(Traditional) Text Mining

Tafseer Ahmed

Presentation Plan

- Text Mining Introduction
- Steps involved in Text Mining
- Applications involved in Text Mining
  - Categorization
  - Clustering

Text Mining - Introduction

Text Mining

- Text Mining deals with unstructured textual information and it discovers previously unknown structure and implicit meanings buried within the large amount of text.
- A huge amount of information is present as unstructured text, so we need a special process to analyze it.
- Text Mining draws on data mining, machine learning, information retrieval and computational linguistics.

Applications of Text Mining (1)

- **Categorization**

  Assigning a new document to one of the defined categories of documents
- **Clustering**

  Finding clusters/categories in a given set of documents
- **Term Extraction**

  Extracting important terms and keywords used in the document
- **Summarization**

Applications of Text Mining (2)

- **Information Retrieval**

  Finding related documents corresponding to a query

- **Information/Feature Extraction**

  Extraction of (processable) information from a given document

- **Thematic Indexing**

  Knowledge about meaning of words to identify broad topics covered in the document

- …..

Applications of Text Mining (3)

- **News for Information Extraction:**

  19 March - A bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported. According to unofficial sources, the bomb - allegedly detonated by urban guerrilla commandos - blew up a power tower in the northwestern part of San Salvador at 0650 (1250 GMT).

- **Information Extracted**

INCIDENT TYPE	bombing
DATE	March 19
LOCATION	San Salvador (city)
PERPETRATOR	urban guerrilla commandos
PHYSICAL TARGET	power tower
HUMAN TARGET	-
EFFECT ON PHYSICAL TARGET	destroyed
EFFECT ON HUMAN TARGET	no injury or death
INSTRUMENT	bomb

