Chapter 2
SDM Principles

The spatial data mining (SDM) method is a discovery process of extracting generalized knowledge from massive spatial data, which builds a pyramid from attribute space and feature space to concept space. SDM is an interdisciplinary subject and is therefore related to, but different from, other subjects. Its basic concepts are presented in this chapter, which include manipulating space, SDM view, discovered knowledge, and knowledge representation.

2.1 SDM Concepts

SDM aims to improve human ability to extract knowledge and insights from large and complex collections of digital data. It efficiently extracts previously unknown, potentially useful, and ultimately understandable knowledge from these huge datasets for a given task with constraints (Li et al. 2001, 2006, 2013; Wang 2002; Han et al. 2012; Wang and Yuan 2014). To implement a credible, innovative, and interesting extraction, the SDM method not only relies on the traditional theories of mathematical statistics, machine learning, pattern recognition, neural networks, and artificial intelligence, but it also engages new methods, such as data fields, cloud models, and decision trees.

2.1.1 SDM Characteristics

SDM extracts abstract knowledge from concrete data. The data are explicit while, in most circumstances, the knowledge is implicit. The extraction may be a repeated process of human–computer interactions between the user and the

D. Li et al., *Spatial Data Mining*, DOI 10.1007/978-3-662-48538-5_2

dataset. The original data are raw observations of spatial objects. Because the data may be dirty, they need to be cleaned, sampled, and converted in accordance with the SDM measurements and the user-designated thresholds. The discovered knowledge consists of generic patterns acting as a set of rules and exceptions. The patterns are further interpreted by professionals when they are utilized for data-referenced decision-making with various requirements.

SDM's source is spatial data, which are real, concrete, and massive in volume. They already exist in the form of digital data stored in spatial datasets, such as databases, data markets, and data warehouses. Spatial data may be structured (e.g., instances in relational data), semi-structured (e.g., text, graphics, images), or non-structured (e.g., objects distributed in a network). Advances in acquisition hardware, storage capacity, and Central Processing Unit (CPU) speeds have facilitated the ready acquisition and processing of enormous datasets, and spatiotemporal changes in networks have accelerated the velocity of spatial data accumulation. Noises and uncertainties exist in spatial data (e.g., errors, incompleteness, redundancy, and sparseness), which can cause problems for SDM; therefore, polluted spatial data are often preprocessed for error adjustment and data cleaning.

SDM aims to capture knowledge. The knowledge may be spatial or non-spatial; it is previously unknown, potentially useful, and ultimately understandable under the umbrella of spatial datasets. This knowledge can uncover the description and prediction of the patterns of spatial objects, such as spatial rules, general relationships, summarized features, conceptual classification, and detected exception. These patterns are hidden in the data along with their internal relationships and developing trends. However, SDM is a much more complex process of selection, exploration, and modeling of large databases in order to discover hidden models and patterns, where data analysis is only one of its capabilities. SDM implements high-performance distributed computing, seamless integration of data, rational knowledge expression, knowledge updates, and visualization of results. SDM also supports hyperlink and media quotas among hierarchical document structures when mining data.

The SDM process is one of discovery instead of proofing. Aided by SDM human–computer interaction, the process is automatic or at least semi-automatic. The methods may be mathematic or non-mathematic, and the reasoning may be deductive or inductive. The SDM process has the following requirements. First, it is composed of multiple mutually influenced steps, which require repeated adjustment to spiral up in order to extract the patterns from the dataset. Second, multiple methods are encouraged, including natural languages, to present the process of discovery and its results. SDM brings together all of the available variables and combines them in different ways to create useful models for the business world beyond the visual representation of the summaries in online analysis and processing (OLAP) applications. Third, SDM looks for the relationships and associations between phenomena that are not known beforehand. Because the discovered knowledge usually only needs to answer a particular spatial question, it is not necessary to determine the universal knowledge, the pure mathematical formula, or a new scientific theorem.

2.1.2 Understanding SDM from Different Views

SDM is a repeated process of spatial data-referenced decision-making. It can be understood best from the following five perspectives:

(1) **Discipline**. SDM is an interdisciplinary subject that matches the multidisciplinary philosophy of human thinking and suitably deals with the complexity, uncertainty, and variety present when informing data and representing rules. SDM is the outcome in the stage at which some technologies develop, such as spatial data access technology, spatial database technology, spatial statistics, and spatial information systems; therefore, it brings together the fruits of various fields. Its theories and techniques are linked with data mining, knowledge discovery, database systems, data analysis, machine learning, pattern recognition, cognitive science, artificial intelligence, mathematical statistics, network technology, software engineering, etc.

(2) **Analysis**. SDM discovers unknown and useful rules from huge amounts of data via a set of interactive, repetitive, associative, and data-oriented manipulations. It mainly utilizes certain methods and techniques to extract various patterns from spatial datasets. The discovered patterns describe the existent rules or predict a developing trend, along with the certainty or credibility to measure the confidence, support, and interest of the conclusions derived from analysis. They can help users to make full use of spatial repositories under the umbrella of various applications, stressing efficient implementation and timely response to user commands.

(3) **Logic**. SDM is an advanced technique of deductive spatial reasoning. It is discovery, but not proofing, in the context of the mined data. As a part of deductive inference, it is a special tool for spatial reasoning that allows a user to supervise or focus on discovering the rules of interest. The reasoning can be automatic or semi-automatic. Induction is used to discover knowledge, while deduction is used to evaluate the discovered knowledge. The mining algorithms are a combination of induction and deduction.

(4) **Actual object operation**. The data model can be hierarchical, network, relational, object-oriented, object-related, semi-structured, or non-structured. The data format may be vector, raster, or vector-raster spatial data. The data repositories are file systems, databases, data markets, data warehouses, etc. The data content may involve locations, graphics, images, texts, video streams, or any other data collections organized together, such as multimedia data and network data.

(5) **Sources**. SDM is implemented on original data in databases, cleaned data in data warehouses, detailed commands, information from users, and background knowledge from applicable fields, which include the raw data provided by spatial databases and the corresponding attribute databases or the manipulated data stored in the spatial data warehouse, the advanced instructions sent by the user to the controller, and the various expert knowledge of different fields stored in the knowledge base. With the help of the network, SDM can

break down the local restrictions of spatial data, using not only the spatial data in its own sector, but larger scopes or even all of the data in the field of space and related fields. It also can be available to discover more universal spatial knowledge and to implement spatial online data mining (SOLAM). To meet the needs of decision-making, SDM makes use of decentralized heterogeneous data sources, with timely and accurately extracted information and knowledge through data analysis using the query and analysis tools of the reporting module.

2.1.3 Distinguishing SDM from Related Subjects

SDM's interdisciplinary nature integrates machine learning, computer visualization, pattern recognition, statistical analysis, database systems, and artificial intelligence. SDM and its related disciplines are related but different.

SDM is a branch of data mining with spatial characteristics. The general mining object of SDM is a conventional structured relational database. SDM objects are spatial datasets in which there are not only attribute and location data (e.g., maps, remote sensing images, and urban spatial planning), but also spatial relations and distance data as well. Unstructured spatial graphics and images may be vector or raster in a number of layers. Moreover, its spatial data storage structures, query methods, data analysis, and database operations are different from a conventional database. As for granularity, data mining is transactional data; and SDM data may be a point, line, polygon, pixel, or tuple. Taking a vector object as granularity, SDM can use the location, form, and spatial correlation of spatial objects to discover knowledge. Taking a raster pixel as granularity, SDM can use the pixel location, multi-spectral value, elevation, slope, and other information to extract image features for fine to coarse granularity. As to scale, SDM adds scale dimension to represent the geometric transformation of spatial data from large scale to small scale. The larger the scale is, the finer are the spatial patterns of the presented object.

Machine learning obtains or reproduces information via a training computer, focusing on the design and development of new algorithms as empirical input. It is a training technique via analysis so the specific data prepared for machine learning need not have significance in the real world (Witten and Frank 2000). In machine learning, pattern recognition is the assignment of a label to a given input value. According to the output, pattern recognition may be supervised or unsupervised. Based on the known properties learned from the training data, machine learning and pattern recognition pay attention to prediction. SDM is an extracted process via interacted discovery, such as task understanding, data conversion, data cleaning, dimension reduction, and knowledge interpretation. SDM also highlights the discovery of the unknown properties of the data.

Artificial intelligence is the academic basis of data mining generation and is based on the study of how human beings acquire and use knowledge. Mainly

based on interpretation, artificial intelligence is a positive step toward understanding the world. While SDM uses machines to simulate human intelligence for the discovery of knowledge from data, machines also are used to reproduce human understanding, which is the reverse way to understanding the world. The artificial intelligence of space-based data mining systems, which have human-like thinking and cognitive ability, can discover new knowledge to complete a new task. Discovering knowledge through SDM is making progress as far as constructing expert systems and generating knowledge bases to achieve a new entity model for the cognitive science of artificial intelligence.

