3-25-2014

# Towards Next Generation Vertical Search Engines

Li Zheng
zhengli8341@gmail.com

**DOI:** 10.25148/etd.FI14071146

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

# TOWARDS NEXT GENERATION VERTICAL SEARCH ENGINES

A dissertation submitted in partial fulfillment of the

requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Li Zheng

2014

To: Dean Amir Mirmiran
College of Engineering and Computing

This dissertation, written by Li Zheng, and entitled Towards Next Generation Vertical Search Engines, having been approved in respect to style and intellectual content, is now referred to you for judgment.

We have read this dissertation and recommend that it be approved.

Zhenmin Chen

Jainendra K. Navlakha

Sundaraja Sitharama Iyengar

Shu-Ching Chen, Co-major Professor

Tao Li, Co-major Professor

Date of Defense: March 25, 2014

The dissertation of Li Zheng is approved.

Dean Amir Mirmiran
College of Engineering and Computing

Dean Lakshmi N. Reddi
University Graduate School

Florida International University, 2014

# ACKNOWLEDGMENTS

I would like to express the deepest appreciation to my committee chair Professor Tao Li. As my major advisor, he continually and persuasively conveyed a spirit of adventure in regard to research, scholarship, leadership, and an excitement in regard to teaching. Without his supervision and help this dissertation would not have been possible. Thanks for his kind guidance which shines my way where the research meets the practice.

I would like to express the deepest appreciation to my co-advisor Professor Shu-Ching Chen, who has shown the great attitude and responsibility of a group leader: he always has influence on me by his passion and conscientiousness. Thanks for his solid support which gives me the opportunity to practice leadership in Disaster Information Management group.

I would like to thank my committee members, Professor S. S. Iyengar, Professor Jainendra K. Navlakha, and Professor Zhenmin Chen, whose works encourage my research and motivate me to grow as a scientist.

In addition, I will give my sincere thanks to all the members in FIU Disaster Information Management Group. A thank you to Steve Luis, who works with and supervises me for all those years in this interest area. A thank to Chao Shen and Liang Tang, from whom I learned quite a lot. I will also convey my thanks to all the members in Professor Tao Li's Knowledge Discovery Research Group.

Finally, I would like to express my appreciation to FIU School of Computing and Information Science with amazing resources, assistances, and nice staff. The past 6 years during which I dedicated myself to the research and life means a new beginning in my career.

ABSTRACT OF THE DISSERTATION

# TOWARDS NEXT GENERATION VERTICAL SEARCH ENGINES

by

Li Zheng

Florida International University, 2014

Miami, Florida

Professor Tao Li, Co-major Professor

Professor Shu-Ching Chen, Co-major Professor

As the Web evolves unexpectedly fast, information grows explosively. Useful resources become more and more difficult to find because of their dynamic and unstructured characteristics. A vertical search engine is designed and implemented towards a specific domain. Instead of processing the giant volume of miscellaneous information distributed in the Web, a vertical search engine targets at identifying relevant information in specific domains or topics and eventually provides users with up-to-date information, highly focused insights and actionable knowledge representation. As the mobile device gets more popular, the nature of the search is changing. So, acquiring information on a mobile device poses unique requirements on traditional search engines, which will potentially change every feature they used to have. To summarize, users are strongly expecting search engines that can satisfy their individual information needs, adapt their current situation, and present highly personalized search results.

In my research, the next generation vertical search engine means to utilize and enrich existing domain information to close the loop of vertical search engine's system that mutually facilitate knowledge discovering, actionable information extraction, and user interests modeling and recommendation. I investigate three problems in which domain taxonomy plays an important role, including taxonomy generation using a vertical search engine, actionable information extraction based on domain taxonomy, and the use of ensemble taxonomy to catch user's interests. As the fundamental theory, ultra-metric, dendrogram, and hierarchical clustering are intensively discussed. Methods on taxonomy generation using my research on hierarchical clustering are developed. The related vertical search engine techniques are practically used in Disaster Management Domain. Especially, three disaster information management systems are developed and represented as real use cases of my research work.

TABLE OF CONTENTS

CHAPTER PAGE

## LIST OF FIGURES

FIGURE PAGE

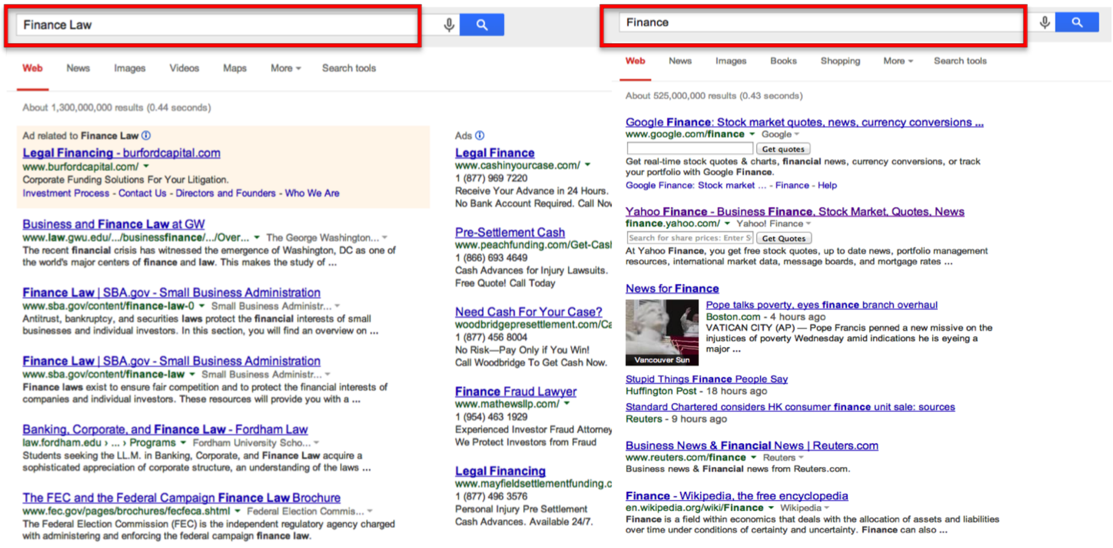
CHAPTER 1

# Introduction

As the Web evolves unexpectedly fast, information grows explosively. Useful resources become more and more difficult to find because of their dynamic and unstructured characteristics. General search engines, such as Google (www.google.com), Yahoo (www.yahoo.com), and Bing (www.bing.com), can largely help people figure out many important resources based on each user's customized queries. However, as the size of indexable Web pages keeps exploding, it is impossible for a search engine to maintain an index with both comprehensiveness and freshness. Google, considered the best search index today, can only catalog a fraction of these massive contents. Even with spiders to crawl millions of web pages per week, Google's current index size is only 40 billion out of the 450+ billion pages estimated to exist, less than 10% of all available internet-served content. In addition, under many circumstances, the general-purpose search engine, such as Google, can easily generate millions of search results, but many of them are not relevant to the users intension or are duplications of each other. For example, when the keywords can be generally used in various situations or areas, the generated results will be highly diversified. Also when the keywords can be presented in several equivalent forms or expressions, site information from a specific domain will not often be included in the top hits.

## 1.1 Next generation vertical search engine

A vertical search engine is designed and implemented towards a specific domain. In terms of functionality, a vertical search engine is a further refinement and extension based on general search engines. Instead of processing the giant volume of miscellaneous information distributed in the Web, a vertical search engine targets at identifying relevant information in specific domains or topics and eventually provides users with up-to-date information, highly focused insights and actionable knowledge representation.



About 1,300,000,000 results (0.44 seconds)

Ad related to **Finance Law**

**Legal Financing** - burfordcapital.com
www.burfordcapital.com/
Corporate Funding Solutions For Your Litigation.
Investment Process - Contact Us - Directors and Founders - Who We Are

Business and **Finance Law** at GW
www.**law**.gwu.edu/.../business**finance**/.../Over... The George Washington...
The recent **financial** crisis has witnessed the emergence of Washington, DC as one of the world's major centers of **finance** and **law**. This makes the study of ...

**Finance Law** | SBA.gov - Small Business Administration
www.sba.gov/content/**finance-law**-0 Small Business Administr...
Antitrust, bankruptcy, and securities **laws** protect the **financial** interests of small businesses and individual investors. In this section, you will find an overview on ...

**Finance Law** | SBA.gov - Small Business Administration
www.sba.gov/content/**finance-law** Small Business Administr...
**Finance laws** exist to ensure fair competition and to protect the financial interests of companies and individual investors. These resources will provide you with a ...

Banking, Corporate, and **Finance Law** - Fordham Law
law.fordham.edu › ... › Programs Fordham University Scho...
Students seeking the LL.M. in Banking, Corporate, and **Finance Law** acquire a sophisticated appreciation of corporate structure, an understanding of the laws ...

The FEC and the Federal Campaign **Finance Law** Brochure
www.fec.gov/pages/brochures/fecfeca.shtml Federal Election Commis...
The Federal Election Commission (FEC) is the independent regulatory agency charged with administering and enforcing the federal campaign **finance law**.

Ads

**Legal Finance**
www.cashinyourcase.com/
1 (877) 969 7220
Receive Your Advance in 24 Hours.
No Bank Account Required. Call Nov

Pre-Settlement Cash
www.peachfunding.com/Get-Casl
1 (866) 693 4649
Cash Advances for Injury Lawsuits.
Free Quote! Call Today

Need Cash For Your Case?
woodbridgepresettlement.com/Ca
1 (877) 456 8004
No Risk—Pay Only if You Win!
Call Woodbridge To Get Cash Now.

**Finance** Fraud Lawyer
www.mathewsllp.com/
1 (954) 463 1929
Experienced Investor Fraud Attorney
We Protect Investors from Fraud

**Legal Financing**
www.mayfieldsettlementfunding.c
1 (877) 496 3576
Personal Injury Pre Settlement
Cash Advances. Available 24/7.

About 525,000,000 results (0.43 seconds)

Google **Finance**: Stock market quotes, news, currency conversions ...
www.google.com/**finance** Google
Get quotes
Get real-time stock quotes & charts, **financial** news, currency conversions, or track your portfolio with Google **Finance**.
Google Finance: Stock market ... - Finance - Help

Yahoo **Finance** - Business **Finance**, Stock Market, Quotes, News
**finance**.yahoo.com/ Yahoo! Finance
Search for share prices: Enter S Get Quotes
At Yahoo **Finance**, you get free stock quotes, up to date news, portfolio management resources, international market data, message boards, and mortgage rates ...

News for **Finance**

Pope talks poverty, eyes **finance** branch overhaul
Boston.com - 4 hours ago
VATICAN CITY (AP) — Pope Francis penned a new missive on the injustices of poverty Wednesday amid indications he is eyeing a major ...

Stupid Things **Finance** People Say
Huffington Post - 18 hours ago

Standard Chartered considers HK consumer **finance** unit sale: sources
Reuters - 9 hours ago

Business News & **Financial** News | Reuters.com
www.reuters.com/**finance** Reuters
Business news & **Financial** news from Reuters.com.

**Finance** - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Finance** Wikipedia
**Finance** is a field within economics that deals with the allocation of assets and liabilities over time under conditions of certainty and uncertainty. **Finance** can also ...

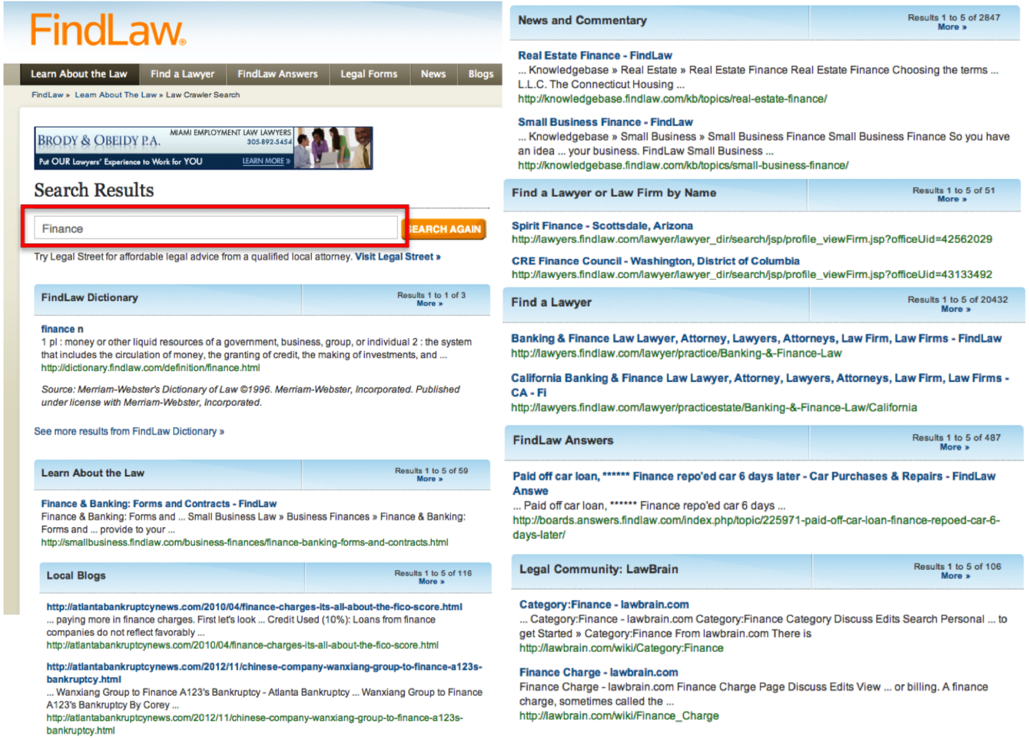
Figure 1.1: Google search results with keyword Finance.

Figure 1.1 and Figure 1.2 show the different top hits between Google and FindLaw by using the keyword "Finance". We can easily conclude that the user who is looking for information about legal assistance definitely will prefer FindLaw's results. The results in Figure 1.2 are all about legal issues relevant to "Finance" and information is well categorized to match user's typical purposes, such as "Find a Lawyer" and "Local Blogs". Even when we change the keyword to "Finance Law", in Figure 1.1, the refined google results are much less comprehensive and actionable than FindLaw.

There are many hot applied areas, such as business, medicine, science, education and job, in which many vertical search engines are already implemented. FindLaw[1] is one of the leading law search engines. You can also access www.theweathernetwork.com to find weather reports, or www.expedia.com to search for available flights.

The nature of search is changing, especially on mobile devices. General search engine used to be the main starting point for consumers looking to buy products, visit sites, or discover hotspots. However, as mobile devices are getting more popular, using handsets to instantly start a search becomes possible from almost anywhere and anytime. Since

[1] http://www.findlaw.com



Search Results

Finance

Try Legal Street for affordable legal advice from a qualified local attorney. **Visit Legal Street »**

**FindLaw Dictionary** Results 1 to 1 of 3 **More »**

**finance n**
1 pl : money or other liquid resources of a government, business, group, or individual 2 : the system that includes the circulation of money, the granting of credit, the making of investments, and ...
http://dictionary.findlaw.com/definition/finance.html

*Source: Merriam-Webster's Dictionary of Law ©1996. Merriam-Webster, Incorporated. Published under license with Merriam-Webster, Incorporated.*

See more results from FindLaw Dictionary »

**Learn About the Law** Results 1 to 5 of 59 **More »**

**Finance & Banking: Forms and Contracts - FindLaw**
Finance & Banking: Forms and ... Small Business Law » Business Finances » Finance & Banking: Forms and ... provide to your ...
http://smallbusiness.findlaw.com/business-finances/finance-banking-forms-and-contracts.html

**Local Blogs** Results 1 to 5 of 116 **More »**

**http://atlantabankruptcynews.com/2010/04/finance-charges-its-all-about-the-fico-score.html**
... paying more in finance charges. First let's look ... Credit Used (10%): Loans from finance companies do not reflect favorably ...
http://atlantabankruptcynews.com/2010/04/finance-charges-its-all-about-the-fico-score.html

**http://atlantabankruptcynews.com/2012/11/chinese-company-wanxiang-group-to-finance-a123s-bankruptcy.html**
... Wanxiang Group to Finance A123's Bankruptcy - Atlanta Bankruptcy ... Wanxiang Group to Finance A123's Bankruptcy By Corey ...
http://atlantabankruptcynews.com/2012/11/chinese-company-wanxiang-group-to-finance-a123s-bankruptcy.html

**News and Commentary** Results 1 to 5 of 2847 **More »**

**Real Estate Finance - FindLaw**
... Knowledgebase » Real Estate » Real Estate Finance Real Estate Finance Choosing the terms ... L.L.C. The Connecticut Housing ...
http://knowledgebase.findlaw.com/kb/topics/real-estate-finance/

**Small Business Finance - FindLaw**
... Knowledgebase » Small Business » Small Business Finance Small Business Finance So you have an idea ... your business. FindLaw Small Business ...
http://knowledgebase.findlaw.com/kb/topics/small-business-finance/

**Find a Lawyer or Law Firm by Name** Results 1 to 5 of 51 **More »**

**Spirit Finance - Scottsdale, Arizona**
http://lawyers.findlaw.com/lawyer/lawyer_dir/search/jsp/profile_viewFirm.jsp?officeUid=42562029

**CRE Finance Council - Washington, District of Columbia**
http://lawyers.findlaw.com/lawyer/lawyer_dir/search/jsp/profile_viewFirm.jsp?officeUid=43133492

**Find a Lawyer** Results 1 to 5 of 20432 **More »**

**Banking & Finance Law Lawyer, Attorney, Lawyers, Attorneys, Law Firm, Law Firms - FindLaw**
http://lawyers.findlaw.com/lawyer/practice/Banking-&-Finance-Law

**California Banking & Finance Law Lawyer, Attorney, Lawyers, Attorneys, Law Firm, Law Firms - CA - Fi**
http://lawyers.findlaw.com/lawyer/practicestate/Banking-&-Finance-Law/California

**FindLaw Answers** Results 1 to 5 of 487 **More »**

**Paid off car loan, ****** Finance repo'ed car 6 days later - Car Purchases & Repairs - FindLaw Answe**
... Paid off car loan, ****** Finance repo'ed car 6 days ...
http://boards.answers.findlaw.com/index.php/topic/225971-paid-off-car-loan-finance-repoed-car-6-days-later/

**Legal Community: LawBrain** Results 1 to 5 of 106 **More »**

**Category:Finance - lawbrain.com**
... Category:Finance - lawbrain.com Category:Finance Category Discuss Edits Search Personal ... to get Started » Category:Finance From lawbrain.com There is
http://lawbrain.com/wiki/Category:Finance

**Finance Charge - lawbrain.com**
Finance Charge - lawbrain.com Finance Charge Page Discuss Edits View ... or billing. A finance charge, sometimes called the ...
http://lawbrain.com/wiki/Finance_Charge

Figure 1.2: FindLaw search results with keyword Finance.

the mobile device is portable (location sensitive) and personal (extremely user-centric), acquiring information on the mobile device poses unique requirements on a traditional search engine, which will potentially change every feature it used to have. To summarize, users are strongly expecting search engines that can satisfy their individual information needs, adapt their current situation, and present highly personalized search results.

A powerful vertical search engine can efficiently identify domain relevant resources, extracting critical information, and adapt the search results to specific user's needs. Therefore, the utilization and enrichment of existing domain information plays an important role in closing the loop of a vertical search engine's system. This mutually facilitates knowledge discovering and results representation as shown in Figure 1.3.

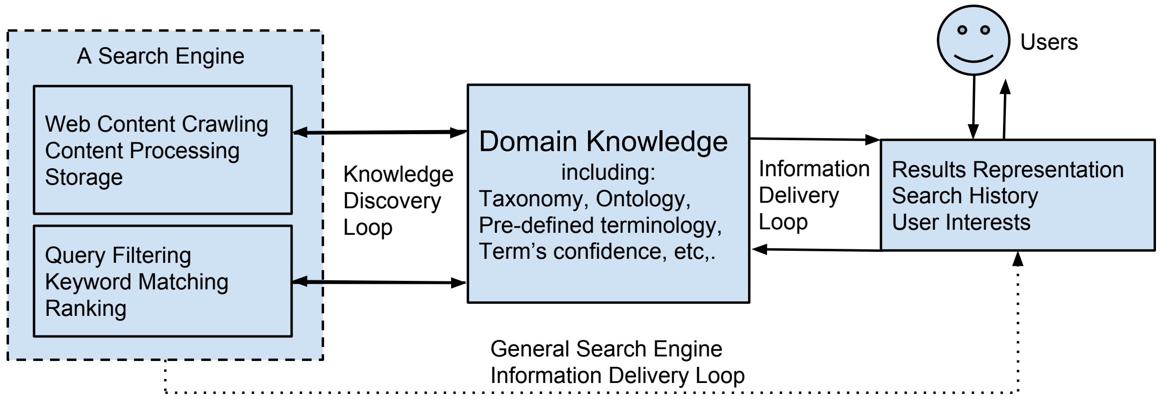


Figure 1.3: An example of vertical search engine system.

## 1.2 Problem statement

Given a set of URLs (seeds) related to a domain or a given topic, a vertical search engine needs to explore and maintain an appropriate amount of indexed Web pages so as to provide users with updated insight knowledge about important aspects of such domain. The provided search results should be based on the user's query and concentrate on the given topic so that the domain professionals can easily gain a deep and comprehensive understanding about some aspects of the topic. To fully utilize the advantages of existing domain knowledge, there are several challenges need to be solved:

- Challenge 1: *How to efficiently build domain taxonomy using vertical search engine?* There are ways of organizing domain related keywords and terms. A term hierarchy is the most popular form to represent the relationship between important concepts in a particular domain. An efficient and effective method to build a domain taxonomy based on textual content in such domain is expected to be stable and flexible. We model this problem as hierarchical clustering with constraints, which generates stable term hierarchy based on hierarchical clustering by transforming domain knowledge into constraints.

- Challenge 2: *How to extract actionable information from un-structured data resources?*

  To deal with un-structured data, general Information Retrieval and Natural Language Processing (NLP) techniques can be utilized to identify named entities and relations contained in a textual Web page. But it brings big challenges to figure out the domain-relevant entity (eg. related people, location, organization) and identify its current status. Domain taxonomy can be utilized to clustering and classify similar entities. On the other hand, such taxonomy can be dynamically adjusted as the domain evolves based on those on-topic resources the crawler collected. Also, models can be trained to identify the status information.

- Challenge 3: *How to efficiently capture the users's interests and deliver right information?*

  General search engine generates ranking list with diversified and repetitive results which does not support decision-making process very well in a specific domain. Since the users of a vertical search engine have focused search intensions, how to understand the user's interests and match the search results with corresponding user's profile is non-trivial task. Our proposed recommendation framework considers both user's historical interests and different user groups' common interests to recommend users the most relevant contents.

- Application Challenge: *How to apply vertical search engine techniques to disaster management domain?*

  Disaster management, as a national priority, gains massive attentions from both research and engineering community. Efforts from various academic areas are made to build a general framework in this domain. It is important to understand the characteristics of this domain and figure out the specific requirements which can be solved by adaptively applying vertical search techniques. Disaster information management domain taxonomy is also developed to support previous three tasks.

## 1.3 Summary of contributions

The following chapters give detailed discussions about critical techniques we proposed to deal with each challenge.

Chapter 2 lists the most recent work relevant to my research. Several major categories are mentioned and necessary discussions are given, such as Web Classification and Crawling, Domain Taxonomy Generation, Information Extraction, and Personalization and Recommendation.

In Chapter 3, we propose a hierarchical ensemble clustering framework which can naturally combine both partitional clustering and hierarchical clustering results. We study three important problems: Dendrogram Description, Dendrogram Combination and Dendrogram Selection. We develop two approaches for dendrogram selection based on tree distances and investigate various dendrogram distances for representing dendrograms. We provide a systematic empirical study of the ensemble hierarchical clustering problem. Experimental results demonstrate the effectiveness of our proposed approaches. The research works were published in [ZLD10a]

In Chapter 4, we propose a novel semi-supervised hierarchical clustering framework based on ultra-metric dendrogram distance. The proposed framework is able to incorporate triple-wise relative constraints. We establish the connection between hierarchical clustering and ultra-metric transformation of dissimilarity matrix and propose two techniques (the constrained optimization technique and the transitive dissimilarity based technique) for semi-supervised hierarchical clustering. Experimental results demonstrate the effectiveness and the efficiency of our proposed methods. The research works were published in [ZL11].

Chapter 5 considers the problem of how to extract useful information from the Web. Two related aspects are discussed: taxonomy generation and information extraction. We model the taxonomy generation problem as document hierarchical clustering with ordered constraints in which the constraints are given as a partially know hierarchy, the domain related concepts extracted from web documents are treated as instance and our solution is to build a term hierarchy which satisfies the relative hierarchical structure in given partial hierarchy. We also utilize techniques from information extraction and natural language processing to build efficient model to quickly extract structured status information from domain documents. A focus crawler prototype is presented in this chapter. The research works were published in [ZST$^+$12, ZST$^+$10].

Chapter 6 discusses our proposed information recommendation framework to satisfy online readers with their own reading preference. A novel personalized news recommendaTion framework using ensemble hierarchical clustering to provide attractive recommendation results. Specifically, given a set of online readers, our approach initially separates readers into different groups based on their reading histories, where each user might be designated to several groups. A document hierarchy is constructed for each user group. When recommending document to a given user, the hierarchies of multiple user groups that the user belongs to are merged into an optimal one. Finally a list of news articles are selected from this optimal hierarchy based on the users personalized information, as the recommendation result. Extensive empirical experiments on a set of news articles collected from various popular news websites demonstrate the efficacy of our proposed approach. The research works were published in [ZLHL12, ZLD10a].

Chapter 7 describes how the research can contribute to the real application in disaster management domain. We have developed techniques to facilitate information sharing and collaboration between both private and public sector participants for major disaster recovery planning and management. We have designed and implemented two parallel systems: a web-based prototype of a Business Continuity Information Network system and an All-Hazard Disaster Situation Browser system that run on mobile devices. Data mining and information retrieval techniques help impacted communities better understand the current disaster situation and how the community is recovering. User studies with more than 200 participants from Emergency Operation Center (EOC) personnel and companies demonstrate that our systems are very useful to gain insights about the disaster situation and for making decisions. The application works were published in [ZST$^+$11, ZST$^+$10, ZST$^+$12, ZST$^+$13, WZLD09].

Finally, Chapter 8 summarizes this dissertation comprehensively. For each important component, my research contributions are provided and future improvements are discussed.

CHAPTER 2

# Related work

This chapter presents related work in building a vertical search engine and domain knowledge generation. Existing techniques are categorized into four aspects. Section 2.1 provides previous research on web classification and crawling strategies; Section 2.2 provides relevant work on taxonomy generation techniques and information retrieval methods for entity recognition and relation extraction; Section 2.3 provides recommendation frameworks that are widely studied in various application areas; Section 2.4 provides relevant researches on various clustering techniques.

## 2.1 Web classification and crawling

### 2.1.1 Link-based algorithm

There are different types of link contextual information which can be evaluated in general focused crawler: *link context*, *ancester pages* and *web graph*. Early algorithms, like Fish search [DBP94], simply follow all links in an on-topic page by assuming the successive relevance from the parent page. Shark search uses a few words around a hyperlink to define more granular context [HJM$^+$98]. Richer information (header, title) is extracted from parent page to obtain more meaningful contextual information [CPS02, PM03]. By introducing the concept of context graphs [DCL$^+$00, HW06], features collected from paths (ancestors) leading up to relevant nodes are utilized to guide the crawler and back-links are used to estimate the link distance form a page to target pages. By considering the detected web graph to identify a ”good” hub[4], the priority of its following hyperlinks can be increased [CvdBD99, PM03].

[4] A page contains links to many relevant pages.

### 2.1.2 Machine learning-based algorithms

Researchers also explore different algorithms from machine learning perspective using various contextual information. Proper predicting models are trained to evaluate the relevance between detected hyperlinks and the topic. In [LJM06], a Hidden Markov Model (HMM) is trained based on user browsing history to predict how likely a page lead to a target page. Algorithms using reinforcement learning [RM99, MNRS99] are designed to learn a mapping performed by Naive Bayes text classifier from the text surrounding a hyperlink to a value function of sum of rewards. The estimation of the number of relevant pages can be obained as the results of following that hyperlink. Genetic Algorithm [JTG03, SCR05] is used to explore the space of potential strategies and evolve good strategies based on the text and link structure of the referring pages. The strategies produce a rank function which is a weighted sum of several scores such as hub, authority and SVM scores of parent pages going back k generations. A population of agents are modeled by an ANNs network to search for relevant pages and decide which links to follow using evolving query vectors [MBC$^+$99, MM99].

### 2.1.3 Ontology-driven crawling strategies

Ontology[1] is defined as a well-organized knowledge scheme that represents high-level background knowledge with concepts and relations. Research work presented in [MEH$^+$02, EM03, SGYL05] utilizes ontology to evaluate the relevance between web pages and topic. Entities in visited pages are processed by calculating the entity distance (simply the linking steps between an entity in the ontology and the crawling topic), thus the concept weights of a page can be generated by a heuristically predefined discount factor raised to the power of the entity distance. The page relevance score is equivalent to the summation of concept weights multiplied by the frequencies of corresponding entities in the visited web

[1] http://en.wikipedia.org/wiki/ontology_(computer_science)

pages. In [ZKK08] combine ontology and an ANN to classify the visited web pages by enhancing the qualification of a concept based on a set of training examples. Such method is claimed to be able to overcome the disadvantages that the relevance score can not optimally reflect the relevance of concepts to the crawling topics proposed by previous work.

In this dissertation, we propose a focused crawling strategy considering the challenges that traditional focused crawlers are trying to solve. To differentiate our work with existing methods, we fully utilize the domain taxonomy for web page classification, prioritizing, and link prediction. Instead of classifying a web page into "related" and "not-related", we assigned to it a concept in our concept hierarchy which is essentially a domain taxonomy. These domain related concepts increase the coherence of the Web pages of a given topic, which plays an important role of bridging different related sites and different domain concepts.

## 2.2 Entity recognition and taxonomy generation

### 2.2.1 Named-entity recognition and relation extraction

Named-entity recognition (NER) is defined as a subtask of information extraction that identify and assign information units in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages [PK01, NS07, WKPU08]. The most straightforward way is using an entity list and part of speech (POS) patterns. A list contains a set of related terms or phrases in a particular domain. The POS patterns allow people to define things like noun phrases, verb phrases, or any POS patterns as regular expression. More complex and effective methods such as regular expression, Conditional Random Field (CRF) model, and Maximum Entropy-based model are well studied by academic community. Using regular expressions, people can define things like address, date ranges, email and have them treated as named entities [RH02, SP01, CC97]. Conditional Random Field model can be trained using labeled named entity types, such as Person, Date, Place. Such model is applied to new lexical documents to label the terms that can be recognized [LMP01, SP03]. Hidden Markov Model is widely used for labeling sequential text. An HMM is essentially a finite state automaton with stochastic state transitions and observations. By modeling the probability of reaching a state given an observation and the previous state. Maximum Entropy-based model can find the best model of the data which satisfies certain constraints and makes the fewest possible assumptions [CN02b, CN02a, MFP00].

A relationship extraction task requires the detection and classification of semantic relationship mentioned within a set of elements typically from text documents. One solution to this problem involves the use of domain ontologies [RTWH99, RKS06]. For instance, the ARCHILES [WLB09] uses only Wikipedia and search engine page count for acquiring relations to construct small-scale ontologies. Many researches on extraction of semantic relationships focus on using syntactic parse trees. [MFRW00] builds generative models for the augmented trees based on semantic information corresponding to entities and relations. [ZAR03] extracts relations by computing kernel functions between parse trees. [CS04] extends this work to estimate kernel functions between augmented dependency trees. Other method,such as [Kam04], builds Maximum Entropy models for extracting relations that combine diverse lexical, syntactic and semantic features.

### 2.2.2 Taxonomy-driven information extraction

Research indicates that even state-of-the-art NER systems are brittle, meaning that NER systems developed for one domain do not typically perform well on other domains [PK01]. [Res11] sets up a probabilistic framework and defines the measurement of semantic similarity in information-theoretic terms in a taxonomy and applies such taxonomy-based measurement to problems of ambiguity in natural language. [YM07, MC06, WHW08] use "ontology-driven information extraction" to define an IE process [CL96] guided by an ontology. [WD10] provides the explicit definition for using ontology to not only extract certain types of information but also present the output. A task for extracting instances and property value with respect to classes and properties of a given ontology is often named as ontology population[VVMD+01, DFKK08, SFMB07]. KIM system proposed in [VFC05, KPT+04] defines a general framework for ontology-supported document retrieval, and integrate full-text search with ontology-based methods. [DK07] proposes a non part-of-speech based method for semantic elements extraction and applies it in health care domain to enhance the ontology and information retrieval performance.

### 2.2.3 Clustering-based taxonomy generation

Traditional hierarchical (divisive or agglomerative) clustering algorithms can be used to construct a hierarchical taxonomy and some heuristic algorithms are applied to optimize the taxonomic structure [CC03a, CC02]. Automatic tools for initial taxonomy construction based on hierarchical clustering of documents are proposed in [LZO03, VD00]. [Bol98] developed a divisive partitioning algorithm PDDP, which iteratively separates a intermedia cluster into two children by testing if its scatter value exceeds a user-defined threshold. COBWEB [CCH02] performs the incremental conceptual clustering to construct class hierarchies. A set of topical summary terms are used for taxonomy construction. These terms are selected by maximizing the joint probability of their topicality and predictiveness [LC03]. DisCover[KLR+04] is developed to maximize the coverage as well as the distinctiveness of the taxonomy incrementally. [SJN06] defines taxonomy learning task as finding the taxonomy that maximizes the probability of certain relationships between terms. Multiple supervised logistic regression models is trained to capture the relations. [YC09] presented a semi-supervised taxonomy induction framework that integrates contextual, co-occurrence, and syntactic dependencies, lexico-syntactic patterns, and other features to learn an ontology metric, calculated in terms of the semantic distance for each pair of terms in a taxonomy. There are also efforts on generating taxonomies by incorporating evidence from multiple classifiers with the knowledge in a preexisting semantic taxonomy, such as WordNet, to optimize the entire structure of the taxonomy [SJN06, YC09]. Some other researches focus on not generating but exploiting existing document taxonomies and enriching them using various classification techniques [AAHM00, AS01, CC03b].

### 2.2.4 Using external source for taxonomy generation

Rule-based approaches use predefined rules or heuristic patterns to extract terms and relations typically based on lexico-syntactic patterns [Hea92]. Such lexico-syntactic patterns can be defined manually [BC99, KRH08] or obtained by bootstrapping techniques [GBM06, PP06], from which certain type of relation can be extracted, such as A is a kind of B. Other rule-based approaches learn a taxonomy by applying heuristics to supporting resources such as Wikipedia [SKW08, PS11], or utilizing computational lexicons such as WordNet [PN09]. WordNet is also widely used as the underlying reference ontology to support the evaluation of semantic similarity [VVR$^+$05]. The category system in Wikipedia can be taken as a conceptual network. [PS11] labels the semantic relations between categories using methods based on connectivity in the network and lexico-syntactic matching to generate a large scale taxonomy. [PN09] presents a knowledge-rich methodology for disambiguating Wikipedia categories with WordNet synsets. A taxonomy automatically generated from the Wikipedia system of categories can be restructured by using external semantic information. Meanwhile, the WordNet is effectively enriched with a large number of instances from Wikipedia.

In this dissertation, we propose a novel framework for generating domain taxonomy using hierarchical clustering with constraints. Domain knowledge is used as the base taxonomy and can be enriched and updated as more meaningful concepts are extracted from crawled documents. Our framework is different from existing work on several aspects. In our approach, information extraction is supported by taxonomy generated from collection of domain related documents. External resources are not necessary to be used. In addition, to generate domain taxonomy, we transform domain knowledge into ordered constraints which are defined especially for hierarchical clustering. Semi-supervised hierarchical clustering with ordered constraints and triple-wise constraints is implemented to integrate domain knowledge.

## 2.3 Personalization and recommendation

### 2.3.1 Content-based

Recommendation purely based on news content is essentially to consider the similarities between the user's profile and the target textual content. Both the user's profile and the news content can be generally represented using vector space model (e.g., TF-IDF) [JMK$^+$00] or topic distributions quantified by language models (e.g., PLSI [Hof99] and LDA [BNJ03]). Many content-based news recommender systems have been proposed in the last decade. For example, News Dude [BP99], is a personal news recommending agent that utilizes TF-IDF combined with the K-Nearest Neighbor algorithm to recommend news items to individual users. Other illustrative content-based methods include YourNews [ABG$^+$07] and Newsjunkie [GDH04], where the former intends to increase the transparency of adapted news delivery by allowing the given user to adapt his/ber profile information, and the latter tries to filter news stories by formal measures of information novelty. Although the content-based approaches is quite straightforward to understand and implement, it is insufficient to construct a single user's profile information by aggregating a bag of words since such model cannot adapt to relative stability of a user's general reading interest and frequent shift of the user's fine-grained preferences.

### 2.3.2 Collaborative filtering

Collaborative filtering systems assume that users with similar rating behaviors in the past usually have similar preferences to new items. Such systems utilize historical user-item-rating combinations to provide recommendation services. Generally, most of systems using collaborative filtering do not use the context or content of items. For personalized news recommendation, news articles are regarded as different items, since there is usually no explicit ratings on news articles. In such case, item ratings are typically binary: a news article be accessed by a reader is assigned the score of 1, otherwise it is assigned a 0 score [DDGR07]. Practically, the usage of users' historical rating behaviors varies based on different mechanisms. Some collaborative filtering systems aggregate the rating behaviors of a group of users "similar" to the given user to predict news ratings [RIS$^+$94, SKKR01, YXT$^+$02], some others model users' behaviors in a probabilistic way [AC09, Hof04, SHB06]. Typically, under two circumstances, collaborative filtering systems can efficiently predict the score of unrated items based on similar users' behaviors: 1. when there is relatively good amount of overlap in historic ratings on the item set; 2. the content universe is almost static [SKR99]. However, there are still challenges in collaborative filtering framework. For example, in many web-based scenarios, the content universe dynamically changes, with content popularity changing over time as well [LCLS10]. Moreover, new relevant items with no historical ratings from users can not receive high predicted scores, which is known as a *cold-start* problem [SPUP02].

### 2.3.3 Hybrid methods

Hybrid solutions to news recommendation by combining two types of methods have also been developed recently. A scalable two-stage personalized news recommendation system was proposed in [LWL$^+$11], which models user preference by building a two-level news hierarchy to enhance the representativeness of each topic cluster. [SHG09] proposed to use the content information in the form of user and item meta data in combination with collaborative filtering information from previous user behavior in order to predict the value of an item for a user. Representative examples include [CP09, Bur05], in which the inability of the collaborative filtering to recommend news items is commonly alleviated by using content-based filtering to solve the cold-start problem. To the best of our knowledge, little research effort has been done to consider the relationship between individual user profile and a profile group simultaneously to achieve a reasonable recommendation result. [MSDR04] utilizes an ontology in a recommender system to solve the cold-start problem and dynamically discover user interests. They first apply the nearest neighbor algorithm to classify documents and user interests, and then identify meaningful communities by ontology-based network analysis. [ZLST04] proposed a taxonomy-based recommender framework for exploiting relationships between father-concepts and child-concepts in the domain taxonomy to address the sparseness issue and diversify the recommendation results to match the specific users interests.

In this dissertation, we propose a novel recommendation framework that captures users' preferences based on not only individual user's reading history, but also the historical consumptions of a group of users with similar reading preferences based on the fact that each user group has its unique preferences to different news topics. Furthermore, the profile of a given user group is not represented using the traditional vector space model, but is characterized by a news hierarchy in which the merged preference between pair of new articles demonstrates their similarity. A consensus hierarchical clustering method is used to combine the news hierarchies associated with the user groups that the user belongs to. The user's interest can be easily captured in a united way. Our proposed framework is beyond content-based methods and collaborative filtering, in which individual user behavior and user group behavior are simultaneously considered for recommendation. Our proposed framework achieves a good balance between the topic coverage and the content diversity of the recommended news list. We model the news selection problem as a budgeted maximum coverage problem, which is more realistic than independently selecting news items.

## 2.4 Clustering methods

### 2.4.1 Hierarchical clustering

Hierarchical clustering algorithms are unsupervised methods to generate tree-like clustering solutions. Many research efforts have been reported on algorithm-level improvements of the hierarchical clustering process and on understanding hierarchical clustering [WXC09][ZK02a]. There are two approaches to hierarchical clustering using bottom-up by grouping small clusters into larger ones or top-down by splitting big clusters into small ones. These are called agglomerative and divisive clusterings, respectively [TSK05a]. Also, other researches define the distance (closeness) between two sub-clusters. There are several basic choices. Single-Link defines the distance between two clusters as the minimum distance between their members; In complete-link, the distance between clusters is the maximum distance between their members; The average-link defines the distance between two clusters as the mean of pairwise distances between members from two clusters; The Ward's method says that the distance between two clusters is how much the sum of squares will increase when they get merged [WJ63, Mur83, TSK05a]. The problem of fitting a tree metric to the (dis-)similarity data on pairs of objects from a given set has been studied quite extensively [AC05a]. Ultra-metric is a special kind of tree metric where all elements of the input dataset are leaves in the underlying tree and all leaves are at the same distance from the root. Ultra-metric naturally corresponds to a hierarchy of clusterings of the data [ABF$^{+}$99][AC05a]. Given a dissimilarity $D$ on pairs of objects, the problem of finding the best ultra-metric $d_u$ such that $||D - d_u||_p$ is minimized is NP-hard for $L_1$ and $L_2$ norms (e.g., when $p = 1$ and $p = 2$) [ABF$^+$99].

### 2.4.2 Ensemble clustering

The problem of ensemble clustering is to find a combined clustering result based on multiple clusterings of a given dataset. There are many ways to obtain multiple clusterings such as applying different clustering algorithms; using re-sampling to get subsamples of the dataset, utilizing feature selection methods to obtain different feature spaces, and exploiting the randomness of the clustering algorithm. Many approaches have been developed to solve ensemble clustering problems over the last few years [AF09][FB04][GMT05][LD08a][MTMG03][TJP05]. However, most of these techniques are designed for partitional clustering methods. The problem of ensemble hierarchical clustering using dendrogram descriptors has been studied in [MRA08]. The key difference is that we present a coherent algorithm to learn the closest ultra-metric solution (matrix $B$ in Problem 1 near Eq.(4.3)) while the approach in [MRA08] requires many parameters that are selected in an ad-hoc manner.

### 2.4.3 Consensus tree

Consensus tree has been widely studied in bioinformatics when comparing the evolution of species to reach a consensus or agreement [Ada86b][III72]. Most techniques for finding consensus tree are based on agreement subtrees (e.g., sub structures that are common to all the trees) [FPT95][Wil94]. It is very difficult for these consensus tree techniques to preserve structural information while including all the existing leaves from the input trees [Swo91]. In supervised classification, different decision trees can be combined using bagging [BB96], boosting [SS99], stacking [Wol92], or random forests [BB01]. Unlike our ensemble hierarchical clustering, these ensemble methods are designed for supervised classification. In addition, most of the decision tree ensembles do not generate a final tree and just combine the output predictions of base trees.