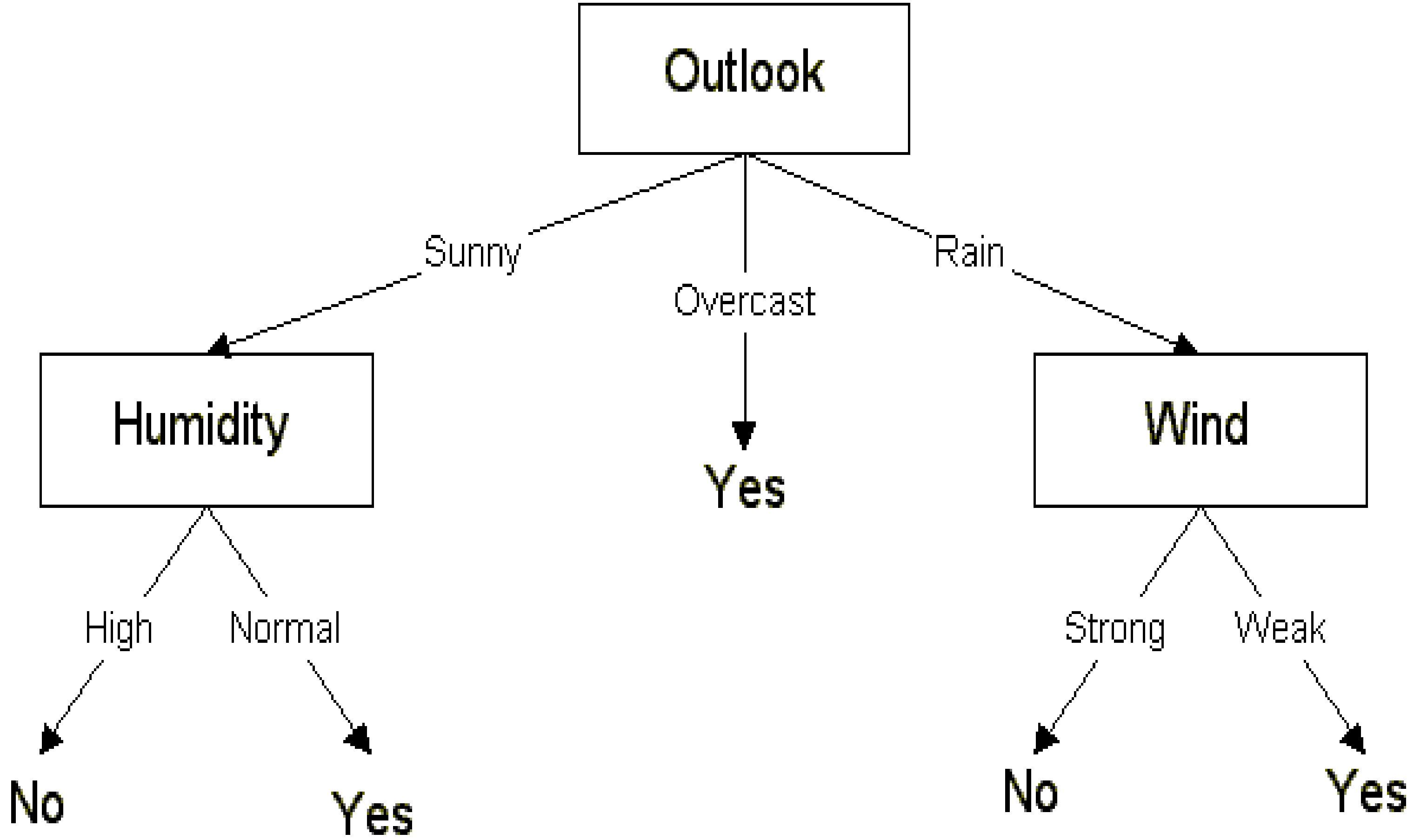
Data Mining

- It seems fairly non-controversial that text mining is a sub-discipline of the broader and slightly older field of data mining, the sub-discipline which deals with textual data.
- Data Mining is the discovery of interesting, unexpected or valuable structures in large data sets.

An Example of Data Mining (1)

Day	Outlook	Temperature	Humidity	Wind	Play outside
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Strong	Mild	Sunny	Normal	Yes
D14	Rain	Mild	High	Strong	No

An Example of Data Mining (2)



Why Text Mining is difficult?

- Text consists of Unstructured Data.

News	Category
The Pakistan Hockey Federation (PHF) has revised the format of the 56th National Hockey Championship while also rescheduling the event dates.	Sports
Karachi will host back to back one-day internationals following the PCB's revision of the Sri Lankan tour schedule.	Sports
Citing surging raw material costs and the falling Rupee, the govt. has allowed increases in the prices of several drugs.	Business
Overseas Pakistani workers sent record remittances in December helping the country to minimise its trade and current account deficits.	Business

Steps involved in Text Mining

Feature based Model (1)

- A subset of document features is selected as the representational model of the document.
- The features can be:
  - Characters n-grams e.g. cha, har, rac, act, ….
  - Words e.g. word, feature, extraction, ….
  - Terms e.g. ‘feature extraction’, ‘text mining’
- Bag of Words Approach:

  a document is simply an unstructured set of words (or terms) appearing in it

Feature based Model (2)

- **Sample Text**

  Text Mining deals with unstructured textual information and it discovers previously unknown structure and implicit meanings buried within the large amount of text. A huge amount of information is present as unstructured text, so we need a special process to analyze it.

- **Feature vector**

  text 3, unstructured 2, information 2, amount 2,

  and 2, it 2, of 2, a 2,

  mining 1, deal 1, with 1, textual 1,discovers 1, previously 1, unknown 1, structure 1, implicit 1, meanings 1, buried 1 within 1, the 1, large 1, huge 1, is 1, present 1, as 1, so 1, we 1, need 1, special 1, process 1, to 1, analyze 1

Feature based Model (3)

- All words in the text are not required to represent the document as feature.
- Stemming and Stop Word Removal are employed to get relevant features only.

