sciences and Technologies* (2006), InSciT2006, pp. 615–619. pages 113, 121

[73] Janssens, F., Zhang, L., De Moor, B., and Glänzel, W. Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing and Management 45*, 6 (2009), 683–702. pages 14, 15, 57, 90, 105, 126

[74] Joachims, T., Cristianini, N., and Shawe-Taylor, J. Composite kernels for hypertext categorisation. In *Proceedings of the Eighteenth International Conference on Machine Learning* (2001), ICML'01, pp. 250–257. pages 31, 44, 113

[75] Kim, M.-Y., Dou, Q., Zaiane, O. R., and Goebel, R. Unsupervised mapping of sentences to biomedical concepts based on integrated information retrieval model. In *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology* (New York, NY, USA, 2010), ACM, pp. 322–329. pages 136

[76] Kolda, T. G., and Bader, B. W. The TOPHITS model for higher-order web link analysis. In *Proceedings of the SIAM Data Mining Conference Workshop on Link Analysis, Counterterrorism and Security* (2006). pages 10, 28

[77] Kolda, T. G., and Bader, B. W. Tensor decompositions and applications. *SIAM Review 51*, 3 (2009), 455–500. pages 10, 28, 35

[78] Kolda, T. G., and Sun, J. Scalable tensor decompositions for multi-aspect data mining. In *Proceedings of the 8th IEEE International Conference on Data Mining* (December 2008), pp. 363–372. pages 66, 84, 157

[79] Krallinger, M., Alonso-Allende, R., and Valencia, A. Text-mining approaches in molecular biology and biomedicine. *Drug Discovery Today 10* (2005), 439–445. pages 135

[80] Kriegel, H., Kröger, P., and Zimek, A. Outlier detection techniques (tutorial). In *Proceedings of 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining* (2009). pages 159

[81] Krings, G., Calabrese, F., Ratti, C., and Blondel, V. D. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment 2009* (2009), L07003. pages 118

[82] Kroonenberg, P., and de Leeuw, J. Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika* (1980). pages 40

[83] Lambiotte, R., and Panzarasa, P. Communities, knowledge creation, and information diffusion. *Journal of Informetrics 3(3)* (2009), 180–190. pages 114, 118

[84] Lanckriet, G. R. G., De Bie, T., Cristianini, N., Jordan, M. I., and Noble, W. A statistical framework for genomic data fusion. *Bioinformatics 20(6)* (2004), 2626–2635. pages 110

[85] Lanckriet, G. R. G., Deng, M., Cristianini, N., Jordan, M. I., and Noble, W. Kernel-based data fusion and its application to protein function prediction in yeast. In *Proceedings of the Pacific Symposium on Biocomputing* (2004), pp. 300–311. pages 110

[86] Leach, S. M., Tipney, H., Feng, W., Baumgartner, W. A., Kasliwal, P., Schuyler, R. P., Williams, T., Spritz, R. A., and Hunter, L. Biomedical discovery acceleration, with applications to craniofacial development. *PLoS computational biology 5*, 3 (2009), e1000215+. pages 135

[87] Leydesdorff, L. Can scientific journals be classified in terms of aggregated journal-journal citation relations using the journal citation reports? *Journal of the American Society for Information Science and Technology 57* (2006), 601–613. pages 87

[88] Leydesdorff, L., and Rafols, I. A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology 60* (2009), 348–362. pages 87, 114

[89] Lin, Y.-R., Sun, J., Castro, P., Konuru, R., Sundaram, H., and Kelliher, A. Metafac: community discovery via relational hypergraph factorization. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (2009), pp. 527–536. pages 10

[90] Liu, X., Yu, S., Janssens, F., Glänzel, W., Moreau, Y., and De Moor, B. Weighted hybrid clustering by combining text mining and bibliometrics on a large-scale journal database. *Journal of the American Society for Information Science and Technology 61*, 6 (2010), 1105–1119. pages 14, 139

[91] Liu, X., Yu, S., Moreau, Y., De Moor, B., Glänzel, W., and Janssens, F. Hybrid clustering of text mining and bibliometrics applied to journal sets. In *Proceedings of SIAM International Conference on Data Mining* (2009), pp. 49–60. pages 27, 28, 31, 48, 57, 59, 88, 92, 101, 110

[92] Long, B., Yu, P. S., and Zhang, Z. M. A general model for multiple view unsupervised learning. In *Proceedings of the 2008 SIAM International Conference on Data Mining* (2008), pp. 822–833. pages 28, 31, 58