SDM has the highest information capacity and is the most difficult to implement. Query and reporting have the lowest information capacity and are the easiest to implement. Query and reporting tools help explore data at various levels. Data retrieval extracts interesting data and information from archives and databases with preset criteria by using preliminary statistical analysis. The query and reporting tools describe what a database contains; however, OLAP, which can create multidimensional reports, is used to explain why certain relationships exist. This suggests a trade-off between information capacity and ease of implementation. The user-made hypothesis about the possible relationships between the available variables is checked and confirmed by analyzing a graphical multi-dimensional hypercube from the observed data. Unlike SDM, the research hypotheses are suggested by the user and are not uncovered from the data. Furthermore, the extrapolation is a purely computerized procedure, and no use is made of modeling tools or summaries provided by the statistical methodology. OLAP can provide useful information for databases with a small number of variables; however, problems arise when there are tens or hundreds of variables. Then, it becomes increasingly difficult and time-consuming to find a good hypothesis and analyze the database with OLAP tools to confirm or deny it. OLAP and SDM are complementary; when used together, they can create useful synergies. OLAP can be used in the preprocessing stages of data mining, which makes understanding the data easier because it becomes possible to focus on the most important data, identifying special cases, or looking for principal interrelations. The final data mining results, expressed using specific summary variables, can be easily represented in an OLAP hypercube. As a web-based authentication of SDM, SOLAM supports multi-dimensional data analysis, verifying the set assumptions under the guidance of users (Han et al. 2012).

2.1.4 SDM Pyramid

In accordance with its basic concept, SDM's process includes data preparation (understanding the prior knowledge in the field of application, generating target datasets, cleaning data, and simplifying data), data mining (selecting the data mining functions and algorithms; searching for the knowledge of interest in the form of certain rules and exceptions: spatial associations, characteristics, classification, regression, clustering, sequence, prediction, and function dependencies), and

post-processing of data mining (interpretation, evaluation, and application of the knowledge). During the SDM process, every move to the next stage deepens the awareness and understanding of the spatial entities, transforming the spatial data into information and then into knowledge. The more abstract, coherent, and general the description is, the more advanced the technologies need to be.

As a result, when explaining the concept of data mining, Piatetsky-Shapiro (1994) proposed the concept of the Data, Information, and Knowledge Pyramid (DIKP). However, the pyramid merely distinguished the specific concept elements of the different levels in data mining and failed to make an association between the concepts of data mining with the process. Han et al. (2012) depicted the process of data mining visually by different graphics, but only the process of data mining was emphasized; therefore, they failed to clearly illustrate the role of all the other elements and their distinctions. Because spatial data are far more complex than common data, it was necessary to combine DIKP and the process of SDM in order to clearly explain the concept and roles of SDM. After they were combined, clearly the SDM pyramid was more specific and complete (Fig. 2.1).

It can be seen in Fig. 2.1 that under the effects of external forces, such as access to the spatial data, storage of the spatial data, network sharing, calibration of target data, data cleaning, data mining, interpretation, and evaluation, the entities of the

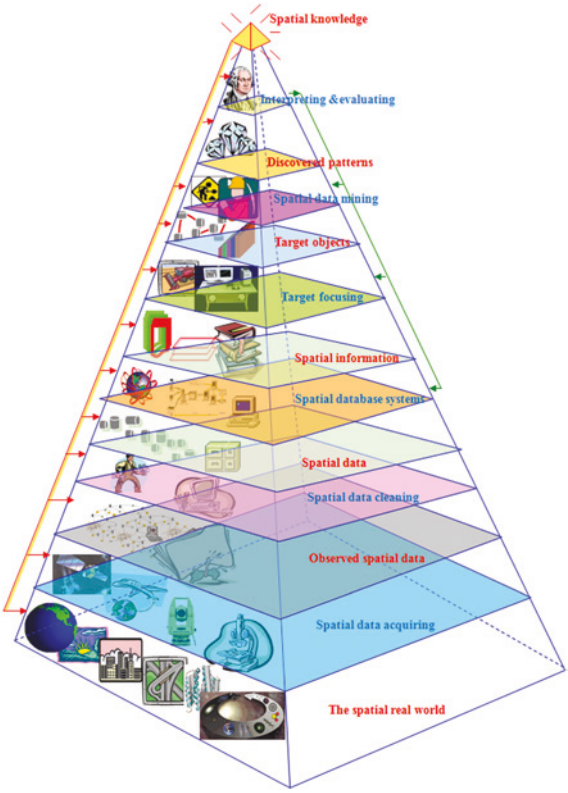

Fig. 2.1 SDM pyramid

SDM pyramid more closely matched the reality of the physical world experience as far as the spatial concept, spatial data, spatial information, changes in spatial size, and increases in spatial scale, which finally become spatial knowledge. In different disciplines, the definitions of some basic concepts in Fig. 2.1, such as spatial data, spatial information, and spatial knowledge, are probably different.

2.1.5 Web SDM

The Web is an enormous distributed parallel information space and valuable information source. First, it offers network platform resources, such as network equipment, interface resources, computing and bandwidth resources, storage resources, and network topology. Second, a variety of data resources use the platform as a carrier, such as text, sound, video data, network software, and application software. The rapid development of network technology provides a new opportunity for a wide range of spatial information sharing, synthesis, and knowledge discovery.

While the high level of network resources makes the information of the network suitable, user-friendly, general, and reusable, how can these distributional, autonomous, heterogeneous data resources be used to obtain, form, and use the required knowledge in a timely manner? With the help of a network, the local restrictions are brought to the spatial dataset and can make use of not only the internal data of a department, but also on a greater scale or even all of the spatial data in the field of space or space-related fields. The discovered knowledge thus becomes more meaningful. Because of the inherent open, distributed, dynamic, and heterogeneous features of a network, it is difficult for the user to accurately and quickly obtain the required information.

Web mining is the extraction of useful patterns and implicit information from artifacts or activities related to the World Wide Web under the internet (Liu 2007). As the internet broadens the available data resources, web mining may include web-content mining, web-structure mining, and web-usage mining. Web-content mining discovers knowledge from the content of web-based data, documents, and pages or their descriptions. Web-structure mining uncovers the knowledge from the structure of websites and the topological relationships among different websites (Barabási and Albert 1999). Web-usage mining extracts Web-user behavior or modeling and predicting how a user will use and interact with the Web (Watts and Strogatz 1998). One of the most promising mining areas being explored is the extraction of new, never-before encountered knowledge from a body of textual sources (e.g., reports, correspondence, memos, and other paperwork) that now reside on the Web. SDM is one of those promising areas. Srivastava and Cheng (1999) provided a taxonomy of web-mining applications, such as personalization, system improvement, site modification, business intelligence, and usage characterization. Internet-based text mining algorithms and information-based Web server log data mining research are attracting increasingly more attention. For example,

some of the most popular Web server log data are growing at a rate of tens of megabits every day; discovering useful model, rules, or the visual structures from them is another area of research and application of data mining. To adapt to the distributed computing environment of networks, SDM systems should also make changes in the architecture and data storage mode, providing different levels of spatial information services such as data analysis, information sharing, and knowledge discovery on the basis of display, service, acquisition, and storage of spatial information on the distributed computing platform. SOLAM and OLAP in multidimensional view with SDM on a variety of data sources (Han et al. 2012), stress the efficient implementation and timely response to user commands for more universal knowledge.

In the course of the development of a network, users come to realize that the internet is a complex, non-linear network system. Next-generation internet will gradually replace the traditional internet and become the information infrastructure of the future by integrating existing networks with new networks that may appear in the future. As a result, the internet and its huge information resources will be the most important basic strategic resource of a country. By efficiently and reasonably using those resources, the massive consumption of energy and material resources can be reduced and sustainable development can be achieved. The authors believe that networked SDM must take into account the possible space, time, and semantic inconsistencies in the databases distributed in networks, as well as the difference in spatiotemporal benchmarks and standards for the semantics. We must make use of spatial information network technology to build a public data assimilation platform on these heterogeneous federal spatial databases. Studies are underway to address solutions to these issues.

2.2 From Spatial Data to Spatial Knowledge

SDM is a gradual process of sublimation from data to knowledge, experiencing spatial numerical values, spatial data, spatial information, and spatial knowledge (see Fig. 2.1). These basic concepts are distinguishable and related. Based on geo-spatial information science, SDM meanings are assigned to these concepts, referencing the existing interdisciplinary definitions from many literature sources (Li et al. 2001; Frasconi et al. 1999; Han et al. 2012).

2.2.1 Spatial Numerical

A spatial numerical value is a numerical value with a measurement unit, which assigns spatial meanings to numbers. The number may be a natural number (1, 2, 3…), a rational number (−1, 0, 6.2, …), or an irrational number (e, π, …). The role of numbers is public; for example, the number 6,000 may represent a

land area of 6,000 m^2, a monthly salary of $6,000, or a distance of 6,000 km. When a number is restricted to a specific meaning, a numerical value appears. Furthermore, when it is restricted in the field of spatial information, a numerical value becomes the symbol to represent a spatial numerical value privately. For example, the number 6,000 is restricted to representing the distance of 6,000 km. Once a numerical value becomes a spatial numerical value, it is distinguished from the common number. The privatized numerical value is able to characterize spatial objects. In the process of verification of various theories and algorithms on computers, a variety of spatial numerical values are actually computerized via numbers. The unit of measurement is used to assign different spatial meanings. A spatial numerical value is a carrier when the objects are collected, transformed, stored, and applied in a computer system.

2.2.2 Spatial Data

Spatial data are important references that help humans to understand the nature of objects and utilize that nature by numerically describing spatial objects with the symbol of attributes, amounts, and positions and their mutual relationships. Spatial data can be numerical values, such as position elevation, road length, polygon coverage, building volume, and pixel grayscale; character strings, such as place name and notation; or multimedia information, such as graphics, images, videos, and voices. Spatial data are rich in content from the microcosmic world to the macroscopic world, such as molecular data, surface data, and universal data. Compared to common data, spatial data contain more specifics, such as spatiotemporal change, location-based distribution, large volume, and complex relationships. In spatial datasets, there are both spatial data and non-spatial data. Spatial data describe a geographic location and distribution in the real world, while non-spatial data consist of all the other kinds of data. Sometimes, a spatial database is regarded as a generic database—one special case of which is a common database. Spatial data can be divided into raw data and processed data or digital data and non-digital data. Raw data may be numbers, words, symbols, graphics, images, videos, language, etc. Generally, SDM utilizes digital datasets after they are cleaned.

