The task of entity resolution is to identify difference representations referring to the same real-world entity. In many applications, users may input data with multiple representations to the same real-world object and the same representation. For example, for the same author named “Wei Wang,” in the publications, the name can be “Wei Wang,” “Wang Wei,” or “W. Wang.” Multiple authors share name “Wei Wang.” During the processing of such data, it is necessary to identify the representations referring to the same real-world entity.

According to the form of results, entity resolution can be classified to pair-wise entity resolution and group-wise entity resolution. The former one is to obtain pairs of data objects with each pair referring to the same real-world entity. The later is to cluster the data set, such that the data objects in each cluster refer to the same real-world entity.

Entity resolution has wide applications such as information integration, e-commerce, and bibliography management. In Web, the same entity has different representation in different data sources. With entity resolution techniques, some representations can be detected, and the information from different data sources can be used more sufficiently. For example, in Web 2.0, on e-commerce Websites, such as eBay, the same product may have multiple names. For the convenience of users to browse and compare the same products from different buyers, entity resolution is in demand.

For its importance, entity resolution has become one the hottest research topics in the research field of database, machine learning, and natural language processing. Many techniques and applications have been proposed. This book attempts to give an introduction to these techniques and applications.

This book has three objectives. Firstly, this book will summarize the state-of-art techniques of entity resolutions. It provides researchers a reference of entity resolution for their research works. The possible areas where researchers will be interested include database, data quality, information system, and information integration. Since entity resolution is an important content of the course of data quality, database or information system, the second objective of this book is to provide a textbook or reference book for students with majors in computer science, information systems, and management. Entity resolution has wide applications in many systems, such as e-commerce, management information systems, and Web information systems. Therefore, the third objective of this book is to provide a reference book for developers for such systems. In summary, the potential audiences would be educators, information system designers, and graduate students.

In particular, the book is divided into four section.

The first section introduces basic principles of entity resolution and gives an overview of the whole area. The second section proposes entity resolution on various types of data including entity resolution on names, context, XML data, graph databases, and complex networks. Since data management has a
natural relationship with entity resolution, the third section discusses the database techniques and entity resolution including basic operators for entity resolution, data cleaning based on entity resolution, and query processing based on entity resolution. The fourth section focuses on four applications of entity resolution including information integration, bibliography management, e-commerce, and healthcare.

The contents of chapters in this book are introduced as follows.

Entity resolution is one of many importation operations for data quality management, information retrieval, and data management. It has wide applications in Web search, e-commerce search, data cleaning, and information integration. Due to its importance, entity resolution has been studied by researchers in multiple fields including database, machine learning, information retrieval, as well as high performance computation. This book contains a number of chapters, which are carefully chosen in order to discuss the broad research issues in entity resolution. In addition, a number of important applications of entity resolution are also covered in the book. The purpose of chapter 1 is to provide an overview of the concepts, applications, and research topics of entity resolution, as well as the coverage of these topics in this book.

Chapter 2 discusses the measures of entity resolution. This is the base of entity resolution. The goal is to evaluate entity resolution techniques. Traditional measures include precision and recall, and F-measure as the combination of precision and recall. Sometimes, such simple measures are not sufficient. In some cases, the entity resolution results sharing same precision and recall may have significant difference in quality. To distinguish the entity resolution techniques, many new measures are proposed, especially for large data. In this chapter, some traditional measures are discussed with some state-of-the-art entity resolution results evaluation measures.

Chapter 3 surveys entity resolution techniques on names. The original goal of entity resolution is to identify the real-world entities according to the names in the flat text. Thus, this is the basic version of entity resolution. The difficulty of this problem is led by synonym, ambiguity, and typos in the text. This problem is related to natural language processing. Additionally, entity resolution on complex data types such as records and XML data requires entity resolution on names. This problem is related to natural language processing, data management, and even programming language. This problem has been studied for more than 50 years, and many solutions have been proposed. This chapter gives a survey of related techniques. Three kinds of methods are introduced. The first is the similarity measure between strings as fundamental techniques. The second is the transformation methods to handle the difference other than the textual differences. The third is the learning algorithm for string transformation rules for more complex cases.

Chapter 4 proposes context-based entity resolution framework. Names are not sufficient for entity resolution when the names have ambiguity or the data structure is complex. With the consideration that in real applications names are related to each other, the context information in data could be extracted to compare for the entity resolution. Additionally, context-based entity resolution provides an effective approach for entity resolution on mixed data in various types, since context information could be extracted according to the structure of various types of data. Thus, in chapter 4, we present Context-Based Entity Description (CED) to make context information help entity resolution. In our framework, each entity is described by a set of CEDs. During entity resolution, objects are only compared with CEDs to determine the corresponding entity. Additionally, we propose efficient algorithms for CED discovery, maintenance, and CED-based entity resolution. From the experimental results on real datasets, the context-based method outperforms existing methods.