### 2.4.4 Semi-supervised clustering

Many researchers have explored the use of instance-level background information, such as pairwise must-link and cannot-link constraints for learning a distance/dissimilarity measure, or modifying the objective criterion, or improving the optimization procedures [BhHSW05, BBM04, KKM02, Wag02, XNJR02, Zhu05]. Other types of knowledge hints (such as size of the clusters, partial labels of the data points, and user-provided external prototypes/representatives) have also been used for clustering [Ped04]. However, most of these works are designed for partitional clustering and few research efforts have been reported on semi-supervised hierarchical clustering methods. We note the very recent works of Zhao et al. [ZQ10] and Bade et al. [BN06] which perform hierarchical clustering with order constraints and partially known hierarchy. There is few previous work particularly focusing on integrating constraints in hierarchial clustering. In detail, new type of constraint [BN06] is used in agglomerative hierarchical clustering. [BN06] also utilizes a metric-based learning framework to adapt the weight associated with each feature by mapping the given constraints to a distance measure. The objective is to maximize the relations specified by each constraint. Simple gradient decent framework can be performed to obtain the feature weights. Heuristic method is also proposed by verifying possible violations which can prevent certain pair of clusters from merging together in each merging step [ZQ10]. Such method can stop at some step when constraints block any possible further merging or force to merge by setting some constraints to be invalid.

In this dissertation, we propose a framework for ensemble hierarchical clustering based on descriptor matrices to preserve the common structures from the input hierarchical clusterings and also generate a full consensus tree. Our framework is different from existing work on several aspects. In our approach, there are no parameters. In addition, we propose a hierarchical ensemble clustering framework which can naturally combine both partitional clustering and hierarchical clustering results and systematically studied the problems related to dendrogram description, selection, and combination. For semi-supervised hierarchical clustering framework, our triple-wise constraints are conceptually special cases of the order constraints. However, different from their works, our proposed semi-supervised hierarchical clustering framework is based on ultra-metric dendrogram distance. Such unified framework integrates both ultra-metric fitting and triple-wise relative constraints and seeks an approximate dissimilarity metric (ultra-metric) which represents a tuned dendrogram that satisfies the given constraints.

CHAPTER 3

# Hierarchical ensemble clustering

Data clustering arises in many disciplines and has a wide range of applications. The general goal of data clustering is to group a finite set of points in a multi-dimensional space into clusters such that points in the same cluster are similar to each other while points in different clusters are dissimilar. The clustering problem has been extensively studied in the data mining, database, and machine learning communities and many different approaches have been developed from various perspectives with various focuses. Based on the way the clusters are generated, these clustering methods can be roughly divided into two categories: partitional clustering and hierarchical clustering [TSK05a]. Generally, **partitional clustering** decomposes the dataset into a number of disjoint clusters which typically represent a local optimum of some predefined objective function. **Hierarchical clustering** groups the data points into a hierarchical tree structure using bottom-up or top-down approaches.

Clustering is an inherently difficult problem. Different clustering algorithms and even multiple trials of the same algorithm may produce different results due to random initializations and stochastic learning methods. Recently, as a way for overcoming the resulting instability and improving clustering performance, ensemble clustering has emerged as an important elaboration of the classical clustering problem. Ensemble clustering refers to the situation in which a number of different (input) clusterings have been obtained for a particular dataset and it is desired to find a single (consensus) clustering which is a better fit in some sense than the existing clusterings [SG03]. Many approaches have been developed to solve ensemble clustering problems over the last few years [AF09][FB04][LOM04][GMT05][LD08a][MTMG03][TJP05][LDHN11].

However, most of these existing ensemble techniques are designed for partitional clustering methods. Few research efforts have been reported for ensemble hierarchical clustering methods. Different from partitional clustering where the clustering results are ”flat” and can thus be easily represented using vectors, clustering indicators, or connectivity matrices [LD08a][SG03], the results of hierarchical clustering are more complex and typically represented as dendrograms or trees. In addition, unlike partitional clustering methods which generally have some predefined objective functions, hierarchical clustering methods have no global objective functions. These properties have made the ensemble hierarchical clustering problem more challenging.

In this chapter, we demonstrate a **hierarchical ensemble clustering** (HEC) framework. In this framework, the input could be both partitional clusterings and hierarchical clusterings. The output is a **consensus** hierarchical clustering. We discuss three cases below.

(1) The input data consists of partitional clusterings. In this case, we first construct the **aggregate consensus distance** from these partitional clusterings. We then generate a consensus clustering using the consensus distance. If we stop here, that would be the usual ensemble clustering. In HEC, we can further build a **structure hierarchy** on top of the consensus clustering using the consensus distance.

First, a structure hierarchy on top of a clustering solution is useful to organize and understand the discovered knowledge (topic or pattern). Second, the cluster structure hierarchy resolves a problem in the usual ensemble clustering when the input partitional clusterings have different number of clusters. In this case, $K$, the number of clusters in the final clustering solution is not uniquely determined (Much research has been done on finding the most appropriate number of clusters in a dataset [FR98][SJ03][TWH01]). Different frameworks have also been proposed to deal with ensemble clusterings [LD08a][SG03][WLDJ11]. In ensemble clustering, we consider input partitional clusterings as meaningful results, including the **number of clusters** in each input partitional clustering. Therefore, if the number of clusters of input partitional clusterings has a range $[K_1, K_2]$, then $K$ of the final ensemble clustering should be $K \in [K_1, K_2]$.

From this analysis, in HEC framework, we can set $K = K_2$ for the bottom clusterings (leaves) of the structure hierarchy. In this way, the ”true” number of clusters is guaranteed to be inside the cluster structure hierarchy.

(2) The input data consists of hierarchical clusterings, i.e., a set of dendrograms. In this case, we first construct the **aggregate dendrogram distance** between objects. From this distance, we then generate a hierarchical clustering as the final solution.

(3) The input data consists of both partitional clusterings and hierarchical clusterings. In this case, we construct the consensus distance from the partitional clusterings and the dendrogram distance from hierarchical clusterings. We combine these two distances into a single distance, and then generate a hierarchical clustering as the final solution. Figure 3.1 illustrates this case. The dataset is shown in Fig.1(A) and their distances are shown in Fig.1(B). K-means clustering is performed with different numbers of clusters in Fig.1(C) and leads to a consensus distance matrix in Fig.1(E). A hierarchical clustering is done in Fig.1(D) and leads to a dendrogram distance matrix in Fig.1(F). The consensus distance matrix of Fig.1(E) and the dendrogram distance matrix in Fig.1(F) are combined in Fig.1(G) and the final hierarchical clustering is generated in Fig.1(H).

Our preliminary work was presented at the International Conference on Data Mining (ICDM) 2010 [ZLD10b] in which we focused on the ensembles of hierarchical clustering and the related computational algorithms. In this journal submission, we extend our previous work by systematically studying the following three important problems:

1. ***Dendrogram Description***: How can we represent the dendrograms so that different hierarchical clustering solutions can be compared and combined?

2. ***Dendrogram Combination***: How can we aggregate different dendrograms and generate final hierarchical solution?

3. ***Dendrogram Selection***: Given a large collection of input hierarchical clusterings, how can we select a subset from the input collection to effectively build an ensem-

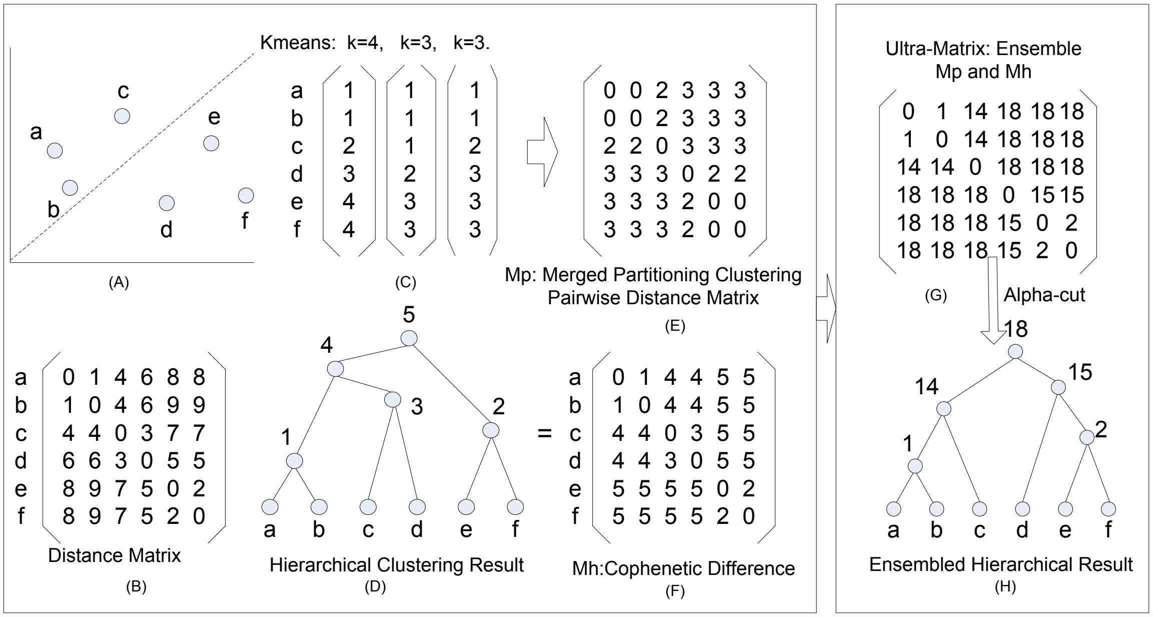


Figure 3.1: An illustrative example of hierarchical ensemble clustering with both partitional and hierarchical clusterings as input. The dataset is shown in (A) and their distances are shown in (B). K-means clustering are performed in (C) and lead to a consensus distance matrix in (E). A hierarchical clustering is done in (D) and leads to a dendrogram distance matrix in (F). The consensus distance matrix of (E) and the dendrogram distance matrix in (F) are combined in (G) and the final hierarchical clustering are generated in (H).

ble solution that performs as well as or even better than using all available clusterings [FL08]?

In particular we investigate various descriptor matrices for representing dendrograms and propose a novel method for deriving a final hierarchical clustering by fitting an ultra-metric from the aggregated descriptor matrix. We also study the problem of combining both hierarchical and partitional clustering results not only focuses on the combination hierarchical clusterings. We develop two approaches for dendrogram selection based on tree distances. Our experimental evaluation also provides a systematic empirical study on the ensemble hierarchical clustering problem. Experimental results have demonstrated the effectiveness of our proposed approaches.

The rest of the chapter is organized as follows: Section 3.1 discusses the related work; Section 3.3 gives the comprehensive discussion about the HEC framework, in which algorithm is provided in Section 3.3.1; Section 3.3.2 describes the distance matrix used for representing partitional clustering results; Section 3.3.3 proposes a novel method for deriving final hierarchical clustering by fitting an ultra-metric from the aggregated distance matrix which is combined with multiple input hierarchical clusterings; Section 3.3.4 presents our approaches for dendrogram selection, i.e., selecting a subset of hierarchical clusterings from the input collection; Section 3.4 shows experimental evaluations and result analysis; and Finally, Section 3.5 concludes the chapter and discusses future work.

## 3.1 Related work

**Hierarchical Clustering:** Hierarchical clustering algorithms are unsupervised methods to generate tree-like clustering solutions. They group the data points into a hierarchical tree structure using bottom-up (agglomerative) or top-down (divisive) approaches [TSK05a]. Many research efforts have been reported on algorithm-level improvements of the hierarchical clustering process and on understanding hierarchical clustering [WXC09][ZK02a].

**Ensemble Clustering:** The problem of ensemble clustering is to find a combined clustering result based on multiple clusterings of a given dataset. There are many ways to obtain multiple clusterings such as applying different clustering algorithms; using resampling to get subsamples of the dataset, utilizing feature selection methods to obtain different feature spaces, and exploiting the randomness of the clustering algorithm. Many approaches have been developed to solve ensemble clustering problems over the last few years [AF09][FB04][GMT05][LD08a][MTMG03][TJP05]. However, most of these techniques are designed for partitional clustering methods. The problem of ensemble hierarchical clustering using dendrogram descriptors has been studied in [MRA08]. The key difference is that we present a coherent algorithm to learn the closest ultra-metric solution (matrix $B$ in Problem 1 near Eq.(4.3)) while the approach in [MRA08] requires many parameters that are selected in an ad-hoc manner. In our approach, there are no parameters.

In addition, we propose a hierarchical ensemble clustering framework which can naturally combine both partitional clustering and hierarchical clustering results and systematically studied the problems related to dendrogram description, selection, and combination.

**Consensus Tree:** Consensus tree has been widely studied in bioinformatics when comparing the evolution of species to reach a consensus or agreement [Ada86b][III72]. Most techniques for finding consensus tree are based on agreement subtrees (e.g., sub structures that are common to all the trees) [FPT95][Wil94]. It is very difficult for these consensus tree techniques to preserve structural information while including all the existing leaves from the input trees [Swo91]. In this chapter, we propose a framework for ensemble hierarchical clustering based on descriptor matrices to preserve the common structures from the input hierarchical clusterings and also generate a full consensus tree.

**Metric Fitting:** The problem of fitting a tree metric to the (dis-)similarity data on pairs of objects from a given set has been studied quite extensively [AC05a]. Ultra-metric is a special kind of tree metric where all elements of the input dataset are leaves in the underlying tree and all leaves are at the same distance from the root. Ultra-metric naturally corresponds to a hierarchy of clusterings of the data [ABF$^+$99][AC05a]. Given a dissimilarity $D$ on pairs of objects, the problem of finding the best ultra-metric $d_u$ such that $||D - d_u||_p$ is minimized is NP-hard for $L_1$ and $L_2$ norms (e.g., when $p = 1$ and $p = 2$) [ABF$^+$99]. In this chapter, we propose a new method for fitting an ultra-metric to the aggregated descriptor matrix.

**Ensemble Decision Trees:** In supervised classification, different decision trees can be combined using bagging [BB96], boosting [SS99], stacking [Wol92], or random forests [BB01]. Unlike our ensemble hierarchical clustering, these ensemble methods are designed for supervised classification. In addition, most of the decision tree ensembles do not generate a final tree and just combine the output predictions of base trees. In this chapter, we generate the final ensemble result as a complete hierarchical clustering result.

**Cluster Ensemble Selection:** The problem of selecting a subset of input clusterings to form a smaller but better performing cluster ensemble than using all available solutions has been studied recently for partitional clustering [AF09][FL08]. In this chapter, we develop cluster ensemble selection methods for hierarchical clustering based on tree distances.

## 3.2 Ultra-metric, dendrogram, and hierarchical clustering

In this section, we explicit establish the relationship between ultra-metric, dendrogram, and hierarchical clustering. By theoretically building the equivalence between a dendrogram and a hierarchical clustering results, the ultra-metric representation can be extensively used as the basis for hierarchical ensemble clustering (discussed in Chapter3) and semi-supervised hierarchical clustering (discussed in Chapter4).

### 3.2.1 Ultra-metric distance

**Definition 1.** *A distance matrix $D = (d_{ij})$ is a* **metric**, *if it has the following properties: nonnegativity*

$$d_{ij} \geq 0,$$

*if $d_{ij} = d(x_i, x_j)$ = 0, then $x_i = x_j$, symmetry*

$$d_{ij} = d_{ji},$$

*and the* **triangle inequality**

$$d_{ij} \geq 0, \;\; d_{ij} \leq d_{ik} + d_{kj}, \; i \neq k \neq j.$$

It should noted that although nonnegativity and symmetry hold for many distance measures in data mining, the triangle inequality does not always hold.

A more restricted version of the triangle inequality is called **ultra-metric inequality**:

$$d_{ij} \leq max(d_{ik}, d_{jk}) \tag{3.1}$$

for all triplets of points $i, j, k$. This is equivalent to saying that for any distinct triple $i, j, k$, the largest two distances among $d_{ij}, d_{ik}, d_{jk}$ are equal and not less than the third.

**Definition 2.** *A distance measure is an ultra-metric if it satisfies the ultra-metric inequality, nonnegativity and symmetry.*

A distance measure automatically satisfies the triangle inequality if it satisfies the ultra-metric inequality. Thus an ultra-metric distance is also a metric distance; But the converse is not true.

### 3.2.2 Ultra-metric and dendrogram reconstruction

A dendrogram is a rooted tree that represents the result of a hierarchical clustering. On the root, leaves represent data objects and internal nodes represent clusters at various levels. The structural information is kept by pairwise cophenetic proximity which measures the level at which two data objects are first merged in the same cluster [JD88].

Given a dendrogram, our task is to assign distances between leaf nodes. This problem has been studied in literature [MRA08][Pod00a]. In Section 3.2.3, we will describe several commonly used dendrogram distances (also called descriptors). We note that each of these dendrogram distance is in fact an ultra-metric distance. This is important because given an ultra-metric distance matrix $D = (d_{ij})$, we can reconstruct the original tree.

In single-link and complete-link hierarchical clustering, a dendrogram is generated by repeatedly picking the closest pair of clusters from the distance matrix, merging these two clusters into one, and updating the distance matrix. Various schemes differ in how the distance between a newly formed cluster and the other clusters is defined. Let $d$ be the final generated distance. It can be easily shown that $d$ is an ultra-metric. To see why, consider three objects $i, j, k$. Without loss of generality, assume $i$ and $j$ merge first. Then we have $d(i, j) \leq d(i, k) = d(j, k)$. More details can be found in [JD88].

Ultra-metric distance plays a critical role in our HEC frame because of the unique reconstruction property. Suppose, we are given a dendrogram $G$ and we construct a dendrogram distance $D$ using a particular method $M$.

The following proposition holds:

**Proposition 1.** *From a given ultra-metric distance $D$, a unique dendrogram $G$ can be constructed, in the sense that if we construct the distance from $G$, we recover $D$ exactly.*

Note that Proposition 1 does not exclude the possibility that two different ultra-metric distances $D_1, D_2$ lead to the same reconstructed dendrogram $G$. In fact, there are several ways to model the pairwise distance matrix between instances in a dendrogram (see Section 3.2.3). Using different dendrogram distance measures leads to different ultra-metric distances. But the reconstructed dendrograms from these different ultra-metric distances are identical.

### 3.2.3 Dendrogram distances

A dendrogram is usually used to represent the hierarchical clustering results for cluster analysis and it is easy to interpret. The ultra-metric information contained in the pairwise distance matrix can be clearly mapped to dendrogram structural information. So, for each dendrogram, there is a ultra-metric matrix which uniquely characterizes it and can be used to recover this dendrogram [MRA08].

For instance, a dendrogram generated by the single-link hierarchical clustering algorithm can be viewed as a weighted dendrogram, in which every internal node is associated with a continuous variable indicating the merge distance within all leaves covered by the internal node. The merge distance is called the *height*. If we replace the height of an internal node with its rank order (i.e., the *level*) which is maintained globally with respect to the whole dendrogram, then a weighted dendrogram becomes a fully ranked dendrogram [Pod00a]. A dendrogram descriptor can be regarded as a distance function describing the relative position of a given pair of leaves in the dendrogram and is used to characterize a corresponding dendrogram.

We now introduce the dendrogram descriptors that will be used in our investigation. The first three dendrogram descriptors are based on a fully ranked dendrogram and use the level information [MRA08][Pod00a]. The other descriptors do not directly consider level information.

- ***Cophenetic Difference*** **(CD)** : the lowest height(i.e., merge distance) of internal nodes in the dendrogram where two specified leaves are joined together. For example, CD between nodes v and x in Figure 3.2 is 30.
- ***Maximum Edge Distance*** **(MED)**: the depth of node in a bottom-up view. All leaf nodes are assigned as depth 0, the depth of any internal nodes is generated in a bottom-up manner. Suppose C3 is the internal node at which C1 and C2 firstly merge, then Depth$(C3) = \max($Depth$(C1),$Depth$(C2)) + 1$. For example, MED of nodes v and x in Figure 3.2 is 2. Nodes v and x firstly merged at internal node c, so Depth$(c) = \max($Depth$(a),$Depth$(x)) + 1 = \max(1, 0) + 1 = 2$, since Depth$(a) = \max($Depth$(v),$Depth$(w)) + 1 = 1$.
- ***Partition Membership Divergence*** **(PMD)**: By utilizing the property that a hierarchical clustering result implies a sequence of nested partitions obtained by cutting the dendrogram at every internal node, the PMD is defined as the number of partitions of the hierarchy in which two specified leaves are not in the same cluster.
- ***Cluster Membership Divergence*** **(CMD)**: the size of the smallest cluster in the hierarchy which contains two specified leaves.
- ***Sub-dendrogram Membership Divergence*** **(SMD)**: the number of sub-dendrograms in which two specified leaves are not included together.

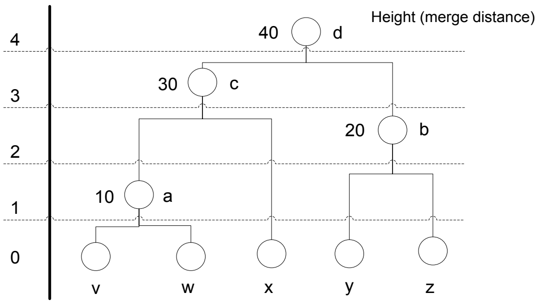


Figure 3.2: A dendrogram example.

## 3.3 Hierarchical ensemble clustering

### 3.3.1 Algorithm

With above discussions on ultra-metric distances and dendrogram, we outline the algorithmic strategy of our hierarchical ensemble clustering. Our central strategy is listed below:

1. Use a dendrogram distance measure to generate an ultra-metric dendrogram distance for each input dendrogram (see Chapter 3.2). The consensus distance matrix for partitional clustering results are discussed in Section 3.3.2.
2. Aggregate the ultra-metric dendrogram distances as well as the consensus distance for partitional clusterings. (see Chapter 3.2)
3. Finding the closest ultra-metric distance from the aggregated distance. (see Section 3.2)
4. Construct the final hierarchical clustering. (see Section 3.2)

### 3.3.2 Distance matrices for partitional clustering results

Our framework can be naturally extended to ensemble both partitional and hierarchical clustering results by representing the partitional clustering results with a distance matrix.

Formally let $X = \{x_1, x_2, \cdots, x_n\}$ be a set of $n$ data points. Given a partitional clustering $C$, consisting of a set of clusters $C = \{C_1, C_2, \cdots, C_k\}$ where $k$ is the number of clusters and $X = \bigcup_{\ell=1}^{k} C_\ell$, we can define the following associated distance matrix $D(C)$ whose $ij$-th entry is defined as

$$d_{ij} = \begin{cases} 0 & (i,j) \in C_\ell \\ 1 & \text{Otherwise} \end{cases} \tag{3.2}$$

where $(i,j) \in C_\ell$ means that $i$-th data point and $j$-th data point are in the same cluster $C_\ell$. In other words, if $i$-th data point and $j$-th data point are in the same cluster, then their distance between them is $0$.

Given a set of $s$ clusterings (or partitions) $\mathcal{P} = \{P^1, P^2, \cdots, P^s\}$ of the data points in $X$, the associated consensus distance matrix $D$ can be represented as

$$D(\mathcal{P}) = \frac{1}{s} \sum_{i=1}^{s} D(P^i). \tag{3.3}$$

In other words, the $ij$-th entry of $D$ indicates the average number of times that the $i$-th data point and the $j$-th data point are not in the same cluster.

Equation 3.3 defines a way to aggregate multiple partitional clustering results into one consensus distance matrix. Also there are many different ways to define the consensus function such as co-associations between data points or based on pairwise agreements between partitions. Some of the criteria are based on the similarity between data points and some of them are based on the estimates of similarity between partitions. The relationship between consensus matrix and other measures is discussed and summarized in [LOM10].

Note that the distance matrix can be combined with the dendrogram descriptors to form the aggregated distance matrix for dendrogram combination. A weight can be assigned to the distance matrix to ensure that it is at the same scale of the dendrogram descriptors.

### 3.3.3 Dendrogram combination

Suppose we have computed consensus distance $D(\mathcal{P})$ from the input partitional clusterings and aggregated dendrogram distances $D(H)$ from the input hierarchical clusterings.

Given any similarity we can do any kind of hierarchical clustering. However, there are many different choices here: single-link, complete-link, average-link, and many other choices. Which one to choose? Our logic is that since the input individual descriptors are ultra-metric, and the consensus matrix is not ultra-metric, the most natural approach is to a find an ultra-metric which is as close to the consensus matrix as possible. Once this ultra-metric is learned, the final hierarchical clustering is uniquely determined. There are other choices here. The entire approach is uniquely deterministic.

Our tasks now are

1. Find an ultra-metric distance $T$ which is the closest to $D = \frac{1}{2} \times (D(\mathcal{P}) + D(H))$.
2. Construct the final hierarchical clustering based on $T$.

For (2) after the ultra-metric $T$ is obtained, we obtain the final hierarchical clustering by performing the alpha-cut [MNB04]. In the rest of this section we will concentrate on (1), i.e., how to compute $T$.

First we note that the aggregated distance $D$ will not be ultra-metric, even if each individual dendrogram distance is an ultra-metric. We compute the ultra-metric distance $T$ which is closest to $D$, instead of using $D$ directly due to the following two reasons. The first reason is for the unique reconstruction of the eventual dendrogram, the final hierarchical clustering, as discussed in Section 3.2.1 in Chapter3.2.

The second reason is an interesting property of our way of constructing $T$, the close approximation of $D$. We use a transitive dissimilarity to construct $T$, which has the tendency that the solution for $T$ attracts nearby data objects into a closer proximity and therefore enhances the cluster separation (thus improves the clustering quality) (See the example in

Section 4.2.2 in Chapter3.2). In the following, we first describe the algorithm to construct $T$ and then demonstrate the clustering separation property.

### 3.3.4 Dendrogram selection

Selecting a subset of input clusterings to form a smaller ensemble has been shown to achieve better performance than using all available solutions for partitional clustering methods[AF09][FL08]. The selection is based on the quality and diversity of each individual clustering solution. For partitional clustering, since the clustering solutions are naturally represented using vectors or matrices [LD08a][SG03], the diversity and quality of the clustering solutions can be easily computed. To perform dendrogram selection, the question is how to compute the diversity and quality of different hierarchical clustering solutions.

We propose two approaches to perform dendrogram selection based on tree distances. First, we introduce the tree distances to measure the similarities/differences between different hierarchies. There are two distances which are frequently used in literature to compute the distance between two evolutionary dendrograms: Branch Score Distance (BSD) of Kuhner and Felsenstein (1994) [KF94], and Symmetric Difference (SD) of Robinson and Foulds (1981) [RF81]. Both distances are computed by considering all possible branches that could exist on the two trees. Note that each branch makes a partition of the given dataset into two groups – the ones connected to one end of the branch and the ones connected to the other. BSD uses branch lengths while SD does not use branch lengths and only uses the tree topologies. For BSD, each partition on a dendrogram has an associated branch length. BSD is then computed by taking the sum of squared differences between the branch lengths of two dendrograms. SD is calculated as the number of partitions of two dendrograms among which such partition exists in exact only one dendrogram.

The goal of dendrogram selection is to select a diverse subset of dendrograms where each of them has good quality. We propose two approaches for dendrogram selection using tree distances. The first approach is to use a modified K-medoids algorithm (with the tree distances) to cluster those dendrograms and then select the medoids for each cluster. The medoid of a cluster is a representative object whose average similarity to all the other objects in the cluster is maximized, thus the medoid dendrogram can be considered to best capture the information contained in the cluster and has good quality. On the other hand, selecting medoids from different clusters achieves diversity.

The second approach is based on the farthest-point heuristic [Gon85]. The approach starts with the medoid of all the input clustering solutions. Then pick a dendrogram that is as far from the selected dendrogram as possible. In general, the approach picks a dendrogram to maximize the distances to the nearest of all dendrograms picked so far. Specifically, if $t_1, t_2, \cdots, t_{i-1}$ denote the selected dendrograms so far, then we pick $t_i$ to maximize

$$\mathit{min}\{\mathit{dist}(s_i, s_1), \mathit{dist}(s_i, s_2), \cdots, \mathit{dist}(s_i, s_{i-1})\}. \tag{3.4}$$

The approach stops after the required number of dendrogram has been selected.

## 3.4 Experiments

### 3.4.1 Experiment setup

To evaluate our ensemble framework, we focus on how well the ensemble hierarchical solution reflects the characteristics of the original dataset. **Co-Phenetic Correlation Coefficient (CPCC)**, which aims to evaluate how faithfully a dendrogram preserves the pairwise distances between the original data samples [RF68][SR62], is used as our performance measure. CPCC can be calculated as

$$c = \frac{\sum_{i<j}(d(i,j) - d)(h(i,j) - h)}{\sqrt{[\sum_{i<j}(d(i,j) - d)^2][\sum_{i<j}(h(i,j) - h)^2]}}, \tag{3.5}$$

where $d(i,j)$ is the distance between the $i$-th and $j$-th data instances, h(i,j) is the height of lowest common ancestor of the $i$-th and $j$-th data instances in ensemble dendrogram, $d$ is the averages of $d(i,j)$ over all pairs, and $h$ is the average of $h(i,j)$. Generally, the higher the CPCC value, the better the clustering performance.

We use six datasets from different domains to conduct the experiments: four datasets (Wine, Parkinson Disease, Libras Movement and Madelon) from UCI Machine Learning Repository[1], and two benchmark text datasets for document clustering (WebACE and Reuters datasets) [LD08a]. The datasets and their characteristics are summarized in Table 3.1. The two text datasets are represented using the vector space model, and they are also pre-processed by removing the stop words and unnecessary tags and headers. All experiments are conducted under the environment of Windows XP operating system plus Intel P4 1.83 GHz CPU and 4 GB of RAM.

| Name | # samples | # attributes | # classes |
|---|---|---|---|
| Wine | 178 | 13 | 3 |
| Parkinson Disease | 195 | 22 | 2 |
| Libras Movement | 360 | 90 | 15 |
| Madelon | 2000 | 500 | 2 |
| WebACE | 2340 | 1000 | 12 |
| Reuters | 2787 | 1000 | 9 |

Table 3.1: Dataset descriptions

### 3.4.2 Ensemble hierarchical clusterings

For experiments on ensemble hierarchical clusterings, 20 input dendrograms are generated for each dataset by using different hierarchical clustering methods on different attribute subsets. In particular, they are generated as follows: 1) 10 different attribute subsets are randomly constructed first, each of which contains 90% of all the attributes; and 2) single-link (SL) and complete-link (CL) algorithms are applied to different attribute subsets.

[1] The datasets can be downloaded from http://archive.ics.uci.edu/ml/.

| Descriptor | ultra | single-link |
|---|---|---|
| CD | **0.392** | 0.381 |
| CMD | **0.443** | 0.273 |
| MED | **0.292** | 0.288 |
| PMD | **0.267** | 0.232 |
| SMD | **0.299** | 0.290 |

Table 3.2: Experimental results on Wine dataset using all input dendrograms. The maximum CPCC value for any input dendrogram is $0.407$ and the average value of all input dendrograms is $0.282$.

| Descriptor | ultra | single-link |
|---|---|---|
| CD | **0.577** | 0.554 |
| CMD | **0.431** | 0.419 |
| MED | **0.485** | 0.428 |
| PMD | 0.402 | **0.417** |
| SMD | 0.448 | **0.491** |

Table 3.3: Experimental results on Parkinson Disease dataset using all input dendrograms. The maximum CPCC value for any input dendrogram is $0.381$ and the average value of all input dendrograms is $0.201$.

| Descriptor | ultra | single-link |
|---|---|---|
| CD | **0.423** | 0.419 |
| CMD | **0.411** | 0.389 |
| MED | 0.36 | **0.363** |
| PMD | **0.279** | 0.266 |
| SMD | **0.45** | 0.438 |

Table 3.4: Experimental results on Libra Movement dataset using all input dendrograms. The maximum CPCC value for any input dendrogram is $0.334$ and the average value of all input dendrograms is $0.25$.

| Descriptor | ultra | single-link |
|---|---|---|
| CD | **0.063** | 0.042 |
| CMD | 0.068 | **0.074** |
| MED | **0.05** | 0.039 |
| PMD | **0.088** | 0.082 |
| SMD | 0.04 | **0.047** |

Table 3.5: Experimental results on Madelon dataset using all input dendrograms. The maximum CPCC value for any input dendrogram is $0.057$ and the average value of all input dendrograms is $0.014$.

| Descriptor | ultra | complete-link |
|---|---|---|
| CD | **0.465** | 0.4637 |
| CMD | **0.4971** | 0.4963 |
| MED | **0.4787** | 0.4699 |
| PMD | 0.4831 | **0.4896** |
| SMD | **0.5056** | 0.4781 |

Table 3.6: Experimental results on WebACE dataset using all input dendrograms. The maximum CPCC value for any input dendrogram is $0.47$ and the average value of all input dendrograms is $0.428$.

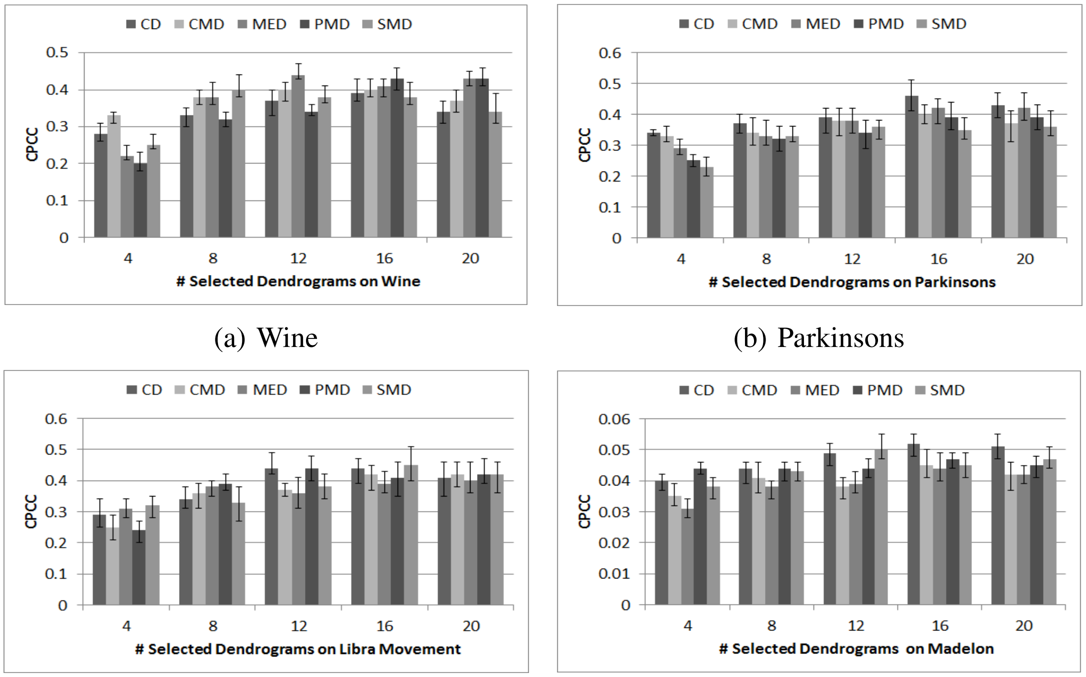
| Descriptor | ultra | complete-link |
|---|---|---|
| CD | **0.7349** | 0.7312 |
| CMD | **0.7822** | 0.7435 |
| MED | **0.7415** | 0.7176 |
| PMD | **0.7624** | 0.6955 |
| SMD | 0.6475 | **0.6479** |

Table 3.7: Experimental results on Reuters dataset using all input dendrograms. The maximum CPCC value for any input dendrogram is $0.7583$ and the average value of all input dendrograms is $0.633$.

We evaluate our proposed method for generating the final hierarchical solution by fitting an ultra-meric using all five dendrogram descriptors (i.e., CD, CMD, MED, PMD, SMD). We also compare our proposed method (denoted as *ultra* in the experimental results) with the method that directly performs single-link and complete-link hierarchical clusterings on the aggregated descriptor matrices (denoted as *single-link/complete-link* or *SL/CL*[2]).

[2] In our work, we apply single-link (*SL*) on the aggregated descriptor matrices for 4 UCI datasets and apply complete-link (*CL*) on the aggregated descriptor matrices for 2 text datasets.

### Results using all input dendrograms



(a) Wine

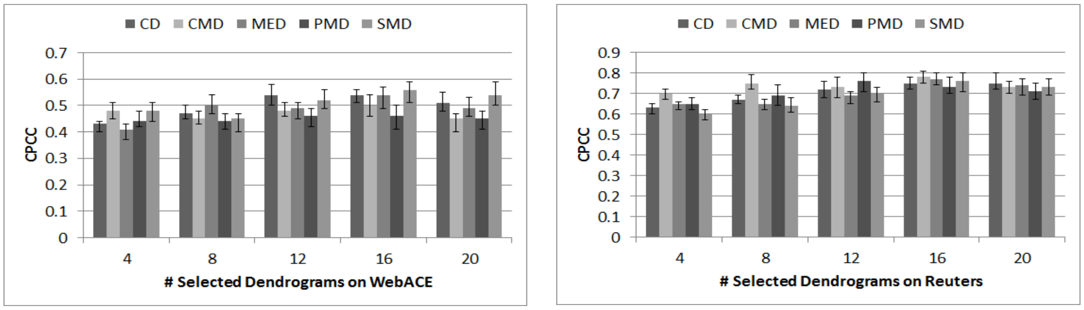
(b) Parkinsons

(c) Libra Movement

(d) Madelon

Figure 3.3: The performance variation on different number of selected dendrograms over 20 trails.

Tables 3.2 to 3.7 present the experimental results on **six** datasets using all input dendrograms, respectively. Note that, unlike ensemble clustering for partitional clustering results, for hierarchical clustering ensemble, once the set of individual hierarchical clustering results is fixed, then the result of ensemble is also determined. From the experimental results, we observe that: 1) Our proposed method *ultra* generally outperforms hierarchical clustering (*single-link* or *complete-link*) across various descriptors on most counts, especially on large datasets (e.g., WebACE and Reuters); and 2) the ensemble solution using all input dendrograms may be worse than the best individual dendrogram, thus demonstrating the need for ensemble selection.



(a) WebACE (b) Reuters

Figure 3.4: The performance variation on different number of selected dendrograms over 20 trails.

### Results on different input dendrograms

In order to provide more insights on our proposed method, we also conduct experiments with different sets of input dendrograms. Figure 3.3 and Figure 3.4 show the experimental results on the four UCI datasets (Wine, Parkinsons, Libra Movement and Madelon) and the two text datasets (WebACE and Reuters) respectively with different sets of input dendrograms. In particular, for a given size, we randomly select a set of input dendrograms, and then perform the experiments. The reported results are averaged over 20 different runs.

Based on our observation, the best performance is often obtained when the number of input dendrograms is around 16. Although this experiment is conducted by randomly selecting input dendrograms, it still demonstrates that using a subset of input dendrograms (rather than using all dendrograms) may improve the ensemble performance. The issue of using dendrogram selection strategies to form the candidate subset are discussed in Section 3.4.2 and Section 3.4.2, respectively.

| Desc. | Sel | Dis | max | ave | ultra | SL |
|---|---|---|---|---|---|---|
| CD | F | B | 0.292 | 0.245 | **0.352** | 0.331 |
|  | K | B | 0.306 | 0.251 | **0.373** | 0.357 |
|  | F | S | 0.281 | 0.229 | **0.329** | 0.292 |
|  | K | S | 0.299 | 0.238 | 0.336 | **0.344** |
| CMD | F | B | 0.292 | 0.245 | **0.387** | 0.378 |
|  | K | B | 0.306 | 0.251 | **0.373** | 0.365 |
|  | F | S | 0.281 | 0.229 | **0.361** | 0.329 |
|  | K | S | 0.299 | 0.238 | **0.35** | 0.337 |
| MED | F | B | 0.292 | 0.245 | **0.369** | 0.348 |
|  | K | B | 0.306 | 0.251 | **0.355** | 0.316 |
|  | F | S | 0.281 | 0.229 | **0.339** | 0.318 |
|  | K | S | 0.299 | 0.238 | **0.357** | 0.323 |
| PMD | F | B | 0.292 | 0.245 | **0.296** | 0.284 |
|  | K | B | 0.306 | 0.251 | 0.315 | **0.331** |
|  | F | S | 0.281 | 0.229 | **0.316** | 0.302 |
|  | K | S | 0.299 | 0.238 | 0.305 | **0.32** |
| SMD | F | B | 0.292 | 0.245 | **0.321** | 0.307 |
|  | K | B | 0.306 | 0.251 | **0.338** | 0.32 |
|  | F | S | 0.281 | 0.229 | **0.317** | 0.293 |
|  | K | S | 0.299 | 0.238 | **0.309** | 0.304 |

Table 3.8: Experimental results on Wine dataset using 16 selected input dendrograms. K and F denote K-medoid and Farthest Neighbor of ensemble selection methods respectively, and B and S denote Branch Length Score Distance and Symmetric Distance of dendrogram distances respectively.

### Experiments on ensemble selection

We also conducted experiments to demonstrate the effects of ensemble selection. Note that dendrogram selection can be performed using two different approaches ( K-medoid and Farthest neighbor, denoted as K and F) with two different distances (Branch Length Score Distance or Symmetric Distance, denoted as B and S). Table 3.8 to Table 3.13 present the experimental results on the **six** datasets using around 16 selected input dendrograms, respectively [3]. In these tables, *Sel* denotes the ensemble selection approaches, *Dis* represents

[3] The value of 16 is chosen based on our experiments on ensemble size selection and it seems to provide good results in our experiments. How to come up with a principled way for ensemble size selection is one of our future works.

the tree distances, *max* represents the maximum CPCC value for any input dendrogram, and ave represents the average CPCC value for the input dendrograms. The experiments show that: 1) with ensemble selection, the results of both *ultra* and hierarchical clustering (*SL* or *CL* have improved; 2) *ultra* still outperforms hierarchical clustering (*SL* or *CL* in most cases; 3) in many cases, the experiment results of *ultra* and hierarchical clustering (*SL* or *CL* outperform the best dendrogram in the candidate set, which means those ensemble dendrograms could be more representative of the original set; and 4) the Farthest Neighbor selection method tends to produce better ensemble results than K-Medoids.

| Desc. | Sel | Dis | max | ave | ultra | SL |
|---|---|---|---|---|---|---|
| CD | F | B | 0.438 | 0.256 | **0.549** | 0.521 |
| | K | B | 0.467 | 0.251 | 0.538 | **0.544** |
| | F | S | 0.493 | 0.273 | **0.537** | 0.505 |
| | K | S | 0.452 | 0.235 | **0.526** | 0.524 |
| CMD | F | B | 0.438 | 0.256 | **0.56** | 0.512 |
| | K | B | 0.467 | 0.251 | **0.572** | 0.542 |
| | F | S | 0.493 | 0.273 | **0.553** | 0.527 |
| | K | S | 0.452 | 0.235 | 0.524 | **0.536** |
| MED | F | B | 0.438 | 0.256 | **0.574** | 0.539 |
| | K | B | 0.467 | 0.251 | **0.595** | 0.532 |
| | F | S | 0.493 | 0.273 | **0.54** | 0.537 |
| | K | S | 0.452 | 0.235 | **0.589** | 0.527 |
| PMD | F | B | 0.438 | 0.256 | **0.517** | 0.492 |
| | K | B | 0.467 | 0.251 | 0.523 | **0.531** |
| | F | S | 0.493 | 0.273 | **0.502** | 0.499 |
| | K | S | 0.452 | 0.235 | **0.544** | 0.507 |
| SMD | F | B | 0.438 | 0.256 | 0.529 | 0.529 |
| | K | B | 0.467 | 0.251 | **0.551** | 0.504 |
| | F | S | 0.493 | 0.273 | **0.547** | 0.516 |
| | K | S | 0.452 | 0.235 | 0.498 | **0.511** |

Table 3.9: Experimental results on Parkinson Disease dataset using 16 selected input dendrograms. K and F denote K-medoid and Farthest Neighbor of ensemble selection methods respectively, and B and S denote Branch Length Score Distance and Symmetric Distance of dendrogram distances respectively.

