Unit 5

Text A

Data Structure

A data structure is a specialized format for organizing and storing data. General data structure types include the array, the file, the record, the table, the tree, and so on. Any data structure is designed to organize data to suit a specific purpose so that it can be accessed and worked with in appropriate ways. In computer programming, a data structure may be selected or designed to store data for the purpose of working on it with various algorithms.

1. Array

(1) In data storage, an array is a method for storing information on multiple devices.

(2) In general, an array is a number of items arranged in some specified way, for example, in a list or in a three-dimensional table.

(3) In computer programming languages, an array is a group of objects with the same attributes that can be addressed individually, using such techniques as subscripting.

(4) In random access memory (RAM), an array is the arrangement of memory cells.

2. File

(1) In data processing, a file is a related collection of records. For example, you might put the records you have on each of your customers in a file. In turn, each record would consist of fields for individual data items, such as customer name, customer number, customer

address, and so forth. By providing the same information in the same fields in each record so that all records are consistent, your file will be easily accessible for analysis and manipulation by a computer program. This use of the term has become somewhat less important with the advent of the database and its emphasis on the table as a way of collecting record and field data. In mainframe systems, the term data set is generally synonymous with file but implies a specific form of organization recognized by a particular access method. Depending on the operating system, files (and data sets) are contained within a catalog, directory, or folder.

(2) In any computer system, especially in personal computers, a file is an entity of data available to system users (including the system itself and its application programs) that is capable of being manipulated as an entity (for example, moved from one file directory to another). The file must have a unique name within its own directory. Some operating systems and applications describe files with given formats by giving them a particular file name suffix. The file name suffix is also known as a file name extension. For example, a program or executable file is sometimes given or required to have an“.exe”suffix. In general, the suffixes tend to be as descriptive of the formats as they can within the limits of the number of characters allowed for suffixes by the operating system.

3. Record

(1) In computer data processing, a record is a collection of data items arranged for processing by a program. Multiple records are contained in a file or data set. The organization of data in the record is usually prescribed by the programming language that defines the record's organization and/or by the application that processes it. Typically, records can be of fixed-length or be of variable length with the length information contained within the record.

(2) In a database, a record, sometimes called a row, is a group of fields within a table that are relevant to a specific entity. For example, in a table called customer contact information, a row would likely contain fields such as: ID number, name, street address, city, telephone number and so on.

4. Table

In computer programming, a table is a data structure used to organize information, just as it is on paper. There are many different types of computer-related tables, which work in a number of different ways. The following are examples of the more common types.

(1) In data processing, a table, also called an array, is an organized grouping of fields. Tables may store relatively permanent data, or may be frequently updated. For example, a table contained in a disk volume is updated when sectors are being written.

(2) In a relational database, a table, sometimes called a file, organizes the information about a single topic into rows and columns. For example, a database for a business would typically contain a table for customer information, which would store customers' account numbers, addresses, phone numbers, and so on as a series of columns. Each single piece of data, such as the account number, is a field in the table. A column consists of all the entries in a single field, such as the telephone numbers of all the customers. Fields, in turn, are organized as records, which are complete sets of information, such as the set of information about a particular customer, each of which comprises a row. The process of normalization determines how data will be most effectively organized into tables.

(3) A decision table, often called a truth table, which can be computer-based or simply drawn up on paper, contains a list of decisions and the criteria on which they are based. All possible situations for decisions should be listed, and the action to take in each situation should be specified. A rudimentary example: For a traffic intersection, the decision to proceed might be expressed as yes or no and the criteria might be the light is red or the light is green.

A decision table can be inserted into a computer program to direct its processing according to decisions made in different situations. Changes to the decision table are reflected in the program.

(4) An HTML table is used to organize Web page elements spatially or to create a structure for data that is best displayed in tabular form, such as lists or specifications.

