Content-Based Access Control

By

Wenrong Zeng

Submitted to the Department of Electrical Engineering and Computer Science and the
Graduate Faculty of the University of Kansas
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Committee members

Dr. Bo Luo, Chairperson

Dr. Arvin Agah

Dr. Jerzy Grzymala-Busse

Dr. Prasad Kulkarni

Dr. Alfred Tat-kei Ho

Date defended: April 3, 2015

The Dissertation Committee for Wenrong Zeng certifies
that this is the approved version of the following dissertation :

Content-Based Access Control

Dr. Bo Luo, Chairperson

Date approved: _____

Abstract

In conventional database, the most popular access control model specifies policies explicitly for each role of every user against each data object manually. Nowadays, in large-scale content-centric data sharing, conventional approaches could be impractical due to exponential explosion of the data growth and the sensitivity of data objects. What's more, conventional database access control policy will not be functional when the semantic content of data is expected to play a role in access decisions. Users are often over-privileged, and *ex post facto* auditing is enforced to detect misuse of the privileges. Unfortunately, it is usually difficult to reverse the damage, as (large amount of) data has been disclosed already. In this dissertation, we first introduce Content-Based Access Control (CBAC), an innovative access control model for content-centric information sharing. As a complement to conventional access control models, the CBAC model makes access control decisions based on the content similarity between user credentials and data content automatically. In CBAC, each user is allowed by a meta-rule to access "a subset" of the designated data objects of a content-centric database, while the boundary of the subset is dynamically determined by the textual content of data objects. We then present an enforcement mechanism for CBAC that exploits Oracles Virtual Private Database (VPD) to implement a row-wise access control and to prevent data objects from being abused by unnecessary access admission. To further improve the performance of the proposed approach, we introduce a content-based blocking mechanism to improve the efficiency of CBAC enforcement to further reveal a more relevant part of the data objects comparing with only using the user credentials and data content. We also utilized several tagging mechanisms for more accurate tex-

tual content matching for short text snippets (e.g. short VarChar attributes) to extract topics other than pure word occurrences to represent the content of data. In the tagging mechanism, the similarity of content is calculated not purely dependent on the word occurrences but the semantic topics underneath the text content. Experimental results show that CBAC makes accurate access control decisions with a small overhead.

Acknowledgements

In this section, I would like to express my gratitude to my advisors, my colleagues, my committee members and my family for their encouragement, support and assistance down along the road of my PhD study.

First of all, I would like to thank my advisor Dr. Bo Luo for his valuable guidance during my thesis. He was a great advisor to work with. He originally led me to the field of access control and patiently explained the fundamental background of what it is, why it is important to database security and its potential impacts on big data platform. He has been very supportive, and enthusiastic in all our discussions. He usually inspires me with his solid background knowledge on database security and kindly provides insights to draw conclusion on experimental results, which help me to push the experiment forwards meanwhile consolidate my work. I would also like to thank my previous advisor Dr. Xue-wen Chen. When I began my PhD, he led me to machine learning field. He directed me to multi-label learning, which is another major part of my PhD work. The guidance from him led me to explore multi-label applications in image analysis with graphical modeling, and theoretical optimization of multi-label improvements. I am also very gratefully thankful to my committee members: Dr. Arvin Agah, Dr. Jerzy Grzymala-Busse, Dr. Prasad Kulkarni, and Dr. Alfred Tat-kei Ho. They offer me professional suggestion on my proposal and dissertation. They kindly have provided their insights on further directions and experiments with my work. Without their help, I cannot finish the entire course of my PhD.

Secondly, I would like to thank all my colleagues in University of Kansas. They began as colleagues and ended to be my best friends. They have provided a lot of happiness,

supports, and assistance in my life and study. Working together with everyone is a memorable moment in the entire course of my study. Dr. Hongliang Fei, Dr. Yi Jia, Dr. Jintao Zhang, and Junyan Li, I have been grateful meeting them in the painful yet rewardful PhD study. When I met problems, they always provide valuable suggestion. Besides, I should owe special thanks to Dr. Jong Cheol Jeong. He is a valuable colleague to work with, thorough in mind and detail oriented in execution. He is willing to help me whenever I feel confused and lost in research. His determination in research has set him up as my role model always.

Last but not least, I want to thank my family. My parents gave me endless courage and love during my PhD stage. They have visited me three times from China, and supported me with all their assistance in my household. They are the best parents one could ask. My husband, the one I am super lucky to have, is the most supportive, patient, generous and humorous man in my life, who has brought the happiness and healed the pain. My baby daughter, Brenda, I would like to thank her for being the best project I have ever done. Although she does cry a lot, she laughs more. Her smile is the best gift after work.

Contents

1 Introduction **1**

1.1 Introduction 1

2 Related Works **6**

2.1 Access Control Models 10

2.1.1 Discretionary Access Control 10

2.1.2 Role-Based Access Control 12

2.1.3 Attribute-Based Access Control 17

2.1.4 Policy-Based Access Control 19

2.1.5 Risk-Adaptabe Access Control 20

2.1.6 Access Control Based on Content 21

2.2 Oracle Virtual Private Database (VPD) 23

3 Text Feature Extraction **27**

3.1 TF-IDF 28

3.1.1 Stop Word 29

3.1.2 Stemming 30

3.2 *n*-Gram 30

3.3 Topic Modeling 32

3.3.1 Latent Dirichlet Allocation 32

3.3.2 Non-negative Matrix Factorization 33

3.4 TAGME 35

4 Content-Based Access Control Model **37**

4.1 Background and Assumptions 37

4.2 Contribution 39

4.3 Model Definition 41

4.4 Content Similarity 44

4.5 Top-K Similarity 45

5 CBAC Enforcement **47**

5.1 CBAC On-the-Fly Enforcement 47

5.1.1 The Basic CBAC Model 47

5.1.2 Experiments 50

5.2 Offline CBAC 56

5.2.1 Unsupervised Nearest Neighbor Offline Training 56

5.2.1.1 Brute Force Algorithm 56

5.2.1.2 K-D tree 57

5.2.1.3 Ball Tree Algorithm 58

5.2.2 Experiments 61

6 CBAC Optimizing Strategies **67**

6.1 Content-Based Blocking 67

6.1.1 Naive k-means Clustering 69

6.1.2 The Advantage of Careful Seeding: k-means++ 70

6.1.3 Scaled k-means++ with mini-batch Strategy 70

6.1.4 Experiments 71

6.2 Content-based labeling 73

6.2.1 Document labeling 73

6.2.2 Soundness of CBAC Enforcement 76

6.2.3 Experiments 78

7 Labeling Improvement with Multi-Label Learning (MLL) 86

7.1 Motivation 86

7.2 Problem Definition and Challenges 87

7.3 Background 88

7.4 Related Work 92

7.5 Methodology 95

7.5.1 Preliminary 95

7.5.2 Objective Function 95

7.5.3 Algorithm 97

7.6 Experiment 99

7.6.1 Data Set Statistics 99

7.6.2 Comparison Methods 102

7.6.3 Evaluation Metric 103

7.6.4 Results 104

7.7 Conclusion 105

8 Discussions 107

8.1 Computational Complexity 107

8.2 Negative Rules and Conflict Resolution 108

8.3 CBAC for XML Data 109

9 Conclusion 110

A The Top 10 Words of Non-Negative Matrix Factorization 126

List of Figures

2.1 Selected Access Control Models (NIST (2009)) 7

2.2 Access Control List 13

2.3 Capability List 14

2.4 Role-Based Access Control Model (Sandhu et al. (1996)) 16

2.5 Risk-Adaptable Access Control Notional Process (McGraw (2009)) 21

3.1 Plate Notation of Latent Dirichlet Allocation 34

3.2 Plate Notation of Smoothed Latent Dirichlet Allocation 34

3.3 TAGME Annotation Example 36

5.1 ABAC Efficiency with QUERY1 56

5.2 ABAC Efficiency with QUERY2 57

5.3 Threshold CBAC Efficiency with QUERY1 58

5.4 Threshold CBAC Efficiency with QUERY2 61

5.5 Threshold CBAC + ABAC Efficiency with QUERY1 62

5.6 Threshold CBAC + ABAC Efficiency with QUERY2 63

5.7 Top-10 CBAC Efficiency 63

5.8 2-D K-D Tree Subspace Splits 64

5.9 K-D Tree Example 65

5.10 Offline Efficiency 65

6.1 Threshold CBAC + Blocking Efficiency with QUERY1 72

6.2 Threshold CBAC + Blocking Efficiency with QUERY2 73
6.3 Threshold CBAC + ABAC + Blocking Efficiency with QUERY1 74
6.4 Threshold CBAC + ABAC + Blocking Efficiency with QUERY2 75
6.5 Top-10 CBAC + Blocking Efficiency 76
6.6 Soundness of CBAC Enforcement 79
6.7 Threshold CBAC + Labeling Efficiency with QUERY1 80
6.8 Threshold CBAC + Labeling Efficiency with QUERY2 81
6.9 Threshold CBAC + ABAC + Labeling Efficiency with QUERY1 81
6.10 Threshold CBAC + ABAC + Labeling Efficiency with QUERY2 82
6.11 Top-10 CBAC + Labeling Efficiency 82
6.12 Top-10 CBAC + Blocking + Labeling Efficiency 83
6.13 Density Fit 83
6.14 Cumulative Probability Fit 84
6.15 NMF 100 Density Fit 84
6.16 NMF 100 Cumulative Probability Fit 85

7.1 Scene of Sunset at Sea 93
7.2 Molecular Function Annotation of P75957 94

List of Tables

2.1 Access Control Matrix Example 11

2.2 VPD Function Example 25

2.3 VPD Policy Example 26

5.1 Schemas 51

5.2 Column Description 51

5.3 CBAC Top-10 Example 52

5.4 CBAC Threshold Example 53

7.1 Multi-Label Example 91

7.2 Binary Relevance Matrix Example 91

7.3 Label Power-Set Example 92

7.4 Label Power-Set Matrix Example 92

7.5 Statistics of Data Sets 100

7.6 Imbalance Rate (%) 100

7.7 Sample Sizes of Data Sets 101

7.8 Macro-Averaging F1 Measure (%) ↑ 102

7.9 Micro-Averaging F1 Measurel (%) ↑ 103

7.10 Subset Accuracy (%) ↑ 105

A.1 The Top 10 Words of Non-Negative Matrix Factorization with 10 Topics 127

A.2 The Top 10 Words of Non-Negative Matrix Factorization with 20 Topics 128

A.3 The Top 10 Words of Non-Negative Matrix Factorization with 50 Topics 129
A.4 The Top 10 Words of Non-Negative Matrix Factorization with 100 Topics 133

Chapter 1

Introduction

1.1 Introduction

Simply put, database access control models and enforcement mechanisms define and enforce "who can access what". Here, "who" represents a set of users/roles, and "what" represents a set of data objects, e.g. tuples or XML nodes, attributes of SQL databases. In conventional database access control models, database administrators (DBAs) or data owners/users explicitly specify access rights of each data object for each role by GRANT or REVOKE certain rights from each role. However, due to the exponential explosion of data, especially for content-centric data, such approaches may not be suitable or even practical. The reason for this is three-folded. Firstly, it is determined by the characteristics of content-centric data. Content-centric data usually contains a lot of free text. For example, electronic health record (EHR) is a content-centric kind of data. In the EHR, doctors other than list the basic information of patients (eg. name, gender, age, etc.) describe the symptoms of every patient. Instead of choosing exact words to describe the patients' symptoms, a doctor usually use more descriptive ways to record the symptoms given by the patients. That's why EHR data is rather free text kind of data than formatted text kind of data. Free text, as the example shown, can express the same semantic meaning with different term distributions. Secondly, in content-centric database, the data content is expected to play a role in making the

access control decisions. Let's continue on the EHR example. Before giving a concluded decision on what the patient's problem is, doctors might have needs to review some other patient's record with similar symptoms, especially for unusual diseases. For this kind of situation, it could be very difficult to explicitly describe access rights for very large amounts of data objects, especially when the decisions are based on content – it is too labor-intensive to require a system administrator to manually examine every record in the database and assign access rights to each user/role. Thirdly, in distributed and dynamic environments, it could be difficult to explicitly define access rights for every user from remote peers, e.g., an organization could easily develop new roles without notifying its collaborators, which happens a lot in information sharing. In this case, access control decisions could be based on remote requestor's knowledge that is dynamically submitted with every query. Meanwhile, in distributed information sharing, data owners may only want to share with people who contribute similar data which might reveal that they have similar interests due to the sensitivity of the data content, but they cannot specify access control rules unless they explore the content of others' data. To further motivate this research, let us see the following examples:

Example 1: A law enforcement agency (e.g. FBI) holds a database of highly sensitive case records. A director *Bob* assigns a case to agent *Alice* for investigation. Naturally, the director also needs to grant Alice access to all related or similar cases. In this scenario, the concept of "related cases" is determined by the semantical content similarity of the records, which could be geological, temporal, motis operandi, or just the similarity in the textual description of the case records. Moreover, when new cases are added to the database, cases that are similar to *Alice*'s should be automatically made accessible to *Alice*, without requiring the director to further intervene. For example, that new added related cases could be a crucial key to the case being investigated. Unfortunately, in the existing database access control paradigm, this type of access control description is not supported. Meanwhile, it is too labor-intensive for the director to manually examine every record to grant/revoke access. In practice, the Multi-Level Security (MLS) model is often adopted and every agent is granted access to a large number of records – everything lower than or equal to his/her security level. Similarly, many content processing companies (e.g. survey processing and

telemarketing firms) allow every employee to access all the (potentially sensitive) customer records in their databases, due to lack of capability to enforce access control based on the textual content of the records. In all these similar scenarios, information sharing could be either too conservative or being abused because unnecessary information leakage.

∎

Example 2: In traditional subscription systems, users pay for access towards entire periodicals. For instance, a researcher interested in "information security" may subscribe to IEEE Transactions on Knowledge and Data Engineering, though he is only interested in a small portion of the papers in the journal. An alternative approach would be that each user subscribes to a set of tags, and each paper (as a record in the database) is tagged by keywords. Thus, access control decisions could be made by matching user's tags with paper's tags. However, such approach suffers two serious drawbacks: (1) tag quality is essential to the approach, but the quality control is a non-trivial problem; (2) the number of accessible papers could be too small or too large, for instance, a paper carrying a tag may be only slightly related to the tag topic. In a desired solution, the subscriber is expected to submit his interests as a textual description or identify some seed articles (e.g. his own papers), and then be granted access to articles with similar content. In the ideal solution, the granted access control policies based on content similarity would further improve the work efficiency of users based on qualified selective articles.

∎

Example 3: In distributed information sharing scenarios, some data owners will only share their records with peers who contribute relevant data, so that the sharing is mutually beneficial. For instance, in a collaborative project with Department of Public Administration studying citizen engagement, surveyees are found to be willing to share their opinions with others who have similar opinions. In this case, opinions are represented by a short paragraph of text. In other scientific research domains, we also see investigators sharing research data (in a shared and access-controlled repository) with colleagues who contributes similar data. Let us revisit Example 1: when FBI

collaborates with other law enforcement agency (say, CIA), they only share "related cases", while the case relationships are accessed by semantical content similarity. Privacy-preserving similar document matching (Murugesan et al. (2010); Scannapieco et al. (2007)) has been used to identify and share similar documents. However, in the scenario that FBI is willing to disclose cases that are similar to a known CIA case, an alternative solution is to employ database access control to allow CIA to access the "similar cases".

∎

Example 4: Healthcare information sharing is strictly governed by HIPAA. Medical records are well protected by healthcare providers, and are only shared under very rigorous rules. However, within the facility, users (doctors, nurses, researchers) are often given broader access privileges, while *ex post facto* auditing is enforced to detect and punish misuse of the privileges (Appari & Johnson (2010); Malin et al. (2007); Boxwala et al. (2011); Rostad & Edsberg (2006)). Another thrust of solutions employs the "break the glass (BTG)" mechanism – to allow users to break access control rules in a controlled manner in special circumstances (Ferreira et al. (2006)). Additional auditing will be performed once a user invokes the BTG policy.

∎

From the examples, we can see that conventional deterministic database access control models fall short in content-centric data sharing scenarios. In such cases, a new access control model is expected to emerge to meet the needs of generating access decision based on the semantical content similarity of the data. Another desirable capability of such content-based access control model is the similarity of semantical content should be measured as native functions provided by RDBMS, and it only requires minimal intervene from database administrators (DBAs). In this dissertation, we present a first attempt towards this endeavor: we present the content-based access control model and enforcement mechanisms, where access rights are granted based on the lexicon similarity between requestor's credentials and the requested records. The new model, as a complement to existing access control approaches, provides an effective and efficient means of access control that exploits content features in content-rich data sharing, and leads to a first effort to

solve the difficulties in content-centric database access control in big data era. In the dissertation, we explore the new needs of security and privacy in distributed information systems and decide to tackle such issues with innovative designs. Therefore, we formally propose a new data-driven access control model called content-based access control (CBAC) model which exploits the data content to achieve more flexible and powerful access control semantics towards conetent-centric databases in information sharing in the dissertation. CBAC is the first attempt to create access control model that introduces the notion of approximate security, and it is capable of dealing with situations where explicit access control policies are not at all available. In CBAC, we decide to use machine learning methods (i.e. text mining techniques) for access control modeling and enforcement. By introducing these methods, access control principle is translated into algorithm implementation, and in the sense, we aim to enhance the dynamic properties, automation and "intelligence" into access control models via all these techniques.

Chapter 2

Related Works

Computer technology has transformed the way of daily life including education, career life, and entertainment of people. It makes convenience for people to seek information for knowledge, find jobs, enjoy fancy music and films. Meanwhile, computer technology also transformed the way of running companies including hunting for suppliers to compete their offers, collecting, storing and broadcasting their information of products, and maintaining their close work with clients. Not only computer technology has improved the efficiency of everyone's daily life, it has also changed how information is created, processed, transferred, stored, and concealed. Nowadays, one of the most important security problem is to prevent unauthorized access to information, which prevents unauthorized people have access to credential information he/she is NOT allowed to. The common risks from unauthorized access include but not limited to:

- Unauthorized disclosure of information
- Disruption of computer services
- Loss of productivity which delaying normal computer activities in time critical applications
- Financial loss such as corruption of information or disruption of services
- Legal implications due to lawsuits from investors, customers, or the public

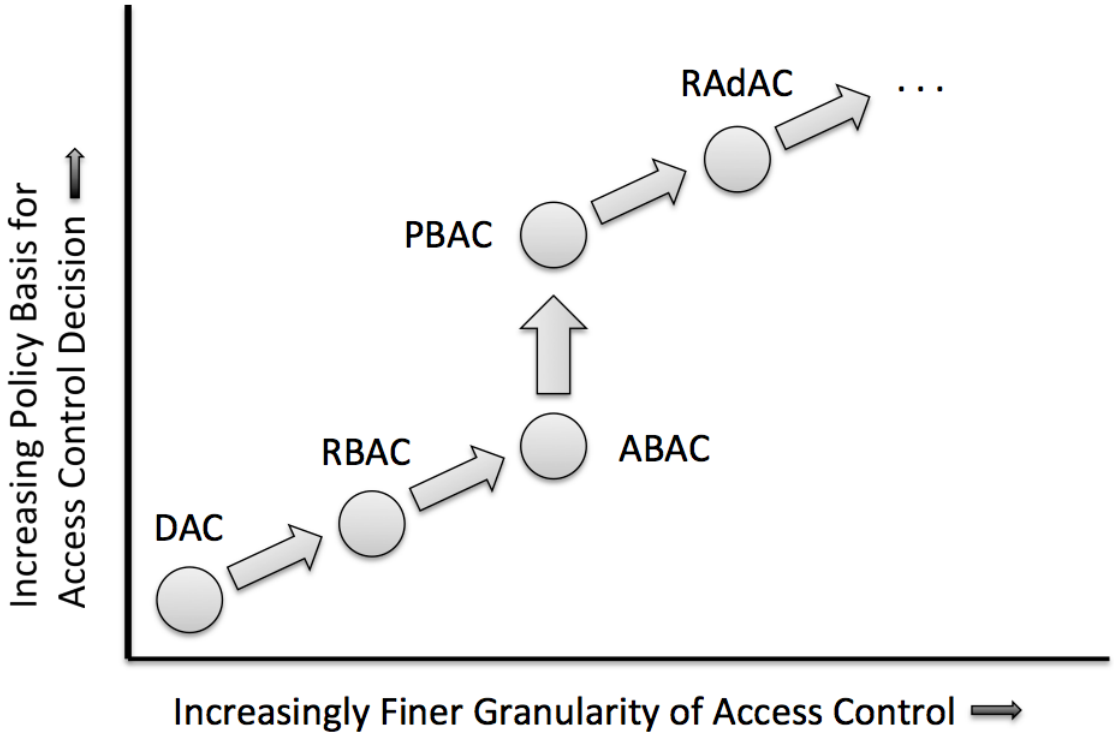

Figure 2.1: Selected Access Control Models (NIST (2009))

- Blackmail intruders extort money from the company by threatening the security system

To avoid these risks, researchers developed different access control models to paradigm of "who" has the authorities to access "what". In this chapter, we select some common access control models for introduction. Figure 2.1 is modified from Figure 1 (NIST (2009)) to show the relationship among these models. We follow the list of access control models (NIST (2009)) and add more details about models which have concrete mathematical definition.

Database access control research could be roughly categorized as *access control models* and *access control enforcement*. Relational access control models can be classified into: *mandatory access control* (Jajodia & Sandhu (1991); Sandhu (1993); Sandhu & Chen (1998); Winslett et al. (1994); McCune et al. (2006); Lindqvist (2006); Thuraisingham (2009); Upadhyaya (2011)), *discretionary access control (DAC)* (Moffett et al. (1990); Thomas et al. (1993); Ahn (2009); Li

(2011); Downs et al. (1985)) and *role-based access control (RBAC)* (Ferraiolo et al. (2001); Osborn et al. (2000); Sandhu et al. (1996)).

Mandatory access control (MAC) emphasizes only the database administrators have the authorities to manage the access control policy and usage. These policies and usage cannot be modified by any other users other than the administrators. Therefore, MAC is most often used in systems or databases when the highest priority is placed on confidentiality. The assignment and enforcement of access control policy under MAC models places strict restrictions on users. The dynamic alteration of any access control policy requires detailed investigation of the policy itself purely by database administrators manually. One obvious shortcoming is any update might introduce dilemmas in the entire access control policies. Also frequent database updates will be labor-intensive for administrators. Another shortcoming of MAC is it can be too protective to unnecessarily overclassify data through “the high-water mark principle" and limit the ability of transfer information between users and databases. On the other hand, most real world RDBMS implement a table/column level DAC or RBAC similar to the one in System R (Griffiths & Wade (1976)).

Discretionary Access Control (DAC) is the type of access control where users has complete authority over all the data they owns. Also they have the authorities to assign GRANT/REVOKE to other users to access or not to their own data. DAC requires the permission assignment between users who hold the data and who want to access the data. Thus, it is commonly known as the “need-to-know" model. Compared to MAC, DAC shows an obvious advantage enabling fine-grained control over system or database objects. Data objects can have access control restrictions with the minimum rights needed. However, security policies are extremely difficult for DAC as the access control right is owned by users. Compromised users could pass potential threats to the database and further them to other users. Thus, DAC has high potential to insecure problems.

Role-based access control (RBAC) is the type of access control model where users are firstly assigned to different roles due to different job functions in an enterprise, and then the permission are not directly assigned to users but to roles. The permission in contrast to the above two methods of access control which GRANT/REVOKE user access on a rigid, object-by-object basis. In

RBAC, users are easily to be granted or revoked accesses due to the change of their work status. In large organizations, to cluster of many users into a single role allows much more convenient management. RBAC also integrates support for least-privilege principle, duty separation, and role membership central administration. Although RBAC shows a great advantage over the above two conventional access control models. Meanwhile it also has its own limitations. In large systems, role membership, hierarchical structure among roles, and the need to maintain least-privilege principle make administration overwhelming. Besides, although RBAC supports data abstraction, it is unable to be used to ensure permissions on sequences of operations need to be controlled as discussed in Section 2.1.2.

To enforce access control policy, view-based approaches is the traditional method to enable row-level access control (Bertino & Haas (1988); Bertino et al. (1983)). Over the years, many models and enforcement mechanisms have been proposed (a survey is available (Samarati & de Vimercati (2001))), such as the Flexible Authorization Manager (FAM) (Jajodia et al. (1997)), temporal DAC and RBAC models (Bertino et al. (1996, 2001a)), credential-based access control (Bertino et al. (2001b); Winslett et al. (1997)), group-centric models (Krishnan et al. (2009)); and more recently: purpose based access control (Byun & Li (2008)), policy-based access control for the semantic web (Bhatti et al. (2007); Kagal et al. (2003)), Oracle VPD (Oracle (2012)), XML access control (Bertino & Ferrari (2002); Damiani et al. (2002, 2000); Yu et al. (2002)), and access control for the Web (Hicks et al. (2010); Park et al. (2001)). In Bertino et al. (2003a), a formal framework for logic-based reasoning of access control model is proposed to analyze the relationships between access control rules. Much efforts have been devoted to facilitate effective management of users, roles, rules in different applications, e.g. role mining and administration (Dekker et al. (2008); Molloy et al. (2010); Takabi & Joshi (2009)), policy integration and user provisioning (Li et al. (2009); Molloy et al. (2009); Ni et al. (2009); Rao et al. (2009)), mediation in distributed systems (Leighton & Barbosa (2010); Rao & Jaeger (2009)).

More relevant to the proposed research, the notion of *content-based access control* has been proposed in relational access control specification (Bertino et al. (1997)) and in the context of

digital libraries (Adam et al. (2002)). In Adam et al. (2002); Bertino et al. (1997), the notion of *content* refers to attribute values or definitive concepts extracted from digital library objects. Access privileges are (statically) specified based on relationships between user credentials and data attributes (concepts). Meanwhile, policy-based access control models (Bhatti et al. (2007); Kagal et al. (2003)) bind access rights with user credentials (attributes); however, the decision is still based on definitive values of the attributes (e.g. users with `title`="physician" could access patient records in his/her `department`). On the other hand, concept-level access control has also been proposed for the semantic web (Qin & Atluri (2003)). However, our notion of content-based access control is significantly different from existing approaches, we refer to the *semantic* content and semantic similarity of data, as well as the notion of *approximation*.

Selected access control models are introduced explicitly in Section 2.1.

2.1 Access Control Models

2.1.1 Discretionary Access Control

The popularity of discretionary access controls (DACs) (Moffett et al. (1990); Thomas et al. (1993)) started at about the late 1970's due to the mergence of multi-user systems. Even nowadays, DACs are most widely used forms of access control in multi-user systems such as Windows and UNIX. The emergence of multi-user systems demands a huge need to limit the access to objects owned by different users on a same file system for sharing information. Under UNIX system, it is well known that three primitive permissions are attached to each object (eg. files and folders, etc) including read, write, and execute to determine which user can do what on one object. The scenario in DACs is that each object (eg. data object) has an owner who has primary controls over the object. This means the owner is able to assign the access control policy on his/her own data to other users under the same system.

The concept behind DACs is very simple: each object (eg. files or folders on file system; data objects or attributes in databases) on systems has its own associated mappings between the

set of users requesting access to it and the set of actions that can be performed on the object. In fact, different DACs show different data structures to handle the associated mappings. For example, access control matrix uses matrices to denote the DACs' deployment of access control by constructing two sets including subject set (users) S and object set O, explicit annotation of access rights (namely, *read*, *write*, and *execute*) and administration rights (eg. *control*). All these components are then constructed as a matrix with the dimension of $|S| \times (|S| + |O|)$

Table 2.1: Access Control Matrix Example

	Alice	Bob	Carl	File1	Folder1	Process1
Alice		control		read, write, own		
Bob			control		write	
Carl						execute

From Table 2.1, we can see for three users *Alice*, *Bob* and *Carl* and three objects *File1*, *Folder1* and *Process1*, access control matrix is of 3-by-6 dimensions. With the increases of subjects and objects, the access control matrix tends to be much larger and sparser. This results an inefficiency in storing access control policy in matrices which is a waste of the storage spaces. Also as before any actions can be performed from a subject to an object, the access permission should be checked first. To search against a large matrix to verify every action like this, is inefficient. Therefore, DACs have two lists to replace the matrix data structure called access control lists (ACLs) and capability lists (CLs). These two lists are different from the standing points. Access control lists are commonly observed in UNIX file system as an appending property of a file, a folder or an application with a string constructed by *r* for read, *w* for write, *x* for execute and - for no access. Thus, the standing point of access control list is from the objects. Contrary to access control lists, capability lists basically show a certain subject the overall access rights under some objects. Examples of ACLs and CLs is illustrated in Figure 2.2 and 2.3. In this sense, ACLs provide access review on a per-obejct basis; while CLs provide access review on a per-subject basis. Additionally, in access assessment process, CLs requires unforgeability and control of propagation of capabilities, while

ACLs just require authentication of subjects. The relative simplicity enhances the popularity of ACLs, which makes ACLs commonly used in all modern operating systems even other contexts. For instance, they have been applied together with network contexts where the target resource access is requested by remote systems. Despite of the widely usage of ACLs, they do have their limitations. To assess permissions, ACLs need to authenticate of subjects every time, which deteriorate the time wastes. In situation where an enterprise requires many people having different levels of access rights to many different resources, the updating of ACLs on each individual objects can be really time consuming and also may invite errors hard to be detected.

2.1.2 Role-Based Access Control

A comprehensive model of role-based access control shown in Figure 2.4 cited from Sandhu et al. (1996). The feature of RBAC models (Ferraiolo et al. (2001); Osborn et al. (2000); Sandhu et al. (1996)) is that the model first construct *User Assignment* to assign users to roles due to different work sessions, and then uses *Permission Assignment* to further assign roles to access rights on different objects. The scenario is a basic RBAC. However, in realistic enterprise situations, users have different job titles and tasks, which forms a hierarchical structure. The hierarchical structure is not based on users themselves, but the roles. Therefore, in Figure 2.4, role hierarchy is also specified. Role hierarchy can also be written: $\geq$ (The notation: $x \geq y$ means that x inherits the permissions of y) To construct RBAC model, the roles of administrators need to interfere with the *User Assignment* and *Permission Assignment*. They form a special group of roles. In the comprehensive model, it is also called administrative role-based access control (ARBAC).