### Experiments on ensemble size

To demonstrate the effect of the size of the ensemble, Figure 3.3 and Figure 3.4 show the performance variation on different number of selected dendrograms on all datasets. We apply K-Medoid selection methods on Symmetric Difference to choose candidate dendrograms. For each dataset, we vary the group size of candidate dendrograms and use CMD as the descriptor to conduct the dendrogram selection.

Figure 3.5 shows the CPCC value for each dendrogram group averaging over 20 runs. Note that for better readability, the plotted value of Madelon dataset is 10 times its actual value. The performance slightly decreases once the number of ensemble dendrograms reaches a certain size. So selecting a relatively smaller subset is likely to produce better ensemble results. It also shows that ensemble selection can influence the ensemble results and can be used to produce better hierarchical solutions.

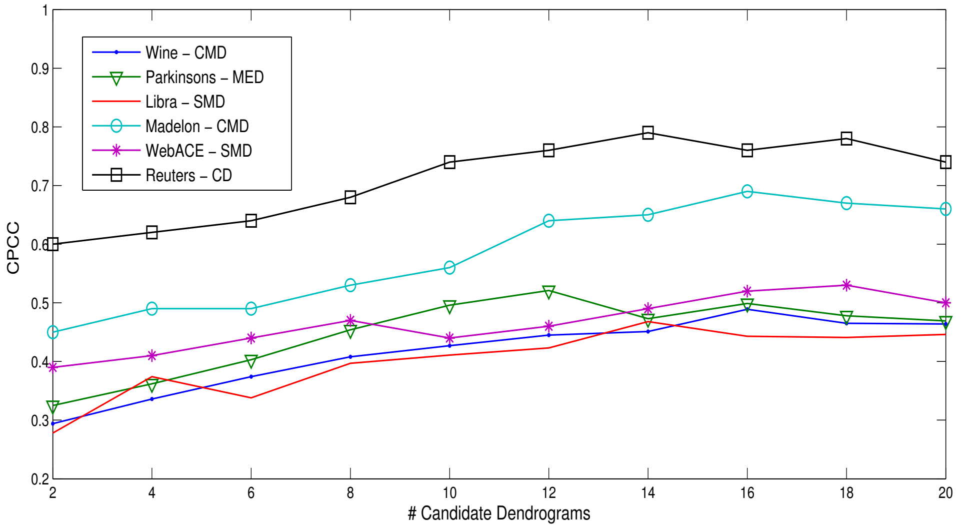
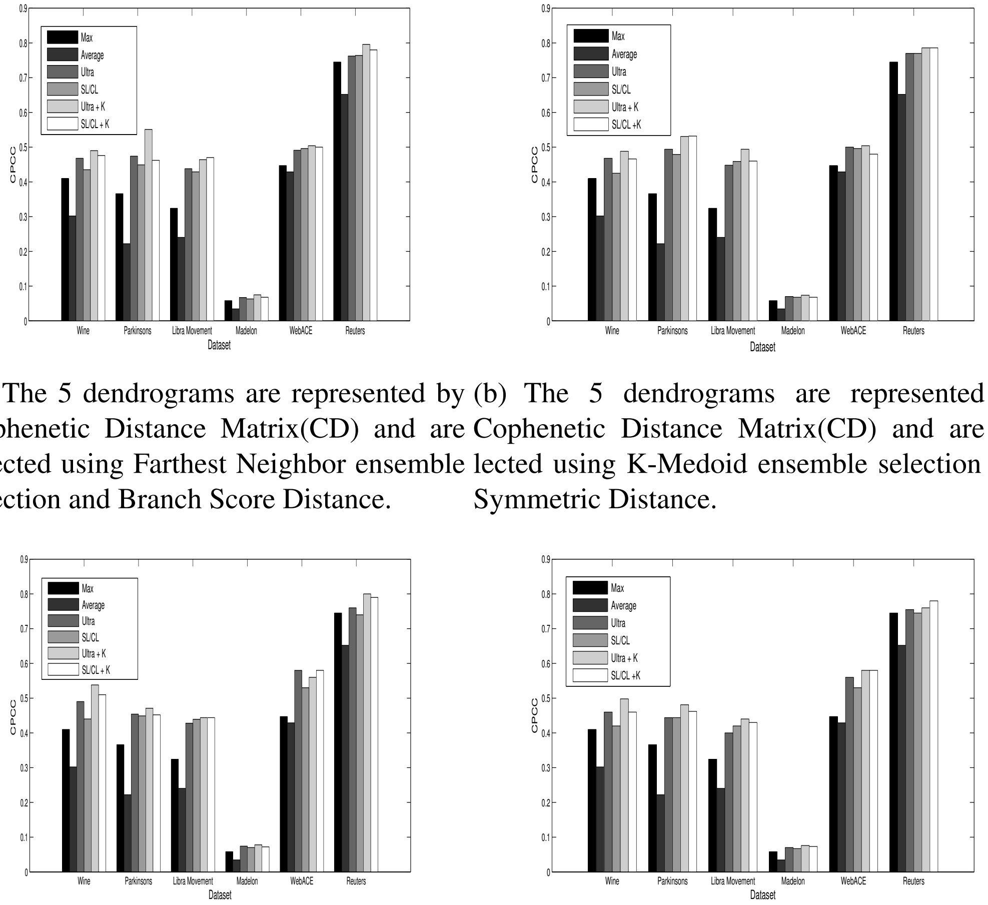


Figure 3.5: The performance variation on all datasets with different numbers of candidate dendrograms.



(a) The 5 dendrograms are represented by Cophenetic Distance Matrix(CD) and are selected using Farthest Neighbor ensemble selection and Branch Score Distance.

(b) The 5 dendrograms are represented by Cophenetic Distance Matrix(CD) and are selected using K-Medoid ensemble selection and Symmetric Distance.

(c) The 5 dendrograms are represented by Cluster Membership Divergence(CMD) and are selected using K-Medoid ensemble selection and Branch Score Distance.

(d) The 5 dendrograms are represented by Cluster Membership Divergence(CMD) and are selected using Farthest Neighbor ensemble selection and Symmetric Distance.

Figure 3.6: The performance comparison of combining 10 partitional clustering results with 10 selected dendrograms. max represents the maximum CPCC value for any input dendrogram, and ave represents the average CPCC value for the input dendrograms. ultra and SL/CL represents the recovery approaches for ensemble dendrograms by using ultra-matrix transformation and hierarchical clustering respectively. ultra+K and SL/CL+K represents the combination of K-means clustering results and previous two methods.

### 3.4.3 Ensemble partitional and hierarchical clusterings

We also conducted experiments to evaluate our proposed method for combining both partitional and hierarchical clusterings on all datasets. For each dataset, 10 partitional clustering results are obtained by running K-means 10 times and they are combined with 5 input dendrograms. The experimental results are shown in Figure 3.6. The results demonstrate that our ensemble framework is able to combine both partitional and hierarchical clusterings and improve the performance on most datasets. The results also show that our proposed method *ultra* clearly outperforms *SL/CL* on all datasets and *ultra+K* generally outperforms *SL/CL+K* in most cases.

| Desc. | Sel | Dis | max | ave | ultra | SL |
|---|---|---|---|---|---|---|
| CD | F | B | 0.287 | 0.199 | 0.392 | **0.433** |
| | K | B | 0.291 | 0.185 | **0.441** | 0.408 |
| | F | S | 0.274 | 0.167 | **0.4** | 0.396 |
| | K | S | 0.303 | 0.158 | **0.398** | 0.385 |
| CMD | F | B | 0.287 | 0.199 | **0.432** | 0.424 |
| | K | B | 0.291 | 0.185 | **0.446** | 0.418 |
| | F | S | 0.274 | 0.167 | **0.410** | 0.402 |
| | K | S | 0.303 | 0.158 | **0.453** | 0.391 |
| MED | F | B | 0.287 | 0.199 | **0.49** | 0.458 |
| | K | B | 0.291 | 0.185 | 0.442 | **0.476** |
| | F | S | 0.274 | 0.167 | **0.483** | 0.472 |
| | K | S | 0.303 | 0.158 | 0.453 | **0.461** |
| PMD | F | B | 0.287 | 0.199 | **0.397** | 0.346 |
| | K | B | 0.291 | 0.185 | **0.383** | 0.315 |
| | F | S | 0.274 | 0.167 | **0.401** | 0.359 |
| | K | S | 0.303 | 0.158 | **0.394** | 0.329 |
| SMD | F | B | 0.287 | 0.199 | **0.437** | 0.384 |
| | K | B | 0.291 | 0.185 | **0.462** | 0.391 |
| | F | S | 0.274 | 0.167 | 0.423 | **0.439** |
| | K | S | 0.303 | 0.158 | **0.468** | 0.379 |

Table 3.10: Experimental results on Libra Movement dataset using 16 selected input dendrograms. K and F denote K-medoid and Farthest Neighbor of ensemble selection methods respectively, and B and S denote Branch Length Score Distance and Symmetric Distance of dendrogram distances respectively.

| Desc. | Sel | Dis | max | ave | ultra | SL |
|---|---|---|---|---|---|---|
| CD | F | B | 0.052 | 0.028 | **0.06** | 0.058 |
| | K | B | 0.046 | 0.031 | 0.054 | 0.054 |
| | F | S | 0.047 | 0.032 | **0.531** | 0.049 |
| | K | S | 0.056 | 0.038 | **0.064** | 0.057 |
| CMD | F | B | 0.052 | 0.028 | **0.066** | 0.062 |
| | K | B | 0.046 | 0.031 | **0.059** | 0.054 |
| | F | S | 0.047 | 0.032 | **0.063** | 0.058 |
| | K | S | 0.056 | 0.038 | 0.06 | **0.069** |
| MED | F | B | 0.052 | 0.028 | **0.067** | 0.63 |
| | K | B | 0.046 | 0.031 | **0.063** | 0.058 |
| | F | S | 0.047 | 0.032 | 0.056 | **0.058** |
| | K | S | 0.056 | 0.038 | **0.069** | 0.056 |
| PMD | F | B | 0.052 | 0.028 | **0.074** | 0.06 |
| | K | B | 0.046 | 0.031 | **0.069** | 0.062 |
| | F | S | 0.047 | 0.032 | **0.071** | 0.058 |
| | K | S | 0.056 | 0.038 | **0.072** | 0.64 |
| SMD | F | B | 0.052 | 0.028 | 0.06 | **0.066** |
| | K | B | 0.046 | 0.031 | **0.054** | 0.051 |
| | F | S | 0.047 | 0.032 | **0.053** | 0.05 |
| | K | S | 0.056 | 0.038 | **0.059** | 0.056 |

Table 3.11: Experimental results on Madelon dataset using 16 selected input dendrograms. K and F denote K-medoid and Farthest Neighbor of ensemble selection methods respectively, and B and S denote Branch Length Score Distance and Symmetric Distance of dendrogram distances respectively.

## 3.5 Conclusion

In this chapter, we introduced and discussed a framework for ensemble hierarchical clusterings based on descriptor matrices and study three important components of the framework: Dendrogram Selection, Dendrogram Description and Dendrogram Combination. We propose two ensemble selection schemes based on tree distances, investigate five different dendrogram descriptor matrices, and develop a novel method for fitting an ultra-metric from the aggregated descriptor matrix. Our descriptor matrices based framework can be naturally generalized to ensemble both partitional clustering and hierarchical clustering results as partitional clustering results can be easily represented using distance matrices. Experiments are performed to evaluate our proposed approaches and the results demonstrate their effectiveness.

| Desc. | Sel | Dis | max | ave | ultra | CL |
|---|---|---|---|---|---|---|
| CD | F | B | 0.483 | 0.41 | **0.491** | 0.49 |
| | K | B | 0.474 | 0.409 | **0.505** | 0.499 |
| | F | S | 0.465 | 0.417 | 0.492 | 0.492 |
| | K | S | 0.487 | 0.405 | **0.501** | 0.494 |
| CMD | F | B | 0.483 | 0.41 | **0.511** | 0.501 |
| | K | B | 0.474 | 0.409 | **0.509** | 0.507 |
| | F | S | 0.465 | 0.417 | 0.498 | **0.503** |
| | K | S | 0.487 | 0.405 | **0.505** | 0.497 |
| MED | F | B | 0.483 | 0.41 | **0.513** | 0.502 |
| | K | B | 0.474 | 0.409 | **0.504** | 0.497 |
| | F | S | 0.465 | 0.417 | **0.5** | 0.497 |
| | K | S | 0.487 | 0.405 | **0.507** | 0.489 |
| PMD | F | B | 0.483 | 0.41 | 0.496 | **0.498** |
| | K | B | 0.474 | 0.409 | 0.492 | **0.497** |
| | F | S | 0.465 | 0.417 | **0.501** | 0.5 |
| | K | S | 0.487 | 0.405 | **0.498** | 0.49 |
| SMD | F | B | 0.483 | 0.41 | **0.503** | 0.491 |
| | K | B | 0.474 | 0.409 | **0.5** | 0.493 |
| | F | S | 0.465 | 0.417 | **0.499** | 0.484 |
| | K | S | 0.487 | 0.405 | **0.507** | 0.495 |

Table 3.12: Experimental results on WebACE dataset using 16 selected input dendrograms. K and F denote K-medoid and Farthest Neighbor of ensemble selection methods respectively, and B and S denote Branch Length Score Distance and Symmetric Distance of dendrogram distances respectively.

There are several avenues for future work. First, we plan to investigate the techniques for scaling up the ensemble process to large-scale datasets. Second, our studies show that selecting a relatively smaller subset is likely to produce better ensemble results. One interesting question is how to determine the ensemble size. Another interesting yet related direction is that rather than picking representative dendrograms, we can associate every generated dendrogram with a weight. So when considering the ensemble, dendrograms with larger weights can contribute more than dendrograms with smaller weights. Third, another aspect of interest is to provide a formal analysis on cluster separation enhancement using transitive dissimilarity.

| Desc. | Sel | Dis | max | ave | ultra | CL |
|---|---|---|---|---|---|---|
| CD | F | B | 0.73 | 0.682 | **0.747** | 0.739 |
| | K | B | 0.741 | 0.635 | 0.785 | **0.794** |
| | F | S | 0.737 | 0.696 | **0.792** | 0.786 |
| | K | S | 0.729 | 0.64 | **0.769** | 0.75 |
| CMD | F | B | 0.73 | 0.682 | **0.793** | 0.767 |
| | K | B | 0.741 | 0.635 | **0.798** | 0.752 |
| | F | S | 0.737 | 0.696 | **0.794** | 0.755 |
| | K | S | 0.729 | 0.64 | **0.782** | 0.751 |
| MED | F | B | 0.73 | 0.682 | **0.779** | 0.754 |
| | K | B | 0.741 | 0.635 | **0.783** | 0.781 |
| | F | S | 0.737 | 0.696 | 0.765 | **0.77** |
| | K | S | 0.729 | 0.64 | **0.752** | 0.75 |
| PMD | F | B | 0.73 | 0.682 | **0.782** | 0.763 |
| | K | B | 0.741 | 0.635 | **0.775** | 0.755 |
| | F | S | 0.737 | 0.696 | **0.787** | 0.761 |
| | K | S | 0.729 | 0.64 | 0.74 | **0.745** |
| SMD | F | B | 0.742 | 0.726 | **0.797** | 0.784 |
| | K | B | 0.744 | 0.727 | **0.782** | 0.753 |
| | F | S | 0.736 | 0.730 | **0.771** | 0.767 |
| | K | S | 0.731 | 0.722 | 0.75 | 0.75 |

Table 3.13: Experimental results on Reuters dataset using 16 selected input dendrograms. K and F denote K-medoid and Farthest Neighbor of ensemble selection methods respectively, and B and S denote Branch Length Score Distance and Symmetric Distance of dendrogram distances respectively.

## CHAPTER 4

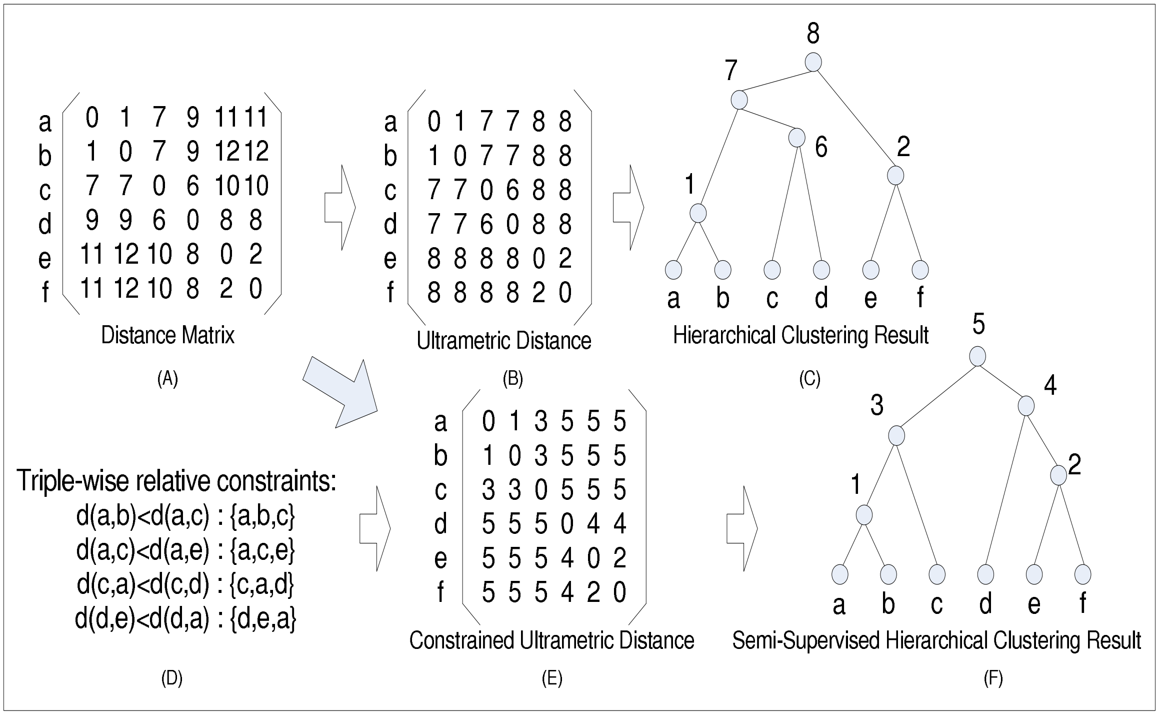
# Semi-supervised hierarchical clustering

The clustering problem arises in many disciplines and has a wide range of applications. Basically clustering aims to group the given samples into clusters such that samples in the same cluster are similar to each other while samples in different clusters are dissimilar [JD88]. Based on the way the clusters are generated, clustering methods can be divided into two categories: partitional clustering and hierarchical clustering [HK06][TSK05a]. Generally *partitional clustering* decomposes the dataset into a number of disjoint clusters which are usually optimal in terms of some predefined objective functions. *Hierarchical clustering* groups the data points into a hierarchical tree-like structure using bottom-up or top-down approaches.

In many situations when we discover new patterns using clustering, there are known prior knowledge about the problem. Recently, semi-supervised clustering (i.e., clustering with knowledge-based constraints) has emerged as an important variant of the traditional clustering paradigms [DR05][LL05]. Given the data representation, existing semi-supervised methods have utilized background knowledge to learn a distance/dissimilarity measure, to modify the objective criterion for evaluating clustering, and to improve the optimization procedures [BhHSW05, BBM04, KKM02, Wag02, XNJR02, Zhu05].

There are two limitations in current studies of semi-supervised clustering. First, most of these existing semi-supervised clustering algorithms are designed for partitional clustering methods and *few research efforts have been reported on semi-supervised hierarchical clustering methods*. Different from partitional clustering where the clustering results can be easily represented using vectors, clustering indicators, or connectivity matrices for optimization [XNJR02], the results of hierarchical clustering are more complex and typically represented as dendrograms or trees. In addition, hierarchical clustering methods have no global objective functions. These properties have made the semi-supervised hierarchical clustering problem more challenging.

Another limitation is *on the types of constraints*. Existing semi-supervised clustering methods have been focused on the use of background information in the form of instance level must-link and cannot-link constraints. A must-link (ML) constraint enforces that two instances must be placed in the same cluster while a cannot-link (CL) constraint enforces that two instances must not be placed in the same cluster. However, *both ML and CL constraints are not suitable for hierarchical clustering methods since objects are linked over different hierarchy levels* [BN06][BN08b].



$$\begin{array}{c} a\\ b\\ c\\ d\\ e\\ f \end{array}\begin{pmatrix} 0 & 1 & 7 & 9 & 11 & 11\\ 1 & 0 & 7 & 9 & 12 & 12\\ 7 & 7 & 0 & 6 & 10 & 10\\ 9 & 9 & 6 & 0 & 8 & 8\\ 11 & 12 & 10 & 8 & 0 & 2\\ 11 & 12 & 10 & 8 & 2 & 0 \end{pmatrix}$$

Distance Matrix

(A)

$$\begin{array}{c} a\\ b\\ c\\ d\\ e\\ f \end{array}\begin{pmatrix} 0 & 1 & 7 & 7 & 8 & 8\\ 1 & 0 & 7 & 7 & 8 & 8\\ 7 & 7 & 0 & 6 & 8 & 8\\ 7 & 7 & 6 & 0 & 8 & 8\\ 8 & 8 & 8 & 8 & 0 & 2\\ 8 & 8 & 8 & 8 & 2 & 0 \end{pmatrix}$$

Ultrametric Distance

(B)

Hierarchical Clustering Result

(C)

Triple-wise relative constraints:

d(a,b)<d(a,c) : {a,b,c}
d(a,c)<d(a,e) : {a,c,e}
d(c,a)<d(c,d) : {c,a,d}
d(d,e)<d(d,a) : {d,e,a}

(D)

$$\begin{array}{c} a\\ b\\ c\\ d\\ e\\ f \end{array}\begin{pmatrix} 0 & 1 & 3 & 5 & 5 & 5\\ 1 & 0 & 3 & 5 & 5 & 5\\ 3 & 3 & 0 & 5 & 5 & 5\\ 5 & 5 & 5 & 0 & 4 & 4\\ 5 & 5 & 5 & 4 & 0 & 2\\ 5 & 5 & 5 & 4 & 2 & 0 \end{pmatrix}$$

Constrained Ultrametric Distance

(E)

Semi-Supervised Hierarchical Clustering Result

(F)

Figure 4.1: An illustrative example of semi-supervised hierarchical clustering with triple-wise relative constraints. The original data dissimilarity matrix is shown in (A). (B) shows a standard transitive dissimilarity matrix obtained from the original dissimilarity and (C) is the corresponding hierarchical clustering result without constraints. The triple-wise relative constraints are given in (D). By combining both (A) and (D), the constrained ultra-metric distance matrix is shown in (E) with its corresponding hierarchial clustering result in (F).

In this chapter, we demonstrate a semi-supervised hierarchical clustering framework based on the ultra-metric dendrogram distance. The characteristics of our proposed framework are summarized below:

1. Triple-wise relative constraints: In the proposed framework, we consider the triple-wise relative constraints in the form of $(x_i, x_j, x_k)$ which indicates the dissimilarity (or the distance) between $x_i$ and $x_j$, noted as $d(x_i, x_j)$, should be smaller than $d(x_i, x_k)$. The relative constraint, also referred as must-link-before (MLB) constraint, specifies the order in which the objects are merged (or linked) and can be naturally incorporated into the hierarchical clustering process.

2. Ultra-metric dendrogram distance: Our proposed framework is based on ultra-metric dendrogram distance. Note that the results of hierarchical clustering can be represented using ultra-metric distance matrices [Pod00b]. Using the ultra-metric distance matrices, we propose two techniques for solving semi-supervised hierarchical clustering problem: the optimization-based technique and the transitive dissimilarity based technique.

3. Effectiveness and efficiency: Extensive experimental results demonstrate the effectiveness and efficiency of our proposed framework.

An illustrative example of semi-supervised hierarchical clustering is given in Figure 4.1. The original dissimilarity is shown in Figure 4.1(A). Its ultra-metric distance matrix is shown in Figure 4.1(B) and the corresponding hierarchical clustering result (without constraints) is shown in Figure 4.1(C). Four triple-wise relative constraints are given in Figure 4.1(D). A constrained ultra-metric distance matrix is obtained in Figure 4.1(E) and its corresponding hierarchical clustering result (with constraints) is shown in Figure 4.1(F).

To sum up, different from existing research efforts on semi-supervised (hierarchical) clustering, in our work, we explicitly establish the equivalence between ultra-metrics and hierarchical clustering and also provide a unified framework integrating both ultra-metric fitting and triple-wise relative constraints, discussed in Chapter3.2. Our framework seeks an approximate dissimilarity metric (ultra-metric) which represents a tuned dendrogram that satisfies the given constraints. Two different solutions based on iterative projection and heuristic (modified Floyd-Warshall) algorithms are proposed and empirically evaluated.

The rest of the chapter is organized as follows: Section 4.1 discusses the related work; Section 4.2 extended the discussion about the transitivity that ultra-metric transformation is preserved; Section 4.3 comprehensively discusses the semi-supervised hierarchical clustering, in which Section 4.3.1 formally defines the semi-supervised hierarchical problem and Section 4.3.2 presents two different techniques for semi-supervised hierarchical clustering based on ultra-metric distance; Section 4.4 describes the experimental results; and finally Section 4.5 concludes the chapter.

## 4.1 Related work

**Hierarchical Clustering** Hierarchical clustering algorithms are unsupervised methods to generate tree-like clustering solutions. They group the data points into a hierarchical structure using bottom-up (agglomerative) or top-down (divisive) approaches [TSK05a]. The typical bottom-up approaches take each data point as a single cluster to start with and then builds bigger clusters by grouping similar clusters together until the entire data set is grouped into one final cluster. The divisive approaches start with all data points in one cluster and then split the largest cluster recursively. Many research efforts have been reported on algorithm-level improvements of the hierarchical clustering process and on understanding of hierarchical clustering [YCWX09][ZK02b].

**Semi-supervised Clustering:** Integrating background knowledge into the clustering process has been investigated extensively. Many researchers have explored the use of instance-level background information, such as pairwise must-link and cannot-link constraints for learning a distance/dissimilarity measure, or modifying the objective criterion, or improving the optimization procedures [BhHSW05, BBM04, KKM02, Wag02, XNJR02, Zhu05]. Other types of knowledge hints (such as size of the clusters, partial labels of the data points, and user-provided external prototypes/representatives) have also been used for clustering [Ped04]. However, most of these works are designed for partitional clustering and few research efforts have been reported on semi-supervised hierarchical clustering methods. We note the very recent works of Zhao et al. [ZQ10] and Bade et al. [BN06] which perform hierarchical clustering with order constraints and partially known hierarchy. Conceptually our triple-wise constraints are special cases of the order constraints. However, different from their works, our proposed semi-supervised hierarchical clustering framework is based on ultra-metric dendrogram distance. Experimental studies demonstrate the effectiveness and efficiency of our proposal.

**Metric Fitting:** The problem of fitting a tree metric to the (dis-)similarity data on pairs of objects from a given set has been studied quite extensively [AC05b]. Ultra-metric is a special kind of tree metric where all elements of the input dataset are leaves in the underlying tree and all leaves are at the same distance from the root. Ultra-metric naturally corresponds to a hierarchy of clusterings of the data. Given a dissimilarity $D$ on pairs of objects, the problem of finding the best ultra-metric $d_u$ such that $||D - d_u||_p$ is minimized is NP-hard for $L_1$ and $L_2$ norms (e.g., when $p = 1$ and $p = 2$) [ABF$^+$99]. In this chapter, we propose two techniques for fitting an ultra-metric using the given relative constraints.

## 4.2 Transitive dissimilarity

In Chapter 3, we establish the explicit relation between ultra-metric, dendrogram, and hierarchical clustering in Section 3.2. In this section, we further the study on the transitive dissimilarity that the ultra-metric transformation preserves and connect the dots between transitivity dissimilarity and clustering.

### 4.2.1 Transitive preservation

First, the nonnegative distance $D$ can be viewed as the edge weight on a graph. Our task is to construct the transitive dissimilarity starting from $D$.

The idea of transitive dissimilarity is to **preserve transitivity** of a graph, more precisely a social network with $n$ persons represented as $(V_1 \cdots, V_n)$. If person $V_1$ knows person $V_2$, and person $V_2$ knows person $V_3$, transitivity implies that person $V_1$ knows person $V_3$. Turning this into distances, the transitivity of $V_1 \to V_2 \to V_3$ can be enforced as

$$d_{13} \leq \max(d_{12}, d_{23}),$$

i.e., the distance $d_{13}$ should be no greater than either $d_{12}$ or $d_{23}$.

Now consider 4 persons. One can see the above enforcement satisfies the associativity: i.e., if both $d_{13} \leq \max(d_{12}, d_{23})$ and $d_{24} \leq \max(d_{23}, d_{34})$ hold, then

$$d_{14} \leq \max(d_{12}, d_{23}, d_{34}).$$

Generalizing to any path $P_{ij}$ between $i$ and $j$, on the graph, the **transitive dissimilarity** on a path $P_{ij}$ (a set of edges connect $V_i$ and $V_j$) can be defined as

$$T(P_{ij}) = \max(d_{i,k_1}, d_{k_1,k_2}, d_{k_2,k_3}, \cdots, d_{k_{n-1},k_n}, d_{k_n,j}). \tag{4.1}$$

So for any given pair of vertices $V_i$ and $V_j$, the transitive dissimilarity varies according to different paths chosen between $V_i$ and $V_j$. The **minimal transitive dissimilarity** is defined as:

$$m_{ij} = \min_{P_{ij}}(T(P_{ij})), \text{for given vertices } V_i \text{ and } V_j. \tag{4.2}$$

It is clear that $m_{ij} \leq d_{ij}, \forall V_i$ and $V_j$, which implies that minimal transitive dissimilarity brings vertices closer than the original distance matrix.

**Lemma 1.** *Triangle Inequality is preserved in consensus similarity if each individual distance satisfies it. But Ultra-metric inequality is not preserved even if each individual dendrogram distance satisfies it.*

*Proof.* The proof of the first part is trivial. To prove the second part, we give a counterexample. We construct two dendrograms on a dataset with three points, denoted as A and B. The distance between points in A are $d_{12}^A = d_{23}^A = 3$ and $d_{13}^A = 2$; The distance between points in B are $d_{12}^B = d_{13}^B = 3$ and $d_{23}^B = 2$. C is the consensus dendrogram of A and B. The distance between points in C is given by $d_{ij}^C = \frac{1}{2}(d_{ij}^A + d_{ij}^B)$, thus $d_{13}^C = d_{23}^C = 5/2$ and $d_{12}^C = 3$. Clearly, $d_{12}^C \leq \max(d_{13}^C, d_{32}^C)$ is violated and the utlra-metric inequality of consensus clustering does not hold. □

**Proposition 2.** *For any weighted dissimilarity graph, the minimal transitive dissimilarity between any pair of vertices satisfies the ultra-metric inequality:*

$$m_{ij} \leq \max(m_{ik}, m_{kj}), \forall i, j, k.$$

*Proof.* Let $P_{ij}$ is a set of all paths in which each element indicates an existing path connecting $V_i$ and $V_j$ as its end points. $(P_{ik}, P_{kj})$ is describing a path starting from $V_i$ to $V_j$ via $V_k$ in a weighted graph. It is clear that $(P_{ik}, P_{kj})$ is a subset of $P_{ij}$. We define $W(P_{ij})$ as edge weights of any directly connected vertices in all possible paths $Pij$.

$$\begin{aligned} m_{ij} &= \min_{P_{ij}} \max[W(P_{ij})] \\ &\leq \min(P_{ik}, P_{kj}) \max(W(P_{ik}, P_{kj})) \\ &= \min(P_{ik}, P_{kj}) \max[\max[W(P_{ik})], \max[W(P_{kj})]] \\ &= \max[\min_{P_{ik}}(\max[W_{P_{ik}}]), \min_{P_{kj}}(\max[W_{P_{kj}}])] \\ &= \max(m_{ik}, m_{kj}). \end{aligned}$$

□

Thus, the problem of obtaining the ultra-metric transformation of a consensus matrix can be formulated as the following optimization problem:

**Problem 1.** *A is the consensus distance matrix; B is the desired ultra-metric to be computed:*

$$\min_B \sum_{ij} |A_{ij} - B_{ij}|, s.t. \;\; B_{ij} \leq A_{ij}. \tag{4.3}$$

The ultra-metric constraint on $B$ is a hard constraint. The optimal solution is given by Algorithm 1. In other words, the desired ultra-metric distance always smaller than input distance.

Then, we use the modified Floyd-Warshall algorithm [DHX$^+$06] to compute the updated transitive dissimilarity of all pair of vertices in the weighted graph. In particular, given input G, the adjacency matrix of a weighted graph with $N$ nodes, the algorithm procedure is described in Algorithm 1.

```
Input: G: Pairwise distance matrix of data set.
Output: M: Minimum Transitive dissimilarity matrix closure of G.
Init: M = G.
for $k \leftarrow 0$ to $N$ do
    for $i \leftarrow 0$ to $N$ do
        for $j \leftarrow 0$ to $N$ do
            $m_{ij} = \min(m_{ij}, \max(m_{ik}, m_{kj}))$;
        end for
    end for
end for
return $M$ ;
```

**Algorithm 1:** Modified Floyd-Warshall algorithm to compute the minimum transitive dissimilarity of weighted graph G

The following propositions are needed to show the correctness of the modified Floyd-Warshall algorithm.

**Proposition 3.** *Suppose the edge weights of given graph satisfy the minimal transitive dissimilarities as defined in Eq.(4.2). The transitive dissimilarities are equal to the edge weights.*

*Proof.* We prove Proposition 3 using dynamic programming. Starting from 2-hop paths $V_i$-$V_k$-$V_j$ between any given vertices $V_i$ and $V_j$. As the edge weights $d$ satisfy the minimal transitive dissimilarities, so $d_{ij}$ must be less or equal to 2-hop transitive weight $T(P_{ikj})$ for any k. Since we have minimal transitive dissimilarity $m_{ij} \leq d_{ij}$ implied by Eq.(4.2), so $m_{ij} \leq d_{ij} \leq T(P_{ikj})$ holds. For 2-hop minimal transitive dissimilarity, we get $m_{ij} = d_{ij}$.

Given any 3-hop path between $V_i$ and $V_j$, denoted as $V_i$-$V_k$-$V_l$-$V_j$, we can change $V_i$-$V_k$-$V_l$ to $V_i$-$V_l$, or change $V_k$-$V_l$-$V_j$ to $V_k$-$V_j$ based on the destination from 2-hop paths. We apply transitive dissimilarity and edge weight equivalence property again on path $V_i$-$V_l$-$V_j$ or $V_i$-$V_k$-$V_j$ again, then we get $m_{ij} = d_{ij}$, for any path $V_i$-$V_k$-$V_l$-$V_j$.

For any n-hop path $(n \geq 2)$, the same process can be applied. Thus Proposition 3 is proved. □

**Proposition 4.** *Given node pair $V_i$ and $V_j$. Let $V_i$-$V_{k1}$-$\cdots$-$V_{km}$-$V_j)$ be the path with the eventual minimal transitive dissimilarity. After successive tightening of edges $V_i$-$V_{k1}$, $V_{k1}$-$V_{k2}$, $\cdots$, $V_{km}$-$V_j$ in order, the transitive dissimilarity achieves the final optimal maximal transitive dissimilarity. This holds no matter what other edge relaxations occur.*

*Proof.* Since the eventual path between $V_i$ and $V_j$ with minimal transitive dissimilarity is given, the length-2 minimal transitive dissimilarity (optimal solution) can be easily obtained. Also, the length-3 minimal transitive dissimilarity can be obtained based on length-2 solution, and it is obviously the optimal solution. The conclusion holds when extending to the last edge of the path. Thus Proposition 4 is proved. □

**Proposition 5.** *Algorithm 1 correctly computes the minimum transitive dissimilarity.*

*Proof.* The outer loop $k = 1$ to $N$ guarantees that all paths between any given vertices $V_i$ and $V_j$ will be considered to achieve the eventual optimal path. Proposition 4 ensures that final correct solution will be reached no matter how internal vertices along the path are involved. Proposition 3 guarantees that any optimal solution obtained before traversing all the possible solutions will be maintained without change in the future. □

Table 4.1: Original distance among the 10 objects shown in Figure 4.2.

| $i, j$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 138 | 288 | 428 | 532 | 262 | 230 | 267 | 335 | 432 |
| 2 | 138 | 0 | 151 | 301 | 419 | 336 | 272 | 240 | 266 | 370 |
| 3 | 288 | 151 | 0 | 163 | 295 | 435 | 356 | 268 | 234 | 326 |
| 4 | 428 | 301 | 163 | 0 | 137 | 504 | 419 | 299 | 214 | 260 |
| 5 | 532 | 419 | 295 | 137 | 0 | 550 | 468 | 341 | 241 | 226 |
| 6 | 262 | 336 | 435 | 504 | 550 | 0 | 85 | 210 | 309 | 356 |
| 7 | 230 | 272 | 356 | 419 | 468 | 85 | 0 | 127 | 227 | 284 |
| 8 | 267 | 240 | 268 | 299 | 341 | 210 | 127 | 0 | 101 | 173 |
| 9 | 335 | 266 | 234 | 214 | 241 | 309 | 227 | 101 | 0 | 104 |
| 10 | 432 | 370 | 326 | 260 | 226 | 356 | 284 | 173 | 104 | 0 |

### 4.2.2 Cluster separation enhancement and transitive dissimilarity

Here we demonstrate the enhancement of cluster separation due to the transitive dissimilarity. We use a small dataset shown in Figure 4.2, where the two clusters are reasonably visible.

The original distance of the dataset is shown in Table 4.1 and the transitive distance is shown in Table 4.2. The distance is computed using Euclidean distance and the value is scaled by multiplying 1000 for readability. It is clear that the distance in Table 4.2 provides an enhanced/improved 2-cluster structure, because the diagonal block (1-5) and (6-10) elements (distances within the same cluster) are visibly reduced. while the distance between the two clusters remain at the fixed value 214.

For example, the original distance between $x_1$ and $x_5$ (they are in the same cluster) $d_{15}^{original} = 532$, while the original distance between $x_1$ and $x_6$ (they are in different clusters) $d_{16}^{original} = 262$. This is not **intuitive** because it implies that members of the same cluster could have **larger** distance than the distance between members of different clusters.

With transitive distance, this counterintuitive situation is **corrected** because now $d_{15}^{transitive} = 163$, while $d_{16}^{transitive} = 214$. The key point is that within-cluster distances shrinked more than between-cluster distances.

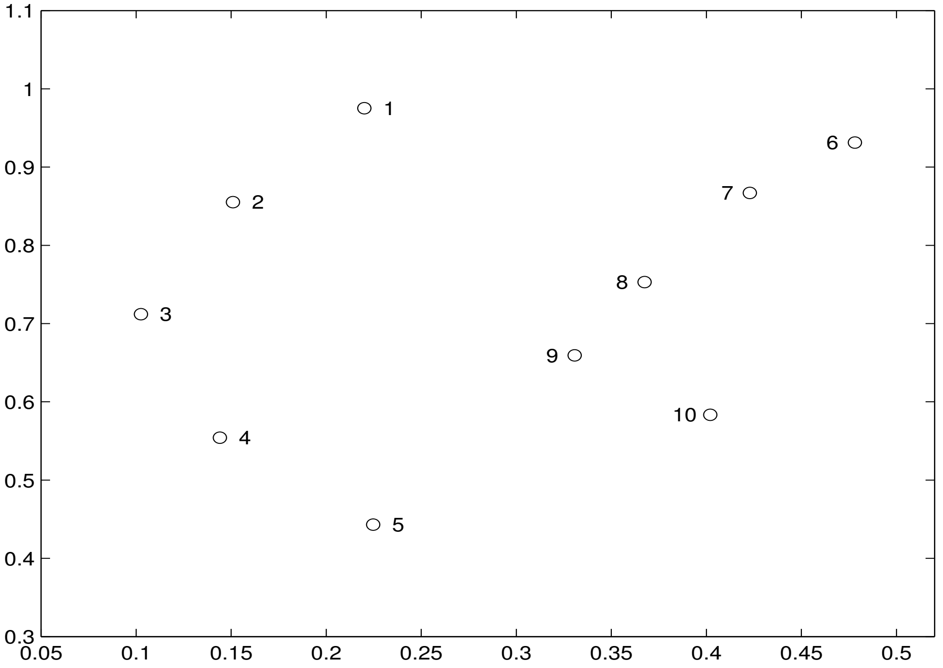


Figure 4.2: Illustration of cluster separation due to the transitive dissimilarity for a simple dataset of 10 points in 2D space.

## 4.3 Semi-supervised hierarchical clustering

### 4.3.1 Problem statement

**Problem Definition**

Given set of instances $X = \{x_1, x_2, \cdots, x_n\}$, their pair-wise dissimilarities $D = \{d(x_i, x_j)|x_i, x_j \in X\}$ and a set of constraints $C = \{(x_i, x_j, x_k)\ |d(x_i, x_j) < d(x_i, x_k)$ $, x_i, x_j, x_k \in X\}$. The semi-supervised hierarchical clustering problem aims to output a clusters hierarchy/dendrogram $\mathcal{H}$ to *satisfy as many constraints as possible* and meanwhile to *maintain the merge order based on sample dissimilarities as close as possible*.