New Words

specialized	[ˈspeʃəlaizd]	*adj.*专用的，专门的
organize	[ˈɔːgənaiz]	*vt.*组织；构成，组成
array	[əˈrei]	*n.*数组，排列
record	[ˈrekɔːd]	*n.*记录
	[riiˈkɔːd]	*vt.*记录；录音
table	[ˈteibl]	*n.*表，表格
appropriate	[əˈprəupriət]	*adj.*正确的，适当的
various	[ˈvɛəriəs]	*adj.*不同的，各种各样的，多方面的，多样的
subscript	[ˈsʌbskript]	*adj.*下标
collection	[kəˈlekʃən]	*n.*集合，收集来的总和
item	[ˈaitəm]	*n.*项目
consistent	[kənˈsistənt]	*adj.*一致的，调和的，相容的
accessible	[əkˈsesəbl]	*adj.*易接近的，可访问的，易受影响的
manipulation	[məˌnipjuˈleiʃən]	*n.*处理，操作
advent	[ˈædvənt]	*n.*出现，到来

emphasis	[ˈemfəsis]	*n.*强调，重点
imply	[imˈplai]	*vt.*暗示，意味
suffix	[ˈsʌfiks]	*n.*后缀；下标
prescribe	[prisˈkraib]	*v.*指示，规定
define	[diˈfain]	*vt.*定义，详细说明
length	[leŋθ]	*n.*长度
row	[rau]	*n.*行，排
relevant	[ˈrelivənt]	*adj.*有关的，相应的
common	[ˈkɔmən]	*adj.*共同的，公共的，公有的，普通的
permanent	[ˈpəːmənənt]	*adj.*永久的，持久的
frequently	[ˈfriːkwəntli]	*adv.*常常，频繁地
sector	[ˈsektə]	*n.*扇区
normalization	[ˌnɔːməlaiˈzeiʃən]	*n.*规范化，正常化，标准化
criteria	[ˈkraiˈtiəriə]	*n.*标准
rudimentary	[ruːdiˈmentəri]	*adj.*基本的，初步的
intersection	[ˌintəːˈsekʃən]	*n.*交集，十字路口，交叉点
spatial	[ˈspeiʃəl]	*adj.*空间的，立体的，三维的

Phrases

data structure	数据结构
and so on	等等
a number of	许多的
three-dimensional table	三维表
memory cell	内存单元
and so forth	等等
data set	数据集
file name extension	文件扩展名
account number	账号
in turn	依次，轮流
decision table	判定表，决策表
truth table	真值表
draw up	草拟

Abbreviations

ID（Identification, Identity）	身份

Notes

[1] Any data structure is designed to organize data to suit a specific purpose so that it can be accessed and worked with in appropriate ways.

本句中，to organize data to suit a specific purpose so that it can be accessed and worked with in appropriate ways 是一个动词不定式短语，作目的状语，修饰 is designed。在该短语中，to suit a specific purpose 也是一个动词不定式短语，作目的状语，修饰 to organize，so that it can be accessed and worked with in appropriate ways 是一个目的状语从句。

[2] In computer programming languages, an array is a group of objects with the same attributes that can be addressed individually, using such techniques as subscripting.

本句中，with the same attributes 是一个介词短语，作定语，修饰和限定 a group of objects。that can be addressed individually 是一个定语从句，也修饰和限定 a group of objects，using such techniques as subscripting 是一个现在分词短语，作方式状语，修饰从句的谓语 can be addressed。

[3] By providing the same information in the same fields in each record so that all records are consistent, your file will be easily accessible for analysis and manipulation by a computer program.

本句中，By providing the same information in the same fields in each record so that all records are consistent 是一个现在分词短语，作方式状语，修饰谓语 will be easily accessible。在该短语中，so that all records are consistent 是一个结果状语从句，修饰谓语 providing。

英语中，so that 既可以引导一个目的状语从句，也可以引导一个结果状语从句。请看下例：

We asked the professor to speak louder so that we could hear him.