2.2.3 Spatial Concept

A spatial concept defines and describes a spatial object, along with its connotation and extension. Connotation refers to the essence reflected by the concept, while extension refers to the scope of the concept. Generally, the definition of spatial concept and its interpretation of connotation and extension are applied to explain the phenomena and the states of spatial objects, as well as to resolve problems

in the real world. Take the concept of remote-sensing image interpretation as an example. The connotation of "green" refers to the type of land covered by green plants, whose gray value of low-level eigenvalue is in a particular range. Its extension includes all the image pixels whose gray values are in that particular range. The scope of concept extension will decrease if its connotation scope increases. For example, if the attribute of "texture characteristics" is added to the connotations of "green," its extension will be greatly reduced. As such, it only refers to the greens that have certain texture features, such as grassland, woodland, and forest land. The concept is hierarchical. Although "green" is a concept and "grassland," "forest land," or "woodland" are also concepts, these concepts are at different hierarchies. The concept layer of "forest land" is lower than that of "green" but higher than that of "woodland."

Spatial objects represented by spatial concepts sometimes are uncertain (e.g., random and ambiguous). Conceptual space refers to the universe of discourse on the same kind of concept. When discussing the different meanings of a conceptual variable, it is necessary to specify their connotation and extension, as well as their mutual similarity or affiliation in the universe of discourse. For example, a variable of "distance" may be "3000 km or so," "around a 12-h train journey," or "very far." In this way, the technology supported by a spatial concept can be meaningful and therefore leap from the unknown world to the known world. Spatial concepts should be defined along with granularity and scale. The same concept can have different spatial meanings of granularity under a different scale, and scale reflects the scaling of the concept granularity. For example, the concept of "the distance from Wuhan to Beijing is 1,100 km" is a close distance on global scale but a long distance on the scale of Hubei Province, China.

2.2.4 Spatial Information

Spatial information characterizes the features of spatial objects in the form of a summarizing dataset. It is the product of data processing that provides a new result and fact in the storage, transmission, and conversion of object data. Wiener, the founder of cybernetics, stated that "information is the content of exchange between human beings and the outer world in the process of their interaction." Shannon, a founder of information theory, argued that "information is what is constant in the process of reversible re-encoding on communication." Philosophically, information indicates the ways of existence and the moving laws of substances, spreading from person to person for communication and interaction. Sometimes, information is treated as a resource to directly help decision-makers solve problems. When it supports the decision-makers only after reprocessing, data are still data or a piece of a message but not information. Spatial information provides useful interpretation to eliminate spatial uncertainty in a meaningful form and order after data processing in specific environments. Incomplete information results in

uncertainty, while complete information results in determination. For example, "South Lake" may remind us of various information, such as "the South Lake in Wuhan City," "the South Lake in Nanning," "the South Lake Airport" and "the area of South Lake," etc. because of incomplete information. After more specific information is added, it is clearly known the information is referring to "the closed South Lake Airport in Wuhan."

A variety of geometric features and spectrum features are extracted from remotely sensed images. Thematic information from investigation and observation can create a multisource and multi-dimensional thematic map of a substantial amount of thematic information for one location, such as location, shape, size, quality, height, slope, noise, pollution, transportation, land cover, moisture, distribution, and relationships of surface objects such as rivers, hills, forests, and soil. The same data under a different background may represent different information; for example, "600 m^2" can be either the area of a land parcel or the construction area of a building. Different data under the same background can represent the same information; for example, "around 1,100 km," "around a 12-h train journey," "far away," etc. can all be information about "distances from Beijing to Wuhan." The quality of spatial information can be represented by a percentage or a rating description of "excellent, good, common, or poor," which makes it possible to sublimate SDM concepts. Spatial information should be collected and updated in a timely manner because its change is dynamically spatiotemporal, such as the rise and fall of sea water, the movement of the desert, the melting of a glacier, and a change of land use. The position identity of spatial information is linked with spatial data. Because satellite positioning and remote sensing systems can uninterruptedly access spatial data for a long period of time, they are and will continue to be an effective and convenient way to collect spatial information.

2.2.5 Spatial Knowledge

Spatial knowledge is a useful structure for associating one or more pieces of information. The spatial knowledge obtained by SDM mainly includes patterns such as correlation, association, classification, clustering, sequence, function, exceptions, etc. As a set of concepts, regulations, laws, rules, models, functions, or constraints, they describe the attributes, models, frequency, and cluster and predict the trend discovered from spatial datasets, such as an association of "IF the road and river intersect, THEN the intersection is a bridge on the river with 80 % possibility." Spatial knowledge is different from isolated information, such as "Beijing, the capital of China." In many practical applications, it is not necessary to strictly distinguish information from knowledge. It should be noted that data processing to obtain professional information is a basis of SDM but is not equal to SDM, such as image processing, image classification, spatial query, and spatial analysis. The conversion from spatial data to spatial information—a process of data

processing—can be achieved by a database system. The conversion from spatial information to spatial knowledge is another cognition process, along with human–machine interaction and specific SDM techniques. The value of information and knowledge is that it can be converted into productive power or used to extract new information and to uncover new knowledge. Spatial information that is not being used, despite being meaningful, is of no value. Only if it is accumulated systematically with a purpose can it become knowledge, just like commodities that are not exchanged have no value.

2.2.6 Unified Action

Spatial numerical values act as digital carriers during the process of object collection, object transmission, and object application. Objects in the spatial world are first transferred to forms of data by macro- and micro-sensors and other equipment based on a certain model of the concept models, approximately described according to some theories and methods, and finally stored as physical models in the physical medium (e.g., hard drives, disks, tapes, videos) of a database in a spatial information system (e.g., GIS) or as a separate spatial database. In fact, during the process of data mining, the numerical data participate in the actual calculations, and the unit of measurement in a certain space is only used to give those numerals different spatial meanings.

A spatial numerical value is a kind of spatial data, while spatial data are the spatial information carriers, which refers to the properties, quantities, locations, and relationships of the spatial entities represented by spatial symbols, such as the numerical values, strings, graphics, images, etc. Spatial data represent the objects, and spatial information looks for the content and interpretation of spatial data. Spatial information is the explanation of the application values of spatial data in a specific environment, and spatial data are the carrier of spatial information. The conversion from spatial objects to spatial data, and then to spatial information, is a big leap in human recognition. The same data may represent different information on different occasions, while different data may represent the same information on the same occasion when closely related to the spatial data. For example, “the landslide displaces 20 mm southward” is spatial data, while “the landslide displaces about 20 mm southward” is a spatial concept.

Spatial data are also the key elements to generate spatial concepts. A spatial concept is closely related to spatial data; for example, “around 3,000 km” is a spatial concept, but “3,000 km” is spatial data. The conversion between spatial concept and spatial data is the cornerstone of the uncertain conversion between the qualitative and the quantitative. A spatial concept is the description and definition of a spatial entity. It is an important method used to represent spatial knowledge. Although spatial data and spatial information are limited, spatial knowledge is infinite. The applicable information structure that is formed by one or more pieces

of associated information is spatial knowledge, which is the relationship among spatial entities on different cognitive levels with a high degree of cohesion and distillation of the spatial data and spatial information. It is more general and abstract and can be used directly by users. Spatial knowledge has different spatial meanings of the granularity on different scales. Different spatial concepts of granularity should be defined under a specific scale.

2.3 SDM Space

With the generalization of a large spatial dataset, SDM runs in different spaces, such as attribute space to recognize attribute data, feature space to extract features, and concept space to represent knowledge. SDM organizes emergent spatial patterns according to data interactions by using various techniques and methods. Fine patterns are discovered in the microcosmic concrete space, and coarse patterns are uncovered in the macrocosmic abstract space.

2.3.1 Attribute Space

Attribute space is a raw space that is composed of the attributes to depict spatial objects. An object in human thinking is described in a nervous system with an entity (e.g., case, instance, record, tuple, event, and phenomenon), and an object in SDM is represented as an entity with attributes in a computerized system. The dimension of attribute space comes from the attributes of spatial objects. Spatial objects with multiple attributes create a multi-dimensional attribute space. Every attribute in the attribute space links to a dimensional attribute of the object. When it is put into attribute space, an object becomes a specific point with a tuple of its attribute data in each dimension (e.g., position, cost, benefit). Thousands of spatial objects are projected to the attribute space as tens of thousands of points. In the context of specific issues, attribute data are primitive, chaotic, shapeless accumulations of natural states, but they are also the sources to generate order and rules. Going through the disorganized and countless appearance of attribute data, SDM uncovers the implied rules, orders, outliers, and relevance.