Note that hierarchical clustering results can be represented graphically on dendrograms as shown in Figure 4.3. The vertical line along with the clustering dendrogram is labeled

Table 4.2: Transitive distance among the 10 objects shown in Figure 4.2.

| $i, j$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 138 | 151 | 163 | 163 | 214 | 214 | 214 | 214 | 214 |
| 2 | 138 | 0 | 151 | 163 | 163 | 214 | 214 | 214 | 214 | 214 |
| 3 | 151 | 151 | 0 | 163 | 163 | 214 | 214 | 214 | 214 | 214 |
| 4 | 163 | 163 | 163 | 0 | 137 | 214 | 214 | 214 | 214 | 214 |
| 5 | 163 | 163 | 163 | 137 | 0 | 214 | 214 | 214 | 214 | 214 |
| 6 | 214 | 214 | 214 | 214 | 214 | 0 | 85 | 127 | 127 | 127 |
| 7 | 214 | 214 | 214 | 214 | 214 | 85 | 0 | 127 | 127 | 127 |
| 8 | 214 | 214 | 214 | 214 | 214 | 127 | 127 | 0 | 101 | 104 |
| 9 | 214 | 214 | 214 | 214 | 214 | 127 | 127 | 101 | 0 | 104 |
| 10 | 214 | 214 | 214 | 214 | 214 | 127 | 127 | 104 | 104 | 0 |

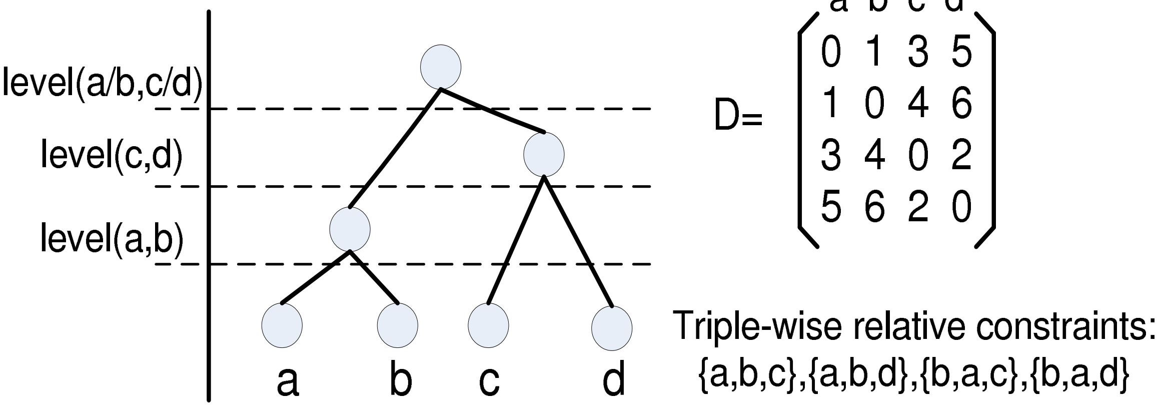


Figure 4.3: Triple-wise relative constraints for samples a and b in 4-point sample set.

by the value of the updated dissimilarity between the merged clusters, which can be treated as a measure of separation of paired samples. The dissimilarity of sample $a$ and $c$ in the dendrogram is noted by $level(a, c)$. Note that some relative constraints (e.g., constraint (a, b, c) in Figure 4.3) are consistent with the given dissimilarity matrix while many constraints are not (e.g., constraint (a, d, b)).

### Constraint Pre-processing

**Transitive Closure:** Constraints given by human experts or by partially known data hierarchy may be *incomplete*. Some constraints are not explicitly given, for example, two given constraints $c_1 = (x_i, x_j, x_k)$ and $c_2 = (x_i, x_k, x_l)$ imply an additional constraint $c_3 = (x_i, x_j, x_l)$ which might not be explicitly stated. In our framework, given the initial constraint set, we perform Floyd-Warshall algorithm to find its transitive closure and extend the constraint set.

**Conflict Removal:** In practice, the given constraints may be *conflicting*. For example $c_1 = (x_i, x_j, x_k)$ and $c_2 = (x_j, x_k, x_i)$ are explicitly conflicting with each other or $c_1 = (x_i, x_j, x_k)$ , $c_2 = (x_i, x_k, x_l)$ and $c_3 = (x_i, x_l, x_j)$ form a circle of merge orders. Conflicts in the constraint set can form deadlocks, and the clustering algorithm may fail to identify a valid merging pair of clusters. To remove conflicts, we randomly and iteratively remove one of the conflicting constraint until there is no conflict.

### 4.3.2 Algorithm

We implementded two approaches for semi-supervised hierarchical clustering based on ultra-metric distance matrices: the optimization-based approach and the transitive dissimilarity based approach. The optimization-based approach models the semi-supervised hierarchical clustering as a constrained optimization problem of constructing an optimal distance matrix satisfying both the ultra-metric constraints and relative constraints. The transitive dissimilarity based approach aims to incorporate the relative constraints into the process of constructing the transitive dissimilarity.

#### Constraint-based Optimization

In semi-supervised hierarchical clustering, besides satisfying ultra-metricity, the clustering results should also consider relative constraints. We assume the dissimilarity matrix is non-negative and symmetric in our proposed algorithm, so we can adopt a vector representation. Suppose we have $n$ samples and $r$ relative constraints. For simplicity, the $n \times n$ symmetric dissimilarity matrix $D$ can be represented by an $m \times 1$ vector $\vec{d}$ with $m = n * (n-1)/2$

entries of $D$'s upper/lower triangle elements. Thus, each relative constraint $(x_i, x_j, x_k) \in C$ can also be represented by an $m \times 1$ vector $\vec{c}$ where the index corresponds to $D_{ij}$ is set to $1$ and the index of $D_{ik}$ is set to $-1$. So, for any constraint $c$ that is not consistent with the given dissimilarity matrix, we have $d^T c \geq 0$. An illustrative example is shown in Figure 4.4. Following the vector representation of dissimilarity and constraints, semi-supervised hierarchical clustering problem can be represented in the form below:

$$\arg\min_{\hat{d}} (\vec{d} - \vec{\hat{d}})^T E (\vec{d} - \vec{\hat{d}}), \tag{4.4}$$

subject to

$$\hat{D}_{ij} \leq \max\{\hat{D}_{ik}, \hat{D}_{jk}\}, \forall x_i, x_j, x_k \in X, \tag{4.5}$$

$$C\vec{d} \leq \vec{0}, \tag{4.6}$$

where $\vec{d}$ and $\vec{\hat{d}}$ are vectors representing pair-wise dissimilarities, E is a $m \times m$ identity matrix, and $C = [c_1^T; c_2^T; \cdots; c_r^T]$ is an $r \times m$ matrix containing all $r$ relative constraints.

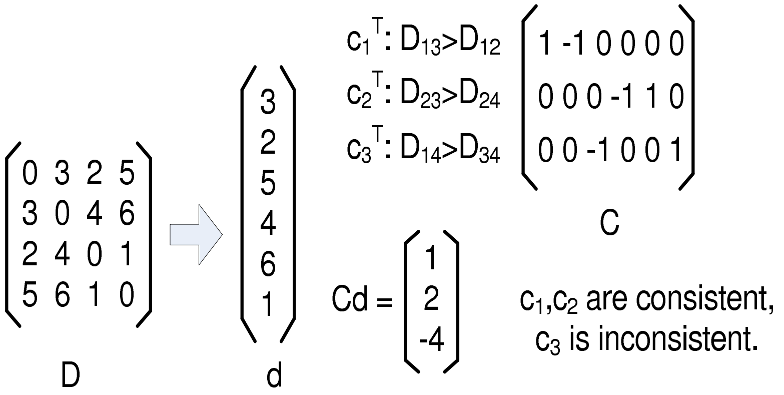


Figure 4.4: Utilizing constraints in the optimization framework.

The above optimization problem can be solved by conducting iterative projection approach which provides optimal solution to minimize the least-square loss function under inequality constraints [HA95][J.00][Soe84]. Different from related approaches, our problem formulation considers both ultra-metric and triple-wise relative constraints and seeks an approximate dissimilarity metric (ultra-metric) that satisfies the given constraints. The ultra-metricity of the dissimilarity is taken as the underlying constraints to generate a tree-like hierarchy. Iterative projection can be generally conducted by repeatedly following the iterative "augmenting" steps. At each iteration, the parameter estimates are first projected onto closed convex sets defined by the inequality constraints $C\vec{d} \leq \vec{0}$, and are then updated by subtracting a vector of the changes made in the previous projection. Iterative projection carried out with this augmentation step is guaranteed to converge to the least squares optimal solution for a given fixed set of constraints [J.01].

Algorithm 2 shows a simple implementation of iterative projection used in [Dyk83]. The procedure simultaneously generates sequence of estimated solutions $a(t)$ and a sequence of Kuhn-Tucker vectors $u(t)$ where $a(t)$ and $u(t)$ denote the $\vec{a}$ and $\vec{u}$ in iteration $t$ [KT50].

```
Input: $\vec{d}, C, E$
Output: $\vec{\hat{d}}$
Init: $\vec{a} = \vec{d}$ and $\vec{u} = \vec{0}$.
while not converge do
    $p = t \bmod r$
    $\vec{s} = \vec{a}(t-1) + E\vec{c_p}\vec{u}(t-1)_p/2$
    for $q = 1$ to $r$ do
        if $q = p$ then
            $\vec{u}(t)_q = \max(0, 2 * \vec{c_q^T}\vec{s}/\vec{c_q^T}E\vec{c_q}))$
        else
            $\vec{u}(t)_q = \vec{u}(t-1)_q$
        end if
    end for
    $\vec{a}(t) = \vec{s} - E\vec{c_q}\vec{u}(t)_q/2$
    $t = t + 1$
end while
return $\vec{\hat{d}} = \vec{a}$ ;
```

**Algorithm 2:** Iterative Projection to minimize least-square error

Note that the iterative projection approach can be extended to an L1 minimization algorithm by using iteratively re-weighted least-squares (IRLS) framework [J.00][B.83].

### Transpositive Dissimilarity

The Floyd-Warshall algorithm can be used to compute the minimum transitive dissimilarity. In this section, we modified the Floyd-Warshall algorithm to fit the original dissimilarity matrix to a ultra-matrix and at the meantime to incorporate the given relative constraints. Algorithm 3 shows the algorithm procedure to incorporate the relative constraints into the ultra-metric transformation process. Its difference from standard Floyd-Warshall algorithm is that the updated value for $m_{ij}$ is not only determined by the pairwise dissimilarities related to $x_i$ and $x_j$, but also restricted by any constraints specifying merge orders about them (see Lines 4-7).

```
Input: G: Pair-wise distance matrix of data set.
       C: Merge order constraints.
Output: M: Minimum Transitive dissimilarity matrix closure of G.
Init: M = G.
 1: for $k \leftarrow 0$ to $N$ do
 2:     for $i \leftarrow 0$ to $N$ do
 3:         for $j \leftarrow 0$ to $N$ do
 4:             for all $c = (x_i, x_j, x_l)$ do
 5:                 $minCon = \min(minCon, d(x_i, x_l));$
 6:             end for
 7:             $m_{ij} = \min\{m_{ij}, \max(m_{ik}, m_{kj}), minCon\};$
 8:         end for
 9:     end for
10: end for
11: return $M$ ;
```

**Algorithm 3:** Modified Floyd-Warshall algorithm to compute the minimum transitive dissimilarity of weighted graph G

## 4.4 Experiments

In this section, we conduct experiments on various datasets to evaluate our proposed semi-supervised hierarchical clustering framework. We compare our proposed techniques in Section 4.3.2 (the iterative projection algorithm (IPoptim) and the transitive dissimilarity transformation algorithm (UltraTran)) with two baseline algorithms: the standard agglomerative hierarchical clustering(HAC) without constraints and the constraint-based HAC (denoted as HACoc) proposed in [ZQ10].

| Name | # samples | # attributes | # classes |
|---|---|---|---|
| Iris | 150 | 4 | 3 |
| Wine | 178 | 13 | 3 |
| Protein | 116 | 20 | 6 |
| Ionosphere | 351 | 34 | 2 |
| CSTR | 475 | 1000 | 4 |
| Log | 1367 | 1000 | 8 |
| WebACE | 2340 | 1000 | 20 |
| Reuters | 2900 | 1000 | 10 |

Table 4.3: Dataset descriptions

| dataset | Algorithm | FScore | Time |
|---|---|---|---|
| Iris | HAC | 0.8906 | 107 |
| | HACoc | **0.96** | 233694 |
| | IPoptim | 0.9293 | 18917 |
| | UltraTran | 0.9211 | **18490** |
| Wine | HAC | 0.7614 | 109 |
| | HACoc | **0.9346** | 573002 |
| | IPoptim | 0.86 | **30034** |
| | UltraTran | 0.8456 | 32636 |
| Protein | HAC | 0.4669 | 196 |
| | HACoc | **0.5131** | 53580 |
| | IPoptim | 0.4730 | 8889 |
| | UltraTran | 0.4669 | **8342** |
| Ionosphere | HAC | 0.7401 | 361 |
| | HACoc | 0.7446 | 1392259 |
| | IPoptim | **0.7503** | 270198 |
| | UltraTran | 0.7501 | **251164** |

Table 4.4: Performance comparison on 4 small datasets.

Table 4.3 shows the summary of the datasets used in the experiments. We use 8 datasets with the number of classes ranges from 2 to 20, the number of samples ranges from 116 to 2900 and the number of dimensions ranges from 4 to 1000. The details of the datasets are: (1) Four datasets (Ionosphere, Iris, Protein and Wine) are from UCI data repository [BM98]. (2) Four datasets (CSTR, Log, Reuters, WebACE) are benchmark text datasets for document clustering. Each document is represented as a term vector using vector space model. All document datasets are pre-processed by removing the stop words and unnecessary tags and headers. More information of the datasets can be found in [LD08b].

| dataset | Algorithm | FScore | Time |
|---|---|---|---|
| CSTR | HAC | 0.653 | 784 |
| | HACoc | 0.6524 | 4911106 |
| | IPoptim | **0.6632** | 577451 |
| | UltraTran | 0.6631 | **570320** |
| Log | HAC | 0.8871 | 3255 |
| | IPoptim | **0.9001** | 1.984e+7 |
| | UltraTran | 0.8973 | **1.9698e+7** |
| WebACE | HAC | 0.5471 | 19580 |
| | IPoptim | 0.5492 | **1.0081e+8** |
| | UltraTran | **0.5514** | 1.0090e+8 |
| Reuter | HAC | 0.6154 | 33000 |
| | IPoptim | **0.6187** | **1.7682e+8** |
| | UltraTran | 0.6178 | 1.7694e+8 |

Table 4.5: Performance comparison on 4 large datasets.

### 4.4.1 Evaluation measures

All the eight datasets have data labels which will be used in clustering performance evaluation. The accuracy of a hierarchical clustering is evaluated by considering the entire hierarchy [ZK02b]. A single cut on the hierarchy produces a possible partition of the data set and such partition can be measured by FScore proposed in [LA99]. Supposing $G_i$ is one of the clusterings generated by cutting on the hierarchy $H$ and $D_j$ is a group of data sharing the same label over $L$ classes, then

$$FScore(G_i, D_j) = \frac{2 * Recall(G_i, L_j) * Precision(G_i, L_j)}{Recall(G_i, L_j) + Precision(G_i, L_j)}. \tag{4.7}$$

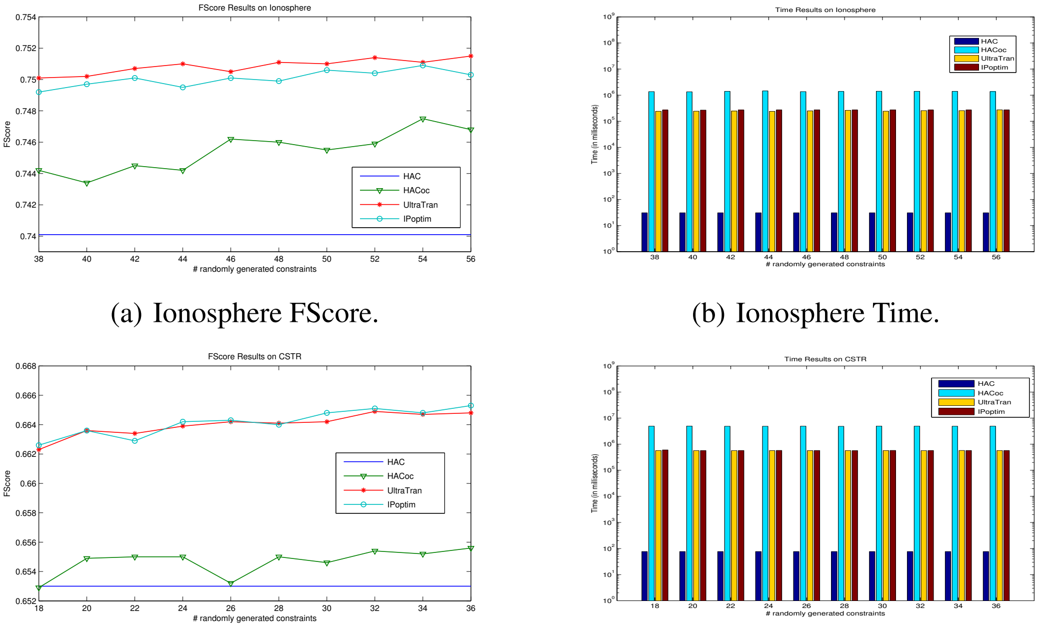
The FScore of group $G_i$ is defined as the maximum FScore over all $L$ classes

$$FScore(G_i) = \max_{j \in L} FScore(G_i, D_j). \tag{4.8}$$

For a hierarchical clustering with $|D|$ samples. Totally $N = \frac{(1+|D|)*|D|}{2}$ possible groups can be generated by cutting at different levels. The FScore defined on the entire hierarchy is computed as the weighted sum of each group's FScore:

$$FScore(H) = \sum_{i=1}^{N} \frac{|G_i|}{|D|} FScore(G_i). \tag{4.9}$$

We also compare the running time of different algorithms. The running time is recorded at milliseconds (1/1000s).



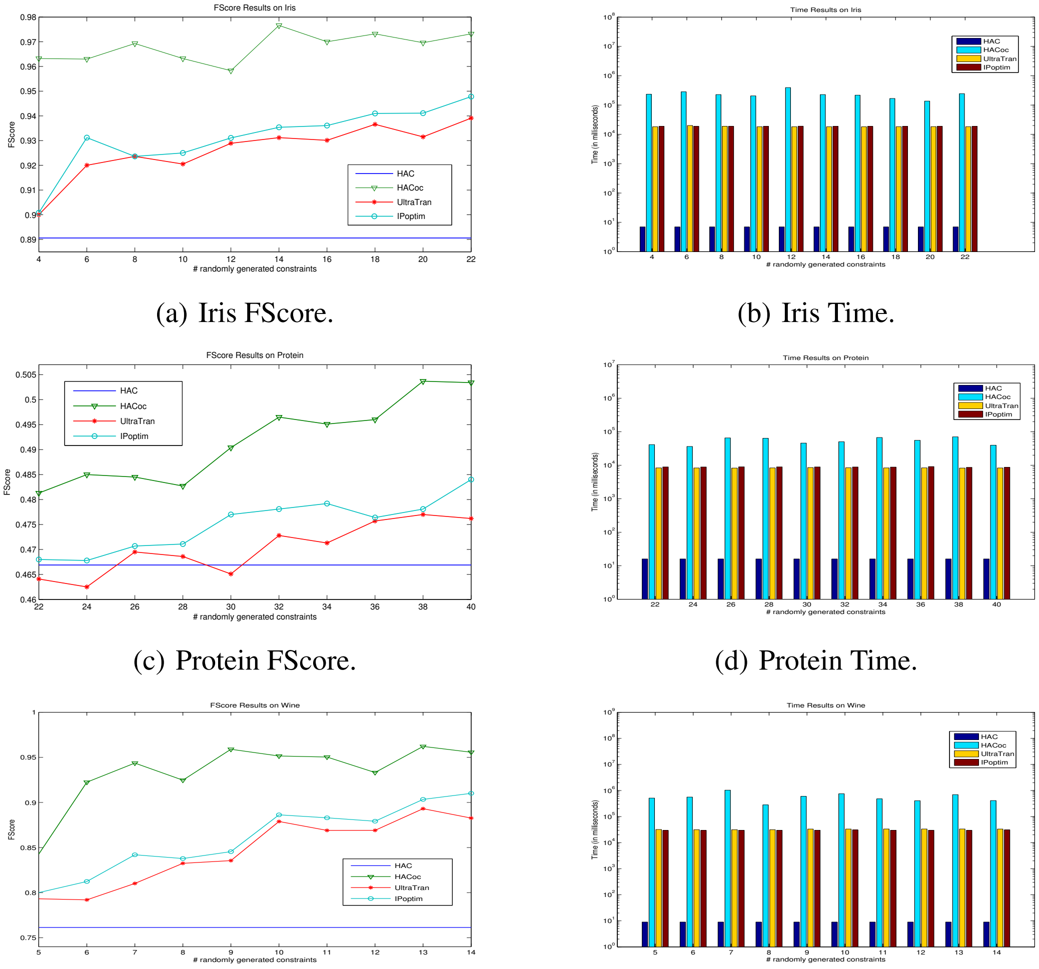
(a) Ionosphere FScore.

(b) Ionosphere Time.

(c) CSTR FScore.

(d) CSTR Time.

Figure 4.5: Results on Ionosphere and CSTR datasets. The performance as a function of the number of constraints.



(a) Iris FScore.

(b) Iris Time.

(c) Protein FScore.

(d) Protein Time.

(e) Wine FScore.

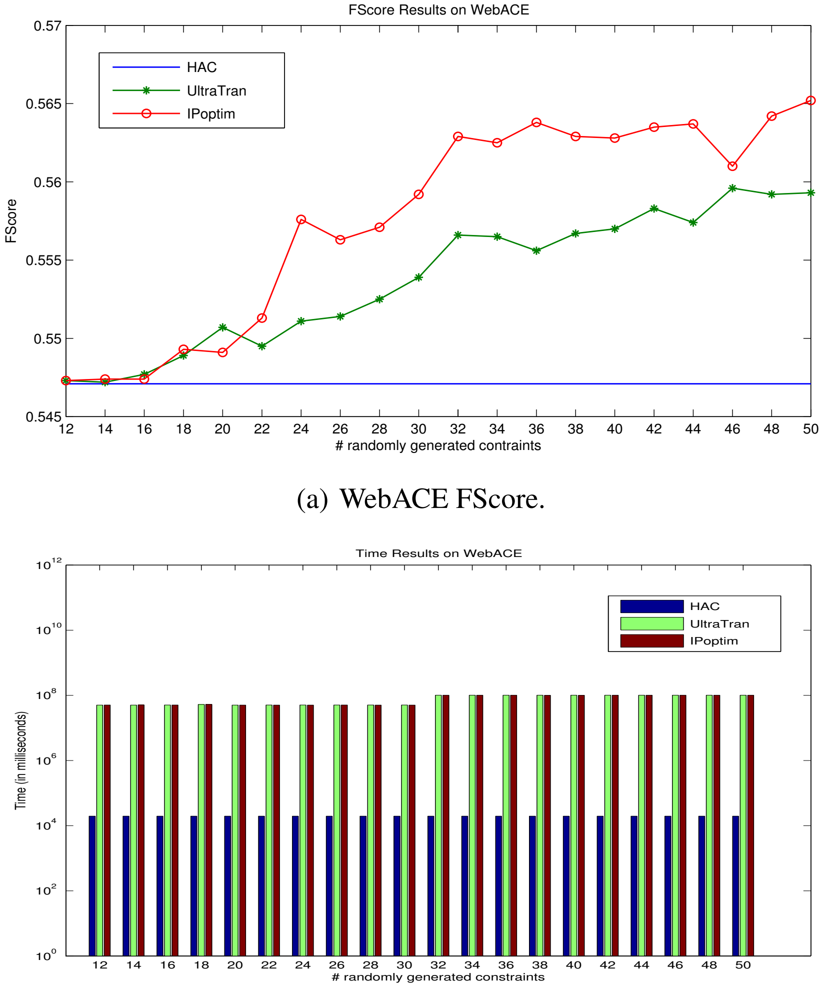
(f) Wine Time.

Figure 4.6: Results on Iris, Protein, and Wine datasets. The performance as a function of the number of constraints.

### 4.4.2 Experiment setup

According to the given class label of each sample, we randomly select three samples from two different classes to generate a constraint. For example, if $x_i, x_j \in Class1$ and $x_k \in Class2$, then $c = (x_i, x_j, x_k)$ is a relative constraint. So each generated constraint is based on the actual class label information and should reflect the domain knowledge. As a result, we can expect the clustering performance should be generally improved when these constraints are utilized. In our experiments, the reported results are computed by averaging

10 runs. For the first five small datasets (the number of samples $\leq$ 1000), we randomly generate 100 constraints for each run. For the other three large datasets, we randomly generate 200 constraints for each run. All constraint sets are preprocessed to eliminate the conflicts. The experiments are conducted under the environment of Linux 2.6 plus 8 Intel(R) Xeon(R) CPU E5420 2.50GHz and 16 GB of RAM.



(a) WebACE FScore.

(b) WebACE Time.

Figure 4.7: Results on WebACE dataset. The performance as a function of the number of constraints.

### 4.4.3 Result analysis

The experimental results are shown in Table 4.4 and 4.5. Note that the running time of HACoc is much longer than the other algorithms, especially on large datasets. So we do not include the results of HACoc on Log, WebACE and Reuter datasets for comparison. From Table 4.4 and 4.5, we observe that:

- By incorporating relative constraints, semi-supervised hierarchical clustering outperforms hierarchical clustering without constraints. In all datasets, the Fscore values of HAC are consistently lower than those of other semi-supervised hierarchical clustering frameworks with constraints. The performance improvement is significant on Iris, Wine and Protein as shown in Figure 4.6.
- Although HACoc achieves the best clustering performance on three small datasets (Iris, Wine, Protein), it is not efficient and needs long execution time.
- Our proposed techniques (IPoptim and UltraTran) are much more efficient than HACoc. In terms of clustering performance, IPoptim and UltraTran outperform HACoc on Ionosphere and CSTR datasets as shown in Figure 4.5.
- In general, IPoptim outperforms UltraTran in clustering performance while UltraTran is more efficient than IPoptim.

To further investigate the performance of semi-supervised hierarchical clustering, we conduct experiments by varying the number of relative constraints. Figure 4.6, Figure 4.5 and Figure 4.7 plot the clustering performance and execution time as a function of the number of constraints on six datasets (Iris, Protein, Wine, Ionosphere, CSTR, and Reuter). Note that the computation time of the algorithms does not increase much as the number of constraints increases. We also observe that the performance enhancement obtained by the semi-supervised clustering is generally greater as the number of constraints increases. However, the performance is not monotonically increasing with the number of constraints.

There are two possible reasons. First, our proposed framework is not aiming to satisfy all constraints but to find a good approximation of the constrained ultra-metric. Second, the clustering performance is also depending on the quality of generated constraints. It is intuitive to say that not all constraints have the same importance to the performance of semi-supervised hierarchical clustering. And the constraints we applied are directly generated from the instance similarities and the true class labels. How to discover important constraints would be a valuable consideration in our future work.

## 4.5 Conclusion

In this chapter, we propose a semi-supervised hierarchical clustering framework based on ultra-metric dendrogram distance. The triple-wise relative constraints are introduced, particularly for hierarchial clustering, to describe the merge preference among instances. Two techniques are developed to solve semi-supervised hierarchical clustering problem. The optimization-based technique minimizes the distance between the original dissimilarity matrix and the target ultra-matrix using the ultra-metricity and relative constraints. The transitive dissimilarity based technique takes those relative constraints into the ultra-metric transformation process. Experiments are conducted to demonstrate the effectiveness and efficiency of our proposed methods.

CHAPTER 5

# Topic I: taxonomy generation and information extraction

To efficiently generate domain taxonomy, traditional vertical search engines can be used to identify domain-related web resources and contents. In order to maintain a effective domain taxonomy, the following three tasks should be considered:

- *Automatically generate domain taxonomy*: Taxonomy has been widely used as a structured organization of domain knowledge. However, most taxonomy is generated manually and cost lots of human efforts. Also, since the information on the internet changes and evolves very fast and hard to be predicted, the developed taxonomy is easy to become stagnant, manually rebuilding the taxonomy will bring more inefficiency under such circumstances. Thus, to automatically generate taxonomy becomes crucial means to deal with the challenging environment.

- *Develop the crawling strategy*: A focused crawler is a web crawler that collects Web pages. Those collected page should satisfy some specific properties by carefully prioritizing the crawl frontier and managing the hyperlink exploration process [Cha09]. A good crawling strategy is able to predict the probability that an unvisited page will be relevant before actually downloading the page. Besides a reasonable prediction algorithm, the performance of a focused crawler is also relevant to the richness of links in a specific topic.

- *Identify entity information from documents*: The crawled on-topic contents are usually un-structured texts. To identify meaningful terms or elements in a domain, entity-recognition techniques are necessary and often the best choice to perform such task. In addition, each entity in a document has contextual information that would be very useful to help people to better understand the document. Therefore, methods to transform the unstructured text into structured formats associated with important entities is highly expected.

In the following sections, we will give details about two approaches which accomplish the above tasks in our research. Taxonomy Generation in Section 5.1 demonstrates a hierarchical clustering with constraints to generate domain taxonomy from crawled documents; A focused crawler implemented to identify domain-specific web resources is discussed in Section 5.2; Section 5.3 shows an actionable information extraction process using information extraction and natural language processing techniques to transform unstructured text into structured contents. Section 5.4 concludes the chapter.

## 5.1 Taxonomy generation

Taxonomy plays significantly important role in most knowledge-based information management systems applied in various application domains. They are designed to provide structurally organized terminologies that are formal, application-independent and with common agreement within a community of practice [Sim09, LWSL10]. However, generating taxonomy from the scratch suffers high-cost, low-efficiency problem. Ensembling several existing taxonomies or incrementally integrating new concepts into existing taxonomy becomes effective and well-accepted approach for taxonomy generation and reuse.

### 5.1.1 Base concept hierarchy generation

There are several taxonomy generation techniques having been implemented as we discussed in Chapter 2. However the performance of those state-of-art techniques can be improved by considering the following two approaches according to my preliminary research: Ensemble multiple hierarchies discussed in Chapter 3 and Constraints-based hierarchy generation discussed in Chapter 4. In our taxonomy generation component, we model this problem as document hierarchical clustering with ordered constraints in which the constraints are given as a partially known hierarchy, the sister related concepts extracted from web documents are treated as instances, and our goal is to build a term hierarchy which satisfies the relative hierarchical structure in given partial hierarchy.

Our initial disaster taxonomy is built manually from the scratch. For example, based on our long cooperation with Miami-Dade Emergency Operational Center (EOC), we extracted thousands of frequent terms in its official announcements and situation reports in the past 5 years. We reasonably assume that those terms with high frequency indicate important concepts in disaster domain. Through careful filtering and organizing those terminologies from our staff and developers, our initial disaster taxonomy is obtained and then verified by our domain experts.

### 5.1.2 Iterative taxonomy generation

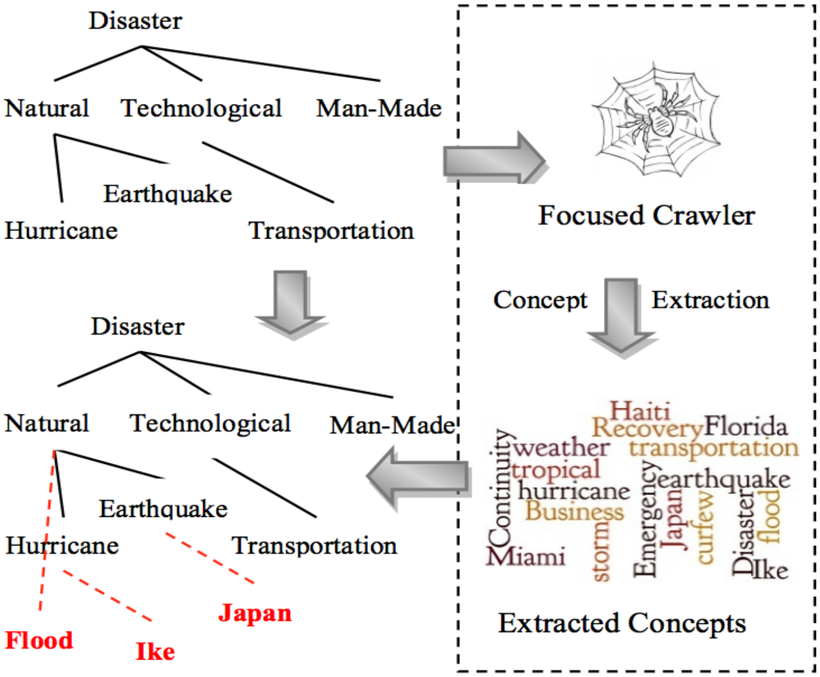


Figure 5.1: Iterative taxonomy generation.

The taxonomy generation process follows an interactive and iterative strategy. The focused crawler utilizes the taxonomy to classify accessed web pages and prioritizes those pages with highest relevance to disaster domain. From the repository of collected data, high quality data will be analyzed and disaster-related concepts without being mentioned in the existing taxonomy, will be extracted. Those extracted concepts are considered as highly popular terms that can extend and enrich the existing taxonomy. After integrating those newly- discovered concepts into existing taxonomy, domain experts can verify the updated knowledge based and provide valuable feedback.

Given the domain of disaster management, Figure 5.1 illustrates the flow of how we interactively construct an domain texonomy. The focused crawler utilizes the taxonomy to classify accessed web pages and prioritizes those pages with highest relevance to disaster domain. From the repository of crawled data, high quality data will be analyzed and disaster-related concepts without being mentioned in the existing taxonomy, will be extracted. Those extracted concepts are considered as highly popular terms that can extend and enrich the existing taxonomy. After integrating those newly discovered concepts into disaster taxonomy, domain experts can verify the updated knowledge based and provide valuable feedback.

### 5.1.3 Hierarchical clustering with constraints

Our developed technique is to build a hierarchical structure to model the basic human understanding of the relationships among disaster relevant concepts. A basic taxonomy/concept hierarchy is given at the very beginning of the generation process. In our work, we use agglomerative hierarchical clustering with constraint to algorithmically integrate newly-discovered terms or concepts into the existing ones.

### Problem definition

All concepts in existing taxonomy are denoted as $T = \{t_1, t_2, \cdots, t_n\}$ and the newly-discovered concepts are denoted as $C = \{c_1, c_2, \cdots, c_n\}$. $H$ is the existing concepts hierarchy formed by terms from $T$. Our goal is to generate an updated concept hierarchy $H'$ that is formed by all terms from both $T$ and $C$. The integration of $T$ and $C$ is non-trivial. There are three important aspects worth mentioning:

1. Each concept in $T$ or $C$ is represented by a set of terms extracted from the web documents repository. So, essentially there is a subset of web documents under each concept.
2. $H$ is essentially a hierarchical clustering on all documents. The hierarchy of the concepts reflects the inclusion or exclusion of documents sets. There is no partial overlap between document sets under different concepts.
3. There is a merging preference/order for each pair of concepts in both $H$ and $H$ which indicates the level of closeness between two document sets. The new concepts in $C$ should not change the relative merging order of existing concepts in $T$. The details are given in the following section.

### Algorithm and partial hierarchy constraint

The merging preferences mentioned above are modeled as relatively ordered constraints when performing hierarchical clustering on document set. Constraints defined in hierarchical clustering are different from constraints, such as instance-level constraints [WCRS01] and prior knowledge [LZS09] in partitional clustering. Several types of constraints that can be applied in hierarchical clustering are defined in the literature [BN08a, WCRS01, ZL11].

In our application, we use Bades algorithm [BN08a] to refine the given disaster concept hierarchy by considering further extracted concepts. The constraint in [BN08a] is named must-link-before (MLB), shown in Figure 5.2, which specifies the order in which objects are linked. When applied to concept hierarchy, such order indicates the merge preference between concepts (document sets). Bades algorithm [BN08a] can utilize the existing concept hierarchy as partially known hierarchy and update it by directly attaching newly-discovery concepts to previous hierarchy. The other two methods do not meet our needs because updated hierarchy requires to be built from the scratch.

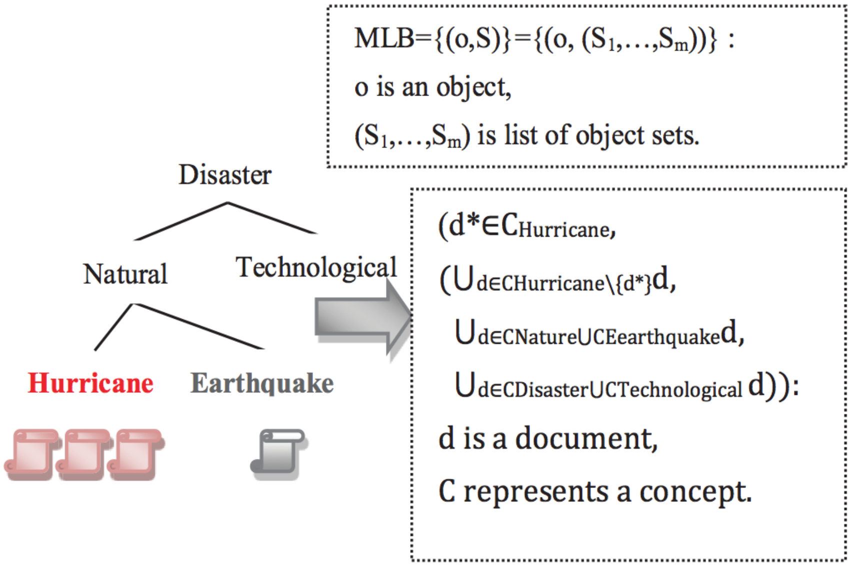


Figure 5.2: Iterative taxonomy generation.

## 5.2 Focused crawler

We adopt focused crawling technique to retrieve the disaster aware information in the Web. In addition, contents also come from subscription of some local news feeds and monitoring announcement from government sites. Compared with a standard focused crawler defined in [CvdBD99, AAGY01], there are some challenges when applied crawling in a specific domain.

Loose cohesion: Except when there is a situation happens which generates massive information in a short time, most domain-relevant information is scattered in the Web. In news websites, stories about an event sometimes embed in other types of news.

Diversity of topic: A common topic often includes many subtopics, for example, disaster domain can be divided into various types of disasters. It is difficult to evaluate a web pages relevance on a consistent scale among all these subtopics. It is very likely that the crawled data will bias towards some of the subtopics and leave some others uncovered. To address the above issues, we utilize the taxonomy we developed.

The following description will take diaster management as a scenario to better explain the application of our approaches.

### 5.2.1 Selection strategy

Best-first approaches are widely used by focused crawlers, selecting the next page to be crawled from all currently assessed candidate page URLs by their scores as

$$l^* = \arg\max_{l \in queue} score(l)$$

where $score(l)$ is calculated based on a classifier indicating whether or not the URL $l$ belongs to the topic. However, the best may bias to some of the subtopics of general disaster topic because of the unbalance of these subtopics and a limited initial training dataset. To get a set of web pages with high diversity for a specific disaster, we simultaneously crawl web pages for each disaster concept based on the concept hierarchy. Our selection strategy considers a disaster concept:

$$l_c^* = \arg\max_{l \in queue} score(l, C),$$

that is, for each disaster concept, select the next page to be crawled from all currently assessed candidate page URLs according to their scores with respect to the concept.

### 5.2.2 Prioritization based on concept relationship

For a web page, instead of classifying it into “Disaster” and “Non-disaster”, we assigned to it a concept in our concept hierarchy, such as “weather”, “government” and “environment protection”. These disaster related concepts increase the coherence of the Web pages of disaster topic, playing a role of bridging between pages of different sites of disaster concepts and pages of different disaster concepts. To calculate the prioritization score of a URL, the concept of the page from which the URL is linked is utilized as follows:

$$score(l, C_d) = P(C_i^* \to C_d) * P(page_l = C_i^*),$$

where $P(page_i = C_i^*)$ is the output of our content classifier indicating the probability the page where the link $l$ is linked from belongs to its optimal concept $C_i^*$ , and $P(C_i \to C_d)$ is the link relationship between concepts, the probability that a page of concept $C_i$ links to a page of concept $C_d$. It can be calculated as

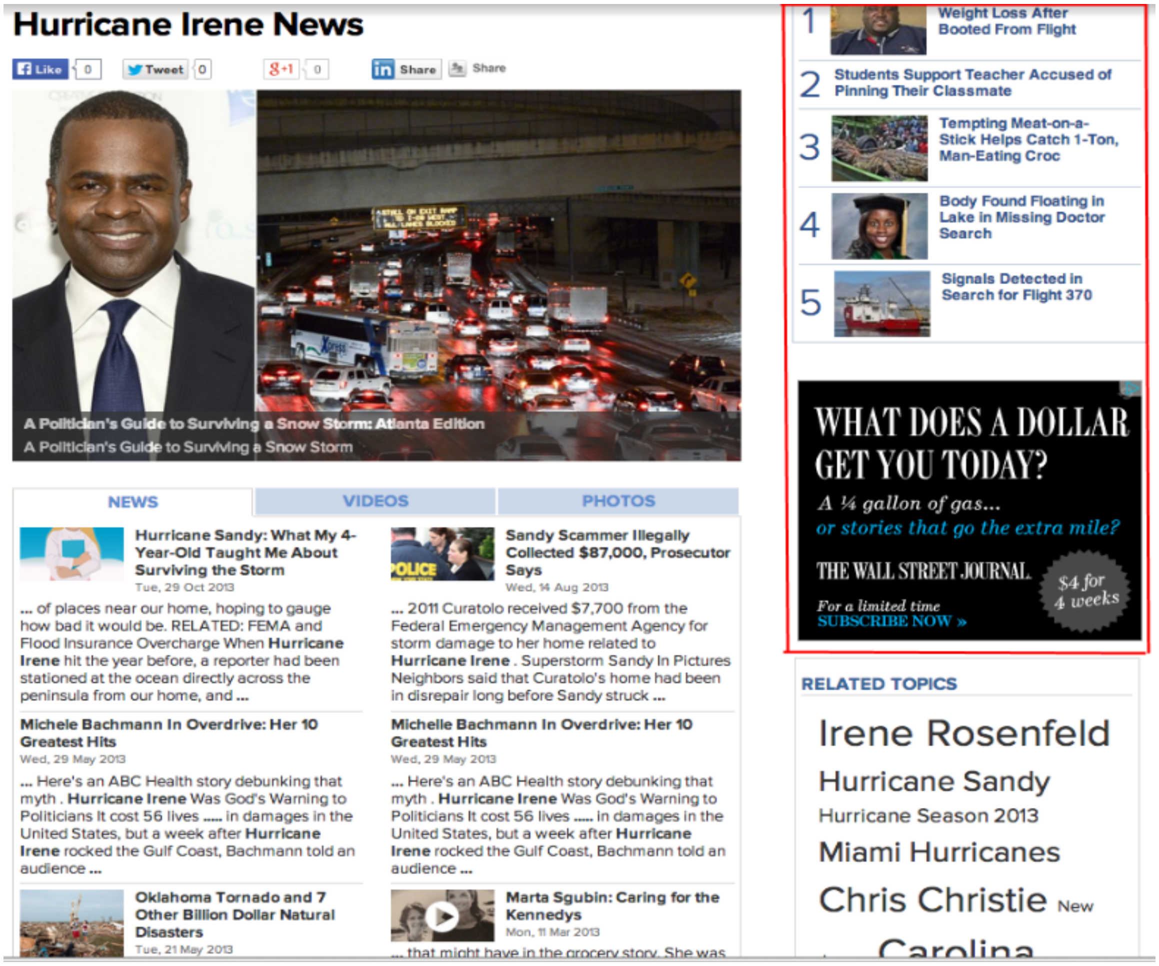
$$P(C_i \to C_d) = \frac{\sum_{p \in C_i} |L_{p,fetched} \bigcap C_d| + \lambda}{\sum_{p \in C_i} |L_{p,fetched} \bigcap C_d| + \lambda \cdot \sum_{p \in C_i} |L_{p,unfetched}|},$$

the ratio of the number of links classified as $C_d$ from pages of $C_i$ to the number of all fetched links from pages of $C_i$, with a Dirichlet smoothing using un-fetched links. Note that with the process of crawling, $P(C_i \to C_d)$ is being updated, so that the scoring of links is also adaptive with more data crawled.

