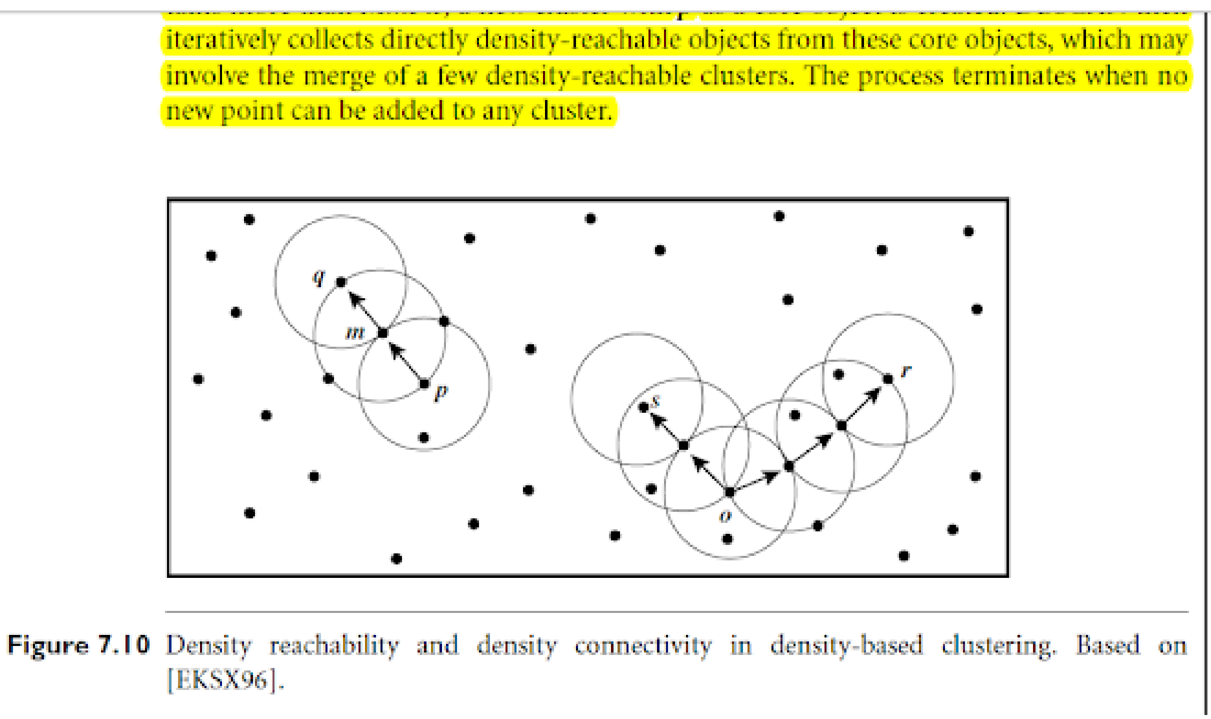
← DM Unit 5

CIE3 Answers

https://drive.google.com/file/u/3/d/1n-9ZyiuMDcSG7sZBCXb_vbB15bHxJZLq/view?usp=sharing

1a)
DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density based
clustering algorithm. The algorithm grows regions with sufficiently high density into clusters and discovers clusters of arbitrary shape in spatial databases with noise.
It defines a cluster as a maximal set of density-connected points.
The basic ideas of density-based clustering involve a number of new definitions. We
intuitively present these definitions, and then follow up with an example.

- The neighborhood within a radius e of a given object is called the e-neighborhood of
the object.
- If the e-neighborhood of an object contains at least a minimum number, MinPts, of objects, then the object is called a core object.
- Given a set of objects, D, we say that an object p is directly density-reachable from
object q if p is within the e-neighborhood of q, and q is a core object.
- An object p is density-reachable fromobject q with respect to e andMinPts in a set of
objects, D, if there is a chain of objects p1, : : : , pn, where p1 = q and pn = p such that
- pi+1 is directly density-reachable from pi with respect to e and MinPts, for 1 i n,
pi 2 D.
- An object p is density-connected to object q with respect to e and MinPts in a set of
objects, D, if there is an object o 2 D such that both p and q are density-reachable
from o with respect to e and MinPts.



iteratively collects directly density-reachable objects from these core objects, which may involve the merge of a few density-reachable clusters. The process terminates when no new point can be added to any cluster.

Figure 7.10 Density reachability and density connectivity in density-based clustering. Based on [EKSX96].

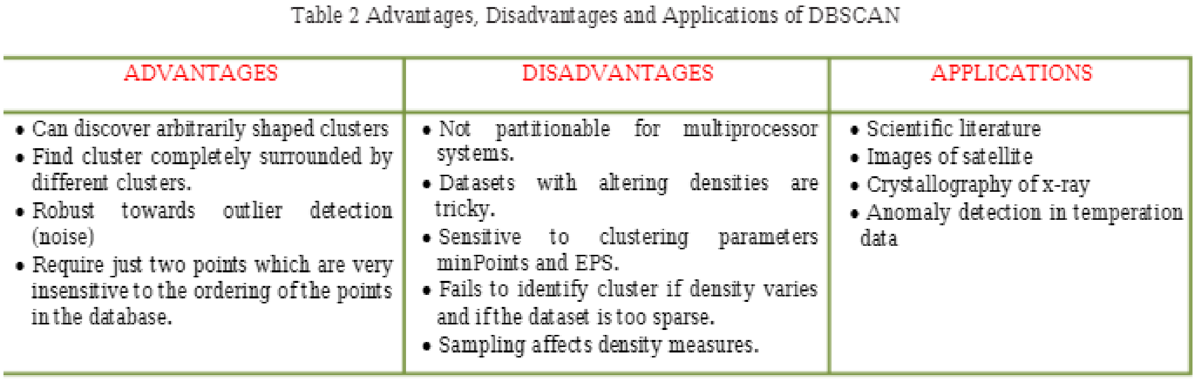


Table 2 Advantages, Disadvantages and Applications of DBSCAN

ADVANTAGES	DISADVANTAGES	APPLICATIONS
• Can discover arbitrarily shaped clusters • Find cluster completely surrounded by different clusters. • Robust towards outlier detection (noise) • Require just two points which are very insensitive to the ordering of the points in the database.	• Not partitionable for multiprocessor systems. • Datasets with altering densities are tricky. • Sensitive to clustering parameters minPoints and EPS. • Fails to identify cluster if density varies and if the dataset is too sparse. • Sampling affects density measures.	• Scientific literature • Images of satellite • Crystallography of x-ray • Anomaly detection in temperation data

(**Strengths and weakness of DBSCAN reference:** https://www.researchgate.net/publication/271520302_Performance_Evaluation_of_Clustering_Algorithm_Using_Different_Datasets)

1b)

The basic structure of a Web page is its DOM4 structure. When a Web page is presented to the user, the spatial and visual cues can help the user unconsciously divide the Web page into several semantic parts. It is possible to automatically segment the Web pages by using the spatial and visual cues.an algorithm called VIsion-based Page Segmentation (VIPS). VIPS aims to extract the semantic structure of a Web page based on its visual presentation.It first extracts all of the suitable blocks from the HTML DOM tree, and then it finds the separators between these blocks. Based on these separators, the semantic tree of the Web page is constructed. A Web page can be represented as a set of blocks

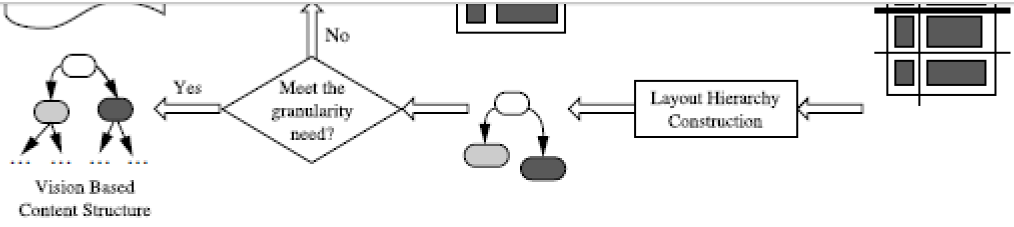
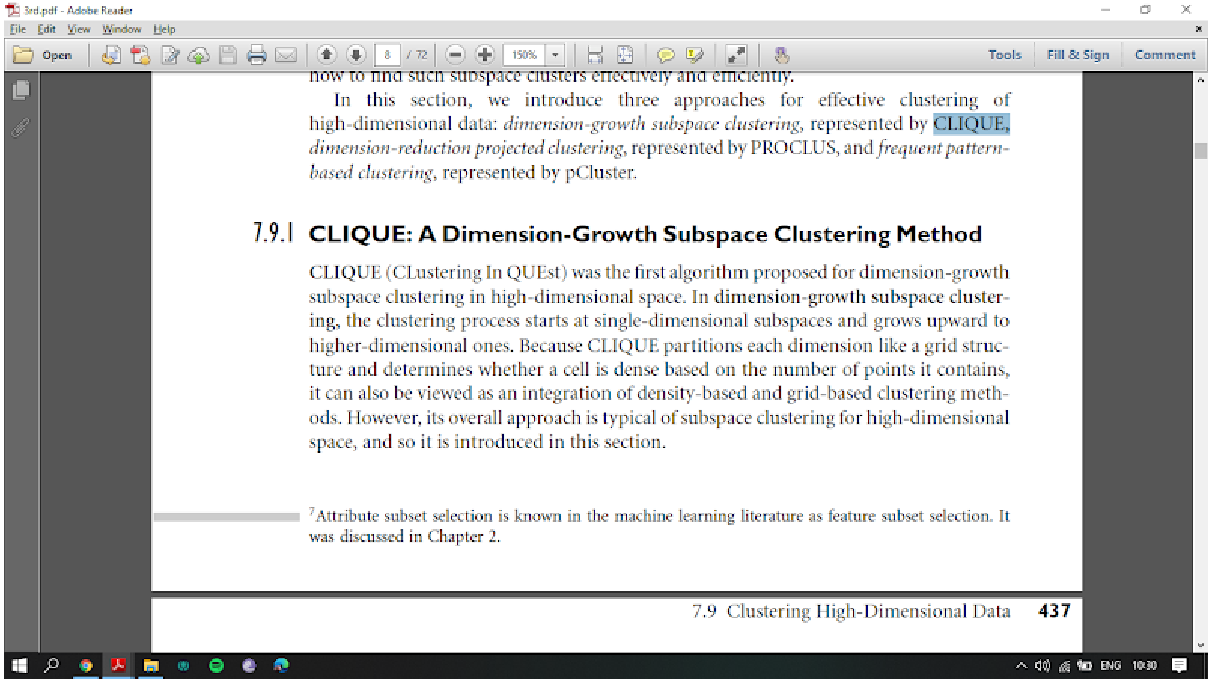

No
Yes
Meet the granularity need?
Layout Hierarchy Construction
Vision Based Content Structure

Figure 10.8 The process flow of vision-based page segmentation algorithm.

2a)



now to find such subspace clusters effectively and efficiently.