To concretely define a RBAC model, the following concepts need to be addressed:

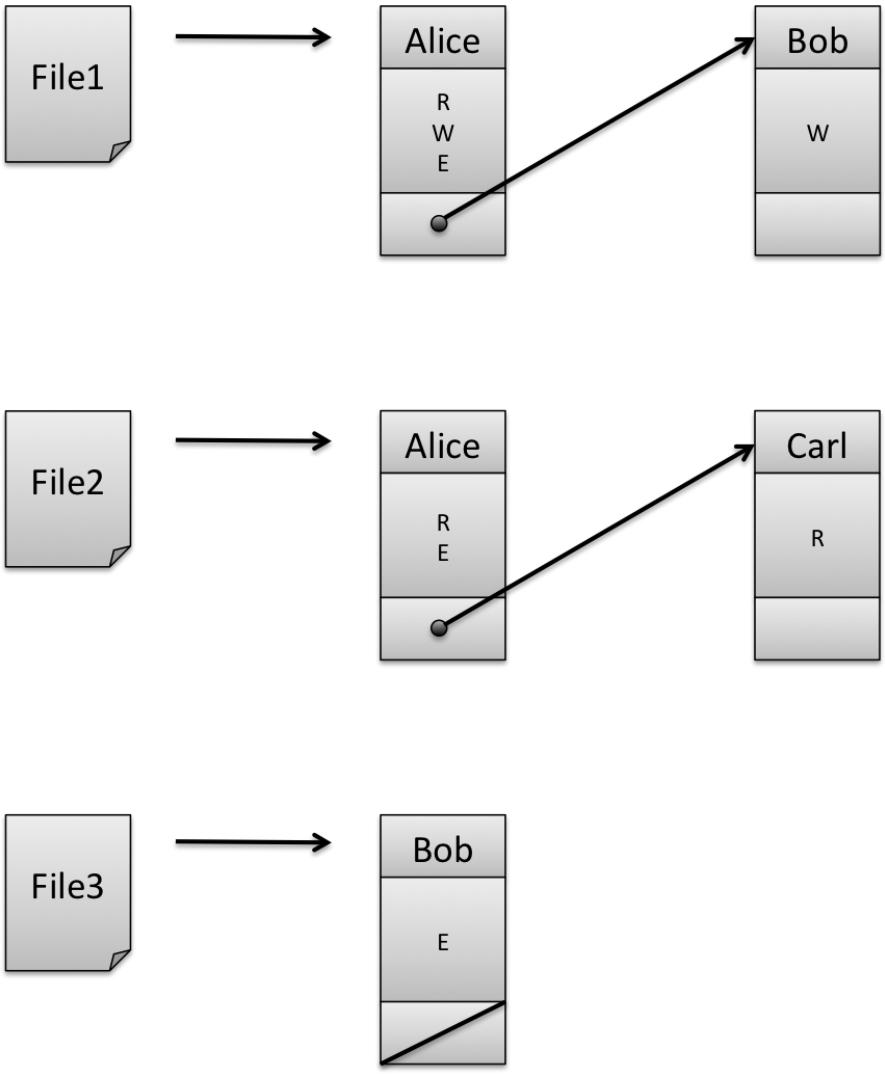

Figure 2.2: Access Control List

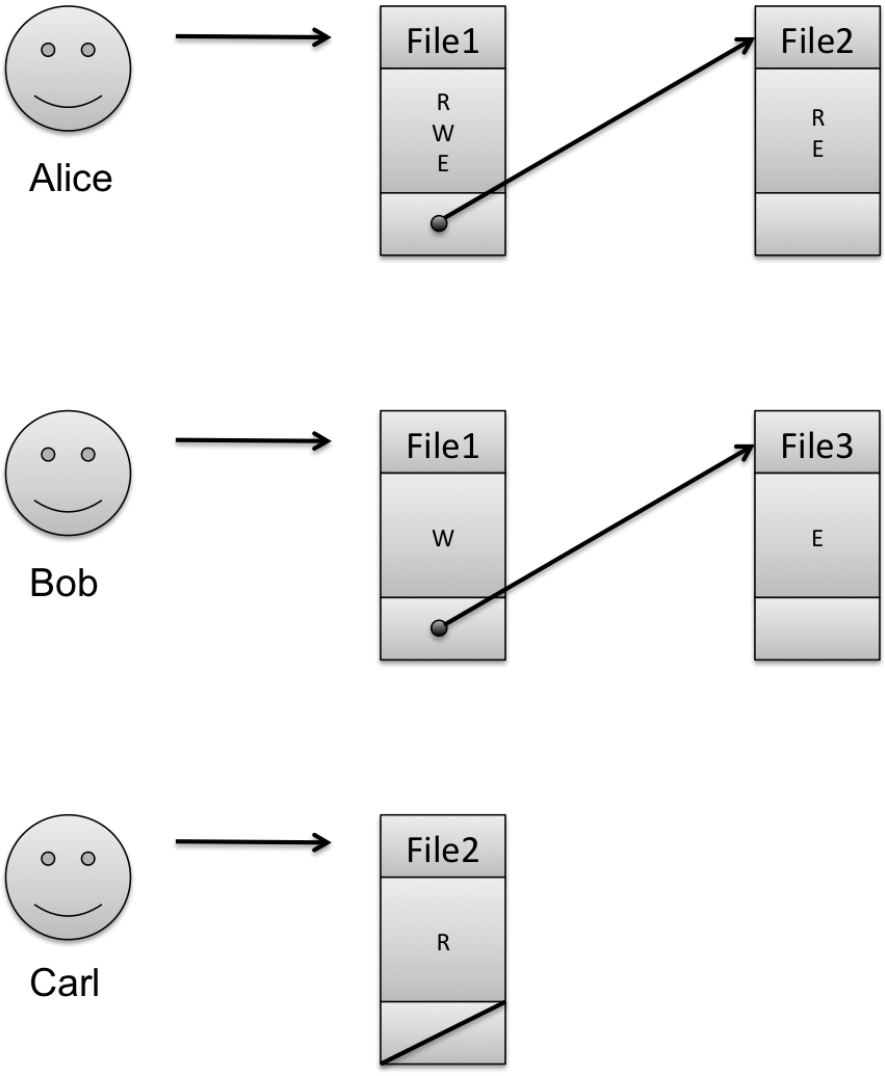

Figure 2.3: Capability List

$$
\begin{aligned}
U &= a\ set\ of\ users \\
R &= roles\ determined\ by\ job\ functions \\
P &= permissions \\
SE &= sessions\ =\ a\ mapping\ involving\ U,\ R\ and/or\ P \\
UA &= user\ assignment \\
PA &= permission\ assignment \\
RH &= role\ hierarchy
\end{aligned}
$$

RBAC emerges later than DACs' paradigm in about mid 1990's. Unlike DACs' rigid object-to-object basis of access control, RBAC consider more on the relationship between subjects and their organizations according to sessions or job tasks in a deployment view. In other words, instead of the subject, the subject function or role determines his/her access rights will be granted or revoked. In RBAC, it successfully addresses some of the limitations in DACs. DACs based on rigid subject-to-object access right assessment suffer a cumbersome process when being applied to large organizations with large amount of resources. RBAC in treating subjects by grouping them into clusters of roles improves the scalability of access control at an enterprise level. Since RBAC determines access rights based on roles, and several people may share the same role (eg. the role of data scientists), RBAC allows a group of people who have the same work function share the one set of access control permissions on a set of resources (eg. databases, source code, etc). This performs better scalability than DACs. Meanwhile, RBAC also allows users to be assigned to different roles at the same time. For example, a data scientist can also be assigned to a role of software engineer to share the access to source codes during projects on-going. Another explanation for multi-role of a user is from the role hierarchical side. One supervisor in a role hierarchy can access resources

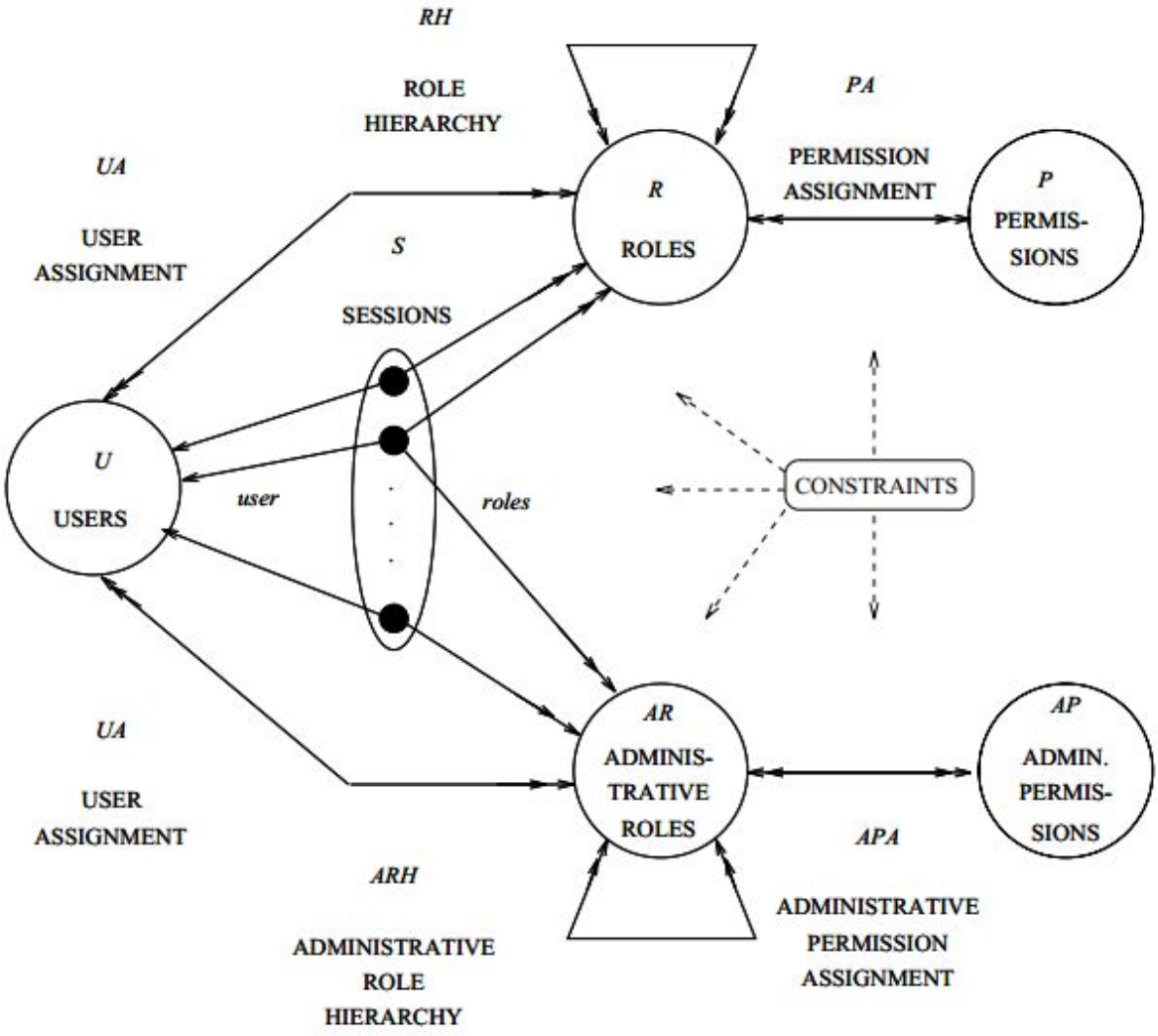

Figure 2.4: Role-Based Access Control Model (Sandhu et al. (1996))

which can be accessed by employees who are in a lower level of role hierarchy than him/her.

The assumptions of RBAC:

- A subject can have multiple roles.
- A role can have multiple subjects.
- A role can have many permissions.
- A permission can be assigned to many roles.
- An operation can be assigned many permissions.
- A permission can be assigned to many operations.

RBAC is increasingly being commonly used at the system and application level in enterprises due to its scalability and versatility. The applications include Microsoft SQL Server, Oracle DBMS, FusionForge, Solaris, SELinux and many other implementions of RBAC. Although RBAC has advantages compared to DACs model, it does have its own limitation. First, RBAC needs infrastructure work to deploy the entire model compared to DACs. Second, one of the most limitation of RBAC is by grouping individuals into a group, the model is difficult to define granular access controls for each one of them. This might require a more specific role to prevent an individual who is in a certain group but does not need full access rights according to the group permission assignments. In this case, the ability to differentiate individual members of a group and selectively grant or revoke access is needed. Therefore, the attribute-based access control (ABAC) is created to deal with the problem.

2.1.3 Attribute-Based Access Control

Attribute-based access control (ABAC) (Hu et al. (2015)) is a type of access control model where the determination of access is made through the evaluation of features or attributes associated with the subjects, the environment and the objects.

To concretely define an ABAC model, the following concepts need to be addressed:

$$
\begin{aligned}
S &= a\ set\ of\ subjects\ or\ users \\
O &= a\ set\ of\ objects\ or\ resources \\
E &= environment \\
SA &= \{SA_k | 1 \leq k \leq K\} \\
OA &= \{OA_m | 1 \leq m \leq M\} \\
EA &= \{EA_n | 1 \leq n \leq N\} \\
attr(s) &\subseteq SA_1 \times SA_2 \times \ldots \times SA_K \\
attr(o) &\subseteq OA_1 \times OA_2 \times \ldots \times OA_M \\
attr(e) &\subseteq EA_1 \times EA_2 \times \ldots \times EA_N
\end{aligned}
$$

$attr(s)$, $attr(o)$ and $attr(e)$ denote the attribute set of subjects, objects, and environments. In general form, whether a subject s can access an object o under a particular environment e is determined by a Boolean function of the attributes of s, o and e.

$$can\_access(s,o,e) \leftarrow f(attr(s), attr(o), attr(e)) \tag{2.1}$$

The attributes in ABAC are distinct fields. Subjects' attributes are considered to be the subjects' characterisitics which can be used as tags to differ one subject from another in access control scenario sense, such as identifier, name, job title, role and etc. Objects' attributes usually include the description of their functionality, content and any other information on the value of objects' usgae in access control, such as modified time, title, tags, file format and etc. Environments' attributes are used to record the status of environments, like current date time, CPU temperature, current threat level, and etc. Therefore, the access rule basically means to access the permission of access right of a subject s to an object o under the environment e.

A key advantage of ABAC is that the subjects are well abstracted by their attributes, which means the system or administrator does not need to know the subjects in advance. Therefore, ABAC is extremely useful to systems which constantly have unanticipated users. It makes arbitrary access of resources more efficient. Unlike RBAC and DACs application popularity. ABAC is usually implemented at an intermediary level to mediate access between a user or an application and the resource to which access is requested like XACML.

The shortcoming of ABAC is how to generate discriminant attributes of subjects, objects, and environment, especially in large scaled system. In large system, to use a complete list o attributes in determining whether a subject can access an object under a specific environment is huge waste. However, in large system, it is often desirable to harmonize the entire system or uniform access control policies enforced. Then there emerges policy-based access control (PBAC) to solve the problem.

2.1.4 Policy-Based Access Control

PBAC is said to be an evolutionary version of ABAC. Similar to ABAC, PBAC generates access control policies using attributes from subjects, objects, and environment. However, as there is a demand from enterprise level that access control policies should be harmonized across different departments of enterprises. ABAC is suitable to perform local access control as there is defined attributes to make decisions on access permissions. In enterprises, however, one department may only require password and job title to determine whether an employee can access a working spreadsheet, while another department may require all of the user's credentials to determine whether he/she can access the spreadsheet. Nevertheless, enterprises usually have some descriptive access control principles, which is hard to apply due to its abstraction. To enforce the harmonization of enterprise level, PBAC is then introduced to derive policies according to enterprises demand on abstracted access control principles by listing them concretely with rules in ABAC.

PBAC as an evolutionary version of ABAC is much more complicated than ABAC. The first problem is to maintain the attribute set over the enterprises. An Authoritative Attribute Source

(AAS) is the one source of sttribute data that can be used to standardize attributes in ABAC and PBAC. The second question is how to transform the abstracted access control principles into rules that can be implemented in application level. One most widely used tool of this transformation is eXtensible Access Control Markup Language (XACML) based upon XML. It is developed to specify the policies into a machine readable format.

2.1.5 Risk-Adaptabe Access Control

Enterprises are constantly evolving with the progressing economic aim and financial realities, and market challenges. The dynamic nature of enterprises and any other organizations requires high adaptive ability of access control policies which can constantly assess the risk of information propagation in enterprise level. Thus, risk-adaptable access control (RAdAC) model emerges, since the previous access control models including ABAC and PBAC cannot adequately meet the need for dynamism in the risk assessment. RAdAC was intended to be created as a real-time, adaptable, risk-assess access control facility to enterprises.

Figure 2.5 (McGraw (2009)) shows the flow chart of RAdAC notional process. Before access requests, RAdAC evaluates the security risks to decide whether the existing policies of access control need to be overridden or not. RAdAC leads a profound shift from static access control model to dynamic access control models. For example, if the computer system is normal, users can access the resources on the computer via username and password. However, if a security level is breached, RAdAC will enforce a much more stricter access control policy to protect the resources being disrupted.

Despite the attractiveness of RAdAC in its dynamic property, the implementation is daunting. First, RAdAC faces the difficulty in sharing the access control model of different organizations. This requires a standard way of interpreting risks among organizations. Second, RAdAC in assessing risks needs to gather as much trustworthy information about subjects, objects, environments and risk factors as possible. However, there is no standard way to extract these useful sets of information. Third, data exchange on recording the information needed to assess risks needs to have

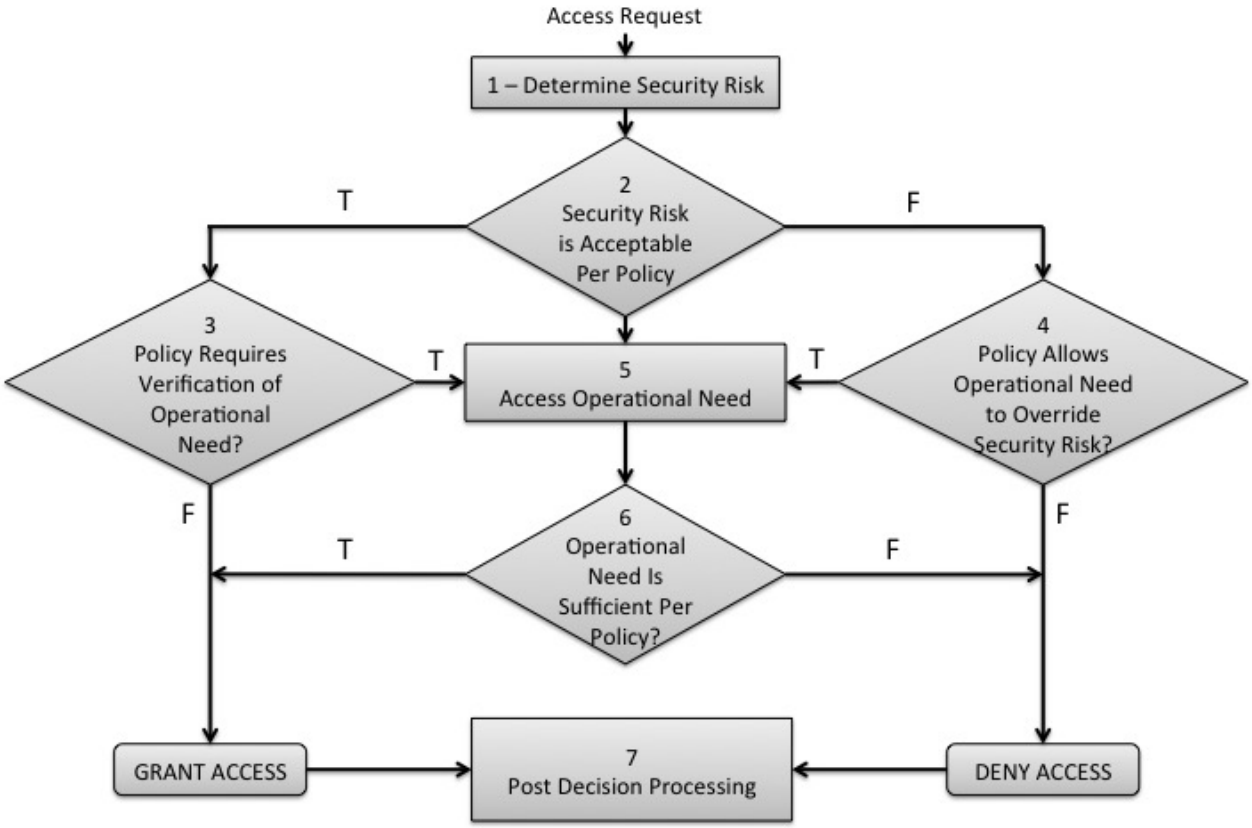

Figure 2.5: Risk-Adaptable Access Control Notional Process (McGraw (2009))

standard formats to ensure that RAdAC's efficiency of adaption to changeable environments. Last but not least, the heuristics of assessing risks relies on advanced techniques, including machine learning, and genetic algorithms which are still open research topics far from being solved and defined.

2.1.6 Access Control Based on Content

Bertino *et al.* pointed out that "*mechanisms for enforcing access control policies based on data contents*" are needed for comprehensive data protection (Bertino et al. (2011)). More relevant to the

proposed research, the notion of *content-based access control* has been used in relational access control specification (Bertino et al. (1997); Giuri & Iglio (1997)), multimedia database (Tzelepi et al. (2001); Tran & Dang (2007)), web 2.0 (Hart et al. (2007); Monte (2010); Hart (2006)), and digital libraries Adam et al. (2002), etc. However, their definition is quite different from ours. In particular, in Bertino et al. (1997); Giuri & Iglio (1997); Adam et al. (2002); Amjad (2007), the notion of *content* refers to attribute values or definitive concepts extracted from digital library objects. Access privileges are *statically* specified based on relationships between user credentials and attributes/concepts. Similarly, policy-based access control models (Kagal et al. (2003); Bhatti et al. (2007); Reddivari et al. (2005)) bind access rights with user credentials, however, the decision is still based on definitive values of the attributes (e.g. users with `title="physician"` could access patient records in his/her `department`). In Tzelepi et al. (2001), RBAC is extended to specify access control policies on image content (captured as attributes). Bertino et al. (2000) and Tran & Dang (2007) enforce access control of video databases based on text annotations on videos, while Bertino et al. (2003b) manages videos in clusters (based on visual content), and supports more flexible access control. In all cases, explicit and static rules are required – user credentials, video content and access control policies are all explicitly defined *a priori*.

More recently, Hart et al. (2007); Monte (2010); Hart (2006) enforces access control in Web 2.0 based on tags of messages, where tags are learned from the message content. Access control is explicitly specified on tags, for instance, there are explicit rules such as: "`[family members]` are allowed to access messages tagged with `[home]`". To handle the dynamics in modern enterprise applications, a few recent proposals attempt to infer access control provisioning from known decisions using supervised learning, when a decision cannot be directly made from available policies Molloy et al. (2012); Ni et al. (2009). This approach is effective when a good number of training samples (known access decisions) are available, and training and testing samples statistically follow the same distribution. On the other hand, concept-level access control has also been proposed for the semantic web (Qin & Atluri (2003)). Last, the terms *context* and *semantic* has been used in various access control approaches. *Context* mostly refers to the operational context of the user

Toninelli et al. (2006), while *semantic* is often used to indicate the semantic of data schema and access control policies, especially in data integration and federation applications Pan et al. (2006); Fabian et al. (2012).

Our notion of content-based access control is significantly different from existing approaches, we refer to the *semantic* content and semantic similarity of data in RDBMS or XML DB, as well as the notion of *approximation* and *implicit access control specification* in access control. In our approach, *content* refers to the meanings of data objects, which is significantly different from those concepts. Last, Oracle's CONTEXT index is essentially an *inverted index* for text retrieval, which is very different from the *context* used in access control literature.

2.2 Oracle Virtual Private Database (VPD)

Oracle offers special supports in database security for database administrators (DBAs) called Virtual Private Databases (VPD). Oracle Virtual Private Database (VPD) enables developers to generate security policies of database access at the row and column level. It essentially adds on a dynamic WHERE clause to a SQL query issued by the users against tables, or views to restrain the result set from being revealed with the entire content of the tables or views. In the sense, VPD enforces security to a finer level of granularity, directly on database tables or views, as it requires developers or database administrators (DBAs) attach the security policy together with its implementation to the database objects via PL/SQL functions, or packages. When users log in the database and issue a query against the database objects, security policy enforced by VPD will automatically applied. The policy is usually determined by the credentials of the users, such as identity, job title, and department. In CBAC, the policy is enforced according to the data objects owned or can be accessed by the user initially. With VPD, there is no way to bypass the enforced the security policy.

VPD utilized appending a dynamic WHERE clause or modifying WHERE condition dynamically to realize the enforced security policy. The modification is returned by a function implement-

ing security policy. Users can apply VPD policies to SELECT, INSERT, UPDATE, INDEX, and DELETE statements together with the implemented function. For example, if a user issued a query as follows to the table OE.ORDERS for which security is enforced by VPD as the user can only see the data objects from his/her own.

```
SELECT * FROM OE.ORDERS;
```

Then VPD would likely appends a WHERE clause to rewrite the query as

```
SELECT * FROM OE.ORDERS
WHERE SALES_REP_ID = SYS_CONTEXT('USERENV','SESSION_USER');
```

There are several benefits of using VPD. First, VPD can enforce security policy to a fine-grained access control level. It also makes the on-the-fly access control possible. Second, the implemented function associated VPD enforced security policy is added once to make the access control enforcement much more simpler. Third, via VPD, developers and databases administrators (DBAs) have the flexibility to add several security policies to a table or a view. The policies can be addressed differently on SELECT, INSERT, UPDATE, and DELETE.

To generate a dynamic WHERE clause, a function should be implemented first to define the concrete steps of policy to be enforced. There are certain restrains on the function. First, it must take a schema name and an object (table, or view) as inputs. Second, the return value should be a VARCHAR2 type to hold the appending WHERE clause string. Table 2.2 shows an example of the function enforcing a security policy example listed above.

In the example, the input arguments take in the schema, and object which the security policy is applied on. The declared return value con provides a string which will be appended to WHERE clause condition in SQL statement. con basically controls the result set belonging to the current session user. Further examples can be found in Chapter 5.

After creating the VPD function, a policy attaching the function to objects should be specified, as in the unction, the schema and object are used as input arguments. To create the policy, DBMS_RLS package in Oracle has to be utilized. Please note the developer of the security policy should be granted with EXECUTE privileges in using this package. Table 2.3 shows a concrete

Table 2.2: VPD Function Example

```
CREATE OR REPLACE FUNCTION hide_other (
 v_schema IN VARCHAR2,
 v_objname IN VARCHAR2)
RETURN VARCHAR2 AS
con VARCHAR2 (200);
BEGIN
 con := 'SALES_REP_ID = SYS_CONTEXT(' ' 'USERENV' ' ',
 ' ' 'SESSSION_USER' ' ')';
 RETURN (con);
END hide_other;
/
```

example of the policy creation. Further examples can be found (Oracle (2012)).

In CBAC, we utilize Oracle VPD to generate on-the-fly and offline access control policy. In the policy, we specify that when a user issue a query against content-centric databases, he/she will get a result set similar to the data objects he/she owns or granted access right initially.

Table 2.3: VPD Policy Example

```
BEGIN
 DBMS_RLS.ADD_POLICY(
 object_schema => 'OE',
 object_name => 'ORDERS',
 policy_name => 'secure_update',
 policy_function => 'hide_other',
 policy_type => DBMS_RLS.SHARED_CONTEXT_SENSITIVE);
END;
/
```

Chapter 3

Text Feature Extraction

In content-centric database, the first question before measuring any similarities between free texts is how we treat the free text, or concretely, how we can convert the free texts into some type of data, which is computable. That is the very question before any type of problems in text mining, usually called text feature extraction. Traditionally in text mining, when we have our text represented as a sequence of characters, one of the usual next step is to convert it into a sequence of words. The other is to treat the text into sequences of characters. The most commonly used word-based feature is term frequency inverse document frequency (TF-IDF) which is viewed as a term-distributed feature totally relying on the word occurrences in documents. This feature entirely neglects the sequences of words. Meanwhile, character-based feature considers sequences of characters by calculating statistics (frequency) of sequences of n characters or n words. Aside of features based purely on term distributions, topic modelings tried to extract semantic features with term distributions and statistical information on documents in an unsupervised way. In this chapter, we first introduce two kinds of term-distributed features of free text. After that, we will introduce two commonly used topic modeling algorithms. Last, we will come to an annotation tool of tagging text with topics of Wikipedia, which goes a step further to bring in authorized database annotation as topic seeds for document tagging.

3.1 TF-IDF

TF-IDF (Joachims (1996)) is abbreviation for term frequency-inverse document frequency which is a statistic to reflect the importance of a word to a document in a corpus of documents. It is commonly used in text mining, and information retrieval. Before TF-IDF became the most popular term-distributed feature in text mining, different term weighting schemes had been explored (Salton & Buckley (1988)). TF-IDF basically considers two factors to determine how important a word can differentiate a document from the others in the corpus, namely, term frequency and document frequency. Term frequency (TF) generally calculates the ratio of a word t appears in one document d over the number of total words in the document. In practice, there are variants of term frequency. If we denote the raw term frequency as $f(t,d)$, boolean term frequency is: $f(t,d) > 0 \Leftrightarrow tf(t,d) = 1$. $f(t,d)$ can also be directly applied in computing term frequency. In realistic application, to prevent the bias towards longer documents, augmented term frequency shown in Equation 3.1 is most commonly used.

$$tf(t,d) = 0.5 + \frac{0.5 \times f(t,d)}{max\{f(t,d) : w \in d\}} \tag{3.1}$$

Document frequency is equal to the ratio of the word t appears in how many documents in the corpus over the entire number of documents in the set. The inverse document frequency (IDF) basically measures how common the term is in the set of documents. The more common the term is, the less important the term is to differentiate a document from others, according to Equation 3.2 calculation.

$$idf(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}|} \tag{3.2}$$

The formula demands $|\{d \in D : t \in d\}|$ to be greater than 0, which means every term in calculation must at least occur once in the set of documents. Based on the principle, we need first to collect all the terms in all documents to form a specified dictionary for TF-ID calculation. Having got both of the two factors ready, the TF-IDF is calculated as a multiplication of TF and IDF.

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D) \tag{3.3}$$

After extracting TF-IDF features from every term of every document, each document is treated as one single sample with a representation of a vector. Therefore, documents are transformed into a mathematical vector space. In Equation 3.4, $w_{j,i}$ denotes the i-th term's TF-IDF value in the j-th document. There are n terms in the specified dictionary.

$$d_j = \{w_{j,1}, w_{j,2}, \ldots, w_{j,n}\} \tag{3.4}$$

The similarity of two documents is therefore equal to the cosine product of two vectors of documents represented as Equation 4.1.

$$Sim(d_j, d_k) = \frac{\sum w_{j,i} \cdot w_{k,i}}{\sqrt{\sum {w_{j,i}}^2 \cdot \sum {w_{k,i}}^2}} \tag{3.5}$$

3.1.1 Stop Word

In text feature extraction, there are some words, which usually occur in documents; however won't contribute to any semantic content in the documents, such as, "a", "an", "the", etc. These words, although having grammatical meanings, are considered as meaningless words in text. Another type of words has very concrete meanings; however is too general to contribute to a specified topic, sometimes is also considered as "meaningless" in text, such as "with", "somehow", etc. They together form a group called "stop words". In different languages, there are different lists of stop words. Even in English itself, there are several different versions of stop word lists. In text mining, nowadays, people tend to use comprehensive stop word list from google search engine. In the experiments, we filter out the stop words according to most search engines, when calculating TF-IDF.

3.1.2 Stemming

Another important step for preprocessing before feature extraction of texts is called stemming (Porter (2001)). Stemming solves the problem that defined calculation of features will innately ignore the same meanings of “apple" and “apples", and treat them as two different terms. In linguistic morphology and information retrieval, stemming is the process for reducing inflected (or sometimes derived) words to their stem, which is the base or root form generally a written word is from. The stem needs not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root. Algorithms for stemming have been studied in computer science since 1968. Many search engines treat words with the same stem as synonyms as a kind of query broadening, a process called conflation. Stemming programs are commonly referred to as stemming algorithms or stemmers. One of the most famous stemmer is Porter’s Stemmer. The Porter stemming algorithm (or Porter stemmer) is a process for removing the commoner morphological and inflexional endings from words in English. Its main use is as part of a term normalization process that is usually done when setting up Information Retrieval systems. This, however, is a problem, as not all parts of speech have such a well formulated set of rules. Thus, the Porter’s Stemmer, although is simple to use, sometimes goes too far to shrink the duplicity of terms. Also there is no clear evidence showing that text feature extraction with stemming boosts the performance of text understanding. Therefore, in the experiments, we keep all variations of words without stemming, when calculating TF-IDF.

3.2 *n*-Gram

Clearly, TF-IDF totally ignores the sequences of words. However, in text mining, sequences of words especially for multi-word phrases play important roles in understanding text content. One of the common ways to address the problem is to use *n*-gram feature instead of TF-IDF. The difference between TF-IDF and *n*-gram is that *n*-gram treats *n* adjacent words ($n > 1$) as one

single “term" to introduce local sequences of words. Thus, for every n-gram feature, the feature is represented as $\{t_i, t_{i+1}, \ldots, t_{i+n-1}\}$. The features are collected for every document in the document set and unified into a specific “dictionary", which contains all n-gram features occurred in all the documents. After n-gram features are extracted, every feature can be treated as a new special “term". Therefore, all the term frequency, inverse document frequency, similarity measurement can be easily applied on n-gram features.