我们请教授讲话声再大一些，以便让我们能听清。（目的状语从句）

Mary didn’t plan her time well, so that she didn’t finish the work in time.

玛丽没有把时间计划好，结果没有按时完成这项工作。（结果状语从句）

[4] In any computer system but especially in personal computers, a file is an entity of data available to system users (including the system itself and its application programs) that is capable of being manipulated as an entity (for example, moved from one file directory to another).

本句中，available to system users 是一个现在分词短语，作定语，修饰和限定 an entity of data。that is capable of being manipulated as an entity 是一个定语从句，也修饰和限定 an entity of data。

[5] The organization of data in the record is usually prescribed by the programming language

that defines the record’s organization and/or by the application that processes it.

本句中，and/or 连接了 by 引导的两个方式状语。that defines the record’s organization 是一个定语从句，修饰和限定 the programming language。

[6] For example, a database for a business would typically contain a table for customer information, which would store customers’ account numbers, addresses, phone numbers, and so on as a series of columns.

本句中，which would store customers’ account numbers, addresses, phone numbers, and so on as a series of columns 是一个非限定性定语从句，对 a table for customer information 进行补充说明。

[7] Fields, in turn, are organized as records, which are complete sets of information, such as the set of information about a particular customer, each of which comprises a row.

本句中，which are complete sets of information 是一个非限定性定语从句，对 records 进行补充说明。such as the set of information about a particular customer 是对 complete sets of information 的举例说明，each of which comprises a row 是一个非限定性定语从句，对 the set of information about a particular customer 进行补充说明。

英语中，定语从句还可以由名词（代词/数词）+ of + which（whom）来引导，表示部分与整体的关系。注意不要误用 which 和 whom。which 指物，whom 用来指人。请看下例：

Peter’s father knows a lot of people, many of whom are professors.

彼得的爸爸认识许多人，其中许多是教授。

She bought many books yesterday, five of which are on ERP.

她昨天买了许多书，其中 5 本是 ERP 方面的。

✍ Exercises

【Ex. 1】 根据课文内容回答问题。

1. What is a data structure?
2. What is an array in computer programming languages?
3. Must the file have a unique name within its own directory?
4. How do some operating systems and applications describe files with given formats?
5. How is the organization of data in the record usually prescribed?
6. What is a table in computer programming?
7. What is a table in data processing?
8. What is a table in a relational database?
9. What is a decision table often called? What does it contain?
10. What is an HTML table used to do?

【Ex. 2】 根据下面的英文解释，写出相应的英文词汇。

1. __________: A signal to a computer that stops the execution of a running program so that another action can be performed.
2. __________: A collection of related, often adjacent items of data, treated as a unit.
3. __________: In word processing, a block of text formatted in aligned rows and columns.
4. __________: A multi-element data structure that has a linear organization but that allows elements to be added or removed in any order.
5. __________: A distinguishing character or symbol written directly beneath or next to and slightly below a letter or number.
6. __________: An affix added to the end of a word or stem.
7. __________: To make or write a definition.
8. __________: A series of objects placed next to each other, usually in a straight line.
9. __________: A bit or a set of bits on a magnetic storage device making up the smallest addressable unit of information.
10. __________: To organize data, typically a set of records, in a particular order.

【Ex. 3】 把下列句子翻译为中文。

1. Star topologies are normally implemented using twisted pair cable, specifically unshielded twisted pair (UTP).
2. A video card is the part of your computer that transforms video data into the visual display you see on your monitor.
3. A multi-user operating system allows many different users to take advantage of the computer’s resources simultaneously.
4. Address is the unique location of an information site on the Internet, a specific file (for example, a Web page), or an E-mail user.
5. Over the years, ARPA has funded many projects in computer science research, many of which had a profound effect on the state of the art.
6. In truth of course by making the creation of more complex software practical, computer languages have merely created new types of software bugs.
7. A computer virus is a program designed to spread itself by first infecting executable files or the system areas of hard and floppy disks and then making copies of itself.
8. When the entire RAM is being used (for example if there are many programs open at the same time) the computer will swap data to the hard drive and back to give the impression that there is slightly more memory.
9. The compiler ignores all comments.
10. You can E-mail your document without ever leaving word.