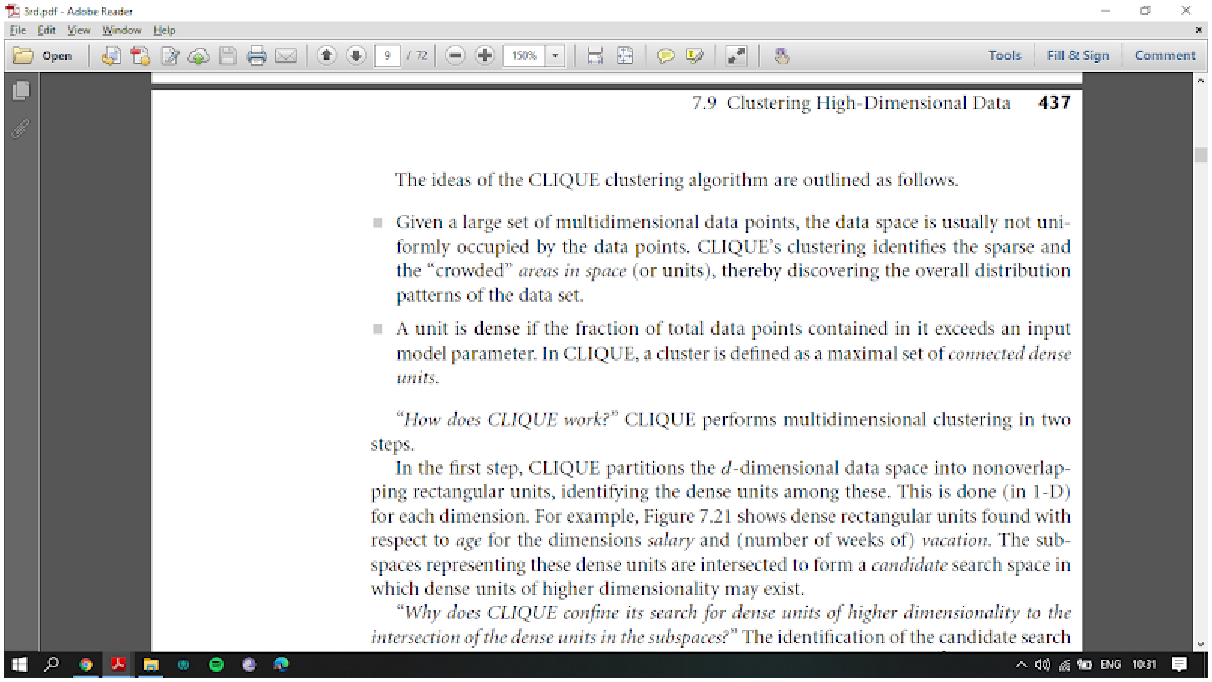
In this section, we introduce three approaches for effective clustering of high-dimensional data: *dimension-growth subspace clustering*, represented by CLIQUE, *dimension-reduction projected clustering*, represented by PROCLUS, and *frequent pattern-based clustering*, represented by pCluster.

7.9.1 CLIQUE: A Dimension-Growth Subspace Clustering Method

CLIQUE (CLustering In QUEst) was the first algorithm proposed for dimension-growth subspace clustering in high-dimensional space. In **dimension-growth subspace clustering**, the clustering process starts at single-dimensional subspaces and grows upward to higher-dimensional ones. Because CLIQUE partitions each dimension like a grid structure and determines whether a cell is dense based on the number of points it contains, it can also be viewed as an integration of density-based and grid-based clustering methods. However, its overall approach is typical of subspace clustering for high-dimensional space, and so it is introduced in this section.

[7] Attribute subset selection is known in the machine learning literature as feature subset selection. It was discussed in Chapter 2.

7.9 Clustering High-Dimensional Data **437**



7.9 Clustering High-Dimensional Data **437**

The ideas of the CLIQUE clustering algorithm are outlined as follows.

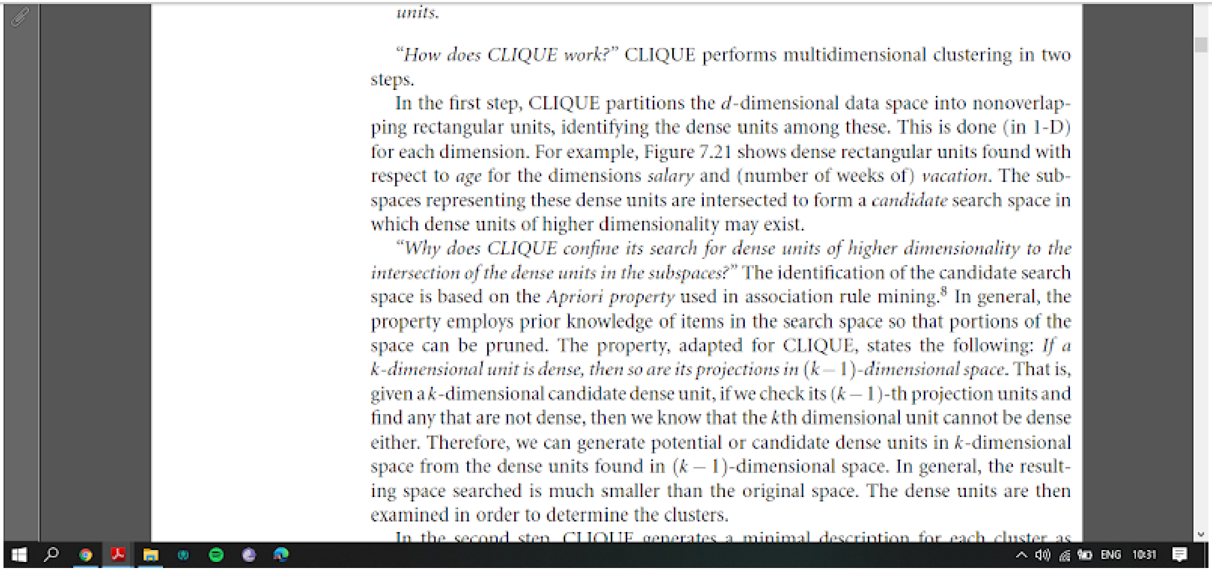
- Given a large set of multidimensional data points, the data space is usually not uniformly occupied by the data points. CLIQUE's clustering identifies the sparse and the "crowded" *areas in space* (or **units**), thereby discovering the overall distribution patterns of the data set.
- A unit is **dense** if the fraction of total data points contained in it exceeds an input model parameter. In CLIQUE, a cluster is defined as a maximal set of *connected dense units*.

"How does CLIQUE work?" CLIQUE performs multidimensional clustering in two steps.

In the first step, CLIQUE partitions the *d*-dimensional data space into nonoverlapping rectangular units, identifying the dense units among these. This is done (in 1-D) for each dimension. For example, Figure 7.21 shows dense rectangular units found with respect to *age* for the dimensions *salary* and (number of weeks of) *vacation*. The subspaces representing these dense units are intersected to form a *candidate* search space in which dense units of higher dimensionality may exist.

"Why does CLIQUE confine its search for dense units of higher dimensionality to the intersection of the dense units in the subspaces?" The identification of the candidate search

DM Unit 5



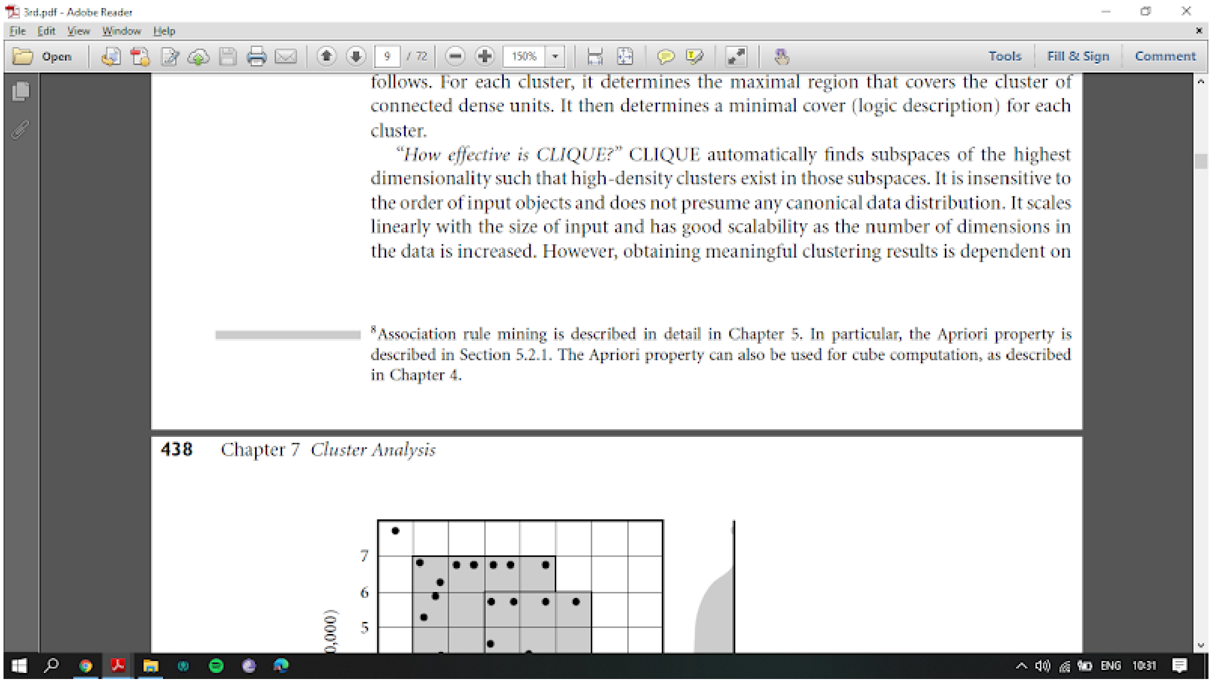
units.

"*How does CLIQUE work?*" CLIQUE performs multidimensional clustering in two steps.

In the first step, CLIQUE partitions the d-dimensional data space into nonoverlapping rectangular units, identifying the dense units among these. This is done (in 1-D) for each dimension. For example, Figure 7.21 shows dense rectangular units found with respect to *age* for the dimensions *salary* and (number of weeks of) *vacation*. The subspaces representing these dense units are intersected to form a *candidate* search space in which dense units of higher dimensionality may exist.

"*Why does CLIQUE confine its search for dense units of higher dimensionality to the intersection of the dense units in the subspaces?*" The identification of the candidate search space is based on the *Apriori property* used in association rule mining.[8] In general, the property employs prior knowledge of items in the search space so that portions of the space can be pruned. The property, adapted for CLIQUE, states the following: *If a k-dimensional unit is dense, then so are its projections in $(k-1)$-dimensional space*. That is, given a k-dimensional candidate dense unit, if we check its $(k-1)$-th projection units and find any that are not dense, then we know that the kth dimensional unit cannot be dense either. Therefore, we can generate potential or candidate dense units in k-dimensional space from the dense units found in $(k-1)$-dimensional space. In general, the resulting space searched is much smaller than the original space. The dense units are then examined in order to determine the clusters.

In the second step, CLIQUE generates a minimal description for each cluster as

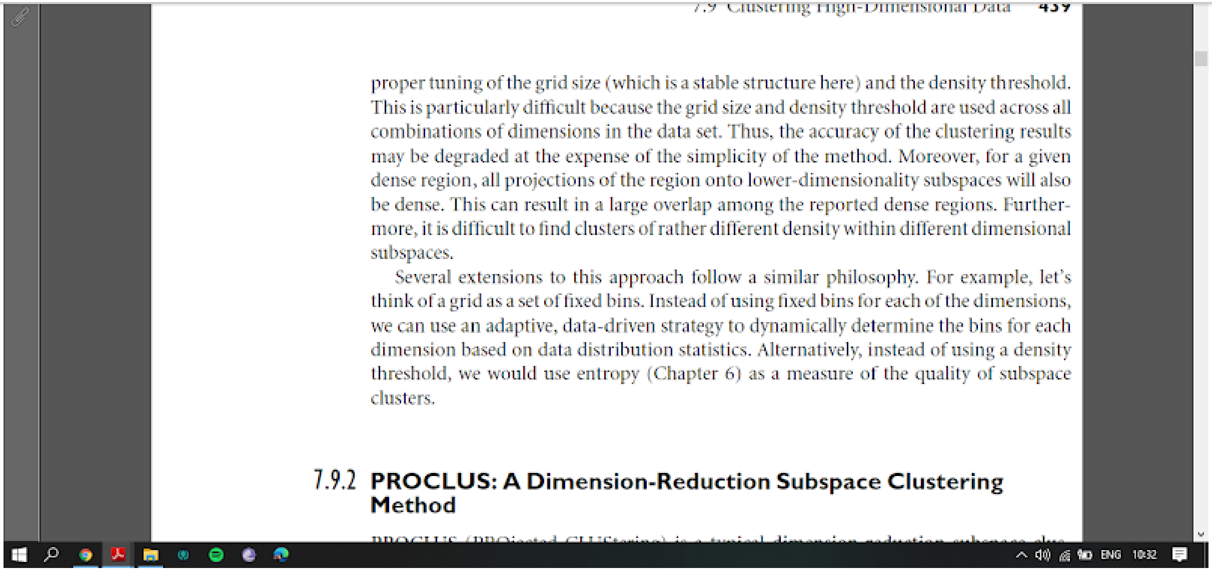


follows. For each cluster, it determines the maximal region that covers the cluster of connected dense units. It then determines a minimal cover (logic description) for each cluster.