2.3.2 Feature Space

Feature space is a generalized space that highlights the object essence on the basis of attribute space. A number of different features of a spatial entity create a multi-dimensional feature space. The feature may be single attribute, composite attribute, or derived attribute. It gradually approximates the nature of a great deal of

objects with many attributes. The more generalized the dataset, the more abstract the feature is. For example, feature-based data reduction is an abstraction of spatial objects by sharply declining dimensions and greatly reducing data processing. Based on the features, attribute-based object points are freely spaced, the whole of which generates several object groups. The objects in the same group are often characterized with the same feature. With gradual generalization, datasets are summarized feature by feature, along with the distribution change of objects in feature space. Diverse distributions result in various object compositions and even re-compositions. The combination varies with different features in the context of the discovery task. Jumping from attribute space to feature space, attribute-based object points are summarized as feature-based object points or clusters. When further jumping from one microscopic feature space to another macroscopic feature space, object points are partitioned into more summarized clusters. Furthermore, the feature becomes more and more generic until the knowledge is discovered (i.e., clustering knowledge).

2.3.3 Conceptual Space

Conceptual space may be generated from attribute space or feature space when concepts are used in SDM. The objective world involves physical objects, and the subjective world reflects the characteristics of the internal and external links between physical objects and human recognition. From the existence of the subjective object to the existence of self-awareness, each thinking activity targets a certain object. The concept is a process of evolution and flow on objects, linking with the external background. When there are massive data distributed in the conceptual space, various concepts will come into being. Obviously, concepts are more direct and better understood than data. Reflecting the intention and extension, all kinds of concepts create a conceptual space. All the data in the conceptual space contributes to a concept. The contribution is related to the distance between the data and the concept and the value of the data. The greater the value of the data is, the larger the contributions of the data are. Given a set of quantitative data with the same scope of attributes or features, how to generalize and represent the qualitative concepts is the basis of knowledge discovery. By combining and recombining the basic concepts in various ways, the cognitive events further uncover the knowledge. In conceptual space, various concepts to summarize the object dataset show different discovery states.

2.3.4 Discovery State Space

Discovery state space is a three-dimensional (3D) operation space for the cognition and discovery activity (Li and Du 2007). Initially, it is composed of the

attribute-oriented dimension, entity-oriented dimension, and template-oriented dimension, in terms of the breadth and depth of data mining. The attribute-oriented operation is for the relationships and rules among various attributes, the entity-oriented operation is for the consistency and difference in the multi-attribute model of entities, and the template-oriented operation is for knowledge discovery that generalizes or details the knowledge, changing from a microscopic knowledge template to a macroscopic template. Both the attribute-oriented operation and the entity-oriented operation create a two-dimensional knowledgebase, which are the operations for a specific knowledge template. A template-oriented operation addresses the attributes and the entities as a whole and promotes the degree of knowledge abstraction under induction. When the degree of abstraction increases, the generality in the dimension of both the attribute and the entity increases and the physical size of the knowledge template becomes increasingly smaller. When introduced into geo-spatial information sciences, scale dimension is appended in discovery state space, and as a result, forms a four-dimensional discovery state space. A scale-oriented operation addresses the measurement transformation of knowledge discovery when there are various scales, such as 1:500, 1:1,000, and 1:10,000.

In addition to the scale of measurement, the discovery state space is also related to granularity in resolution and hierarchy in cognition. In fact, an increase in the degree of abstraction is an increase of granularity size, a decrease in the measuring scale, and/or the increase of the degree of induction of the cognition hierarchy. Specifically, if each dimension of attributes and entities in the knowledge-template tends toward generalization and the physical size of the knowledge template decreases, then the data of the spatial entity are making combinations or enrichment differently (Fig. 2.2) according to their different tasks in attribute space, conceptual space, or feature space.

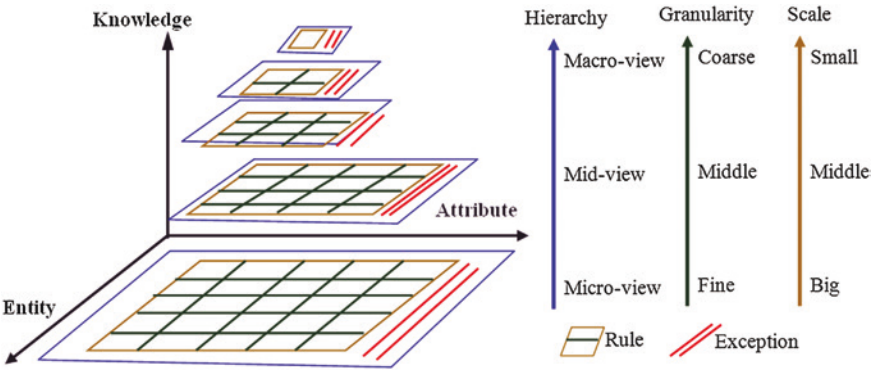

Fig. 2.2 Discovery state space = {attribute space → conceptual space → feature space | cognitive level (granularity and/or scale)}. Reprinted from Wang and Shi (2012), with kind permission from Springer Science+Business Media

2.4 SDM View

SDM view is a mining perspective which assumes that different knowledge may be discovered from the same spatial data repositories for various mining purposes. That is, different users with different backgrounds may discover different knowledge from the same spatial dataset in different applications by using different methods when changing measurement scales at different cognitive hierarchies and under different resolution granularity.

View-angle enables the dataset to be illuminated by the purpose lights from different angles in order to focus on the difference when SDM creates patterns innovatively or uncovers unknown rules. The difference is accomplished via the elements of SDM view, which include the internal essential elements that drive SDM (i.e., the user, the application, and the method) and the external factors that have an impact on SDM (i.e., hierarchy, granularity, and scale). The composition of the elements and their changes result in various SDM views. As a computerized simulation of human cognition by which one may observe and analyze the same entity from very different cognitive levels, as well as actively moving between the different levels, SDM can discover the knowledge not only in worlds with the same view-angle but also in worlds with different view-angles from the same datasets for various needs. The choice of SDM views must also consider the specific needs and the characteristics of the information system.

2.4.1 SDM User

A user is interested in SDM with personality. There are all kinds of human users, such as a citizen, public servant, businessman, student, researcher, or SDM expert. The user also may be an organization, such as a government, community, enterprise, society, institute, or university. In an SDM organization, the user may be an analyst, system architect, system programmer, test controller, market salesman, project manager, innovative researcher, or president. The background knowledge context of the SDM user may be completely unfamiliar, somewhat knowledgeable, familiar, or proficient. The realm indicates the level of human cognition of the world; different users with different backgrounds have different interests. When a SDM-referenced decision is made, the users are hierarchical. Top decision-makers macroscopically master the entire dataset for a global development direction and therefore ask for the most generalized knowledge. Middle decision-makers take over from the above hierarchy level and introduce and manage information. Bottom decision-makers microscopically look at a partial dataset for local problem resolution and therefore ask for the most detailed knowledge.

2.4.2 SDM Method

When users input external data into a system, summarize datasets, and build a knowledge base, the conventional methods encounter problems because of the complexity and ambiguity of the knowledge and the difficulties in representing it. Fortunately, this is not the case for and is the major advantage of SDM. The SDM model mainly includes dependency relationship analysis, classification, concept description, and error detection. The SDM methods are closely related to the type of discovered knowledge, and the quality of the SDM algorithms directly affects the quality of the discovered knowledge. The operation of the SDM algorithms is supported by techniques that include rule induction, concept cluster, and association discovery. In practice, a variety of algorithms are often used in combination. The SDM system can be an automatic human–computer interaction using its own spatial database or external databases from GIS databases. Development can be stand-alone, embedded, attached, etc. Various factors need to be taken into account in SDM; therefore, SDM's theories, methods, and tools should be selected in accordance with the specific needs of the user. SDM can handle many technical difficulties, such as massive data, high dimension, contaminated data, data uncertainty, a variety of view-angles, and difficulties in knowledge representation (Li and Guan 2000).

2.4.3 SDM Application

SDM is applied in a specific field with constraints and relativity. SDM uncovers the specific knowledge in a specific field for spatial data-referenced decision-making. The knowledge involves spatial relationships and other interesting knowledge that is not stored in external storage but is easily accepted, understood, and utilized. SDM specifically supports information retrieval, query optimization, machine learning, pattern recognition, and system integration. For example, SDM will provide knowledge guidance and protection to understand remote sensing images, discover spatial patterns, create knowledge bases, reorganize image databases, and optimize spatial queries for accelerating the automation, intelligence, and integration of image processing. SDM also allows the rational evaluation of the effectiveness of a decision based on the objective data available.

SDM is therefore a kind of decision-making support technology. The knowledge in the decision-making system serves data applications by helping the user to maximize the efficient use of data and to improve the accuracy and reliability of their production, management, operation, analysis, and marketing processes. SDM supports all spatial data-referenced fields and decision-making processes, such as GIS, remote sensing, GPS, transportation, police, medicine, transportation, navigation, and robotics.

2.4.4 SDM Hierarchy

Hierarchy depicts the cognitive level of human beings when dealing with a dataset in SDM, reflects the level of cognitive discovery, and describes the summarized transformation from the microscopic world to the macroscopic world (e.g., knowledge with different demands). Human thinking has different levels based on the person's cognitive stature. Decision-makers at different levels and under different knowledge backgrounds may need different spatial knowledge. Simultaneously, if the same set of data is mined from dissimilar view-angles, there may be some knowledge of different levels. SDM thoroughly analyzes data, information, and concepts at various layers by using roll-up and drill-down. Roll-up is for generalizing coarser patterns globally, whereas drill-down is for detecting finer details locally. Hierarchy conversion in SDM sets up a necessary communication bridge between hardware platforms and software platforms. Their refresh and copy technology include communication and reproduction systems, copying tools defined within the database gateway, and data warehouse designated products. Its data transmission and transmission networks include network protocol, network management framework, network operating system, type of network, etc. Middleware includes the database gateway, message-oriented middleware, object request broker, etc.