[93] Lu, H., Plataniotis, K. N., and Venetsanopoulos, A. N. Multilinear principal component analysis of tensor objects for recognition. *IEEE Transactions on Neural Networks 19(1)* (2008), 18–39. pages 61

[94] Luxburg, U. A tutorial on spectral clustering. *Statistics and Computing 17(4)* (2007), 395–416. pages 28, 32, 33, 52, 61, 115

[95] Marshakova, I. System of connections between documents based on references (as the science citation index). *Nauchno-Tekhnicheskaya Informatsiya Seriya 2*, 6 (1973), 3–8. pages 87

[96] Mirkin, B. Reinterpreting the category utility function. *Machine Learning 45* (2001), 219–228. pages 99

[97] Modha, D. S., and Spangler, W. S. Clustering hypertext with applications to web searching. In *Proceedings of the 7th ACM on Hypertext and Hypermedia* (2000), New York:ACM Press, pp. 143–152. pages 88, 113

[98] Moya-Anegón, F. d., Vargas-Quesada, B., Chinchilla-Rodríguez, Z., Corera-Álvarez, E., Munoz-Fernández, F. J., and Herrero-Solana, V. Visualizing the marrow of science. *Journal of the American Society for Information Science and Technology 58* (2007), 2167–2179. pages 87

[99] Mucha, P. J., Richardson, T., Macon, K., Porter, M. A., and Onnela, J.-P. Community structure in time-dependent, multiscale, and multiplex networks. *Science 328(5980)* (2010), 876–878. pages 11, 27, 31, 45

[100] Mullins, N., and Snizek, W.and Oehler, K. *Handbook of quantitative studies of science and technology.* Elsevier Science, New York, 1988, ch. The structural analysis of a scientific paper, pp. 81–105. pages 113

[101] Muralidhara, C., Gross, A. M., Gutell, R. R., and Alter, O. Tensor decomposition reveals concurrent evolutionary convergences and divergences and correlations with structural motifs in ribosomal RNA. *Public Library of Science One (PLoS ONE) 6*, 4 (2011), e18768. pages 28

[102] Nallapati, R. M., Ahmed, A., Xing, E. P., and Cohen, W. W. Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (New York, NY, USA, 2008), KDD '08, ACM, pp. 542–550. pages 60

[103] Névéol, A., Kim, W., Wilbur, W. J., and Lu, Z. Exploring two biomedical text genres for disease recognition. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing* (Stroudsburg, PA, USA, 2009), Association for Computational Linguistics, pp. 144–152. pages 136

[104] Névéol, A., Shooshan, S., Humphrey, S., Rindflesh, T., and Aronson, A. Multiple approaches to fine-grained indexing of the biomedical literature. In *Pac Symp Biocomput* (2007), World Scientific, pp. 292–303. pages 136

[105] Newman, M. E. J. Analysis of weighted networks. *Physical Review E 70(5)* (2004), 056131+. pages 116

[106] Newman, M. E. J. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E 74(3)* (2006), 036104. pages 69, 70, 119

[107] Newman, M. E. J. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences 103(23)* (2006), 577–8582. pages 97, 116

[108] Ng, A., Jordan, M., and Weiss, Y. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems* (2001), MIT Press, pp. 849–856. pages 32, 33

[109] Noble, W. S., and Ben-hur, A. *Bioinformatics-From Genomes to Therapies*. Wiley-VCH Verlag GmbH, Weinheim, Germany, 2008, ch. Integrating Information for Protein Function Prediction, pp. 1297–1314. pages 9, 109

[110] Omberg, L., Golub, G. H., and Alter, O. A tensor higher-order singular value decomposition for integrative analysis of dna microarray data from different studies. *Proceedings of the National Academy of Sciences USA (PNAS) 104*, 47 (2007), 18371–18376. pages 28, 38

[111] Oseledets, I., and Tyrtyshnikov, E. Recursive and tensor-train decompositions in higher dimensions. In *Proceedings of The 9th Hellenic European Research on Computer Mathematics and Conference its Applications* (2009), HERCMA 2009, pp. 1–5. pages 157

[112] Overton, M. L., and Womersley, R. S. Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices. *Mathematical Programming 62*, 2 (1993), 321–357. pages 33

[113] Page, L., Brin, S., Motwani, R., and Winograd, T. The pagerank citation ranking: Bringing order to the web. Tech. rep., Stanford Digital Library Technologies Project, 1998. pages 5

[114] Pan, S. J., and Yang, Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering 22* (2010), 1345–1359. pages 158