"*How effective is CLIQUE?*" CLIQUE automatically finds subspaces of the highest dimensionality such that high-density clusters exist in those subspaces. It is insensitive to the order of input objects and does not presume any canonical data distribution. It scales linearly with the size of input and has good scalability as the number of dimensions in the data is increased. However, obtaining meaningful clustering results is dependent on

[8] Association rule mining is described in detail in Chapter 5. In particular, the Apriori property is described in Section 5.2.1. The Apriori property can also be used for cube computation, as described in Chapter 4.

438 Chapter 7 *Cluster Analysis*



7.9 Clustering High-Dimensional Data 439

proper tuning of the grid size (which is a stable structure here) and the density threshold. This is particularly difficult because the grid size and density threshold are used across all combinations of dimensions in the data set. Thus, the accuracy of the clustering results may be degraded at the expense of the simplicity of the method. Moreover, for a given dense region, all projections of the region onto lower-dimensionality subspaces will also be dense. This can result in a large overlap among the reported dense regions. Furthermore, it is difficult to find clusters of rather different density within different dimensional subspaces.

Several extensions to this approach follow a similar philosophy. For example, let's think of a grid as a set of fixed bins. Instead of using fixed bins for each of the dimensions, we can use an adaptive, data-driven strategy to dynamically determine the bins for each dimension based on data distribution statistics. Alternatively, instead of using a density threshold, we would use entropy (Chapter 6) as a measure of the quality of subspace clusters.

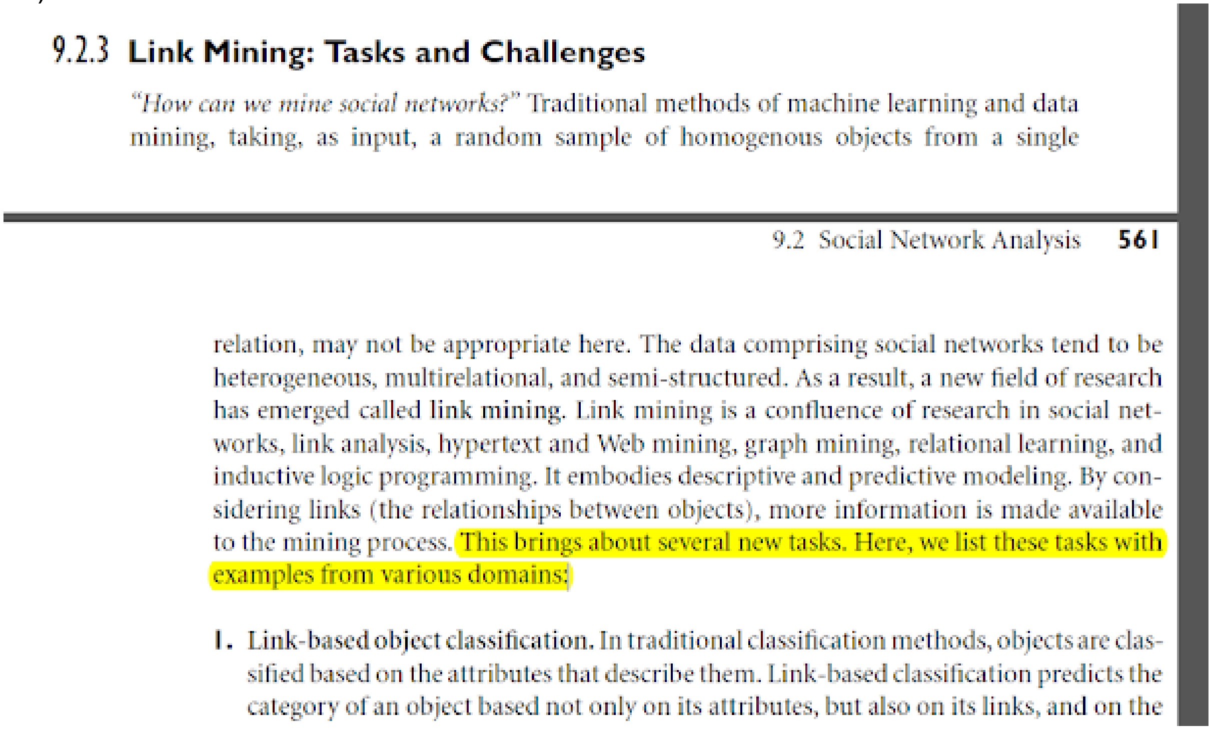
7.9.2 PROCLUS: A Dimension-Reduction Subspace Clustering Method

PROCLUS (PROjected CLUStering) is a typical dimension-reduction subspace clus-

2b)

The secrecy of authority is hiding in Web page linkages. The Web consists not only of pages, but also of hyperlinks pointing from one page to another.These hyperlinks contain an enormous amount of latent human annotation that can help automatically infer the notion of authority. When an author of a Web page creates a hyperlink pointing to another Web page, this can be considered as the author's endorsement of the other page. The collective endorsement of a given page by different authors on the Web may indicate the importance of the page and may naturally lead to the discovery of authoritativeWeb pages. Therefore, the tremendous amount ofWeb linkage information provides rich information about the relevance, the quality, and the structure of theWeb's contents, and thus is a rich source forWeb mining.

3a)



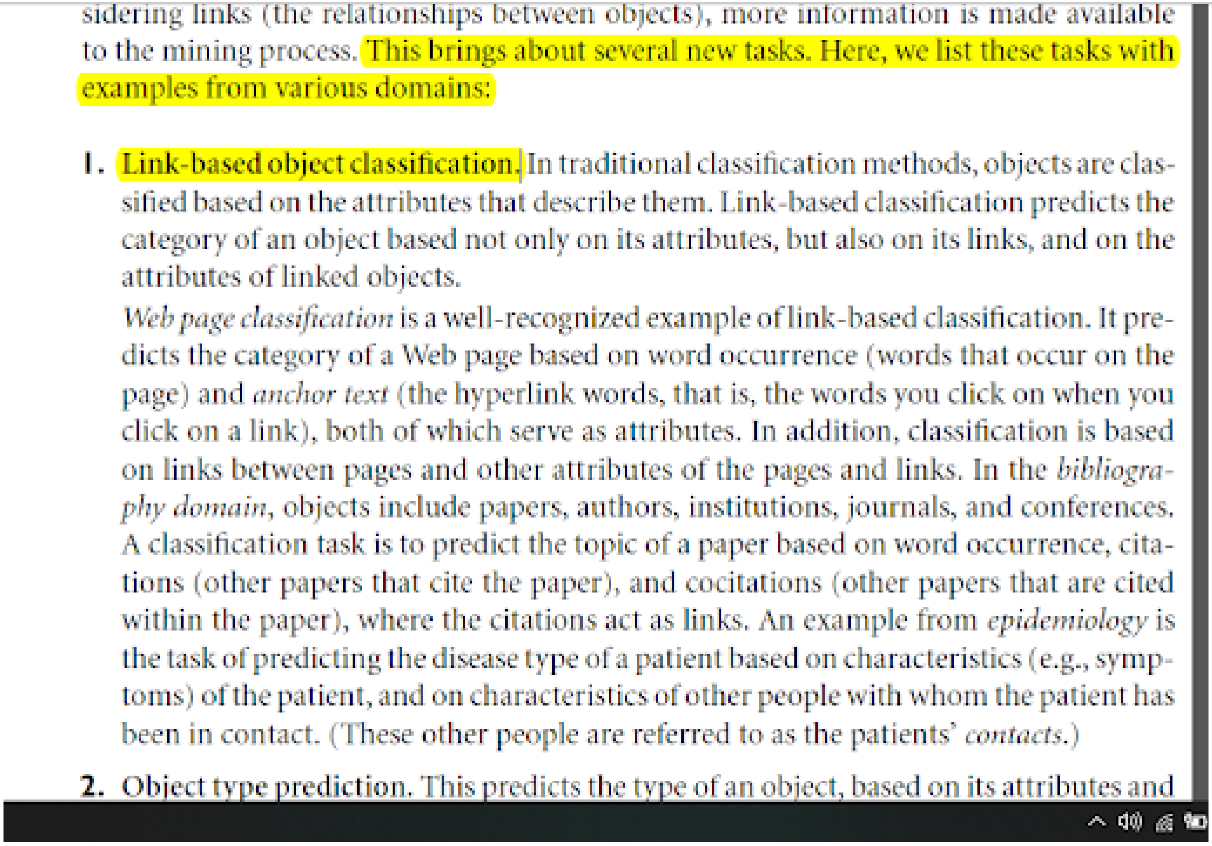
9.2.3 Link Mining: Tasks and Challenges

"How can we mine social networks?" Traditional methods of machine learning and data mining, taking, as input, a random sample of homogenous objects from a single

9.2 Social Network Analysis 561

relation, may not be appropriate here. The data comprising social networks tend to be heterogeneous, multirelational, and semi-structured. As a result, a new field of research has emerged called **link mining.** Link mining is a confluence of research in social networks, link analysis, hypertext and Web mining, graph mining, relational learning, and inductive logic programming. It embodies descriptive and predictive modeling. By considering links (the relationships between objects), more information is made available to the mining process. This brings about several new tasks. Here, we list these tasks with examples from various domains:

1. **Link-based object classification.** In traditional classification methods, objects are classified based on the attributes that describe them. Link-based classification predicts the category of an object based not only on its attributes, but also on its links, and on the

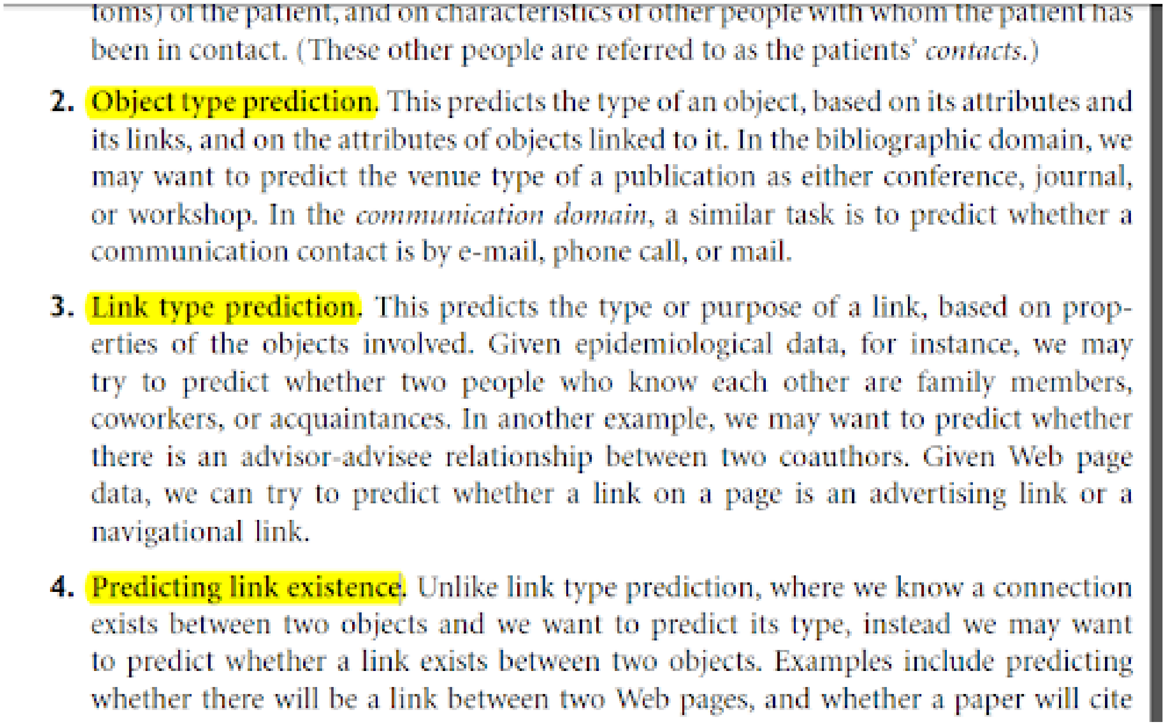


sidering links (the relationships between objects), more information is made available to the mining process. This brings about several new tasks. Here, we list these tasks with examples from various domains:

1. **Link-based object classification.** In traditional classification methods, objects are classified based on the attributes that describe them. Link-based classification predicts the category of an object based not only on its attributes, but also on its links, and on the attributes of linked objects.