Human thinking is hierarchal. The cognitive activity of human beings may arouse some physical, chemical, or electrical changes in their bodies. Reductionism in life science supposes that thinking activities can be divided into such brain hierarchies as biochemistry and neural structure. However, the certain relationships among thinking activities and sub-cellular chemistry and electrical activities cannot be built up, nor can the kind of neural structure be determined that will produce a certain kind of cognitive model. A good analogy here is the difficulty of monitoring e-mail activities on computer networks by merely detecting the functions of the most basal silicon CMOS (Complementary metal-oxide-semiconductor) chip in a computer. As a result, reductionism is questioned by system theory, which indicates that the characteristics of the system as a whole are not the superposition of low-level elements. An appropriate unit needs to be found to simulate human cognition activities. The composition levels of the matter can be seen as the hierarchy. For example, if the visible objects are macroscopic, celestial bodies are cosmologic. Objects that are smaller than atoms and molecules are called microscopic objects. The atomic hierarchy is very important because the physical model of atoms is one of the five milestones for human beings to recognize the world. Cognitive level is also related to resolution granularity and measurement scale.

Depending on the objects to be mined, SDM has the hierarchies of objects distributed throughout the world, from the analysis of protein interactions in live cells to global atmospheric fluctuations. Spatial objects may be stars or satellites distributed in the universe; various natural or manmade features on the Earth's surface are also projected reflected in computerized information. They also can be

the structure of proteins or chromosomes in molecular biology or the trajectory of electrons that move around the nucleus in atoms.

Various concepts may create a hierarchical structure on a linguistic variable. In the conceptual space, the concepts may be large or small, coarse or fine, as well as the equivalent relationship or subordinate relationships between them. A concept is divided into a number of sub-concepts. All of the concepts in the same linguistic variable naturally form a hierarchal structure—that is, the conceptual tree. The traditional conceptual tree has a clear concept boundary, and each leaf sub-concept only belongs to one father concept. However, it neglects the uncertainty between a qualitative concept and its quantitative data. Sometimes, one concept may interact with another concept and there may be an indeterminate boundary between mutual concepts. A pan-conceptual tree is presented in Fig. 2.3 to match the actual characteristic of concepts (Wang 2002). Comparatively, the concepts interact with each other in the same hierarchy, showing the randomness and ambiguity of qualitative concepts. When the hierarchy jumps up or down, each sub-concept may belong to more than one father concept. In a pan-conceptual tree considering uncertainty, the structure is also pan-hierarchy due to the concepts with different granularity sizes. Figure 2.3 is such a pan-conceptual tree of the linguistic variable of "landslide displacement" at different cognitive hierarchies in the conceptual space.

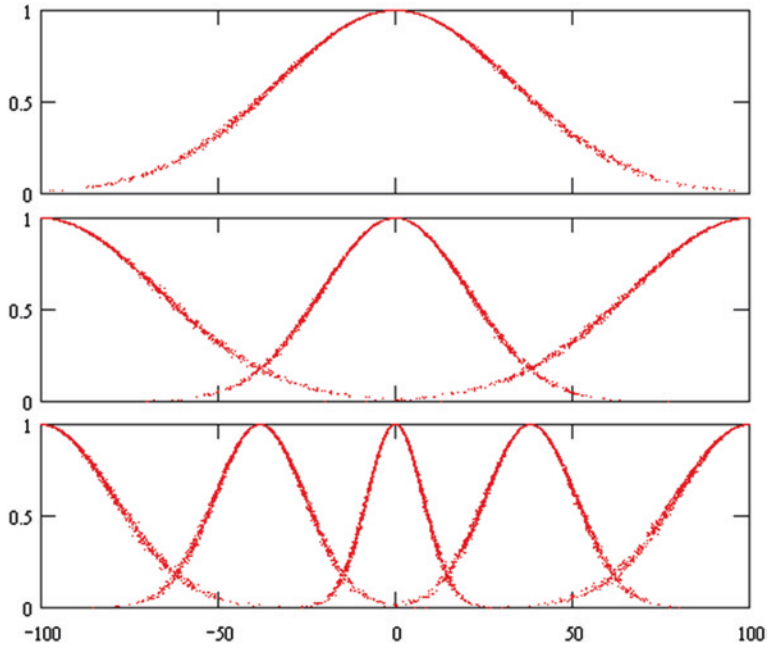

Fig. 2.3 A pan-conceptual tree of "displacement"

2.4.5 SDM Granularity

Granularity depicts the resolution of a dataset on spatial objects in SDM. It reflects the precision of SDM's interior detailing and describes the combined transformation from the fine world to the coarse world (e.g., an image with different pixels). Granularity, which is the average metric of particle or size in physics, can be descriptions of the pixel size, database tuples, the division of program modules, cognitive level for spatial data, etc. Granularity may be macro-size, meso-size, and micro-size under the abstract degree of spatial information and spatial knowledge. Based on granularity, knowledge can be grouped as generalized knowledge with coarse granularity, multiple knowledge with various granularities, and raw knowledge with fine granularity.

The granularity of general data mining is directly obtained from the field or derived by simple mathematical or logical operators. Spatial relationships, graphic features, and other granular properties, in general, are not directly stored in the database but rather are implied in images with multiple layers. Such properties can be obtained by spatial operation and spatial analysis using vector or raster. For example, the altitude of spatial objects is determined from the overlay analysis, the adjacent objects from topology analysis, the distance between the objects from the buffer analysis and distance analysis, and the slope and direction of the DEM terrain analysis.

Humans think in granular terms. In artificial intelligence and cognitive science, researchers recognize that human intelligence can freely observe and analyze the same problem in worlds of different granularities. The granularities may be very different from each other, but each granularity matches a specific purpose. Not only can people solve problems a world of the same granularity, but they also can change their thinking quickly and freely from a world of one granularity to the world of another. Even problems in worlds of different granularities can be analyzed at the same time. The size of granularity that should be used to discover knowledge from spatial data is not permanent, as there is no permanent optimal distance of observation and it depends on the research problem being addressed. For example, the granularity size of a spatial data warehouse is multiple. How much information will be reserved in the world of fine granularity after abstraction? If there is no solution in the world of coarse granularity, is there any corresponding solution in the world of fine granularity? It is an issue worthy of further research. Generally, the size of granularity reflects the difference between spatial objects in thickness, level, and point of view. That is to say, for the observation of an abstracted target, high-level problems from whole angles of view are a reflection of coarse granularity, and the observation of a concrete target, low-level problems from single angle of view is a reflection of fine granularity. In the process of spatial information processing, the size of granularity can be linked with spatial information processing units. If robust function is needed in the processing, then it is in the large granularity; otherwise, it is in the small granularity.

2.4.6 SDM Scale

Scale depicts the measurement of a dataset on spatial objects in SDM. It reflects the measurement of the exterior geometry of DMKD and describes the zoomed transformation from the big world to the small world, such as observing a map with the manipulation of zoom-in or zoom-out. Scale helps SDM extract the generalities and personalities from datasets. Generality masters objects at a high cognitive hierarchy and is more profound than personality. Personality understands objects at a low cognitive hierarchy and provides more information than generality. When SDM runs from big scale to small scale, little of the main personality is kept in generality, and much more is generalized or abandoned. However, neither generality nor personality can replace each other as each one meets a specific demand. For example, when land use is surveyed from a big scale to a small scale, the results change from personality to generality, each of which has a distinct application (i.e., land parcel, farmland, village land, town land, county land, city land, province land, or national land). A small scale helps data miners increase the distance to observe objects via datasets. Ignoring the nuances, datasets are generalized to determine the generality; a large scale helps miners shorten the distance to observe objects via datasets. Datasets are distinguished in order to determine the personality; the nuances hidden in complex phenomena are uncovered carefully and accurately.

Scale is closely related to the geometric transformation of multiple scales. The larger the scale is, the more detailed the SDM runs. With a large scale, spatial data are summarized to obtain the detailed knowledge of some objects, which may decrease the mining efficiency because the space required for data storage and computerization becomes enormous. On the other hand, it may keep the object details, resulting in the problem interpreting and resolving large amounts of data under spatial knowledge. The smaller the scale is, the more generalized SDM runs, and tiny attributes or features become invisible or abandoned. With small scale, spatial data are summarized to discover the common knowledge of all objects, which may increase the mining efficiency because less space is required for data storage and computerization. However, details may be lost, which decreases the problem of large amounts of data to interpret and resolve under spatial knowledge. In discovery state space, the scale-oriented operation changes from fine data to coarse knowledge (e.g., map generalization in cartography). Houses and rivers, for example, in a large-scale map are turned into points and lines, respectively, but these features become too tiny to keep in a small-scale map.

2.4.7 Discovery Mechanism

Humans think by searching the outer appearance to perceive the essence of an object. SDM operates in the same way through a computer by discovering patterns

from datasets. To simulate human thinking activities, SDM must find approaches to establish relationships between the human brain and the computer. Logically, SDM is an inductive process that goes from concrete data to abstract patterns, from special phenomena to general rules. The process of discovering the implicit knowledge from the explicit dataset is similar to the human cognitive process of uncovering the internal nature from the external phenomena. As a discovery process, SDM changes from concrete data to abstract patterns and also from particular phenomena to general laws. Therefore, the feature interactions among objects may help knowledge discovery.