### 5.2.3 Link prediction

Although a page is disaster relevant, the links of the page may not necessarily lead to other pages of disasters. Figure 5.3 shows an example news page about hurricane Irene in which links in red block is irrelevant to hurricane information.

To further distinguish the links in a page, a link classifier is trained, using the prediction of the content classifier for crawled pages as training data. The rationale is that many links



Hurricane Irene News

A Politician's Guide to Surviving a Snow Storm: Atlanta Edition

A Politician's Guide to Surviving a Snow Storm

NEWS | VIDEOS | PHOTOS

**Hurricane Sandy: What My 4-Year-Old Taught Me About Surviving the Storm**
Tue, 29 Oct 2013

... of places near our home, hoping to gauge how bad it would be. RELATED: FEMA and Flood Insurance Overcharge When **Hurricane Irene** hit the year before, a reporter had been stationed at the ocean directly across the peninsula from our home, and ...

**Michele Bachmann In Overdrive: Her 10 Greatest Hits**
Wed, 29 May 2013

... Here's an ABC Health story debunking that myth . **Hurricane Irene** Was God's Warning to Politicians It cost 56 lives ..... in damages in the United States, but a week after **Hurricane Irene** rocked the Gulf Coast, Bachmann told an audience ...

**Oklahoma Tornado and 7 Other Billion Dollar Natural Disasters**
Tue, 21 May 2013

**Sandy Scammer Illegally Collected $87,000, Prosecutor Says**
Wed, 14 Aug 2013

... 2011 Curatolo received $7,700 from the Federal Emergency Management Agency for storm damage to her home related to **Hurricane Irene** . Superstorm Sandy In Pictures Neighbors said that Curatolo's home had been in disrepair long before Sandy struck ...

**Michelle Bachmann In Overdrive: Her 10 Greatest Hits**
Wed, 29 May 2013

... Here's an ABC Health story debunking that myth . **Hurricane Irene** Was God's Warning to Politicians It cost 56 lives ..... in damages in the United States, but a week after **Hurricane Irene** rocked the Gulf Coast, Bachmann told an audience ...

**Marta Sgubin: Caring for the Kennedys**
Mon, 11 Mar 2013

... that might have in the grocery story. She was

1 Weight Loss After Booted From Flight

2 Students Support Teacher Accused of Pinning Their Classmate

3 Tempting Meat-on-a-Stick Helps Catch 1-Ton, Man-Eating Croc

4 Body Found Floating in Lake in Missing Doctor Search

5 Signals Detected in Search for Flight 370

RELATED TOPICS

Irene Rosenfeld
Hurricane Sandy
Hurricane Season 2013
Miami Hurricanes
Chris Christie New
Carolina

Figure 5.3: An example page of hurricane Irene.

contain a description of the content of the linked page. Another observation we find is about link structure, that for a pair of link which are in the sibling nodes of the HTML DOM tree, e.g. in a list of the page, they tend to be of a similar topic. We follow the work of [16] and build a link classifier based on Native Bayes. To apply the link prediction:

1. The prioritization score can be extended as: $score(l, C_d) = P(C_i^* \rightarrow C_d) * P(page_l = C_i^*) * P(C_d|l)$, where $P(C_d|l)$ is the output of the link classifier, probability that link l leads to a page under concept $C_d$.

2. To reduce the redundancy, we first divide the links into clusters, and constrain the

crawler such that links in the same cluster are not fetched at same time. Once a link is fetched, the prediction of links in the same cluster will be updated.

### 5.2.4 Architecture of the focused crawler

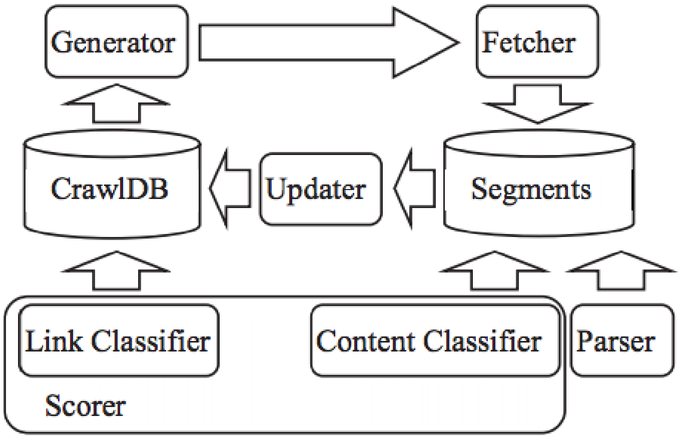


Figure 5.4: Focused crawler architechture.

We build our crawler based on Nutch[1], which is a distributed general crawling tool running on Hadoop[2] clusters. We customize the scoring module and generator module in Nutch. The current architecture is shown in Figure 5.4. In each iteration, the Fetcher fetches page content of a list of URLs, and stores them as a segment. The updater updates CrawDB, where the crawled data is associated with a URL. The scoring module assigns a prioritization score to each URL indicating the importance of the URL. The generator module generates a set of URL, covering all disaster concepts in the concept hierarchy. The Fetcher fetches the web page content.

[1] Apache Nutch. https://nutch.apache.org

[2] Apache Hadoop/ http://hadoop.apache.org

## 5.3 Actionable information extraction

Actionable information means having the necessary information immediately available in order to deal with the situation at hand [3]. There are several important factors for an actionable information: entity, time, location, and status. Those information can direct to immediate answers to typical questions like, when, how, where. We define actionable information as a triple relationship, <entity, time, status>revealing the status information of the entity at a certain time, which needs to be extracted from given text set. Actionable information extraction not only provides ways to identify those important factors from unstructured text but also automatically build relations between those factor to transform unstructured text to structured records.

| Time: October 21, 2005 12:30 p.m. |
|---|
| Miami-Dade Emergency Operations Center is currently activated at a level II and officials and emergency managers are carefully monitoring Hurricane Wilma.<br>Residents are urged to finalize their personal hurricane preparations.<br>On Monday, October 24, Miami-Dade County offices, public schools, and courts will be closed.<br>Currently, transit bus and rail service continues, including Metrobus, Metrorail and Metromover.<br>Miami International Airport is open. However, if you have travel plans please check with your airline for flight information.<br>Tomorrow afternoon, the American Red Cross will open hurricane evacuation centers for residents who do not feel safe in their homes or live in low-lying areas. |

Table 5.1: An example of EOC report.

[3] http://en.wikipedia.org/wiki/Actionable_information_logistics

### 5.3.1 Structured information extraction from text

The BCIN system is an information sharing system for companies and government agencies. To provide a user-friendly interface to all these users, we do not request a unified format for them to submit the reports. Instead, we use information extraction methods to integrate reports from different sources. For example, Table 5.1 shows an example of EOC reports.

For any events, the key information is “What was/is/will be the status of Facilities/Services/ at the time of ”. From the EOC reports, we need to extract such information in the form of a triple: (entity, time, status), which reveals the status information of the entity at a certain time. Take disaster management domain as an example, the entity in reports may be a facility or public service like “Miami International Airport”, “schools”, “bus”, and an order like “curfew”. If the entity is referred to an order, the triple means whether the order is in effect or not at that specific time. We extract these triples through two steps: first, we extract entities and time expressions, then, we classify a pair of (service, time) to a proper category, “no relation”/ “open” / “close” / “unclear”. We assume that the information of one event will not span on different sentences, so we process every sentence individually to extract an event. To extract those triples, both entity and relation extraction will be performed.

| Time: October 21, 2005 12:30 p.m. |
|---|
| Miami-Dade Emergency Operations Center is currently<br>activated at a level II and officials and emergency managers<br>are<br>carefully monitoring Hurricane Wilma.<br>Residents are urged to finalize their personal hurricane<br>preparations.<br>On $\langle T\rangle$Monday, October 24 $\langle /T\rangle$, $\langle E\rangle$Miami-Dade County<br>offices$\langle /E\rangle$, $\langle E\rangle$public schools$\langle /E\rangle$, and $\langle E\rangle$courts$\langle /E\rangle$<br>will be closed.<br>$\langle T\rangle$Currently$\langle /T\rangle$, $\langle E\rangle$transit bus$\langle /E\rangle$ and $\langle E\rangle$rail<br>service$\langle /E\rangle$ continues, including $\langle E\rangle$Metrobus$\langle /E\rangle$,<br>$\langle E\rangle$Metrorail$\langle /E\rangle$ and $\langle E\rangle$Metromover$\langle /E\rangle$.<br>$\langle E\rangle$Miami International Airport$\langle /E\rangle$ is open. However, if<br>you<br>have travel plans please check with your airline for flight<br>information.<br>$\langle T\rangle$Tomorrow afternoon$\langle /T\rangle$, the American Red Cross will<br>open $\langle E\rangle$hurricane evacuation centers$\langle /E\rangle$ for residents<br>who<br>do not feel safe in their homes or live in low-lying areas. |

Table 5.2: Entity extraction result of the report in Table 5.1.

### 5.3.2 Entity extraction

For each report, sentence segmentation is conducted first, and each sentence is POS-tagged. To extract entities and time expressions, we manually label some news and train a linear chain conditional random fields (CRF) model to tag all words of sentences, using “BIO” annotation [LMP01, SP03]. A word tagged as [TYPE-B]/[TYPE-I] means it is the beginning/continuing word of the phrase of the TYPE, and the ones tagged as O means it is not in any phrase. Here TYPE can be E or T, referring to the entity and time expression. Using CRF, given the sentence X, the probability of its tags Y is as follows:

$$p(Y|X) = \frac{1}{Z_X}\exp(\sum_{i,k}\lambda_k f_k(y_{i-1}, y_i, X) + \sum_{i,l}\mu_l g_l(y_{i-1}, y_i, X)),$$

where $Z_x$ is the normalization constant that makes the probability of all state sequences sum to one; $f_k(y_{i-1}, y_i, X)$ is an arbitrary feature function over the entire observation sequence and the states at positions $i$ and $i-1$ while $g_l(y_{i-1}, y_i, X)$ is a feature function of the states at position i and the observation sequence; $k$ and $k$ are the weights learned for the feature functions $f_k$ and $g_l$ , reflecting the confidence of feature functions by maximum likelihood procedure. The most probable labels can be obtained as

$$Y^* = \arg\max\nolimits_Y P(Y|X),$$

by Viterbi-like dynamic programming algorithm [LMP01]. We use for features the local lexicons and POS tags, and plus the dictionary composed of the existent entity names in the database. Table 5.2 shows the result of the entity extraction result of the report in Table 5.1.

### 5.3.3 Relation extraction

| Feature Name | Description |
|---|---|
| DistanceBetween(e,t) | number of words between entity and time |
| WordBetween(e,t) | what is the words between entity and time |
| TenseOfSentence(e.t) | the tense of the sentence |
| NegativeVerbsInSentence(e,t) | number of negative verbs in the sentence |
| PositiveVerbsInSentence(e,t) | number of positive verbs in the sentence |
| ContainDate(t) | whether the sentence contains time |
| PrepositionBefore(t) | what is the preposition |
| FromDocument(t) | document the sentence belongs |

Table 5.3: Features used to classify whether the entity e is associated with the time expression t.

If a sentence contains an entity but no time expression, the time associated with the report will attached to the end of the sentence. To generate the triple by connecting the entity with the time expression with a proper status label, we train a multi-category SVM [HL02] to classify each pair of (entity, time) to a proper category, "no relation"/ "open"/ "close"/ "unclear". Table 5.3 shows the features we used for classification. Among them, TenseOfSentence(s,t), NegativeVerbsInSentence(s,t), and PositiveVerbsInsentence(s,t) are extracted by heuristic rules to indicate the tense of the sentence, the verbs with and without negative modifier semantically in the sentence, respectively. Note that FromDocument(t) indicates whether the time is the time associated with document or not.

Finally, we extract those pairs of entity and time expression in the "open" or "close" categories to form the triple. Meanwhile the time expressions are formatted into an absolute form of expression from relative time expression such as "next Monday", "this afternoon" and etc. using the time of report as a benchmark. The structured information extracted from the report in Table 5.1 is shown in Table 5.4.

| Service | Time | Status |
|---|---|---|
| Miami-Dade County offices | October 24, 2005 | close |
| public schools | October 24, 2005 | close |
| courts | October 24, 2005 | close |
| transit bus | October 22, 2005 6:30 p.m. | open |
| Rail service | October 22, 2005 6:30 p.m. | open |
| ⋮ | ⋮ | ⋮ |
| Miami International Airport | October 22, 2005 6:30 p.m. | open |
| hurricane evacuation centers | October 23, 2005 afternoon | open |

Table 5.4: Information extracted from the EOC report shown in Table 5.1.

## 5.4 Conclusion

In this chapter, we discussed the design and implementation of a focused crawler by utilizing domain taxonomy. To generate domain taxonomy, enhanced information extraction techniques are used to identify meaningful domain-specific terms. Those terms can be used to enrich exiting taxonomy. Such iterative process is conducted as the crawler is running, so the taxonomy can be refreshed. Both crawling and extracting components are using the enhanced hierarchical clustering techniques (discussed in Chapter 3) and considering constraints transformed from domain knowledge (discussed in Chapter 4).

The system can be improved in both accuracy and efficiency. In particular, the time cost of taxonomy generation can be further reduced by utilizing distributed frameworks or using different agglomeration strategy to quickly generate similar hierarchies. We also plan to incorporate the dynamic seed selection components into the crawler to maintain more authoritative and informative seed sites.

CHAPTER 6

# Topic II: modeling user interest and recommendation

Web search is generally motivated by an information need. Since asking well-formulated questions is the fastest and the most natural way to obtain information for human beings, almost all the queries posed to search engines correspond to some underlying questions, which represent the information need. Accurate determination of these questions may substantially improve the quality of search results and usability of search interfaces. Moreover, in case of imprecise or ambiguous queries, automatically generated questions can naturally engage the users into feedback cycles to refine their information need and guide them towards their search goals. Implementation of the proposed strategy raises new challenges in content indexing, question generation, ranking and feedback.

## 6.1 User profile modeling using hierarchical ensemble clustering

News recommendation has becoming one of the most important applications for major content providers, such as *Google News* and *Yahoo! News*. It deals with the information explosion problem which prevents readers from obtaining the most important information. Recommendation services can largely improve the efficiency and accuracy of information acquiring, and recommender systems are designed to filter the critical news, key events and meaningful items. However, such crucial information cannot be of the same importance for the global set of users. The news personalization and localization have emerging quickly to fit the reading preferences for each individual and geographical region respectively. Therefore, a high-quality news recommender system should be able to provide personalized "important" news reading lists according to each user's preference.

News personalization has been extensively studied from many perspectives. However, there are three major issues remaining challenging in personalized recommendation task: 1) how to capture each **user's reading interest** according to his/her historical consumption? 2) how to model the **relations between news content and user profiles** for better reading preference matching? and 3) how to make good predictions with regard to the **quality and diversity** of the recommended result?

In this chapter, we develop a novel PErsonalized NEws recommendaTion framework using ensemble hieRArchical clusTEring (*PENETRATE*) to systematically address the aforementioned issues in news recommendation. *PENETRATE* captures users' preferences based on not only individual user's reading history, but also the *historical consumptions of a group of users with similar reading preferences* based on the fact that each user group has its unique preferences to different news topics. Furthermore, the profile of a given user group is not represented using the traditional vector space model, but is characterized by a *news hierarchy* in which the merged preference between pair of new articles demonstrates their similarity. By combining the news hierarchies associated with the user groups that the user belongs to using a *consensus hierarchical clustering* method, the user's interest can be easily captured in a united way. We then can identify news groups that the user might be interested in by cutting the consensus hierarchy at different levels, and finally recommend news articles within each group according to the user's reading behavior. The framework of *PENETRATE* is described in Figure 6.1.

In summary, the contribution of our work is three-fold:

- Our proposed framework is beyond content-based methods and collaborative filtering, in which ***individual user behavior and user group behavior*** are simultaneously considered for recommendation.
- We provide a novel method to ***integrate multiple group-oriented news hierarchies***, by which the general reading preference of individual users can be effectively captured. The proposed framework achieves a good balance between ***the topic coverage and the content diversity*** of the recommended news list.

- We observe that ***the interestedness of news articles with respect to a user is regressive***, and based on this "submodularity" property, we then model the news selection problem as a budgeted maximum coverage problem, which is more realistic than independently selecting news items.

The rest of the section is organized as follows. Chapter 2 presents a brief summary of prior approaches relevant to personalized news recommendation. In Section 6.1.1, the system framework will be introduced, and the algorithmic details for major components are presented in Section 6.1.2,6.2.1,6.2, respectively. Extensive experimental results are reported in Section 6.3. Finally Section 6.4 concludes the chapter.

### 6.1.1 Recommendation framework

Figure 6.1 illustrates the framework of our system. The system is composed of three components described as follows:

I. **Profiling on Users and Groups** (See Section 6.1.2): In this module, individual user's profile is enriched by taking into consideration the profiles of users similar to the given user. The user pool is first divided into multiple groups under the "guideline" of latent topics lying between users and their preferred words. Then the users' profiles in each group are integrated in a weighted way, where the user who likes more the topic category contributes more to the group profile. When newly-published news articles come, the news set is organized into a group-oriented hierarchy, as the ensemble element for further personalization.

II. **Ensemble news hierarchies** (See Section 6.2.1): Since each profile group has its own news hierarchy, the ensemble hierarchical clustering component is designed to combine multiple hierarchies when there are several profile groups related to a given user. Provided that each news hierarchical clustering associates with a dendrogram, we utilize the dendrogram descriptor to define the similarities between all pairs of leaf nodes. Such descriptor can preserve the dendrogram structure (merging order of each pair of sub-cluster).

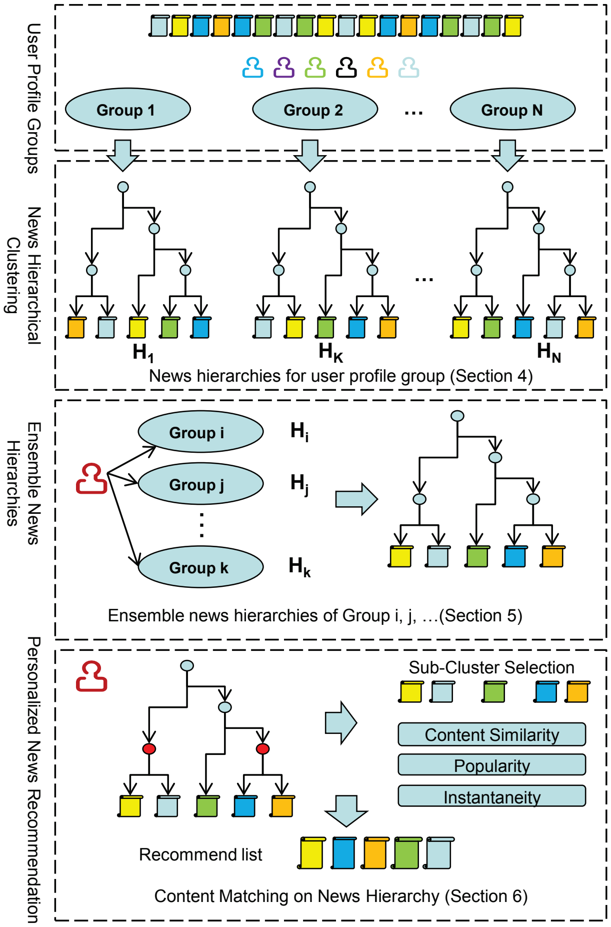

User Profile Groups
Group 1
Group 2
...
Group N
News Hierarchical Clustering
$H_1$
$H_K$
$H_N$
News hierarchies for user profile group (Section 4)
Ensemble News Hierarchies
Group i
Group j
Group k
$H_i$
$H_j$
$H_k$
Ensemble news hierarchies of Group i, j, …(Section 5)
Personalized News Recommendation
Sub-Cluster Selection
Content Similarity
Popularity
Instantaneity
Recommend list
Content Matching on News Hierarchy (Section 6)

Figure 6.1: System framework of PENETRATE.

The ensemble result is an aggregated dendrogram descriptor that can be easily recovered to a consensus news hierarchy.

III. **Personalized News recommendation** (See Section 6.2): Based on the ensembled news hierarchy and the given user's profile, we compare the topic distributions of each intermediate cluster and the user's accessed news, and then sequentially pick up the intermediate clusters based on the similarity score, as the first level of the result. In each cluster, we compare the similarities between each small news group and the user's accessed news, and select the most similar group as the base of the second level. In the selected group, we model personalized news recommendation as a budgeted maximum coverage problem [KMN99] (details in Section 6.2.3), and solve it by selecting news items greedily.

### 6.1.2 Profiling on users and groups

In order to capture a user's reading interests on news articles, news recommendation systems start with constructing the user's profile. Traditionally, a user's profile can be defined by keeping track of what articles the user has read so far (or called consumption history), mainly based on news content. A survey of various user profile construction techniques is provided in [GSCM07]. However, in many cases a user's reading history might not be enough to construct a comprehensive profile representing the exact reading preference of the user. In order to handle this issue, we propose to enrich individual user's profile by taking into consideration the profiles of users similar to the given user. In this way, we can easily capture the general reading interest of users by profiling the user groups that the user belongs to.

**User profiling**

In our system, we consider to construct profiles, using well-known topic models, Probabilistic Latent Semantic Indexing (PLSI) [Hof99] and Latent Dirichlet Allocation (LDA) [BNJ03]. The PLSI model and the LDA model are similar in terms of probabilistic language models, except that in LDA the topic distribution is assumed to have a *Dirichlet prior*. Note that the PLSI model is equivalent to the LDA model under a uniform *Dirichlet prior* information, whereas the LDA model is essentially the *Bayesian* version of the PLSI model [GK03]. *Bayesian* formulation tends to perform better on small data sets because *Bayesian* methods can avoid overfitting [BNJ03]. In reality, the reading history of a specific user might not involve too many news articles. Therefore, we choose LDA as the topic model to detect the possible topics, and represent the topic distribution of the user's profile as a topic vector, each entry of which denotes the weight of the representative words in each topic[1].

[1] The topic vector is built based on the entire vocabulary.

### Group profiling

As is mentioned above, individual user's profile might not be representative enough in terms of the general topics that the user prefers, and therefore, we propose to cluster users into different groups and then characterize the user's general interest using group profiles. Notice that a specific user may have several interested topic categories; in other words, it is not quite reasonable to classify a user into a single group. Thus, in our system, we employ soft clustering algorithms for user clustering task, and then integrate all the users' profiles of the generated group into an aggregated group profile.

### Profile clustering

To obtain online reader groups, we employ PLSI model to quantitatively characterize user profile clustering task. Formally, we have a set of users $\mathcal{U} = \{u_1, u_2, \cdots, u_m\}$, and a set of representative terms obtained from the users' profiles $\mathcal{T} = \{t_1, t_2, \cdots, t_n\}$. The profiles data can be conceptually viewed as a $m \times n$ user-word matrix $UT = [w(u_i, t_j)]_{m \times n}$, where $w(u_i, t_j)$ represents the weight of the term $t_j$ in the profile of the user $u_i$. Note that $w(u_i, t_j)$ is calculated using LDA language model, introduced in Section 6.1.2. We use a set of hidden (unobserved) variables $\mathcal{Z} = \{z_1, z_2, \cdots, z_l\}$, which in our system PENETRATE, correspond to the general topic categories existing in the reading histories of users. Our goal is to automatically discover and characterize user groups based on the user-word matrix.

The user's preference on a specific word, represented by an entry of $UT$ matrix, can be modeled as

$$Pr(u_i, t_j) = \sum_{k=1}^{l} Pr(z_k) \cdot Pr(u_i|z_k) \cdot Pr(t_j|z_k), \tag{6.1}$$

where all possible choices of $z_k$ from which the observation could have been generated are summed up. Here we are only interested in the probability $Pr(u_i|z_k)$, i.e., the possibility that the user $u_i$ belongs to a topic group $z_k$, and use Expectation-Maximization algorithm to estimate it. Finally, we can obtain the probability distribution that the user belongs to groups identified by the hidden variable $z_k$. Here a threshold is needed to filter the unrelated groups; in our system, we empirically set this threshold as 4 (See Section 6.3.4). In this way, the user pool is separated into multiple groups in a soft way, where each user might belongs to several distinct groups.

**Profile aggregation**

After obtaining the clustering result of the user set, our system automatically aggregates all the users' profiles in each group to quantify the group profile. Here we utilize a user weighted aggregation scheme to achieve this goal. Each user has his/her own preference on groups, indicated by the weights $Pr(u_i|z_k)$. For simplicity, we integrate all the users' profiles in this group by a linear combination, where the user with higher preference weight would contribute more to the final group profile. Formally, for a group $\hat{G} = \{\langle u_1, w_1\rangle, \langle u_2, w_2\rangle, \cdots, \langle u_n, w_n\rangle\}$, where each pair $\langle u_i, w_i\rangle$ represents the $i$-th user's profile and the corresponding preference weight on $\hat{G}$, the aggregation can be described as

$$\mathcal{P}_{\hat{G}} = \frac{w_1}{\hat{w}}u_1 + \frac{w_2}{\hat{w}}u_2 + \cdots + \frac{w_n}{\hat{w}}u_n, \tag{6.2}$$

where $\mathcal{P}_{\hat{G}}$ denotes the group profile and $\hat{w} = \sum_{i=1}^{n} w_i$. In this way, the recomputation of LDA on the reading histories of all the users in group $\hat{G}$ can be avoided, and such aggregation scheme provides us reasonable representation of the group profiles.

### 6.1.3 Group-oriented hierarchy generation

Up to this point, we have obtained a list of profile groups related to different hidden topic categories. Given a collection of newly-published news articles, our system *PENETRATE* automatically generates the group-specific news hierarchies for further ensemble processing. Formally, the news set $\mathcal{N}$ contains $m$ news items, $\{n_1, n_2, \cdots, n_m\}$, and the profile groups $\mathcal{G}$ includes $r$ group profiles, $\{g_1, g_2, \cdots, g_r\}$. Note that each news item is summarized by LDA similar to the group profile, represented by a topic distribution. To generate the group-specific news hierarchies, we calculate the *Conditional Mutual Information* (CMI) [CTW$^+$91] of two news items $n_i$ and $n_j$ given a group profile $g_k$ as follows:

$$CMI(n_i; n_j|g_k) = H(n_i, g_k) + H(n_j, g_k) - H(n_i, n_j, g_k) - H(g_k), \tag{6.3}$$

where $H(\cdot)$ denotes the Entropy or joint Entropy of the corresponding variables.

After computing all the CMI values for any pair of news items given a specific group profile, we can obtain a CMI matrix of the original news set, where each entry represents the CMI value of topic distributions of the corresponding two news articles. We then transform the CMI matrix into a news-pair similarity matrix. The transformation procedure is as follows: (1) $l_1$-normalize the CMI matrix; and (2) substitute the value on the main diagonal by 1. The generated similarity matrix is utilized to construct a group-specific news hierarchy using hierarchical clustering, for the purpose of ensemble hierarchical clustering on multiple news hierarchies.

## 6.2 Personalized recommendation

Personalized news recommendation is oriented from exploring the relations between newly-published news articles and the user's profile. In *PENETRATE*, a two-level recommendation hierarchy is provided, where the first level shows a brief summary for each topic category the user might prefer, and the second level gives a specific list of news articles similar to the user's reading interest. Further, we model personalized news selection as a budgeted maximum coverage problem by exploring the "submodularity" hidden in multiple aspects of news personalization, and then resolve it greedily. In this way, the recommended news list can achieve an elegant balance of the topic coverage and the content diversity, as well as the satisfaction of online readers.

### 6.2.1 Ensemble news hierarchies

Given a news set, each user profile cluster corresponds to a specific hierarchical clustering of the news set. Such news hierarchy reflects the reading preferences of the group of users. To capture the user's interest without losing the diversity of user's preferences, we propose an ensemble clustering framework to combine various news hierarchical clustering results associated with profile groups.

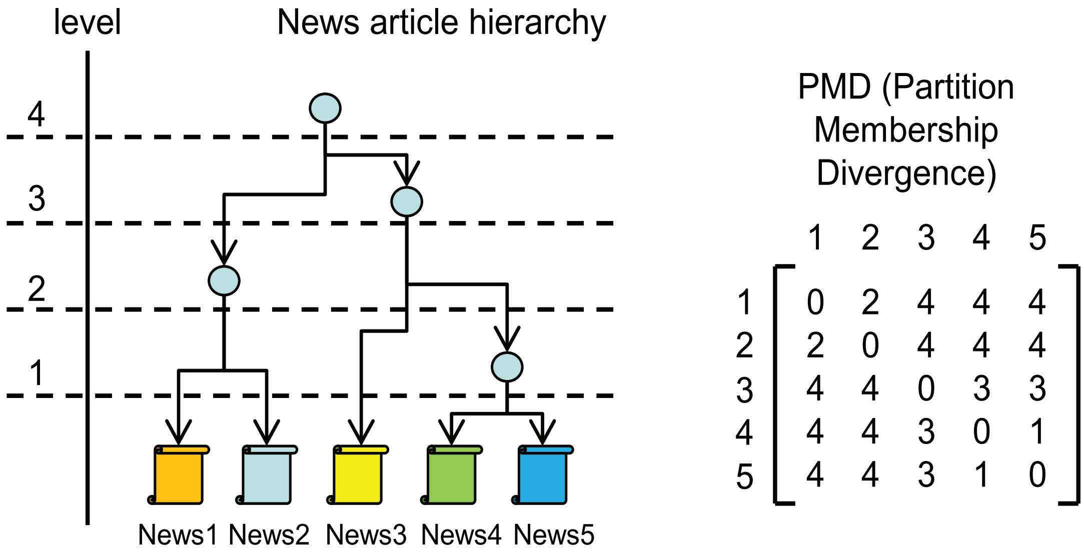
#### Problem formulation

The task of ensemble news hierarchical clustering is to obtain a single consensus news hierarchy from multiple hierarchical clustering results. Formally, let $X = \{x_1, x_2, \cdots, x_n\}$ be a set of $n$ pieces of news in the given set. A set of $T$ hierarchical clusterings $\mathcal{SP} = \{H^1, H^2, \cdots, H^T\}$ building on data points in $X$ is given to demonstrate various merging preferences among news. The dendrogram descriptor, which is defined to preserve the structural information of hierarchies, are used to represent a hierarchical clustering result as a dis-similarity matrix [Ada72][Ada86a]. We use the following descriptor to represent a hierarchial clustering.

- ***Partition Membership Divergence*** **(PMD)**: By utilizing the property that a hierarchical clustering result implies a sequence of nested partitions obtained by cutting the dendrogram at every internal node, the PMD is defined as the number of partitions of the hierarchy in which two specified leaves are not in the same cluster.

Figure 6.2 gives a simple example about how PMD describes the structural distance preserved in hierarchical clustering. The consensus news hierarchy is a single news hierarchical clustering result by aggregating all clusterings in $\mathcal{SP}$. It could be different from any of $H^i, i \in T$. The ensemble procedure is illustrated in Figure 6.3.

In general, our ensemble hierarchy framework utilizes the representative power of ultrametrics to integrate multiple group-oriented news hierarchies into one consensus hierarchy.


level
News article hierarchy
4
3
2
1
News1 News2 News3 News4 News5
PMD (Partition Membership Divergence)

$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & 2 & 4 & 4 & 4 \\ 2 & 2 & 0 & 4 & 4 & 4 \\ 3 & 4 & 4 & 0 & 3 & 3 \\ 4 & 4 & 4 & 3 & 0 & 1 \\ 5 & 4 & 4 & 3 & 1 & 0 \end{array}$$

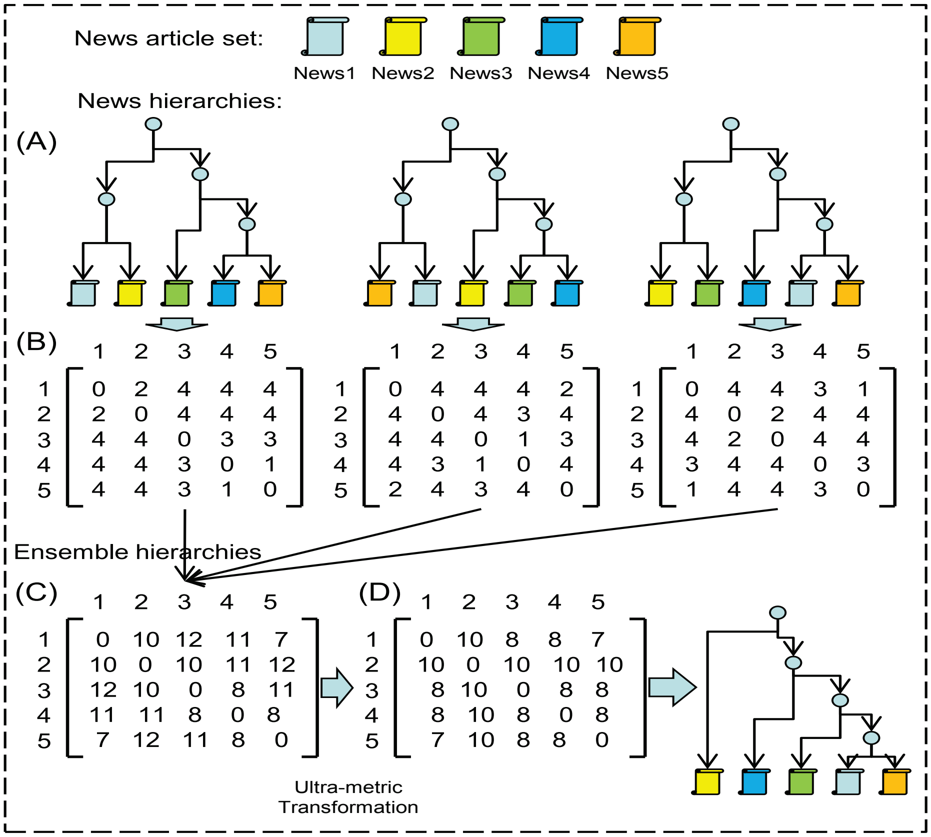
Figure 6.2: A dendrogram descriptor example.

The obtained single news hierarchy achieves a good balance between the topic coverage and the content diversity of the recommended news list.

### 6.2.2 Interest matching for representation Lv.1

After obtaining the group-specific news hierarchy, we use Dunn's Validity Index [Dun73] as the metric to generate groups of news articles. This validity measure is based on the idea that high-quality clustering produces well-separated compact clusters. In general, the larger Dunn's Index, the better the clustering. Therefore, our method tries to maximize the Dunn's Index. In this way, we do not have to specify the number of clusters when performing clustering on news articles. Note that each news group is summarized using LDA, similar to the user's profile.

Once we generate news groups and the user's profile, the first representation level can be obtained by sequentially matching the user's profile onto the news groups. For simplicity, we only consider the cosine similarity between topic distributions of each intermediate



$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & 2 & 4 & 4 & 4 \\ 2 & 2 & 0 & 4 & 4 & 4 \\ 3 & 4 & 4 & 0 & 3 & 3 \\ 4 & 4 & 4 & 3 & 0 & 1 \\ 5 & 4 & 4 & 3 & 1 & 0 \end{array} \quad \begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & 4 & 4 & 4 & 2 \\ 2 & 4 & 0 & 4 & 3 & 4 \\ 3 & 4 & 4 & 0 & 1 & 3 \\ 4 & 4 & 3 & 1 & 0 & 4 \\ 5 & 2 & 4 & 3 & 4 & 0 \end{array} \quad \begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & 4 & 4 & 3 & 1 \\ 2 & 4 & 0 & 2 & 4 & 4 \\ 3 & 4 & 2 & 0 & 4 & 4 \\ 4 & 3 & 4 & 4 & 0 & 3 \\ 5 & 1 & 4 & 4 & 3 & 0 \end{array}$$

$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & 10 & 12 & 11 & 7 \\ 2 & 10 & 0 & 10 & 11 & 12 \\ 3 & 12 & 10 & 0 & 8 & 11 \\ 4 & 11 & 11 & 8 & 0 & 8 \\ 5 & 7 & 12 & 11 & 8 & 0 \end{array} \quad \begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & 10 & 8 & 8 & 7 \\ 2 & 10 & 0 & 10 & 10 & 10 \\ 3 & 8 & 10 & 0 & 8 & 8 \\ 4 & 8 & 10 & 8 & 0 & 8 \\ 5 & 7 & 10 & 8 & 8 & 0 \end{array}$$

Figure 6.3: An illustrative example of ensemble news hierarchical clusterings. The original candidate hierarchical clusterings are shown in (A). (B) shows their corresponding PMD for each news hierarchy and (C) is the aggregated hierarchical clustering. (D) is the ultra-metric transformation on aggregation matrix obtained in (C) with its corresponding hierarchial clustering result.

cluster and the user's reading history. In practice, people tend to have their preference on news categories, i.e., not interested in all the categories. Therefore, we choose the categories whose corresponding similarity is greater than a dynamic threshold[2]. After selecting the appropriate categories that roughly satisfy the user's general preference, we dig into each category and choose the news articles as the second level representation.

[2] The dynamic threshold is set to be the median of all similarity scores with respect to a specific user's profile.

### 6.2.3 News selection for representation Lv.2

To facilitate the selection of specific news items, we summarize each news article using LDA, as the profile of news item. We then model news selection as a budgeted maximum coverage problem, and solve it by a greedy algorithm. Intuitively, *the interestedness of news articles with respect to a user is regressive*, i.e., after he/she clicks the first piece of news she is interested in, the interest value might decrease when she clicks the second one or more.

**Introduction to submodularity**

Let $E$ be a finite set and $f$ be a real valued nondecreasing function defined on the subsets of $E$ that satisfies

$$f(T \cup \{\varsigma\}) - f(T) \leqslant f(S \cup \{\varsigma\}) - f(S), \tag{6.4}$$

where $S \subseteq T$, $S$ and $T$ are two subsets of $E$, and $\varsigma \in E \setminus T$. Such a function $f$ is called a **submodular** function [NWF78]. Intuitively, by adding one element to a larger set $T$, the value increment of $f$ can never be larger than that by adding one element to a smaller set $S$. This intuitive diminishing property exists in different areas. For example, in social network, adding one new friend cannot increase more social influence for a more social group than for a less social group. The similar scenario holds in personalized news recommendation: *the interestedness of news articles with respect to a user is regressive*, i.e., after he/she clicks the first piece of news she is interested in, the interest value might decrease when she clicks the second one or more.

The budgeted maximum coverage problem is then described as: given a set of elements $E$ where each element is associated with an influence and a cost defined over a domain of these elements and a budget $B$, the goal is to find out a subset of $E$ which has the largest possible influence while the total cost does not exceed $B$. This problem is NP-hard [KMN99]. However, [KMN99] proposed a greedy algorithm which picks up the element that increases the largest possible influence within the cost limit each time and it guarantees the influence of the result subset is $(1 - 1/e)$-approximation. Submodularity resides in each "pick up" step. A key observation is that submodular functions are closed under nonnegative linear combinations [LKG$^{+}$07].

### Submodular model for recommendation

In a particular news group, most of news articles concentrate on similar or even the same topic, with minor difference on major aspects of the corresponding topic. For example, given a news group talking about a popular movie "*Inception*", one piece of news may focus on the actor cast of this movie, while another may describe the high-end techniques used in this movie. Typically, a news reader is interested in some specific aspects of the given topic, but not all of them. Under this intuition, our news selection strategy can be described as follows (note that $\mathcal{N}$ denotes the original news group, $\mathcal{S}$ represents the selected news set, and $\varsigma$ is the news item being selected). After selecting $\varsigma$,

I. $\mathcal{S}$ should be similar to the general topic in $\mathcal{N} \setminus \mathcal{S}$;

II. The topic diversity should not deviate much in $\mathcal{S}$;

III. $\mathcal{S}$ should provide more satisfaction to the given user's reading preference.

Per the above strategy, we define a quality function $f$ to evaluate the current selected news set $\mathcal{S}$ over the whole news group $\mathcal{N}$ as

$$\begin{aligned} f(\mathcal{S}) = & \frac{1}{|\mathcal{N} \setminus \mathcal{S}| \cdot |\mathcal{S}|} \sum_{n_1 \in \mathcal{N} \setminus \mathcal{S}} \sum_{n_2 \in \mathcal{S}} sim(n_1, n_2) \\ & + \frac{1}{\binom{|\mathcal{S}|}{2}} \sum_{n_1, n_2 \in \mathcal{S} \& n_1 \neq n_2} -sim(n_1, n_2) + \frac{1}{|\mathcal{S}|} \sum_{n_1 \in \mathcal{S}} sim(u, n_1), \end{aligned} \tag{6.5}$$

where $n_1$ and $n_2$ denote news items, $u$ represents the given user, and $sim(\cdot, \cdot)$ represents the similarity between two profiles, either the user profile or the news profile.

In Eq.(6.5), three components are involved, corresponding to the news selection strategy we list above. $f(\mathcal{S})$ balances the contributions of different components Note that all these three components are naturally submodular functions. Based on the linear invariability of the submodular function [LKG+07], $f(\mathcal{S})$ is also a submodular function. Suppose $\varsigma$ is the candidate news article, the quality increase is therefore represented as follows:

$$I(\varsigma) = f(\mathcal{S} \cup \{\varsigma\}) - f(\mathcal{S}). \tag{6.6}$$

The goal is to select a list of news articles which provide the largest possible quality increase within the budget[3]. Hence, personalized news recommendation is transformed to the budgeted maximum coverage problem [KMN99].

In each news group, we adopt a greedy algorithm to solve the budgeted maximum coverage problem: sequentially pick up the news article which provides the largest quality increase based on the selected news set until the budget is reached. To integrate recommended news items from different news groups into the final recommendation list, we sequentially select top ranking items within each group, and the number of items selected in one group is proportional to the interest weight of the user on the corresponding topic category. Finally, the recommendation list is adjusted based on the popularity and recency of the selected news articles, and presented to the user.

## 6.3 Empirical evaluation

### 6.3.1 Real world dataset

For evaluation purpose, we gather news articles along with users' access history from several popular news websites[4], ranging from July 15th, 2010 to July 16th, 2011. It contains the details of news articles (e.g., news title, content, published time, etc.) and user access history (e.g., anonymous users, accessed news items, accessed time, etc.). After obtaining

[3] Here the budget can be regarded as the maximum number of recommended items in each news group.

[4] The data is collected from commercial parties.

the whole dataset, we preprocess the data by removing news articles that are rarely accessed (i.e., the accessed frequency is less than 10 times per day) and by storing users with frequent online reading behaviors (i.e., users who read news articles every day and read more than 1 piece of news each day). By doing this, we can somewhat avoid introducing user bias and item bias into user profiling. After preprocessing, a total of 1,042,200 news items are stored, along with 52,630 users, each day in average with 2,848 news articles. Notice that in the experiment, we are not concerned with the live traffic to a news website, but focus on the recommendation quality of our proposed method over the collected dataset.

### 6.3.2 Experimental setup

To evaluate our proposed system, we implement it based on the architecture introduced in Section 6.1.1. The entire system contains the following three major components: (a) An offline component responsible for periodically clustering the user pool, and updating group profiles and individual profiles; (b) An offline component of periodically clustering news articles published within a time range based on different group profiles; and (c) An online component to recommend news articles to individual users. From the experimental perspective, we verify our system components in an integrated manner, where all these three components are tested under a unified online environment.