In n-gram, stop words removing is vital. Even though n-gram incorporates local sequences of words to prevent the loss of sequence information in TF-IDF. However, n-gram usually suffers the problem of maintaining a lot of fake phrases and noisy “terms". Keeping stop words will make the problem even worse and it will make the “dictionary" even longer than it could be. The general problem of term-distributed text feature is the sparseness of the vectors in the feature spaces. A longer “dictionary" list will increase the vector dimensionality which will make the text representative feature even sparser.

Another problem of term-distributed features is although they maintain the appearances of words, and terms or even phrases, it does not mean that using these features makes the understanding of text much easier. There still stands the gap of semantic meaning behind the words in the documents due to the flexibility in explaining the same meaning with different words and addressing different meaning with similar words. Topic modeling was raised to tackle the problem. Researchers tried to extract “topics" based on statistical analysis of term distributions. In the next section, two commonly used topic modeling algorithms are introduced. However, intuitively, the model can be just as good as the data inputs. These “topics" are generated by pure term distributions, by which means it will only reflect what term distributions will tell us. Therefore, the “true" semantic representation of documents is a difficult topic interested lots of researcher in academy and industry. Nowadays, researchers start to use annotation tagging schemes on text to extract not direct words, terms or phrases, but related topics. One of the annotation tool is TAGME (Ferragina & Scaiella (2010)). It uses the Wikipedia as a backup data source to annotate topics on text content. The motivation of TAGME is to address the text understanding in short text where it is hard to

find two documents share the same word in document set. That is the exact situation where term-distributed feature would fail. We will further our discussion on TAGME, an annotation tagging tool on text understanding in the last part of this chapter.

3.3 Topic Modeling

Topic modeling is a statistical model in natural language processing for extracting“topics" in a collection of documents in the statistical sense. The algorithm behind topic modeling either treats documents as mixtures of topic (distributions) or treats documents as linear combination of topics, which are the basis of the documents' space. The representation of “topics" in topic modeling is collections of words/terms. Here we introduce two models: Latent Dirichlet Allocation (Blei et al. (2003)), and Non-negative Matrix Factorization (Lee & Seung (2001)), which represent two different strategies of topic modelings.

3.3.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative model, which is capable to discover “topics" from document repository in an unsupervised way. It treats documents as mixtures of “topics" and “topics" are represented as Dirichlet distributions. Innately, LDA is a non-parametric Bayesian inference model.

Figure 3.1 is the plate notation for LDA. The outer plate denotes the document repository; the inner plate denotes the topics and words within each document. M is the number of documents in the repository. Therefore, we have the follows:

- α is the parameter of the Dirichlet prior on topic distribution across documents,
- β is the parameter of the Dirichlet prior on word distribution on topics,
- θ_i is the topic distribution for the i-th document,
- ϕ_k is the word distribution for the k-th topic,

- z_{ij} is the topic for the j-th word in i-th document, and
- w_{ij} is the j-th word in the i-th document.

The problem of the basic LDA is that a new coming document is very likely to contain words that did not exist in the training document repository. In the basic LDA, when using maximum likelihood of multinomial parameters, zero probability will be assigned to such words, and thus zero probability to new documents. Therefore, a smoothed model of LDA was raised to extend the Bayesian inference model, which treats β as random variables. The plate notation of smoothed LDA is shown in Figure 3.2, in which K denotes the number of topics, and ϕ is a $K \times V$ matrix, where each row denotes the word distribution of a topic (V is the dimension of vocabulary).

The generative process of LDA treats documents as random mixtures over latent “topics", and “topics" are viewed as multinomial distributions over words. Assumptions behind LDA are as follows.

Given a corpus D of M documents each with length N_i:

- Choose $\theta_i \sim \text{Dir}(\alpha)$, where $i \in \{1,\ldots,M\}$ and $\text{Dir}(\alpha)$ is the Dirichlet distribution.
- Choose $\phi_k \sim \text{Dir}(\beta)$, where $k \in \{1,\ldots,K\}$.
- For each word $w_{i,j}$, where $i \in \{1,\ldots,N_i\}$, and $j \in \{1,\ldots,M\}$.
  - Choose a topic $z_{i,j} \sim \text{Multinomial}(\theta_i)$.
  - Choose a word $w_{i,j} \sim \text{Multinomial}(\phi_{z_{i,j}})$.

3.3.2 Non-negative Matrix Factorization

Non-negative matrix factorization factorizes a matrix V into two matrices W and H, where the elements of these three matrices are greater than or equal to zero.

Formally, we have that given a non-negative matrix V, find two non-negative matrices W and H such that:

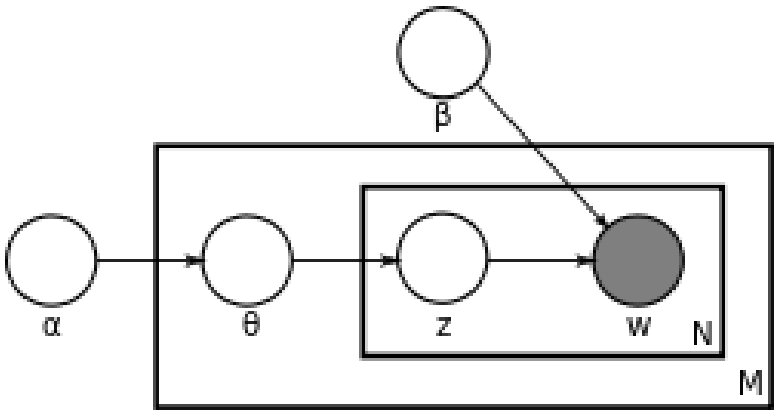

Figure 3.1: Plate Notation of Latent Dirichlet Allocation

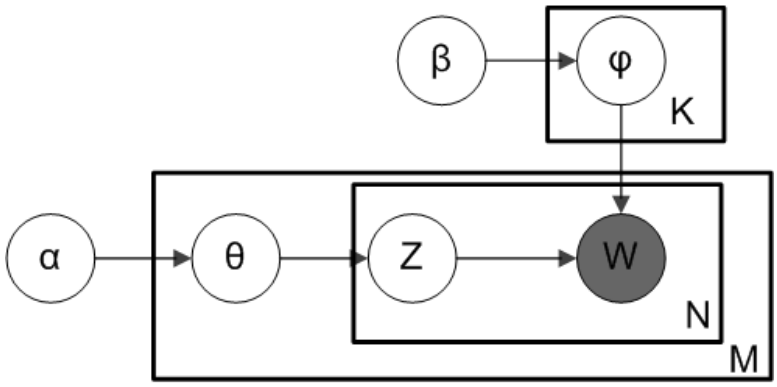

Figure 3.2: Plate Notation of Smoothed Latent Dirichlet Allocation

$$V \approx W \times H \tag{3.6}$$

Consider if V is a $m \times n$ non-negative matrix which containing TF-IDF values for each word in the vocabulary of each document. Every row represents the feature vector of a document. W (a $m \times k$ non-negative matrix) is the topic weights for each document; while H (a $k \times n$ non-negative matrix) is the topic distribution across the vocabulary. That being said, innately NMF tries to reconstruct V by linear combination (W) of “topics" (H).

3.4 TAGME

Although topic modelings make their efforts to extract semantic representations of documents based on statistical assumption and analysis, it still embeds the disadvantages of TF-IDF. The “topics" are either weighted sum or mixture of terms in the existing term distributions, even given the smoothing version of LDA. Therefore, researchers tried to bring in some external annotation database. Here comes TAGME with the topic annotation from Wikipedia. Before TAGME (Ferragina & Scaiella (2010)), text annotation has been studied (Carpineto et al. (2009); Kulkarni et al. (2009); Mihalcea & Csomai (2007); Milne & Witten (2008)). TAGME first implemented a web-based API tool motivated to address annotating short texts. It uses Wikipedia anchor texts as spots and the pages linked to them in Wikipedia as their possible senses. By finding collective agreement among these pages, TAGME is able to calculate out a new score to verify the importance of the topic (Wikipedia page title) in the input text. TAGME considering the sparseness of the anchors in short texts, combines the relatedness function among concepts (Witten & Milne (2008)) and probabilistic statistics drawn from Wikipedia. TAGME is available on `http://tagme.di.unipi.it`. Figure 3.3 shows an example posted on the web site. The above text box is used to get the users' inputs of text. The right bar is used to tune the threshold of annotation. The higher the bar sets, the more related topics are annotated to the input text. The box below shows the annotated results. The topics are linked to Wikipedia web pages via blue phrases. Innately, TAGME uses an XML ile to represent the annotated topics with the title of the linked Wikipedia pages, and for each topic, there is a parameter call ρ in the range of 0 to 1 to denote the importance or goodness for the topic annotation.

Figure 3.3: TAGME Annotation Example

Chapter 4

Content-Based Access Control Model

4.1 Background and Assumptions

In this chapter, let's first introduce the basic assumptions of problems specified in the dissertation. To satisfy the new needs of enforcing access control without explicitly identifying every subject and object in the access control policies, we propose *content-based access control* (CBAC). CBAC, as an addition to the existing database access control models, works for content-centric data sharing scenarios that satisfy the following assumptions:

1. Subjects and objects: We assume there are large amounts of users, who are authenticated with some level of trust. Each user is expected to get access to a subset of the data objects in databases. In practice, we can treat such set of users as a special role in the database. There are also large amounts of data objects (e.g. records), and the data is content-rich in nature. Each data object features with a block of unstructured textual content (e.g. a CLOB type attribute), which is a key part in every data object.

2. Data-driven access control decisions: In CBAC, it is assumed that the access control decision for each user against each data object is expected to be data (or content)-driven. In particular, the decision is supposed to be determined by the content similarity of the textual data with the data objects held by each user him/her self, as we have illustrated in previous examples.

3. Lack of explicit authorization: Another important assumption is that there might be situations that explicit authorizations (for each user against each data object) are not available, since it requires excessive labor to examine the textual content for each record and make a verdict for each user. Moreover, the data set is dynamic that new records are added constantly, hence, access control verdicts need to be made on-the-fly.

4. Approximation: In contrast to the conventional data confidentiality notions, approximation is allowed in this scenario. That says, it is acceptable if a user (of the special role) accesses a few more (or a few less) records than it would have been assigned by an administrator. This assumption states there should be approximation tolerance in the demands of access control policies.

Example 5: If we revisit Example 1: assume that only 15 cases in the database are relevant to Alice's case, that says, a careful and accurate director would only allow Alice to access those 15 records. However, if an automated mechanism blocks a small portion of these 15 records, or allows Alice to access a few other records, it is considered to be acceptable, especially comparing with the current practice which gives Alice access to all the records.

∎

In a nutshell, in CBAC, each user is allowed by a meta-rule to access a subset of the designated data objects, but the boundary of the subset will be dynamically determined during query processing, and the decision is data-driven.

Our goal is to design an access control model that considers the textual content of data objects in making access control decisions, and to develop a mechanism that enforces this model efficiently. Meanwhile, the access control enforcement mechanism is expected to have the following features:

- Autonomous: CBAC enforcement mechanism is expected to require minimal intervention from the system administrators and data owners.
- Transparent: Users of the CBAC role are expected to issue queries as usual – without being affected by the existence of the access control mechanism.
- Efficient: Although content similarity assessment could be computationally expensive, we

still expect the CBAC enforcement mechanism to return answers promptly.

- Off-the-shelf: The CBAC enforcement mechanism is expected to employ native access control capabilities from off-the-shelf database management systems, so that the proposed model and mechanisms could be easily adopted.

4.2 Contribution

Let's revisit Example 1 and 4. Both examples demonstrate applications where explicit access control specifications at record level are too labor-intensive; therefore, users become significantly over-privileged due to the nonexistence of record-level content-based access control. The excessive privilege is somehow mitigated with two controls: (1) RBAC or MLS is enforced so that users have basic clearance to access the database; and (2) *ex post facto* auditing is enforced to punish misuse of the privileges. However, with the size of data, the basic clearance still allows a user to access an unacceptable amount of records. Meanwhile, *ex post facto* auditing does not reverse the damage, since the suspicious user has already committed the misfeasance, and it is impractical to revoke disclosed data. Ideally, we expect a more restrictive and automated access control model, instead of allowing users to be significantly over-privileged or requiring excessive human intervention. That is, the new model is expected to intelligently identify a smaller subset of records that are relevant to the user's task, and only grant access to this subset.

Attribute-based access control (ABAC) could be employed to partially mitigate the problem. For instance, in Example 2, we can specify access control based on a combination of doctors' and patients' attributes: a doctor may access records of patients that have ever been treated in his/her department. However, attribute-based access control may not work with unstructured text (free text) content. Moreover, when the database structure and the attributes are very complicated, it may be difficult to obtain closed-form expressions for ABAC policies. That is, in managing content-rich data, it may be difficult to describe administrators' access control intensions with a small set of explicit, closed-form policies. In such applications, "hard security" requires a high

price of excessive human labor and degradation of usability (e.g. waiting for manual authorization in BTG). From the technology perspective, there does not exist a computational model to precisely describe semantic content, or to model this human cognitive process – the rationale behind the decision is too vague and complicated.

In such use cases, it is expected to have an access control model that extends ABAC to make access decisions based on the *semantic content* of the data. It is also desired that such content-based access control capability to be provided by RDBMS as native functions, and only requires minimal intervention from administrators. In this paper, we present our first attempt towards this endeavor: we introduce the *content-based access control* model and enforcement mechanisms. In particular, we propose a two-phase hybrid solution: (1) the data owner or administrator manually identifies a small *base set* of records – the core of the set of records that are accessible to the user; and (2) at runtime, CBAC extends the base set and makes access verdicts according to specified CBAC rules, which are based on the *lexicon similarity* between the base set and the requested records. The new model, as an extension to ABAC and a complement to legacy access control approaches, provides an effective and efficient means of access control that exploits content features in content-rich data sharing.

We would like to emphasize that **content-based access control does not imply weakened or relaxed security**. Rather, it enforces an additional layer of access control on top of existing "precise" access control methods. CBAC allows approximation – it does not provide a static boundary for the accessible set of records. However, allowing the user to access a small set (size of n) of roughly (and automatically) selected records is **more secure** than allowing the user to access all the records in the pool (size of N), especially considering that n is usually orders of magnitude smaller than N.

Our contributions are three-fold: first, we formally propose a data-driven access control model that exploits the data content to achieve flexible and powerful access control semantics. Second, we develop an effective enforcement mechanism of CBAC utilizing native functions from off-the-shelf database systems. Last but not least, we further develop a blocking mechanism and labeling

mechanisms to improve the efficiency for CBAC enforcement, and to improve the accuracy of textual content matching.

4.3 Model Definition

A simple access control policy could be specified as:

$$ACR = \{subject,\ object,\ action,\ sign\}$$

where the *subject* denotes a user, and the *data object* could be a table, an attribute, or a tuple in the relational data model, or an XML node in the XML model. The action identifies an *operation* on the data objects, such as read, delete, update, etc, and the *sign* denotes whether the operation is allowed or denied.

In conventional database access control models, the *subject* could be identified by a user ID, a role (in RBAC (Zhang et al. (2002); Sandhu et al. (1996); Yang et al. (2004))), or an attribute (in attribute/credential based access control). In content-based access control, we assume that all CBAC users belong to a special role that is allowed to access "some data", as we have introduced. A CBAC user is further represented by a set of records owned by the *subject*. Particularly, a CBAC user is able to access his/her own data initially. Alternatively, the *subject* could also be a short description, which is modeled as unstructured textual content.

The *data object* could be a table, an attribute, or a tuple in the relational data model, or an XML node in the XML model. We assume *fine-grained access control* in CBAC: access control in enforced at record or node level. Hence, we consider each tuple in the relational model as a *data object*. Hereafter, we assume relational data, and we terms "record" and "data object" interchangeably. We will discuss the handling of XML data in Chapter 8.

The similarity between two *data objects* are defined as the weighted sum of the similarities across all *N* attributes interested:

$$Sim_d(\mathbf{d}_i, \mathbf{d}_j) = \sum_{x=1}^{N} w_{a_x} \times sim_{a_x}(d_{i,x}, d_{j,x}) \tag{4.1}$$

where sim_{a_x} is the normalized similarity unction defined on the domain of attribute a_x, and w_{a_x} is the weight on the attribute. Similarity functions on simple types (e.g. integer) could be relatively trivial, and the access control functions could be implemented using views. For instance, it is non-fancy to enforce "user *Alice* could access all the cases in 'San Francisco' that took place after 2010" using a view. On the other hand, we are more interested in content-rich unstructured text types, such as VarChar, CLOB or TEXT types, as we have introduced. In the big data era, unstructured text type is becoming more and more popular, meanwhile they carry much more information. For example, there are news, blogs, twitters, and facebook comments exploded every day in daily life which carry the rich information not just about what happened in the realistic world, but also the trend of technology, the opinions of citizens towards a piece of news, the possible interests on investigation from big agencies and so on. However, despite of the advantages and research interests researchers have in unstructured text type of data (free text data), it is also known as a difficult source of data to extract meaningful and useful information or being understood/assessed due to its unformatted characteristics and its exponential explosion in quantity.

In content-based access control, the static, binary notion of $sign \in \{true, false\}$ is extended to be a content-based access control function, which is evaluated on-the-fly during access control enforcement. The CBAC policy could be represented as:

$$ACR = \{subject,\ object,\ action,\ f(u, \mathbf{d}_i)\} \tag{4.2}$$

where *u* represents the user, and **d** represents the data object. For each query, the function evaluates to a value $f(u, \cdot) \in \{true, false\}$ for each $\mathbf{d}_i$ in the candidate result set. Access is granted when $f(u, \mathbf{d}_i) == true$, and hence $\mathbf{d}_i$ is included in the result to the query. In most cases, the decision function $f()$ consists a similarity function and a threshold to compare with:

$$f(u, \mathbf{d}_i) = Sim_d(\mathbf{u}, \mathbf{d}_i) \geq T \tag{4.3}$$

Moreoever, if the user is represented by a set of data objects (records) owned by him/her (as in Examples 1 and 3), the decision function will get the maximum similarity between the data object and the owner's records, and compare it with a reset threshold:

$$f(u, \mathbf{d_i}) = max(Sim_d(\mathbf{d}_{u,j}, \mathbf{d}_i)) \geq T \tag{4.4}$$

where $\mathbf{d}_{u,j}$ denotes the set of records owned by the user u, the similarity function $Sim_d(\cdot,\cdot)$ represents the content-based record similarity function we described above, and $max_j(Sim_d(\cdot,\cdot))$ will return the largest similarity for all j.

Example 6: Let us revisit Example 1. The content-based access control rule for *Alice* and other agents will be defined as:

$$\begin{aligned} ACR &= \{u, \mathbf{d}_i, read, f(\mathbf{D}_u, \mathbf{d}_i)\} \\ \mathbf{D}_u &= \{\mathbf{d}_j | Access(u, \mathbf{d}_j) = 1\} \\ f(\mathbf{D}_u, \mathbf{d}_i) &= Sim_d(\mathbf{d}_j, \mathbf{d}_i) > T, \exists \mathbf{d}_j \in \mathbf{D}_u \end{aligned}$$

That says, agent u is granted "read" access to record $\mathbf{d}_i$, when the content similarity between $\mathbf{d}_i$ and any one of u's records is greater than a preset threshold T. In particular, the director (administrator) first manually grants agent u access to a set of records $\mathbf{D}_u$ (mimicking the scenario that the director assigns cases to *Alice*), and agent u is thus represented by $\mathbf{D}_u$ in CBAC policies. The CBAC policy further allows agent u to read records that are similar to the ones in $\mathbf{D}_u$. This policy needs to be enforced by DBMS without requiring further attention from the administrators.

∎

4.4 Content Similarity

In the CBAC model, the similarity function $Sim_{a_x}(d_{i,x}, d_{j,x})$ for attribute a_x is respectively defined for different data types. As we have shown in the examples in Chapter 1, the CBAC model is mostly designed for content-rich unstructured text types, as as VarChar, CLOB or TEXT types. Ideally, we are expected to model such types by their *linguistic semantics*, and measure record-wise similarity based on the semantical content. However, natural language understanding remains an open problem, and it is very difficult to provide a reliable similarity assessment purely based on the linguistic content (Kao & Poteet (2007); Sharma (2010); Wu & Chen (2011)). For domain-specific data, customized semantic similarity measurements could be adopted, e.g. Pedersen et al. (2012) for medical data. While the specific attribute similarity measurement is not the focus of CBAC, in this paper, we present a generic measure that works for unstructured text attributes.

We model unstructured text by the statistical distribution of terms. The terms (words) from the selected textual attribute for all tuples are collected to construct a feature space (term space). Each cell is represented as a vector ($\mathbf{d}_i$) in the term space:

$$\mathbf{d}_i = [w_{1,i}, w_{2,i}, \ldots, w_{N,i}]$$

where $w_{t,i}$ is the TF-IDF weight of document i on term t. The original TF-IDF weight is defined as:

$$w_{t,i} = tf_{t,i} \times idf_t = tf_{t,i} \times \log \frac{N}{df_t}$$

where $tf_{t,i}$ is the frequency of term t in document i, and df_t is the number of documents that contain term t. Many variations of TF-IDF weight have been used in the research community (Manning et al. (2008)). Furthermore, the similarity between two documents could be calculated as the cosine similarity of two document vectors:

$$Sim_d(\mathbf{d}_i, \mathbf{d}_j) = \frac{\mathbf{d}_i \cdot \mathbf{d}_j}{|\mathbf{d}_i| \times |\mathbf{d}_j|}$$

Please note that the choice of content modeling and similarity measurement is not determined by the CBAC model, rather, it is the choice of the system administrators or data owners, who specify the policies. In Chapter place holder, we present a CBAC enforcement mechanism that exploits the content model and text similarity measurement embedded in the Oracle DBMS.

4.5 Top-K Similarity

In the basic CBAC model, content similarity is compared with a preset threshold, and the user is granted access to all of the "similar records". A potential problem is that the number of accessible records heavily depends on the preset threshold. In particular, it is difficult to estimate the range of pair-wise similarities (e.g. TF-IDF) among the records, especially for different seed records. Therefore, even for experts who are familiar with the similarity measurement, it could be difficult to give a reasonable threshold for any arbitrary user. With a bad threshold, the subject may be granted access to too many or too few records. Meanwhile, we also observed that using same threshold for different seed record will result in very different number of records in the positive set.

To tackle the problem, *top-K similarity* could be used. Instead of setting a threshold for record similarity values, the administrator could preset the number of data objects to grant access. For instances, in Example 1, we may define that Agent *Alice* is allowed to access 30 cases that are most similar to each of her own cases. The top-K similarity measurement provides flexible and intuitive control to database administrators and data owners, especially to users who are not familiar with the content model or the similarity measurement.

Enforcing top-K similarity is very similar to enforcing basic CBAC. An additional step is required to sort the similarity values for all records, and select the K largest. This step could be optimized by sorting only the top K elements, instead of sorting all records. Selecting the top

K elements could be done at linear time (e.g. QuickSelect). When sorting of the K records is not needed, the overall complexity is $O(N)$, where N denotes the total number of records in the database. On the other hand, when the top K records are to be sorted, the overall computational complexity of the overhead is $O(N + K \log K)$ (Hoare (1961)).

Chapter 5

CBAC Enforcement

In this chapter, we set up the basic content-based access control model with two different strategies: top-K similarity and threshold methods. First, we establish the framework of on-the-fly similarity assessment with Oracle's VPD modula. Second, we discuss offline content-based access control training, especially for the databases which are not updated that frequently.

5.1 CBAC On-the-Fly Enforcement

5.1.1 The Basic CBAC Model

In content-based access control models, we exploit Oracle's VPD modula on content-centric databases to perform row-wise access control. In the scenario, there are two basic assumptions. The first one is that the content of the database is contributed by multiple users (eg. *user1*, *user2*, and *etc*.) where everyone owns a certain amount of data objects and maintain by a special user (e.g. *John Doe*) or a database administrator. Second, the users are granted to access their own data initially and based on the data they owned, they are able to explore similar data objects.

In content-based access control model, the overall aim is to rewrite the user's query by appending a dynamic WHERE clause in which restrains the query in an accessible range other than the entire database. The accessible range is determined by the initial data objects the user can access.

Algorithm 1 CBAC threshold strategy

Require: a threshold score $T > 0$.

Require: A user u in a database of **D**.

Ensure: An array C_u which contains the d's in **D** that u can access.

1: u logs in **D** with his/her password and issues a query q against **D**,
2: **for all** d_i in **D** which can be accessed by u **do**
3: $Access(d_i, u) = 1$
4: **end for**
5: Let $\mathbf{D}_u = \{d_i | Access(d_i, u) = 1\}$.
6: $C_u \leftarrow \mathbf{D}_u$
7: **for all** d_i in $\mathbf{D}_u$ **do**
8: Calculate $sim(d_j, d_i), \forall d_j \in \{d_j | Access(d_j, u) = 0\}$
9: Select the ID's of d_j whose $sim(d_j, d_i) \geq T$, and add them into C_u.
10: **end for**
11: Append a dynamic where clause to q, such that the range of q is restrained within the ID's in C_u.

Algorithm 2 CBAC top-K strategy

Require: An integer $N > 0$.

Require: A user u in a database of **D**.

Ensure: An array C_u which contains the d's in **D** that u can access.

1: u logs in **D** with his/her password and issues a query q against **D**,
2: **for all** d_i in **D** which can be accessed by u **do**
3: $Access(d_i, u) = 1$
4: **end for**
5: Let $\mathbf{D}_u = \{d_i | Access(d_i, u) = 1\}$.
6: $C_u \leftarrow \mathbf{D}_u$
7: **for all** d_i in $\mathbf{D}_u$ **do**
8: Calculate $sim(d_j, d_i), \forall d_j \in \{d_j | Access(d_j, u) = 0\}$
9: Sort $sim(\cdot, d_i)$ descendingly, select the ID's of top N, and add them into C_u.
10: **end for**
11: Append a dynamic where clause to q, such that the range of q is restrained within the ID's in C_u.

In the scenario of content-based access control, we consider the situation when a user logs in the database and issues a query against the entire database, CBAC model will first check the accessible data objects of the user according to the data objects owned and feedback the relevant instances according to the content similarity between the user's data objects and the rest part of the database. It seems as before handling any query from the user's input, the system first submit "pre-queries" to collect similar data objects which the user is able to access currently. The two strategies of content-based access control model are concretely described in Algorithm 2 and Algorithm 1. The basic model shows the aim to limit the every user's accessible range and prevent the entire database from being revealed to unauthorized users.

5.1.2 Experiments

In CBAC enforcement, we exploit Oracle's VPD modula to implement record-level access control on content-centric databases. To enforce CBAC model in VPD, we rewrite the user's query by appending a dynamic predicate, which represents the content-based access control semantics. The range of accessible records is determined by the user's base set, as well as the similarity-based access control function. As we have introduced, there are two types of CBAC policies: (1) threshold-based CBAC, and (2) Top-K CBAC. In this section, we assume that similarity assessments are performed on-the-fly.

Settings. In the experiments, we utilize the NSF research awards data set from UCI KDD repository (Bache & Lichman (2013)). The original data set contains 129,000 instances. Records containing empty abstracts are removed, and thus leaves 102,508 awards. Awards are extracted, parsed and loaded into three tables: `Award_Basic(A_ID, Title, A_Instr, Div, abs, S_date, E_date, Ex_tol_amt)`; `Aw_Intr(A_ID, I_ID)`; `Investigator(I_ID, I_Name, I_Email)`. In particular, attribute `AWARD_BASIC.abs` contains full-text abstracts of NSF awards, representing the content-rich information. To demonstrate the scalability of CBAC enforcement, we increase the number of records in the database by adding synthetic dummy records. We employ an automatic

Table 5.1: Schemas

TABLE	COLUMNS
AW_INTR	(A_ID, I_ID)
INVESTIGATOR	(I_ID, I_NAME, I_EMAIL)
AWARD_BASIC	(A_ID, TITLE, A_INSTR, ABS, S_DATE, E_DATE, EX_TOL_AMT)

Table 5.2: Column Description

COLUMN	DATA TYPE	DESCRIPTION
A_ID	NUMBER	Award ID
I_ID	VARCHAR2(20)	Investigator's ID
I_NAME	VARCHAR2(100)	Investigator's Name
I_EMAIL	VARCHAR2(100)	Investigator's Email
TITLE	VARCHAR2(100)	Award Title
A_INSTR	VARCHAR2(100)	Award Institution
ABS	CLOB	Award Abstract
S_DATE	DATE	Start Date
E_DATE	DATE	Expiration Date
EX_TOL_AMT	NUMBER	Expected Funding

CS paper generator *SCIgen*[1] to generate very large amount of content-rich but meaningless records. Eventually, we have constructed a database with 2,714,025 records for the experiments. Note that the content in this database is not sensitive thus it does not require access control, however, it mimics content-centric databases, for which CBAC is designed. Table 5.1 describes the columns of each relevant table, and Table 5.2 gives description of the columns in the tables.

We use Oracle 11g for the experiments, and apply CONTEXT indexing on the `AWARD_BASIC.ABS` attribute. In order to optimize the similarity calculation, we remove the common English stop

[1] Available at: `http://pdos.csail.mit.edu/scigen/`

Table 5.3: CBAC Top-10 Example

```
SELECT A_ID FROM
(SELECT ROWNUM m, score(1), A_ID
FROM JOHNDOE.AWARD_BASIC
WHERE ID NOT IN
(SELECT A_ID FROM JOHNDOE.AW_INTR
WHERE I_ID=SYS_CONTEXT('USERENV','SESSION_USER'))
AND CONTAINS (ABS,
 '<query><textquery grammar="CONTEXT" lang="english">
 Markov chain Monte Carlo (MCMC)
 methods are an important algorithmic
 device in a variety of fields.
</textquery>
<score datatype>="float" algorithm="DEFAULT"/>
</query>',1)>=0 ORDER BY score(1) desc)
WHERE m BETWEEN 1 and 10;
```

words which are ignored by most search engines. We follow the theory of TF-IDF, which considers the term frequency against document frequency of words, and ignores the sequence of word appearance in the content for simplicity. Accordingly, we apply ACCUMulate operator (,) to join words in unstructured text featured content into a query against CONTEXT full text search in Oracle. The following shows an example of how the pre-query collects similarity data objects of one of the user's data object. Suppose the user has the access authority of the following document.

```
Markov chain Monte Carlo (MCMC) methods are an important
algorithmic device in a variety of fields.
```

Then one example of the Oracle SQL statements which use pre-query to collect the top-10 similar document in the database is shown in Table 5.3.

Also the threshold statement example which returns the documents in the database with above

Table 5.4: CBAC Threshold Example

```
SELECT A_ID
FROM JOHNDOE.AWARD_BASIC
WHERE ID NOT IN
(SELECT A_ID FROM JOHNDOE.AW_INTR
WHERE I_ID=SYS_CONTEXT('USERENV','SESSION_USER'))
AND CONTAINS (ABS,
 '<query><textquery grammar="CONTEXT" lang="english">
 Markov chain Monte Carlo (MCMC)
 methods are an important algorithmic
 device in a variety of fields.
</textquery>
<score datatype>="float" algorithm="DEFAULT"/>
</query>',1)>=20;
```

20 similarity score is shown in Table 5.4.

As the instance of the "pre-queries" used to search against the database affects the parsing step directly and is also used to restrain the number of unique terms in TF-IDF. Knowing this, two optimizing strategies are extended in Chapter 6 and 7. Before extending to Chapter 6 and 7, A naive solution is raised to boost the efficiency by skipping the sort step with threshold method. The restrain to return the same amount of data objects is used to remove the bias of IO time costs. In the experiments, the thresholds of 10, 20 and 30 are tried separately, where the score is in the range from 0 to 100. In the result, another factor which affects the efficiency is revealed: the threshold. The threshold controls the overall returned similar instances. The lower the threshold is, the more the returned similar instances. The whole process of searching relevant instances in threshold method is implemented with a cursor in PL/SQL. Therefore, it costs more in opening a "large" cursor.

The database runs on a 64-bit Windows 7 system, with Intel® Core™ 2 Duo CPU E8500 @

3.16GHz and 4.0GB RAM. To mimic the real-world use case, queries are issued from SQL-Plus, and the query evaluation time includes all I/O (e.g. network I/O).