Chapter 5 is a survey for entity resolution on a single relation. Many data are in relational databases due to the success of relational databases in the market. The relational model is also used in many in-
formation integration systems such as federal databases and mediate-wrapper-based systems. Hence, entity resolution on relational databases is in demand. The challenges of entity resolution on relational database include not only the synonym, ambiguity, and typos in attributes but also the determination of importance of them. Many techniques have been proposed. Chapter 5 surveys related work. The basic operation is attribute similarity computation. Based on the attribute similarity computation methods, many techniques for different areas are proposed to fulfill the process of entity resolution. Among the techniques, we focus on record similarity computation, rule-based approach, similarity threshold computation, and blocking. The former three are the framework of entity resolution on relations, and the latter one is the acceleration strategy for entity resolution on a single relation.

Chapter 6 discusses the methods of entity resolution on multiple relations to use join to get more information from multiple relations to support more accurate entity resolution. The complex structure for prime-foreign keys in multiple relations brings difficulty to this problem. Chapter 6 defines the problem of entity resolution on multiple relations. Then the similarity measures and reasonable algorithm are proposed to solve the problem. The efficiency and effectiveness of the proposed algorithm are verified with extensive experiments.

Chapter 7 summarizes state-of-art techniques for entity resolution on XML data. Currently, XML is the standard of the Web, e-commerce, and e-government due of its flexibility. Additionally, since XML data is in semi-structured data model, it has been widely used for data organization in databases and information integration systems. Entity resolution on XML data is helpful to improve database quality in both XML databases and information integration based on XML. The challenge of entity resolution brought by XML is its hierarchy structural information. In chapter 7, we survey entity resolution on XML data, the concrete applications of which include XML document management in highly dynamic applications such as the Web and peer-to-peer systems, detection of duplicate elements in nested XML data, and finding similar identities among objects from multiple Web sources. We survey techniques of pair-wise and group-wise entity resolution for XML data, respectively. We summarize the representations for XML structure and content including as a tree, Bayesian network, and set. We also introduce some well-known entity resolution algorithms for XML data based on these various structures.

Chapter 8 surveys the algorithms for entity resolution for graphs. Many data could be naturally represented as graph, such as organic molecular in chemistry and protein interaction network. Data quality management for graph-structured data requires entity resolution for graphs. The difficulties of entity resolution on graph data set include the measure of similarity between graphs and the computational difficulty in similarity comparison problems for popular measures such as graph edit distance and the largest common subgraphs. In chapter 8, we introduce the distance measures between graphs as well as their applications, which require approximation algorithms for some difficult problems. The approximate graph matching algorithms may be index-based like the NH-Index method, or kernel function-based like G-hash method. Other methods concentrate on providing new definitions of similar graphs that are easier to compute than traditional methods, like the Web-collection method and the Grafil method. To increase the resolution ability of traditional methods, researchers provide some methods to recognize similar graphs, such as graph bounded simulation and p-homomorphism.

Chapter 9 summaries and compares entity resolution algorithms on complex networks. In complex networks, the basic unit for entity resolution is vertex. Entity resolution on a complex network should consider two kinds of information, tags on vertices and the structure of the network. The applications of entity resolution on complex network include the detection of mirror Websites, name recognition in social network, and information searching on the Internet. The difficulties are from the definition of similarity
vertex according to both of structural and content information and make the clustering algorithm suitable for large networks. This chapter will mainly introduce some applications including the detection of mirror Websites and name recognition in social network in detail, as well as node similarity description and clustering method including SimRank, PSimRank, and other related methods.

Chapter 10 presents an entity resolution algorithm on cloud for big data. With the increasing of data, entity resolution is required to be executed on large data sets in many applications. For two reasons, entity resolution is quite suitable to be performed in the cloud. One is that entity resolution is expensive in computation for large data sets. The other is entity resolution could be processed offline for most of the applications. Entity resolution on cloud is not straightforward since it often requires computing the similarities among data objects while cloud computation requires the partition of the data objects. Chapter 10 shows that it is necessary to use wave of strings to compute records similarity in cloud computing and provides a method based on wave of strings of entity resolution. Theoretical analysis and experimental results show that the method proposed in this chapter is efficient and effective.

Chapter 11 introduces basic data operators for entity resolution with implementation algorithms. From the aspect of data management, entity resolution can be considered as a series of basic data operators. Data operators for entity resolution are different from traditional relational data operators in two aspects. One aspect is that those for entity resolution are similar entity resolution due to possible errors in data while relational operators are accurate operators. The other is that the conversion and knowledge should be embedded in the operators for entity resolution since entity resolution requires conversion and knowledge due to ambiguity. The major concern of the basic data operators for entity resolution is efficiency. For efficient implementation of the operators, some database ideas are adopted to them such as index. In chapter 11, we first introduce the solution of similarity search, covering gram-based algorithms and sketch-based algorithms. Then we turn to the solution of similarity join, covering both exact and approximate algorithms. At last, we deal with the problem of clustering similar strings in a set that can be applied to duplicate detection in databases.