### 6.3.3 Profiling evaluation

As discuss in Section 6.3.3, LDA would be more beneficial when the news dataset is small. In order to verify this claim, we design the experiment as follows: (1) use 1 hour, 12 hours, 1 day, 2 days, and 3 days as the time periods; (2) for each time period, randomly select 10 time ranges, extract news articles in these time ranges, and treat the articles published earlier than these time ranges as the reading history of the users; (3) perform PLSI and LDA on the generated ensemble hierarchical clustering results of these news sets, respectively; (4) perform top @10 news recommendation to 2000 users randomly selected from the users' pool, where user profiles, along with group profiles, are constructed by virtue of PLSI and LDA, respectively; (5) compute the averaged F1-score (over both time range and all 2000 users) for PLSI-based and LDA-based systems. We also use the simple term frequency (TF) to construct profiles, as the comparison baseline. The result is shown as in Figure 6.4.

From the result, we observe that: (i) LDA-based system has steady recommendation performance in terms of F1-score, regardless of different size of news corpus; and (ii) PLSI-based recommender system has comparable results when the news corpus becomes larger. However, when the dataset is relatively small, the performance of PLSI-based system is comparatively lower than LDA-based system. Essentially, it results from *overfitting* when the dataset is small. Therefore, ***LDA is more applicable to our recommender system***.

### 6.3.4 Profile clustering evaluation

In our system PENETRATE, newly-published news articles are hierarchically clustered based on different profile groups, under the assumption that the profile of the user group can be more representative and useful than individual profiles when clustering news articles. In order to verify our assumption, we design a series of experiments on examining the behavior of user groups, described as follows.

#### Group profile V.S. individual profile

It is straightforward that individual profile has more personalized property, whereas group profile can better describe the general reading preference of individual users. To better capture this claim, we adopt the experimental setting similar to the one introduced in Section 6.3.3, and compare the performance based on the following profile schemes: (1) using

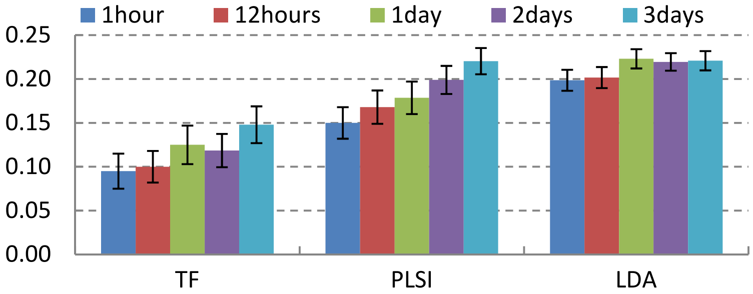
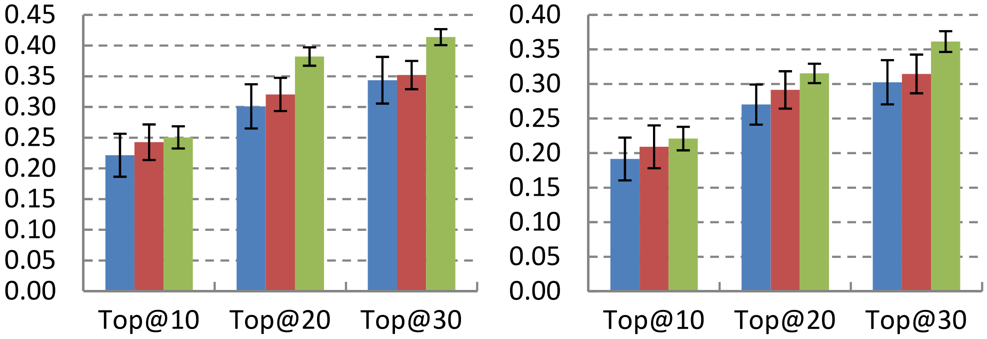


Figure 6.4: Performance comparison of TF-, PLSI- and LDA-based systems.



(a) Recall.

(b) F1-score.

Figure 6.5: Comparison among recommender systems using different profile schemes. Remark: Blue – IP, Red – SP, Green – PENETRATE.

individual profile of the given user to filter news groups obtained from the hierarchical clustering result on news articles, and to filter news items in each news group, denoted as "IP"; (2) using the integrated profile of users similar to the given user to filter news groups, and the given user's profile to filter news items in each group, denoted as "SP"; and (3) constructing ensemble news hierarchy using the profiles of user groups that the given user belongs to, and then using the given user's profile to filter news groups and news items in the resulted hierarchy, which is exactly the scheme applied in our recommender system, PENETRATE. We recommend news items (top @10, top @20 and top @30) to the selected 2000 users, and compared the averaged recall and F1-score of recommendation results over 5 time ranges, where each time range contains 3 days. Figure 6.5 shows the comparison results.

From the result, we observe that: (1) our proposed profile scheme outperforms the other two baselines, which verifies our previous assumption; and (2) simply using individual profile to filter news groups and news items does not provide promising performance. The reason behind this might result from the fact that ***individual profile cannot reasonably capture the general topics of the user's reading history***, and therefore fails to extract appropriate news groups from the news hierarchy.

### Different soft clusterings on profiles

In general, different online users might have different reading preference, and therefore the important words extracted from their reading histories may differ. For a given user, he/she might prefer several general topics. As described in Section 6.3.3, we use topic modeling, e.g., PLSI and LDA, to handle the relationship between users and representative words; in other words, there might be several implicit topics that can classify users and words into different groups simultaneously, where each user may belong to several groups. To examine the effectiveness of soft clustering techniques in our recommender system, we utilize the experimental setting similar to the previous procedures, and compare the performance of different soft clustering algorithms based on our proposed recommendation framework. The soft clustering algorithms being considered include PLSI, LDA and Fuzzy K-means (Fuz). We choose the possible clusters that a user might belong to in a range of [2,10], and recommend top @30 news articles to 2000 randomly selected users in 5 times periods (3 days in one period). The F1-score is averaged and plotted in Figure 6.6.

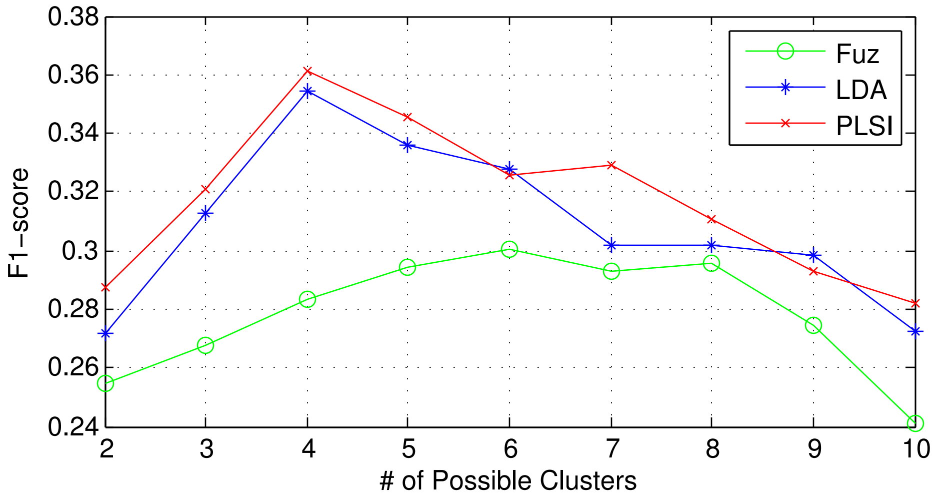


Figure 6.6: Recommendation F1-score based on different soft clustering algorithms.

As is evident in Figure 6.6, PLSI and LDA give comparatively better performance under our experimental setting. We choose PLSI as the soft clustering technique in our system, since PLSI requires less parameter estimations than LDA. In addition, from the result we can observe that when the possible number of clusters for users is 4, PLSI-based recommender system achieves the best performance.

### The effect of ensemble hierarchy

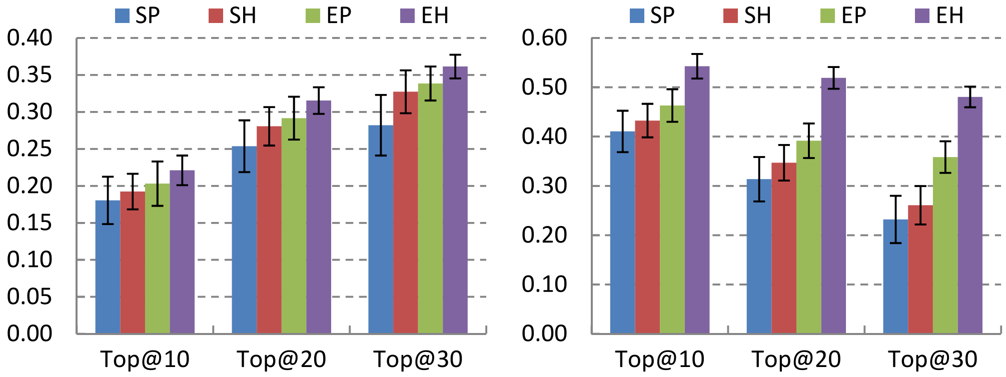
When recommending news items to individual users, our strategy replies on the selection of news groups from a news hierarchy. A couple of intuitive ways to construct the basis for news selection involve: (1) Single Partition (SP): To cluster newly-published news articles based on the most promising profile group by employing partition-based clustering techniques; (2) Single Hierarchy (SH): To cluster newly-published news articles based on the most promising profile group by employing hierarchical clustering techniques; (3) Ensemble Partition (EP): To integrate the partition-based clustering results based on multiple profile groups preferred by the user; and (4) Ensemble Hierarchy (EH): To integrate the hierarchical clustering results based on multiple profile groups preferred by the user.

Note that in our work, we use EH to build the recommendation base. As shown in [LWL$^+$11], the performance of news recommenders based on hierarchical clustering is superior to the one by partition-based clustering (e.g., K-means). In addition, *only adopting a single profile group might result in the dearth of the general topics, which renders the final recommendation result less diverse*. Alternatively, in our system PENETRATE, all the profile groups preferred by the user are taken into account, and are used as a prior to construct an ensemble news hierarchy on newly-published articles. In this way, the generated news hierarchy might involve a couple of *distinct topics* that the user might be interested in, and therefore make the result more diverse.

In order to demonstrate our observation, we compare the recommendation result based on the aforementioned 4 methods in terms of accuracy and diversity. For single hierarchy, we utilize the priority queue implemented using a binary heap to speed up the hierarchical clustering process. For partition-based methods, we conduct K-means clustering 10 times to obtain the best partition, by which eliminating the over-dependency on random seeds initialization. The experimental setting is the same as the previous procedures. Particularly, to evaluate how diverse the recommendation result is, we compare the set diversity described in [ZH08] between the results of SH and EH. The news set diversity is defined as the *average dissimilarity* of all pairs of news items in the recommendation list. Specifically, given a news set $\mathcal{N}$, the *average dissimilarity* of $\mathcal{N}$, $f_d(\mathcal{N})$, is defined as

$$f_d(\mathcal{N}) = \frac{2}{p(p-1)} \sum_{n_i \in \mathcal{N}} \sum_{n_j \in \mathcal{N}, n_j \neq n_i} (1 - Sim(n_i, n_j)) \tag{6.7}$$

where $|\mathcal{N}| = p$, and the dissimilarity of a news pair is represented as $1 - Sim(n_i, n_j)$, in which $Sim(n_i, n_j)$ denotes the news profile similarity between the news item $n_i$ and $n_j$.



(a) F-score comparison.

(b) Diversity comparison

Figure 6.7: Comparison between partition-based and hierarchical-based recommender systems.

The experimental result of the quality of top @10, @20, @30 news items is shown in Figure 6.7. From the result, we observe that:

- The diversity decreases as the recommendation news list enlarges. It is straightforward that when more news articles are selected, the topic distribution of the news list becomes closer to the user's reading interest, and therefore the selected news items are more similar.

- The results of ensemble-based methods are superior to the ones of single profile group based methods. ***By ensemble, a user's profile can be enriched to a great extent, due to the distinguishable interests originated from multiple profile groups***.

- The diversity of the recommendation list provided by single group based method drops dramatically as the list size increases due to *the restricted topics of single hierarchy or single partition.*

- EH-based method shows promising performance, and since we intentionally consider the requirement of news readers via ensemble hierarchy, the diversity decreases very smoothly when we recommend more news items to individual users.

### 6.3.5 News selection strategy evaluation

In order to verify the effectiveness of our proposed news selection strategy, we provide detailed comparison between ours and the general greedy selection strategy simply based on pairwise similarities. Also, we implement a recommender system that models the recommendation as a contextual bandit problem [LCLS10], as the comparison base. For each approach, we randomly select 2000 users to provide recommendations for them. We plot the precision and recall pair for each user on top @10, @20, and @30 news items recommended to these users. Figure 6.8 shows the comparison results. From the result, we observe that besides the higher precision and recall, the performance distribution of PENETRATE is more compact than the other methods. The reason behind this phenomenon is that our proposed "submodularity-based modeling" tries to select news articles by ***considering the representativeness of news items, the diversity of news lists, and the satisfaction of online news readers, simultaneously***. This demonstrates the stability of our proposed news selection strategy.

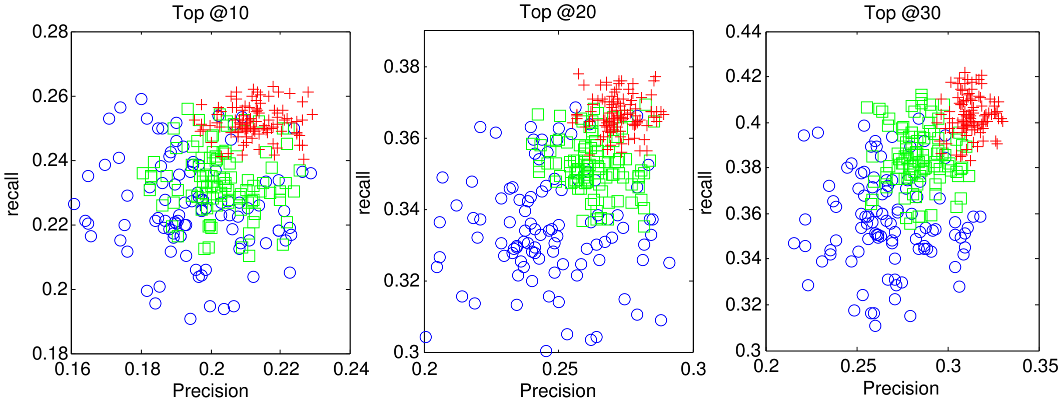


Figure 6.8: Precision-recall plot for different news selection strategies. Remark: ○ represents users using the general greedy-based recommender system; □ denotes users using the bandit-based recommender system; and + represents users using PENETRATE.

| Methods | $N \leq 20$ | | | $20 < N \leq 50$ | | | $N > 50$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | Top @10 | Top @20 | Top @30 | Top @10 | Top @20 | Top @30 | Top @10 | Top @20 | Top @30 |
| Goo | 0.1845 | 0.2633 | 0.2901 | 0.2002 | 0.2957 | 0.3203 | 0.2206 | 0.3126 | 0.3365 |
| ClickB | 0.1730 | 0.2516 | 0.2872 | 0.1874 | 0.2831 | 0.2907 | 0.2119 | 0.2957 | 0.3147 |
| Bilinear | 0.1860 | 0.2587 | 0.2921 | 0.1923 | 0.2809 | 0.3115 | 0.2153 | 0.3073 | 0.3218 |
| Bandit | 0.1716 | 0.2409 | 0.2762 | 0.1837 | 0.2765 | 0.3087 | 0.2037 | 0.2984 | 0.3109 |
| PENETRATE | **0.2102** | **0.2981** | **0.3272** | 0.2182 | 0.3013 | **0.3426** | 0.2213 | 0.3185 | **0.3640** |

Table 6.1: Comparison on F1-score of different algorithms for three distinct user groups. The bold numbers indicate that the corresponding results significantly outperform the others under $p < 0.005$.

### 6.3.6 Overall evaluation

In the above experiments, all the users are equally treated as the experimental subject. In reality, users with different news access patterns, such as different reading frequency every day, may have distinct preferences on news topics, and therefore the dynamic interest on news articles may vary a lot. In addition, many news recommendation systems cannot address the so-called "cold-start" problem. In order to verify the performance of our proposed algorithm on different user groups, we separate the selected users into three groups based on their reading habits. Suppose a user reads $N$ news articles per day, then the three groups are: (i) $N \leq 20$ $(25\%)$; (ii) $20 < N \leq 50$ $(38\%)$; (iii) $N > 50$ $(37\%)$. We apply different algorithms on these three users groups with top @10, top @20 and top @30 recommended news, and record the F1-score respectively. Here, the comparison base includes four existing approaches of different frameworks: [DDGR07] (Goo, collaborative filtering), [LDP10b] (ClickB, content-based), [CP09] (Bilinear, probabilistic model) and [LCLS10] (Bandit, hybrid). Table 6.1 shows the comparison results. It demonstrates that our system PENETRATE can achieve a reasonable recommendation result when it is subject to the "cold-start" problem because our proposed method ***considers the group behavior instead of individual behavior*** when finding the general topics that the user might be interested in; in other words, even if the given user is a new user, his/her profile can be ***enriched by the profiles of users similar to the given user***.

## 6.4 Conclusion

In this chapter, a novel personalized news recommendation system, PENETRATE, is proposed to provide attractive news reading lists to online readers. Our system takes into consideration the reading behaviors of both individual user and a group of users when performing recommendation. The group behavior shows us the general topics that the user might be interested in, whereas the individual behavior provides us personalized information for further filtering news articles. Extensive empirical results demonstrate the efficacy of our system.

The system can be improved in terms of both accuracy and efficiency. In particular, the time cost of ensemble hierarchical clustering (as introduced in Section 6.2.1) can be further reduced by carefully design, e.g., utilizing distributed frameworks or Map-Reduce programming model. We also plan to incorporate the temporal information into the recommendation paradigm (as introduced in Section 6.2), i.e., the recommendation should be biased to more recent preference of online users.

## CHAPTER 7

## Topic III: application in disaster management domain

Business closures caused by disasters can cause millions of dollars in lost productivity and revenue. A study in Contingency Planning and Management shows that 40% of companies that were shut down by a disaster for three days failed within 36 months. Thin margins and a lack of a well-designed and regularly tested disaster plan can make companies, particularly small businesses, especially vulnerable[ZST$^+$10]. We believe that the solution to better disaster planning and recovery is one where the public and private sectors work together to apply computing tools to deliver the right information to the right people at the right time to facilitate the work of those working to restore a communitys sense of normalcy. While improved predictive atmospheric and hydrological models and higher quality of building materials and building codes are being developed, more research is also necessary for how to collect, manage, find, and present disaster information in the context of disaster management phases: preparation, response, recovery, and mitigation[HCL$^+$10, McE].

In the United States, the Federal Emergency Management Agency (FEMA) has recognized the importance of the private sector as a partner in addressing regional disasters. The State of Florida Division of Emergency Management has created a Business and Industry Emergency Support Function designed to facilitate logistical and relief missions in affected areas. Four counties, Palm Beach, Broward, Miami-Dade, and Monroe, which constitute the Southeastern population of South Florida and include over 200 000 business interests, are developing Business Recovery Programs to help facilitate faster business community recovery through information sharing and collaboration.

Disaster management researchers at Florida International University have collaborated with the Miami-Dade Emergency Operations Center (EOC), South Florida Emergency Management and industry partners including Wal-Mart, Office Depot, Wachovia, T-Mobile, Ryder Systems, and IBM to understand how South Florida public and private sector entities manage and exchange information in a disaster situation. The efficiency of sharing and management of information plays an important role in the business recovery in a disaster[SLD$^+$08]. Users are eager to find valuable information to help them understand the current disaster situation and recovery status. The community participants (the disaster management officials, industry representatives, and utility agents) are trying to collaborate to ex- change critical information, evaluate the damage, and make a sound recovery plan. For example, it is critical that companies receive information about their facilities, supply chain, and city infrastructure. They seek this information from media outlets like television/radio newscasts, employee reports, and conversations with other companies with which they have a relationship. With so many sources of information, with different levels of redundancy and accuracy, possibly generated by a variety of reports (structured and unstructured), it is difficult for companies to quickly assimilate such data and understand their situation.

We have learned that a large-scale regional disaster may cause a disruption in the normal information flow, which in turn affects the relationships between information producers and consumers. Effective communication is critical in a crisis situation. What is not very well known is how to effectively discover, collect, organize, search, and disseminate real-time disaster in- formation.

Our study of the hurricane disaster information management domain has revealed two interesting yet crucial information management issues that may present similar challenges in other disaster management domains. The first issue is that reconstructing or creating information flow becomes intractable in domains where the stability of information networks is fragile and can change frequently. However, important information networks often carry and store critical information between parties, which dominates the flow of resources and information exchanges. The consequence is that the ability and the efficiency of communication degrade once critical networks are disrupted by the disaster and people may not have alternative paths to transfer information. For example, once power is disabled and uninterruptable power supplies fail after a hurricane, computing and networking equipment will fail unless preventative measures are taken. However, maintaining a fuel-consuming generator is not always possible.

Another issue is the large volume of disaster situation information. Reading and assimilating situational information are very time consuming and may involve redundant information. Thus, to quickly reassemble or create information flows for multiparty coordination activities during disaster situations, technologies that are capable of extracting information from recent updates, delivering that information without conflict or irrelevance, and representing preferential information are needed.

This research is mainly focused on the second issue. Research in disaster management addresses the needs and challenges of information management and decision making in disaster situations[BBP$^+$09, BBPK08, LDP$^+$10a]. We have developed an understanding of those needs for hurricane scenarios. The information delivery should support users complex information needs tailored to the situation and the tasks; and the information should be synthesized from heterogeneous sources and tailored to specific contexts or tasks at hand. It should be summarized for effective delivery and immediate usefulness for making decision.

## 7.1 Challenges

The approaches and the tools that are used for information sharing vary based on the task and scale of the participating agencies or the types of information exploration platforms.

Commercial systems, such as WebEOC[Web] and E-Teams[ET] used by Emergency Management departments located in urban areas, can access multiple resources. A Disaster Management Information System developed by the Department of Homeland Security is available to county emergency management offices and participating agencies to provide an effective reports/document sharing software system. The National Emergency Management Network[Net] allows local government to share resources and information about disaster needs; The RESCUE Disaster Portal is a web portal for emergency management and disseminating disaster information to the public[HCL+10]; The Puerto Rico Disaster Decision Support Tool is an Internet-based tool for disaster planners, responders, and related officials at the municipal, zone, and state level for access to a variety of geo-referenced information [Too].

Efforts, such as GeoVISTA [MRJ+11], facilitate the information distribution process in disasters. GeoVISTA monitors tweets to form situation alerts on a map-based user interface according to the geo-location associated with the tweets. Such a system applies geographic information sciences to scientific, social, and environmental problems by analyzing geospatial data [MRJ+11].

These useful situation-specific tools provide query interfaces, and GIS and visualization capabilities to simplify the users interaction and convey relevant information. The primary goal of these systems are message routing, resource tracking, and document management for the purpose to support situation awareness, demonstrate limited capabilities for automated aggregation, data analysis, and mining[HCL+10].

Through careful study of existing disaster information management systems and close cooperation with domain experts and local departments, we have identified four key design challenges for disaster information sharing platforms and tools.

1. Effective techniques to capture the status information: Participants need to communicate status through many channels, including email, mailing lists, web pages, press releases, and conference calls. It is desirable to capture such status information when it is available and to prevent redundant reporting. To facilitate the reuse of such materials, users should be able to update status information via unstructured documents

such as plain text, Adobe PDFs, and documents. It is necessary to identify the useful information in the documents.

2. Effective and interactive information summarization methods: It is important to build a summarized view to support under- standing the situation from reports. Multi-document summarization provides users with a tool to effectively extract important and related ideas of current situations. Previous text summarization techniques gave users a fixed set of sentences based on the user query. An interactive summarization interface is needed to help users navigate collected information at different granularities, and locate their target information more efficiently.

3. Intelligent information delivery techniques: Data can be collected through different channels and may belong to different categories. During disaster preparation and recovery, users do not have the time to go through the system to find the in- formation they want. Structured information can help people make decisions by providing them with actionable and concrete information representation and exploration. However, navigating large datasets on a mobile device is particularly inefficient. An interactive tabular interface can help users filter useful information by adaptively changing query conditions and user feedback.

4. Dynamic community generation techniques: In information sharing tasks, identifying a group of recipients to which a certain type of information is conveyed can improve the efficiency of communication. In addition, identifying how participants interact with these communities in a disaster situation may reveal information helpful in a recovery scenario. User recommendation techniques can automatically and interactively generate potential recipients for different pieces of information. In addition, user recommendation techniques can help to dynamically organize user groups according to various information sharing tasks.

We created an information-rich service on both web-based and mobile platforms in the disaster management domain to address the design challenges. In particular, to address the first challenge, we apply information extraction to automatically extract the status information from documents. To address the second challenge, we apply hierarchical summarization to automatically extract the status information from a large document set and also provide a hierarchical view to help users browse information at different granularities. To address the third, we create a user interface capability called the dynamic dashboard to improve information quality to match users interests, and use document summarization techniques to give users fast access to multiple reports. In addition, a dynamic query form is designed to improve information exploration quality on mobile platforms. It captures users interests by interactively allowing them to refine and update their queries. To address the fourth challenge, for community discovery, we adopt spatial clustering techniques to track assets like facilities, or equipment, which are important to participants. The geo-location of such participants can be organized into dynamic communities, and these communities can be informed about events or activities relevant to their spatial footprints. For user recommendation, we use transactional recommendation history combined with textual content to explore the implicit relationship among users.

## 7.2 Data-driven techniques for disaster information management

### 7.2.1 Spatial clustering with constraints

Spatial data clustering identifies clusters, or densely populated regions, according to some distance measurement in a large, multidimensional data set [TSK05b, Han05]. Many spatial clustering techniques [ZFLW02, EKSX96, ZL02b] have been developed for identifying clusters with arbitrary shapes of various densities and with different physical constraints. In practice, communities formed by geographically related entities can be of various shapes.

So we extend DBScan [EKSX96], a well- known density-based clustering algorithm, which is capable of identifying arbitrary shape of clusters, to generate dynamic communities.

We consider the method of spatial clustering with constraints [ZST$^+$10]. Generally, there are three types of constraints [HKP06]: 1). Constraints on individual objects: Such constraints are non-spatial instance level constraints that can be preprocessed before performing clustering algorithms. 2). Constraints as clustering parameters: Such constraints are usually confined to the algorithm itself. Usually, user-specified parameters are given through empirical studies. 3). Constraints as physical obstacles: Such constraints are tightly intertwined with clustering process. It is clear that physical obstacles are such constraints which prevent two geographically close entities from being clustered together. In real case, the bridge, highway and rivers are of this type. We focus on object constraints and physical constraints. Object constraints: We have two ways to obtain object constraints: 1) users submit formatted reports through report interface. Those reports are immediately recorded in the database; 2) our system extracts entity status from reports. For example, Table 7.1 shows the information extracted from Emergency Operational Center (EOC) internal reports, which can be used as object constraints.

| Service | Time | Status |
|---|---|---|
| Miami-Dade County | Oct 24, 2005 | Close |
| Public Schools | Oct 24, 2005 | Close |
| Rail Service | Oct 22, 2005 6:30 p.m. | Open |
| Miami International Airport | Oct 22, 2005 6:30 p.m. | Open |
| ... | ... | ... |
| I-94 North Entry | Oct 26 - 28, 2005 | Close |
| Hurricane Evacuation Center | Oct 23, 2005 afternoon | Close |

Table 7.1: Information extracted from the EOC reports

Obstacle constraints: Polygon is a typical structure in spatial analysis to model objects. Obstacles modeled by a polygon can be represented as a set of line segments after performing polygon reduction [ZL02a].

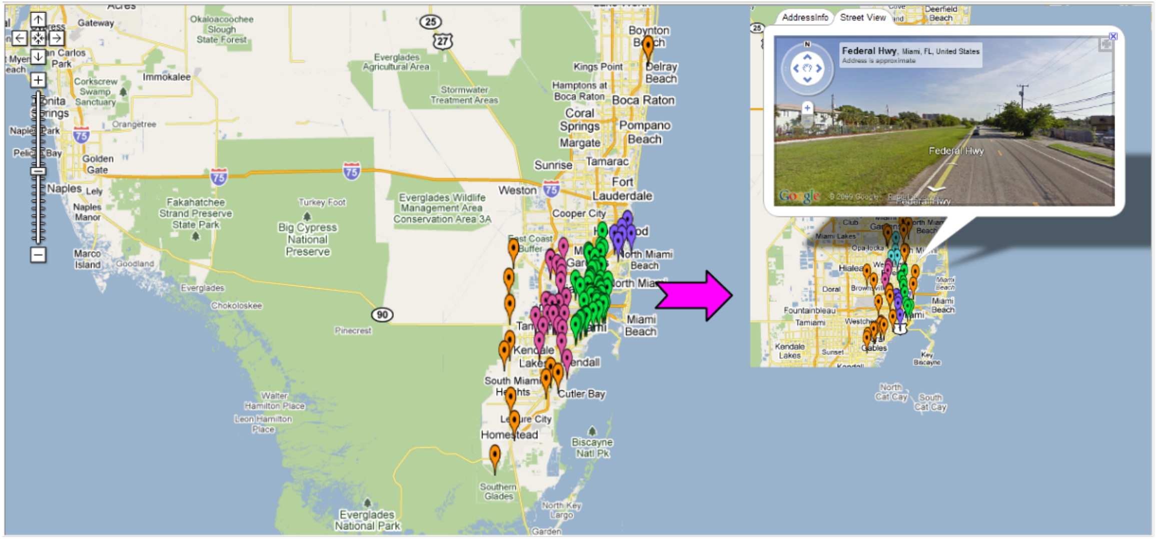


Figure 7.1: Hierarchical representation of spatial clustering with constraints.

Figure 7.1 shows the communities generated by clustering all open facilities and companies in Miami with the constraint: “I75 closed.” In order to deal with un- balanced size of clusters, we provide users with an interactive mechanism to track the subcommunity information within a large size community. Further clustering process will be trig- gered in the runtime when a user selects a larger community and wants to see the cluster information within such a commu- nity at a finer granularity. By using this mechanism, users can obtain clusters with different granularities and more meaningful results. Figure 7.1 shows the interactive clustering results within the largest cluster.

### 7.2.2 User recommendation

To formalize user recommendation service, an interaction or transaction is defined as the process of a user sharing a report with one or more other users [ZST$^+$11]. So, the reports sharing transaction database can be treated as a hypergraph with each node representing a registered user and a set of edges created at the same time from one node to a set of nodes representing an occurred transaction.

There are three important factors associated with each edge: Time: The time that the transaction happened. It indicates the importance of recency. In general, the more recently a transaction happens, the more important the report is to those users involved. Direction: The relation of an interaction. An edge pointed from node A to node B indicating that A shares some information with a set of users including B. The direction indicates that the shared information is more important to the sender than to receivers. Textual Content: Each transaction is associated with some certain textual content, so the content of an edge means that someone thinks such content is important or related to some group of users. In practice, a personalized user recommendation requires the algorithm to identify potential users who have frequent and active interactions with the sender and are also interested in some certain topics. In completion of two recommendation tasks, we extend both [CC08] and [RBDD$^+$10] by taking the direction, timeliness and textual content of the interaction into consideration to generate: 1) a suggested user list for specific report and 2) a suggested user list for specified seeds (users).

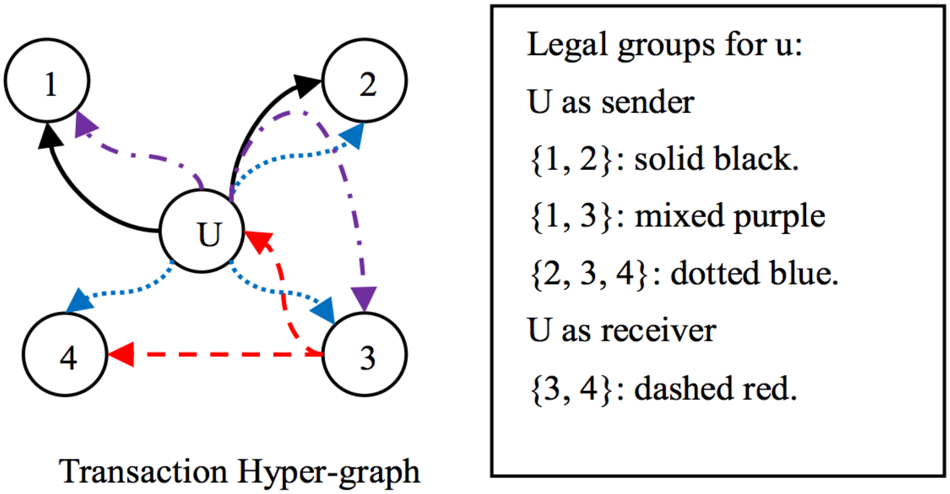


Figure 7.2: Transactional user group.

There could be multiple transactions associated with a specified user and each transaction involves a group of users, as shown in Figure 7.2. Even though transactions may include the same sender and receivers, they are treated as unique in the transactional hyper graph since they are associated with unique timestamps. Despite the textual content of each transaction, the contribution of each group made to current user seeds can be easily evaluated by Interaction Rank proposed in [RBDD$^+$10].

**To build the user profile**, we consider textual content in all transactions related to the user. Carvalho[CC08] introduced a centroid vector-based representation which aggregates all related documents to build a user profile. In our method, we consider transaction directions and assign document sending weight $W_s$ or receiving weight $W_r$ respectively. We use term frequency-inverse document frequency (TF-IDF) transformation to represent textual content as a vector. So the user profile can be represented as:

$$profile(u) = W_s \cdot \sum_{d \in S(u)} tfidf(d) + W_r \cdot \sum_{d \in R(u)} tfidf(d), \tag{7.1}$$

where $tfidf(d)$ is defined as

$$tfidf(d)_i = TFIDF(d)_i^t, \tag{7.2}$$

where $t = \frac{time(now) - time(n)}{\lambda}$ indicates an over-time exponential decay of each document's contribution. $S(u)$, $R(u)$ are sets of documents which sent and received by $u$ respectively. So, for a report $d$, user $u$ s preference to this report can be simply generated by computing the cosine similarity between the users profile and the TF-IDF vector of $d$ as:

$$preference(u, d) = cos(profile(u), ts_{tfidf}(d)). \tag{7.3}$$

Practically, user profile is stored separately and will not be updated in each calculation. Typically, it will be updated when there is new event announced or regularly every few days.

We extended the friend-finding algorithm proposed in [RBDD$^+$10] to generate a list of user recommendations by aggregating the groups contribution to a user and considering

Figure 7.3: Suggesting user routine

**Input:** u, the user; d, the report; S, the seeds
**Output:** R, recommended user list

```
$G \leftarrow GetTransactionlGroups(u)$
**for** $group\ g \in G$ **do**
    **for** $user\ c \in G, c \notin S$ **do**
        **if** $c \notin R$ **then**
            $R[c] \leftarrow 0$
        **end if**
        $R[c] \leftarrow R[c] + GroupScore(c, S, g, d)$
        or $R[c] \leftarrow R[c] + CommunityScore(c, S, g)$
    **end for**
**end for**
```

the relevance between users and reports. Algorithm is described in Figure 7.3. Score of each user in the list represents the interaction preference with respect to the given user and report

The group score or community contribution used in From the algorithm described in Figure 7.3, the interaction preference of a user is the aggregated value of the contribution that each transaction made to the user. There are two types of contribution measurements with respect to different tasks. We use group score and community score to represent contributions for report sharing and community user recommendation respectively.

The group contribution **GC** described below represents the contribution that a user group contributes on the user. There are two situations considered, 1) In order to suggest users related to a document, we consider the preference (similarity) between the document and a user; 2) In order to help user form a meaningful group, we consider the similarity between users. We defined $GC$ as an aggregated score of users' preferences to a specific document considering the direction and timeliness of each interaction.

For the first situation, we use similarities between each user in a group with report $d$:

$$GC(d, g) = W_s \cdot \sum_{i \in O(u,g)} s(i, d)^t + W_r \cdot \sum_{i \in I(u,g)} s(i, d)^t, \tag{7.4}$$

where $s(i,d) = \sum_{u \in i} preference(u,d)$.

For the second situation, we simply modified the $GC(d,g)$ as $GC(c,g)$ and $s(i,d)$ as

$$GC(d,g) = \sum_{u \in i} cos(profile(u), profile(c)), \tag{7.5}$$

to calculate similarity without document information. In both situations, $O(u,g)$ and $I(u,g)$ are sets of sending and receiving interactions/transactions which user $u$ was involved.

**Recommending a report to group of users** involves historic recommendation transactions and the reports textual content. The score that a transaction contributes to a user is the aggregation of preferences of a group of users to the given report:

$$GroupScore(c,S,g,d) = \begin{cases} GC(d,g) & \text{if } S \cap g \notin \phi; \\ 0 & \text{otherwise.} \end{cases}$$

**Recommending users to form communities** involves historic transactions without textual information. The score that a transaction contributes to a user is the aggregation of similarities between the user and users in the group:

$$CommunityScore(c,S,g) = \begin{cases} GC(c,g) & \text{if } S \cap g \notin \phi; \\ 0 & \text{otherwise.} \end{cases}$$

By specifying a couple of users as seeds, our recommendation components can dynamically generate more users related to the given textual content and list of users with high concurrence.

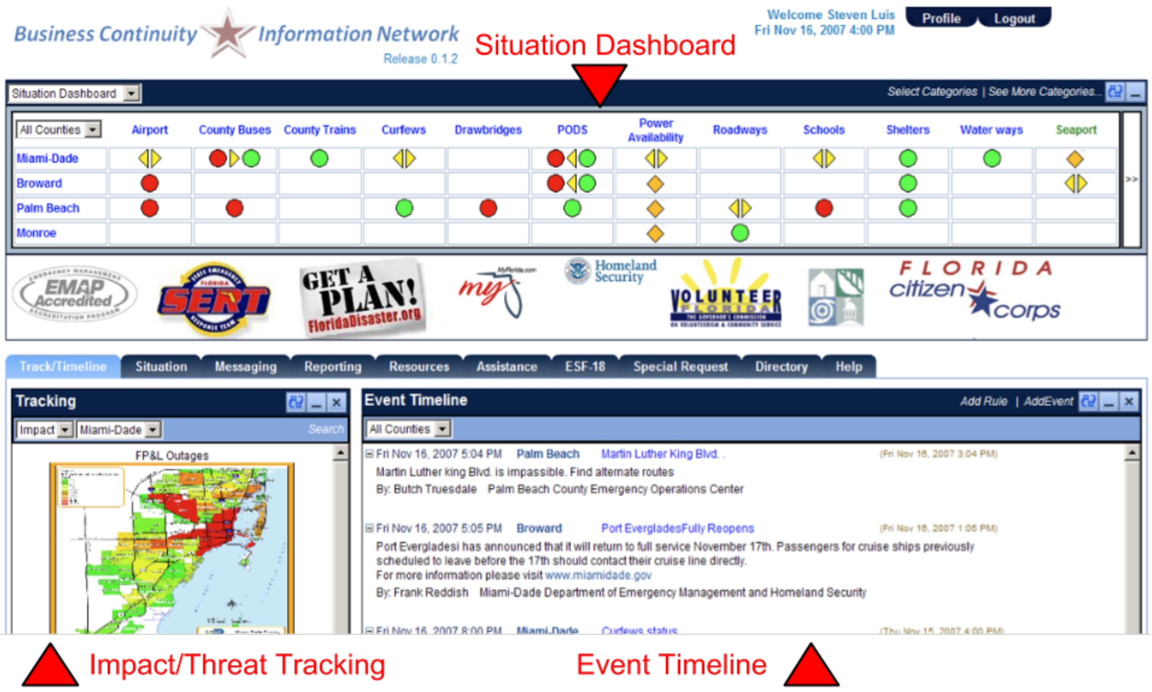
## 7.3 System development

We designed and implemented a web-based prototype of a Business Continuity Information Network (BCiN) that is able to link participating companies into a community network, provide businesses with effective and timely disaster recovery information, and facilitate collaboration and information exchange with other businesses and government agencies. We also designed and implemented an All-Hazard Disaster Situation Browser (ADSB) system that runs on Apples mobile operating system (iOS), and iPhone and iPad mobile devices. Both systems utilize the data processing power of advanced information technologies for disaster planning and recovery under hurricane scenarios. They can help people discover, collect, organize, search, and disseminate real-time disaster information [ZST$^+$10, HCL$^+$10]. This study introduces a unified framework that systematically integrates the different techniques developed in [ZST$^+$10] and [ZST$^+$11]. The idea is that such a framework can be utilized when dealing with different systems or applications separately (e.g., BCiN and ADSB), and hopefully can be easily applied to other scenarios having critical information sharing and management needs.

### 7.3.1 Business continuity information network (BCiN)

BCiN is a platform of information sharing, integration, extraction, and processing for disaster management and recovery. It is also a data mining solution for disaster management and recovery that is able to process and analyze the data from diverse and heterogeneous information sources of different types (categorical events and continuous data) with different formats (structured and unstructured: database records, document news, reports) [ZST$^+$10]. BCiN system demo is shown in Figure 7.4.

Based on observations we have made during our preliminary research, we have identified several key problems that inhibit better information sharing and collaboration among both private and public sector participants for disaster management and recovery. In this project, we will focus on these problems. ***1. How can the system quickly capture the status report information?*** Participants will communicate status reports through many channels, including direct emails, mailing lists, web pages, press releases, conference calls. It is desirable to capture such status information the minute it is available and prevent redundant



Situation Dashboard

| All Counties | Airport | County Buses | County Trains | Curfews | Drawbridges | PODS | Power Availability | Roadways | Schools | Shelters | Water ways | Seaport |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Miami-Dade | ◀▶ | 🔴▶🟢 | 🟢 | ◀▶ |  | 🔴◀🟢 | ◀▶ |  | ◀▶ | 🟢 | 🟢 | 🔶 |
| Broward | 🔴 |  |  |  |  | 🔴◀🟢 | 🔶 |  |  | 🟢 |  | ◀▶ |
| Palm Beach | 🔴 | 🔴 |  | 🟢 | 🔴 | 🟢 | 🔶 | ◀▶ | 🔴 | 🟢 |  |  |
| Monroe |  |  |  |  |  |  | 🔶 | 🟢 |  |  |  |  |

Event Timeline

All Counties

⊟ Fri Nov 16, 2007 5:04 PM Palm Beach Martin Luther King Blvd. . (Fri Nov 16, 2007 3:04 PM)
Martin Luther king Blvd. is impassible. Find alternate routes
By: Butch Truesdale Palm Beach County Emergency Operations Center

⊟ Fri Nov 16, 2007 5:05 PM Broward Port EvergladesFully Reopens (Fri Nov 16, 2007 1:05 PM)
Port Evergladesi has announced that it will return to full service November 17th. Passengers for cruise ships previously scheduled to leave before the 17th should contact their cruise line directly.
For more information please visit www.miamidade.gov
By: Frank Reddish Miami-Dade Department of Emergency Management and Homeland Security

⊟ Fri Nov 16, 2007 8:00 PM Miami-Dade Curfews status (Thu Nov 15, 2007 4:00 PM)

Figure 7.4: BCiN system components.

reporting. To facilitate the reuse of such materials, users can upload status information in the form of unstructured documents such as plain text, Adobe PDFs, and Microsoft DOC. It is thus necessary to identify the useful information in the documents.