[115] Phan, A., and Cichocki, A. Tensor decompositions for feature extraction and classification of high dimensional datasets. *Nonlinear Theory and Its Applications, IEICE in print* (2010). pages 67

[116] Porter, M. A., Onnela, J.-P., and Mucha, P. J. Communities in networks. *Notices of the American Mathematical Society 56 (9)* (2009), 1082–1097, 1164–1166. pages 115, 118

[117] Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics 20* (1987), 53–65. pages 97

[118] Rüping, S., and Scheffer, T. Learning from multiple views. In *Proceedings ICML workshop on Learning with Multiple Views* (2005). pages 2, 6

[119] Salton, G., and Buckley, C. Term-weighting approaches in automatic text retrieval. *Information Processing and Management 24* (1988), 513–523. pages 138

[120] Salton, G., and McGill, M. J. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc, New York, NY, USA, 1986. pages 115

[121] Savas, B., and Eldén, L. Krylov-type methods for tensor computations. *arXiv:1005.0683v2 [math.NA]* (2010). pages 31, 66, 84, 157

[122] Selee, T. M., Kolda, T. G., Kegelmeyer, W. P., and Griffin, J. D. Extracting clusters from large datasets with multiple similarity measures using IMSCAND. In *CSRI Summer Proceedings 2007, Technical Report SAND2007-7977, Sandia National Laboratories, Albuquerque, NM and Livermore, CA* (2007), M. L. Parks and S. S. Collis, Eds., pp. 87–103. pages 10, 28, 32, 157

[123] Shi, J., and Malik, J. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8)* (2000), 888–905. pages 32, 33

[124] Small, H. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science 24*, 4 (1973), 265–269. pages 87

[125] Small, H. G. Cited documents as concept symbols. *Social Studies of Science 8*, 3 (1978), 327–340. pages 87

[126] Snizek, W.and Oehler, K., and Mullins, N. Textual and nontextual characteristics of scientific papers: Neglected science indicators. *Scientometrics 20*, 1 (1991), 25–35. pages 113

[127] Strehl, A., and Ghosh, J. Cluster ensembles- a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research 3* (2002), 583–617. pages 10, 28, 32, 44, 45, 58, 71, 92, 93, 96, 98, 122, 143

[128] Sun, J., Tao, D., and Faloutsos, C. Beyond streams and graphs: dynamic tensor analysis. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (New York, NY, USA, 2006), ACM, pp. 374–383. pages 10, 28, 32

[129] Tan, P.-N., Steinbach, M., and Kumar, V. *Introduction to Data Mining*. Addison-Wesley, 2005. pages 143

[130] Tang, L., Wang, X., and Liu, H. Uncovering groups via heterogeneous interaction analysis. In *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining* (Washington, DC, USA, 2009), IEEE Computer Society, pp. 143–152. pages 11, 45, 58

[131] Tang, L., Wang, X., and Liu, H. Community detection in multi-dimensional networks. Technical Report TR10-006, School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AS, USA, 2010. pages 28, 31, 44, 114

[132] Tang, W., Lu, Z., and Dhillon, I. S. Clustering with multiple graphs. In *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining* (Washington, DC, USA, 2009), IEEE Computer Society, pp. 1016–1021. pages 28, 31, 44, 58

[133] Topchy, A., Jain, A. K., and Punch, W. Clustering ensembles: Models of consensus and weak partitions. *IEEE Transactions on Pattern Analysis and Machine Intelligence 27* (2005), 1866–1881. pages 98

[134] Tucker, L. The extension of factor analysis to three-dimensional matrices. In *Contributions to mathematical psychology*, H. Gulliksen and N. Frederiksen, Eds. Holt, Rinehart & Winston, NY, 1964, pp. 109–127. pages 37

[135] Verma, D., and Meila, M. A comparison of spectral clustering algorithms. Tech. rep., Department of CSE University of Washington Seattle, WA, 2003. pages 28

[136] Wang, F., Ding, C., and Li, T. Integrated KL (K-means - Laplacian) clustering: A new clustering approach by combining attribute data and pairwise relations. In *Proceedings of the 9th SIAM Conference on Data Mining (SDM)* (Sparks, Nevada, USA., 2009), SIAM, pp. 38–49. pages 60

[137] Wang, H., and Ahuja, N. A tensor approximation approach to dimensionality reduction. *International Journal of Computer Vision 76* (2008), 217–229. pages 61, 157