Execution. We mimic the scenario in Example 1. In initial authorization, the *base set* of each user is defined as the award records PI-ed by the user. Each user is explicitly granted access to such records. Next, we simulate the following access control scenarios: (R1) an attribute-based access control (ABAC) rule: the user is allowed to access records in a division where he/she has PI-ed an award; (R2) a content-based access control (CBAC) rule: the user is only allowed to access awards that have similar abstracts with the awards in his/her base set; and (R3) a combined (ABAC+CBAC) rule: R1 AND R2. All three scenarios are implemented with Oracle VPD to show CBAC is capable to either work with existing access control model or work independently.

In the experiments, we login as 60 randomly selected users to issue the following queries.

```
QUERY1: SELECT TITLE, ABS FROM johndoe.AWARD_BASIC
        WHERE S_DATE >= TO_DATE('1996/01/01', 'yyyy/mm/dd')
        AND   ROWNUM<=10;
QUERY2: SELECT COUNT(*) FROM johndoe.AWARD_BASIC
        WHERE S_DATE >= TO_DATE('1996/01/01', 'yyyy/mm/dd');
```

The end-to-end query evaluation time for ABAC (R1) is shown in Figure 5.1 and 5.2. For threshold-based CBAC, the end-to-end query evaluation time is shown in Figure 5.3 and 5.4. The end to end query evaluation of threshold based CBAC + ABAC is presented in Figure 5.5 and 5.6. Note that the rightmost bar in this group indicates the "w/o-CBAC" case, where no access control is enforced. We have performed the experiment with different thresholds – a larger threshold means a stricter constraint, which requires higher similarity between the queried records and the seeds. As shown, query processing for Query1 with CBAC is very efficient. A larger threshold leads to slower query processing, since Oracle needs to scan through more records to identify first 10 records that satisfy the stricter CBAC condition. Query evaluation slows down with R3, with the overhead required by both ABAC and CBAC semantics. On the other hand, Query2 forces

Oracle to go through all records. As shown in Figure 5.5 and 5.6, the overhead is acceptable, especially consider that CBAC models data content in a high-dimensional vector space, which requires excessive computation.

Top-K CBAC. We have developed two implementations of top-K CBAC.

Naive implementation. In a naive implementation, we simply included the top-K semantics in the dynamic predicate. Unfortunately, query performance was very slow, since the top-k ranking in VPD predicate was repeatedly evaluated.

Optimized implementation. To improve query performance, we split the top-K semantics into two steps: (1) in PL/SQL, we select the top K records that are most similar to the base set; (2) we identify the similarity score (T_s) of the K-th record, and generate a threshold-based predicate with threshold T_s. The average end-to-end query processing time is significantly reduced, comparing with the naive implementation. Note that query evaluation is still relatively slow comparing with threshold-based CBAC, mainly due to the size of the database (see Figure 5.7). Especially, Oracle does not provide native support for selecting first K records – Oracle sorts the entire table to return the top K records (complexity: $O(N \log N)$). However, as shown in Chapter 5, the computation of selecting and ranking top K records could be as low as $O(N + K \log K)$. In Chapter 6 and 7, we will optimize top-K CBAC performance using blocking and tagging.

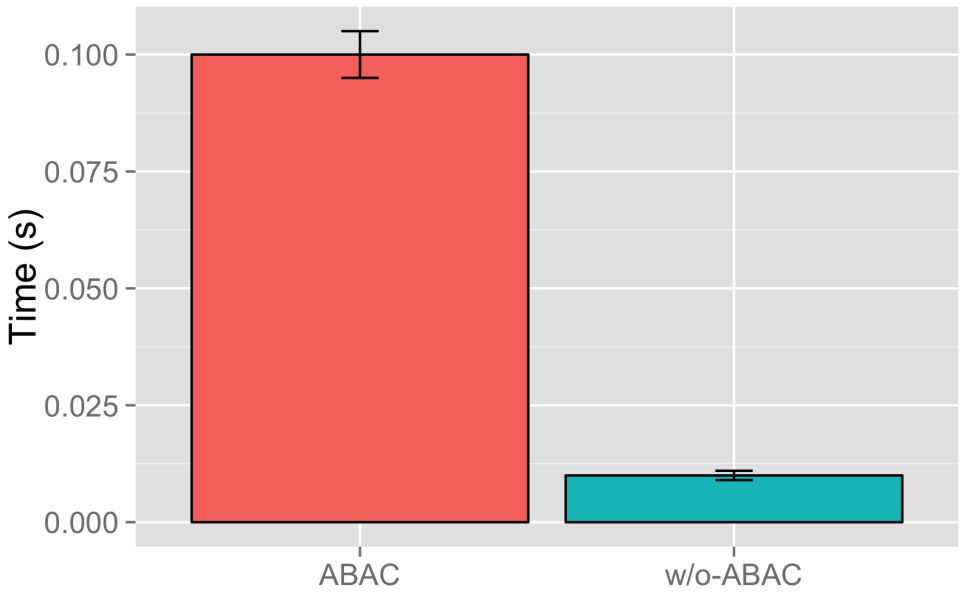

Figure 5.1: ABAC Efficiency with QUERY1

5.2 Offline CBAC

For databases which are not updated frequently, offline training is a much more efficient way to perform content-based access control model. In a naive offline training, the accessible range is calculated and hard coded in VPD policy for every user. Here we adopted unsupervised nearest neighbor algorithm for ranking out the top-K documents. In Section 5.2.1, three different strategies, namely, brute force, kd-tree and ball tree algorithms, of nearest neighbor algorithm are introduced. As shown (Kumar et al. (2008)), ball tree in k nearest neighbor search yields excellent result in both the sense of accuracy and efficiency.

5.2.1 Unsupervised Nearest Neighbor Offline Training

5.2.1.1 Brute Force Algorithm

Brute-force search, also known as exhaustive search, is the basic problem-solving technique for nearest neighbor algorithm, which enumerates all possible candidates for the solution and validates

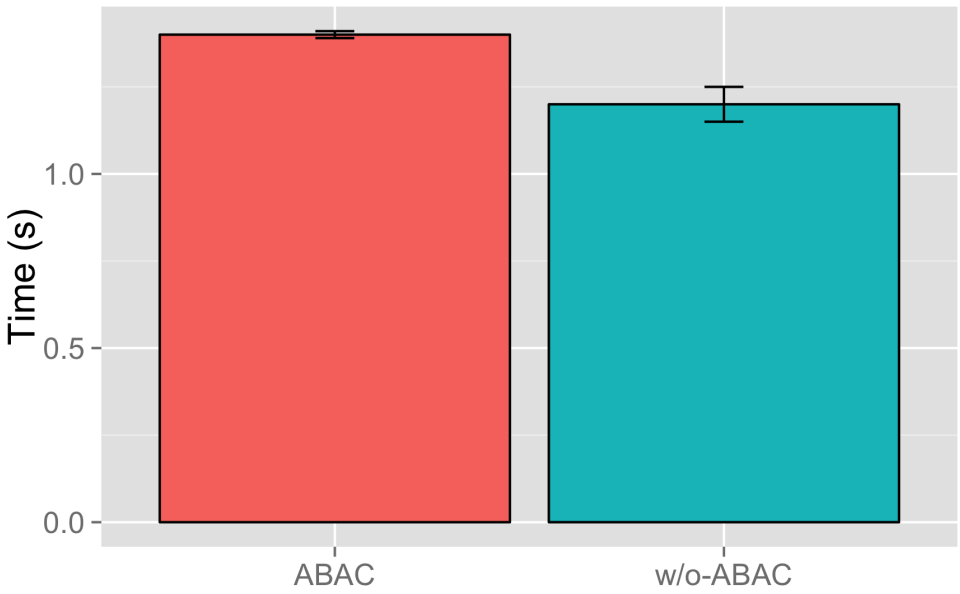

Figure 5.2: ABAC Efficiency with QUERY2

whether every candidate satisfies the problem criteria.

The algorithm for nearest neighbor for ranking iterates in the following way:

As shown above, brute-force algorithm is fairly simple to implement, and will converge to its solution as always if there is one. However, the cost of it is proportional to $O(|X| \times |Z|)$, which will grow fast for many realistic problems. Therefore, brute-force algorithm is recommended for problems with limited size, or the accuracy is emphasized over speed. Other than that, brute-force is usually used as a baseline for benchmarking other algorithms or heuristics.

5.2.1.2 K-D tree

K-D tree follows the pinciple of binary search (Bentley (1975)). It uses a binary tree to store the training data X, where every node in the binary tree is a k-dimensional point. Every non-leaf point generates a hyper-plane to separate the interested feature space into half. Points to the left of the hyper-plane represent the left subtree of the node; while points to the right of the hyper-plane represent the right subtree of the node. Then testing data Z follows Algorithm 5 to get the closest

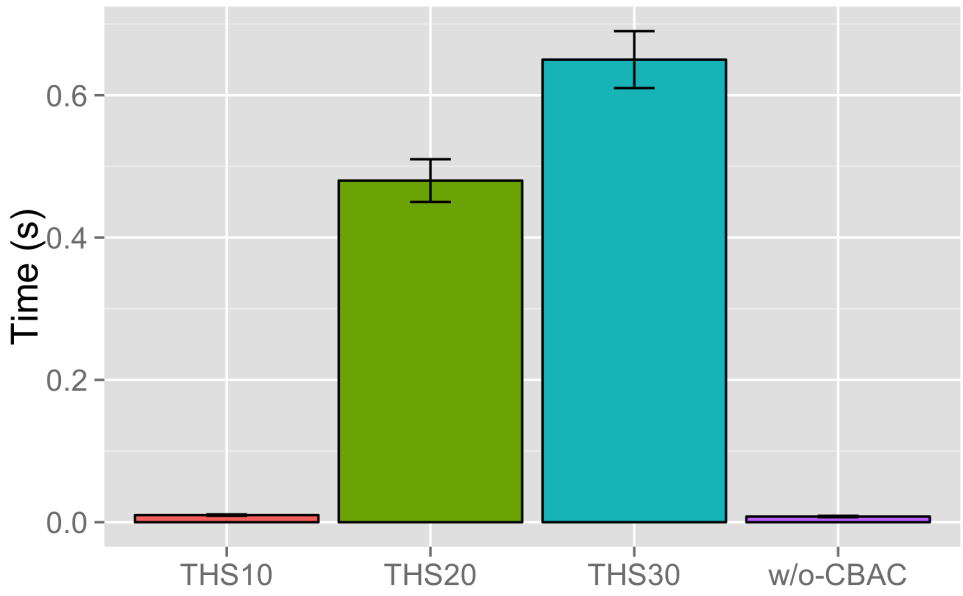

Figure 5.3: Threshold CBAC Efficiency with QUERY1

point from X for every $\mathbf{z}_i \in Z$. Figure 5.8 and 5.9 illustrate a 2-D K-D tree example. K-D tree guarantees $O(n \log n)$ complexity.

5.2.1.3 Ball Tree Algorithm

Ball tree algorithm also know as metric tree algorithm is a variation from K-D tree (Omohundro (1989); Uhlmann (1991)). Instead of using hyper planes, ball tree algorithm utilizes hyper spheres to split the space (sub-space). Algorithm 6 explains the construction of ball tree. Algorithm 7 explains how to search and get k nearest neighbors, given a testing data $\mathbf{z}$. Given a testing data set Z, Algorithm 7 is easy to iterate all points in Z. With the improvement of dimension selection for splits, ball tree is considered as a better choice compared to K-D tree.

Algorithm 3 Brute Force Algorithm

Require: Training Data X, and Testing Data Z, and k.

1: Initialize Y with an empty matrix.
2: **for all** $\mathbf{z}_i \in Z$ **do**
3: **for all** $\mathbf{x}_j \in X$ **do**
4: Compute the distance: $d(\mathbf{z}_i, \mathbf{x}_i)$
5: **end for**
6: Rank the $d(\mathbf{z}_i, \cdot)$ acsendingly.
7: Append the indices of X with top-k $d(\mathbf{z}_i, \cdot)$ as a column to the right of Y.
8: **end for**
9: **return** Y

Algorithm 4 Calculate $T = kdtree(X, d = 0)$

Require: $d \geq 0$

Ensure: X is an $n \times k$ matrix.

Ensure: $T = kdtree(X, d)$.

1: $axis \leftarrow d \ mod \ k$
2: Sort points according to *axis* and choose median as pivot element.
3: Initialize an empty tree node N.
4: $N.value \leftarrow median$
5: $X.left \leftarrow kdtree($ points in X before median, $d+1)$
6: $X.right \leftarrow kdtree($ points in X after median, $d+1)$
7: $N.leftChild \leftarrow kdtree(X.left, d+1)$
8: $N.rightChild \leftarrow kdtree(X.right, d+1)$
9: **return** N

Algorithm 5 Seach in K-D Tree

Require: Binary Tree T, and Testing Data Z, and k.

Initialize Y with an empty matrix.
for all $\mathbf{z}_i \in Z$ **do**
  Iniitialize $\mathbf{y}$ with an empty vector.
  $B \leftarrow T$
  for all $i \in \{1, \cdots, k\}$ **do**
    Search $\mathbf{z}_i$ against B basing on the split dimension.
    Once reach a leaf node $\mathbf{b}_j$, store the node as current best $b_c = b_j$.
    Calculate $d(\mathbf{z}_i, \mathbf{b}_j)$ to be the radius of a hyper sphere centered at $\mathbf{z}_i$.
    if The hyper sphere has an intersection with the existing hyper planes **then**
      Compute the nodes represented in the intersected hyper sub-spaces to get the closest point b_c.
    end if
    Add b_c in $\mathbf{y}$.
    Delete t_c in T
  end for
  Append $\mathbf{y}$ as a column to the right of Y.
end for
return Y.

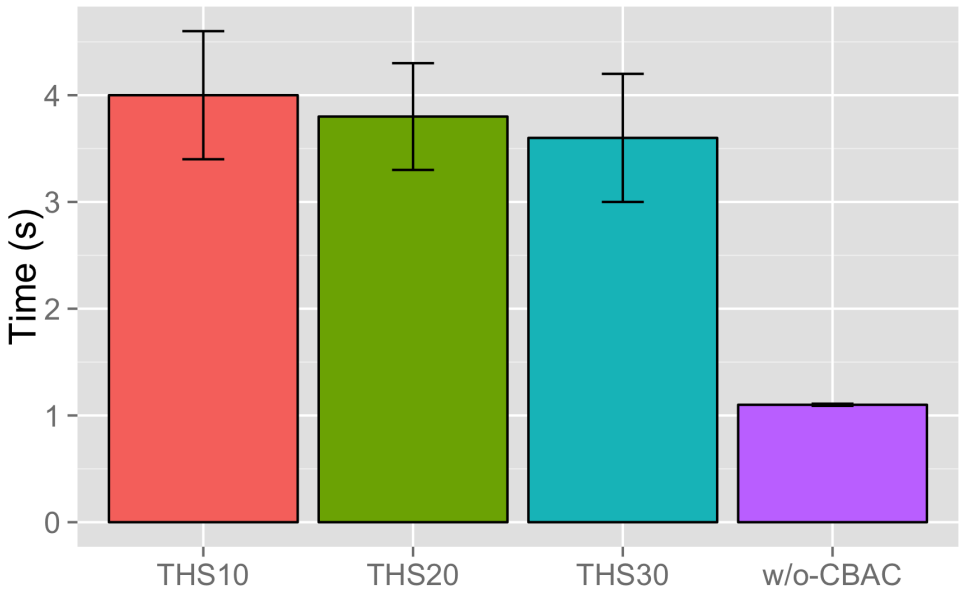

Figure 5.4: Threshold CBAC Efficiency with QUERY2

5.2.2 Experiments

In the experiment, we utilized *k*-nn with ball tree algorithm. The naive offline CBAC efficiency has been tested against the baseline where there is no constraint of accessible range of users. That is to say the baseline is the situation where no access control policy is applied to the entire database and users can query against the entire database without constraints. In the test, we force the baseline returning the same amount of instances as CBAC does for every user. Figure 5.10 shows that the offline CBAC model performs nearly the same with queries without any content-based access control policy.

The other option for offline similarity assessment is to build up a table which is used to store the pair-wise similarity of instances all over the database. The table holds information of the user's ID, the instance's ID which belongs to him or her, the other instance's ID, and the similarity score. This option is suitable for small and medium size of databases. An advantage for this option is the databases can be updated frequently. The only thing needs to be changed is the scores associated with the updated data objects. The overall advantage for the offline similarity assessment is it

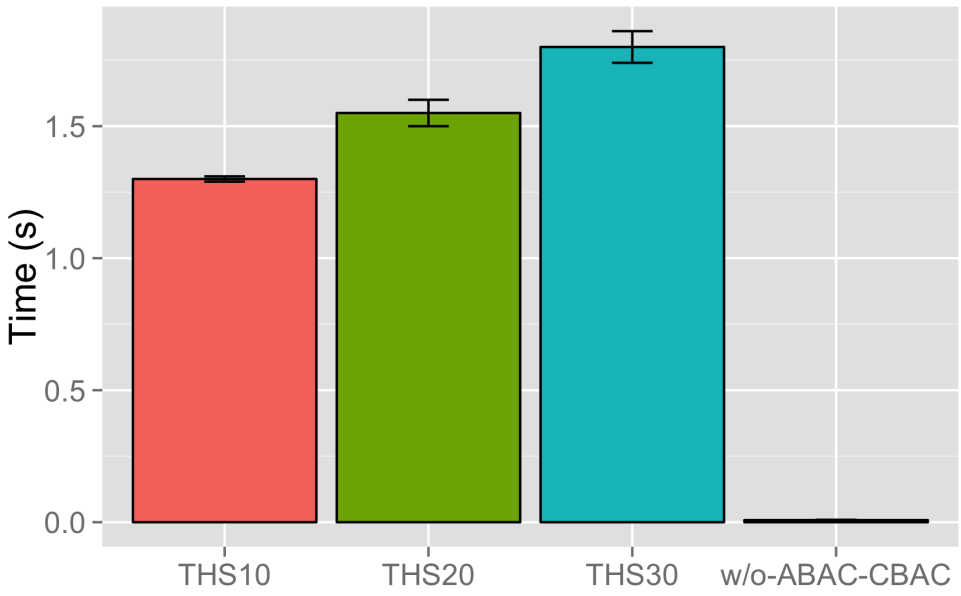

Figure 5.5: Threshold CBAC + ABAC Efficiency with QUERY1

makes indexing the scores possible for the pair-wise similarity. This will further boost the sorting efficiency of CBAC model and other time efficiency even for thresholding.

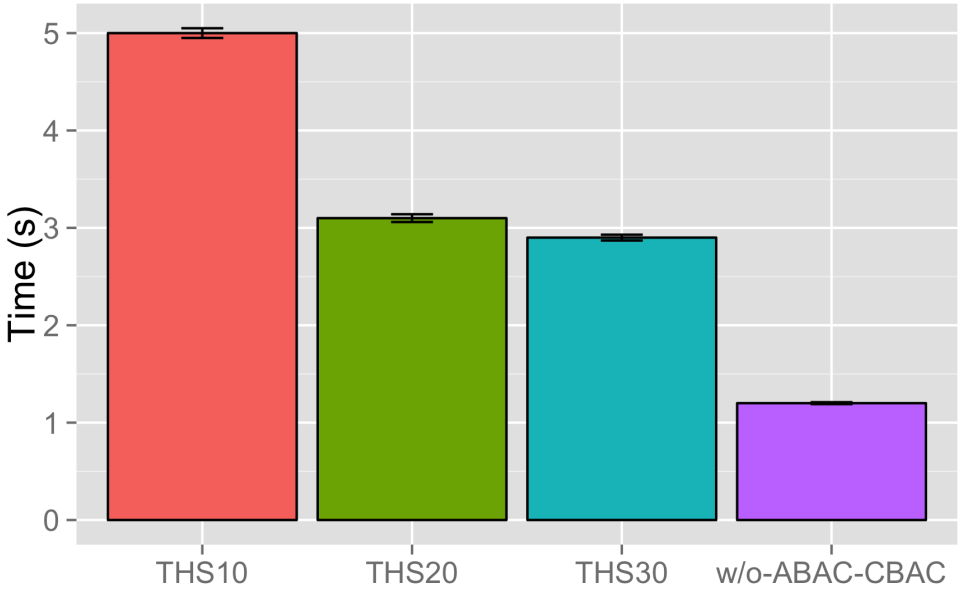

Figure 5.6: Threshold CBAC + ABAC Efficiency with QUERY2

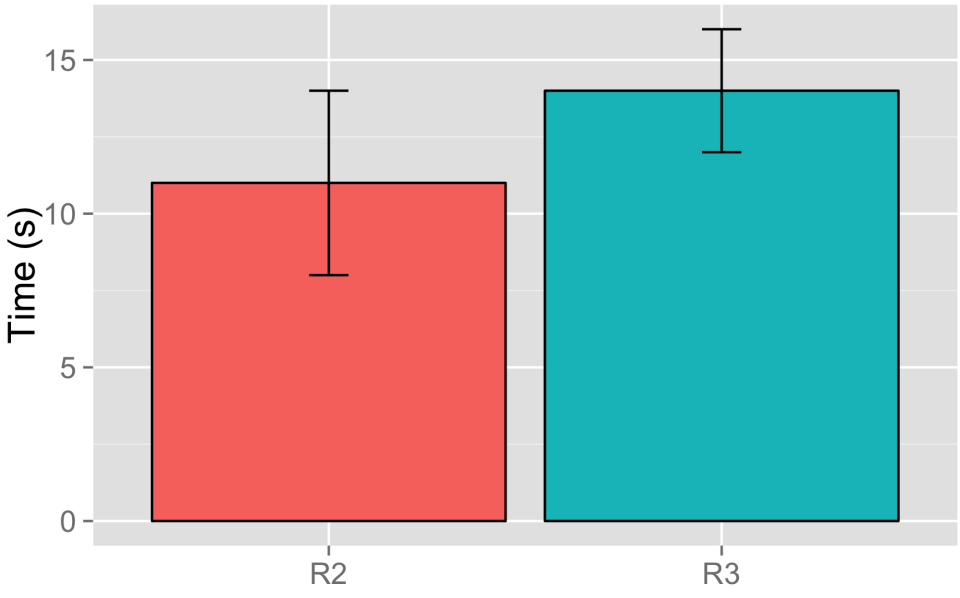

Figure 5.7: Top-10 CBAC Efficiency

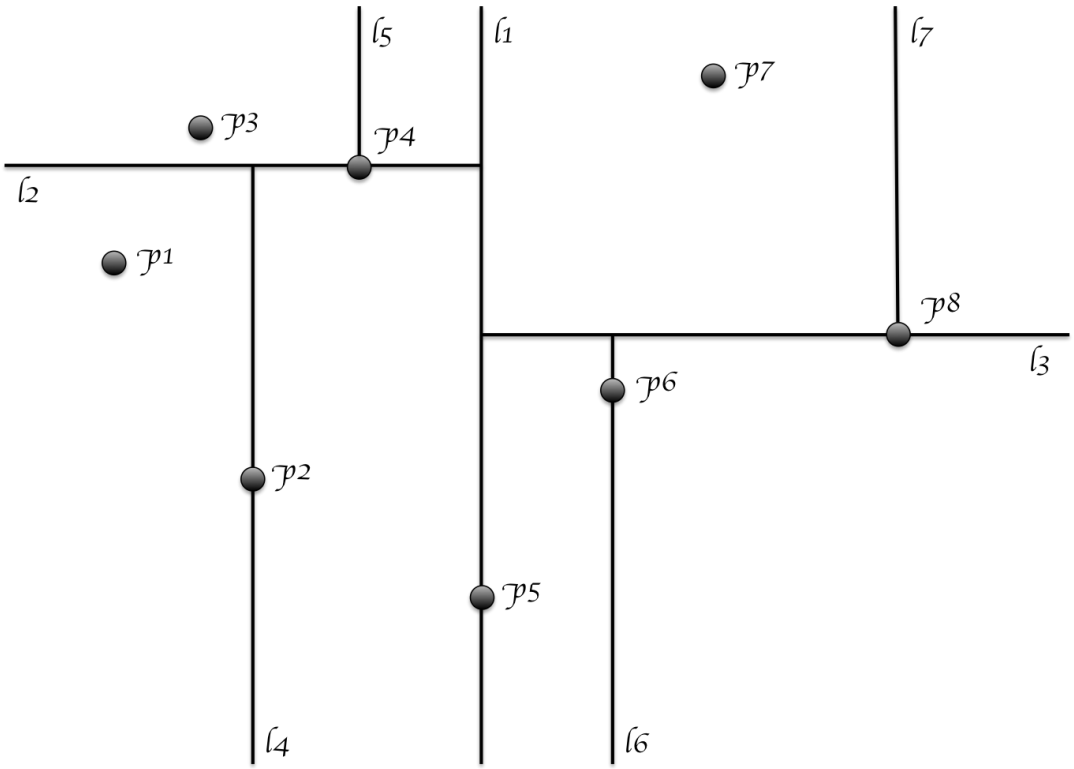

Figure 5.8: 2-D K-D Tree Subspace Splits

Algorithm 6 Calculate $T = balltree(X)$

Ensure: X is an $n \times k$ matrix.

Ensure: $T = kdtree(X, d)$.

1: $c \leftarrow$ the dimension of greatest spread
2: Sort points according to *axis* and choose median as pivot element.
3: Initialize an empty tree node N.
4: $N.value \leftarrow median$
5: $X.left \leftarrow kdtree($ points in X before median, $d+1)$
6: $X.right \leftarrow kdtree($ points in X after median, $d+1)$
7: $N.leftChild \leftarrow balltree(X.left)$
8: $N.rightChild \leftarrow balltree(X.right)$
9: **return** N

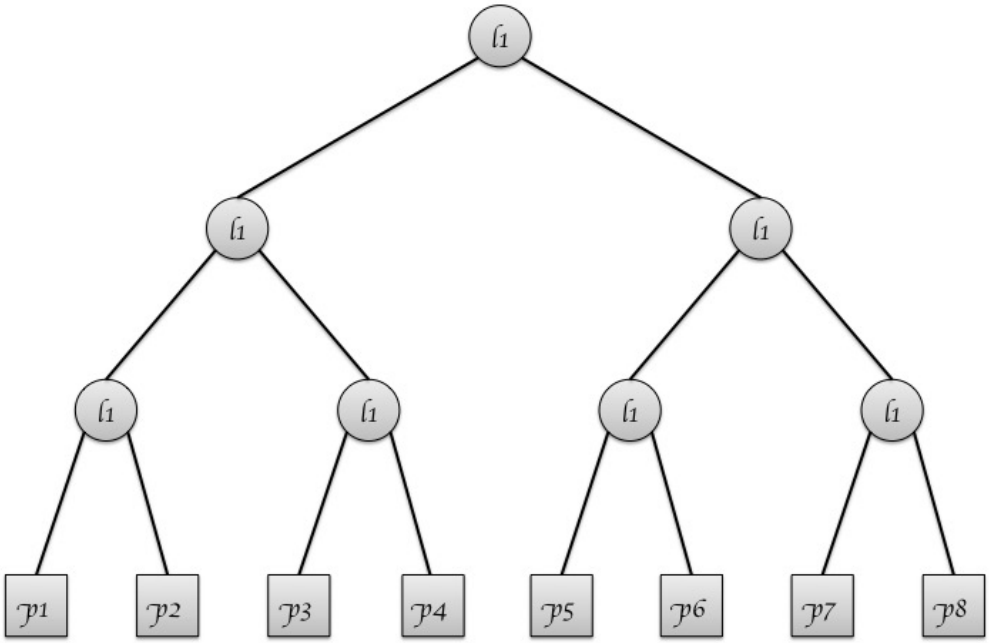

Figure 5.9: K-D Tree Example

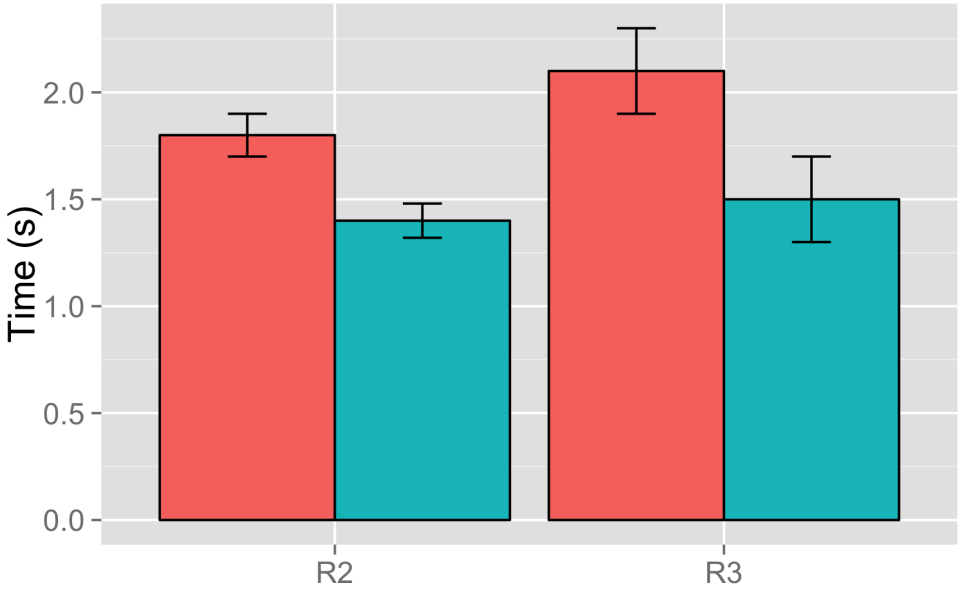

Figure 5.10: Offline Efficiency

Algorithm 7 $knn\_search = (\mathbf{z}, k, Q, T)$

Ensure: $\mathbf{z}$ is a d dimensional vector.

Ensure: k is the number of nearest neighbor of $\mathbf{z}$ to search for.

Ensure: Q is max-first queue with length as k.

Ensure: T be the constructed ball tree

if $distance(\mathbf{z}, T.pivot) \geq distance(\mathbf{z}, Q.first)$ **then**
  return Q unchanged
else if T is a leaf node **then**
  for all node p in T **do**
    if $distance(\mathbf{z}, p) < distance(\mathbf{z}, Q.first)$ **then**
      Add p to Q
      if $size(Q) > k$ **then**
        Remove the furthest neighbor from Q
      end if
    end if
  end for
else
  Let $T.child1$ be the child node closest to $\mathbf{z}$.
  Let $T.child2$ be the child node furthest from $\mathbf{z}$.
  $knn\_search = (\mathbf{z}, k, Q, T.child1)$
  $knn\_search = (\mathbf{z}, k, Q, T.child2)$
end if

Chapter 6

CBAC Optimizing Strategies

In this chapter, we discuss two optimization approaches in enforcing content-based access control. First, we aim to improve the efficiency of the approach, especially when similarity assessment is performed on-the-fly using blocking/clustering technique. Next, we also try to improve the accuracy of content similarity assessment, especially for short text content, where term-distribution based approaches would fail.

6.1 Content-Based Blocking

In the previous chapter, we have shown that content-based access control could be efficiently enforced with offline similarity assessment. However, in some scenarios, the user credentials is consistently updated, or provided with the query. Hence, it is not possible to pre-compute similarities offline – the content similarity assessment needs to be performed on-the-fly for every record. To expedite query processing, we introduce the content-based blocking scheme.

When the similarity function $Sim_d(\cdot,\cdot)$ is provided with the access control policies, we can pre-partition the records into non-overlapping blocks, based on the content similarity of the records. That is, we pre-cluster the records into c clusters, so that records with similar contents are labeled in the same cluster: $Sim_d(\mathbf{d}_i,\mathbf{d}_i)$ is large when $\mathbf{d}_i$ and $\mathbf{d}_j$ belong to the same cluster; while $Sim_d(\mathbf{d}_i,\mathbf{d}_i)$ is small when $\mathbf{d}_i$ and $\mathbf{d}_j$ belong to different clusters. The centroid of cluster C_k is defined as the

average of the documents in the class:

$$\mu(C_k) = \frac{1}{|C_k|} \sum_{d_i \in C_k} \mathbf{d}_i$$

where $|C_k|$ denotes the number of documents in the cluster. The centroid vectors of all clusters are stored separately. Note that the centroid vectors are not human comprehensible. Alternatively, it is also possible to store a document that is closest to the centroid, so that administrators could easily estimate the content of the clusters. On average, each cluster will have $\frac{N}{c}$ documents. However, most of the clustering approaches does not guarantee that cluster sizes are balanced.