The essence of SDM is to observe and analyze the same dataset by reorganizing datasets at various distances with mining views. The change from the elements of view results in various SDM views. For example, the change of one or all of the hierarchy, granularity, and scale will make SDM different in the angle aspects of high cognition or low cognition, coarse resolution or fine resolution, small measurement or big measurement. In the SDM process, the spatial concept first is extracted from its corresponding dataset, and then the preliminary features are extracted from the concept; finally, the characteristic knowledge is induced from feature space. At a close distance, the knowledge template is at a low cognitive hierarchy with fine granularity and a large scale; however, the discovered knowledge is a detailed personality for microscopically distinguishing object differences carefully. The tiny features may be uncovered for looking at the typical examples, such as sharpening an image for local features. At a far distance, the knowledge template is at a high cognitive hierarchy with coarse granularity and a small scale; the discovered knowledge is summarized generally for macroscopically mastering all of the objects as a whole. The subtle details are neglected for grasping the key problems, such as smoothing an image for local features.

Regular rules and exceptional outliers are discovered simultaneously. A spatial rule is a pattern showing the intersection of two or more spatial objects or space-dependent attributes according to a particular spacing or set of arrangements (Ester et al. 2000). In addition to the rules, during the discovering process of description or prediction, there may be some exceptions (also called outliers) that deviate very much from other data observations (Shekhar et al. 2003). The following approaches identify and explain exceptions (surprises). For example, spatial trend predictive modelling first discovers the centers that are local maximal of a certain non-spatial attribute and then determines its theoretical trend when moving away from the centers. Finally, a few deviations are found in that some data are far from the theoretical trend, which may arouse the suspicion that they are noise or are generated by a different mechanism. How are these outliers explained? Traditionally, the detection of outliers has been studied using statistics. A number of discordancy tests have been developed, most of which treat outliers as noise and then try to eliminate their effects by removing them or by developing some outlier-resistant method (Hawkins 1980). These outliers actually prove the rules; in the context of data mining, they are meaningful input signals rather than noise. In some cases, outliers represent unique characteristics of the objects that are

important to an organization. Therefore, a piece of generic knowledge is virtually in the form of a rule plus exception.

An exception (outlier) is also a useful pattern in SDM. In the implementation process of SDM, one or more specific elements in the target dataset may be selected (e.g., object attributes, personality circumstances). There also will be some special circumstances when generalizing attributes, promoting concepts, or introducing features from lower reorganization levels to higher reorganization levels. That is, there are always some spatial patterns that are described in the world of a fine knowledge template, which cannot be involved in the world of a coarse knowledge template and therefore exists as an exception.

Thus, SDM is the process of discovering and extracting patterns from spatial data and converting them to spatial knowledge with view-angles. The knowledge is the pattern of "rule plus exception" or "class plus outlier." SDM view utilizes its integrated human cognition to change individual objects to the general knowledge-base and from concrete data to abstract patterns. It enables the user to thoroughly manipulate a spatial dataset as a whole for grasping data essence.

2.5 Spatial Knowledge to Discover

A variety of knowledge can be discovered from spatial datasets (Table 2.1). The discovered knowledge follows the pattern of "rule plus exception" with different SDM views (Fig. 2.2). Rules mainly include both generality and individuality rules, such as association, characteristics, discrimination, clustering, classifications, serials, predictions, and functional dependence. Exceptions refer to the bias of the rules. At the same time, the type of knowledge also depends on the type of SDM task (i.e., the problem that SDM can solve). The knowledge is not isolated from either. A variety of rules are required at the same time when solving practical problems.

2.5.1 General Geometric Rule and Spatial Association Rule

Generally, some types of geometric rules can be discovered from GIS databases, such as geometric shape, distribution, evolution, and bias (Di 2001). This general knowledge of geometry refers to the common geometric features of a certain group of target objects (e.g., volume, size, shape). The objects can be divided into three categories: points (e.g., independent tree, settlements on a small-scale map), lines (e.g., rivers, roads), and polygons (e.g., residents, lakes, squares). The numbers or sizes of these objects are calculated by using the methods of mathematical probability and statistics. The size of linear objects is expressed by length and width, and the size of a polygon object is represented by its area and perimeter. The morphological features of objects are expressed by a quantitative eigenvalue

Table 2.1 Spatial knowledge to be discovered

Knowledge	Interpretations	Examples
Association rule	A logic association among different sets of entities that associate one or more objects with other objects for studying the frequency of items occurring together in databases	Rain (location, amount of rain) ⇒ Landslide (location, occurrence), support is 76 %, confidence is 98 %, interest is 51 %
Characteristics rule	A common feature of a kind of entity, or several kinds of entities, for summarizing similar features of objects in a target class	Characterize similar ground objects in a large set of remotely sensed images
Discriminate rule	A different feature that distinguishes one entity from another entity and for comparing the core features of objects between a target class and a contrasting class	Compare land prices in a suburban area with land prices in urban center
Clustering rule	A segmentation rule that groups a set of objects by virtue of their similarity without any knowledge of what causes the grouping and how many groups exist. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters	Group crime locations to find distribution patterns
Classification rule	A rule that defines whether an entity belongs to a particular class with a defined kind and variables, or a set of classes	Classify remotely sensed images based on spectrum and GIS data
Serial rule	A temporal constrained rule that relates entities or the functional dependency among the parameters in a time sequence, for analyzing the sequential pattern, regression, and similar sequences	During summer, landslide disasters often occur
Predictive rule	An inner trend that forecasts future values of some variables when the temporal or spatial center is moved to another one, or predicts some unknown or missing attribute values based on other seasonal or periodical information	Forecast the movement trend of landslide based on available monitoring data. Identify the segments of a population likely to respond similarly to given events
Exception	An outlier that is isolated from the common rules or is derived from other data observations substantially, used for identifying anomalous attributes and objects	A monitoring point detecting exceptional movement that predicts landslides or detecting fraudulent credit card transactions

that is easily achieved by a computer through an intuitive and visual graph. The morphological features of the linear objects are characterized by twists and turns degree (complexity) and the direction; polygon objects are characterized by the intensity, twists and turns degree of the boundaries, and the major axis direction; and point objects have no morphological features. However, the morphological features of point objects gathered together as a cluster can be determined by methods similar to those for polygon objects. Generally, GIS databases only store some geometric features (e.g., length, area, perimeter, geometric center), while the calculation of morphological features requires special algorithms. Some statistical values of geometric features (e.g., minimum, maximum, mean, variance, plural) can be calculated; if there are enough samples, the feature histogram data can be used as priori probability. Therefore, general geometric knowledge at a higher level can be determined according to the background knowledge.

Spatial association refers to the internal rules among spatial entities that are present at the same time and describes the conditional rules that frequently appear in the feature data of spatial entities in a given spatial database. First, an association rule may be adjacent, connective, symbiotic, and contained. A one-dimensional association includes a single predicate, and a multi-dimensional association includes two or more spatial entities or predicates. Second, the association may be general and strong. General association is a common correlation among spatial entities, and a strong association appears frequently (Koperski 1999). The meanings of strong association rules are more profound and their application range is broader. They also are known as generalized association rules. Third, association rules are descriptive rules that provide a measurement of the support, confidence, and interest. For example, "(x, road) $\rightarrow$ close to (x, river) (82 %)" is a description of Chengde associated with roads and rivers. Describing association rules in a form that is similar to structured query language (SQL) brings SDM into standard language and engineering. If the attributes of objects in SDM are limited to Boolean type, the association rules can be extracted through the conversion type in the database that contains the categorical attributes by combining some information of the same objects. Fourth, association rules are temporal and transferable and may ask for additional information, such as a valid time and transferable condition. For example, the association rules on Yellow River water—"IF it rains, THEN the water level rises (spring and summer)" and "IF it rains, THEN the water level recedes (autumn and winter)"—cannot be exchanged when they are used to prevent and reduce the flooding of the Yellow River. For example, the association rules obtained in the first round of election may be different from or even contrary to the rules determined in the second round of election as the voters' willingness may be transferred under the given conditions. Rationally predicating this transfer can help candidates adjust the strategy of election, thereby increasing the possibility of winning. Finally, an association rule is a simple and practical rule in SDM that attracts many researchers for normalization, query optimization, minimizing a decision tree, etc. in spatial databases.

2.5.2 Spatial Characteristics Rule and Discriminate Rule

Spatial characteristic rules summarize the generality of distribution and attribute characteristics of a single kind or multiple kinds of spatial entities and are the generalized description of the concept and spatial class. If there are enough samples, the map, histogram, pie chart, bar graph, curve, data cube, and data table of spatial characteristics can be converted to the priori probability knowledge. For example, "the direction of most roads in Beijing is east–west or north–south" and "most roads in Beijing are relatively straight" are two spatial characteristics rules that describe the common features of Beijing's roads, which are also general geometry knowledge.

Spatial discrimination distinguishes the feature difference between two types of spatial objects or among multiple types of spatial objects. The feature difference determines the target class and the contrastive class under the given feature from spatial datasets. Discriminate rules are discovered by comparing the characteristics between the target class and the contrastive class. For example, "the direction of most roads in Beijing are east–west or north–south, while the direction of most roads in Tianjin are parallel or perpendicular to the rivers" and "most roads in Beijing are relatively straight, while most roads in Tianjin are flexural" are two discriminate rules that describe the general differences in the direction and shape of roads in Beijing and Tianjin. Discriminately, spatial objects are distributed in vertical, horizontal, or vertical-horizontal directions, such as the vertical distribution of alpine vegetation, the difference between the infrastructures of a city and a village, or exotic features distributed along a slope and their exposure. For example, "the majority of Chinese rice-growing areas are located south of the Huaihe River and the Qinling Mountains" or "most of the Chinese wheat-growing areas are located southeast of the mountain, north of the Huaihe River and the Qinling Mountains" are two spatial distribution rules of crops.