***2. How can the system effectively understand the situation from a large collection of reports?*** In larger organizations, or in cases where there is a large accumulation of companies in an area, like a corporate park, reports about a particular area can be redundant. It is important to build a summarized view to understand the situation users are interested from these reports.

***3. How can we automatically capture user interests and effectively deliver the relevant information to the users?*** The status reports are collected through many different channels and are concerned about different categories. During disaster preparation and recovery, users typically don't have the time and patience to go through the system to find the information they want.

*4. How can we take advantage of the community information for disaster recovery?* Participating companies and organizations interact in different communities, such as being members of the same industry sector, or using the same shipping company. Identifying how participants interact with these communities in a disaster situation is very important since it may reveal information that would be helpful in a recovery scenario.

**BCiN system overview**

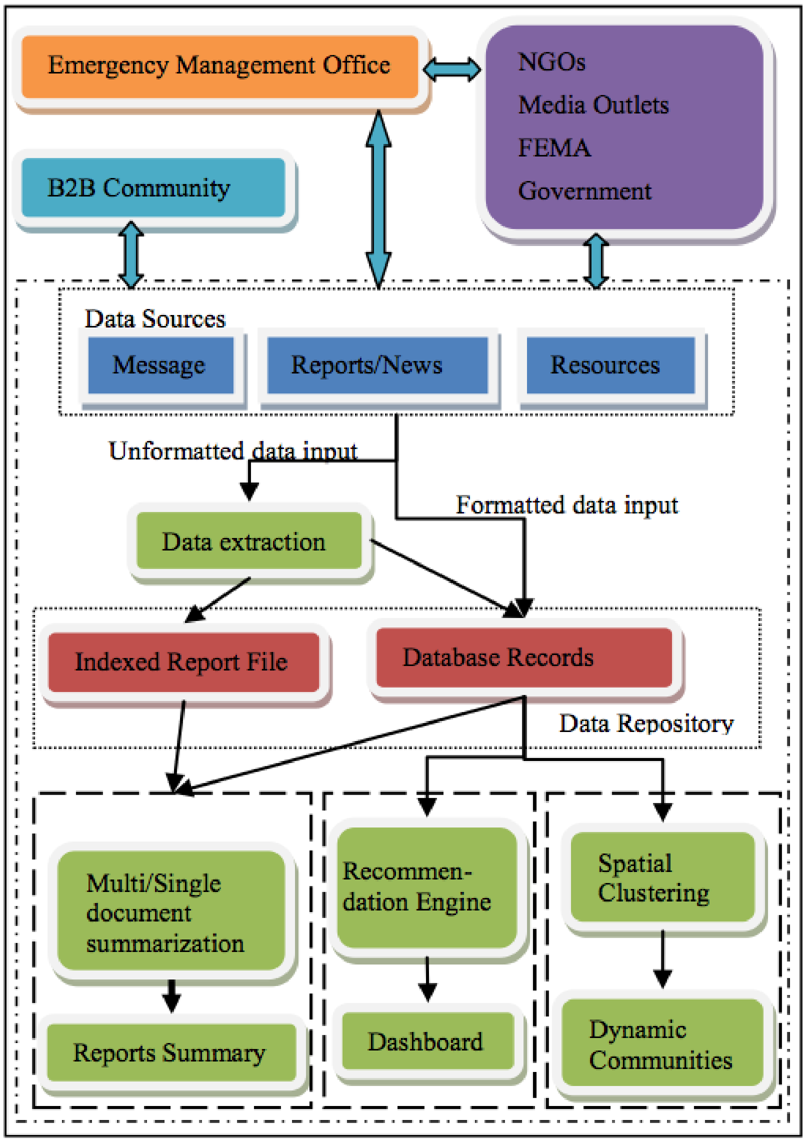


Figure 7.5: BCiN system architecture.

The BCiN system allows company users to submit reports related to their own business, and government users to make announcements on the public issues. To collect more information during the disaster, BCIN can monitor the news published on the websites and takes the news as its input. Like traditional information systems, these reports and news, and the status information of entities they contain can be retrieved and accessed by queries. For example, reports can be viewed according to alert categories or geo- locations, and resources can be viewed according to status or usages. Furthermore, BCIN not only displays users-submitted information but also conducts necessary and meaningful data processing work. BCIN makes recommendations based on the current focus and dynamically adapts based on the users interests. BCIN summarizes reports and news to provide users with brief and content-oriented stories, preventing users from being troubled when searching in huge amount of information. By introducing the concept of Community, BCIN offers users a hierarchical view of important reports or events around them.

In this system, we discuss the following four main information processing and representation components: **Information Extraction**, **Dynamic Dashboard**, **Report Summarization**, and **Dynamic Community Generation**. These four components utilize and develop data mining and machine learning techniques and apply them to disaster management and recovery. The system architecture is shown in Figure 7.5.

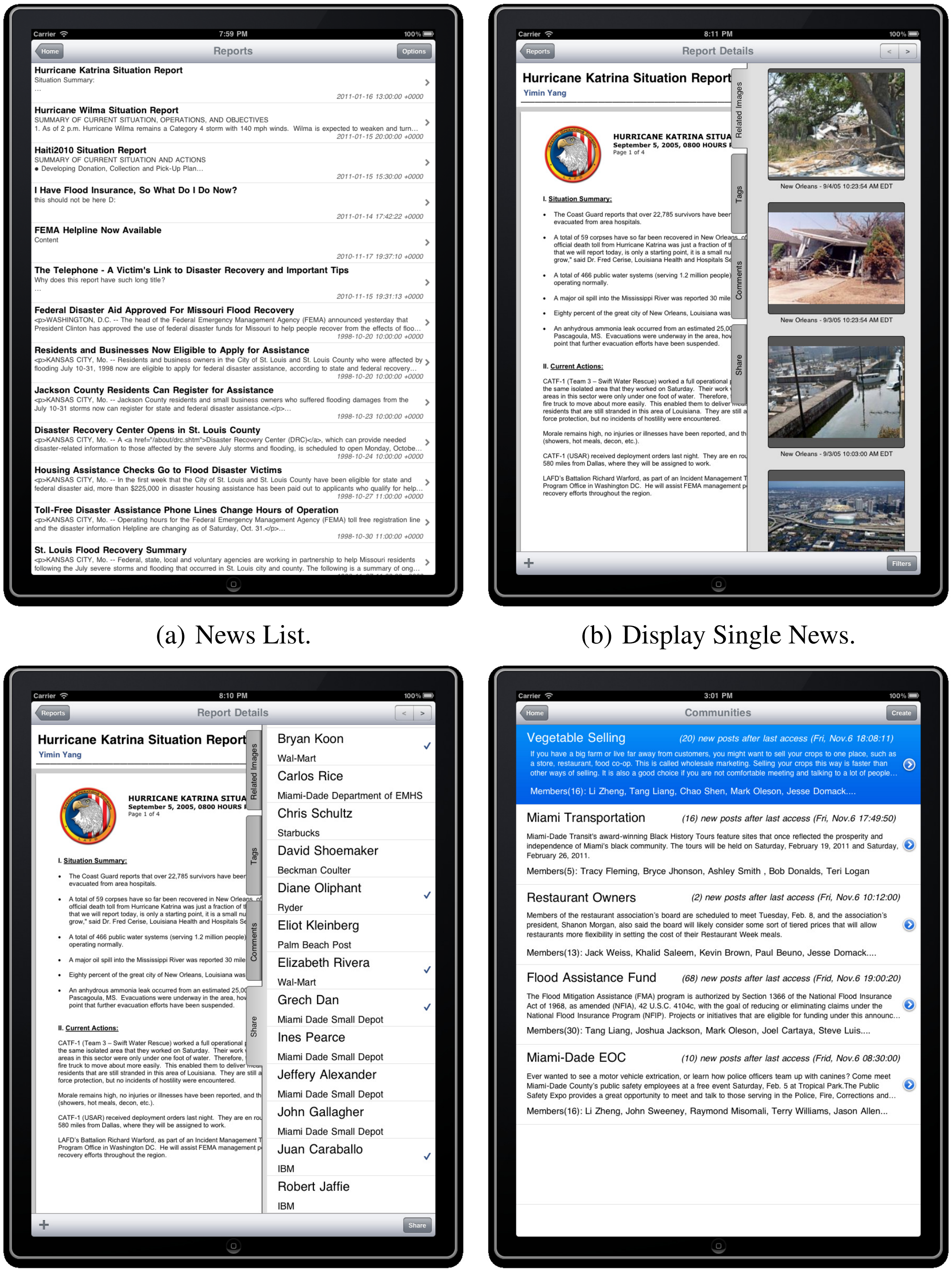
**Information Extraction:** As a data pre-processing component, we adopt sequence tagging and classification methods to extract the structured information from text to integrate different input without a unified format. The detailed approaches used for information extraction are presented in Chapter 5.

**Dynamic Dashboard:** In order to improve the relevance of information to match the users interests we have created a user interface capability called the Dynamic Dashboard. The dynamic dashboard analyzes user interactions such as what kinds of reports the participant has submitted and viewed and automatically recommends similar information to display on the dashboard. The dynamic dashboard provides with the users a convenient and fast approach to obtain the disaster information that they probably want during the emergent time. The dynamic dashboards content is personalized with the concerns of different users. The main contributions of the dynamic dashboard lie in two parts. 1) It automatically removes the redundant companies reports, news and other information by clustering methods. 2) It ranks the information by both the relevance to the current user and the importance of information. The details is beyond the scope of my dissertation.

**Report Summarization:** the BCIN system provides users a report summary which is generated from multiple reports to show the updated changes about the process of the disaster. In the summarization process, structured information extracted from text and stored in the database is used to generate the summary to reflect the latest and changed status of an entity. Details of summarization approaches are beyond the scope of my dissertation.

**Community Generation:** Participating companies and organizations interact in different communities, such as being members of the same industry sector, or using the same shipping company. Identifying how participants interact with these communities in a disaster situation is very important since it may reveal information that would be helpful in a recovery scenario. Using spatial relationship techniques we can track assets like facilities, or equipment, which are important to the participants. The geolocation of such participants can be organized into dynamic communities, and these communities can be informed about events or activities relevant to their spatial footprints. By generating dynamic communities, users can directly select those events happening around them and make more efficient and accurate decision. We adapt spatial clustering in an interactive way to provide users a multi-level view of related communities. We apply spatial clustering with constraints to generate communities that are geographically related, the details are described in Section 7.2.

These components are tightly integrated to provide a cohesive set of services and constitute a holistic data-driven solution for disaster management and recovery.



**Hurricane Katrina Situation Report**
Situation Summary:
...
*2011-01-16 13:00:00 +0000*

**Hurricane Wilma Situation Report**
SUMMARY OF CURRENT SITUATION, OPERATIONS, AND OBJECTIVES
1. As of 2 p.m. Hurricane Wilma remains a Category 4 storm with 140 mph winds. Wilma is expected to weaken and turn...
*2011-01-15 20:00:00 +0000*

**Haiti2010 Situation Report**
SUMMARY OF CURRENT SITUATION AND ACTIONS
● Developing Donation, Collection and Pick-Up Plan...
*2011-01-15 15:30:00 +0000*

**I Have Flood Insurance, So What Do I Do Now?**
this should not be here D:
*2011-01-14 17:42:22 +0000*

**FEMA Helpline Now Available**
Content
*2010-11-17 19:37:10 +0000*

**The Telephone - A Victim's Link to Disaster Recovery and Important Tips**
Why does this report have such long title?
...
*2010-11-15 19:31:13 +0000*

**Federal Disaster Aid Approved For Missouri Flood Recovery**
<p>WASHINGTON, D.C. -- The head of the Federal Emergency Management Agency (FEMA) announced yesterday that President Clinton has approved the use of federal disaster funds for Missouri to help people recover from the effects of floo...
*1998-10-20 10:00:00 +0000*

**Residents and Businesses Now Eligible to Apply for Assistance**
<p>KANSAS CITY, Mo. -- Residents and business owners in the City of St. Louis and St. Louis County who were affected by flooding July 10-31, 1998 now are eligible to apply for federal disaster assistance, according to state and federal recovery...
*1998-10-20 10:00:00 +0000*

**Jackson County Residents Can Register for Assistance**
<p>KANSAS CITY, Mo. -- Jackson County residents and small business owners who suffered flooding damages from the July 10-31 storms now can register for state and federal disaster assistance.</p>...
*1998-10-23 10:00:00 +0000*

**Disaster Recovery Center Opens in St. Louis County**
<p>KANSAS CITY, Mo. -- A <a href="/about/drc.shtm">Disaster Recovery Center (DRC)</a>, which can provide needed disaster-related information to those affected by the severe July storms and flooding, is scheduled to open Monday, Octobe...
*1998-10-24 10:00:00 +0000*

**Housing Assistance Checks Go to Flood Disaster Victims**
<p>KANSAS CITY, Mo. -- In the first week that the City of St. Louis and St. Louis County have been eligible for state and federal disaster aid, more than $225,000 in disaster housing assistance has been paid out to applicants who qualify for help...
*1998-10-27 11:00:00 +0000*

**Toll-Free Disaster Assistance Phone Lines Change Hours of Operation**
<p>KANSAS CITY, Mo. -- Operating hours for the Federal Emergency Management Agency (FEMA) toll free registration line and the disaster information Helpline are changing as of Saturday, Oct. 31.</p>...
*1998-10-30 11:00:00 +0000*

**St. Louis Flood Recovery Summary**
<p>KANSAS CITY, Mo. -- Federal, state, local and voluntary agencies are working in partnership to help Missouri residents following the July severe storms and flooding that occurred in St. Louis city and county. The following is a summary of ong...

(a) News List.

**Hurricane Katrina Situation Report**
Yimin Yang

**HURRICANE KATRINA SITUA**
**September 5, 2005, 0800 HOURS F**
Page 1 of 4

I. **Situation Summary:**

- The Coast Guard reports that over 22,785 survivors have beer evacuated from area hospitals.
- A total of 59 corpses have so far been recovered in New Orleans, of official death toll from Hurricane Katrina was just a fraction of t that we will report today, is only a starting point, it is a small nu grow," said Dr. Fred Cerise, Louisiana Health and Hospitals Se
- A total of 466 public water systems (serving 1.2 million people) operating normally.
- A major oil spill into the Mississippi River was reported 30 mile
- Eighty percent of the great city of New Orleans, Louisiana was
- An anhydrous ammonia leak occurred from an estimated 25,00 Pascagoula, MS. Evacuations were underway in the area, how point that further evacuation efforts have been suspended.

II. **Current Actions:**

CATF-1 (Team 3 – Swift Water Rescue) worked a full operational p the same isolated area that they worked on Saturday. Their work w areas in this sector were only under one foot of water. Therefore, t fire truck to move about more easily. This enabled them to deliver mea residents that are still stranded in this area of Louisiana. They are still a force protection, but no incidents of hostility were encountered.

Morale remains high, no injuries or illnesses have been reported, and th (showers, hot meals, decon, etc.).

CATF-1 (USAR) received deployment orders last night. They are en rou 580 miles from Dallas, where they will be assigned to work.

LAFD's Battalion Richard Warford, as part of an Incident Management T Program Office in Washington DC. He will assist FEMA management p recovery efforts throughout the region.

New Orleans - 9/4/05 10:23:54 AM EDT

New Orleans - 9/3/05 10:23:54 AM EDT

New Orleans - 9/3/05 10:03:00 AM EDT

(b) Display Single News.

**Hurricane Katrina Situation Report**
Yimin Yang

**HURRICANE KATRINA SITUA**
**September 5, 2005, 0800 HOURS F**
Page 1 of 4

I. **Situation Summary:**

- The Coast Guard reports that over 22,785 survivors have been evacuated from area hospitals.
- A total of 59 corpses have so far been recovered in New Orleans, of official death toll from Hurricane Katrina was just a fraction of t that we will report today, is only a starting point, it is a small nu grow," said Dr. Fred Cerise, Louisiana Health and Hospitals Se
- A total of 466 public water systems (serving 1.2 million people) operating normally.
- A major oil spill into the Mississippi River was reported 30 mile
- Eighty percent of the great city of New Orleans, Louisiana was
- An anhydrous ammonia leak occurred from an estimated 25,00 Pascagoula, MS. Evacuations were underway in the area, how point that further evacuation efforts have been suspended.

II. **Current Actions:**

CATF-1 (Team 3 – Swift Water Rescue) worked a full operational p the same isolated area that they worked on Saturday. Their work w areas in this sector were only under one foot of water. Therefore, t fire truck to move about more easily. This enabled them to deliver mea residents that are still stranded in this area of Louisiana. They are still a force protection, but no incidents of hostility were encountered.

Morale remains high, no injuries or illnesses have been reported, and th (showers, hot meals, decon, etc.).

CATF-1 (USAR) received deployment orders last night. They are en rou 580 miles from Dallas, where they will be assigned to work.

LAFD's Battalion Richard Warford, as part of an Incident Management T Program Office in Washington DC. He will assist FEMA management p recovery efforts throughout the region.

| | | |
|---|---|---|
| Bryan Koon | Wal-Mart | ✓ |
| Carlos Rice | Miami-Dade Department of EMHS | |
| Chris Schultz | Starbucks | |
| David Shoemaker | Beckman Coulter | |
| Diane Oliphant | Ryder | ✓ |
| Eliot Kleinberg | Palm Beach Post | |
| Elizabeth Rivera | Wal-Mart | ✓ |
| Grech Dan | Miami Dade Small Depot | ✓ |
| Ines Pearce | Miami Dade Small Depot | |
| Jeffery Alexander | Miami Dade Small Depot | |
| John Gallagher | Miami Dade Small Depot | |
| Juan Caraballo | IBM | ✓ |
| Robert Jaffie | IBM | |

(c) Share News with Community.

**Vegetable Selling** *(20) new posts after last access (Fri, Nov.6 18:08:11)*
If you have a big farm or live far away from customers, you might want to sell your crops to one place, such as a store, restaurant, food co-op. This is called wholesale marketing. Selling your crops this way is faster than other ways of selling. It is also a good choice if you are not comfortable meeting and talking to a lot of people...
Members(16): Li Zheng, Tang Liang, Chao Shen, Mark Oleson, Jesse Domack....

**Miami Transportation** *(16) new posts after last access (Fri, Nov.6 17:49:50)*
Miami-Dade Transit's award-winning Black History Tours feature sites that once reflected the prosperity and independence of Miami's black community. The tours will be held on Saturday, February 19, 2011 and Saturday, February 26, 2011.
Members(5): Tracy Fleming, Bryce Jhonson, Ashley Smith , Bob Donalds, Teri Logan

**Restaurant Owners** *(2) new posts after last access (Fri, Nov.6 10:12:00)*
Members of the restaurant association's board are scheduled to meet Tuesday, Feb. 8, and the association's president, Shanon Morgan, also said the board will likely consider some sort of tiered prices that will allow restaurants more flexibility in setting the cost of their Restaurant Week meals.
Members(13): Jack Weiss, Khalid Saleem, Kevin Brown, Paul Beuno, Jesse Domack....

**Flood Assistance Fund** *(68) new posts after last access (Frid, Nov.6 19:00:20)*
The Flood Mitigation Assistance (FMA) program is authorized by Section 1366 of the National Flood Insurance Act of 1968, as amended (NFIA), 42 U.S.C. 4104c, with the goal of reducing or eliminating claims under the National Flood Insurance Program (NFIP). Projects or initiatives that are eligible for funding under this announc...
Members(30): Tang Liang, Joshua Jackson, Mark Oleson, Joel Cartaya, Steve Luis....

**Miami-Dade EOC** *(10) new posts after last access (Frid, Nov.6 08:30:00)*
Ever wanted to see a motor vehicle extrication, or learn how police officers team up with canines? Come meet Miami-Dade County's public safety employees at a free event Saturday, Feb. 5 at Tropical Park.The Public Safety Expo provides a great opportunity to meet and talk to those serving in the Police, Fire, Corrections and...
Members(16): Li Zheng, John Sweeney, Raymond Misomali, Terry Williams, Jason Allen...

(d) Community Generation.

Figure 7.6: ADSB system components.

### 7.3.2 All-hazadous disaster situation browser (ADSB)

ADSB is a collaborative solution on mobile platform designed for information sharing, integration, extraction, and processing. It can help the user efficiently identify, organize, and deliver important information. In ADSB, registered users can submit reports by typing plain texts as well as attach resources of other formats such as PDF and Doc. The system users can also tag those reports to manage their interested information or post comments to interact with other users. ADSB provides hierarchical summaries generated from user specified keywords to briefly capture important information. Also, a set of suggested query forms helps the users efficiently refine the query results. At last, users can also organize their important friends into groups according to different information management tasks. Figure 7.6 illustrates the major components of the system. A video demonstration [1] accompanying the dissertation is available for obtaining details of system functionalities.

During prototyping ADSB to integrate those critical features into the mobile platform, we have identified the following three key tasks to fully utilize the advantages and overcome the limitations of major mobile devices.

1. ***Design and develop effective and interactive information summarization methods to help users understand large collection of reports.*** It is typically difficult for readers to extract useful information from a large quantity of documents. Multi-document summarization provides users with a tool to effectively extract important and related ideas of current situations. However, previous text summarization techniques gave users a set of sentences based on user query. The summarization is fixed once the query is determined. Note that mobile devices are generally with a small display and limited input capabilities. An interactive summarization interface is needed to help users navigate collected information at different granularities, and locate their target information more efficiently

[1] http://users.cis.fiu.edu/ lzhen001/demo/demo.htm

2. ***Design and develop intelligent information delivery techniques to help users quickly identify the information they need.*** The data is collected through many different channels and belongs to different categories. During disaster preparation and recovery, users do not have the time and patience to go through the system to find the information they want. Structured information can be of important value to help people make decisions by providing them with actionable and concrete information representation and exploration. However, navigating the large result set on the mobile device is particularly inefficient. An interactive tabular interface can largely help users filter useful information by adapting changing query conditions and user feedbacks.

3. ***Design and develop dynamic community generation techniques for reports recommendation and user group organization.*** In information sharing tasks, identifying a group of recipients to which a certain type of information is conveyed to can highly improve the efficiency of communication and gain valuable feedback. But on mobile device, managing the groups of friends within the limited display often makes user miss highly related friends. User recommendation techniques can offer a user such convenience by automatically and interactively generating potential recipients for different pieces of information. In addition, user recommendation techniques can help users effectively and dynamically organize user groups according to various information sharing tasks.

### ADSB system overview

ADSB adopts the open source REST (REpresentational State Transfer) framework named Restlet which is a lightweight, comprehensive and fully Java implemented web architecture model designed for both server and client Web applications[Wik, Fie00]. The implementation of the ADSB API is entirely HTTP-based and follows CRUD (Create, Read, Update and Delete) rules by specifying a corresponding HTTP response code. As a Restful resource, ADSB API supports both XML and JSON formats. Due to the simplicity and flexibility of Restlet framework, ADSB API allows us to: 1) Conveniently interact with multiple information domains; 2) Quickly create components and functions based on information management processes; 3) Improve end-user programmability and configurability; 4) Can be easily released to third party clients to embed our data service into different application.

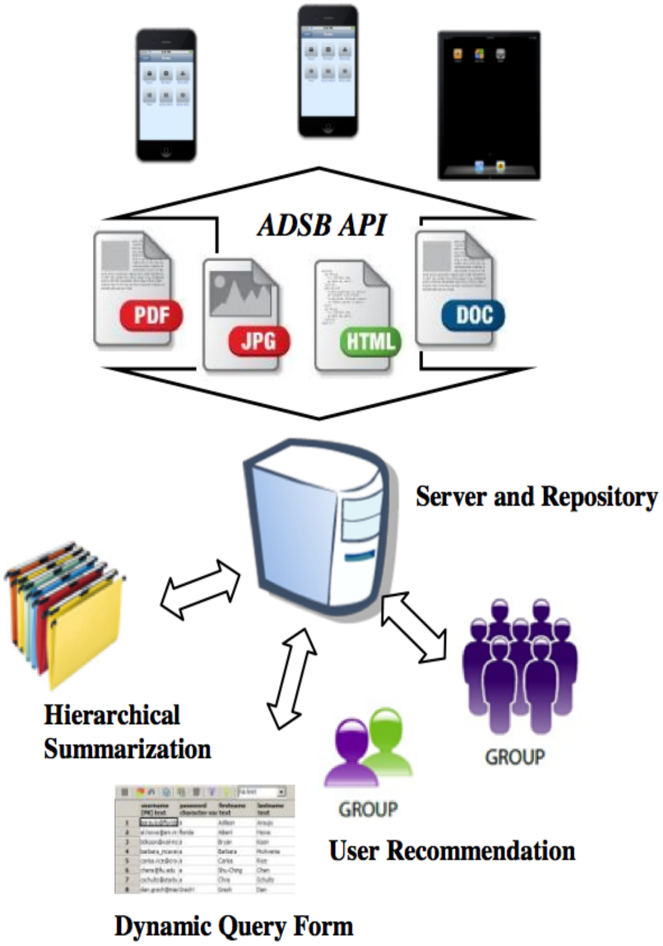


Figure 7.7: ABSB system architecture.

The above-mentioned system information processing and representation functionalities are integrated with the following three critical modules: **Hierarchical Summarization**,

**Dynamic Query Form**, and **User Recommendation**. The system architecture is shown in Figure 7.7.

**Hierarchical Summarization:** ADSB system provides users with reports summaries which are generated from multiple reports. The Affinity propagation method is applied on the sentence similarity graph to build hierarchical summaries in an agglomerative way. The exemplar generated by affinity propagation for each sub- cluster can be used as a summary of that cluster. Details of summarization approaches are beyond the scope of this dissertation.

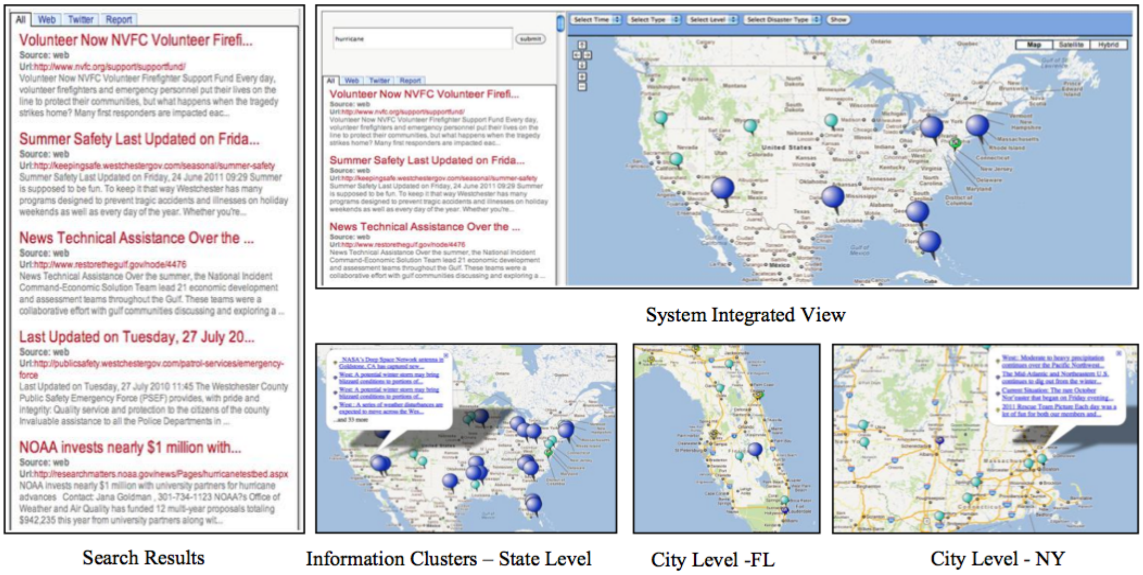
**Dynamic Query Form:** After obtaining document graph and attribute graph which represent relationships among document set and attribute set respectively, we iteratively calculate similarities between documents and attributes separately by running the random walk model. The suggested query condition can be generated based on each given document and previously selected attributes. Details of dynamic query form is beyond the scope of this dissertation.

**User Recommendation:** ADSB provides an interface for users to share a single report with other people. Such sharing transactions are good indications of users preferences and can help us identify meaningful users groups. We utilize the transactional hyper-graph and the textual content to generate the suggested user list by ranking the interaction preference of each user based on the given report and the selected user seeds. The details are described in Section 7.2.

These modules are tightly integrated to provide a cohesive set of services and constitute a holistic effort on developing a data- driven solution for disaster management and recovery.

### 7.3.3 Disaster situation browser (SitRep)

Disaster Situation Reporting System (Disaster SitRep), shown in Figure 7.8, is essentially a disaster information collecting, integration, and presentation platform. It is implemented to address three critical tasks that can facilitate information acquisition, integration and presentation by utilizing domain knowledge as well as public and private web resources for major disaster recovery planning and management [ZST$^+$12].



All | Web | Twitter | Report

Volunteer Now NVFC Volunteer Firefi...
Source: web
Url:http://www.nvfc.org/support/supportfund/
Volunteer Now NVFC Volunteer Firefighter Support Fund Every day, volunteer firefighters and emergency personnel put their lives on the line to protect their communities, but what happens when the tragedy strikes home? Many first responders are impacted eac...

Summer Safety Last Updated on Frida...
Source: web
Url:http://keepingsafe.westchestergov.com/seasonal/summer-safety
Summer Safety Last Updated on Friday, 24 June 2011 09:29 Summer is supposed to be fun. To keep it that way Westchester has many programs designed to prevent tragic accidents and illnesses on holiday weekends as well as every day of the year. Whether you're...

News Technical Assistance Over the ...
Source: web
Url:http://www.restorethegulf.gov/node/4476
News Technical Assistance Over the summer, the National Incident Command-Economic Solution Team lead 21 economic development and assessment teams throughout the Gulf. These teams were a collaborative effort with gulf communities discussing and exploring a ...

Last Updated on Tuesday, 27 July 20...
Source: web
Url:http://publicsafety.westchestergov.com/patrol-services/emergency-force
Last Updated on Tuesday, 27 July 2010 11:45 The Westchester County Public Safety Emergency Force (PSEF) provides, with pride and integrity: Quality service and protection to the citizens of the county Invaluable assistance to all the Police Departments in ...

NOAA invests nearly $1 million with...
Source: web
Url:http://researchmatters.noaa.gov/news/Pages/hurricanetestbed.aspx
NOAA invests nearly $1 million with university partners for hurricane advances Contact: Jana Goldman , 301-734-1123 NOAA?s Office of Weather and Air Quality has funded 12 multi-year proposals totaling $942,235 this year from university partners along wit...

System Integrated View

Search Results

Information Clusters – State Level

City Level -FL

City Level - NY

Figure 7.8: Disaster SitRep system components.

The following three key tasks have been identified to fully utilize the advantages and overcome the shortcomings of traditional general search and information management platform that have never been applied to disaster management domain.

1. ***Design and develop effective and dynamic concept hierarchy generation and reuse methods in disaster management domain to help the domain experts, the crawler and search engine behave efficiently in situation.*** Concept hierarchy, as means of formalizing and sharing knowledge, provides domain experts and knowledge engineers support for modeling specific domain of the world and can be applied in various areas to implement intelligent knowledge and information management system. However, building the hierarchy from scratch is a costly process that requires massive human labor, so automatically improving concept hierarchy generation and

reuse becomes a challenging but critical task. Combining existing hierarchy with concepts extracted from Semantic Web contents largely helps to extend and enrich existing structural concepts in a given domain.

2. ***Design and develop intelligent focused web crawling techniques to manage the data acquiring process and to increase the information coverage and relevance in disaster domain.*** Heterogeneous data collected from various sources bring difficulties to assimilate information at different levels. The strategies for general purpose search engine will lead to many irrelevant web pages being indexed and also the seeds set will be expanded unexpectedly. Intelligent crawling strategies are needed to systematically control the crawling process to guarantee the indexed web contents with high quality and relevance. Also the given seeds can be expanded to a certain level and finally converge to a good seeds list. On the other hand, the query results are required to be personalized to remove duplicity and increase diversity.

3. ***Design and develop data integration techniques for disaster events identification and extraction.*** In disaster situation, many recovery processes are running in a confused mass. Undergoing activities and important situations are hard to detect from many information channels in unformatted patterns. How to understand the information and organize useful knowledge in a unified manner becomes especially helpful for government officials, disaster management agents, business continuity staff, and even public users suffer from disorders during disaster recovery phases. After getting related information from the web, particular techniques need to be designed to integrate the raw data into certain format that are ready to be used by the search engine and topic visualization modules.

### Disaster SitRep system overview

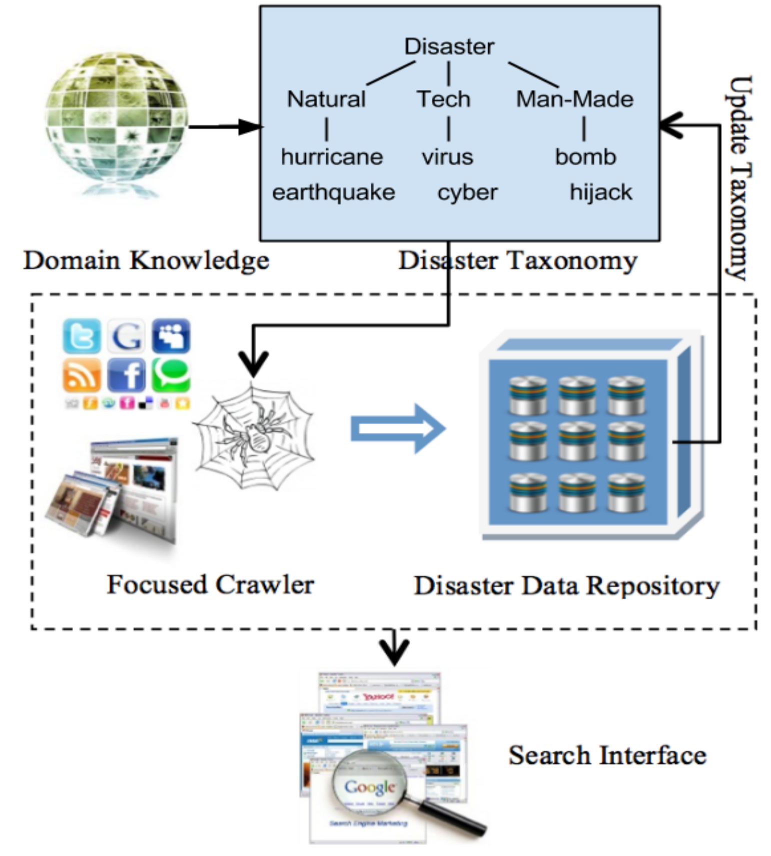


Figure 7.9: Disaster SitRep system architecture.

The complete disaster management domain vertical search engine will be decomposed into 3 major components shown in Figure 7.9:

**Taxonomy Generation**: Based on our cooperation with domain experts, we initialize fundamental disaster taxonomy from disaster expertise. As the system keeps running, more web contents are crawled and extracted from unforeseen sources and new disaster terminologies are dynamically generated and are appended to the existing taxonomy. We propose a semi-supervised hierarchical clustering algorithm to enrich and modified previous taxonomy. Details of taxonomy generation and extension approaches are discussed in Section 5.1 in Chapter 5.

**Foucsed Crawling**: Our focused crawler is implemented to discover more disaster information by intelligently traversing the web contents based on their relevance to ongoing disasters. Usually, the more a web page is related to a certain topic, the higher probability it contains more resources (including hyperlinks to other web pages or possibly relevant concepts) in the same domain. The disaster taxonomy in the previous stage can be utilized to classify web pages into various disaster categories. In general, there are two levels of judgments that help scoring the relevance of a page:

- Web Page Classifier: The classifier adopts hierarchical classification strategy to automatically categorize a crawled web page into different aspects according to the disaster taxonomy or simply report that current web page is irrelevant to any disaster topic.
- Queue Prioritizer: From the categorization results, the focused crawler adjusts the priority of each web page in the queue to guarantee that the most related web resource will be accessed earliest during the crawling process.

Combining these two functionalities, the focused crawling module attempts to assign the most relevant web page with the highest score to make sure such resource can be downloaded earliest. By properly designing those two parts, the crawler can access more related web resources by accessing fewer web pages. Also, as we crawl more disaster related content, it can largely contribute to extend our current taxonomy by including more concepts. Details are discussed in Section 5.2 in Chapter 5.

**Event Extraction**: Textual documents and situation reports crawled from the websites do not usually provide actionable information immediately, such as time, location, status, etc. The replication of information from various resources also challenges the search engine to provide highly related and diversified content to users. To gain further insight about the disaster event rather than a collection of textual documents, we need a domain- oriented skeleton for each type of disasters. The domain- oriented skeleton is the set of structural attributes that we try to extract from disaster documents. This part is beyond the scope of this disseration.

## 7.4 System operation and evaluation

Through a series of interviews with public and private sector partners we identified the specific information both side could share and needed as part of their preparedness and recovery processes. The system then functionally established four key capabilities: Messaging, Reporting, Resources, Situational Browsing, so we can do things like alert a user via messages that a particular resource has been reported available at a local business. The proposed enhancements to the base system we have discussed in this chapter. The system provides new ways to connect reports, with resources, and the people/communities that need it.

FIU has spent over $600K in the development of the application and has received over $400K in sponsored research or industry donation. The system is monitored 24/7 via scripts that verify application, database, web server, and hardware availability. The system is managed in a revision control system and is run through a test suite that validates key functionality such as report submission, field validation, and role based access control. Over 100 companies (local and national) and government agencies in the south Florida area are utilizing the system,working closely with County emergency managers to collaborate on their mutual interest of disaster preparedness, response and recovery. The private sector benefits by receiving timely, accurate, information which impacts business operations and has the ability to report in situational information regarding disaster impact and infrastructure needs which are a priority for their business resumption. The public sector benefits by helping the business community receive and better understand disaster related information and can use disaster related situational reports from private sector to make better assessment of disaster impact.

Before the deployment of the BCIN, in a disaster situation, simultaneous reports from thousands of participants would overwhelm participants, making it very difficult to assess the status without dedicating a significant amount of time by all parties to process this potentially huge volume of information.

Using the proposed information extraction and report summarization techniques, the flooding status of an important commerce area such as Dadeland can be determined even if there are 1,000 companies providing status information. For instance, if these Dadeland based companies each logs into the system and enters a Flood report, or uploads a relevant document that contains relevant information, in an unstructured format, such as flooding area, depth, and public safety issues (nearby canals, down powerlines). The exercise has shown that the proposed techniques are able to identify critical common features of the flooding and summarize these, providing situational reporting in the Dynamic Dashboard. Further, if many of these companies are displaced by the damage, Dynamic Community Generation can inform community members about logistical concerns or assistance opportunities available.

Up to now, BCIN has been exercised at Miami-Dade County Emergency Management for the hurricane disaster management and recovery for three times. Miami-Dade, Florida is a very concentrated urban area (4th largest in the US), with tens of thousands of commercial concerns in a 25 square mile area. Miami-Dade County Emergency Management is interested in assisting this large, diverse business ecosystem to prepare and recovery quickly from hurricane impact.

| Date | Description of the Exercise |
|---|---|
| Jun. 01 2009 | In Florida Dept. of Emergency management's Statewide Hurricane Exercises, BCiN was utilized in a scenario where Miami-Dade County Emergency Management Business Recovery Desk facilitated the logistics to deploy portable ATMs at Shelters and PODs in Miami-Dade County. |
| Jun. 29 2009 | In Maimi-Dade UASI exercise, BCiN supported communicating and collaborating with several companies that participated in the event as observers. |
| Aug. 20 2009 | In a full scale company BCiN training, about 30 companies were given injects to provide information to resolve different information requests. |
| May 10 2010 | In Maimi-Dade Dept. of Emergency Management's Statewide Hurricane Exercise, out systems were responsible for disseminating and responding to injects during the source of the exercise for both government and company users. |
| Jul. 29 2010 | In Miami-Dade company exercises, over 50 company attendees used our systems for a training exercise. |
| May 12 2011 | In the county of West Palm Beach exercise, we demonstrated the system to WPB Dept. of Emergency Management and companies. |

Table 7.2: Evaluation Exercises.

Our system evaluation process consists of presenting the system to emergency managers, business continuity professionals, and other stakeholders for feedback and performing community exercises. The exercises involve a real-time simulation of a disaster event integrated into an existing readiness exercise conducted each year. This evaluation exposes information at different time intervals and asks the community to resolve different scenarios by using the tool. The evaluation is a form of a table-top exercise in which injected information provides details about the current disaster situation and specifies potential goals and courses of action. Participant use the system to gather information to assess the situation and provide details about the actions they will take. We gather information about what information they found to derive their conclusions (or lack thereof). This information allows us to better understand how those techniques improve the information effectiveness.

Table 7.2 describes the exercises. In a regional disaster such as a hurricane, business continuity professionals are under extreme pressure to execute their continuity of operation plans because many of the usual sources of information and services about the community and supply chain are completely disconnected, sporadic, redundant, and many times lack actionable value. The system focuses user input and collaboration around actionable information that both public and private sector can use.

To validate the usability and performance of our system, the participants and the EOC personnel at Miami-Dade participated in the questionnaire session after the exercise. A set of ten questions was designed to evaluate our system where nine of them are multiple choice questions with a five-level scale (strongly agree, agree, not sure, disagree, and strongly disagree) and the last one is an open-ended question. Some of the multiple choice questions are: Are you able to identify related reports that you are interested in? Are you able to identify the correct modules for your tasks? Are you able to switch between different modules? Are the system generated summaries useful? The open-ended question is about feedback and suggestions from the users. On average, about four EOC personnel and 30 participants attended each exercise. The evaluation demonstrated that most of participants are satisfied with the performance of the tools. Specifically, seven out of nine multiple choice questions received "strongly agree" or "agree" from over 90% of the participants, implying a high level of satisfaction with our system.

The feedback from our users is positive and suggests that our system can be used not only to share the valuable action- able information but to pursue more complex tasks like business planning and decision making. There are also many collaborative missions that can be undertaken on our system, which allows public and private sector entities to leverage their local capacity to serve the recovery of the community. We summarized the feedback as follows.

1) Positive feedback: a) the system is easy to use; b) re-lated reports are well orga-

nized based on personalized user groups; and c) reports summarization is representative and interesting.

2) Some suggestions: a) related multimedia information, including images and video, could be shown during navigation; b) report summaries could be organized based on some points of interests.

## 7.5 Conclusion

We identified four key design challenges to support multi-party coordination during disaster situations. We proposed a unified framework that systematically integrates the different techniques that are developed in our previous work [ZST$^+$11, ZST$^+$10]. Such a framework can be utilized when dealing with different systems or applications separately (e.g., BCiN and ADSB), and they are essentially collaborative platforms for preparedness and recovery that helps disaster impacted communities to better understand what the current disaster situation is and how the community is recovering. The system evaluation results demonstrate the effectiveness and efficiency of our proposed approaches.

During the system implementation and assessment process, the users provided suggestions, limitations and possible enhancements. Our future efforts will be focusing on the following tasks: developing efficient tools to automatically crawl related information from public resources including news portals, blogs, and social Medias; capturing the current users interests and construct appropriate query form; and understanding users intends to provide them with actionable answers to their information inquiries.