Chapter 12 proposes truth discovery methods for data cleaning based on entity resolution. During data cleaning, entity resolution could be applied at first. Then, if the values in the same attribute in different data objects referring to the same real-world entity are different, conflicts occur. To clean the data, it is necessary to resolve the conflicts, which requires truth discovery, which is to find true facts from a large amount of conflicting information provided. The challenge brought by truth discovery is to find sufficient information to determine the truth. In chapter 12, we review state-of-the-art approaches to processing truth discovery including trivial method, fixpoint method, approaches based on copying detection, and semi-supervised learning approaches.

In chapter 13, we propose query-processing strategies based on entity resolution including query semantics, basic data operator, and query optimization algorithms. Organizing the tuples in a database according to the referred real-world entity has three benefits: (1) duplicated tuples are merged together and the storage space and disk IO during query processing are saved; (2) the data quality of original data is identified according to the entity-based organization and the dirty-data tolerant operations could be applied on such data is convenient; (3) the query results on such data are organized based on referred entity and such interface could increase the efficiency for users to use the query results. As a result, it is a good choice to organize data based on referred real-world entity. The semantic of query on entity resolution is different from those in traditional data model. The basic data operators are also different from traditional relational data operators and so are query optimization strategies. Chapter 13 presents EntityManager, a dirty data management system with entity as the basic unit, and keep conflicts in data
as uncertain attributes. Even though the query language is also SQL, the query in the system has different semantics on entity-based organized data. To process queries efficiently, this chapter introduces index structure, data operator implementation, and query optimization algorithms for entity-based data management.

Chapter 14 discusses the application of entity resolution in information integration systems. Information integration brings many data quality problems due to the schema mismatch and ambiguous description in autonomous data sources. Thus, quality assurance mechanism is required in information integration system. During quality assurance in information integration, entity resolution is a basic operation. Entity resolution technique could be applied to different components such as schema mapping and integrated data cleaning. For schema mapping, entity resolution is embedded into the schema matching rules as a function. Entity resolution in integrated data cleaning step should handle the conflict in not only attribute value but also the schema. In chapter 14, we focus on these two steps. To represent the similarity between two records from different data sources with different schemas, the optimal bipartite graph matching is adopted on the attributes of them, and the similarity is measured as the weight of such matching. Based on similarity estimation, the basic idea in this chapter is to estimate the range of the records similarity and to determine whether they are duplicate records according to the estimation. When data integration is performed on XML data, there are many problems because of the flexibility of XML. One of the current implementations is to use entity resolution to carry out the above operations. This chapter proposes the concept of quality assurance mechanisms besides the data integrity and reliability.

In chapter 15, the techniques for entity resolution in bibliography management are proposed. Bibliography information may be collected from various data sources. As a result, bibliography management system may contain different representations for the same reference as well as synonym in the author name. This is a classic problem of entity resolution. Many entity resolution techniques focus on this problem. Chapter 15 introduces EIF as a framework suitable for bibliography information management system. Such framework considers both kinds of confusion, ambiguous and synonym. In this framework, effective clustering techniques, approximate string matching algorithms, and a flexible mechanism of knowledge integration are involved. Extensive experimental results are presented to verify the effectiveness and efficiency of the proposed framework.

Chapter 16 discusses the systems for entity resolution on products in e-commerce. There are huge amounts of commodity data on the e-commerce Websites on the Internet. Such Websites often provide search interface to retrieve suitable products with given keywords. Since the information of products is from autonomous data sources, even various individual users, the same product may have many different descriptions. In some Web 2.0 e-commerce Websites, such as eBay or Taobao, such descriptions may have even larger diversity. To make users browse and use such retrieved products information effectively, if the products are classified according to referred real-world entity, the efficiency of purchasing could be improved. Due to frequently missing or wrong values, and subjective difference in description, traditional method of entity resolution may not have a good result on e-commerce data. Therefore, in this chapter a set of algorithms are proposed in data cleaning, attribute and value tagging, and entity resolution, which are specialized for e-commerce data. Additionally, user’s actions are collected to improve the classification result. We evaluate the effectiveness of the proposed algorithms with real-life datasets from e-commerce sites. The experimental results demonstrate that the proposed method is effective and suitable for improve the experiences for users.

Chapter 17 discusses the entity resolution techniques for healthcare information management systems. In healthcare information management systems, some medical documents may contain data quality
problems. A significant example is that the names for the same patients may be different in various documents, especially for the documents from different data sources. For effective management of medical documents in healthcare information management systems, entity resolution should be applied. In this section, some techniques for entity resolution on medical documents, especially for Word Sense Disambiguation are described. We compare the effectiveness of a variety of knowledge sources of WSD in the biomedical domain. These include features that have been commonly used for WSD of general text as well as information derived from domain-specific resources. One of these features is MeSH terms, which we find to be particularly effective when combined with generic features.

In a nutshell, the book provides a comprehensive summary from both the algorithmic and the applied perspectives. It will provide the reader with a better understanding of how entity resolution on data can be efficiently and effectively performed for different applications.

Hongzhi Wang
Harbin Institute of Technology, China