【Ex. 4】 将下列词填入适当的位置（每词只用一次）。

records	leaves	beyond	random	database
two	power	depth	nodes	end

A binary tree is a method of placing and locating files (called records or keys) in a __(1)__, especially when all the data is known to be in __(2)__ access memory (RAM). The algorithm finds data by repeatedly dividing the number of ultimately accessible __(3)__ in half until only one remains.

In a tree, records are stored in locations called __(4)__. This name derives from the fact that records always exist at __(5)__ points; there is nothing __(6)__ them. Branch points are called __(7)__. The order of a tree is the number of branches (called children) per node. In a binary tree, there are always __(8)__ children per node, so the order is 2. The number of leaves in a binary tree is always a __(9)__ of 2. The number of access operations required to reach the desired record is called the __(10)__ of the tree.

Text B

Structured Data, Semi-structured Data and Unstructured Data

1. Structured Data

Structured data refers to any data that resides in a fixed field within a record or file. This includes data contained in relational databases and spreadsheets.

1.1 Characteristics of Structured Data

Structured data first depends on creating a data model – a model of the types of business data that will be recorded and how they will be stored, processed and accessed. This includes defining what fields of data will be stored and how that data will be stored: data type (numeric, currency, alphabetic, name, date, address) and any restrictions on the data input (number of characters; restricted to certain terms such as Mr., Ms. or Dr.; M or F).

Structured data has the advantage of being easily entered, stored, queried and analyzed. At one time, because of the high cost and performance limitations of storage, memory and processing, relational databases and spreadsheets using structured data were the only ways to effectively manage data. Anything that couldn’t fit into a tightly organized structure would have to be stored on paper in a filing cabinet.

1.2 Managing Structured Data

Structured data is often managed using Structured Query Language (SQL) – a programming language created for managing and querying data in relational database management systems. Originally developed by IBM in the early 1970s and later developed commercially by Relational Software, Inc. (now Oracle Corporation).

Structured data was a huge improvement over strictly paper-based unstructured systems, but life doesn’t always fit into neat little boxes. As a result, the structured data always had to be supplemented by paper or microfilm storage. As technology performance has continued to improve, and prices have dropped, it was possible to bring into computing systems unstructured and semi-structured data.

1.3 Structured Data Technology Standards

SQL has been a standard of the American National Standards Institute since 1986. It is managed by InterNational Committee for Information Technology Standards (INCITS) Technical Committee DM 32 – Data Management and Interchange. The committee has two task groups, one for databases and the other for metadata. HP, CA, IBM, Microsoft, Oracle, Sybase (SAP) and Teradata all participate, as well as several federal government agencies. Both of the committee project documents have links to further information on each project. SQL became an International Organization for Standards (ISO) standard in 1987. The published standards are available for purchase from the ANSI eStandards Store, under the INCITS/ISO/IEC 9075 classification.

2. Semi-structured Data

Semi-structured data is a form of structured data that does not conform with the formal structure of data models associated with relational databases or other forms of data tables, but nonetheless, contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data. Therefore, it is also known as self-describing structure.

In semi-structured data, the entities belonging to the same class may have different attributes even though they are grouped together, and the attributes’ order is not important.

Semi-structured data are increasingly occurring since the advent of the Internet where full-text documents and databases are not the only forms of data anymore, and different applications need a medium for exchanging information. In object-oriented databases, one often finds semi-structured data.

2.1 Types of Semi-structured data

2.1.1 XML

XML, other markup languages, email, and EDI are all forms of semi-structured data. OEM (Object Exchange Model) was created prior to XML as a means of self-describing a data structure. XML has been popularized by web services that are developed utilizing SOAP principles.