   Web page classification is a well-recognized example of link-based classification. It predicts the category of a Web page based on word occurrence (words that occur on the page) and *anchor text* (the hyperlink words, that is, the words you click on when you click on a link), both of which serve as attributes. In addition, classification is based on links between pages and other attributes of the pages and links. In the *bibliography domain*, objects include papers, authors, institutions, journals, and conferences. A classification task is to predict the topic of a paper based on word occurrence, citations (other papers that cite the paper), and cocitations (other papers that are cited within the paper), where the citations act as links. An example from *epidemiology* is the task of predicting the disease type of a patient based on characteristics (e.g., symptoms) of the patient, and on characteristics of other people with whom the patient has been in contact. (These other people are referred to as the patients' *contacts.*)

2. **Object type prediction.** This predicts the type of an object, based on its attributes and



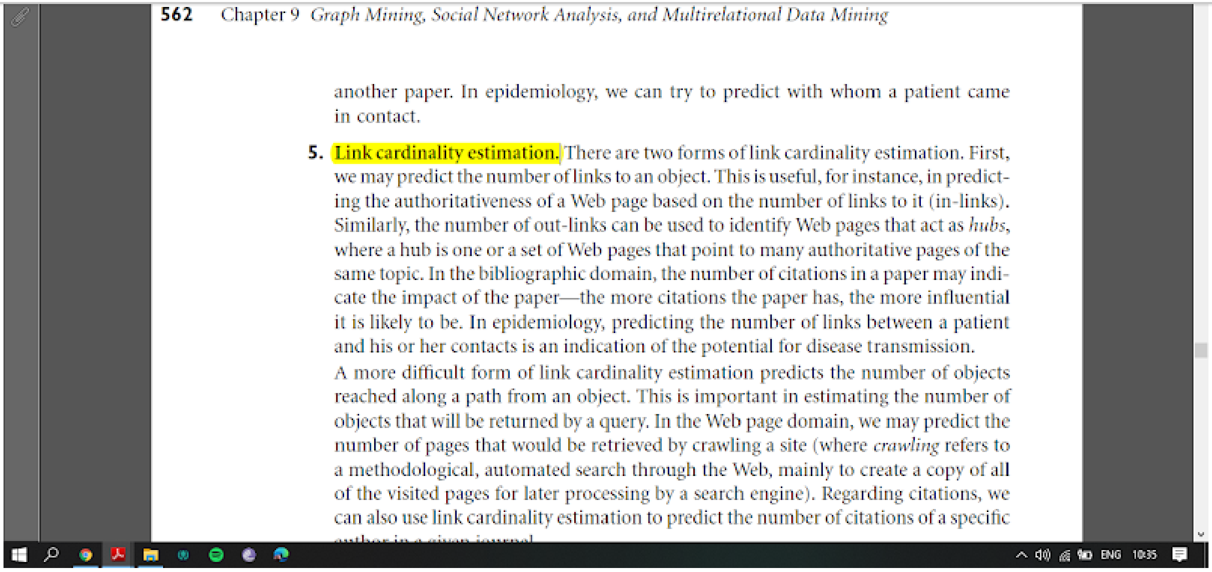
toms) of the patient, and on characteristics of other people with whom the patient has been in contact. (These other people are referred to as the patients' *contacts.*)

2. **Object type prediction.** This predicts the type of an object, based on its attributes and its links, and on the attributes of objects linked to it. In the bibliographic domain, we may want to predict the venue type of a publication as either conference, journal, or workshop. In the *communication domain*, a similar task is to predict whether a communication contact is by e-mail, phone call, or mail.

3. **Link type prediction.** This predicts the type or purpose of a link, based on properties of the objects involved. Given epidemiological data, for instance, we may try to predict whether two people who know each other are family members, coworkers, or acquaintances. In another example, we may want to predict whether there is an advisor-advisee relationship between two coauthors. Given Web page data, we can try to predict whether a link on a page is an advertising link or a navigational link.

4. **Predicting link existence.** Unlike link type prediction, where we know a connection exists between two objects and we want to predict its type, instead we may want to predict whether a link exists between two objects. Examples include predicting whether there will be a link between two Web pages, and whether a paper will cite

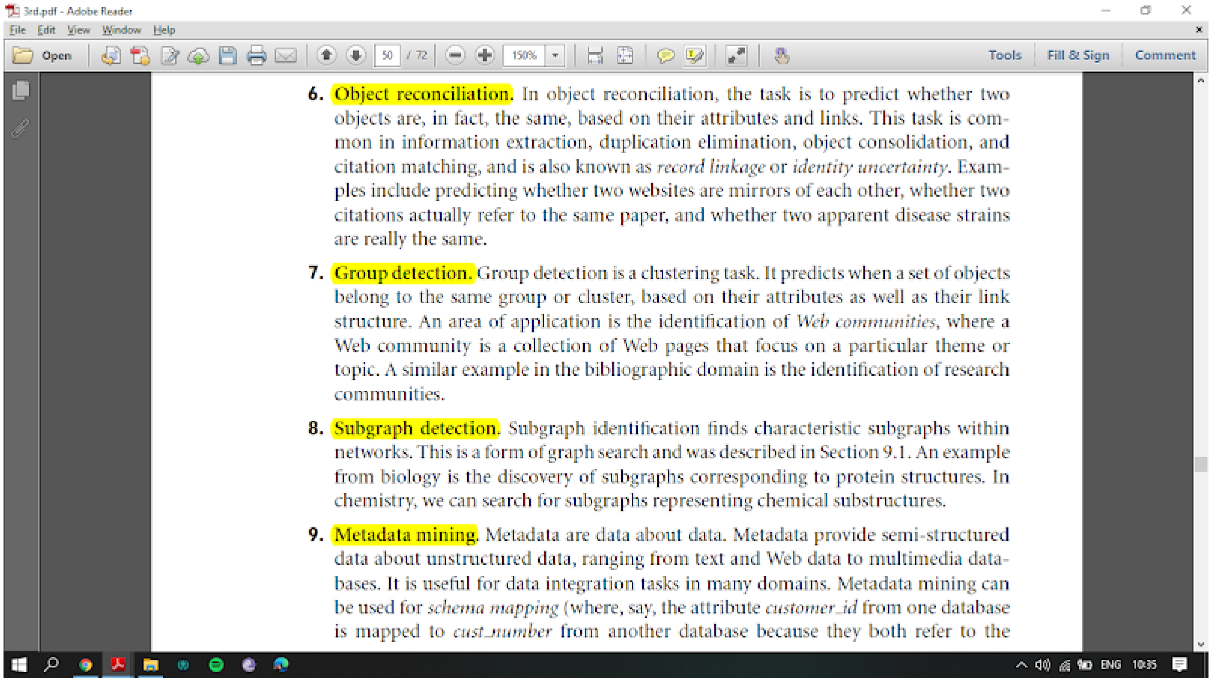
DM Unit 5



another paper. In epidemiology, we can try to predict with whom a patient came in contact.

5. **Link cardinality estimation.** There are two forms of link cardinality estimation. First, we may predict the number of links to an object. This is useful, for instance, in predicting the authoritativeness of a Web page based on the number of links to it (in-links). Similarly, the number of out-links can be used to identify Web pages that act as *hubs*, where a hub is one or a set of Web pages that point to many authoritative pages of the same topic. In the bibliographic domain, the number of citations in a paper may indicate the impact of the paper—the more citations the paper has, the more influential it is likely to be. In epidemiology, predicting the number of links between a patient and his or her contacts is an indication of the potential for disease transmission.

   A more difficult form of link cardinality estimation predicts the number of objects reached along a path from an object. This is important in estimating the number of objects that will be returned by a query. In the Web page domain, we may predict the number of pages that would be retrieved by crawling a site (where *crawling* refers to a methodological, automated search through the Web, mainly to create a copy of all of the visited pages for later processing by a search engine). Regarding citations, we can also use link cardinality estimation to predict the number of citations of a specific author in a given journal.



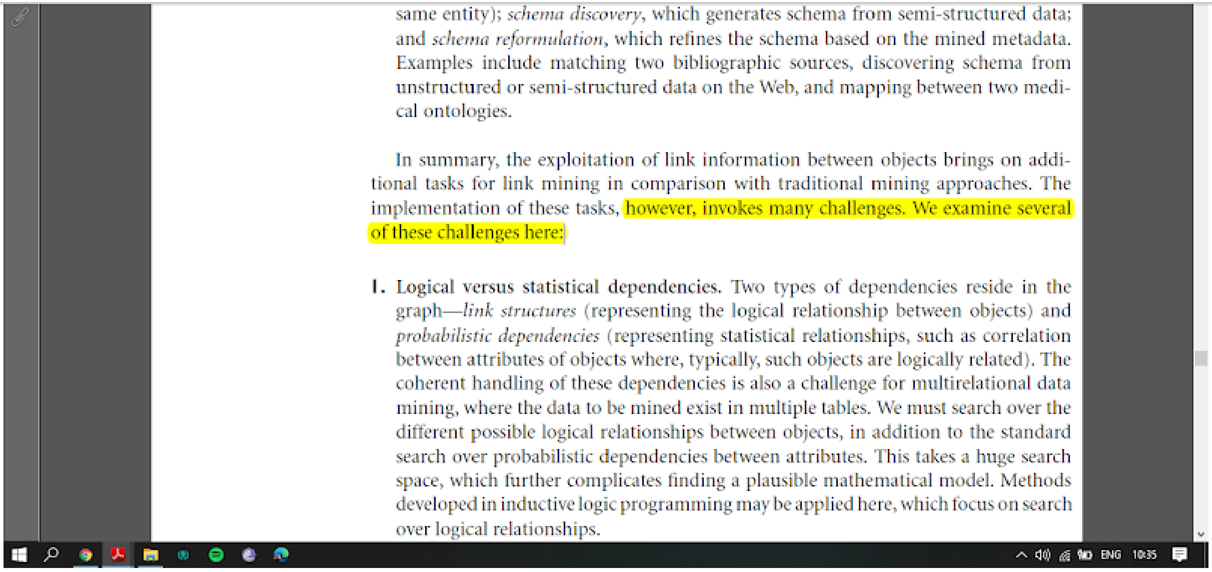
6. **Object reconciliation.** In object reconciliation, the task is to predict whether two objects are, in fact, the same, based on their attributes and links. This task is common in information extraction, duplication elimination, object consolidation, and citation matching, and is also known as *record linkage* or *identity uncertainty*. Examples include predicting whether two websites are mirrors of each other, whether two citations actually refer to the same paper, and whether two apparent disease strains are really the same.

7. **Group detection.** Group detection is a clustering task. It predicts when a set of objects belong to the same group or cluster, based on their attributes as well as their link structure. An area of application is the identification of *Web communities*, where a Web community is a collection of Web pages that focus on a particular theme or topic. A similar example in the bibliographic domain is the identification of research communities.