[138] Wang, Y., and Kitsuregawa, M. Evaluating contents-link coupled web page clustering for web search results. In *Proceedings of the eleventh international conference on Information and knowledge management* (2002), CIKM '02, pp. 499–506. pages 88, 113

[139] Yan, R., and Hauptmann, A. A review of text and image retrieval approaches for broadcast news video. *Information Retrieval 10*, 4 (2007), 445–484. pages 155

[140] Yang, J., Zhang, D., Frangi, A. F., and Yang, J.-y. Two-dimensional pca: A new approach to appearance-based face representation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell. 26* (January 2004), 131–137. pages 61

[141] Yang, T., Jin, R., Chi, Y., and Zhu, S. Combining link and content for community detection: a discriminative approach. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (New York, NY, USA, 2009), KDD '09, ACM, pp. 927–936. pages 60

[142] Yarowsky, D. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics* (Stroudsburg, PA, USA, 1995), ACL '95, Association for Computational Linguistics, pp. 189–196. pages 83

[143] Ye, J. Generalized low rank approximations of matrices. *Machine Learning 61* (2005), 167–191. pages 61, 67

[144] Ye, J., Janardan, R., and Li, Q. GPCA: an efficient dimension reduction scheme for image compression and retrieval. In *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery*

and data mining (New York, NY, USA, 2004), ACM, pp. 354–363. pages 29, 65

[145] Yu, S., Liu, X., Tranchevent, L. C., Glänzel, W., Suykens, J., De Moor, B., and Moreau, Y. Optimized data fusion for k-means Laplacian clustering. *Bioinformatics 27(1)* (2010), 118–126. pages 60, 70, 110, 143

[146] Yu, S., Tranchevent, L.-C. C., De Moor, B., and Moreau, Y. Gene prioritization and clustering by multi-view text mining. *BMC bioinformatics 11* (2010), 28. pages 135, 136, 137, 138, 139

[147] Yu, S., Van Vooren, S., Tranchevent, L., De Moor, B., and Moreau, Y. Comparison of vocabularies, representations and ranking algorithms for gene prioritization by text mining. *Bioinformatics 24*, 16 (2008), i119–i125. pages 17, 140

[148] Zaïane, O. R. *Principles of Knowledge Discovery in Databases*. Department of Computing Science, University of Alberta,, Edmonton, Alberta, Canada, 1999. pages 7

[149] Zhang, L., Liu, X., Janssens, F., Linag, L., and Glänzel, W. Subject clustering analysis based on ISI category classification. *Journal of Informetrics 4*, 2 (2010), 185–193. pages 114, 116

[150] Zhou, D., and Burges, C. J. C. Spectral clustering and transductive learning with multiple views. In *Proceedings of the 24th international conference on Machine learning* (New York, NY, USA, 2007), ACM, pp. 1159–1166. pages 1, 28, 31, 58

[151] Zhou, D., Zhu, S., Yu, K., Song, X., Tseng, B. L., Zha, H., and Giles, C. L. Learning multiple graphs for document recommendations. In *Proceedings of the 17th international conference on World Wide Web* (New York, NY, USA, 2008), WWW '08, ACM, pp. 141–150. pages 60, 83

[152] Zhu, S., Yu, K., Chi, Y., and Gong, Y. Combining content and link for classification using matrix factorization. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (New York, NY, USA, 2007), SIGIR '07, ACM, pp. 487–494. pages 60, 83

[153] Zitt, M., and Bassecoulard, E. Development of a method for detection and trend analysis of research fronts built by lexical or cocitation analysis. *Scientometrics 30*, 1 (1994), 333–351. pages 88

Curriculum vitae

Xinhai Liu was born in Henan, China in August, 1976. He received the degree in Electrical Engineering from Wuhan University of Science and Technology (WUST), Hubei, China, in 1999 and Master degree in Computer Science from Wuhan University of Science and Technology. Afterwards, he worked as a teaching assistant and then a lecture in School of Information Science and Engineering of Wuhan University of Science and Technology.

With the joint scholarship between Katholieke University Leuven and China Scholarship Council (CSC), Xinhai began his study in Department of Electrical Engineering (ESAT) of Katholieke University Leuven since 2006. Afterwards, he entered a PhD program in the lab of Signals, Identification, System Theory and Automation (SCD/SISTA). Since then, he worked in the Systems, Models and Control (SMC) subgroup, under the supervision of Professor Bart De Moor. Xinhai's research interests include multilinear analysis, network analysis, information retrieval and text mining. His research was supported by

1. K.U.Leuven-China Scholarship Council (CSC): Excellence scholarship (CSC, No. 2006153005),
2. Research Council K.U.Leuven: GOA Ambiorics, GOA MaNet, CoE EF/05/006 Optimization in Engineering (OPTEC),
3. EU porject: ERNSI; FP7-HD-MPC (INFSO-ICT-223854), COST intelli-CIS, FP7-EMBOCON (ICT-248940),
4. FWO: PhD/postdoc grants, projects: G0226.06 (cooperative systems and optimization), G0321.06 (Tensors), G.0302.07 (SVM/Kernel),
5. Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, Dynamical systems, control and optimization, 2007-2011).