Some types of data described here as“semi-structured”, especially XML, suffer from the impression that they are incapable of structural rigor at the same functional level as Relational Tables and Rows. Indeed, the view of XML as inherently semi-structured (previously, it was referred to as “unstructured”) has handicapped its use for a widening range of data-centric applications. Even documents, normally thought of as the epitome of semi-structure, can be designed with virtually the same rigor as database schema, enforced by the XML schema and processed by both commercial and custom software programs without reducing their usability by human readers.

In view of this fact, XML might be referred to as having “flexible structure” capable of human-centric flow and hierarchy as well as highly rigorous element structure and data typing.

2.1.2 JSON

JSON or JavaScript Object Notation, is an open standard format that uses human-readable text to transmit data objects consisting of attribute–value pairs. It is used primarily to transmit data between a server and web application, as an alternative to XML. JSON has been popularized by web services developed utilizing REST principles.

There is a new breed of databases such as MongoDB and Couchbase that store data natively in JSON format, leveraging the pros of semi-structured data architecture.

2.2 Pros and Cons of Using a Semi-structured Data Format

2.2.1 Advantages

- Programmers persisting objects from their application to a database do not need to worry about object-relational impedance mismatch, but can often serialize objects via a light-weight library.
- Support for nested or hierarchical data often simplifies data models representing complex relationships between entities.
- Support for lists of objects simplifies data models by avoiding messy translations of lists into a relational data model.

2.2.2 Disadvantages

- The traditional relational data model has a popular and ready-made query language, SQL.
- Prone to “garbage in, garbage out” ; by removing restraints from the data model, there is less fore-thought that is necessary to operate a data application.

3. Unstructured Data

Unstructured data (or unstructured information) refers to information that either does not have a pre-defined data model or is not organized in a pre-defined manner. Unstructured information is typically text-heavy, but may contain data such as dates, numbers, and facts as well. This results in irregularities and ambiguities that make it difficult to understand using traditional programs as compared to data stored in fielded form in databases or annotated (semantically tagged) in documents.

In 1998, Merrill Lynch cited a rule of thumb that somewhere around 80%-90% of all potentially usable business information may originate in unstructured form. This rule of thumb is not based on primary or any quantitative research, but nonetheless is accepted by some.

IDC and EMC project that data will grow to 40 zettabytes by 2020, resulting in a 50-fold growth from the beginning of 2010. The Computer World magazine states that unstructured information might account for more than 70%-80% of all data in organizations.

3.1 Background

The earliest research into business intelligence focused in on unstructured textual data, rather than numerical data. As early as 1958, computer science researchers like H.P. Luhn were particularly concerned with the extraction and classification of unstructured text. However, only since the turn of the century has the technology caught up with the research interest. In 2004, the SAS Institute developed the SAS Text Miner, which uses Singular Value Decomposition (SVD) to reduce a hyper-dimensional textual space into smaller dimensions for significantly more efficient machine-analysis. The mathematical and technological advances sparked by machine textual analysis prompted a number of business to research applications, leading to the development of fields like sentiment analysis, voice of the customer mining, and call center optimization. The emergence of Big Data in the late 2000s led to a heightened interest in the applications of unstructured data analytics in contemporary fields such as predictive analytics and root cause analysis.

3.2 Issues with terminology

The term is imprecise for several reasons:

(1) Structure, while not formally defined, can still be implied.

(2) Data with some form of structure may still be characterized as unstructured if its structure is not helpful for the processing task at hand.

(3) Unstructured information might have some structure (semi-structured) or even be highly structured but in ways that are unanticipated or unannounced.

3.3 Dealing with unstructured data

Techniques such as data mining, natural language processing (NLP), and text analytics provide different methods to find patterns in, or otherwise interpret, this information. Common techniques for structuring text usually involve manual tagging with metadata or part-of-speech tagging for further text mining-based structuring. The Unstructured Information Management Architecture (UIMA) standard provided a common framework for processing this information to extract meaning and create structured data about the information.