CHAPTER 8

# Conclusion and future work

In this dissertation, several algorithms on domain taxonomy generation based on ensemble and semi-supervised hierarchical clustering are discussed and important vertical search engine components including focused crawler, information extraction and user interest modeling are provided. Those implemented techniques can be used to build a domain specific, intelligent, and personalized search engine system.

Particularly, to generate taxonomy efficiently, we introduced and discussed a framework for ensemble hierarchical clusterings based on descriptor matrices and a semi-supervised hierarchical clustering framework based on ultra-metric dendrogram distance. Three important components of the framework are studied, including Dendrogram Selection, Dendrogram Description and Dendrogram Combination. The triple-wise relative constraints are introduced, particularly for hierarchial clustering, to describe the merge preference among instances. Our contributions include the following two aspects:

- We propose two ensemble selection schemes based on tree distances, investigate five different dendrogram descriptor matrices, and develop a novel method for fitting an ultra-metric from the aggregated descriptor matrix. Our descriptor matrices based framework can be naturally generalized to ensemble both partitional clustering and hierarchical clustering results as partitional clustering results can be easily represented using distance matrices.

- Two techniques are developed to solve semi-supervised hierarchical clustering problem. The optimization-based technique minimizes the distance between the original dissimilarity matrix and the target ultra-matrix using the ultra-metricity and relative constraints. The transitive dissimilarity based technique takes those relative constraints into the ultra-metric transformation process.

There are several avenues for future work. First, techniques for scaling up the ensemble process to large-scale datasets will be investigated. Second, our studies show that selecting a relatively smaller subset is likely to produce better ensemble results. One interesting question is how to determine the ensemble size. Another interesting yet related direction is that rather than picking representative dendrograms, we can associate every generated dendrogram with a weight. So when considering the ensemble, dendrograms with larger weights can contribute more than dendrograms with smaller weights. Third, another aspect of interest is to provide a formal analysis on cluster separation enhancement using transitive dissimilarity.

To capture user's interests, we implement a novel personalized news recommendation system, PENETRATE, to provide attractive news reading lists to online readers. Our system takes into consideration the reading behaviors of both individual user and a group of users when performing recommendation. The group behavior shows us the general topics that the user might be interested in, whereas the individual behavior provides us personalized information for further filtering news articles. Extensive empirical results demonstrate the efficacy of our system.

The system can be improved in terms of both accuracy and efficiency. In particular, the time cost of ensemble hierarchical clustering (as introduced in Section 6.2.1) can be further reduced by carefully design, e.g., utilizing distributed frameworks or Map-Reduce programming model. We also plan to incorporate the temporal information into the recommendation paradigm (as introduced in Section 6.2), i.e., the recommendation should be biased to more recent preference of online users.

Finally, we apply techniques in disaster information management domain. We identified four key design challenges to support multi-party coordination during disaster situations. We proposed a unified framework that systematically integrates the different techniques that are developed in our work [ZST$^+$11, ZST$^+$10]. Such a framework can be utilized when dealing with different systems or applications separately (e.g., BCiN and ADSB), and they are essentially collaborative platforms for preparedness and recovery that helps disaster impacted communities to better understand what the current disaster situation is and how the community is recovering. The system evaluation results demonstrate the effectiveness and efficiency of our proposed approaches.

During the system implementation and assessment process, the users provided suggestions, limitations and possible enhancements. Our future efforts will be focusing on the following tasks: developing efficient tools to automatically crawl related information from public resources including news portals, blogs, and social Medias; capturing the current users interests and construct appropriate query form; and understanding users intends to provide them with actionable answers to their information inquiries.

# BIBLIOGRAPHY

[AAGY01] Charu C. Aggarwal, Fatima Al-Garawi, and Philip S. Yu. Intelligent crawling on the world wide web with arbitrary predicates. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pages 96–105, New York, NY, USA, 2001. ACM.

[AAHM00] Eneko Agirre, Olatz Ansa, Eduard Hovy, and David Martínez. Enriching very large ontologies using the www. *arXiv preprint cs/0010026*, 2000.

[ABF$^+$99] Richa Agarwala, Vineet Bafna, Martin Farach, Mike Paterson, and Mikkel Thorup. On the approximability of numerical taxonomy (fitting distances by tree metrics). *SIAM J. Comput.*, pages 1073–1085, 1999.

[ABG$^+$07] J. Ahn, P. Brusilovsky, J. Grady, D. He, and S.Y. Syn. Open user profiles for adaptive news systems: help or harm? In *Proc. of WWW*, pages 11–20. ACM, 2007.

[AC05a] Nir Ailon and Moses Charikar. Fitting tree metrics: Hierarchical clustering and phylogeny. In *In Proceedings of the Symposium on Foundations of Computer Science*, pages 73–82, 2005.

[AC05b] Nir Ailon and Moses Charikar. Fitting tree metrics: Hierarchical clustering and phylogeny. In *In Proceedings of the Symposium on Foundations of Computer Science*, pages 73–82, 2005.

[AC09] D. Agarwal and B. C. Chen. Regression-based latent factor models. In *Proc. of SIGKDD*, pages 19–28, 2009.

[Ada72] E. Adams. Consensus techniques and the comparison of taxonomic trees. In Syst. Zool., pages 390–397, 1972.

[Ada86a] E. Adams. N-trees as nestings: complexity, similarity, and consensus. In Journal of Classification., pages 299–317, 1986.

[Ada86b] Edward N. Adams. N-trees as nestings: Complexity, similarity, and consensus. *Journal of Classification*, 3:299–317, 1986. 10.1007/BF01894192.

[AF09] Javad Azimi and Xiaoli Fern. Adaptive cluster ensemble selection. In *IJCAI'09*, pages 992–997, 2009.

[AS01] Rakesh Agrawal and Ramakrishnan Srikant. On integrating catalogs. In *Proceedings of the 10th international conference on World Wide Web*, pages 603–612. ACM, 2001.

[B.83] Rubin Donald B. Iteratively Reweighted Least Squares. *Encyclopedia of Statistical Sciences*, pages 272–275, 1983.

[BB96] Leo Breiman and Leo Breiman. Bagging predictors. *Machine Learning*, pages 123–140, 1996.

[BB01] Leo Breiman and Leo Breiman. Random forests. *Machine Learning*, pages 5–32, 2001.

[BBM04] Sugato Basu, Mikhail Bilenko, and Raymond J. Mooney. A probabilistic framework for semi-supervised clustering. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 59–68, New York, NY, USA, 2004. ACM.

[BBP$^+$09] Ellen J. Bass, Leigh. Baumgart, Brenda. Philips, Kevin. Kloesel, Kathleen. Dougherty, Havidan. Rodriguez, Walter. Diaz, William R. Donner, Jenniffer. Santos, and Michael Zink. Incorporating emergency management needs in the development of weather radar networks. *Journal of emergency management*, 7(1):45–52, 2009.

[BBPK08] Leigh A. Baumgart, Ellen J. Bass, Brenda Philips, and Kevin Kloesel. Emergency management decision making during severe weather. *Weather and Forecasting*, 23(6):1268–1279, 2014/01/31 2008.

[BC99] Matthew Berland and Eugene Charniak. Finding parts in very large corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 57–64. Association for Computational Linguistics, 1999.

[BhHSW05] Aharon Bar-hillel, Tomer Hertz, Noam Shental, and Daphna Weinshall. Learning a mahalanobis metric from equivalence constraints. *Journal of Machine Learning Research*, 6:937–965, 2005.

[BM98] C. L. Blake and C. J. Merz. UCI repository of machine learning databases, 1998.

[BN06] Korinna Bade and Andreas Nurnberger. Personalized hierarchical clustering. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on*

*Web Intelligence*, WI '06, pages 181–187, Washington, DC, USA, 2006. IEEE Computer Society.

[BN08a] Korinna Bade and Andreas Nrnberger. Creating a cluster hierarchy under constraints of a partially known hierarchy. In *Proceedings of the SIAM International Conference on Data Mining, SDM 2008, April 24-26, 2008, Atlanta, Georgia, USA*, pages 13–24. SIAM, 2008.

[BN08b] Korinna Bade and Andreas Nrnberger. Creating a cluster hierarchy under constraints of a partially known hierarchy. In *SDM'08*, pages 13–24, 2008.

[BNJ03] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[Bol98] DL Boley. Hierarchical taxonomies using divisive partitioning. Technical report, Technical Report TR-98-012, Department of Computer Science, University of Minnesota, Minneapolis, 1998.

[BP99] D. Billsus and M.J. Pazzani. A personal news agent that talks, learns and explains. In *Proc. of AA*, pages 268–275. ACM, 1999.

[Bur05] R. Burke. Hybrid systems for personalized recommendations. *Intelligent Techniques for Web Personalization*, pages 133–152, 2005.

[CC97] Charles LA Clarke and Gordon V Cormack. On the use of regular expressions for searching text. *ACM Transactions on Programming Languages and Systems (TOPLAS)*, 19(3):413–426, 1997.

[CC02] Shui-Lung Chuang and Lee-Feng Chien. Towards automatic generation of query taxonomy: A hierarchical query clustering approach. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pages 75–82. IEEE, 2002.

[CC03a] Pu-Jen Cheng and Lee-Feng Chien. Auto-generation of topic hierarchies for web images from users' perspectives. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 544–547. ACM, 2003.

[CC03b] Shui-Lung Chuang and Lee-Feng Chien. Enriching web taxonomies through subject categorization of query terms from search engine logs. *Decision Support Systems*, 35(1):113–127, 2003.

[CC08] Vitor R. Carvalho and William W. Cohen. Ranking users for intelligent message addressing. In *Proceedings of the IR research, 30th European conference on Advances in information retrieval*, ECIR'08, pages 321–333, Berlin, Heidelberg, 2008. Springer-Verlag.

[CCH02] Patrick Clerkin, Pádraig Cunningham, and Conor Hayes. Ontology discovery for the semantic web using hierarchical clustering. 2002.

[Cha09] Soumen Chakrabarti. Focused web crawling. In *Encyclopedia of Database Systems*, pages 1147–1155. Springer, 2009.

[CL96] Jim Cowie and Wendy Lehnert. Information extraction. *Communications of the ACM*, 39(1):80–91, 1996.

[CN02a] Hai Leong Chieu and Hwee Tou Ng. A maximum entropy approach to information extraction from semi-structured and free text. *AAAI/IAAI*, 2002:786–791, 2002.

[CN02b] Hai Leong Chieu and Hwee Tou Ng. Named entity recognition: a maximum entropy approach using global information. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics, 2002.

[CP09] W. Chu and S.T. Park. Personalized recommendation on dynamic content using predictive bilinear models. In *Proc. of WWW*, pages 691–700. ACM, 2009.

[CPS02] Soumen Chakrabarti, Kunal Punera, and Mallela Subramanyam. Accelerated focused crawling through online relevance feedback. In *Proceedings of the 11th international conference on World Wide Web*, WWW '02, pages 148–159, New York, NY, USA, 2002. ACM.

[CS04] Aron Culotta and Jeffrey Sorensen. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 423. Association for Computational Linguistics, 2004.

[CTW$^+$91] T.M. Cover, J.A. Thomas, J. Wiley, et al. *Elements of information theory*, volume 6. Wiley Online Library, 1991.

[CvdBD99] Soumen Chakrabarti, Martin van den Berg, and Byron Dom. Focused crawling: a new approach to topic-specific web resource discovery. In *Proceed-*

*ings of the eighth international conference on World Wide Web*, WWW '99, pages 1623–1640, New York, NY, USA, 1999. Elsevier North-Holland, Inc.

[DBP94] P. M. E. De Bra and R. D. J. Post. Information retrieval in the world-wide web: making client-based searching feasible. In *Selected papers of the first conference on World-Wide Web*, pages 183–192, Amsterdam, The Netherlands, The Netherlands, 1994. Elsevier Science Publishers B. V.

[DCL$^+$00] Michelangelo Diligenti, Frans Coetzee, Steve Lawrence, C. Lee Giles, and Marco Gori. Focused crawling using context graphs. In *Proceedings of the 26th International Conference on Very Large Data Bases*, VLDB '00, pages 527–534, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[DDGR07] A.S. Das, M. Datar, A. Garg, and S. Rajaram. Google news personalization: scalable online collaborative filtering. In *Proc. of WWW*, pages 271–280. ACM, 2007.

[DFKK08] Thierry Declerck, Christian Federmann, Bernd Kiefer, and Hans-Ulrich Krieger. Ontology-based information extraction and reasoning for business intelligence applications. In *KI 2008: Advances in Artificial Intelligence*, pages 389–390. Springer, 2008.

[DHX$^+$06] Chris Ding, Xiaofeng He, Hui Xiong, Hanchuan Peng, and Stephen R. Holbrook. Transitive closure and metric inequality of weighted graphs: vdetecting protein interaction modules using cliques. *Int. J. Data Min. Bioinformatics*, 1:162–177, September 2006.

[DK07] Tran Quoc Dung and Wataru Kameyama. Ontology-based information extraction and information retrieval in health care domain. In *Data Warehousing and Knowledge Discovery*, pages 323–333. Springer, 2007.

[DR05] Ian Davidson and S. S. Ravi. Clustering with constraints: Feasibility issues and the k-means algorithm, 2005.

[Dun73] J.C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Cybernetics and Systems*, 3(3):32–57, 1973.

[Dyk83] Richard L. Dykstra. An algorithm for restricted least squares regression. *Journal of the American Statistical Association*, 78(384):837–842, 1983.

[EKSX96] Martin Ester, Hans-Peter Kriegel, Jrg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. of 2nd International Conference on Knowledge Discovery and*, pages 226–231, 1996.

[EM03] Marc Ehrig and Alexander Maedche. Ontology-focused crawling of web documents. In *Proceedings of the 2003 ACM symposium on Applied computing*, SAC '03, pages 1174–1178, New York, NY, USA, 2003. ACM.

[ET] E-Teams. by nc4.

[FB04] Xiaoli Zhang Fern and Carla E. Brodley. Solving cluster ensemble problems by bipartite graph partitioning. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, pages 36–, New York, NY, USA, 2004. ACM.

[Fie00] Roy Thomas Fielding. *Architectural styles and the design of network-based software architectures.* PhD thesis, University of California, 2000.

[FL08] Xiaoli Z. Fern and Wei Lin. Cluster ensemble selection. *Stat. Anal. Data Min.*, 1:128–141, November 2008.

[FPT95] Martin Farach, Teresa M. Przytycka, and Mikkel Thorup. On the agreement of many trees. *Inf. Process. Lett.*, 55:297–301, September 1995.

[FR98] C. Fraley and A. E. Raftery. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *The Computer Journal*, 41:578–588, 1998.

[GBM06] Roxana Girju, Adriana Badulescu, and Dan Moldovan. Automatic discovery of part-whole relations. *Computational Linguistics*, 32(1):83–135, 2006.

[GDH04] E. Gabrilovich, S. Dumais, and E. Horvitz. Newsjunkie: providing personalized newsfeeds via analysis of information novelty. In *Proc. of WWW*, pages 482–490. ACM, 2004.

[GK03] M. Girolami and A. Kabán. On an equivalence between PLSI and LDA. In *Proc. of SIGIR*, pages 433–434. ACM, 2003.

[GMT05] Aristides Gionis, Heikki Mannila, and Panayiotis Tsaparas. Clustering aggregation. In *In Proceedings of the 21st International Conference on Data Engineering (ICDE)*, pages 341–352, 2005.

[Gon85] T. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985.

[GSCM07] S. Gauch, M. Speretta, A. Chandramouli, and A. Micarelli. User profiles for personalized information access. *The adaptive web*, pages 54–89, 2007.

[HA95] Lawrence Hubert and Phipps Arabie. Iterative projection strategies for the least-squares fitting of tree structures to proximity data. *Br J Math Stat Psychol*, 48:281–317, 1995.

[Han05] Jiawei Han. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[HCL$^+$10] Vagelis Hristidis, Shu-Ching Chen, Tao Li, Steven Luis, and Yi Deng. Survey of data management and analysis in disaster situations. *J. Syst. Softw.*, 83(10):1701–1714, October 2010.

[Hea92] Marti A Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics, 1992.

[HJM$^+$98] Michael Hersovici, Michal Jacovi, Yoelle S. Maarek, Dan Pelleg, Menanchem Shtalhaim, and Sigalit Ur. The shark-search algorithm. An application: tailored Web site mapping. In *Proceedings of the seventh conference on World Wide Web*, pages 317–326, Brisbane, Australia, April 1998. Elsevier Science.

[HK06] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, second edition, 2006.

[HKP06] Jiawei Han, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques*. Morgan kaufmann, 2006.

[HL02] Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on*, 13(2):415–425, 2002.

[Hof99] T. Hofmann. Probabilistic latent semantic analysis. In *Proc. of UAI*, page 21, 1999.

[Hof04] T. Hofmann. Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems*, 22(1):89–115, 2004.

[HW06] Ching-Chi Hsu and Fan Wu. Topic-specific crawling on the web with the measurements of the relevancy context graph. *Inf. Syst.*, 31(4):232–246, June 2006.

[III72] Edward N. Adams III. Consensus techniques and the comparison of taxonomic trees. *Systematic Zoology*, 21(4):pp. 390–397, 1972.

[J.00] Smith Thomas J. L1 Optimization under Linear Inequality Constraints. *Journal of classification*, 17:225–242, 2000.

[J.01] Smith Thomas J. Constructing ultrametric and additive trees based on the L1 norm. *Journal of classification*, 18:185–207, 2001.

[JD88] A.K. Jain and R.C. Dubes. *Algorithms for clustering data*. Prentice Hall advanced reference series. Prentice Hall, 1988.

[JMK$^+$00] D. Jurafsky, J.H. Martin, A. Kehler, K. Vander Linden, and N. Ward. *Speech and language processing*. Prentice Hall, 2000.

[JTG03] Judy Johnson, Kostas Tsioutsiouliklis, and C. Lee Giles. Evolving strategies for focused web crawling. In Tom Fawcett and Nina Mishra, editors, *ICML*, pages 298–305. AAAI Press, 2003.

[Kam04] Nanda Kambhatla. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 22. Association for Computational Linguistics, 2004.

[KF94] M K Kuhner and J Felsenstein. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution*, 11(3):459–68, 1994.

[KKM02] D. Klein, S. Kamvar, and C. Manning. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering, 2002.

[KLR$^+$04] Krishna Kummamuru, Rohit Lotlikar, Shourya Roy, Karan Singal, and Raghu Krishnapuram. A hierarchical monothetic document clustering algorithm for summarization and browsing search results. In *Proceedings of the 13th international conference on World Wide Web*, pages 658–665. ACM, 2004.

[KMN99] S. Khuller, A. Moss, and J.S. Naor. The budgeted maximum coverage problem. *Information Processing Letters*, 70(1):39–45, 1999.

[KPT$^+$04] Atanas Kiryakov, Borislav Popov, Ivan Terziev, Dimitar Manov, and Damyan Ognyanoff. Semantic annotation, indexing, and retrieval. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2(1):49–79, 2004.

[KRH08] Zornitsa Kozareva, Ellen Riloff, and Eduard H Hovy. Semantic class learning from the web with hyponym pattern linkage graphs. In *ACL*, volume 8, pages 1048–1056, 2008.

[KT50] H. W. Kuhn and A. W. Tucker. Nonlinear programming. In Jerzy Neyman, editor, *Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probability*, pages 481–492. University of California Press, Berkeley, CA, USA, 1950.

[LA99] Bjornar Larsen and Chinatsu Aone. Fast and effective text mining using linear-time document clustering. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '99, pages 16–22, New York, NY, USA, 1999. ACM.

[LC03] Dawn J Lawrie and W Bruce Croft. Generating hierarchical summaries for web searches. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 457–458. ACM, 2003.

[LCLS10] L. Li, W. Chu, J. Langford, and R.E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proc. of WWW*, pages 661–670. ACM, 2010.

[LD08a] Tao Li and Chris Ding. Weighted consensus clustering. In *SIAM International Conference on Data Mining*, pages 798–809, 2008.

[LD08b] Tao Li and Chris Ding. Weighted consensus clustering. In *Proceedings of 2008 SIAM International Conference on Data Mining*, SDM 2008, pages 798–809, 2008.

[LDHN11] Dijun Luo, Chris Ding, Heng Huang, and Feiping Nie. Consensus spectral clustering in near-linear time. In *Proceedings of the 2011 IEEE 27th International Conference on Data Engineering*, ICDE '11, pages 1079–1090, Washington, DC, USA, 2011. IEEE Computer Society.

[LDP$^+$10a] Cedar E. League, Walter Daz, Brenda Philips, Ellen J. Bass, Kevin Kloesel, Eve Gruntfest, and Alex Gessner. Emergency manager decision-making and tornado warning communication. *Meteorological Applications*, 17(2):163–172, 2010.

[LDP10b] J. Liu, P. Dolan, and E.R. Pedersen. Personalized news recommendation based on click behavior. In *Proc. of IUI*, pages 31–40. ACM, 2010.

[LJM06] Hongyu Liu, Jeannette Janssen, and Evangelos Milios. Using hmm to learn user browsing patterns for focused web crawling. *Data and Knowledge Engineering*, 59(2):270 – 291, 2006.

[LKG$^+$07] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *Proc. of SIGKDD*, pages 420–429. ACM, 2007.

[LL05] Zhengdong Lu and Todd K. Leen. Semi-supervised learning with penalized probabilistic clustering. In *Advances in Neural Information Processing Systems 17*, pages 849–856. MIT Press, 2005.

[LMP01] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[LOM04] Tao Li, Mitsunori Ogihara, and Sheng Ma. On combining multiple clusterings. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, CIKM '04, pages 294–303, New York, NY, USA, 2004. ACM.

[LOM10] T. Li, M. Ogihara, and S. Ma. On combining multiple clusterings: an overview and a new perspective. *Applied Intelligence*, 33(2):207–219, October 2010.

[LWL$^+$11] Lei Li, Dingding Wang, Tao Li, Daniel Knox, and Balaji Padmanabhan. Scene: a scalable two-stage personalized news recommendation system. In *Proc. of SIGIR*, pages 125–134. ACM, 2011.

[LWSL10] Lei Li, Dingding Wang, Chao Shen, and Tao Li. Ontology-enriched multi-document summarization in disaster management. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 819–820, New York, NY, USA, 2010. ACM.

[LZO03] Tao Li, Shenghuo Zhu, and Mitsunori Ogihara. Topic hierarchy generation via linear discriminant projection. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 421–422. ACM, 2003.

[LZS09] Tao Li, Yi Zhang, and Vikas Sindhwani. A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 244–252, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[MBC$^+$99] Filippo Menczer, Richard K. Belew, Jaime Carbonell, Yiming Yang, and William Cohen. Adaptive retrieval agents: Internalizing local context and scaling up to the web. In *Machine Learning*, pages 200–0, 1999.

[MC06] Luke K McDowell and Michael Cafarella. *Ontology-driven information extraction with ontosyphon*. Springer, 2006.

[McE] D.A. McEntire. *The Status of Emergency Management Theory: Issues, Barriers, and Recommendations for Improved Scholarship*. University of North Texas. Department of Public Administration. Emergency Administration and Planning.

[MEH$^+$02] Alexander Maedche, Marc Ehrig, Siegfried Handschuh, Raphael Volz, and Ljiljana Stojanovic. Ontology-focused crawling of documents and relational metadata. In *Proceedings of the 11th International World Wide Web Conference, WWW 2002, Honolulu, Hawaii*, May 2002.

[MFP00] Andrew McCallum, Dayne Freitag, and Fernando CN Pereira. Maximum entropy markov models for information extraction and segmentation. In *ICML*, pages 591–598, 2000.

[MFRW00] Scott Miller, Heidi Fox, Lance Ramshaw, and Ralph Weischedel. A novel use of statistical parsing to extract information from text. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 226–233. Association for Computational Linguistics, 2000.

[MM99] F. Menczer and A. E. Monge. Scalable Web search by adaptive online agents: An InfoSpiders case study. In M. Klusch, editor, *Intelligent In-*

*formation Agents: Agent-Based Information Discovery and Management on the Internet*, pages 323–347. Springer, Berlin, 1999.

[MNB04] H. De Meyer, H. Naessens, and B. De Baets. Algorithms for computing the min-transitive closure and associated partition tree of a symmetric fuzzy relation. *European Journal of Operational Research*, 155(1):226 – 238, 2004.

[MNRS99] Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Building domain-specific search engines with machine learning techniques. In *AAAI Spring Symposium on Intelligent Agents in Cyberspace 1999*, 1999.

[MRA08] Abdolreza Mirzaei, Mohammad Rahmati, and Majid Ahmadi. A new method for hierarchical clustering combination. *Intell. Data Anal.*, 12:549–571, December 2008.

[MRJ$^+$11] Alan M. MacEachren, A. C. Robinson, A. Jaiswal, Scott Pezanowski, A. Savelyev, Justine I Blanford, and P. Mitra. Geo-twitter analytics: Applications in crisis management. *25th International Cartographic Conference*, July 3-8 2011.

[MSDR04] Stuart E Middleton, Nigel R Shadbolt, and David C De Roure. Ontological user profiling in recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):54–88, 2004.

[MTMG03] Stefano Monti, Pablo Tamayo, Jill Mesirov, and Todd Golub. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 52:91–118, 2003.

[Mur83] Fionn Murtagh. A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, 26(4):354–359, 1983.

[Net] National Emergency Management Network.

[NS07] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[NWF78] GL Nemhauser, LA Wolsey, and ML Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14(1):265–294, 1978.

[Ped04] Witold Pedrycz. Fuzzy clustering with a knowledge-based guidance. *Pattern Recognition Letters*, 25:469–480, 2004.

[PK01] Thierry Poibeau and Leila Kosseim. Proper name extraction from non-journalistic texts. *Language and computers*, 37(1):144–157, 2001.

[PM03] Gautam Pant and Filippo Menczer. Topical Crawling for Business Intelligence. *Research and Advanced Technology for Digital Libraries*, 2769, 2003.

[PN09] Simone Paolo Ponzetto and Roberto Navigli. Large-scale taxonomy mapping for restructuring and integrating wikipedia. In *IJCAI*, volume 9, pages 2083–2088, 2009.

[Pod00a] Jnos Podani. Simulation of random dendrograms and comparison tests: Some comments. *Journal of Classification*, 17:123–142, 2000.

[Pod00b] Jnos Podani. Simulation of random dendrograms and comparison tests: some comments. *Journal of Classification*, 17(1):123C142, 2000.

[PP06] Patrick Pantel and Marco Pennacchiotti. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 113–120. Association for Computational Linguistics, 2006.

[PS11] Simone Paolo Ponzetto and Michael Strube. Taxonomy induction based on a collaboratively built knowledge repository. *Artificial Intelligence*, 175(9):1737–1756, 2011.

[RBDD$^+$10] Maayan Roth, Assaf Ben-David, David Deutscher, Guy Flysher, Ilan Horn, Ari Leichtberg, Naty Leiser, Yossi Matias, and Ron Merom. Suggesting friends using the implicit social graph. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 233–242, New York, NY, USA, 2010. ACM.

[Res11] Philip Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *arXiv preprint arXiv:1105.5444*, 2011.

[RF68] F. James Rohlf and David R. Fisher. Tests for hierarchical structure in random data sets. *Systematic Zoology*, 17(4):407 – 412, 1968.

[RF81] D. F. Robinson and L. R. Foulds. Comparison of phylogenetic trees. *Math. Biosci.*, 53:131–147, 1981.

[RH02] Deepak Ravichandran and Eduard Hovy. Learning surface text patterns for a question answering system. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 41–47. Association for Computational Linguistics, 2002.

[RIS$^+$94] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. GroupLens: an open architecture for collaborative filtering of netnews. In *Proc. of CSCW*, pages 175–186. ACM, 1994.

[RKS06] Cartic Ramakrishnan, Krys J Kochut, and Amit P Sheth. A framework for schema-driven relationship discovery from unstructured text. In *The Semantic Web-ISWC 2006*, pages 583–596. Springer, 2006.

[RM99] Jason Rennie and Andrew Kachites McCallum. Using reinforcement learning to spider the web efficiently. In *Proceedings of the 16th International Conference on Machine Learning (ICML)*, pages 335–343, 1999.

[RTWH99] Thomas C Rindflesch, Lorraine Tanabe, John N Weinstein, and Lawrence Hunter. Edgar: extraction of drugs, genes and relations from the biomedical literature. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 517–528. NIH Public Access, 1999.

[SCR05] Milad Shokouhi, Pirooz Chubak, and Zaynab Raeesy. Enhancing focused crawling with genetic algorithms. In *Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'05) - Volume II - Volume 02*, ITCC '05, pages 503–508, Washington, DC, USA, 2005. IEEE Computer Society.

[SFMB07] Horacio Saggion, Adam Funk, Diana Maynard, and Kalina Bontcheva. Ontology-based information extraction for business intelligence. In *The Semantic Web*, pages 843–856. Springer, 2007.

[SG03] Alexander Strehl and Joydeep Ghosh. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3:583–617, March 2003.

[SGYL05] Chang Su, Yang Gao, Jianmei Yang, and Bin Luo. An efficient adaptive focused crawler based on ontology learning. In *Hybrid Intelligent Systems, 2005. HIS '05. Fifth International Conference on*, pages 6 pp.–, 2005.

[SHB06] G. Shani, D. Heckerman, and R.I. Brafman. An MDP-based recommender system. *Journal of Machine Learning Research*, 6(2):1265, 2006.

[SHG09] D.H. Stern, R. Herbrich, and T. Graepel. Matchbox: large scale online bayesian recommendations. In *Proceedings of the 18th international conference on World wide web*, pages 111–120. ACM, 2009.

[Sim09] Elena Simperl. Reusing ontologies on the semantic web: A feasibility study. *Data Knowl. Eng.*, 68(10):905–925, October 2009.

[SJ03] Catherine A. Sugar and Gareth M. James. Finding the number of clusters in a data set: An information theoretic approach. *Journal of the American Statistical Association*, 98:750–763, 2003.

[SJN06] Rion Snow, Daniel Jurafsky, and Andrew Y Ng. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 801–808. Association for Computational Linguistics, 2006.

[SKKR01] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl. Item-based collaborative filtering recommendation algorithms. In *Proc. of WWW*, pages 285–295. ACM, 2001.

[SKR99] J.B. Schafer, J. Konstan, and J. Riedi. Recommender systems in e-commerce. In *Proc. of EC*, pages 158–166. ACM, 1999.

[SKW08] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A large ontology from wikipedia and wordnet. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(3):203–217, 2008.

[SLD$^+$08] Khalid Saleem, Steven Luis, Yi Deng, Shu-Ching Chen, Vagelis Hristidis, and Tao Li. Towards a business continuity information network for rapid disaster recovery. In *Proceedings of the 2008 International Conference on Digital Government Research*, dg.o '08, pages 107–116. Digital Government Society of North America, 2008.

[Soe84] Geert Soete. Ultrametric tree representations of incomplete dissimilarity data. *Journal of Classification*, 1(1):235–242, December 1984.

[SP01] Reetinder Sidhu and Viktor K Prasanna. Fast regular expression matching using fpgas. In *Field-Programmable Custom Computing Machines, 2001. FCCM'01. The 9th Annual IEEE Symposium on*, pages 227–238. IEEE, 2001.

[SP03] Fei Sha and Fernando Pereira. Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 134–141. Association for Computational Linguistics, 2003.

[SPUP02] A.I. Schein, A. Popescul, L.H. Ungar, and D.M. Pennock. Methods and metrics for cold-start recommendations. In *Proc. of SIGIR*, pages 253–260. ACM, 2002.

[SR62] Robert R. Sokal and F. James Rohlf. The Comparison of Dendrograms by Objective Methods. *Taxon*, 11(2), 1962.

[SS99] Robert E. Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37:297–336, 1999.

[Swo91] D. Swofford. When are phylogeny estimates from molecular and morphological data incongruent? In M. M. Miyamoto and J. Cracraft, editors, *Phylogenetic analysis of DNA sequences*, chapter 14, pages 295–333. Oxford University press, 1991.

[TJP05] A. Topchy, A.K. Jain, and W. Punch. Clustering ensembles: models of consensus and weak partitions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(12):1866 –1881, dec. 2005.

[Too] The Puerto Rico Disaster Decision Support Tool.

[TSK05a] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005.

[TSK05b] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005.

[TWH01] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal Of The Royal Statistical Society Series B*, 63(2):411–423, 2001.

[VD00] Shivakumar Vaithyanathan and Byron Dom. Model-based hierarchical clustering. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, pages 599–608. Morgan Kaufmann Publishers Inc., 2000.

[VFC05] David Vallet, Miriam Fernández, and Pablo Castells. An ontology-based information retrieval model. In *The Semantic Web: Research and Applications*, pages 455–470. Springer, 2005.

[VVMD+01] Maria Vargas-Vera, Enrico Motta, John Domingue, S Buckingham Shum, Mattia Lanzoni, et al. Knowledge extraction by using an ontology-based annotation tool. In *K-CAP 2001 workshop on Knowledge Markup and Semantic Annotation*, 2001.

[VVR+05] Giannis Varelas, Epimenidis Voutsakis, Paraskevi Raftopoulou, Euripides GM Petrakis, and Evangelos E Milios. Semantic similarity methods in wordnet and their application to information retrieval on the web. In *Proceedings of the 7th annual ACM international workshop on Web information and data management*, pages 10–16. ACM, 2005.

[Wag02] Kiri Lou Wagstaff. *Intelligent clustering with instance-level constraints*. PhD thesis, Cornell University, Ithaca, NY, USA, 2002. AAI3059148.

[WCRS01] Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schrödl. Constrained k-means clustering with background knowledge. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 577–584, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[WD10] Daya C Wimalasuriya and Dejing Dou. Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science*, 36(3):306–323, 2010.

[Web] WebEOC. Manufactured by esi acquisition, inc.

[WHW08] Fei Wu, Raphael Hoffmann, and Daniel S Weld. Information extraction from wikipedia: Moving down the long tail. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 731–739. ACM, 2008.

[Wik] Wikipedia.

[Wil94] Mark Wilkinson. Common cladistic information and its consensus representation: Reduced adams and reduced cladistic consensus trees and profiles. *Systematic Biology*, 43(3):pp. 343–368, 1994.

[WJ63] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[WKPU08] Casey Whitelaw, Alex Kehlenbeck, Nemanja Petrovic, and Lyle Ungar. Web-scale named entity recognition. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 123–132. ACM, 2008.

[WLB09] Wilson Wong, Wei Liu, and Mohammed Bennamoun. Acquiring semantic relations using the web for constructing lightweight ontologies. In *Advances in Knowledge Discovery and Data Mining*, pages 266–277. Springer, 2009.

[WLDJ11] Pu Wang, Kathryn B. Laskey, Carlotta Domeniconi, and Michael Jordan. Nonparametric bayesian co-clustering ensembles. In *SDM*, pages 331–342, 2011.

[Wol92] David H. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992.

[WXC09] Junjie Wu, Hui Xiong, and Jian Chen. Towards understanding hierarchical clustering: A data distribution perspective. *Neurocomputing*, 72(10-12):2319 – 2330, 2009.

[WZLD09] Dingding Wang, Li Zheng, Tao Li, and Yi Deng. Evolutionary document summarization for disaster management. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 680–681. ACM, 2009.

[XNJR02] Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, and Stuart Russell. Distance metric learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems 15*, pages 505–512. MIT Press, 2002.

[YC09] Hui Yang and Jamie Callan. A metric-based framework for automatic taxonomy induction. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 271–279. Association for Computational Linguistics, 2009.

[YCWX09] Hua Yuan, Guoqing Chen, Junjie Wu, and Hui Xiong. Towards controlling virus propagation in information systems with point-to-group information sharing. *Decision Support Systems*, 48(1):57–68, 2009.

[YM07] Burcu Yildiz and Silvia Miksch. ontox-a method for ontology-driven information extraction. In *Computational Science and Its Applications–ICCSA 2007*, pages 660–673. Springer, 2007.

[YXT$^+$02] K. Yu, X. Xu, J. Tao, M. Ester, and H.P. Kriegel. Instance selection techniques for memory-based collaborative filtering. In *Proc. of SDM*, pages 59–74, 2002.

[ZAR03] Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. Kernel methods for relation extraction. *The Journal of Machine Learning Research*, 3:1083–1106, 2003.

[ZFLW02] Osmar R. Zaïane, Andrew Foss, Chi-Hoon Lee, and Weinan Wang. On data clustering analysis: Scalability, constraints, and validation. In *Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, PAKDD '02, pages 28–39, London, UK, UK, 2002. Springer-Verlag.

[ZH08] M. Zhang and N. Hurley. Avoiding monotony: improving the diversity of recommendation lists. In *Proc. of RS*, pages 123–130. ACM, 2008.

[Zhu05] Xiaojin Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.

[ZK02a] Ying Zhao and George Karypis. Evaluation of hierarchical clustering algorithms for document datasets. In *Proceedings of the eleventh international conference on Information and knowledge management*, CIKM '02, pages 515–524, New York, NY, USA, 2002. ACM.

[ZK02b] Ying Zhao and George Karypis. Evaluation of hierarchical clustering algorithms for document datasets. In *CIKM*, pages 515–524, 2002.

[ZKK08] Hai-Tao Zheng, Bo-Yeong Kang, and Hong-Gee Kim. An ontology-based approach to learnable focused crawling. *Inf. Sci.*, 178(23):4512–4522, December 2008.

[ZL02a] Osmar R Zaïane and Chi-Hoon Lee. Clustering spatial data when facing physical constraints. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pages 737–740. IEEE, 2002.

[ZL02b] Osmar R. Zaane and Chi-Hoon Lee. Clustering spatial data in the presence of obstacles: a density-based approach. In *Sixth International Database Engineering and Applications Symposium (IDEAS 2002*, pages 17–19, 2002.

[ZL11] Li Zheng and Tao Li. Semi-supervised hierarchical clustering. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 982 –991, dec. 2011.

[ZLD10a] Li Zheng, Tao Li, and Chris Ding. Hierarchical ensemble clustering. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, ICDM '10, pages 1199–1204, Washington, DC, USA, 2010. IEEE Computer Society.

[ZLD10b] Li Zheng, Tao Li, and Chris H. Q. Ding. Hierarchical ensemble clustering. In *ICDM'10*, pages 1199–1204, 2010.

[ZLHL12] Li Zheng, Lei Li, Wenxing Hong, and Tao Li. Penetrate: Personalized news recommendation using ensemble hierarchical clustering. *Expert Systems with Applications*, (0):–, 2012.

[ZLST04] Cai-Nicolas Ziegler, Georg Lausen, and Lars Schmidt-Thieme. Taxonomy-driven computation of product recommendations. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 406–415. ACM, 2004.

[ZQ10] Haifeng Zhao and Zijie Qi. Hierarchical agglomerative clustering with ordering constraints. In *WKDD*, pages 195–199, 2010.

[ZST$^+$10] Li Zheng, Chao Shen, Liang Tang, Tao Li, Steve Luis, Shu-Ching Chen, and Vagelis Hristidis. Using data mining techniques to address critical information exchange needs in disaster affected public-private networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 125–134, New York, NY, USA, 2010. ACM.

[ZST$^+$11] Li Zheng, Chao Shen, Liang Tang, Tao Li, Steve Luis, and Shu-Ching Chen. Applying data mining techniques to address disaster information management challenges on mobile devices. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 283–291, New York, NY, USA, 2011. ACM.

[ZST$^+$12] Li Zheng, Chao Shen, Liang Tang, Chunqiu Zeng, Tao Li, Steven Luis, Shu-Ching Chen, and Jainendra K. Navlakha. Disaster sitrep - a vertical search engine and information analysis tool in disaster management domain. In Chengcui Zhang, James Joshi, Elisa Bertino, and Bhavani M. Thuraisingham, editors, *IRI*, pages 457–465. IEEE, 2012.

[ZST$^+$13] Li Zheng, Chao Shen, Liang Tang, Chunqiu Zeng, Tao Li, S. Luis, and Shu-Ching Chen. Data mining meets the needs of disaster information management. *Human-Machine Systems, IEEE Transactions on*, 43(5):451–464, Sept 2013.

# VITA

## LI ZHENG

| | |
|---|---|
| May 13, 1983 | Born, Sichuan, China |
| 2004 | B.A., Computer Science<br>Sichuan University<br>Chengdu, China |
| 2007 | M.S., Computer Science<br>Sichuan University<br>Chengdu, China |
| 2008–Current | Ph.D. Candidate<br>Florida International University<br>Miami, Florida |

## PUBLICATIONS AND PRESENTATIONS

Li, L., Zheng, L., Yang, F., Li, T., (2014). Modeling and Broadening Temporal User Interest in Personalized News Recommendation. Journal of Expert Systems with Applications, Volume 41, Issue 7, Pages 3168-3177.

Zheng, L., Shen, C., Tang, L., Zeng, C. Q., Li, T., Luis, S., Chen, S. C.,(2013). *Data Mining Meets the Needs of Disaster Information Management*. IEEE Transactions on Human-Machine Systems (THMS), Volume:43, Issue:5, Pages 451-464 .

Zeng, C. Q., Jiang, Y. X., Zheng, L., Li, J. X., Li, L., Li, H. T., Shen, C., Zhou, W. B., Li, T., Duan, B., Lei, M., Wang, P. N., (2013). *FIU-Miner: A Fast, Integrated, and User-Friendly System for Data Mining in Distributed Environment*. In SIGKDD, Pages 1506-1509.

Zheng, L., Lei Li, Wenxing Hong, Li, T., (2012). *PENETRATE: Personalized News Recommendation Using Ensemble Hierarchical Clustering*. Journal of Expert Systems with Applications, Volume 40, Issue 6, Pages 2127-2136.

Zheng, L., Chao Shen, Liang Tang, Chunqiu Zeng, Li, T., Steve Luis, Shu-Ching Chen, Navlakha K. J., (2012). *Disaster SitRep - A Vertical Search Engine and Information Analysis Tool in Disaster Management Domain*. The 13th IEEE International Conference on Information Integration and Reuse, Pages 457-465.

Zheng, L., Li, T., (2011). *Semi-supervised Hierarchical Clustering*. In Proceedings of 2011 IEEE International Conference on Data Mining, Pages 982-991.

Li L., Zheng, L., and Li, T., (2011). *LOGO: A Long-Short User Interest Integration in Personalized News Recommendation*. In Proceedings of the 5th ACM Conference on Recommender Systems, Pages 317-320.

Zheng, L., Shen, C., Tang, L., Li, T., Luis, S., Chen, S. C.,(2011). *Applying Data Mining Techniques to Address Disaster Information Management Challenges on Mobile Devices*. ACM SIGKDD Conference, Pages 283-291.

Zheng, L., Li, T., and Ding, H. Q., (2010). *Hierarchical Ensemble Clustering*. In Proceedings of 2010 IEEE International Conference on Data Mining, Pages 1199-1204.

Zheng, L., Shen, C., Tang, L., Zeng, C. Q., Li, T., Luis, S., Chen, S. C., Hristidis, V., (2013). *Using Data Mining Techniques to Address Critical Information Exchange Needs in Disaster Affected Public-Private Networks*. ACM SIGKDD Conference, Pages 125-134.

Wang, D. D., Zheng, L., Li, T., Deng, Y., (2009). *Evolutionary Document Summarization for Disaster Management*. ACM SIGIR Conference, Pages 680-681.