During query processing, each incoming query is first compared with the cluster centroids, to identify the most similar x clusters, where $x << c$. In practice, x is usually a very small number, which is 1% of c or smaller. Next, the query is only evaluated against the records in the selected x clusters. That is, a predicate is added to the query that requires the records to have one of the x labels. In this way, in content-based access control enforcement, similarity assessment is only performed between $\mathbf{u}$ and the records from x most similar clusters.

The assumption is that: when $Sim_d(\mathbf{d}_i, \mathbf{d}_j)$ is high,
$Sim_d(\mathbf{u}, \mathbf{d}_i)$ and $Sim_d(\mathbf{u}, \mathbf{d}_j)$ are expected to be similar. In fact, for vector space model and Euclidian distance, we always have: $|\mathbf{u} - \mathbf{d}_j| < |\mathbf{u} - \mathbf{d}_i| + |\mathbf{d}_j - \mathbf{d}_i|$. Therefore, the clusters that are most similar to $\mathbf{u}$ are more likely to contain records that are most similar to $\mathbf{u}$. For the same reason, we can eliminate clusters that are different from $\mathbf{u}$, since their records are supposed to be different from $\mathbf{u}$.

Any vector space clustering method could be employed in this approach. Since clustering is computed offline, the computation is not a big concern. On the other hand, there are clustering approaches that allows one item to belong to multiple clusters: *overlapping clustering*. When overlapping clustering is employed, each record will be attached with multiple labels. Moreover, during query processing, a smaller value could be picked for x, e.g. 3.

6.1.1 Naive k-means Clustering

We utilize the conception of clustering from machine learning to group similar documents into clusters. When comparing the query and original access range of the session user against the entire database, only related clusters will be considered as result candidates, and unrelated clusters will be blocked.

k-means clustering is a method to partition the feature space in k voronoi diagrams (Aurenhammer (1991)). Algorithm 8 illustrates the process of basic k-means clustering. It starts with random selection of k points/centroids in training data X as seeds, and iterates assignments of points of X to k centroids of clusters and calculation of new k centroids based on the recent assignments. k-means although is very intuitive, it cannot guarantee the convergence of global optimum (Vattani (2011)) and it is very expensive at the cost of $O(n^{dk+1}\log n)$, where n is the number of entities to be clustered and d is the feature dimensionality (Inaba et al. (1994)). The next two subsections introduce the methodology of careful selection of k centroids, and scaled algorithm for k-means clustering.

Algorithm 8 Basic k-means Clustering

Require: The training data set X, and the cluster number k

1: Randomly select k points in X to be the initial k centroids of clusters: C
2: Calculate each point in X and assign it to the nearest centroid in C to get the clustering partition V.
3: Re-calculate the new k centroids of clusters from the means of each partition in V as C'.
4: **while** $C' \neq C$ **do**
5: $C \leftarrow C'$
6: Calculate each point in X and assign it to the nearest centroid in C to get the clustering partition V.
7: Re-calculate the new k centroids of clusters from the means of each partition in V as C'.
8: **end while**
9: **return** C and V

6.1.2 The Advantage of Careful Seeding: k-means++

In the previous subsection, we have introduced k-means clustering and its two major shortcoming. In this subsection, we would like to address the first one: its incapability to guarantee the convergence of global optimum. The key to the disadvantage of k-means of local convergence is due to the arbitrary selection of the initial k centroids. k-means++ was introduced to initialize a careful seeding of the initial k centroids by finding the most representative $k-1$ centroids given one initial arbitrary selection of one centroid (Arthur & Vassilvitskii (2007)). Algorithm 9 illustrates the process of the part of careful seeding. A standard k-means clustering is following the seeding procedure. Note $D(x)$ denote the shortest distance from a data point to the closest cluster.

Algorithm 9 The Seeding Algorithm of k-means++

Require: The training data set X, and the cluster number k

1: Randomly select a point c_1 in X to be one initial centroid of clusters and add it in C.
2: **for** $|C| < k$ **do**
3: Take a new centroid c_i, bu choosing $x \in X$ with probability $\frac{D(x)^2}{\sum_{x \in X} D(x)^2}$
4: Add c_i in C.
5: **end for**
6: **return** C

6.1.3 Scaled k-means++ with mini-batch Strategy

k-means++ provides a heuristic method to seed k centroids as shown in the previous subsection; however, k-means++ is still very expensive to implement for large scale data with high dimensionality. In this subsection, we introduce mini-batch strategy for k-means clustering to improve the scalability of k-means. Mini-batch strategy utilizes a random sampling with uniform distribution to generate a small scale of "training data set batch". With the assumption of random sampling with uniform distribution, the algorithm views the mini-batch proportionally reflects the structure of true training data set and by using stochastic gradient descent (SGD), it assumes mini-batch training will converge to the most possible k voronoi diagrams given the mini-batch b is generated

from X. As a golden standard, model can just be as good as input data, a mini-batch solution is an approximation to the solution with the entire training data set.

Algorithm 10 Mini-batch k-means

Require: The training data set X, the cluster number k, mini-batch size b, and iterations t.

```
Select k centroids from X based either on randomization or on seeding.
v ← 0
for i ∈ {1, ··· , t} do
    Randomly select b points in X, and M ← b.
    for x ∈ M do
        d[x ← f(C, x)]
    end for
    for x ∈ M do
        c ← d[x]
        v[c] ← v[c] + 1
        η ← 1/v[c]
        c ← (1 − η)c + ηx
    end for
end for
return C
```

6.1.4 Experiments

In the experiments, we perform offline clustering training on the entire repository. For document clustering, *k-means* clustering is more suitable for vector-space training, as the features in the vector has the same physical conception, so no weighting schemes is needed for *-means* clustering, and sphere distance or cosine similarity, compared to other distance or similarity function is much more proper. Here we use non-overlapping clustering, i.e., each record only belong to one cluster according to the top similarity score between each centroid and the record itself. Non-overlapping performs exclusive clustering compared to mixture modelings.

k-means clustering, although has been judged for its lacking of capability of inferring k in ad-

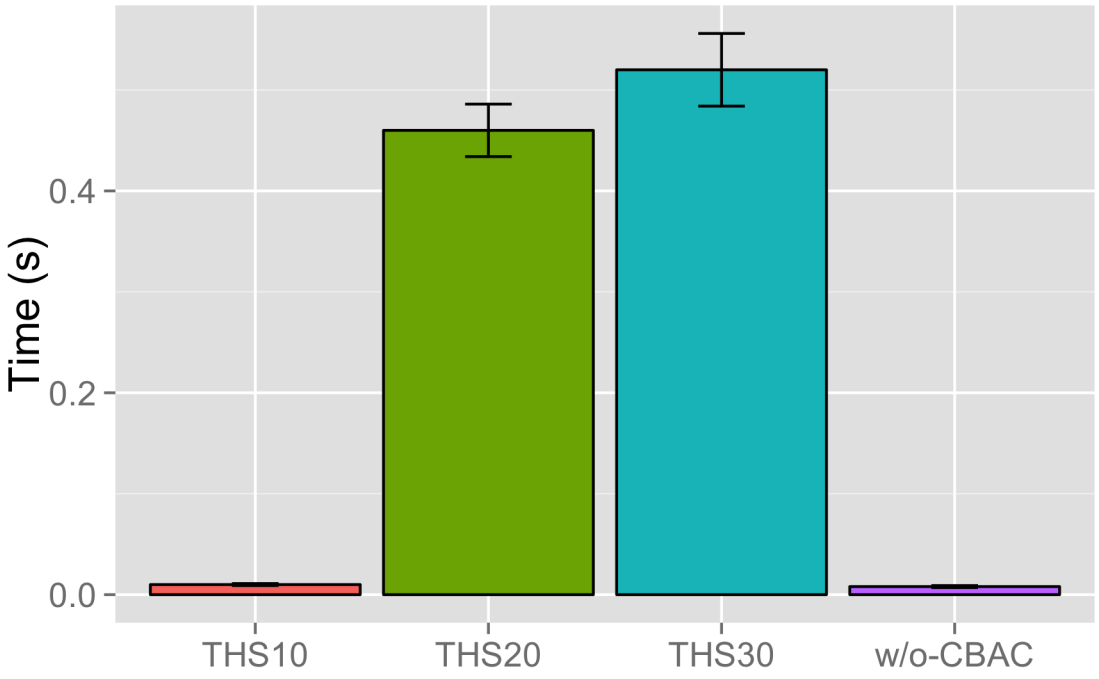

Figure 6.1: Threshold CBAC + Blocking Efficiency with QUERY1

vance, it is simple and easy to implement and generalize in a broad range of application, including document clustering, and multimedia clustering. Per the setting of k, we refer to one simple rule of thumb (Mardia et al. (1979)) that to partition n data objects into k clusters, $k = \sqrt{\frac{n}{2}}$. Concretely, in the experiments, $k = 227$. Figure 6.1, 6.2, 6.3, 6.4, and 6.5 show the results of blocking, which boosts the efficiency in the sense of restraining the searching size of the entire database.

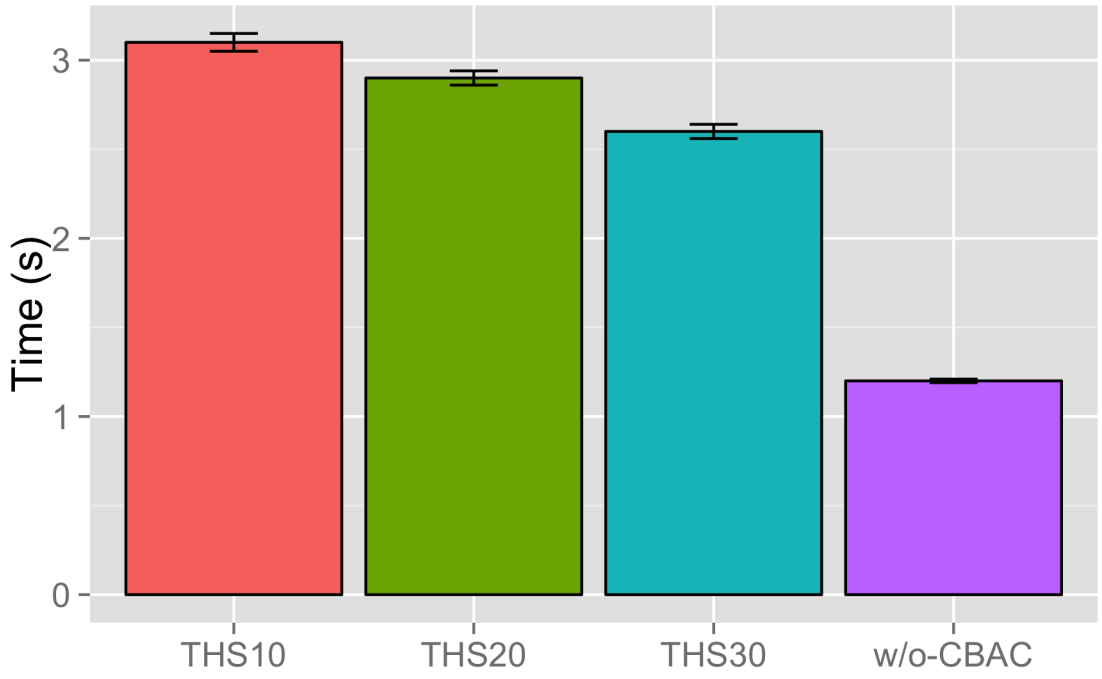

Figure 6.2: Threshold CBAC + Blocking Efficiency with QUERY2

6.2 Content-based labeling

6.2.1 Document labeling

As we have introduced in Chapter 5, the content-based access control model could take any attribute-similarity measurement. So far, we have heavily employed the vector space model to assess document similarity based on distribution of terms. While this model is the most popular method in information retrieval applications, it suffers some drawbacks since it relies on the bag-of-words model, not the meanings behind the terms. More precisely, the bag-of-words model cannot reflect the semantic concepts beyond the word occurrences. This is a typical drawback targeted in bag-of-words model. It cannot bridge the semantic gap in understanding the true meaning in unstructured text. Nowadays, in particular, the vector space model is especially ineffective for short documents. For instance, if we have a table of Twitter style short text snippets, enforcing content-based access control with vector space model would be impractical. Since short text seg-

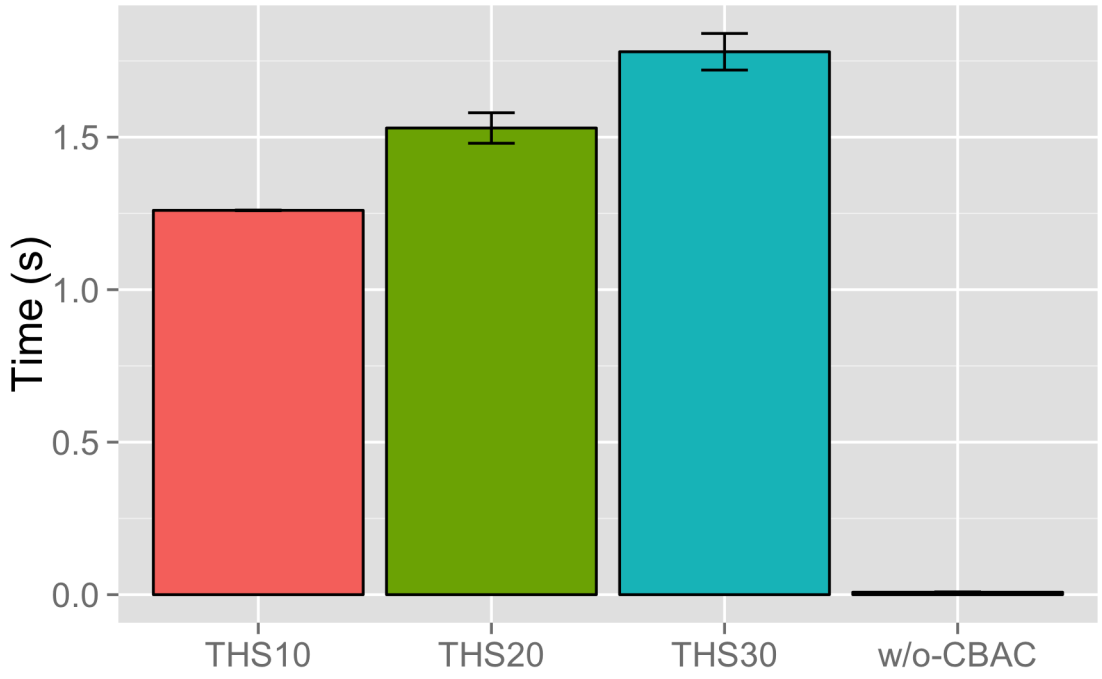

Figure 6.3: Threshold CBAC + ABAC + Blocking Efficiency with QUERY1

ments do not contain enough terms for term-distribution-based content modeling to be effective. As an example, let us look at the following two short documents:

```
D1: privacy preserving similarity assessment for
    semi-structured data


D2: private XML document matching
```

It is clear that `D1` and `D2` are both about the same topic. However, in vector space model, `D1` and `D2` are likely to be orthogonal (if "private" and "privacy" are considered to be different terms, as in most IR systems). Unfortunately, this type of short texts are heavily used in databases, and it is expected to effectively enforce content-based access control on such short text attributes.

To tackle this problem, we need to employ information understanding methods that works better than the term distribution based models. In particular, a group of annotation-based approaches

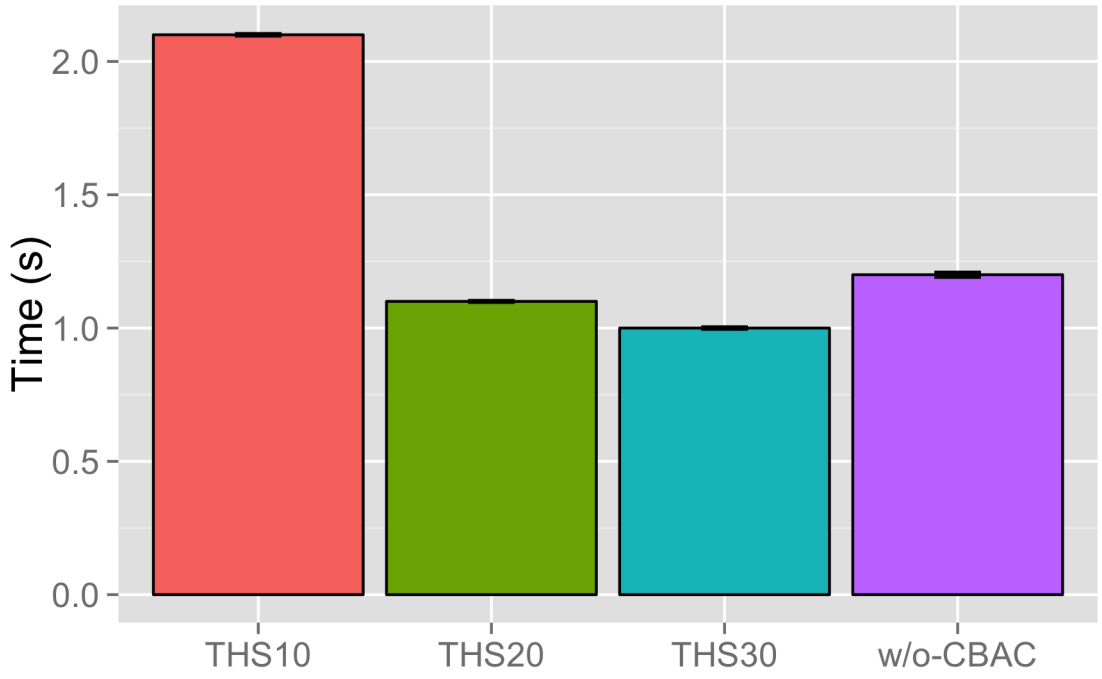

Figure 6.4: Threshold CBAC + ABAC + Blocking Efficiency with QUERY2

have been proposed in the information retrieval literature. The basic idea is to label short text documents with a set of pre-selected unambiguous terms (a.k.a. topics), so that documents are represented as a vector in the new unambiguous "topic space" instead of a "term space", and document-wise similarity could be measured in this topic space. The major concerns on these approaches are: (1) quality of the constructed topic space, and (2) accuracy of document tagging.

We employ both non-negative factorization, and TAGME (Ferragina & Scaiella (2010)) annotation to every abstract in the database to transform the content representation from bag-of-words to bag-of-topics. In our experiment, we utilized python scikit-learn package () to implement non-negative matrix factorization to extract different numbers of topics from NSF corpus. The top 10 words of 10 topics, 20 topics, 50 topics and 100 topics are shown in Table A.1, A.2, A.3, and A.4 in Appendix. We use RESTFUL API calls to exact TAGME annotation. TAGME majorly relies on the topics collected from Wikipedia, which reveals relatively high quality of the constructed topic. For the accuracy of document tagging, TAGME uses a weighted strategy to generate a soft

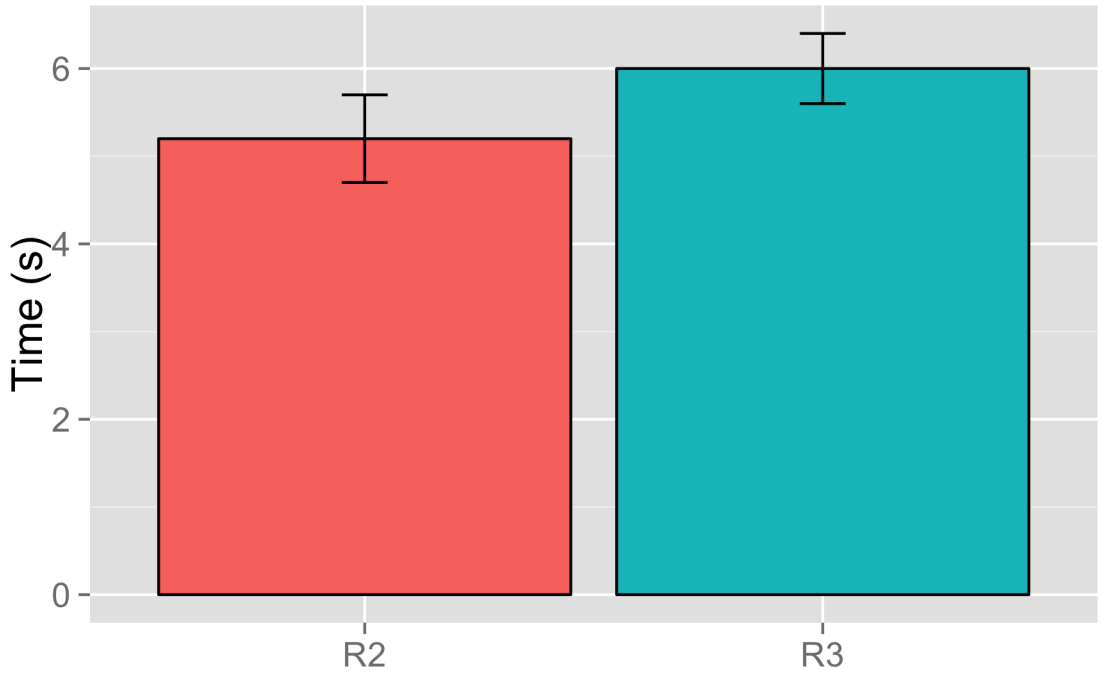

Figure 6.5: Top-10 CBAC + Blocking Efficiency

boundary of choices to the users. It introduced a parameter ρ to represent the accuracy or the goodness of the topic tagged onto the document. By examining these ρ's, a ceretain cut-off can be determined according to the usage of different applications.

6.2.2 Soundness of CBAC Enforcement

CBAC is said to be *sound*, when a CBAC enforcement mechanism makes access control decisions that are consistent with users' decisions. For instance, if we evaluate CBAC in Example 1, we would like to ask: "*when the CBAC enforcement mechanism allows Alice to access a record, will the supervisor agree with this decision?*" Practically, when CBAC enforcement allows users to access records that are semantically similar to the seed records in the base set, the enforcement mechanism is sound. Hence, the soundness of CBAC enforcement depends on the soundness of the content model and the similarity assessment. In practice, understanding semantic meanings

from unstructured text is a very difficult task. In this paper, we follow the notions in information retrieval community to evaluate the precision of the proposed approaches.

It is impractical to evaluate the relevance of a record against 100K records. Therefore, we attempt to evaluate the top-100 records identified by CBAC to assess if a DBA would agree with CBAC's access decision. In the experiments, we first use three rules to coarsely identify "relevant records", and manually examine the content of these records to make adjustments. We noticed that every record in NSF database is assigned with a set of *field identification numbers*. If two records share one or more field identification number(s), they are initially considered to be relevant. Besides, if the two records share two closely related field identification numbers, they are considered to be relevant. Last, if the seeds' content show a close relationship to the record's field name, they are considered to be relevant. Finally, we manually examine all the "relevant documents" identified by these rules, and eliminate the ones that appear to be irrelevant to us. We have tested queries from 60 different users in different disciplines including biology, chemistry, mechanical engineering, mathematics etc, and measures the precision of top-K results (e.g. $K = 10 - 100$) for all the queries. As shown in Figure 6.6, our CBAC enforcement is sound, as the user would agree with approximately 80% of CBAC's (positive) decisions, except for non-negative matrix factorization (NMF) tagging. The reason why NMF tagging does not improve the precision of top-K results is because NMF is actually an approximation of the base document repository, without other resources facility (e.g. Wikipedia). Although it provides "semantic" representation of text, it also compresses the information which has impact on the accuracy of similarity assessment. With the facility of Wikipedia annotation, TAGME approach is able to improve both query efficiency and CBAC accuracy. The efficiency results are shown in Figure 6.7, 6.8, 6.9, 6.10, and 6.11. The efficiency result for combining blocking and labeling in top-K scenario is shown in Figure 6.12.

We do not evaluate *recall*, which is another popular metric in information retrieval. With the amount of data, it is hard to manually identify all the relevant documents for each user. Meanwhile, with the *approximation* assumption for CBAC (Section 2), it is tolerable if a small fraction of relevant records are not accessible to the user, or a small number of irrelevant records are made

accessible.

Please note that the accuracy of the content similarity measurement is not a research problem in the security community. Rather, we are utilizing the methods from information retrieval and NLP communities. Any content modeling and similarity assessment method could be used in CBAC.

6.2.3 Experiments

In the experiments, when using TAGME, each topic is associated with an attribute called ρ in the range of 0 to 1, which reflects the quality of the annotation in the text of input. The overall topics annotated to the entire database is over 7 millions. That is 70 topic annotations per abstract. Considering the entire annotations is not necessary, so we use a threshold of ρ as a filter to pick out qualified topic annotation. In estimating the threshold, we fit all the collected ρ into a non-parametric distribution. Figure 6.13 shows the fitting of the density function. Figure 6.14 shows the fitting of cumulative probability. In Figure 6.14, it shows that by setting the threshold into 0.2, 80% topics are removed. Clearly, we have used the Pareto principle (80-20 rule) to determine our cutoff on the thresholded *rho*. Therefore, in the experiment, we use 0.2 as the cutline to filter topics. The similar filtering is also done with non-negative matrix factorization (NMF) annotation. In NMF, we start the filtering with NMF of 100 topics, and obtain the curve as Figure 6.15 and 6.16.

The filtered topics are then constructed with ACCUMulate operator (,) and add into a new predicate `TAG` in the table `AWARD_BASIC` of data type `CLOB`. A new CONTEXT indexing is added on the predicate. Thus, in the model here instead of calculating similarity upon words, the model calculates the similarity upon topics. The content labeling boosts the efficiency as shown in Table 6.6 in the sense of distilling the contents into the most relevant topics. In general cases, content-based labeling makes the content much shorter compared to bag-of-words model and also it removes ambiguities caused by word appearances.

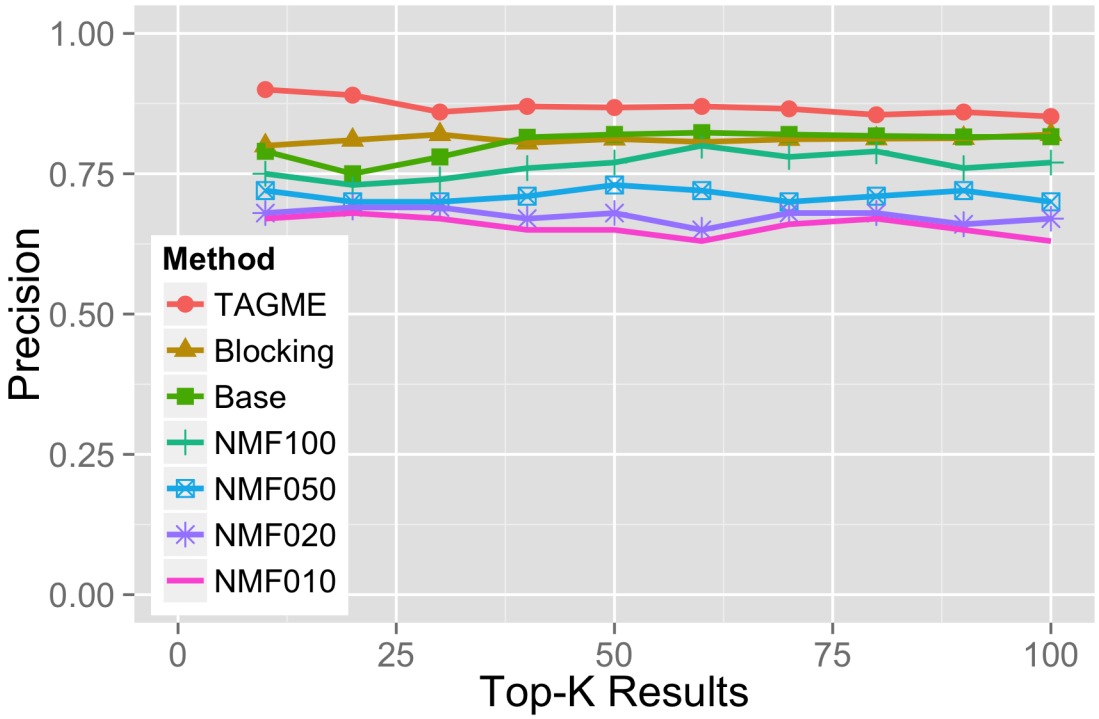

Figure 6.6: Soundness of CBAC Enforcement

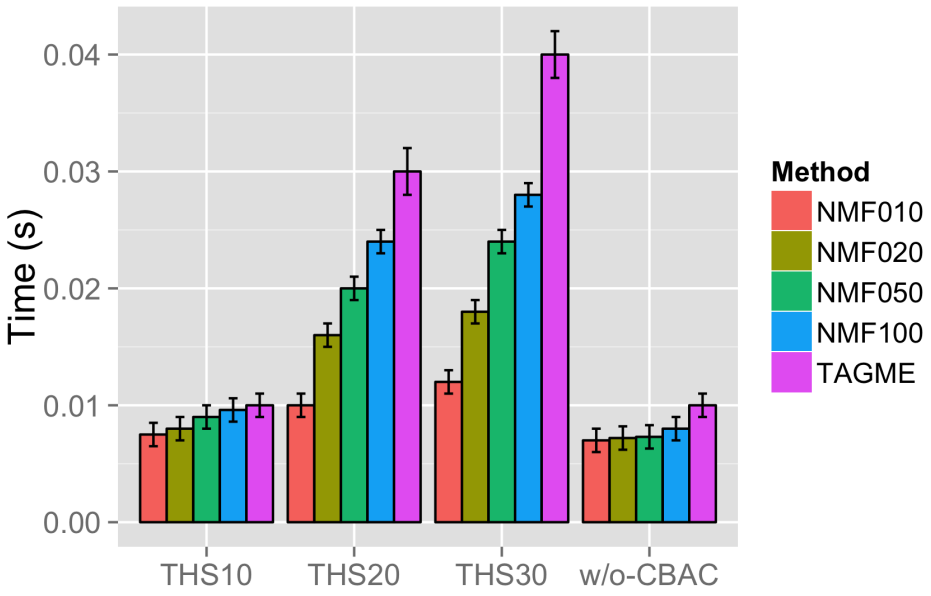

Figure 6.7: Threshold CBAC + Labeling Efficiency with QUERY1

Supervised labeling which is in scenario similar to TAGME usually provides better tagging on documents; however, the labels are usually imbalanced, which means only a small fraction of the document repository could be labeled given a pre-defined label. Due to the reason, in the next chapter, a supervised labeling method for multi-label classification is raised. We try to expand the application field, so that our focus includes document tagging, but not limited to it.