Software that creates machine-processable structure can utilize the linguistic, auditory, and visual structure that exist in all forms of human communication. Algorithms can infer this inherent structure from text, for instance, by examining word morphology, sentence syntax, and other small- and large-scale patterns. Unstructured information can then be enriched and tagged to address ambiguities and relevancy-based techniques then used to facilitate search and discovery. Examples of “unstructured data” may include books, journals, documents, metadata, health records, audio, video, analog data, images, files, and unstructured text such as the body of an e-mail message, Web page, or word-processor document. While the main content being conveyed does not have a defined structure, it generally comes packaged in objects (e.g. in files or documents) that themselves have structure and are thus a mix of structured and unstructured data, but collectively this is still referred to as“unstructured data”. For example, an HTML web page is tagged, but HTML mark-up typically serves solely for rendering. It does not capture the meaning or function of tagged elements in ways that support automated processing of the information content of the page. XHTML tagging does allow machine processing of elements, although it typically does not capture or convey the semantic meaning of tagged terms.

Since unstructured data commonly occurs in electronic documents, the use of a content or document management system which can categorize entire documents is often preferred over data transfer and manipulation from within the documents. Document management thus provides the means to convey structure onto document collections.

Search engines have become popular tools for indexing and searching through such data, especially text.

New Words

characteristic	[ˌkæriktəˈristik]	*n.*特性，特征
		*adj.*特有的，表示特性的，典型的
model	[ˈmɔdəl]	*n.*模型，原型
		*vt.*模仿
		*v.*模拟
memory	[ˈmeməri]	*n.*存储器，内存
commercially	[kəˈməːʃəli]	*adv.*商业上，贸易上
huge	[hjuːdʒ]	*adj.*巨大的，极大的，无限的
supplement	[ˈsʌplimənt]	*n.* & *v.*补充
microfilm	[ˈmaikrəufilm]	*n.*缩影胶片
		*v.*缩微拍摄
interchange	[ˌintəˈtʃeindʒ]	*vt.*交换
committee	[kəˈmiti]	*n.*委员会
participate	[pɑːˈtisipeit]	*vi.*参与，参加，分享，分担
conform	[kənˈfɔːm]	*vt.*使一致，使遵守，使顺从
		*vi.*符合
tag	[tæg]	*n.*标签，标记符
		*vt.*加标签于
marker	[ˈmɑːkə]	*n.*标记
separate	[ˈsepəreit]	*adj.*分开的，分离的；个别的，单独的
		*v.*分开，隔离，分散
hierarchy	[ˈhaiərɑːki]	*n.*层次，层级
entity	[ˈentiti]	*n.*实体
increasingly	[inˈkriːsiŋli]	*adv.*日益，愈加
medium	[ˈmiːdjəm]	*n.*媒体，媒介
		*adj.*中间的，中等的
popularize	[ˈpɔpjuləraiz]	*v.*普及
rigor	[ˈrigə]	*n.*严格，严密，精确
inherently	[inˈhiərəntli]	*adv.*天性地，固有地
epitome	[iˈpitəmi]	*n.*摘要
virtually	[ˈvəːtjuəli]	*adv.*事实上，实质上
schema	[ˈskiːmə]	*n.*模式，方案