8. **Subgraph detection.** Subgraph identification finds characteristic subgraphs within networks. This is a form of graph search and was described in Section 9.1. An example from biology is the discovery of subgraphs corresponding to protein structures. In chemistry, we can search for subgraphs representing chemical substructures.

9. **Metadata mining.** Metadata are data about data. Metadata provide semi-structured data about unstructured data, ranging from text and Web data to multimedia databases. It is useful for data integration tasks in many domains. Metadata mining can be used for *schema mapping* (where, say, the attribute *customer_id* from one database is mapped to *cust_number* from another database because they both refer to the

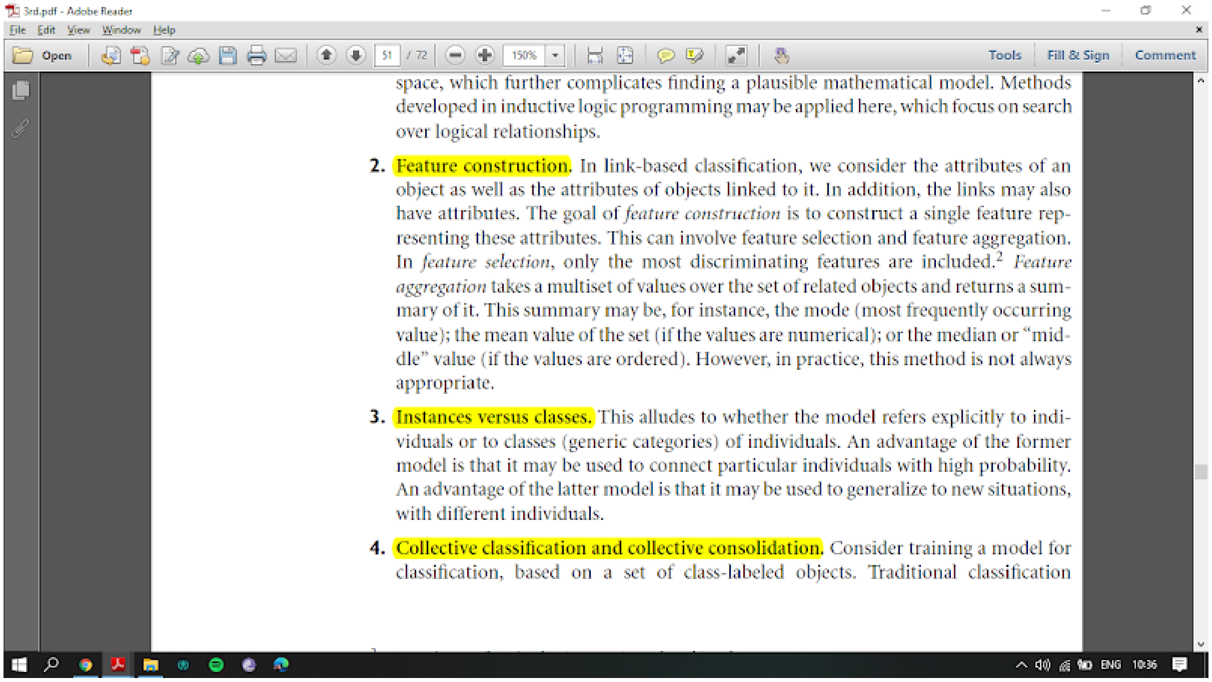
DM Unit 5



same entity); *schema discovery*, which generates schema from semi-structured data; and *schema reformulation*, which refines the schema based on the mined metadata. Examples include matching two bibliographic sources, discovering schema from unstructured or semi-structured data on the Web, and mapping between two medical ontologies.

In summary, the exploitation of link information between objects brings on additional tasks for link mining in comparison with traditional mining approaches. The implementation of these tasks, however, invokes many challenges. We examine several of these challenges here:

1. **Logical versus statistical dependencies.** Two types of dependencies reside in the graph—*link structures* (representing the logical relationship between objects) and *probabilistic dependencies* (representing statistical relationships, such as correlation between attributes of objects where, typically, such objects are logically related). The coherent handling of these dependencies is also a challenge for multirelational data mining, where the data to be mined exist in multiple tables. We must search over the different possible logical relationships between objects, in addition to the standard search over probabilistic dependencies between attributes. This takes a huge search space, which further complicates finding a plausible mathematical model. Methods developed in inductive logic programming may be applied here, which focus on search over logical relationships.

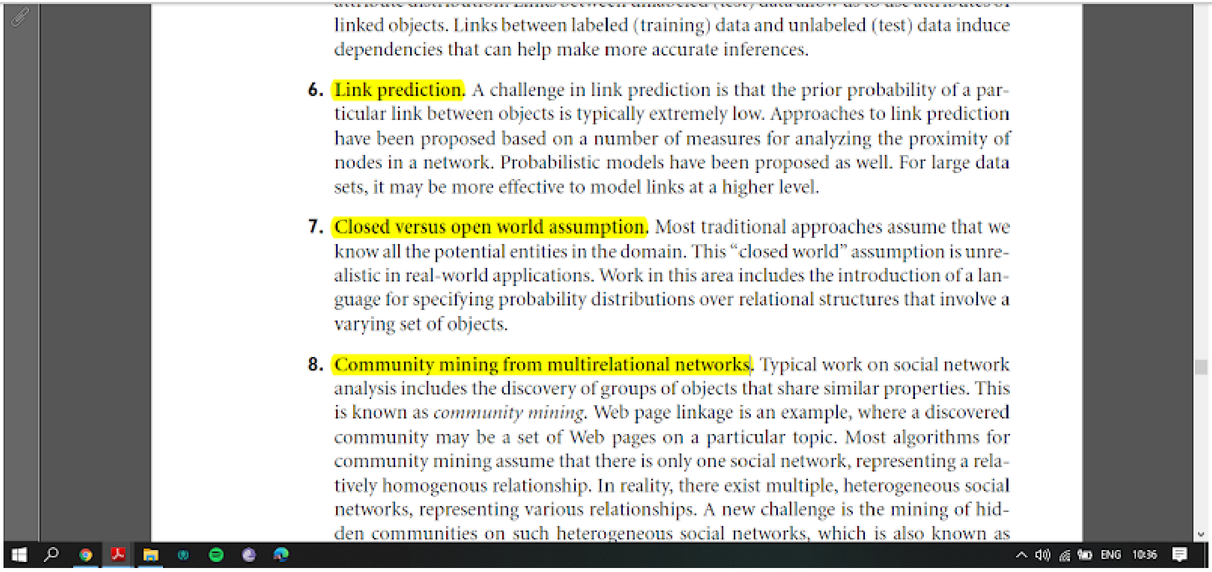


space, which further complicates finding a plausible mathematical model. Methods developed in inductive logic programming may be applied here, which focus on search over logical relationships.

2. **Feature construction.** In link-based classification, we consider the attributes of an object as well as the attributes of objects linked to it. In addition, the links may also have attributes. The goal of *feature construction* is to construct a single feature representing these attributes. This can involve feature selection and feature aggregation. In *feature selection*, only the most discriminating features are included.[2] *Feature aggregation* takes a multiset of values over the set of related objects and returns a summary of it. This summary may be, for instance, the mode (most frequently occurring value); the mean value of the set (if the values are numerical); or the median or "middle" value (if the values are ordered). However, in practice, this method is not always appropriate.

3. **Instances versus classes.** This alludes to whether the model refers explicitly to individuals or to classes (generic categories) of individuals. An advantage of the former model is that it may be used to connect particular individuals with high probability. An advantage of the latter model is that it may be used to generalize to new situations, with different individuals.

4. **Collective classification and collective consolidation.** Consider training a model for classification, based on a set of class-labeled objects. Traditional classification



attribute distribution. Links between unlabeled (test) data allow us to use attributes of linked objects. Links between labeled (training) data and unlabeled (test) data induce dependencies that can help make more accurate inferences.

6. Link prediction. A challenge in link prediction is that the prior probability of a particular link between objects is typically extremely low. Approaches to link prediction have been proposed based on a number of measures for analyzing the proximity of nodes in a network. Probabilistic models have been proposed as well. For large data sets, it may be more effective to model links at a higher level.

7. Closed versus open world assumption. Most traditional approaches assume that we know all the potential entities in the domain. This "closed world" assumption is unrealistic in real-world applications. Work in this area includes the introduction of a language for specifying probability distributions over relational structures that involve a varying set of objects.

8. Community mining from multirelational networks. Typical work on social network analysis includes the discovery of groups of objects that share similar properties. This is known as *community mining*. Web page linkage is an example, where a discovered community may be a set of Web pages on a particular topic. Most algorithms for community mining assume that there is only one social network, representing a relatively homogenous relationship. In reality, there exist multiple, heterogeneous social networks, representing various relationships. A new challenge is the mining of hidden communities on such heterogeneous social networks, which is also known as

3b)
Retail data mining can help identify customer buying behaviors,discover customer shopping patterns and trends, improve the quality of customer service, achieve better customer retention and satisfaction, enhance goods consumption ratios, design more effective goods transportation and distribution policies, and reduce the cost of business.

- **Design and construction of data warehouses based on the benefits of data mining:**The outcome of preliminary data mining exercises can be used to help guide the design and development of data warehouse structures. This involves deciding which dimensions and levels to include and what preprocessing to perform in order to facilitate effective data mining.
- **Multidimensional analysis of sales, customers, products, time, and region:** It is therefore important to provide powerful multidimensional analysis and visualization tools,including the construction of sophisticated data cubes according to the needs of data analysis. The multi feature data cube, introduced in, is a useful data structure in retail data analysis because it facilitates analysis on aggregates with complex conditions.

- **Analysis of the effectiveness of sales campaigns:** Multidimensional analysis can be used for this purpose by comparing the amount of sales and the number of transactions association analysis may disclose which items are likely to be purchased together with the items on sale, especially in comparison with the sales before or after the campaign.

- **Customer retention—analysis of customer loyalty**:Sequential pattern
mining can then be used to investigate changes in customer consumption or loyalty and suggest adjustments on the pricing
- **Product recommendation and cross-referencing of items:** Collaborative recommender systems use data mining techniques to make personalized product recommendations during live customer transactions, based on the opinions of other customers.

Unit 4 left out portions Pang Ning book pg no.622-633

UNIT 5

1. Discuss the use of data mining in intrusion detection. **658**

2. Write note on: **628**
i. Web content mining ii. Mining web link structures. **631**

3. Discuss in detail data mining for the Retail Industry.**651**
4. How can web page layout structure be mined? Elaborate **630**

5. What is information retrieval? Discuss any one application of information retrieval.**615**
6. Explain the various measures used to show the relationship between the set of relevant documents and the set of retrieved documents. **616(see the image) (measures is precision, recall , fscore)**

7. Elaborate on the areas in which data mining technology may be applied or developed for intrusion detection. **658(same as q no1)**
8. Explain at least five examples of data mining in the retail industry**651(same ans of q no. 3)**
9. Write note on Mining the Web's link structures to identify authoritative web pages.**631 (same as q 2.ii)**

10. What is Mining the World Wide Web? What are the challenges for effective resource and knowledge discovery in the web? **(same as q 2.i)**
Describe the ways of resolving these challenges.**628**

11. Discuss in detail data mining for the financial sector. **649**

12. Explain the following:

- Web usage mining **640**
- Basic measures for text retrieval. **616(same as q 6)**

13. Explain the steps of vision-based page segmentation algorithm with process flow. **631(second para), 632 diagram**

15. Write a short note on Spatial Data mining. [illegible]

16. What are the major text mining approaches based on the kinds of the data taken as input? Explain any 2. **624**

17. Discuss any four typical cases of how data analysis and data mining is used in banking and financial industry. **649 (same as q10)**

1) Data mining in Intrusion Detection - 658 - 65

Current traditional intrusion detection systems face many limitations. This has led to an increased interest in data mining for intrusion detection. The following are areas in which data mining technology may be applied or further developed for intrusion detection:

• **Development of data mining algorithms for intrusion detection:**Data mining algorithms can be used for misuse detection and anomaly detection. In misuse detection,training data are labeled as either "normal" or "intrusion." A classifier can then be derived to detect known intrusions.