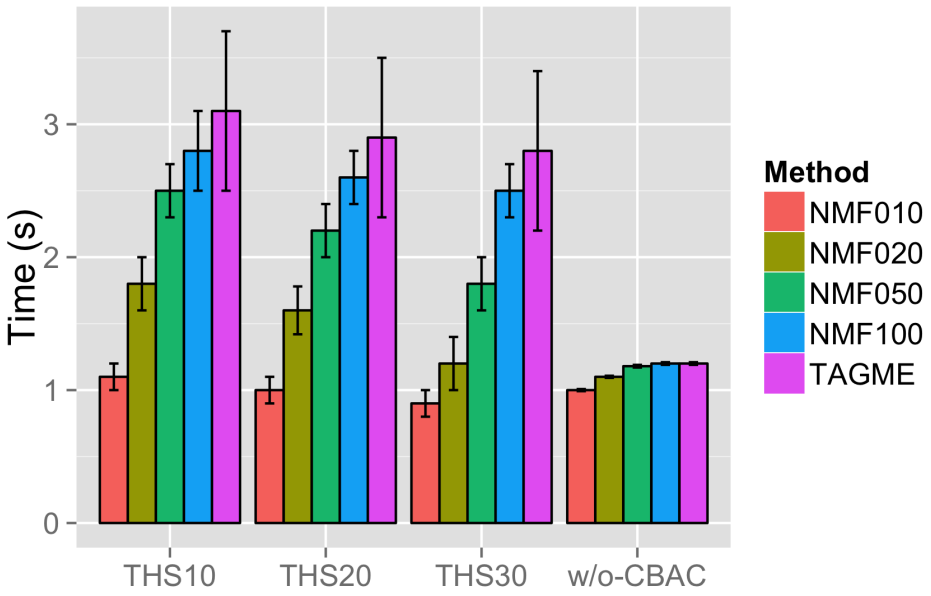

Figure 6.8: Threshold CBAC + Labeling Efficiency with QUERY2

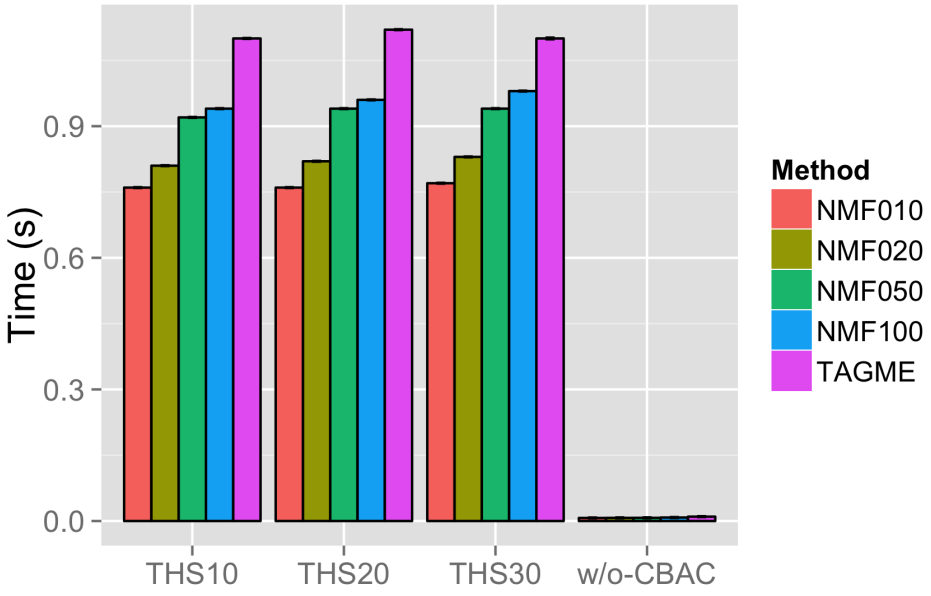

Figure 6.9: Threshold CBAC + ABAC + Labeling Efficiency with QUERY1

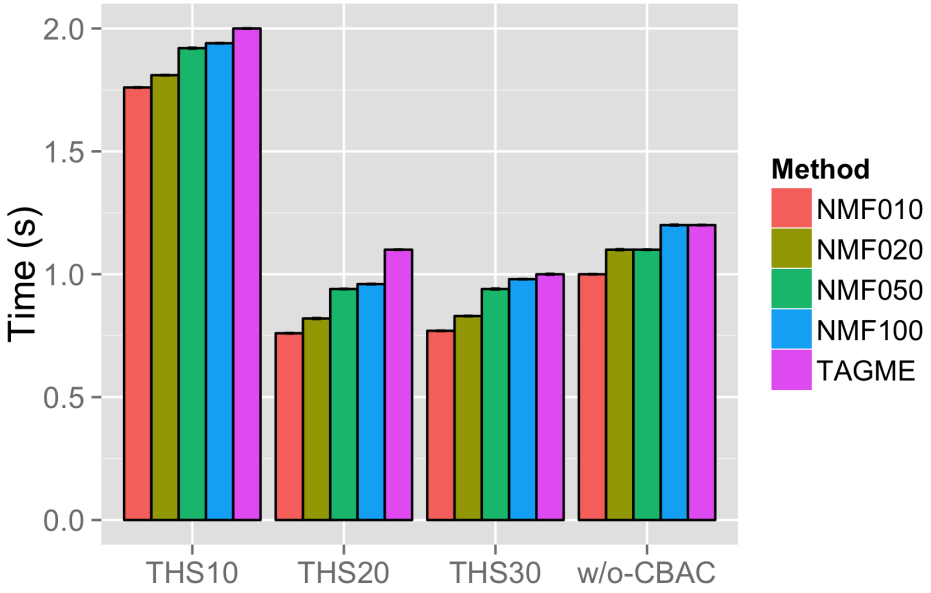

Figure 6.10: Threshold CBAC + ABAC + Labeling Efficiency with QUERY2

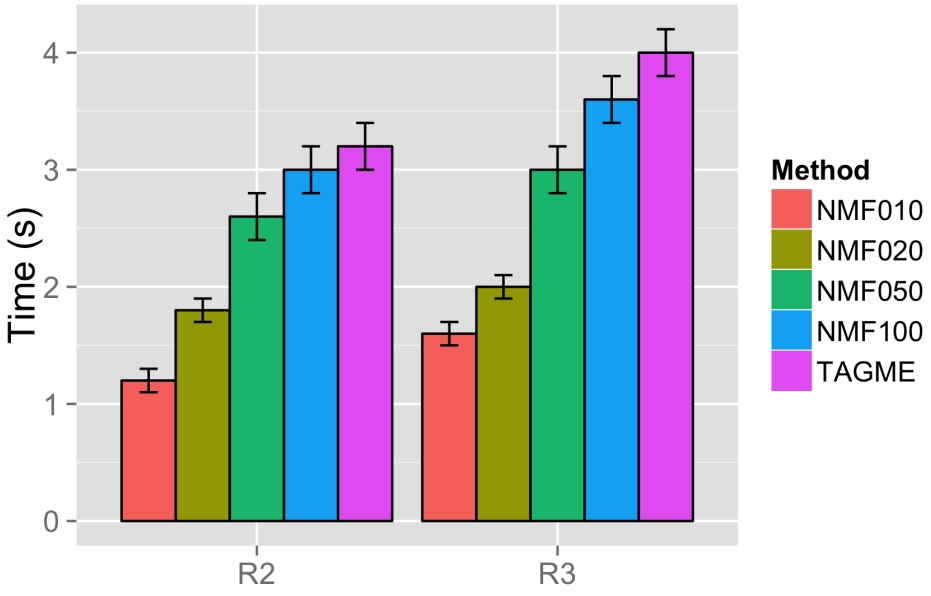

Figure 6.11: Top-10 CBAC + Labeling Efficiency

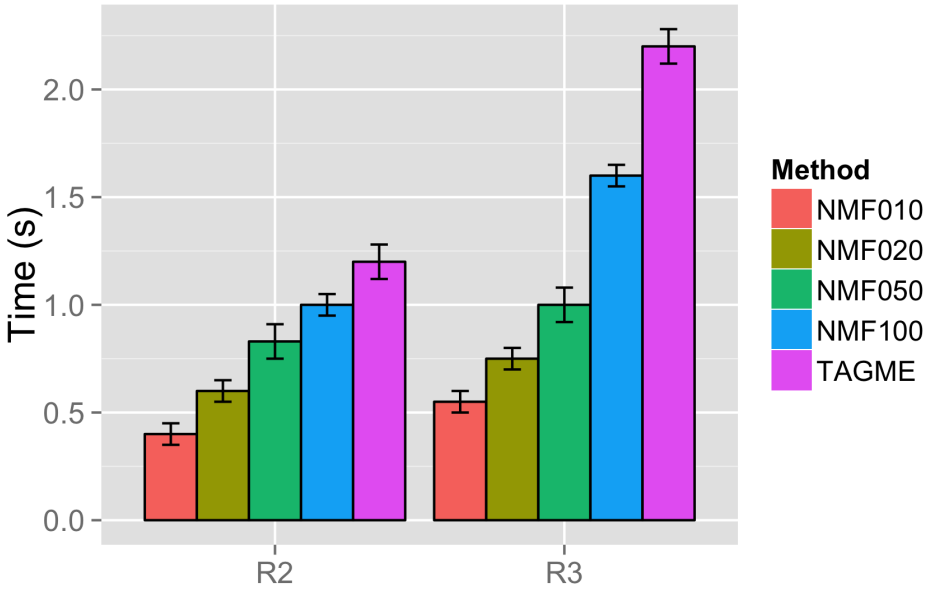

Figure 6.12: Top-10 CBAC + Blocking + Labeling Efficiency

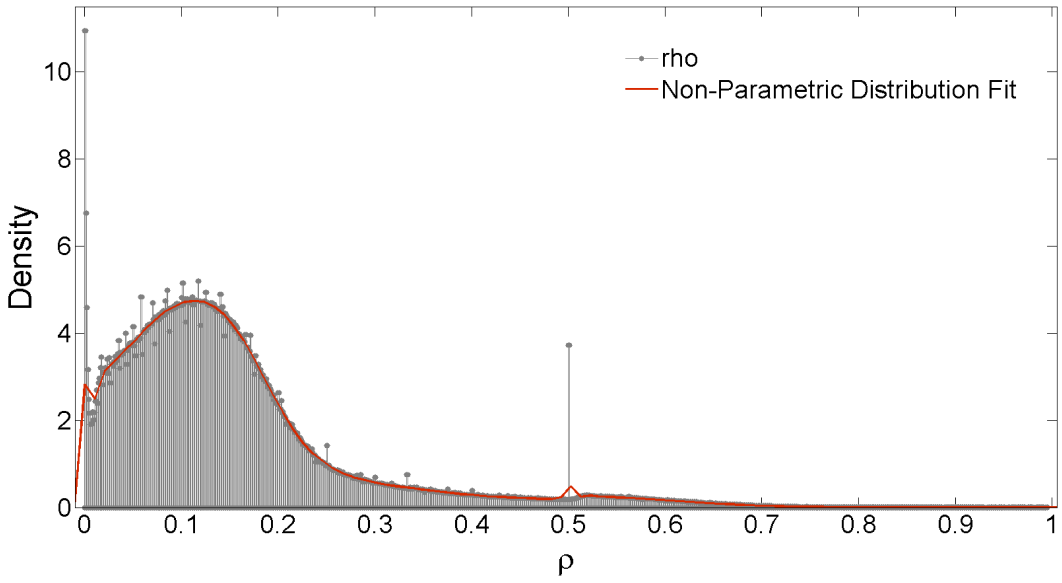

Figure 6.13: Density Fit

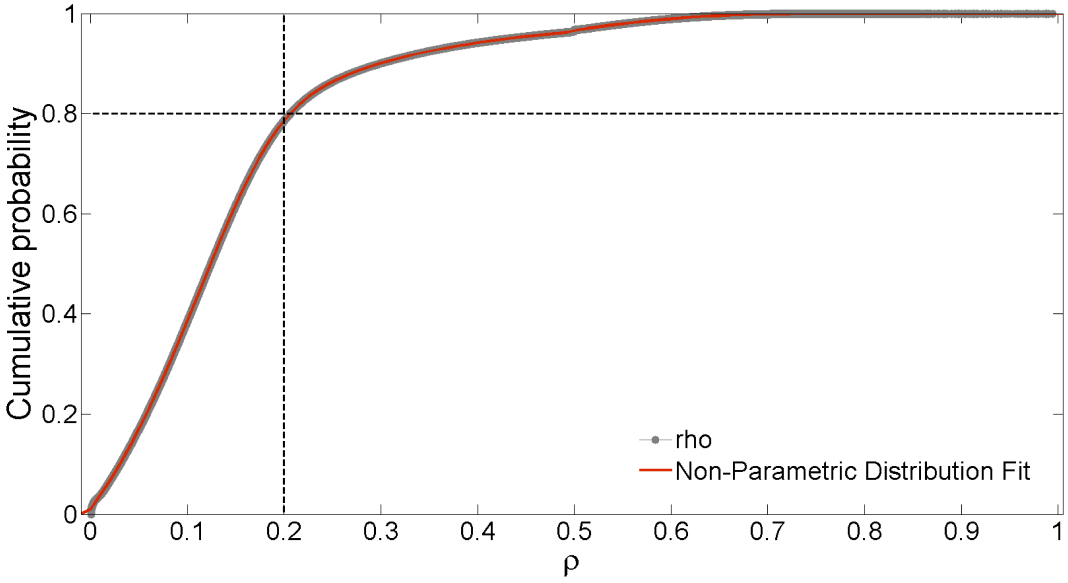

Figure 6.14: Cumulative Probability Fit

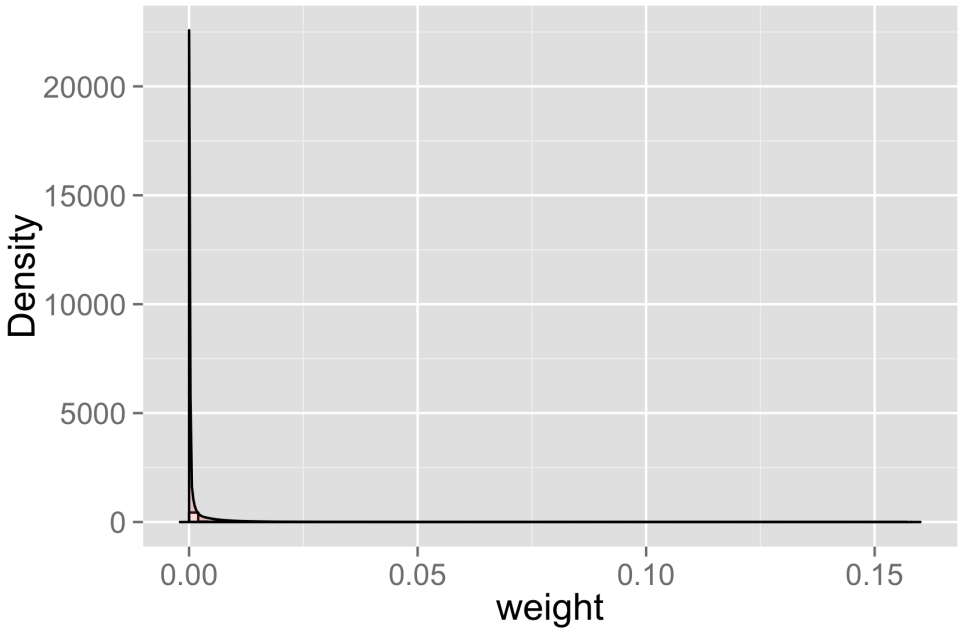

Figure 6.15: NMF 100 Density Fit

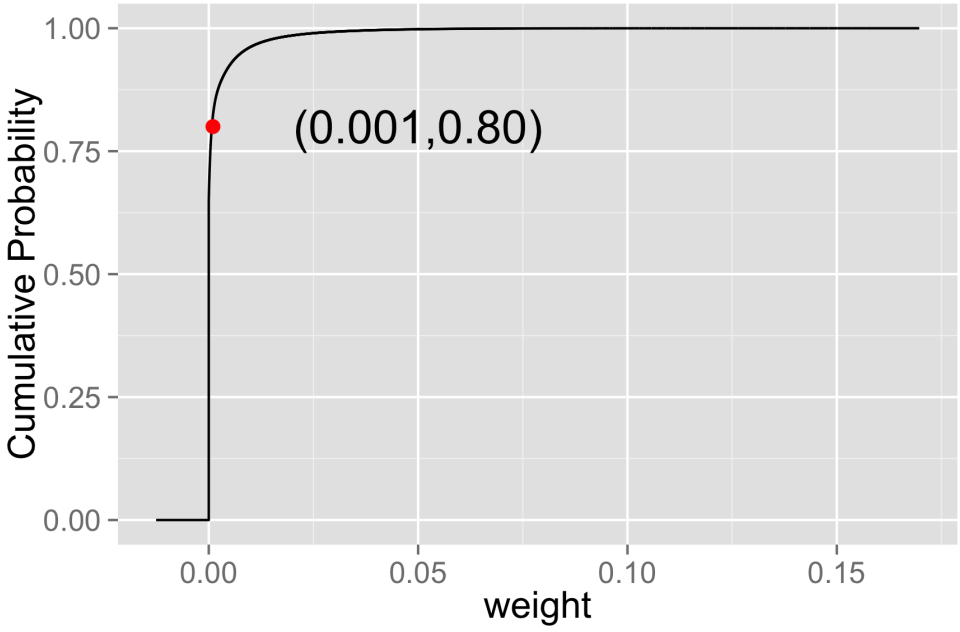

Figure 6.16: NMF 100 Cumulative Probability Fit

Chapter 7

Labeling Improvement with Multi-Label Learning (MLL)

7.1 Motivation

In Chapter 6, both unsupervised labeling, and TAGME with the facility of Wikipedia annotation are shown to provide relevant tags per documents. Specially, TAGME with the external content-rich resources improves both efficiency and accuracy for CBAC. The reason that TAGME is able to annotate documents with relevant tags is restrained with our feature of the experimental data set. NSF data set is considered to be an academic document data set. Wikipedia facilitates the academy fields for years and have expanded to most disciplines. Consider the following example where TAGME will fall short in making relevant tags.

Example 7: A large manufacturer initialized a project to annotate products with potential category tags, which are internally defined. The project need employees to read the product description and then make annotation for every product. Given millions of different products, and a finite set of categories, it is an extremely labor-extensive work for employees to read the description and manually annotate all the products. Plus the quality of tagging, therefore, is highly restrained by the employees' experience and understandings per the manufacturer's domain. The noise of the

tagging therefore is unknown.

∎

TAGME will not be functional for Example 7. First, since the tags are defined internally, even if Wikipedia has similar "terms" for the tags, they are most likely to mean totally different things. Second, the so-called "tags" has no well-defined content associated with them; while Wikipedia, under each web page (topic), there is rich content to explain what it is about, and which other web pages (topics), it is related with. In the scenario of Example 7, supervised labeling will be a possible effective alternative to replace TAGME. Supervised labeling facilitate labeling by lowering down the manual annotation cost. Given a sample of labeled data objects, supervised learning is able to make predictions of tags/labels to the rest data objects in the same knowledge domain. That being said is, first experienced employees will manually label a fraction of products, and second a supervised learning machine is trained to learn how to label products from the guidance provided the experience employees in their labelings. Thus, in this chapter, we want to theoretically study the cases with manual annotation: how well we could recover the labels for the rest part of samples thourgh supervised labelings.

7.2 Problem Definition and Challenges

Labeling for documents is classified as multi-label learning (Tsoumakas et al. (2010)), where one sample can be annotated with none to L labels, given L pre-defined labels. In other words, learning with instances which can be tagged with any of the 2^L possible subsets from the pre-defined L labels is called multi-label learning. Multi-label learning is commonly applied in domains, such as text, multimedia, web and biological data analysis, including automatic tagging, and function/topic predictions.

Multi-label learning, in ideal situations, cannot ignore the correlation between several labels in the L label set. However, the main challenge is actually the dilemma of optimizing label correlations over exponentially large label power-set and the ignorance of label correlations using binary

relevance strategy (1-vs-all heuristic). That means multi-label learning either treats "multi-label" as one single label in the label power-set, or treats the label as binary relevance where extends multi-label learning as L parallel binary learnings. The shortcomings from two different scenarios are obvious. The classification with label power-set usually encounters with highly skewed data distribution, called imbalanced problem (Chawla et al. (2004)). While binary relevance strategy reduces the problem from exponential to linear, it totally neglects the label correlations. However, binary relevance will keep the imbalance rate as is, which is, most of the times, highly skewed.

In this chapter, we propose a novel strategy of introducing Balanced Pseudo-Labels (BPL) which build more robust classifiers for imbalanced multi-label classification, which embeds imbalanced data in the problems innately. By incorporating the new balanced labels we aim to increase the average distances among the distinct label vectors with larger discriminant power. In this way, we also compress the label correlation implicitly in BPL. Another obvious advantage of the proposed method is that it is capable to combine with any classifier and it is proportional to linear label transformation.

In the experiment, we choose five multi-label benchmark data sets including two text data sets, one image data set, one biology data set, and one music data set, and compare our algorithm with the most state-of-art algorithms. Experimental results have shown our algorithm outperforms them in standard multi-label evaluation in most scenarios. The reason why we choose extra domain of data sets is because although the motivation of the algorithm is to facilitate text tagging and labeling, the algorithm is powerful enough to be applied on other kinds of data sets. Therefore, in the rest of the chapter, we expand our focus on multi-label learning not limited to multi-label learning in text domain.

7.3 Background

Multi-label learning (MLL) again refers to the problem to classify instances which can be tagged by 2^L possible subsets from the predefined L labels. Intuitively, MLL is widely applied in a variable

range of domains, such as text mining, multimedia classification, biological data analysis and many more. With the rapid pace of multimedia application development and the accumulation of massive biological data, the amount of data sources such as documents, images, music, and videos in personal collections, public biological data sets, and web data stream is rapidly growing. To this point, the popularity of MLL is based on three uncommon features of multi-labeled data, namely, multi-components, multi-facets, and the structural label taxonomy. That means, one single sample in learning problems has at least of one of the following characteristics, is a multi-label sample:

- The sample has multiple components, and each component has at least one label to be mapped to.
- The sample has multiple facets, and each facet has at least one label to be mapped to.
- The sample is from a hierarchical taxonomy structure, where the parent nodes of the taxonomy are also treated as labels to the sample.

For example, automatic image tagging is a process to assign multiple keywords to a digital image. Figure 7.1 shows a typical image labeled with sunset and sea, containing both a setting sun and a sea sight. That being said, the image has at least two components: a sun and a sea sight, each of which is mapped to one of the labels: *sun* and *sea sight*. Another typical example is the news for David Beckham tagged by *entertainment* and also *sports*. *Entertainment* and *Sports* are two different facets for *David Beckham*, which are mapped to two different types of news. The third example is the gene function categorization shown in Figure 7.2. The protein *P75957* is first annotated with the grey blocks. According to the taxonomy of gene ontology (GO) molecular function annotation, the structure represents a *forest* data structure, each protein gets its own node annotation from GO plus all its parent nodes annotation. Based on the principle of molecular function annotation of GO, The protein *P75957* is annotated with the entire forest shown in Figure 7.2. As shown above, multi-label learning is omnipresent in our daily life. Based on the diversity and broadness of applications which MLL applies, it becomes a hot topic with theoretical and applicative interests.

One of the commonly used MLL strategies is called problem transformation, which transforms MLL into multiple binary classification. The most popular problem transformation strategy is binary relevance (BR) (Table 7.1 and 7.2). Boutell *et al.* addressed BR transformations using SVM in scene analysis (Boutell et al. (2004)). Zhang and Zhou used an algorithm called ML-KNN to combine maximum a posterior (MAP) principle with k-Nearest Neighbors (kNN) upon each individual label (Zhang & Zhou (2007)). Lin and Chen showed the situation where using BR style kNN, multi-label samples might be viewed as outliers (Lin & Chen (2010)).

They proposed a kNN-based MLL approach called voting Margin-Ratio kNN (Mr. kNN) to prevent the false negative situation happened in ML-KNN. While it is simple to implement, the common judgment for BR-based method is that it neglects label correlation. However, the label correlation is crucial in many applications. For example, an image labeled with beach may likely be labeled with sea, while an image labeled with mountain is unlikely labeled with indoor. Importantly in bioinformatics, in digging out gene function correlations, a functional pathway will probably be revealed. Here comes another means called label power-set (LP) (Table 7.1, 7.3, and 7.4). Given L predefined labels, any possible subset in the data is combined as a new label in LP. The advantage of LP is it codes the label correlation into the classification process. However, it will make the complexity from linear as BR to exponential. What's more, the imbalanced rate will be deteriorated to extreme case. Tsoumakas and Vlahavas developed RAKEL strategy to consider the label correlation under a random k label combinations (Tsoumakas & Vlahavas (2007)). As a strategy between BR and LP, RAKEL neglects the problems lying beneath MLL other than label correlation, including imbalanced problems and the ambiguity in label combinations. Other methods have been raised in dealing with the label correlation problem. Wang *et al.* extended Green's function into MLL (Wang et al. (2009)). They used kernel-based method to modify the optimization function by introducing a penalty term constructed by the label correlation matrix. It is related to the algorithm (Ji et al. (2010)), in the sense of extracting shared subspace for MLL by adding a term of label dependency. Kang *et al.* raised correlated label propagation in MLL to introduce label dependency (Kang et al. (2006)).

In this chapter, we propose a strategy to introduce balanced pseudo-labels to improve MLL with skewed data distribution. Compared to others' work, the key contributions of this paper are highlighted as follows:

(1) We introduced the pseudo-labels to increase the distances between the new label vectors, so as to reduce the ambiguities lying in the original label space.

(2) We put emphasis on the difference of the pseudo-classifiers (with the pseudo-labels), so as to make the pseudo-classifiers to work efficiently together with little redundancy.

(3) We considered the balance rate of the pseudo-labels, in order to make more robust pseudo-classifiers.

The rest of the paper is organized as follows. In Section 7.4, we will discuss the relationship between our framework and existing methods. In Section 7.5, we will describe the proposed method with concrete mathematical definition. In Section 7.6, we will decribe the data sets and compare our methods with other state-of-art multi-label algorithms.

Table 7.1: Multi-Label Example

No.	Feature	Label
1	$\mathbf{X}_1$	l_1, l_2
2	$\mathbf{X}_2$	l_3
3	$\mathbf{X}_3$	l_2
4	$\mathbf{X}_4$	l_1, l_2, l_3
5	$\mathbf{X}_5$	l_1, l_2

Table 7.2: Binary Relevance Matrix Example

Feature	l_1	l_2	l_3
$\mathbf{X}_1$	1	1	0
$\mathbf{X}_2$	0	0	1
$\mathbf{X}_3$	0	1	0
$\mathbf{X}_4$	1	1	1
$\mathbf{X}_5$	1	1	0

Table 7.3: Label Power-Set Example

Label	Label Power-Set
l_1, l_2	$\mathbf{Z}_1$
l_3	$\mathbf{Z}_2$
l_2	$\mathbf{Z}_3$
l_1, l_2, l_3	$\mathbf{Z}_4$

Table 7.4: Label Power-Set Matrix Example

Feature	Label Power-Set
$\mathbf{X}_1$	$\mathbf{Z}_1$
$\mathbf{X}_2$	$\mathbf{Z}_2$
$\mathbf{X}_3$	$\mathbf{Z}_3$
$\mathbf{X}_4$	$\mathbf{Z}_4$
$\mathbf{X}_5$	$\mathbf{Z}_1$

7.4 Related Work

The proposed method is inspired by error-correction coding (ECC). ECC was a technique developed in 1950's for detecting and correcting the errors due to channel noises in signal communication (Hamming (1950)). In ECC, the original message is coded into a longer code-word by adding more binary digits. The longer code-word is transferred via a signal channel from the transmitter to the receiver. In the receiving terminal, the longer code-word is decoded back to the original length of the initial message.

The idea had been successfully applied to multi-class classification by transferring every label into a binary vector (Dietterich & Bakiri (1995)). Recently, researchers have applied ECC method to MLL (Kouzani & Nasireding (2009); Ferng & Lin (2011)). It has been shown that the ECC method boosts the performance of MLL in Hamming loss and subset accuracy (Ferng & Lin

Figure 7.1: Scene of Sunset at Sea

(2011)). However, limited explanation has been given to why ECC method can boost the performance. The possible explanation would be the original label space might be tight to discriminate the distinct label vectors. By increasing the dimensionality of label space, ECC will help to enlarge the distances between label vectors in the new space. In the proposed method, we formularize the objective function to make the distances increased during optimization iteration. Meanwhile, as the imbalanced problem will make the binary classifier unstable, we try to make the newly added pseudo-labels balanced.

Other methods have been applied to transform the label space. Compressed sensing (Hsu et al. (2009)), principle label space transformation (Tai & Lin (2012)), canonical relation analysis (Zhang & Schneider (2011); Sun et al. (2011)) have been used in MLL as to transform the label vectors into a real-valued scale. Then the MLL problem will be transformed into a regression problem. Thresholds of prediction should be learnt in these methods. Stable regression model usually needs more samples than classification. Therefore, to transfer MLL to a regression method compared to BR-based methods can usually make the model unstable. Classifier chains (Read et al. (2011)) embeds every previous labels in the feature space in MLL. The method considers

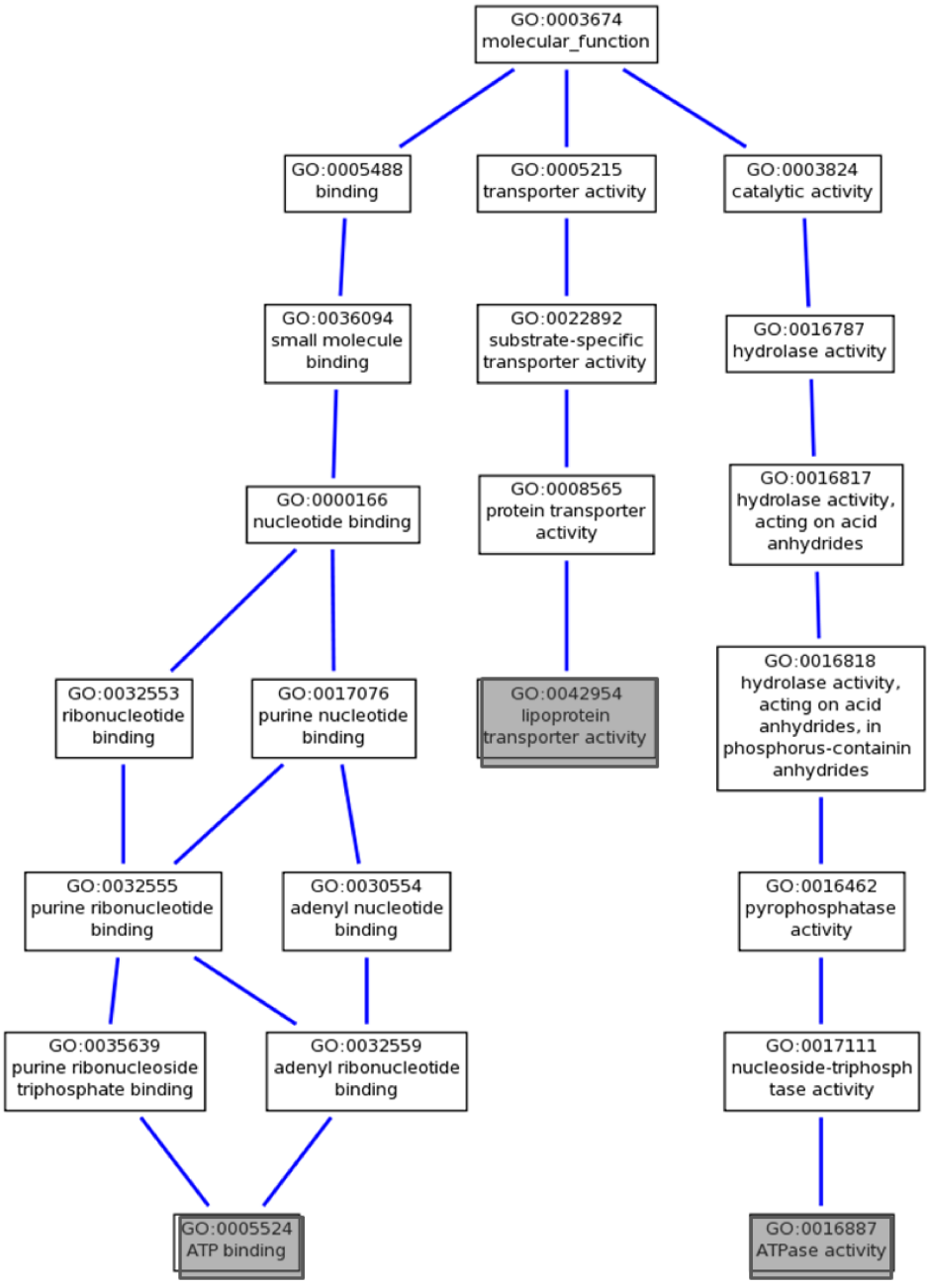

Figure 7.2: Molecular Function Annotation of P75957

the label correlation in serial order. In our proposed method, we consider the label correlation in a batch style rather than a serial order, since there is no clue that the label correlation comes in a sequential order.

In a word, in aim of increasing the distances between label vectors, and deriving discriminant pseudo-classifiers, the proposed method inspired by ECC makes the first effort to add balanced pseudo-labels to boost the performance of imbalanced MLL.