Publications by author

Journal Papers

- Liu X., Yu S., Janssens F., Glänzel W., Moreau Y., De Moor B.. Weighted hybrid clustering by combining text mining and bibliometrics on large-scale journal database, Journal of the American Society for Information Science and Technology, vol. 61, no. 6, Jun. 2010, pp. 1105-1119.
- Liu X., De Moor B., Glänzel W.. Optimal and hierarchical clustering of large-scale hybrid networks for scientific mapping, Scientometrics, 2011, in press.
- Yu S., Liu X., Tranchevent L., Glänzel W., Suykens J., De Moor B., Moreau Y.. Optimized data fusion for K-means Laplacian clustering, Bioinformatics, vol. 27, no. 21, Jan. 2011, pp. 118-126.
- Zhang L., Liu X., Janssens F., Liang L., Glänzel W.. Subject clustering analysis based on ISI category classification, Journal of Informetrics, vol. 4, no. 2, Apr. 2010, pp. 185-193.
- Yu S., Tranchevent L., Liu X., Glänzel W., Suykens J., De Moor B., Moreau Y.. Optimized data fusion for kernel k-means clustering, Under review of IEEE Transactions on Pattern Analysis and Machine Intelligence.
- Liu X., Gevaert O., Tranchevent L., Moreau Y., De Moor B.. A web portal of multi-view text mining for multi-modality analysis and vertical searches, submitted to BMC Bioinformatics.
- Liu X., De Lathauwer L., Glänzel W., De Moor B.. Multi-view partitioning via tensor methods, Under review of IEEE Transactions on Knowledge and Data Engineering.

- Liu X., De Lathauwer L., Glänzel W., De Moor B.. Optimal clustering and joint dimension reduction of multiple graphs, in preparation.

Conference Papers

- Liu X., De Moor B., Glänzel W.. A hierarchical and optimal clustering of the WoS journal database by hybrid information, in Proceedings of 13th International Conference on Scientometrics and Informetrics (ISSI2011), Durban, South Africa, July 2011, pp. 485-496.
- Liu X., De Lathauwer L., Janssens F., De Moor B.. Hybrid clustering on multiple information sources via HOSVD, in Proceedings of the 7th International Symposium on Neural Networks (ISNN 2010), Shanghai, China, Jun. 2010, pp. 337-345.
- Liu X., Yu S., Moreau Y., De Moor B., Glänzel W., Janssens F.. Hybrid clustering by integrating text and citation based graphs in journal database analysis, in Proceedings of the 2009 IEEE International Conference on Data Mining Workshops (ICDMW2009), Miami, Florida, Dec. 2009, pp. 521-526.
- Liu X., Yu S., Moreau Y., De Moor B., Glänzel W., Janssens F.. Hybrid clustering of text mining and bibliometrics applied to journal sets, in Proceedings of the SIAM Data Mining Conference 09 (SIAM DM 09), Sparks, Nevada USA, May 2009, pp. 46-60.

Abstracts

- Liu X., De Lathauwer L., Glänzel W., De Moor B.. Hybrid clustering of multi-view data via MLSVD. Workshop on Tensor Decompositions and Applications (TDA 2010), Monopoli, Bari, Italy, 2010.
- Liu X., Glänzel W., De Moor B.. Graph model based community detection by incorporating text mining and link analysis, Workshop on Advances in Bio Text Mining, Ghent, Belgium, 2010.
- Liu X., Gevaert O., Tranchevent L., Moreau Y., De Moor B.. Biomedical text mining from multiple views: information fusion and vertical search, 19th Annual International Conference on Intelligent Systems for Molecular Biology and 10th European Conference on Computational Biology (ISMB/ECCB 2011), Vienna, Austria, 2011.

Arenberg Doctoral School of Science, Engineering & Technology
Faculty of Engineering
Department of Electrical Engineering
System, Control and Model group SCD/SISTA
Kasteelpark Arenberg 10
Heverlee-Leuven, B-3001, Belgium