2.5.3 Spatial Clustering Rule and Classification Rule

Spatial clustering rules group a set of data in a way that maximizes the feature similarity within clusters and minimizes the feature similarity between two different clusters. Sequentially, spatial objects are partitioned into different groups via the feature similarity by making the difference in data objects between different groups as large as possible and the difference between data objects in the same group as small as possible (Grabmeier and Rudolph 2002). According to the different criteria of similarity measurement and clustering evaluation, the commonly used clustering algorithms may be based on partition, hierarchy, density, and grid (Wang et al. 2011). Clustering rules further help SDM to discretize data, refine patterns, and amalgamate information. For example, continuous datasets are partitioned into discrete hierarchical clusters and multisource information are amalgamated for the same object.

Spatial classification classifies an object by mapping it to a specific class according to its discriminate rules. It attempts to assign each input value to one of a given set of classes. Classification reflects the generality to characterize the difference among different kinds of objects and the similarity inside the same kind of objects. Classification rules may be described in the form of a decision tree, concept lattice, and predicate logic. For example, the value searching of a decision tree from the root to the branches and leaf nodes will be able to determine the category or predict the unknown value of an object. In GIS, object-oriented classification rules refer to the subclass structure of the objects and the general knowledge.

Clustering rules for spatial datasets without cluster labels are different from classification rules with class labels. Clustering rules do not preset cluster labels. Before clustering datasets, the clusters are unknown as far as amount, content, name, etc. At the same time, partitioned clusters may run the preparation of classification generation, and classified classes may supervise clustering.

2.5.4 Spatial Predictable Rule and Serial Rule

Predictable spatial rules can forecast an unknown value, label, attribute, or trend by assigning a spatial-valued output to each input. These rules can determine the internal dependence between spatial objects and their impact variables in the future, such as a regressive model or a decision tree. Before forecasting, correlation analysis can be used to identify and exclude attributes or entities that are useless or irrelevant. For example, the future occurrence of a volcano is extracted from the structure, the plate movement, and the gravity field of Earth in the database. When predicting future values of data based on the trends changing with time, the specificity of the time factor should be fully considered.

A spatial serial rule summarizes the spatiotemporal pattern of spatial objects changing during a period of time. It links the relationships among spatial data and time over a long period of time. For example, in a city over the years, banks and their branches store their operating income and expenditure accounting records and the police department records security cases, from which financial and social trends can be discovered under their geographical distribution. The time constraints can be depicted with a time window or adjacent sequences. Time-constrained serial rules also are called evolution rules. Although they are related to other spatial rules, serial rules mining concentrates more on historical datasets of the same object in different times, such as series analysis, sequence match, time reasoning, etc. Spatial serial rule mining is utilized when the user wants to obtain more refined information excavated for only a period of implicit models. Only by using a series of values of the existing data changing with time can SDM better predict future trends based on the mined results. When little change has occurred in a database, gradual sequence rule mining may speed up the SDM process by taking advantage of previous results.

2.5.5 Spatial Exception or Outlier

Outlier detection, in addition to the commonly used rules, is used to extract interesting exceptions from datasets in SDM via statistics, clustering, classification, and regression (Wang 2002; Shekhar et al. 2003). Outlier detection can also identify system faults and fraud before they escalate with potentially catastrophic consequences. Although outlier detection has been used for centuries to detect and remove anomalous observations from data, there is no rigid mathematical definition of what constitutes an outlier. Ultimately, it is a subjective exercise to determine whether or not an observation is an outlier. There are three fundamental approaches to outlier detection (Hodge and Austin 2004):

(1) Determine the outliers without prior knowledge of the data, which processes the data as a static distribution, pinpoints the most remote points, and flags them as potential outliers. Essentially, it is a learning approach analogous to unsupervised clustering.
(2) Model both normality and abnormality, which is analogous to supervised classification and requires pre-labeled data tagged as normal or abnormal.
(3) Model only normality (or in a few cases, abnormality), which may be considered semi-supervised as the normal class is taught but from which the algorithm learns to recognize abnormality. It is analogous to a semi-supervised recognition or detection task.

Normal distribution of the data is assumed in order to identify observations that are deemed unlikely on the basis of the mean and standard deviations. Distance-based methods frequently use the distance to the nearest neighbors to label observations as outliers or non-outliers (Ramaswamy et al. 2000). In the sequence rule, outlier detection is a heuristic method, which recognizes data that cause a sudden severe fluctuation in the sequential data as an exception by using linear deviation detection. Lee (2000) used a fuzzy neural network to estimate the rules for dealing with distribution abnormality in spatial statistics.

Spatial exceptions or outliers are the deviations or independent points beyond the common features of the most spatial entities. An exception is an abnormality. If manmade factors have been ruled out, an exception is often the presence of sudden changes (Barnett 1978). Deviation detection—a heuristic approach to data mining—can identify the points that have sudden fluctuations in the sequence data as exceptions (Shekhar et al. 2003). Spatial exceptions are the object features that are inconsistent with the general actions or universal models of the data in spatial datasets. They are the descriptions of analogical differences, such as the special case in a standard class, the isolated points out of various classifications, the difference between a single attribute value and a set of attribute values in time serials, and a significant difference between the actual value of an observation and the system forecasting value. A lot of data mining methods ignore and discard exceptions as noise or abnormality. Although excluding such exceptions may be conducive to highlighting the generality, some rare spatial exceptions may be much more

significant than normal spatial objects (Hawkins 1980). For example, near a notable feature of the displacement in a large landslide point, there may be a potential landslide hazard, which is the decisive knowledge of landslide prediction. Spatial exceptional knowledge can be discovered with data fields, statistical hypothesis testing, or identifying feature deviations.

2.6 Spatial Knowledge Representation

Knowledge representation is a key issue in SDM. At present, the common knowledge representation methods include natural language (e.g., "the Baota of the Three Gorges landslide moved towards the micro-west south during the monitoring, and was accompanied by a small amount of settlement"), predicates logic, a function model, a characteristic table (relationship table), a generalized rule, a semantic network, a framework, a script, a process, a Petri net, and visualization, among others. The introduction of natural language in knowledge representation is a general recognition of the uncertainty in thinking and perception. The language value increases the flexibility of knowledge, making the discovered knowledge more reliable and easier to understand. In actual applications, they all have their advantages and disadvantages. Different methods are suitable for different knowledge. The same knowledge generally can be represented by a number of methods, which can be mutually converted.

2.6.1 Natural Language

Natural language is one of the best methods to describe datasets based on human thinking and communicating with each other. As a carrier of human thinking, natural language helps to achieve a powerful tool for thinking to display and retain the subject for thought and to organize the process of thinking. It is the foundation of a variety of other formal systems or languages, which are derived from a special language, such as computer language, or some specific symbol languages, such as mathematical language. The formal systems constituted by these symbols further become a new formal system.

The basic language value of natural language is a qualitative concept, corresponding to a group of quantitative data. Seen from the process of the atomic model that evolved from the Kelvin model, the Thomson model, the Lenard model, the Nagaoka model, and the Nicholson model to Rutherford's atomic model with nuclei, it is a universal and effective methodology to work out the model of material composition. The concept maps the object from the objective world to subjective cognition. As far as concept generation, regardless of whether the characteristic table theory or the prototype theory is used, all conceptual samples are reflected by a set of data. The smallest unit of natural language is the language value to describe

the concepts. The most basic language value represents the most basic concept—that is, the linguistic atom. As a result, the linguistic atom forms the atomic model when human thinking is modeled with the help of natural language.

The difference between spatial knowledge and non-spatial knowledge lies in spatial knowledge having spatial concepts and spatial relationships; furthermore, the abstract representation of these spatial concepts and spatial relationships is most appropriately expressed by language values. Mastering the quantitative dataset with qualitative language values conforms to human cognitive rules. Obtaining qualitative concepts from a quantitative dataset reflects the essence of objects more profoundly, and subsequently fewer resources are spent to deliver adequate information and make efficient judgments and reasoning of complex things. When representing the definitive properties of discovered knowledge, soft natural language is more universal, more real, more distinct, more direct, and easier to understand than exact mathematical language. A lot of knowledge obtained by SDM is qualitative knowledge after induction or abstraction, or a combination of qualitative and quantitative knowledge. The more abstract the knowledge is, the more suitable is natural language. However, the concept represented by natural language inevitably has uncertainty commonly, is even blind and undisciplined, and therefore is a bottleneck to freely transform between quantitative data and qualitative concept.

2.6.2 Conversion Between Quantitative Data and Qualitative Concept

A conversion model inevitably needs to be set up between qualitative language values and quantitative numerical values in order to realize the conversion between numerical values and symbol values at any time, to establish the interrelated and interdependent map relationship between quality and quantity, and to reflect the uncertainty of mapping between quality and quantity (particularly randomness and fuzziness). At present, the commonly used qualitative and quantitative conversion methods include the analysis hierarchy process, quantizing weightings, experts grading, qualitative analysis combined with some mathematical models, and quantitative calculation. The basic of cybernetics in dealing with uncertainty is to eliminate errors depending on the actual goal and practice. Therefore, none of the above methods are perfect because they fail to take both randomness and ambiguity into consideration.