7.5 Methodology

In this section, we will describe the proposed methods with concrete mathematical definition.

7.5.1 Preliminary

Given an input domain $\mathscr{X} \subset \mathbf{R}^d$ and a binary output domain $\mathscr{Y} \subset \mathbf{B}^K$, where $\mathbf{B}$ consists -1 and +1. Let $\mathbf{T}$ represent the training data, where $\mathbf{T} = \{\mathbf{X}, \mathbf{Y}\} = \{(\mathbf{x}_1, \mathbf{y}_1),\ (\mathbf{x}_2, \mathbf{y}_2),\ \cdots,\ (\mathbf{x}_n, \mathbf{y}_n)\}$. Let $\mathbf{U} \subset \mathbf{B}^{m \times K}$ denote the unique label vectors in $\mathbf{T}$, and $\mathbf{w} \subset \mathbf{R}^m$ is the occurrence weight vector of $\mathbf{U}$. $\mathbf{w}$ is normalized so that the summation of its elements is equal to one, where m is the number of unique labels in $\mathbf{T}$. BPL's first goal is to find a $\mathbf{Z} \subset \mathbf{B}^{m \times p}$, by adding $\mathbf{Z}$ to the right of $\mathbf{U}$, to maximize the average distances between the newly formed label vectors. In the next subsection, we are going to discuss how we form the objective function to find $\mathbf{Z}$. It should be noted that $m \leq 2^p - 2$, since the new added pseudo-labels should be at least sufficient to represent the original unique label vectors with at least one +1's and one -1's.

7.5.2 Objective Function

In seeking $\mathbf{Z}$, we consider the row-separation to increase the distances between label vectors, and column-separation to make the pseudo-classifiers distinct from each other. Therefore, we use $\mathbf{Z} = \{\mathbf{z}_{r1},\ \mathbf{z}_{r2}, \cdots,\ \mathbf{z}_{rm}\}$ to denote the m row vectors in $\mathbf{Z}$ and $\mathbf{U} = \{\mathbf{u}_{r1},\ \mathbf{u}_{r2}, \cdots,\ \mathbf{u}_{rm}\}$ to denote the m row vectors in $\mathbf{U}$. Similarly, we use $\mathbf{Z} = \{\mathbf{z}_{c1},\ \mathbf{z}_{c2}, \cdots,\ \mathbf{z}_{cp}\}$ to denote the p column vectors

in $\mathbf{Z}$. Row-wisely, as we are trying to increase the distance between the new label vectors $(\mathbf{u}_{ri}, \mathbf{z}_{ri})$ and $(\mathbf{u}_{rj}, \mathbf{z}_{rj})$, where $i \neq j$ and $i, j = 1, 2, \cdots, m$. The distance comes from two parts, including the distance between $\mathbf{u}_{ri}$ and $\mathbf{u}_{rj}$, and the distance between $\mathbf{z}_{ri}$ and $\mathbf{z}_{rj}$. The first part is fixed in the training data. The second part is what we need to work on. The distance between any two rows in $\mathbf{Z}$ is as follows.

$$\begin{aligned} d(\mathbf{z}_{ri},\, \mathbf{z}_{rj}) &= (\mathbf{z}_{ri} - \mathbf{z}_{rj})(\mathbf{z}_{ri} - \mathbf{z}_{rj})^T \\ &= \mathbf{z}_{ri}\mathbf{z}_{ri}^T - \mathbf{z}_{ri}\mathbf{z}_{rj}^T - \mathbf{z}_{rj}\mathbf{z}_{ri}^T + \mathbf{z}_{rj}\mathbf{z}_{rj}^T \\ &= \mathbf{z}_{ri}\mathbf{z}_{ri}^T - 2\mathbf{z}_{ri}\mathbf{z}_{rj}^T + \mathbf{z}_{rj}\mathbf{z}_{rj}^T \end{aligned} \tag{7.1}$$

From the above derivation, if the inner product of a row in $\mathbf{Z}$ itself is a constant number, the distance is inversely proportional to the inner product of the two rows of $\mathbf{Z}$. Since $\mathbf{Z} \subset \mathbf{B}^{m \times p}$, the inner product of a row in $\mathbf{Z}$ itself is equal to p. Then to maximize the distance of $\mathbf{Z}$ row-wisely is to minimize the part as follows.

$$P_1 = \sum\sum \mathbf{Z}\mathbf{Z}^T \tag{7.2}$$

Column-wisely, we need to build up distinct pseudo-classifiers. In the mathematical sense, we also need to maximize the distance of columns in $\mathbf{Z}$. Therefore, similarly, we get P_2 of minimization as follows.

$$P_2 = \sum\sum \mathbf{Z}^T\mathbf{Z} \tag{7.3}$$

The last part of BPL is to bring the balance into pseudo-labels. The balance rate is guaranteed by minimizing the different counts between positive samples and negative samples. Therefore, we have the balance rate vector β of the pseudo-labels calculated as follows.

$$\beta = 1_{1\times m} \times ((\mathbf{w} \times 1_{1\times p}) \cdot \mathbf{Z})) \tag{7.4}$$

where $1_{1\times m}$ denote the matrix of all one's with one row and m columns. Thus, the P_3 of minimization is as follows.

$$P_3 = \beta\beta^T \tag{7.5}$$

In order to consider the three minimization parts all together, a trade-off parameter needs to be introduced as λ. Then the objective function is as follows. The format of the following optimization objective function follows quadratic integer programming. As known, integer programming is quite expensive (Papadimitriou (1981)). Therefore, it should be noticed that the objective function cannot guarantee converge at polynomial time. Thus, we use greedy searching for improved label space shown in the later algorithms

$$Q = \sum\sum \mathbf{Z}\mathbf{Z}^T + \sum\sum \mathbf{Z}^T\mathbf{Z} + \lambda \cdot \beta\beta^T \tag{7.6}$$

In the following section, we will introduce the learning and the predicting processes.

7.5.3 Algorithm

In this section, we will describe the details of learning and predicting processes of BPL.

In BPL, the first step is to generate the pseudo-labels. The generation of pseudo-labels is described in Algorithm 11.

In the experiment, the *maxNum* is set to 10^6. Although the *maxNum* seems large, the permulation step is quick. The range of λ is $\{2^{-10}, 2^{-9}, 2^{-8}, \cdots, 2^{10}\}$.

The second step is to learn individual label parameters in maximizing F1 measure. Here, parameters are tuned independently to maximize the performance of the individual classifiers.

In the prediction part, a testing sample will be first predicted by the model to generalize the

Algorithm 11 Pseudo-Label Generation

Require: **T**, p, λ and maximum iteration M

Compute **U** and **w** from **T**.
$Q \leftarrow \infty$
bV with size $(2^p - 2) \times p$ columns is generated by excluding all -1's and +1's of the comprehensive combination of p bits.
for all $i \in \{1, 2, \cdots, M\}$ **do**
    Permute the rows of **bV** to select m rows to form **Z** and calculate the objective function value Q_{curr}.
    if Q_{curr} is less than Q **then**
        Record the current **Z**.
    end if
end for
return **U**, **w** and **Z**

Algorithm 12 Individual Label Learning

Require: **T**, **Z**, k, p and Classifier C

Compute the new label vectors $\mathbf{L}_T$ for samples in **T**.
for all $i \in \{1, 2, \cdots, K + p$ **do**
    Train C's model f with k-fold cross-validation.
end for
return f

predicted $K+p$ binary digits $\mathbf{l}'$. Then, the distances of $\mathbf{l}'$ and $\mathbf{L}$ in Algorithm 14 are calculated to find the shortest distance and the corresponding label vector(s) in $\mathbf{L}$. A weighted averaging of the nearest label vectors (if more than one) is computed and thresholded by 0 to return a label vector with 1's and -1's. Then, the first K digits of the returned label vector is given out as the final prediction $\mathbf{l}$. In the next section, we are going to discuss the experiment with description of data sets, comparison algorithms, and metrics of evaluation.

7.6 Experiment

In the experiment, we applied our algorithm with random forest (Liaw & Wiener (2002)) and SVM (Cristianini & Shawe-Taylor (2000)) of linear and RBF kernels in five multi-label benchmark data sets, including *emotions*, *enron*, *medical*, *scene* and *yeast*.

These data sets come from different domains, such as music categorization, sight scene image analysis, and gene function classification. They differ in training and testing sizes, category numbers, features and other statistics related to MLL are presented in this section. In the following subsection, we will present the statistics of these data sets.

Algorithm 13 Learning Process

Require: $\mathbf{T}$, $\mathbf{Z}$, p_{max}, k, Λ and Classifier C

1: $p_{min} \leftarrow \lceil \log_2(m+2) \rceil$
2: **for all** $(p,\lambda) \in \{p_{min} \leq p \leq p_{max}, \lambda \in \Lambda\}$ **do**
3: Perform Pseudo-Label Generation.
4: Perform Individual Label Learning.
5: **end for**
6: Train individual models with k-fold cross-validation, and tune p and λ to get the final p and λ.
7: **for all** $i \in \{1,2,\cdots,K+p\}$ **do**
8: Learn C's model h_i with f_i from $\mathbf{X}$ to $\mathbf{L}_{\mathbf{T_i}}$.
9: **end for**
10: **return** p, λ, $\mathbf{U}$, $\mathbf{w}$, $\mathbf{Z}$ and h

Table 7.5: Statistics of Data Sets

Data Set	CATE	CARD	DENS	$\|\mathbf{U}\|$	COVR
emotions	6	1.87	0.31	26	0.41
enron	53	3.39	0.06	545	0*
medical	45	1.2	0.03	61	0*
scene	6	1.0	0.18	14	0.22
yeast	14	4.2	0.30	164	0.01

Algorithm 14 Prediction

Require: the testing sample $\mathbf{x}$, h, $\mathbf{U}$, $\mathbf{w}$ and $\mathbf{Z}$

1: Form $\mathbf{L} = (\mathbf{U}, \mathbf{Z})$.
2: **for all** $i \in \{1, 2, \cdots, K+p\}$ **do**
3: Predict $\mathbf{l}'_i$ via h_i.
4: **end for**
5: Find the shortest distances from the predicted label vector $\mathbf{l}'$ to $\mathbf{L}$.
6: Use weighted average of the label vectors with the shortest distances and threshold the values with zero to make the final prediction $\mathbf{t}$. The weight comes from the corresponding elements in the weight vector $\mathbf{w}$.
7: The predicted label vector is the first K digits from $\mathbf{t}$ denoting as $\mathbf{l}$.
8: **return** $\mathbf{l}$

7.6.1 Data Set Statistics

Table 7.7 shows the statistics of sizes of samples and features of the data sets. We also computed the category size (CATE), label carginality (CARD), label density (DENS), unique label size $|\mathbf{U}|$ in training data and label coverage (COVR) for each data set. 0* represents a number less than 10^{-10}. The carginality, and density which commonly used in MLL are defined as follows.

Table 7.6: Imbalance Rate (%)

	Average Positive Rate
emotions	30.22
enron	6.39
medical	2.79
scene	17.70
yeast	30.20

Table 7.7: Sample Sizes of Data Sets

Data Set	**Training Size**	**Testing Size**	**Feature Size**
emotions	391	202	72
enron	1123	579	1001
medical	333	645	1449
scene	1211	1196	294
yeast	1500	917	103

$$CARD = \frac{\sum_{i=1}^{n} |\mathbf{y}_i|}{n} \tag{7.7}$$

$$DENS = \frac{\sum_{i=1}^{n} |\mathbf{y}_i|}{n \cdot K} \tag{7.8}$$

We generate a new statistical metric in MLL called label coverage and define it as follows.

$$COVR = \frac{|\mathbf{U}|}{2^K} \tag{7.9}$$

Label cardinality shows the average label size per sample. It determines the degree of multi-label extent from sample perspective. Label density shows the average label rate over the samples and the categories. It represents the label utility from sample perspective. Label coverage reflects how crowd the original label space is.

Table 7.6 represents the imbalance rate of the five datasets. As shown in Table 7.6, every data set embeds different rate of imbalance.

The following subsection will introduce the comparison methods in the experiment.

Table 7.8: Macro-Averaging F1 Measure (%) ↑

	emo	enr	med	sce	yea
BR-LSVM	60.57	13.22	35.54	68.80	35.36
BPL-LSVM	65.92	13.57	33.63	70.85	39.68
BR-RSVM	58.85	12.08	37.21	60.06	25.46
BPL-RSVM	62.71	13.86	35.21	64.88	39.19
BR-RF	66.82	21.68	38.69	70.86	43.87
BPL-RF	**70.46**	**23.39**	36.32	**78.14**	**50.44**
IBLR	62.12	11.24	-	73.95	46.97
LS-CCA	63.14	12.78	35.86	58.36	36.99
LS-CCA-L1	62.87	20.34	**42.46**	63.69	36.90
LS-CCA-L2	62.80	21.53	39.75	64.89	37.76
L-M3L	64.77	17.80	-	67.04	36.14
R-M3L	57.81	22.69	-	77.17	49.83

7.6.2 Comparison Methods

We applied three different algorithms in the comparison part, namely IBLR (Cheng & Hüllermeier (2009)), M3L (Hariharan et al. (2010)), and LS-CCA (Sun et al. (2011)).

IBLR is an instance-based logistic regression method to combine logistic regression with k-nearest neighbours (kNN). In the experiment, we calculate out the parameter α in maximizing the likelihood function supposed (Cheng & Hüllermeier (2009)). M3L is a max-margin multi-label fomulation method. It codes the priors of label correlation via a correlation matrix. In the experiment, the correlation matrix is calculated from training data via Pearson's Correlation. Linear and RBF kernels have been applied. LS-CCA is a canonical relation analysis via least square formulation. In the experiment, LS-CCA both with and without regulation are applied. The regulation of L1 and L2 have been implemented. The parameters and thresholds in these experiments are tuned with 3-fold cross-validation. In LS-CCA with L1 and L2 regulation, the regulation coefficient is in the range of $\{2^{-10}, 2^{-9}, \cdots, 2^{9}, 2^{10}\}$. In M3L with linear kernel, we

Table 7.9: Micro-Averaging F1 Measurel (%) ↑

	emo	enr	med	sce	yea
BR-LSVM	62.44	53.60	78.08	68.29	63.45
BPL-LSVM	66.26	48.30	74.84	70.22	63.62
BR-RSVM	60.09	49.70	78.17	60.19	63.50
BPL-RSVM	64.71	45.02	72.50	63.22	63.87
BR-RF	67.03	**58.11**	**80.65**	70.64	64.40
BPL-RF	**71.15**	53.65	78.39	**77.52**	**67.76**
IBLR	62.60	43.17	-	70.99	67.02
LS-CCA	53.04	28.29	46.87	44.57	58.08
LS-CCA-L1	52.50	49.14	62.51	47.25	57.94
LS-CCA-L2	52.02	47.74	59.20	47.85	64.27
L-M3L	65.27	55.58	-	66.44	63.57
R-M3L	59.46	56.92	-	76.56	66.43

tuned the bias with $\{0, 0.1, \cdots, 1\}$ and the cost with $\{2^{-6}, 2^{-4}, \cdots, 2^{6}\}$. For RBF kernel, we tuned the bias and the cost as in linear kernel and γ with $\{2^{-6}, 2^{-4}, \cdots, 2^{6}\}$. For all the thresholds, we learn the value of each individual label from the minimum scale S_{min} to the maximum scale S_{max} we got from the cross-validation with the interval of $(S_{max} - S_{min})/100$.

In the experiment, we applied random forest and SVM of both linear and RBF kernels. The node number is tuned with $\{10, 20, \cdots, 100\}$ percent of feature size in random forest with the package of random forest (Jaiantilal (2009)). We tuned the cost with $\{2^{-6}, 2^{-4}, \cdots, 2^{6}\}$, and γ with $\{2^{-6}, 2^{-4}, \cdots, 2^{6}\}$ with linear and RBF SVMs with libsvm package (Chang & Lin (2011)). In BPL, we tune $p = \{\lceil \log_2(m+2) \rceil, \lceil \log_2(m+2) \rceil + 1, \cdots, 16\}$ and $\Lambda = \{2^{-10}, 2^{-9}, \cdots, 2^{9}, 2^{10}\}$

7.6.3 Evaluation Metric

In evaluating the experiment, we used standard metrics defined as follows, excluding some inaccurate metrics such as accuracy, and hamming loss.

In the follows, TP_l, FN_l, and FP_l denote true positive, false negative, and false positive rates for label l. *MI* represents micro-averaging, while *MA* represents macro-averaging. As recall and

precision have a reciprocal relationship, F1 measure calculates the balance of them.

$$recall_{MI} = \frac{\sum_{l=1}^{K} TP_l}{\sum_{l=1}^{K}(TP_l + FN_l)} \tag{7.10}$$

$$precision_{MI} = \frac{\sum_{l=1}^{K} TP_l}{\sum_{l=1}^{K}(TP_l + FP_l)} \tag{7.11}$$

$$F1_{MI} = 2 \cdot \frac{recall_{MI} \cdot precision_{MI}}{recall_{MI} + precision_{MI}} \tag{7.12}$$

$$recall_{MA} = \frac{1}{K}\sum_{i=1}^{K} \frac{TP_l}{TP_l + FN_l} \tag{7.13}$$

$$precision_{MA} = \frac{1}{K}\sum_{i=1}^{K} \frac{TP_l}{TP_l + FP_l} \tag{7.14}$$

$$F1_{MA} = 2 \cdot \frac{recall_{MA} \cdot precision_{MA}}{recall_{MA} + precision_{MA}} \tag{7.15}$$

$$subset\text{-}accuracy = \frac{\sum_{j=1}^{N} \delta(\mathbf{y}_j = \mathbf{t}_j)}{N} \tag{7.16}$$

In subset accuracy, δ function is a boolean expression. If $\mathbf{y}_j = \mathbf{t}_j$ is true, the result is one; otherwise, the result is zero. Subset accuracy is considered to be the strictest evaluation metric in multi-label classification, because it measures the exact matching of predictions.

7.6.4 Results

In this part, we are going to show the result of our methods with comparison methods. In the tables, *emo*, *enr*, *med*, *sce* and *yea* represent *emotions*, *enron*, *medical*, *scene*, and *yeast*.

In Table 7.8, macro-averaging metric is presented. Macro-averaging metrics focus on the average performance from label perspective. That means the metrics give equal weights for labels regardless to different positive sample sizes in MLL. As stated before, we consider F1 measure as the most important metric compared to recall and precision. Macro-averaging F1 measure in

Table 7.10: Subset Accuracy (%) ↑

	emo	enr	med	sce	yea
BR-LSVM	25.25	12.44	63.10	51.92	16.03
BPL-LSVM	32.67	14.68	64.81	65.64	21.81
BR-RSVM	25.25	7.94	62.48	43.65	16.03
BPL-RSVM	29.21	11.40	61.71	58.19	20.72
BR-RF	29.21	14.68	66.05	55.18	17.78
BPL-RF	**36.63**	**16.75**	**68.68**	**71.49**	**25.52**
IBLR	13.86	8.46	-	56.10	16.36
LS-CCA	0.05	0	12.40	11.04	0.11
LS-CCA-L1	0.05	0	21.86	9.45	0.11
LS-CCA-L2	0.05	0	21.24	8.19	0
L-M3L	29.21	14.51	-	50.33	16.36
R-M3L	19.80	15.37	-	63.63	18.97

Table 7.8 shows *BPL-RF* out-performs the naive BR-based methods and the other methods except for *enron*.

In Table 7.9, micro-averaging metric is presented. Micro-averaging metrics focus on the average performance from sample perspective. That means the metrics give equal weights for samples in MLC. It shows random forest-based methods including *BR-RF* and *BPL-RF*, out-perform the others in micro-averaging F1 measure.

In Table 7.10, the subset accuracy is listed. The metric is viewed as the strictest evaluation criterion as it rates for samples with exact matching of labels. In the result, *BPL-RF* out-performs all the other methods in this metric. This goes with the method proposed. Although we applied BR strategy after generating pseudo-labels, we aim to find the closest label vectors via BPL.

7.7 Conclusion

As presented in the experimental results, BPL is able to boost the accuracy of labeling with training manual annotation by lowering down the imbalance rates and increase the discrimination power between label vectors. Although the labeling training is expensive in time cost, it is performed

offline. Therefore, CBAC with the content-centric database from specific domain is capable to train their tagging by initializing a set of manual tags covering the basic tags in the specific domain with BPL and scalable classification algorithm like random forest.

Chapter 8

Discussions

8.1 Computational Complexity

The efficiency and scalability issues are major metrics in evaluating any database access control approach. As we have shown in the experiments, the content-based access control model could be efficiently enforced using native functions from Oracle DBMS, especially with blocking and labeling. Meanwhile, we would also give a formal analysis of computational complexity.

First, without any indexing and blocking, the DBMS needs to perform a pairwise comparison between every record and the user's records, in order to make access control decisions. The computational complexity is $O(N \cdot m)$, where N denotes the total number of records (usually very large), and m denotes the number of records that represent the user. Assuming that we adopt the vector space model, the complexity for each comparison would be $O(D)$, where D denotes the dimensionality of the vector space (*a.k.a.* the size of the dictionary). D does not grow linearly with the growth of N. Indeed, as long as the a good number of records cover most of the words used in the context, D increases very slowly when new records are added. Meanwhile, the computation for each comparison could be easily reduced to $O(d)$ (by only including terms in user's records in computing cosine similarity), where d being the number of distinct terms in the user's record. Hence, the overall computational complexity for a query would be $O(N \cdot m \cdot d)$. As we see, query

processing time is linear to the number of records, while m and d are relatively small.

Next, with blocking, the DBMS first selects x clusters of records from the total c clusters, and then perform pairwise comparison between user's records and records within the c clusters. Assuming that the size of clusters are relatively balanced, and each record only belongs to one cluster, there will be N/c records in each cluster on average. Note that content-based clustering is performed offline; hence, the computation for clustering is not concerned in our approach. The computation for selecting top x clusters is $O(c \cdot m \cdot d)$, while the computation for enforcing CBAC for records within the x clusters would be $O((N/c) \cdot m \cdot d \cdot x)$. Hence, the total computation would be:

$$O(c \cdot m \cdot d + \frac{N}{c} \cdot m \cdot d \cdot x) \geq O(2m \cdot d\sqrt{N \cdot x}) \tag{8.1}$$

Hence, the blocking mechanism reduces the overall computation to $O(m \cdot d \cdot sqrtN \cdot x)$. It could be further reduced to $O(m \cdot d \cdot \log(N \cdot x))$, if we implement multi-level blocking.

8.2 Negative Rules and Conflict Resolution

In database access control, negative rules are employed to prevent the user (or role) from accessing specified records. Usually, positive rules allows the user to access a (relatively large) set of records, while negative rules excludes some particular records from the set. It is still possible to use negative rules in CBAC. For instance, to specify that "Agent *Alice* cannot access records similar to Record *X*" in Example 1. Meanwhile, in the case of conflict rules (e.g. a positive rule grants access to a record, while a negative rule forbids it), the negative rule usually takes precedence. In some access control model, the rule with a smaller scope takes precedence. In CBAC, we also have the capability to specify that the rule with higher content-similarity takes precedence. However, details on this topic is mostly the choice of the administrator, and is outside of the scope of this dissertation.

8.3 CBAC for XML Data

Last but not least, CBAC could be effectively applied on XML data as well. We only need to redesign the record-wise similarity function in Equation 4.1 to adapt to XML data. The new function will take two XML nodes, traverse the entire subtrees, and return their similarity value. XML similarity assessment is more complicate than relational data, due to the semi-structured nature o the data. In particular, both structural similarity and textual similarity need to be considered in comparing two XML documents or nodes. For more details about XML similarity comparison, a survey is available (Tekli et al. (2009)). Meanwhile, all the content similarity assessment techniques discussed in Chapter 4, 5 and 6 are still valid for XML data.

Chapter 9

Conclusion

In this dissertation, we introduce the content-based access control and enforcement mechanisms. As a complimentary of the conventional access control approaches, the CBAC model is most suitable for content-centric information sharing scenarios, when content plays a major role in access decision making, and approximation is allowed by the application. CBAC makes access control decisions based on the semantical similarity between the requester's credentials (often represented by his/her own records) and the content of the data. We formally present the content-based access control model, and demonstrate an enforcement mechanism of this model on Oracle VPD. Meanwhile, to improve the computational efficiency of the enforcement mechanism, we introduce an offline similarity assessment approach (like an index), and a content-based blocking approach. We further improve the accuracy of semantic content matching with a content-based tagging mechanism. Experimental results show that the access control decision made by CBAC are reasonable, and the overhead is also acceptable.

Finally yet importantly, the CBAC model provides no restrictions on user and content modeling. We have presented a proof-of-concept implementation of the CBAC model with vector space and tag-based models. In practice, more complicated user and content modeling methods could be employed. For instance, it will be helpful to include advanced content models such as opinion extraction (Ku et al. (2006)), and sentiment analysis (Pang & Lee (2008)). However, understanding

the semantic content of unstructured text content is a very difficult problem, which is outside of the scope of this paper. It is one of the main tasks of our future work.

References

Adam, N. R., Atluri, V., Bertino, E., & Ferrari, E. (2002). A content-based authorization model for digital libraries. *Knowledge and Data Engineering, IEEE Transactions on*, 14(2), 296–315.

Ahn, G.-J. (2009). Discretionary access control. In *Encyclopedia of Database Systems* (pp. 864–866). Springer.

Amjad, H. (2007). A context aware content based federated access control system for healthcare domain. *ECE Masters Theses*, (pp.1̃3).

Appari, A. & Johnson, M. E. (2010). Information security and privacy in healthcare: current state of research. *International journal of Internet and enterprise management*, 6(4), 279–314.

Arthur, D. & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* (pp. 1027–1035).: Society for Industrial and Applied Mathematics.

Aurenhammer, F. (1991). Voronoi diagrams?a survey of a fundamental geometric data structure. *ACM Computing Surveys (CSUR)*, 23(3), 345–405.

Bache, K. & Lichman, M. (2013). UCI machine learning repository.

Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9), 509–517.

Bertino, E., Bettini, C., Ferrari, E., & Samarati, P. (1996). A temporal access control mechanism for database systems. *Knowledge and Data Engineering, IEEE Transactions on*, 8(1), 67–80.

Bertino, E., Bonatti, P. A., & Ferrari, E. (2001a). Trbac: A temporal role-based access control model. *ACM Transactions on Information and System Security (TISSEC)*, 4(3), 191–233.

Bertino, E., Castano, S., & Ferrari, E. (2001b). Securing xml documents with author-x. *Internet Computing, IEEE*, 5(3), 21–31.

Bertino, E., Catania, B., Ferrari, E., & Perlasca, P. (2003a). A logical framework for reasoning about access control models. *ACM Transactions on Information and System Security (TISSEC)*, 6(1), 71–127.

Bertino, E., Fan, J., Ferrari, E., Hacid, M.-S., Elmagarmid, A. K., & Zhu, X. (2003b). A hierarchical access control model for video database systems. *ACM Transactions on Information Systems (TOIS)*, 21(2), 155–191.

Bertino, E. & Ferrari, E. (2002). Secure and selective dissemination of xml documents. *ACM Transactions on Information and System Security (TISSEC)*, 5(3), 290–331.

Bertino, E., Ghinita, G., & Kamra, A. (2011). *Access control for databases: concepts and systems*, volume 8. Now Publishers Inc.

Bertino, E. & Haas, L. M. (1988). Views and security in distributed database management systems. In *Advances in Database Technology—EDBT'88* (pp. 155–169). Springer.

Bertino, E., Haas, L. M., & Lindsay, B. G. (1983). View management in distributed data base systems. In *VLDB* (pp. 376–378).

Bertino, E., Hammad, M. A., Aref, W. G., & Elmagarmid, A. K. (2000). An access control model for video database systems. In *Proceedings of the ninth international conference on Information and knowledge management* (pp. 336–343).: ACM.

Bertino, E., Samarati, P., & Jajodia, S. (1997). An extended authorization model for relational databases. *Knowledge and Data Engineering, IEEE Transactions on*, 9(1), 85–101.

Bhatti, R., Sanz, D., Bertino, E., & Ghafoor, A. (2007). A policy-based authorization framework for web services: Integrating xgtrbac and ws-policy. In *Web Services, 2007. ICWS 2007. IEEE International Conference on* (pp. 447–454).: IEEE.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3, 993–1022.

Boutell, M. R., Luo, J., Shen, X., & Brown, C. M. (2004). Learning multi-label scene classification. *Pattern recognition*, 37(9), 1757–1771.

Boxwala, A. A., Kim, J., Grillo, J. M., & Ohno-Machado, L. (2011). Using statistical and machine learning to help institutions detect suspicious access to electronic health records. *Journal of the American Medical Informatics Association*, 18(4), 498–505.

Byun, J.-W. & Li, N. (2008). Purpose based access control for privacy protection in relational database systems. *The VLDB Journal*, 17(4), 603–619.

Carpineto, C., Osiński, S., Romano, G., & Weiss, D. (2009). A survey of web clustering engines. *ACM Computing Surveys (CSUR)*, 41(3), 17.

Chang, C.-C. & Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3), 27.

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Editorial: special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter*, 6(1), 1–6.

Cheng, W. & Hüllermeier, E. (2009). Combining instance-based learning and logistic regression for multilabel classification. *Machine Learning*, 76(2-3), 211–225.

Cristianini, N. & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

Damiani, E., De Capitani di Vimercati, S., Paraboschi, S., & Samarati, P. (2002). A fine-grained access control system for xml documents. *ACM Transactions on Information and System Security (TISSEC)*, 5(2), 169–202.

Damiani, E., di Vimercati, S. D. C., Paraboschi, S., & Samarati, P. (2000). Securing xml documents. In *Advances in Database Technology—EDBT 2000* (pp. 121–135). Springer.

Dekker, M., Crampton, J., & Etalle, S. (2008). Rbac administration in distributed systems. In *Proceedings of the 13th ACM symposium on Access control models and technologies* (pp. 93–102).: ACM.

Dietterich, T. G. & Bakiri, G. (1995). Solving multiclass learning problems via error-correcting output codes. *arXiv preprint cs/9501101*.

Downs, D. D., Rub, J. R., Kung, K. C., & Jordan, C. S. (1985). Issues in discretionary access control. In *2012 IEEE Symposium on Security and Privacy* (pp. 208–208).: IEEE Computer Society.

Fabian, B., Kunz, S., Konnegen, M., Müller, S., & Günther, O. (2012). Access control for semantic data federations in industrial product-lifecycle management. *Computers in Industry*, 63(9), 930–940.

Ferng, C.-S. & Lin, H.-T. (2011). Multi-label classification with error-correcting codes. In *ACML* (pp. 281–295).

Ferragina, P. & Scaiella, U. (2010). Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 1625–1628).: ACM.

Ferraiolo, D. F., Sandhu, R., Gavrila, S., Kuhn, D. R., & Chandramouli, R. (2001). Proposed nist standard for role-based access control. *ACM Transactions on Information and System Security (TISSEC)*, 4(3), 224–274.

Ferreira, A., Cruz-Correia, R., Antunes, L., Farinha, P., Oliveira-Palhares, E., Chadwick, D. W., & Costa-Pereira, A. (2006). How to break access control in a controlled manner. In *Computer-Based Medical Systems, 2006. CBMS 2006. 19th IEEE International Symposium on* (pp. 847–854).: IEEE.

Giuri, L. & Iglio, P. (1997). Role templates for content-based access control. In *Proceedings of the second ACM workshop on Role-based access control* (pp. 153–159).: ACM.

Griffiths, P. P. & Wade, B. W. (1976). An authorization mechanism for a relational database system. *ACM Transactions on Database Systems (TODS)*, 1(3), 242–255.

Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell System technical journal*, 29(2), 147–160.

Hariharan, B., Zelnik-Manor, L., Varma, M., & Vishwanathan, S. (2010). Large scale max-margin multi-label classification with priors. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)* (pp. 423–430).

Hart, M., Johnson, R., & Stent, A. (2007). More content-less control: Access control in the web 2.0. *IEEE Web*, 2.

Hart, M. A. (2006). Content-based access control.