flexible	[ˈfleksəbl]	*adj.*灵活的，柔软的，能变形的
capable	[ˈkeipəbl]	*adj.*有能力的，能干的，有可能的
alternative	[ɔːlˈtəːnətiv]	*n.*二中择一，可供选择的办法、事物
		*adj.*选择性的，二中择一的
breed	[briːd]	*n.*品种，种类
natively	[ˈneitivli]	*adv.*本机地，本地地
mismatch	[misˈmætʃ]	*vt.*使配错，使配合不当
	[ˈmismætʃ]	*n.*错配
nested	[ˈnestid]	*adj.*嵌套的
simplify	[ˈsimplifai]	*vt.*单一化，简单化
restraint	[risˈtreint]	*n.*抑制，制止，克制
irregularity	[iˌregjuˈlæriti]	*n.*不规则，无规律
ambiguity	[ˌæmbiˈgjuːiti]	*n.*含糊，不明确
potentially	[pəˈtenʃəli]	*adv.*潜在地
background	[ˈbækgraund]	*n.*背景，后台
hyperdimensional	[ˌhaipəˌdaiˈmenʃənəl]	*adj.*多维的
spark	[spɑːk]	*v.*发动，触发； 激起，鼓舞
emergence	[iˈməːdʒəns]	*n.*浮现，露出，出现
terminology	[ˌtəːmiˈnɔlədʒi]	*n.*术语学
imprecise	[ˌimpriˈsais]	*adj.*不严密的，不精确的
implied	[imˈplaid]	*ad.*暗指的，含蓄的
unanticipated	[ˌʌnænˈtisipeitid]	*ad.*不曾预料到的
auditory	[ˈɔːditəri]	*ad.*耳的，听觉的
visual	[ˈvizjuəl]	*ad.*看的，视觉的，形象的
infer	[inˈfəː]	*v.*推断
morphology	[mɔːˈfɔlədʒi]	*n.*词法，词态学
relevancy	[ˈreləvənsi]	*n.*关联
capture	[ˈkæptʃə]	*n.* & *vt.*捕获

Phrases

structured data	结构化数据
semi-structured data	半结构化数据
unstructured data	非结构化数据
fixed field	固定字段
data type	数据类型
data input	数据输入

fit into	适合
filing cabinet	档案柜
database management system	数据局管理系统
American National Standards Institute	美国国家标准协会
semantic element	语义元素
belong to	属于
full-text document	全文本文档
object-oriented database	面向对象数据库
markup language	标记语言
suffer from	忍受，遭受
data-centric application	以数据为中心的应用
garbage in，garbage out	垃圾进、垃圾出；无用数据入、无用数据出
rule of thumb	经验法则，大拇指规则
somewhere around	大约
result in	导致，产生
account for	占据
focused in on	着重于，关注
be concerned with	注重
predictive analytic	预测分析
root cause analysis	根源分析法
may be characterized as	可以称为
at hand	在手边，在附近，即将到来
deal with	处理，涉及，安排
part-of-speech tagging	词性标记，词类标识，词类标注
sentence syntax	句法，语句结构，句子结构
analog data	模拟数据
serve for	充当，用作
search engine	搜索引擎
search through …	把……仔细搜寻一遍

Abbreviations

INCITS (InterNational Committee for Information Technology Standards)	国际信息技术标准委员会
EDI (Electronic Data Interchange)	电子数据交换
OEM (Object Exchange Model)	对象交换模型
SOAP (Simple Object Access Protocol)	简单对象访问协议

JSON (Java Script Object Notation)	Java 脚本对象符号
REST (Representational State Transfer)	表述性状态传递
IDC（International Data Corporation）	国际数据公司
SVD (Singular Value Decomposition)	奇异值分解
NLP (Natural Language Processing)	自然语言处理
UIMA (Unstructured Information Management Architecture)	非结构化信息管理体系结构
HTML (HyperText Markup Language)	超文本标记语言
XHTML (Extensible HyperText Markup Language)	扩展超文本标记语言

Exercises

【Ex. 5】 根据课文内容回答以下问题。

1. What does structured data refer to?
2. What advantage does structured data have?
3. How is structured data often managed?
4. What is SQL? When did it become an International Organization for Standards (ISO) standard?
5. What is semi-structured data?
6. What are the types of semi-structured data mentioned in the text?
7. What are the disadvantages of using a semi-structured data format?
8. What does unstructured data refer to?
9. What techniques are used to deal with unstructured data?
10. Why is the use of a content or document management system which can categorize entire documents often preferred over data transfer and manipulation from within the documents?