Stop Word Removal (1)

- Stop words are used to eliminate words that bear no content or relevant semantics.
- Generally, a stop word list includes articles, pronouns, adjectives, adverbs and prepositions.
- Examples:

about above across after afterwards again against all almost alone along already also although always am among amongst amount an and another any anyhow anyone anything anyway anywhere are around as at

Stop Word Removal (2)

- **Sample Text**

  Text Mining deals with unstructured textual information and it discovers previously unknown structure and implicit meanings buried within the large amount of text. A huge amount of information is present as unstructured text, so we need a special process to analyze it.

- **Feature vector**

  text 3, unstructured 2, information 2, amount 2,

  and 2, it 2, of 2, a 2,

  mining 1, deal 1, with 1, textual 1,discovers 1, previously 1, unknown 1, structure 1, implicit 1, meanings 1, buried 1 within 1, the 1, large 1, huge 1, is 1, present 1, as 1, so 1, we 1, need 1, special 1, process 1, to 1, analyze 1

Stemming

- Striping the word to its basic form. The stem can be different from word’s linguistic root.
- Advantage: ‘read’, ‘reading’ and ‘reads’ all have same stem ‘read’.
- Disadvantage: Two different words can have same stem (Example: ‘international’ and ‘internal’ may have same stem ‘intern’.)

Stemming Algorithm for English

- Output of Potter’s and Lovin’s Algorithm

Word	**Lovin’s Stem**	**Potter’s Stem**
happier	hap	happier
effectiveness	effect	effect
happy	hap	happi
Genetic	genet	genet
Easy	ea	Easi
Invisible	inv	Invis
printed	print	print

Stemming Example (1)

- **Sample Text**

  Text Mining deals with unstructured textual information and it discovers previously unknown structure and implicit meanings buried within the large amount of text. A huge amount of information is present as unstructured text, so we need a special process to analyze it.

- **Feature vector**

  text 3, unstructured 2, information 2, amount 2,

  mining 1, deal 1, textual 1,discovers 1, previously 1, unknown 1, structure 1, implicit 1, meanings 1, buried 1, large 1, huge 1, present 1, need 1, special 1, process 1, analyze 1

Stemming Example (2)

- **Sample Text**

  Text Mining deals with unstructured textual information and it discovers previously unknown structure and implicit meanings buried within the large amount of text. A huge amount of information is present as unstructured text, so we need a special process to analyze it.

- **Feature vector**

  text **4**, structur **3**, information 2, amount 2,

  min 1, deal 1, discover 1, previous 1, known 1, implicit 1, mean 1, bur 1, larg 1, hug 1, present 1, need 1, special 1, process 1, analyz 1

TF*IDF

- Term Frequency * Inverse Document Frequency
- TF_{ik} = Frequency of the Term T_i in Document D_k

  More frequent terms in a document are more important (for discrimination)
- $IDF_i = \log (N/n_i)$

  N = total number of documents

  $n_i =$ the number of documents that contain T_i

  A term common in more documents is less important (for discrimination)

Applications of Text Mining

Categorization

- Assigning a new document to one of the defined categories of documents
- Supervised Learning

Naive Bayes Method (1)

- C1, C2, … Cn are categories. A document D belongs to Category Ci that gives maximum P(Ci| D).
- Bayes Theorem:

$$P(Ci|D) = P(D|Ci)\, P(Ci)/P(D)$$

- P(D) is constant for all categories.
- $P(Ci|D) = P(D|Ci)\, P(Ci)$
- P(Ci) = (no. of documents in Ci) / (total no. of documents)

Naive Bayes Method (2)

- T1, T2 ... Tm is the sequence of terms in D.
- Assumption: a term T in the jth place of D is conditionally independent of all the other terms in D and of the position j.
- P(D|Ci) = P(T1|Ci)*P(T2|Ci) * ... * P(Tm|Ci)
- P(Tj |Ci) = (number of occurrences of Tj in Ci) / (total number of words in Ci)
- P(Ci|D) = P(Ci)*P(T1|Ci)*P(T2|Ci)* ...* P(Tm|Ci)