Hicks, B., Rueda, S., King, D., Moyer, T., Schiffman, J., Sreenivasan, Y., McDaniel, P., & Jaeger, T. (2010). An architecture for enforcing end-to-end access control over web applications. In *Proceedings of the 15th ACM symposium on Access control models and technologies* (pp. 163–172).: ACM.

Hoare, C. A. R. (1961). Algorithm 65: find. *Communications of the ACM*, 4(7), 321–322.

Hsu, D., Kakade, S., Langford, J., & Zhang, T. (2009). Multi-label prediction via compressed sensing. In *NIPS*, volume 22 (pp. 772–780).

Hu, V. C., Kuhn, D. R., & Ferraiolo, D. F. (2015). Attribute-based access control. *Computer*, (2), 85–88.

Inaba, M., Katoh, N., & Imai, H. (1994). Applications of weighted voronoi diagrams and randomization to variance-based k-clustering. In *Proceedings of the tenth annual symposium on Computational geometry* (pp. 332–339).: ACM.

Jaiantilal, A. (2009). Random forest package (matlab version).

Jajodia, S., Samarati, P., Subrahmanian, V., & Bertino, E. (1997). A unified framework for enforcing multiple access control policies. In *ACM Sigmod Record*, volume 26 (pp. 474–485).: ACM.

Jajodia, S. & Sandhu, R. (1991). Toward a multilevel secure relational data model. In *ACM SIGMOD Record*, volume 20 (pp. 50–59).: ACM.

Ji, S., Tang, L., Yu, S., & Ye, J. (2010). A shared-subspace learning framework for multi-label classification. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(2), 8.

Joachims, T. (1996). *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization.* Technical report, DTIC Document.

Kagal, L., Finin, T., & Joshi, A. (2003). A policy based approach to security for the semantic web. In *The Semantic Web-ISWC 2003* (pp. 402–418). Springer.

Kang, F., Jin, R., & Sukthankar, R. (2006). Correlated label propagation with application to multi-label learning. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2 (pp. 1719–1726).: IEEE.

Kao, A. & Poteet, S. R. (2007). *Natural language processing and text mining*. Springer.

Kouzani, A. Z. & Nasireding, G. (2009). Multilabel classification by bch code and random forests. *International journal of recent trends in engineering*, 2(1), 113–116.

Krishnan, R., Sandhu, R., Niu, J., & Winsborough, W. H. (2009). Foundations for group-centric secure information sharing models. In *Proceedings of the 14th ACM symposium on Access control models and technologies* (pp. 115–124).: ACM.

Ku, L.-W., Liang, Y.-T., & Chen, H.-H. (2006). Opinion extraction, summarization and tracking in news and blog corpora. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 100107.

Kulkarni, S., Singh, A., Ramakrishnan, G., & Chakrabarti, S. (2009). Collective annotation of wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 457–466).: ACM.

Kumar, N., Zhang, L., & Nayar, S. (2008). What is a good nearest neighbors algorithm for finding similar patches in images? In *Computer Vision–ECCV 2008* (pp. 364–378). Springer.

Lee, D. D. & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems* (pp. 556–562).

Leighton, G. & Barbosa, D. (2010). Access control policy translation and verification within heterogeneous data federations. In *Proceedings of the 15th ACM symposium on Access control models and technologies* (pp. 173–182).: ACM.

Li, N. (2011). Discretionary access control. *Encyclopedia of Cryptography and Security*, (pp. 353–356).

Li, N., Wang, Q., Qardaji, W., Bertino, E., Rao, P., Lobo, J., & Lin, D. (2009). Access control policy combining: theory meets practice. In *Proceedings of the 14th ACM symposium on Access control models and technologies* (pp. 135–144).: ACM.

Liaw, A. & Wiener, M. (2002). Classification and regression by randomforest. *R news*, 2(3), 18–22.

Lin, X. & Chen, X.-w. (2010). Mr. knn: soft relevance for multi-label classification. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 349–358).: ACM.

Lindqvist, H. (2006). Mandatory access control. *Master's Thesis in Computing Science, Umea University, Department of Computing Science, SE-901*, 87.

Malin, B., Airoldi, E., et al. (2007). Confidentiality preserving audits of electronic medical record access. *Studies in health technology and informatics*, 129(1), 320.

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

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. Academic press.

McCune, J. M., Jaeger, T., Berger, S., Caceres, R., & Sailer, R. (2006). Shamon: A system for distributed mandatory access control. In *Computer Security Applications Conference, 2006. ACSAC'06. 22nd Annual* (pp. 23–32).: IEEE.

McGraw, R. (2009). Risk-adaptable access control (radac). In *Privilege (Access) Management Workshop. NIST–National Institute of Standards and Technology–Information Technology Laboratory*.

Mihalcea, R. & Csomai, A. (2007). Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* (pp. 233–242).: ACM.

Milne, D. & Witten, I. H. (2008). Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management* (pp. 509–518).: ACM.

Moffett, J., Sloman, M., & Twidle, K. (1990). Specifying discretionary access control policy for distributed systems. *Computer Communications*, 13(9), 571–580.

Molloy, I., Dickens, L., Morisset, C., Cheng, P.-C., Lobo, J., & Russo, A. (2012). Risk-based security decisions under uncertainty. In *Proceedings of the second ACM conference on Data and Application Security and Privacy* (pp. 157–168).: ACM.

Molloy, I., Li, N., Li, T., Mao, Z., Wang, Q., & Lobo, J. (2009). Evaluating role mining algorithms. In *Proceedings of the 14th ACM symposium on Access control models and technologies* (pp. 95–104).: ACM.

Molloy, I., Li, N., Qi, Y. A., Lobo, J., & Dickens, L. (2010). Mining roles with noisy data. In *Proceedings of the 15th ACM symposium on Access control models and technologies* (pp. 45–54).: ACM.

Monte, S. (2010). *Access control based on content*. Technical report, TKK Technical Reports in Computer Science and Engineering, B. TKK-CSE-B10. http://www. cse. tkk. fi/en/publications/B/10/papers/Monte final. pdf.

Murugesan, M., Jiang, W., Clifton, C., Si, L., & Vaidya, J. (2010). Efficient privacy-preserving similar document detection. *The VLDB Journal–The International Journal on Very Large Data Bases*, 19(4), 457–475.

Ni, Q., Lobo, J., Calo, S., Rohatgi, P., & Bertino, E. (2009). Automating role-based provisioning by learning from examples. In *Proceedings of the 14th ACM symposium on Access control models and technologies* (pp. 75–84).: ACM.

NIST (2009). A survey of access control models.

Omohundro, S. M. (1989). *Five balltree construction algorithms*. International Computer Science Institute Berkeley.

Oracle (2012). Oracle database security guide 10g release 2 (10.2).

Osborn, S., Sandhu, R., & Munawer, Q. (2000). Configuring role-based access control to enforce

mandatory and discretionary access control policies. *ACM Transactions on Information and System Security (TISSEC)*, 3(2), 85–106.

Pan, C.-C., Mitra, P., & Liu, P. (2006). Semantic access control for information interoperation. In *Proceedings of the eleventh ACM symposium on Access control models and technologies* (pp. 237–246).: ACM.

Pang, B. & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2), 1–135.

Papadimitriou, C. H. (1981). On the complexity of integer programming. *Journal of the ACM (JACM)*, 28(4), 765–768.

Park, J. S., Sandhu, R., & Ahn, G.-J. (2001). Role-based access control on the web. *ACM Transactions on Information and System Security (TISSEC)*, 4(1), 37–71.

Pedersen, T., Pakhomov, S., McInnes, B., & Liu, Y. (2012). Measuring the similarity and relatedness of concepts in the medical domain: Ihi 2012 tutorial overview. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium* (pp. 879–880).: ACM.

Porter, M. F. (2001). Snowball: A language for stemming algorithms.

Qin, L. & Atluri, V. (2003). Concept-level access control for the semantic web. In *Proceedings of the 2003 ACM workshop on XML security* (pp. 94–103).: ACM.

Rao, P., Lin, D., Bertino, E., Li, N., & Lobo, J. (2009). An algebra for fine-grained integration of xacml policies. In *Proceedings of the 14th ACM symposium on Access control models and technologies* (pp. 63–72).: ACM.

Rao, V. & Jaeger, T. (2009). Dynamic mandatory access control for multiple stakeholders. In *Proceedings of the 14th ACM symposium on Access control models and technologies* (pp. 53–62).: ACM.

Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2011). Classifier chains for multi-label classification. *Machine learning*, 85(3), 333–359.

Reddivari, P., Finin, T., & Joshi, A. (2005). Policy-based access control for an rdf store. In *Proceedings of the Policy Management for the Web workshop*, volume 120 (pp. 78–83).

Rostad, L. & Edsberg, O. (2006). A study of access control requirements for healthcare systems based on audit trails from access logs. In *Computer Security Applications Conference, 2006. ACSAC'06. 22nd Annual* (pp. 175–186).: IEEE.

Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5), 513–523.

Samarati, P. & de Vimercati, S. C. (2001). Access control: Policies, models, and mechanisms. In *Foundations of Security Analysis and Design* (pp. 137–196). Springer.

Sandhu, R. & Chen, F. (1998). The multilevel relational (mlr) data model. *ACM Transactions on Information and System Security (TISSEC)*, 1(1), 93–132.

Sandhu, R. S. (1993). Lattice-based access control models. *Computer*, 26(11), 9–19.

Sandhu, R. S., Coyne, E. J., Feinstein, H. L., & Youman, C. E. (1996). Role-based access control models. *Computer*, 29(2), 38–47.

Scannapieco, M., Figotin, I., Bertino, E., & Elmagarmid, A. K. (2007). Privacy preserving schema and data matching. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data* (pp. 653–664).: ACM.

Sharma, D. M. (2010). On the role of nlp in linguistics. In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground* (pp. 18–21).: Association for Computational Linguistics.

Sun, L., Ji, S., & Ye, J. (2011). Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(1), 194–200.

Tai, F. & Lin, H.-T. (2012). Multilabel classification with principal label space transformation. *Neural Computation*, 24(9), 2508–2542.

Takabi, H. & Joshi, J. B. (2009). An efficient similarity-based approach for optimal mining of role hierarchy. In *Proceedings of the 16th ACM Conference on Computer and Communications Security*.

Tekli, J., Chbeir, R., & Yetongnon, K. (2009). An overview on xml similarity: background, current trends and future directions. *Computer science review*, 3(3), 151–173.

Thomas, R. K., Sandhu, R. S., et al. (1993). Discretionary access control in object-oriented databases: Issues and research directions. In *Proc. 16th National Computer Security Conference* (pp. 63–74).

Thuraisingham, B. (2009). Mandatory access control. In *Encyclopedia of Database Systems* (pp. 1684–1685). Springer.

Toninelli, A., Montanari, R., Kagal, L., & Lassila, O. (2006). A semantic context-aware access control framework for secure collaborations in pervasive computing environments. In *The Semantic Web-ISWC 2006* (pp. 473–486). Springer.

Tran, N. A. & Dang, T. K. (2007). A novel approach to fine-grained content-based access control for video databases. In *Database and Expert Systems Applications, 2007. DEXA'07. 18th International Workshop on* (pp. 334–338).: IEEE.

Tsoumakas, G., Katakis, I., & Vlahavas, I. (2010). Mining multi-label data. In *Data mining and knowledge discovery handbook* (pp. 667–685). Springer.

Tsoumakas, G. & Vlahavas, I. (2007). Random k-labelsets: An ensemble method for multilabel classification. In *Machine learning: ECML 2007* (pp. 406–417). Springer.

Tzelepi, S. K., Koukopoulos, D. K., & Pangalos, G. (2001). A flexible content and context-based access control model for multimedia medical image database systems. In *Proceedings of the 2001 workshop on Multimedia and security: new challenges* (pp. 52–55).: ACM.

Uhlmann, J. K. (1991). Satisfying general proximity/similarity queries with metric trees. *Information processing letters*, 40(4), 175–179.

Upadhyaya, S. (2011). Mandatory access control. In *Encyclopedia of Cryptography and Security* (pp. 756–758). Springer.

Vattani, A. (2011). K-means requires exponentially many iterations even in the plane. *Discrete & Computational Geometry*, 45(4), 596–616.

Wang, H., Huang, H., & Ding, C. (2009). Image annotation using multi-label correlated green's function. In *Computer Vision, 2009 IEEE 12th International Conference on* (pp. 2029–2034).: IEEE.

Winslett, M., Ching, N., Jones, V., & Slepchin, I. (1997). Using digital credentials on the world wide web. *Journal of Computer Security*, 5(3), 255–267.

Winslett, M., Smith, K., & Qian, X. (1994). Formal query languages for secure relational databases. *ACM Transactions on Database Systems (TODS)*, 19(4), 626–662.

Witten, I. & Milne, D. (2008). An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA* (pp. 25–30).

Wu, C. & Chen, Y. (2011). A survey of researches on the application of natural language processing in internet public opinion monitor. In *2011 International Conference on Computer Science and Service System (CSSS)* (pp. 1035–1038).

Yang, L., Ege, R. K., Ezenwoye, O., & Kharma, Q. (2004). A role-based access control model for information mediation. In *Information Reuse and Integration, 2004. IRI 2004. Proceedings of the 2004 IEEE International Conference on* (pp. 277–282).: IEEE.

Yu, T., Srivastava, D., Lakshmanan, L. V., & Jagadish, H. (2002). Compressed accessibility map: efficient access control for xml. In *Proceedings of the 28th international conference on Very Large Data Bases* (pp. 478–489).: VLDB Endowment.

Zhang, L., Ahn, G.-J., & Chu, B.-T. (2002). A role-based delegation framework for healthcare information systems. In *Proceedings of the seventh ACM symposium on Access control models and technologies* (pp. 125–134).: ACM.

Zhang, M.-L. & Zhou, Z.-H. (2007). Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7), 2038–2048.

Zhang, Y. & Schneider, J. G. (2011). Multi-label output codes using canonical correlation analysis. In *International Conference on Artificial Intelligence and Statistics* (pp. 873–882).

Appendix A

The Top 10 Words of Non-Negative Matrix Factorization

Table A.1: The Top 10 Words of Non-Negative Matrix Factorization with 10 Topics

TOPIC	**WORDS**
Topic #1	chemistry,organic,reactions,molecules,chemical,nmr,compounds, department,synthesis,metal
Topic #2	students,science,teachers,mathematics,school,education,program, faculty,courses,laboratory
Topic #3	theory,problems,equations,geometry,mathematical,algebraic, mathematics,differential,study,physics
Topic #4	species,plant,genetic,populations,plants,evolutionary,population, evolution,diversity,variation
Topic #5	data,social,project,information,economic,research,analysis, political,study,policy
Topic #6	research,conference,support,university,workshop,award,international, scientists,program,researchers
Topic #7	protein,cell,proteins,cells,gene,genes,expression,molecular,dna,function
Topic #8	materials,properties,magnetic,optical,high,films,phase,devices,surface, electron
Topic #9	ocean,ice,climate,water,carbon,sea,global,processes,arctic,atmospheric
Topic #10	design,systems,control,system,algorithms,software,performance,computer, engineering,research

Table A.2: The Top 10 Words of Non-Negative Matrix Factorization with 20 Topics

TOPIC	WORDS
Topic #1	students,program,research,faculty,undergraduate,graduate,summer,reu, undergraduates,student
Topic #2	protein,cell,proteins,cells,gene,genes,expression,molecular,dna,function
Topic #3	chemistry,organic,reactions,chemical,compounds,metal,molecules,synthesis, reaction,complexes
Topic #4	theory,geometry,algebraic,mathematics,groups,spaces,geometric,algebra, study,number
Topic #5	laboratory,students,courses,equipment,computer,curriculum,experiments, undergraduate,biology,software
Topic #6	science,teachers,mathematics,school,education,teacher,learning,project, schools,teaching
Topic #7	ice,ocean,climate,water,carbon,sea,arctic,global,atmospheric,circulation
Topic #8	physics,quantum,theoretical,energy,particle,electron,systems,experimental, matter,experiments
Topic #9	research,months,support,university,dr,fellowship,award,sciences,postdoctoral, scientific
Topic #10	problems,equations,methods,models,nonlinear,solutions,differential,numerical, mathematical,algorithms
Topic #11	materials,properties,films,optical,devices,surface,thin,high,phase,polymer
Topic #12	mantle,seismic,earthquake,crust,rocks,deformation,crustal,continental,plate,data
Topic #13	network,college,internet,connection,access,resources,networks,services,nsfnet, information
Topic #14	workshop,participants,held,researchers,scientists,international,workshops,discuss, bring,report
Topic #15	engineering,design,education,technology,science,engineers,manufacturing,industry, mechanical,university
Topic #16	nmr,magnetic,spectrometer,molecules,resonance,nuclear,instrument,instrumentation, spectroscopy,mhz
Topic #17	species,plant,genetic,populations,plants,evolutionary,population,evolution,diversity, variation
Topic #18	systems,design,control,system,performance,research,software,algorithms,development, power
Topic #19	conference,held,international,meeting,travel,support,symposium,scientists,researchers, attend
Topic #20	data,social,project,information,research,economic,political,analysis,policy,study

Table A.3: The Top 10 Words of Non-Negative Matrix Factorization with 50 Topics

TOPIC	**WORDS**
Topic #1	laboratory,courses,students,experiments,undergraduate,curriculum,majors, introductory,exercises,laboratories
Topic #2	students,program,faculty,summer,graduate,undergraduate,reu,student,minority, undergraduates
Topic #3	genes,gene,expression,dna,genetic,regulation,transcription,regulatory,genome, sequences
Topic #4	protein,proteins,binding,structure,membrane,function,enzyme,rna,enzymes, structural
Topic #5	ocean,carbon,water,marine,sea,organic,circulation,production,arctic,pacific
Topic #6	theory,geometry,algebraic,groups,mathematics,spaces,algebra,geometric, algebras,topology
Topic #7	software,computer,parallel,computing,programming,distributed,hardware, computational,tools,simulation
Topic #8	conference,held,international,meeting,travel,support,symposium,scientists, researchers,attend
Topic #9	cell,cells,membrane,growth,cellular,calcium,signaling,receptor,tissue,cycle
Topic #10	chemistry,organic,department,chemical,physical,analytical,division,biochemistry, instrumentation,professor
Topic #11	equipment,support,scientific,vessel,instrumentation,transceivers,acquisition, operated,retrieval,purchase
Topic #12	months,fellowship,support,postdoctoral,mathematical,sciences,fellowships, awards,option, doctoral

Topic #13	problems,algorithms,methods,problem,computational,optimization,efficient, techniques,solution, solving
Topic #14	ice,sea,antarctic,sheet,core,cores,arctic,glacial,west,record
Topic #15	data,analysis,information,database,sets,collection,collected,statistical,project,set
Topic #16	workshop,participants,held,researchers,workshops,discuss,bring,meeting, international,scientists
Topic #17	investigator,principal,proposal,pi,abstract,investigators,study,award,planning,young
Topic #18	species,populations,genetic,evolutionary,population,diversity,variation,evolution, relationships,phylogenetic
Topic #19	magnetic,field,properties,spin,fields,resonance,superconducting,superconductors, temperature,magnetization
Topic #20	physics,particle,nuclear,matter,particles,elementary,theoretical,condensed,atomic, experiments
Topic #21	surface,surfaces,adsorption,microscopy,scanning,analytical,atomic,interfaces, interface,adsorbed
Topic #22	nmr,molecules,spectrometer,nuclear,resonance,spectroscopy,chemists,mhz, structure,studies
Topic #23	social,political,economic,policy,project,study,cultural,public,dissertation, archaeological
Topic #24	reactions,metal,compounds,reaction,complexes,organic,chemical,synthesis, molecules,transition
Topic #25	solar,waves,measurements,energy,wind,atmospheric,plasma,wave,atmosphere, observations
Topic #26	learning,teaching,student,students,education,project,knowledge,skills,technology, concepts
Topic #27	models,model,modeling,statistical,develop,developed,simulation,theoretical,

	mathematical,methods
Topic #28	quantum,theoretical,theory,dynamics,systems,electron,states,mechanics,electronic, electrons
Topic #29	systems,control,system,power,nonlinear,dynamical,performance,adaptive,feedback, dynamic
Topic #30	network,college,internet,connection,access,networks,resources,services,nsfnet, community
Topic #31	teachers,mathematics,school,teacher,schools,project,teaching,middle,education, districts
Topic #32	engineering,education,technology,engineers,mechanical,chemical,electrical,civil, women,industry
Topic #33	phase,technology,business,small,ii,innovation,high,commercial,project,process
Topic #34	mantle,rocks,crust,isotopic,continental,crustal,evolution,samples,ridge,seismic
Topic #35	optical,laser,lasers,optics,devices,light,nonlinear,fiber,spectroscopy,imaging
Topic #36	films,thin,film,growth,deposition,devices,silicon,properties,semiconductor,metal
Topic #37	climate,change,global,lake,records,climatic,variability,record,arctic,regional
Topic #38	science,education,national,scientific,technology,activities,computer,programs, program,sciences
Topic #39	equations,differential,nonlinear,solutions,partial,mathematical,problems,equation, analysis,numerical
Topic #40	contract,abstract,required,contracts,services,support,nsf,agreement,firms,contracting
Topic #41	stars,galaxies,star,stellar,galaxy,formation,evolution,clusters,observations,universe
Topic #42	flow,fluid,flows,transport,numerical,heat,fluids,dynamics,turbulent,turbulence
Topic #43	brain,system,neurons,behavior,neural,visual,nervous,information,mechanisms, animals
Topic #44	plant,plants,soil,growth,arabidopsis,environmental,resistance,nitrogen,responses, host

Topic #45	earthquake,seismic,fault,earthquakes,hazard,structures,deformation,damage, california,ground
Topic #46	research,program,training,award,facilities,activities,projects,undergraduate,center, areas
Topic #47	biology,molecular,biological,dna,genetics,training,techniques,evolutionary, fellowship,cellular
Topic #48	design,manufacturing,process,tools,product,performance,designs,development, optimization,integrated
Topic #49	university,dr,award,collaboration,professor,state,cooperative,expertise,institute, center
Topic #50	materials,properties,polymers,polymer,material,processing,characterization, synthesis,electronic,composites

Table A.4: The Top 10 Words of Non-Negative Matrix Factorization with 100 Topics

TOPIC	**WORDS**
Topic #1	research,training,facilities,award,graduate,area,areas,activities,infrastructure, proposed
Topic #2	students,graduate,student,undergraduate,experience,minority,skills,school,project, careers
Topic #3	chemistry,department,physical,division,chemical,analytical,biochemistry,supported, instrumentation,areas
Topic #4	genes,gene,expression,dna,genetic,regulation,transcription,regulatory,genome, sequences
Topic #5	university,center,state,award,cooperative,california,collaboration,professor, department,carolina
Topic #6	theory,geometry,algebraic,spaces,geometric,topology,algebra,number,algebras, space
Topic #7	solar,measurements,atmospheric,wind,observations,atmosphere,radar,plasma, particles,cloud
Topic #8	laboratory,courses,experiments,undergraduate,curriculum,exercises,majors, laboratories,introductory,lab
Topic #9	conference,held,gordon,conferences,researchers,participants,support,sessions, speakers,attend
Topic #10	teachers,school,teacher,schools,project,middle,districts,year,teaching,elementary
Topic #11	equipment,scientific,transceivers,support,retrieval,satellite,dedicated,acquisition, including,purchase
Topic #12	months,fellowship,support,postdoctoral,sciences,mathematical,fellowships,awards, option,choose

Topic #13	species,evolutionary,phylogenetic,relationships,diversity,morphological,genus, evolution,patterns,ecological
Topic #14	equations,differential,solutions,partial,nonlinear,mathematical,equation,numerical, boundary,elliptic
Topic #15	engineering,engineers,mechanical,education,civil,electrical,chemical,industrial, biomedical,disciplines
Topic #16	reu,summer,projects,site,undergraduates,undergraduate,faculty,experiences, participants,ten
Topic #17	connection,internet,access,network,nsfnet,resources,bits,libraries,supercomputers, midlevel
Topic #18	environmental,environment,sciences,conditions,management,pollution,natural, change,ecological,monitoring
Topic #19	physics,particle,nuclear,matter,theoretical,particles,elementary,condensed,atomic, experiments
Topic #20	methods,method,computational,statistical,techniques,developed,develop,numerical, applied,development
Topic #21	workshop,participants,held,researchers,discuss,bring,workshops,report,future,issues
Topic #22	nmr,molecules,nuclear,resonance,spectroscopy,chemists,mhz,spectrometer,studies, elucidation
Topic #23	learning,teaching,student,concepts,knowledge,interactive,modules,skills,courses, environment
Topic #24	control,nonlinear,adaptive,feedback,controllers,optimal,robust,controller,dynamic, realtime
Topic #25	design,designs,tools,performance,optimization,product,methodology,development, designers,project
Topic #26	plant,plants,arabidopsis,resistance,host,responses,seed,insect,crop,pollen
Topic #27	language,languages,programming,speech,linguistic,japanese,children,american,

	spoken,grammar
Topic #28	mathematics,mathematical,calculus,courses,reform,algebra,sciences,applied, mathematicians,statistics
Topic #29	ice,sea,antarctic,sheet,core,cores,glacial,west,snow,ross
Topic #30	data,sets,collected,set,database,statistical,base,acquisition,collect,analyze
Topic #31	brain,neurons,neural,cells,nerve,nervous,visual,activity,mechanisms,sensory
Topic #32	ocean,circulation,pacific,sea,atlantic,experiment,deep,southern,hydrographic, variability
Topic #33	growth,crystal,rates,rate,crystals,factors,nucleation,conditions,epitaxial, development
Topic #34	archaeological,sites,site,region,mr,remains,conduct,period,ms,societies
Topic #35	metal,complexes,transition,metals,compounds,ligands,clusters,ions,atoms,catalytic
Topic #36	high,temperature,low,pressure,temperatures,thermal,superconductors,performance, measurements,resolution
Topic #37	water,groundwater,deep,column,sea,quality,treatment,waters,vapor,oxygen
Topic #38	biology,biological,training,genetics,fellowship,dna,postdoctoral,physiology, developmental,ecology
Topic #39	cell,cells,cellular,calcium,signaling,cycle,tissue,division,receptor,differentiation
Topic #40	protein,proteins,binding,enzyme,rna,enzymes,function,folding,amino,acid
Topic #41	electron,transfer,energy,microscopy,microscope,electrons,scanning,atomic,charge, scattering
Topic #42	genetic,populations,population,variation,selection,natural,traits,reproductive, individuals,evolutionary
Topic #43	quantum,mechanics,theoretical,classical,theory,dots,states,atoms,matter,spin
Topic #44	parallel,computing,performance,memory,distributed,computational,programming, applications,processors,computation
Topic #45	phase,ii,small,business,innovation,commercial,project,applications,feasibility,gas

Topic #46	systems,dynamical,complex,distributed,biological,techniques,chaotic,hybrid, intelligent,embedded
Topic #47	devices,device,semiconductor,electronic,fabrication,circuits,silicon,mems, integrated,sensors
Topic #48	models,modeling,mathematical,statistical,stochastic,spatial,estimation,simulation, random,empirical
Topic #49	surface,surfaces,adsorption,analytical,adsorbed,interface,interfaces,layer,atomic, scanning
Topic #50	climate,change,global,variability,regional,atmospheric,climatic,records,vegetation, tropical
Topic #51	groups,group,lie,representation,finite,representations,relationships,algebras, symmetries,symmetry
Topic #52	reactions,chemical,reaction,intermediates,kinetics,catalytic,molecules,reactive, oxidation,studies
Topic #53	women,careers,gender,girls,minorities,minority,female,participation,career, academic
Topic #54	polymer,polymers,properties,polymerization,blends,chain,liquid,composites,chains, mechanical
Topic #55	system,nervous,override,components,error,prototype,acquisition,abstract,realtime, monitoring
Topic #56	behavior,experiments,behavioral,males,females,experimental,effects,animals,male, reproductive
Topic #57	dr,award,postdoctoral,work,collaboration,complementary,expertise,professor, scientists,visit
Topic #58	films,thin,film,deposition,properties,coatings,substrates,vapor,diamond,substrate
Topic #59	algorithms,efficient,optimization,algorithm,computational,complexity,graph, combinatorial,develop,techniques

Topic #60	problems,problem,optimization,solving,solution,solve,mathematical,inverse, solutions,optimal
Topic #61	analysis,techniques,tools,analytical,quantitative,harmonic,analyses,include,fourier, operators
Topic #62	college,faculty,community,education,colleges,institutions,programs,courses, curriculum,workshops
Topic #63	investigator,principal,proposal,pi,abstract,investigators,award,young,planning,title
Topic #64	collection,collections,specimens,museum,database,project,resource,history,natural, storage
Topic #65	earthquake,seismic,earthquakes,hazard,damage,fault,ground,reduction,california, national
Topic #66	technology,development,technologies,industry,education,center,institute,technical, technological,project
Topic #67	molecular,molecules,dna,level,techniques,genetics,atomic,evolution,molecule, experimental
Topic #68	waves,wave,nonlinear,propagation,gravity,gravitational,numerical,scattering, phenomena,acoustic
Topic #69	materials,properties,material,characterization,processing,composite,advanced, composites,electronic,synthesis
Topic #70	stars,galaxies,star,stellar,galaxy,formation,evolution,clusters,observations,universe
Topic #71	mantle,seismic,crust,melting,melt,upper,ridge,crustal,isotopic,earth
Topic #72	soil,forest,soils,ecosystem,ecosystems,forests,nitrogen,tropical,nutrient,vegetation
Topic #73	contract,abstract,required,contracts,services,agreement,nsf,support,firms, contracting
Topic #74	marine,microbial,organisms,bacteria,coastal,sea,food,communities,nitrogen, phytoplankton
Topic #75	symposium,meeting,international,scientists,travel,held,support,american,society,

	researchers
Topic #76	oceanographic,vessel,instrumentation,support,rv,operated,vessels,fleet,owned, aboard
Topic #77	carbon,dioxide,nitrogen,production,oxygen,organic,global,dissolved,cycle,flux
Topic #78	optical,laser,lasers,optics,light,fiber,spectroscopy,nonlinear,wavelength,pulses
Topic #79	transport,membrane,membranes,ion,channels,plasma,channel,separations, separation,sediment
Topic #80	organic,synthesis,compounds,synthetic,professor,focus,chiral,molecules,products, macromolecular
Topic #81	arctic,heat,basin,global,polar,warming,abstract,ocean,canadian,shelf
Topic #82	study,results,case,studied,studies,relationship,understanding,important,examine, part
Topic #83	flow,fluid,flows,heat,numerical,fluids,turbulent,turbulence,particle,particles
Topic #84	program,programs,year,minority,nsf,academic,career,support,activities,summer
Topic #85	science,education,national,scientific,activities,foundation,computer,earth,scientists, public
Topic #86	power,electric,energy,transmission,voltage,fuel,generation,electronics,low,electrical
Topic #87	deformation,plate,fault,zone,strain,crustal,continental,faults,gps,rocks
Topic #88	processes,understanding,physical,chemical,biological,fundamental,process, formation,interactions,important
Topic #89	social,political,economic,policy,project,public,cultural,interviews,dissertation, policies
Topic #90	model,modeling,developed,develop,based,results,test,simulations,approach, parameters
Topic #91	magnetic,field,fields,spin,properties,resonance,magnetization,superconducting, superconductors,ferromagnetic
Topic #92	instrument,mass,spectrometer,facility,instrumentation,spectrometry,acquisition,

	samples,resolution,microscope
Topic #93	information,provide,database,management,knowledge,gis,decision,users, geographic,web
Topic #94	sediment,lake,record,isotopic,sediments,history,basin,isotope,samples,cores
Topic #95	computer,software,hardware,tools,graphics,computers,simulation,visualization, workstations,interactive
Topic #96	processing,image,imaging,images,signal,digital,video,visual,vision,motion
Topic #97	manufacturing,process,production,industry,industrial,product,planning,quality, products,machining
Topic #98	dynamics,theoretical,dynamical,computational,simulations,interactions, experimental,population,studies,understanding
Topic #99	structure,structures,properties,structural,function,crystal,diffraction,understanding, studies,determination
Topic #100	network,networks,wireless,communication,traffic,neural,routing,performance, networking,distributed