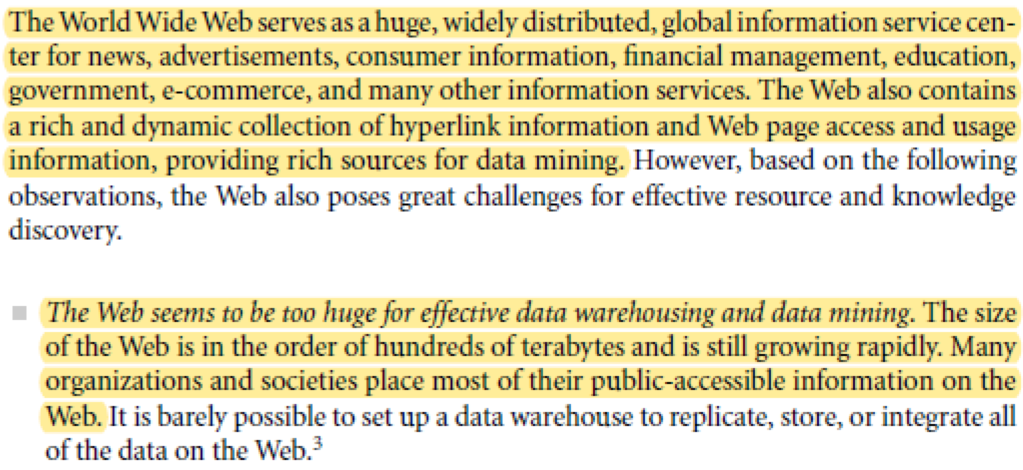
• **Association and correlation analysis, and aggregation to help select and build discriminating attributes:**Such information can provide insight regarding the selection of useful attributes for intrusion detection.

• **Analysis of stream data:** Due to the transient and dynamic nature of intrusions and malicious attacks, it is crucial to perform intrusion detection in the data stream environment.

• **Distributed data mining:** Intrusions can be launched from several different locations and targeted to many different destinations.

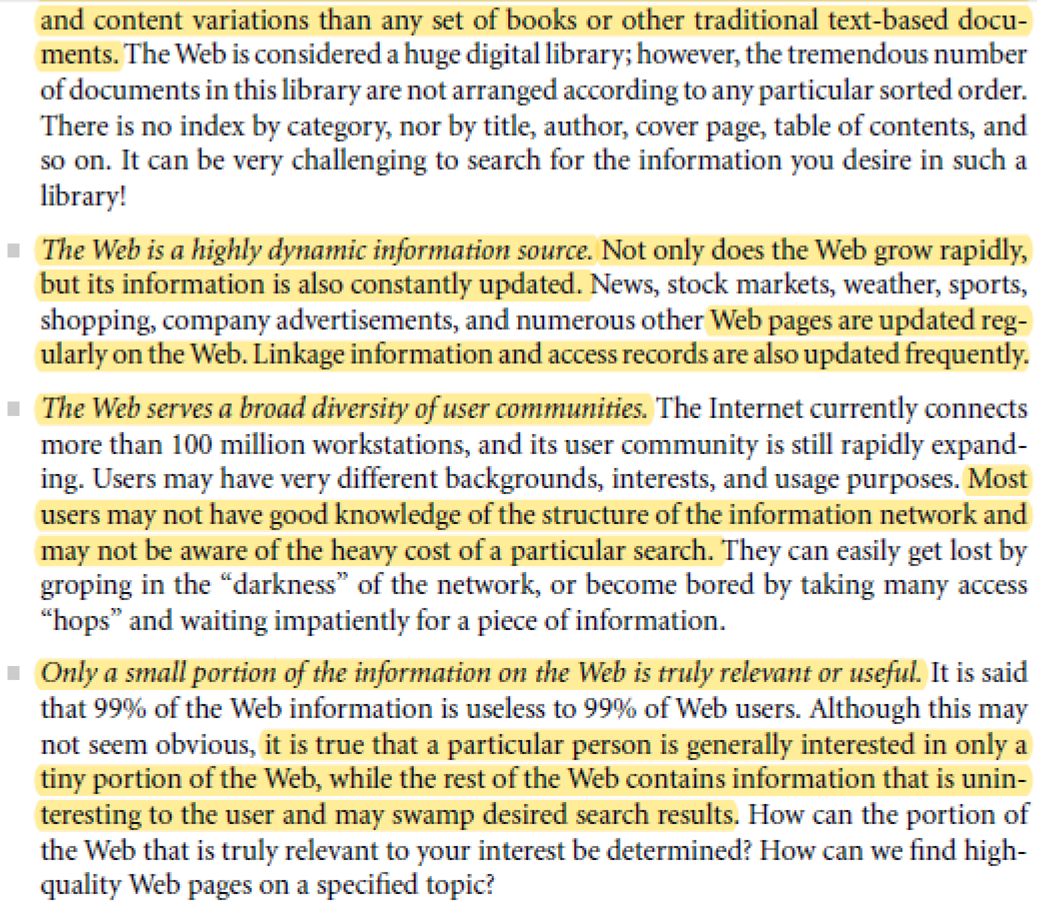
• **Visualization and querying tools:** Visualization tools should be available for viewing any anomalous patterns detected. Such tools may include features for viewing associations,clusters, and outliers. Intrusion detection systems should also have a graphical user interface that allows security analysts to pose queries regarding the network data or intrusion detection results

2. I) Web content mining



The World Wide Web serves as a huge, widely distributed, global information service center for news, advertisements, consumer information, financial management, education, government, e-commerce, and many other information services. The Web also contains a rich and dynamic collection of hyperlink information and Web page access and usage information, providing rich sources for data mining. However, based on the following observations, the Web also poses great challenges for effective resource and knowledge discovery.

- *The Web seems to be too huge for effective data warehousing and data mining.* The size of the Web is in the order of hundreds of terabytes and is still growing rapidly. Many organizations and societies place most of their public-accessible information on the Web. It is barely possible to set up a data warehouse to replicate, store, or integrate all of the data on the Web.[3]



and content variations than any set of books or other traditional text-based documents. The Web is considered a huge digital library; however, the tremendous number of documents in this library are not arranged according to any particular sorted order. There is no index by category, nor by title, author, cover page, table of contents, and so on. It can be very challenging to search for the information you desire in such a library!

- *The Web is a highly dynamic information source.* Not only does the Web grow rapidly, but its information is also constantly updated. News, stock markets, weather, sports, shopping, company advertisements, and numerous other Web pages are updated regularly on the Web. Linkage information and access records are also updated frequently.
- *The Web serves a broad diversity of user communities.* The Internet currently connects more than 100 million workstations, and its user community is still rapidly expanding. Users may have very different backgrounds, interests, and usage purposes. Most users may not have good knowledge of the structure of the information network and may not be aware of the heavy cost of a particular search. They can easily get lost by groping in the "darkness" of the network, or become bored by taking many access "hops" and waiting impatiently for a piece of information.
- *Only a small portion of the information on the Web is truly relevant or useful.* It is said that 99% of the Web information is useless to 99% of Web users. Although this may not seem obvious, it is true that a particular person is generally interested in only a tiny portion of the Web, while the rest of the Web contains information that is uninteresting to the user and may swamp desired search results. How can the portion of the Web that is truly relevant to your interest be determined? How can we find high-quality Web pages on a specified topic?

2. II) Mining web link structures.

The secrecy of authority is hiding in Web page linkages. The Web consists not only of pages, but also of hyperlinks pointing from one page to another.These hyperlinks contain an enormous amount of latent human annotation that can help automatically infer the notion of authority. When an author of a Web page creates a hyperlink pointing to another Web page, this can be considered as the author's endorsement of the other page. The collective endorsement of a given page by different authors on the Web may indicate the importance of the page and may naturally lead to the discovery of authoritativeWeb pages. Therefore, the tremendous amount ofWeb linkage information provides rich information about the relevance, the quality, and the structure of theWeb's contents, and thus is a rich source forWeb mining.

3) Discuss in detail data mining for the Retail Industry.651

Retail data mining can help identify customer buying behaviors,discover customer shopping patterns and trends, improve the quality of customer service, achieve better customer retention and satisfaction, enhance goods consumption ratios, design more effective goods transportation and distribution policies, and reduce the cost of business.

- **Design and construction of data warehouses based on the benefits of data mining:**The outcome of preliminary data mining exercises can be used to help guide the design and development of data warehouse structures. This involves deciding which dimensions multidimensional analysis and visualization tools,including the construction of sophisticated data cubes according to the needs of data analysis. The multi feature data cube, introduced in, is a useful data structure in retail data analysis because it facilitates analysis on aggregates with complex conditions.

• **Analysis of the effectiveness of sales campaigns:** Multidimensional analysis can be used for this purpose by comparing the amount of sales and the number of transactions association analysis may disclose which items are likely to be purchased together with the items on sale, especially in comparison with the sales before or after the campaign.

• **Customer retention—analysis of customer loyalty**:Sequential pattern
mining can then be used to investigate changes in customer consumption or loyalty and suggest adjustments on the pricing
• **Product recommendation and cross-referencing of items:** Collaborative recommender systems use data mining techniques to make personalized product recommendations during live customer transactions, based on the opinions of other customers.

4) Web Page Layout Structure Mining 630

The basic structure of a Web page is its DOM4 structure. When a Web page is presented to the user, the spatial and visual cues can help the user unconsciously divide the Web page into several semantic parts. It is possible to automatically segment the Web pages by using the spatial and visual cues.an algorithm called VIsion-based Page Segmentation (VIPS). VIPS aims to extract the semantic structure of a Web page based on its visual presentation.It first extracts all of the suitable blocks from the HTML DOM tree, and then it finds the separators between these blocks. Based on these separators, the semantic tree of the Web page is constructed. A Web page can be represented as a set of blocks

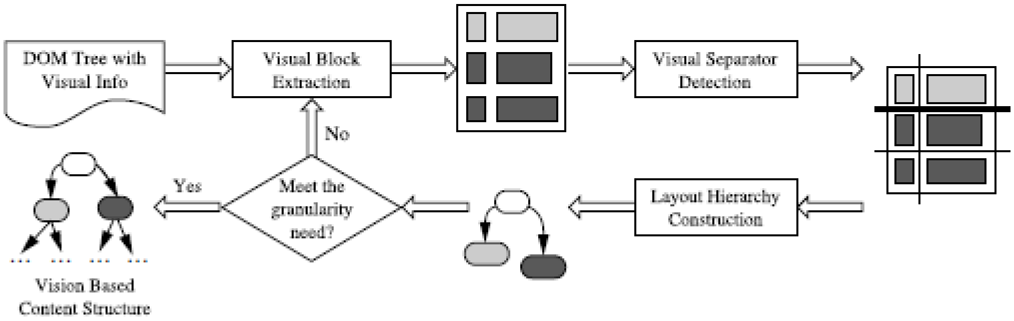
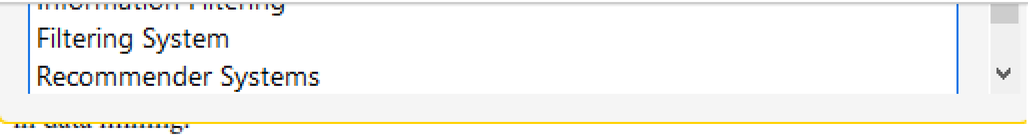

DOM Tree with Visual Info
Visual Block Extraction
Visual Separator Detection
Layout Hierarchy Construction
Meet the granularity need?
No
Yes
Vision Based Content Structure