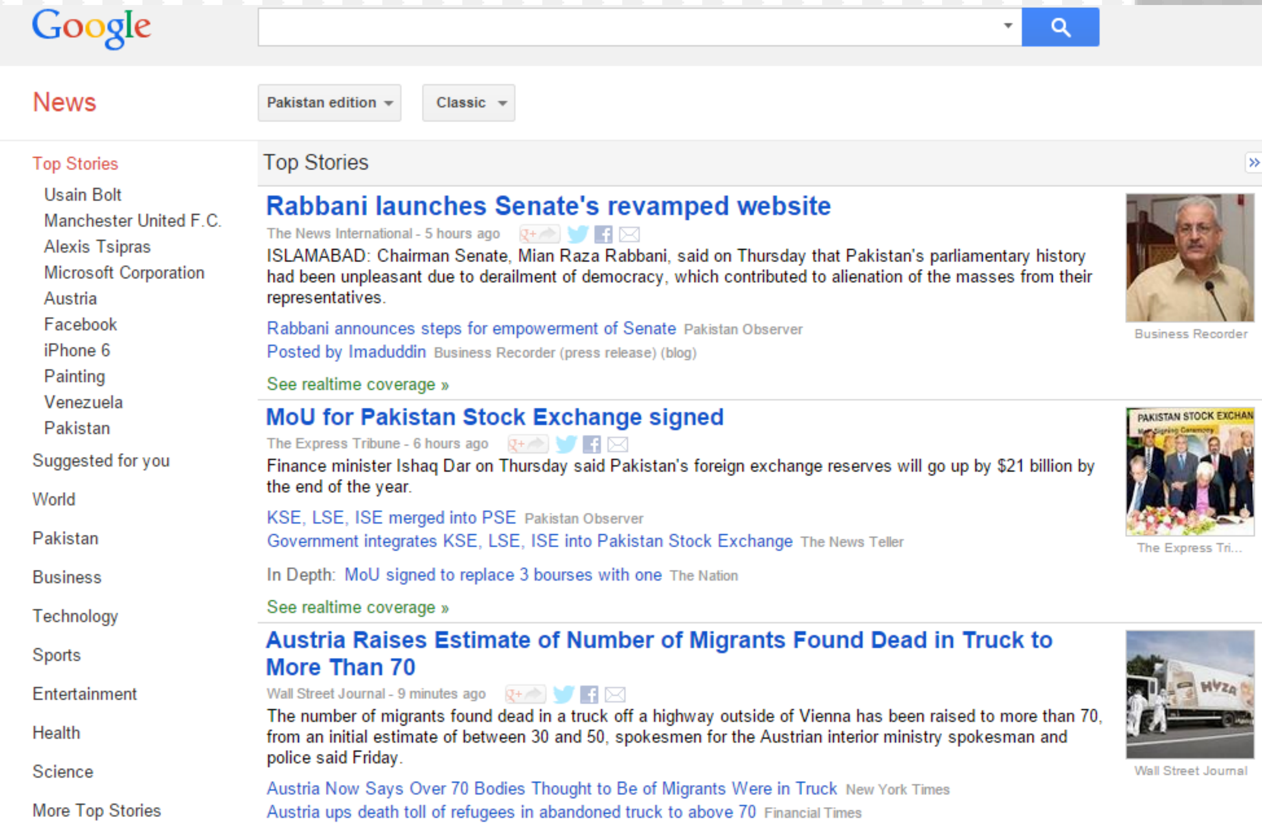
Rocchio Method

- It uses a subset of features in its feature vector.
- Learner Algorithm:
  - Find normalized feature vector Vj of each document Dj.
  - For each category Ci, compute the centroid of all the documents in Ci.
- Categorization Algorithm:
  - For a new document D, find the closest centroid (or a similar measure) and put D into the corresponding category.