参考译文

数据结构

数据结构是组织和存储数据的特殊格式。一般数据结构类型包括数组、文件、记录、表、树等。所有数据结构的设计都是为了达到某一特定目的而组织数据，以便可以用适当的方式访问和工作。在计算机编程中，为了可以用多种算法工作，也可以选择或设计数据结构来存储数据。

1. 数组

（1）在数据存储中，数组是在多种设备上存储信息的方法。

（2）一般来说，数组是按照特定方法（例如，以列表或三维表）排列的许多项目。

（3）在计算机编程语言中，数组是具有相同属性、可以使用如加下标这样的技术分别访问的一组对象。

（4）在随机访问存储器中，数组是许多内存单元的排列。

2. 文件

（1）在数据处理中，文件是一些相关记录的集合。例如，可以把每个客户的记录放到一个文件中。依次地，每个记录由用于独立数据项的域组成，如客户姓名、客户编号、客户地址等。通过在每个记录相同域中提供同类信息（这样所有记录都一致），文件可方便地被计算机程序访问和处理。随着数据库的出现，使用这些术语已经不太重要了，而且它的重点在于用某一方法集合记录和域数据的表。在主机系统中，术语数据集通常与文件同义，但意味着它是可以由特定访问方式辨认的特定组织格式。取决于不同的操作系统，文件（和数据集）可以包含在一个类目、目录或文件夹中。

（2）在任一计算机系统，尤其是个人计算机中，文件是系统用户（包括系统自身及其应用程序）可用的数据实体，可以将其作为实体来处理（例如，从一个文件目录移动到另一个目录）。在自己的目录中，文件必须有唯一的名字。某些操作系统和应用程序通过给特定格式的文件特定的文件名后缀格式来描述文件。文件名后缀也称作文件扩展名。例如，程序或可执行文件有时给定或必须有“.exe”后缀。一般情况下，后缀往往在操作系统允许的字符数以内尽可能地描述文件的格式。

3. 记录

（1）在计算机数据处理中，记录是排列的、以备程序处理的数据项的集合。多项记录可以组成文件或数据集。以记录形式组织的结构数据通常由定义记录的组织结构的编程语言规定，并/或由处理数据的应用程序来定义。通常，记录可以有固定的长度，或带有包含在记录内的长度信息的可变长度。

（2）在数据库中，记录——有时也称作行——是与特定实体相关的表中的一组域。例如，在一个客户联系信息表中，一行中通常包含这样的域：ID 号、姓名、街道地址、城市、电话号码等。

4. 表

在计算机编程语言中，表是用来组织信息的数据结构，就像在纸上一样。有多种不

同的计算机关系表，用许多不同的方式来工作。下面列出比较普通的类型。

（1）在数据处理中，表也称作数组，是组织好的一组域。表可以存储相对不变的数据，也可被频繁更新。例如，包含在磁盘卷号中的表在写扇区时就被更新。

（2）在关系数据库中，表有时也称作文件，它把单一标题的信息组织到行和列中。例如，商业数据库通常包含客户信息表，该表中会用许多列来存储客户的账号、地址、电话号码等。数据的每个单一段（如账号）是表中的一个域。一列包含单个域中的全部项，如全部客户的电话号码。依次地，域被组织成记录，是信息的完整集合，如特定客户的信息集合，每条信息一行。这个规范化处理决定了如何有效地把数据组织到表中。

（3）决策表，通常称作真值表，可以用计算机或在纸上简单画出，它包含一系列的决策和做出决策所依据的标准。用于决策的各种可能出现的情况都必须列出，每种情况下采用的行动都应该被指定。一个简单的例子是：对于交通路口，通行决策也许可以表达为是与否，标准也许是红灯亮或绿灯亮。

决策表可以插入计算机程序中，以便根据不同的情况做出不同的决策。决策表的改变会反映在程序中。

（4）HTML 表用来在空间上组织网页元素，或建立可以按照表格形式更好地显示数据的数据结构，如列表或清单。