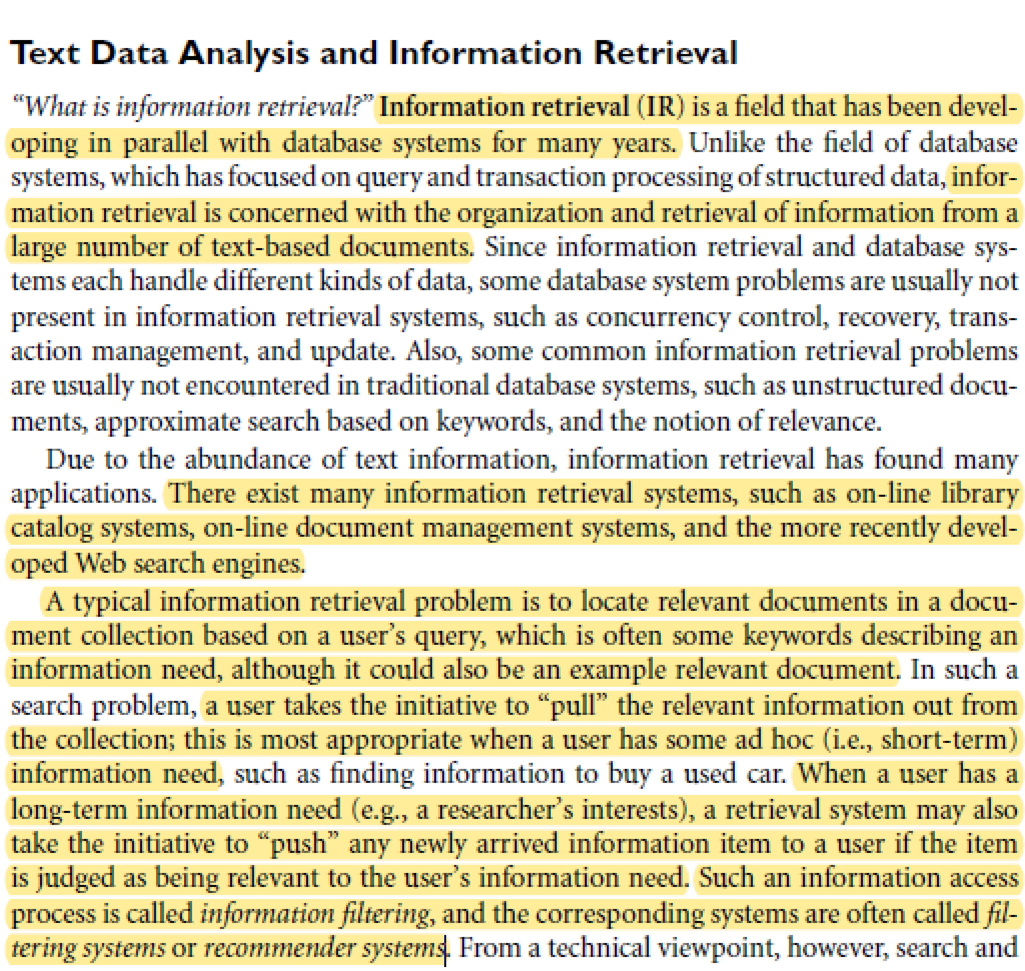
Figure 10.8 The process flow of vision-based page segmentation algorithm.

5) What is information retrieval? Discuss any one application of information retrieval.615



Information Filtering
Filtering System
Recommender Systems

in data mining.



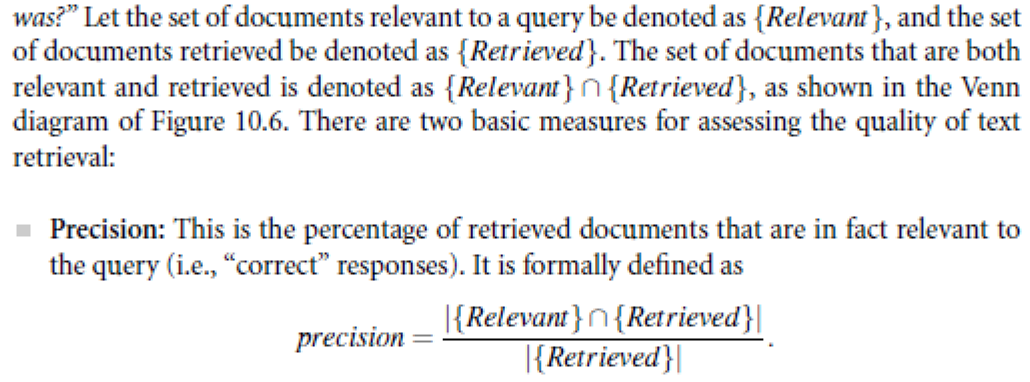
Text Data Analysis and Information Retrieval

"What is information retrieval?" **Information retrieval** (IR) is a field that has been developing in parallel with database systems for many years. Unlike the field of database systems, which has focused on query and transaction processing of structured data, information retrieval is concerned with the organization and retrieval of information from a large number of text-based documents. Since information retrieval and database systems each handle different kinds of data, some database system problems are usually not present in information retrieval systems, such as concurrency control, recovery, transaction management, and update. Also, some common information retrieval problems are usually not encountered in traditional database systems, such as unstructured documents, approximate search based on keywords, and the notion of relevance.

Due to the abundance of text information, information retrieval has found many applications. There exist many information retrieval systems, such as on-line library catalog systems, on-line document management systems, and the more recently developed Web search engines.

A typical information retrieval problem is to locate relevant documents in a document collection based on a user's query, which is often some keywords describing an information need, although it could also be an example relevant document. In such a search problem, a user takes the initiative to "pull" the relevant information out from the collection; this is most appropriate when a user has some ad hoc (i.e., short-term) information need, such as finding information to buy a used car. When a user has a long-term information need (e.g., a researcher's interests), a retrieval system may also take the initiative to "push" any newly arrived information item to a user if the item is judged as being relevant to the user's information need. Such an information access process is called *information filtering*, and the corresponding systems are often called *filtering systems* or *recommender systems*. From a technical viewpoint, however, search and

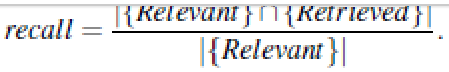
6) Explain the various measures used to show the relationship between the set of relevant documents and the set of retrieved documents.616



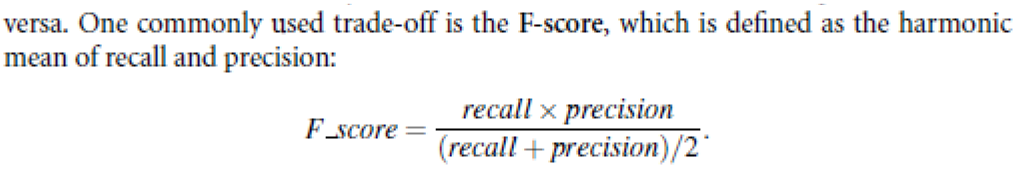
was?" Let the set of documents relevant to a query be denoted as $\{Relevant\}$, and the set of documents retrieved be denoted as $\{Retrieved\}$. The set of documents that are both relevant and retrieved is denoted as $\{Relevant\} \cap \{Retrieved\}$, as shown in the Venn diagram of Figure 10.6. There are two basic measures for assessing the quality of text retrieval:

- **Precision:** This is the percentage of retrieved documents that are in fact relevant to the query (i.e., "correct" responses). It is formally defined as

$$precision = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Retrieved\}|}.$$



$$recall = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Relevant\}|}.$$

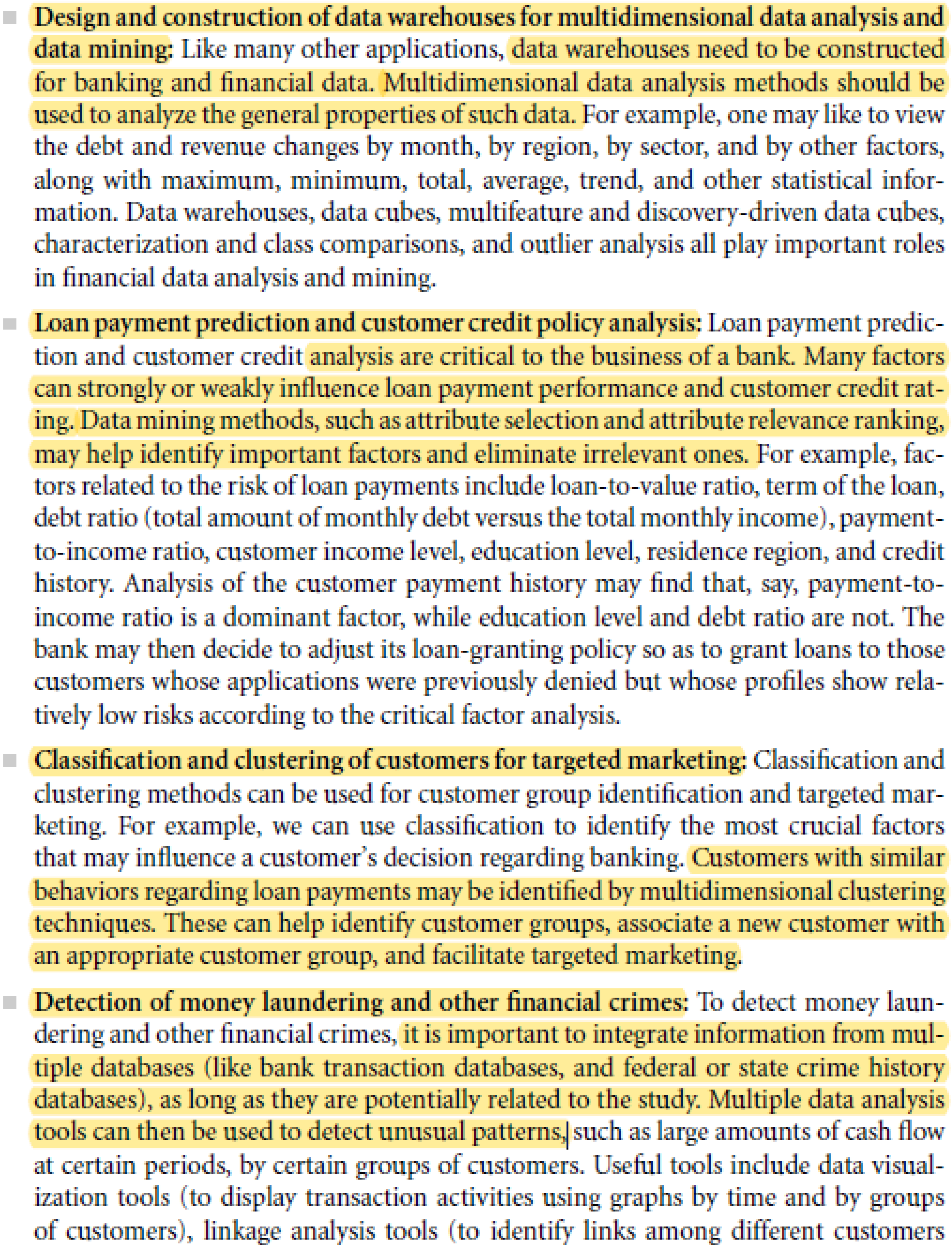


versa. One commonly used trade-off is the **F-score**, which is defined as the harmonic mean of recall and precision:

$$F\_score = \frac{recall \times precision}{(recall + precision)/2}.$$

10)