The cloud model is an uncertainty conversion model between quantitative data and qualitative concepts. The numerical characteristics of the cloud fully integrate the fuzziness and randomness; together, they constitute a mapping between quality and quantity, which can be used as the basis of knowledge representation. When expanding from one-dimensional to two-dimensional and multi-dimensional, the cloud model can express spatial concepts and spatial relationships. For example, the “far,” “close” language value uses a one-dimensional model of cloud to express

"northeast" and "southwest" language value used in a two-dimensional model. A one-dimensional cloud model is used to express quantity language values such as "a small number of," "basically," "the majority," "almost all," etc. (Wang 2002). The language values indicated by the cloud model can be used to express the qualitative concept. The integration of the cloud model and the traditional expression method is an enhancement of the qualitative concept expression and the qualitative, quantitative conversion of these knowledge expression methods.

2.6.3 Spatial Knowledge Measurement

Spatial knowledge measurement is the certainty or credibility of knowledge, such as support, confidence, expected confidence, lift, and interest (Agrawal and Srikant 1994; Reshef et al. 2011). Because not all the discovered patterns are interesting or meaningful, measurements can be used to supervise or limit the progress of SDM. Here, association rules are taken as an example to introduce measurements.

Support that dataset D is a collection of T (Transaction), $I = (i_1, i_2, \ldots, i_n)$ is the set of Items, e.g., object, entity, attribute, feature, and $T \subseteq I$. Each T has an identifier of *TID*. A, B is the item set, $A \subseteq T$, $B \subseteq T$, and $A \cap B = \emptyset$. An association rules $A \Rightarrow B$ is given with measurements in Eq. (2.1).

$$((A_1 \wedge A_2 \wedge \ldots \wedge A_m) \Rightarrow (B_1 \wedge B_2 \wedge \ldots \wedge B_n))|([s], [c], [ec], [l], [i]) \tag{2.1}$$

Where, $A_1, A_2, \ldots, A_m$; $B_1, B_2, \ldots, B_n$ is a set of spatial or non-spatial predicates, […] is optional, [s], [c], [ec], [l], [i] measure the certainty of support, confidence, expected confidence, lift, and interest, respectively.

- *Support measures the level of certainty that describes the probability $P(A \cup B)$ of the union of A and B happening in D, i.e., $s(A \Rightarrow B) = P(A \cup B)$*
- *Confidence measures the level of certainty that describes the conditional probability $P(B/A)$ of the intersection of A and B happening in D; i.e., $c(A \Rightarrow B) = P(B/A)$ or $c(A \Rightarrow B) = P(A \cap B)$*
- *Expected Confidence measures the certainty that describes the expected probability $P(B)$ of B happening in D; i.e., $ec(A \Rightarrow B) = P(B)$. It describes the probability that B happens without any conditions.*
- *Lift measures the certainty of the ratio of confidence and expected confidence; i.e., $l(B/A) = c(A \Rightarrow B)/ec(B)$. It indicates how the probability that A happens impacts the probability that B happens. The bigger Lift is, the more the happening of A will accelerate the happening of B.*
- *Interest measures the level of certainty that users are interested in SDM patterns when making decisions. According to its definition, association rules exist between any two sets of items. If users do not consider whether or not they are useful, a lot of rules may be uncovered but not all will be meaningful.*

Support and confidence are two important indicators that reflect the usable level of the rules: support is the measurement of a rule's importance, whereas confidence is the measurement of a rule's accuracy. The support illustrates how it is representative of the rule in all transactions; the larger the support is, the more important the associations rule becomes. A rule with a high confidence level but very small support indicates that it rarely happens, along with having little practical value.

In the actual SDM, it is important to define the thresholds of spatial knowledge measurements (e.g., minimum support and minimum confidence). Under normal circumstances, when the lift of the useful association rule is more than 1, it suggests that the confidence of the association rules is greater than the expected confidence. If its lift is not more than 1, the association rule is meaningless. The general admission threshold value is based on experience, but users can also be determined by statistics. Only when their measurements are larger than the defined thresholds can the rules be accepted as the interesting ones (also called strong rules) for application. Moreover, the defined thresholds should be appropriate under the given situations. If the thresholds are too small, a large number of useless rules will be discovered, which not only affect the efficiency of the implementation and waste system resources, but also may flood the main objective. If the value is too large, the results may not be rules at all, the number of rules is too small, or the expected rules may be filtered out.

2.6.4 Spatial Rules Plus Exceptions

SDM knowledge can be expressed in a number of ways. A more reasonable expression idea is "spatial rules plus exception," accompanied with measurements at different levels. The patterns of "rule plus exceptions" have extensive applications in intelligent spatial analysis and interpretation. To match different tasks, a group of spatial datasets can be understood from different mining views, enabling many kinds of knowledge on spatial objects to be summarized. The summarized knowledge may technically support decision-making to resolve and interpret natural and human problems at personalized levels. They identify the links between mutual records to generate a summary of a spatial database and to create a prediction model and a classification model for spatial expert systems or decision support systems. They also can restrict, support, and supervise the interpretation of remote sensing images to address the phenomenon in the same spectrum with different objects, and different objects with the same spectrum, in order to reduce doubt in the classification results. SDM enhances the reliability, accuracy, and speed of the interpretation; for example, the rules of a road connected to a town or village or a bridge at the intersection of a road and a river can improve the accuracy of the classification and update the spatial database when they are applied in image classification. The "grass $\Rightarrow$ forest" rule from the image database states that grass and forest often appear at the same time, but the grass on the forest boundary is assigned a larger weight to provide a defined boundary in order to improve the accuracy of classification.

References

Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: Proceedings of international conference on very large databases (VLDB), Santiago, Chile, pp 487–499

Barnett V (1978) Outliers in statistical data. Wiley, New York

Barrabbasi AL, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512

Di KC (2001) Spatial data mining and knowledge discovering. Wuhan University Press, WuHan

Ester M et al (2000) Spatial data mining: databases primitives, algorithms and efficient DBMS support. Data Min Knowl Disc 4:193–216

Frasconi P, Gori M, Soda G (1999) Data categorization using decision trellises. IEEE Trans Knowl Data Eng 11(5):697–712

Grabmeier J, Rudolph A (2002) Techniques of clustering algorithms in data mining. Data Min Knowl Disc 6:303–360

Han JW, Kamber M, Pei J (2012) Data mining: concepts and techniques, 3rd edn. Morgan Kaufmann Publishers Inc., Burlington

Hawkins D (1980) Identifications of outliers. Chapman and Hall, London

Hodge VJ, Austin J (2004) A survey of outlier detection methodologies. Artif Intell Rev 22(2):85–126

Koperski K (1999) A progressive refinement approach to spatial data mining. Ph.D. thesis, Simon Fraser University, British Columbia

Lee ES (2000) Neuro-fuzzy estimation in spatial statistics. J Math Anal Appl 249:221–231

Li DR, Guan ZQ (2000) Integration and Implementation of spatial information system. Wuhan University Press, Wuhan

Li DR, Wang SL, Shi WZ, Wang XZ (2001) On spatial data mining and knowledge discovery (SDMKD). Geomatics Inf Sci Wuhan Univ 26(6):491–499

Li DR, Wang SL, Li DY (2006) Theory and application of spatial data mining, 1st edn. Science Press, Beijing

Li DR, Wang SL, Li DY (2013) Theory and application of spatial data mining, 2nd edn. Science Press, Beijing

Li DY, Du Y (2007) Artificial intelligence with uncertainty. Chapman and Hall/CRC, London

Liu B (2007) Web data mining: exploring hyperlinks, contents, usage data, 2nd edn. Springer, Heidelberg

Piatetsky-shapiro G (1994) An overview of knowledge discovery in databases: recent progress and challenges. In: Ziarko Wojciech P (ed) Rough sets, fuzzy sets and knowledge discovery. Springer, Berlin, pp 1–10

Ramaswamy S, Rastogi R, Shim K (2000) Efficient algorithms for mining outliers from large data sets. In: Proceeding SIGMOD ‘00 proceedings of the 2000 ACM SIGMOD international conference on management of data, pp 427–438

Reshef DN et al (2011) Detecting novel associations in large data sets. Science 334:1518

Shekhar S, Lu CT, Zhang P (2003) A unified approach to detecting spatial outliers. GeoInformatica 7(2):139–166

Srivastava J, Cheng PY (1999) Warehouse creation-a potential roadblock to data warehousing. IEEE Trans Knowl Data Eng 11(1):118–126

Wang SL (2002) Data field and cloud model based spatial data mining and knowledge discovery, PhD thesis, Wuhan University, Wuhan

Wang SL, Shi WZ (2012) Data mining and knowledge discovery. In: Kresse Wolfgang, Danko David (eds) Handbook of geographic information. Springer, Berlin

Wang SL, Yuan HN (2014) Spatial data mining: a perspective of big data. Int J Data Warehouse Min 10(4):50–70

Wang SL, Gan WY, Li DY, Li DR (2011) Data field for hierarchical clustering. Int J Data Warehouse Min 7(4):43–63

Watts DJ, Strogatz SH (1998) Collective dynamics of ‘small world’ networks. Nature 393:400–442

Witten I, Frank E (2000) Data mining, practical machine learning tools and techniques with java implementation. San Francisca: Morgan Kaufman Publishers

http://www.springer.com/978-3-662-48536-1

Spatial Data Mining
Theory and Application
Li, D.; Wang, S.; Li, D.
2015, XXVIII, 308 p. 103 illus., 81 illus. in color.,
Hardcover
ISBN: 978-3-662-48536-1