11) Discuss in detail data mining for the financial sector. 649



- **Design and construction of data warehouses for multidimensional data analysis and data mining:** Like many other applications, data warehouses need to be constructed for banking and financial data. Multidimensional data analysis methods should be used to analyze the general properties of such data. For example, one may like to view the debt and revenue changes by month, by region, by sector, and by other factors, along with maximum, minimum, total, average, trend, and other statistical information. Data warehouses, data cubes, multifeature and discovery-driven data cubes, characterization and class comparisons, and outlier analysis all play important roles in financial data analysis and mining.
- **Loan payment prediction and customer credit policy analysis:** Loan payment prediction and customer credit analysis are critical to the business of a bank. Many factors can strongly or weakly influence loan payment performance and customer credit rating. Data mining methods, such as attribute selection and attribute relevance ranking, may help identify important factors and eliminate irrelevant ones. For example, factors related to the risk of loan payments include loan-to-value ratio, term of the loan, debt ratio (total amount of monthly debt versus the total monthly income), payment-to-income ratio, customer income level, education level, residence region, and credit history. Analysis of the customer payment history may find that, say, payment-to-income ratio is a dominant factor, while education level and debt ratio are not. The bank may then decide to adjust its loan-granting policy so as to grant loans to those customers whose applications were previously denied but whose profiles show relatively low risks according to the critical factor analysis.
- **Classification and clustering of customers for targeted marketing:** Classification and clustering methods can be used for customer group identification and targeted marketing. For example, we can use classification to identify the most crucial factors that may influence a customer's decision regarding banking. Customers with similar behaviors regarding loan payments may be identified by multidimensional clustering techniques. These can help identify customer groups, associate a new customer with an appropriate customer group, and facilitate targeted marketing.
- **Detection of money laundering and other financial crimes:** To detect money laundering and other financial crimes, it is important to integrate information from multiple databases (like bank transaction databases, and federal or state crime history databases), as long as they are potentially related to the study. Multiple data analysis tools can then be used to detect unusual patterns, such as large amounts of cash flow at certain periods, by certain groups of customers. Useful tools include data visualization tools (to display transaction activities using graphs by time and by groups of customers), linkage analysis tools (to identify links among different customers

visual presentation.Such semantic structure is a tree structure: Each node will be assigned a value (Degree of Coherence) to indicate how coherent is the content in the block based on visual perception. The VIPS algorithm makes full use of the page layout feature. It first extracts all of the suitable blocks from the HTML DOM tree, and then it finds the separators between these blocks.Based on these separators, the semantic tree of the Web page is constructed. A Web page can be represented as a set of blocks). Compared with DOM-based methods, the segments obtained by VIPS are more semantically aggregated. Contents with different topics are distinguished as separate blocks.

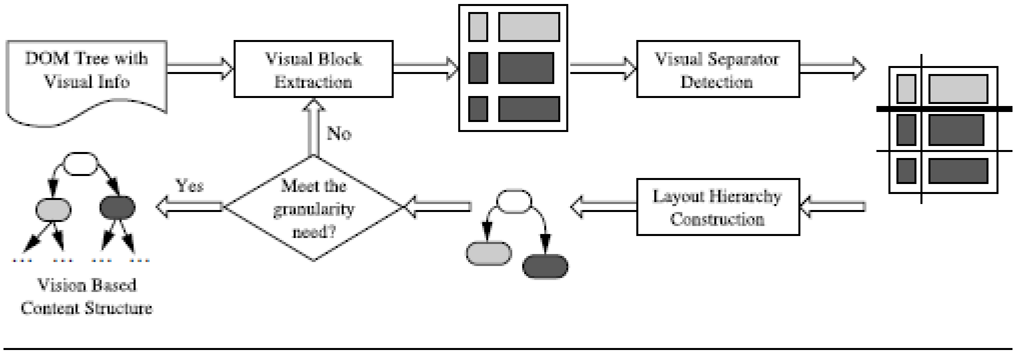
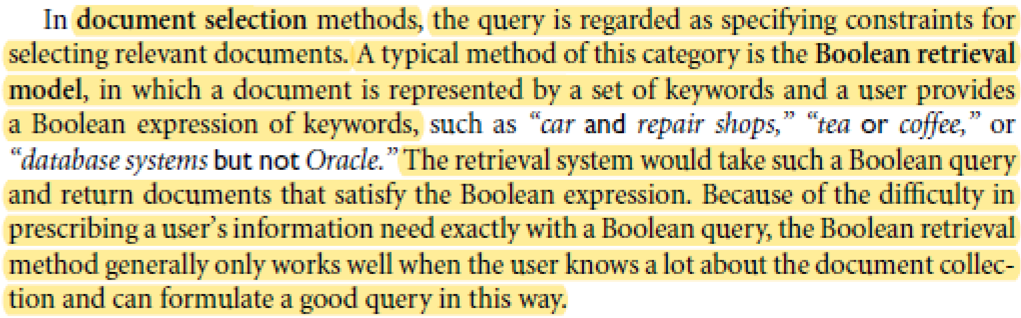

DOM Tree with Visual Info
Visual Block Extraction
Visual Separator Detection
No
Yes
Meet the granularity need?
Layout Hierarchy Construction
Vision Based Content Structure

Figure 10.8 The process flow of vision-based page segmentation algorithm.

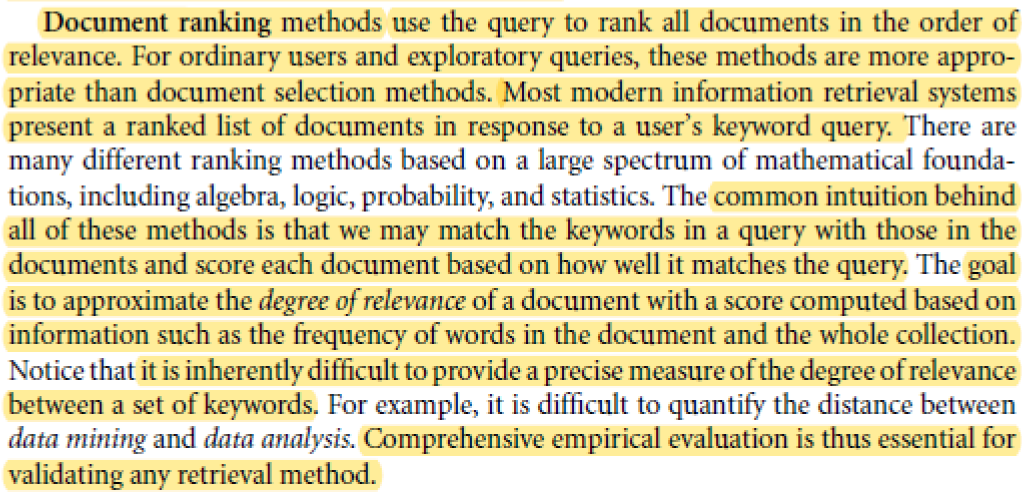
14) part-2) Discuss the Text retrieval methods in detail.

Document Selection



In **document selection** methods, the query is regarded as specifying constraints for selecting relevant documents. A typical method of this category is the **Boolean retrieval model**, in which a document is represented by a set of keywords and a user provides a Boolean expression of keywords, such as "*car* and *repair shops*," "*tea* or *coffee*," or "*database systems* but not *Oracle*." The retrieval system would take such a Boolean query and return documents that satisfy the Boolean expression. Because of the difficulty in prescribing a user's information need exactly with a Boolean query, the Boolean retrieval method generally only works well when the user knows a lot about the document collection and can formulate a good query in this way.

Document Ranking

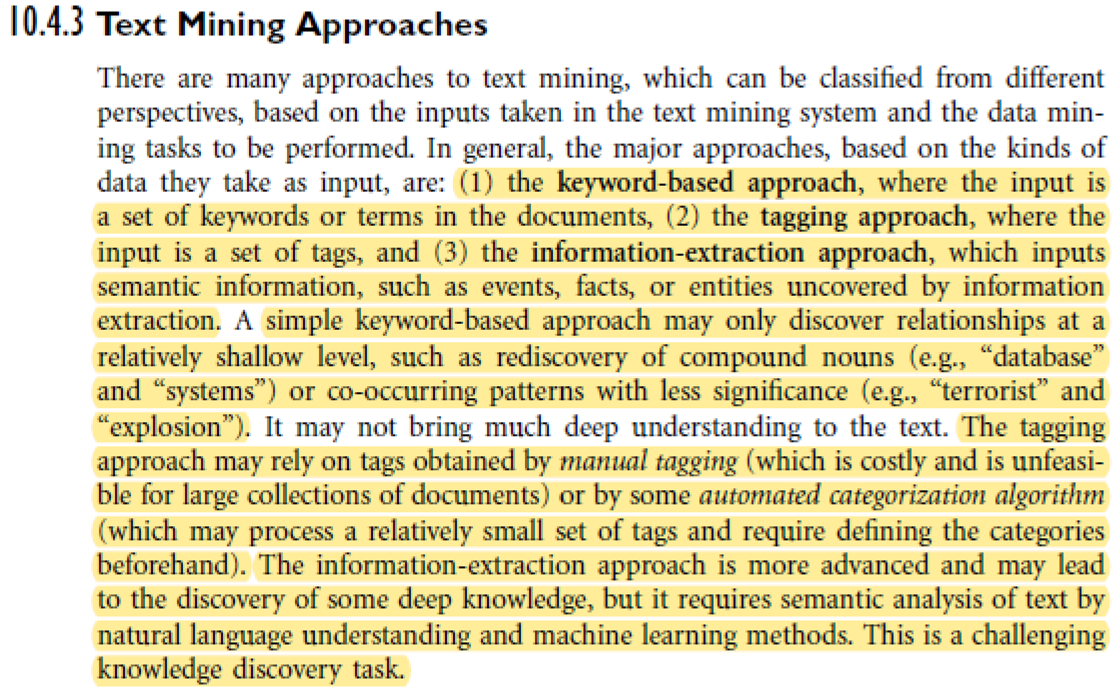


Document ranking methods use the query to rank all documents in the order of relevance. For ordinary users and exploratory queries, these methods are more appropriate than document selection methods. Most modern information retrieval systems present a ranked list of documents in response to a user's keyword query. There are many different ranking methods based on a large spectrum of mathematical foundations, including algebra, logic, probability, and statistics. The common intuition behind all of these methods is that we may match the keywords in a query with those in the documents and score each document based on how well it matches the query. The goal is to approximate the *degree of relevance* of a document with a score computed based on information such as the frequency of words in the document and the whole collection. Notice that it is inherently difficult to provide a precise measure of the degree of relevance between a set of keywords. For example, it is difficult to quantify the distance between *data mining* and *data analysis*. Comprehensive empirical evaluation is thus essential for validating any retrieval method.

15) Write a short note on Spatial Data mining. 600

Spatial data mining refers to the extraction of knowledge, spatial relationships, or other interesting patterns not explicitly stored in spatial databases. Such mining demands an integration of data mining with spatial database technologies. It can be used for under- standing spatial data, discovering spatial relationships and relationships between spatial and control, environmental studies, and many other areas where spatial data are used. A crucial challenge to spatial data mining is the exploration of efficient spatial data mining techniques due to the huge amount of spatial data and the complexity of spatial data types and spatial access methods.

16) What are the major text mining approaches based on the kinds of the data taken as input? Explain any 2. 624



10.4.3 Text Mining Approaches

There are many approaches to text mining, which can be classified from different perspectives, based on the inputs taken in the text mining system and the data mining tasks to be performed. In general, the major approaches, based on the kinds of data they take as input, are: (1) the **keyword-based approach**, where the input is a set of keywords or terms in the documents, (2) the **tagging approach**, where the input is a set of tags, and (3) the **information-extraction approach**, which inputs semantic information, such as events, facts, or entities uncovered by information extraction. A simple keyword-based approach may only discover relationships at a relatively shallow level, such as rediscovery of compound nouns (e.g., "database" and "systems") or co-occurring patterns with less significance (e.g., "terrorist" and "explosion"). It may not bring much deep understanding to the text. The tagging approach may rely on tags obtained by *manual tagging* (which is costly and is unfeasible for large collections of documents) or by some *automated categorization algorithm* (which may process a relatively small set of tags and require defining the categories beforehand). The information-extraction approach is more advanced and may lead to the discovery of some deep knowledge, but it requires semantic analysis of text by natural language understanding and machine learning methods. This is a challenging knowledge discovery task.