Clustering

- Finding clusters in a given set of documents.
- Unsupervised Learning
- Example:

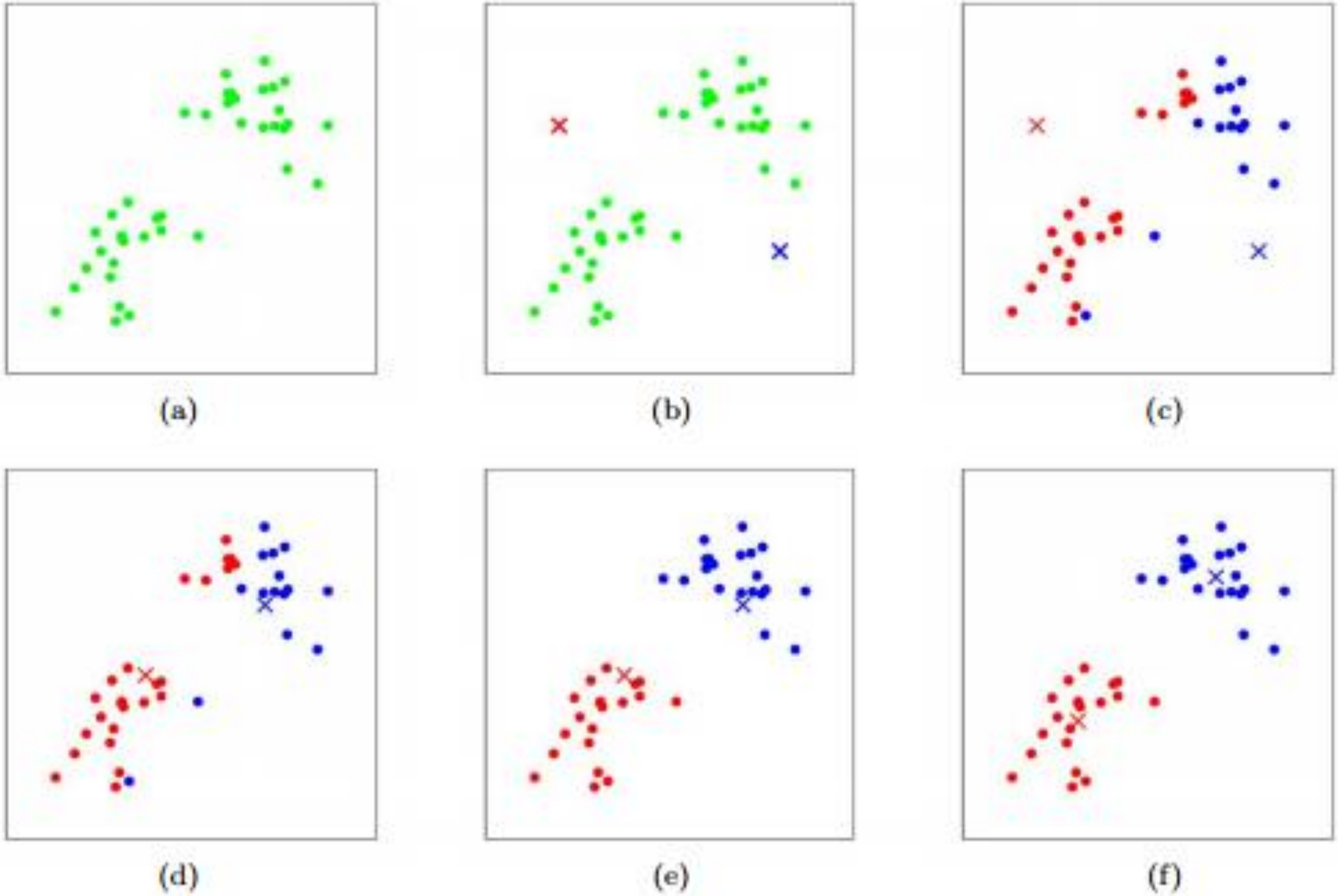
  Google News



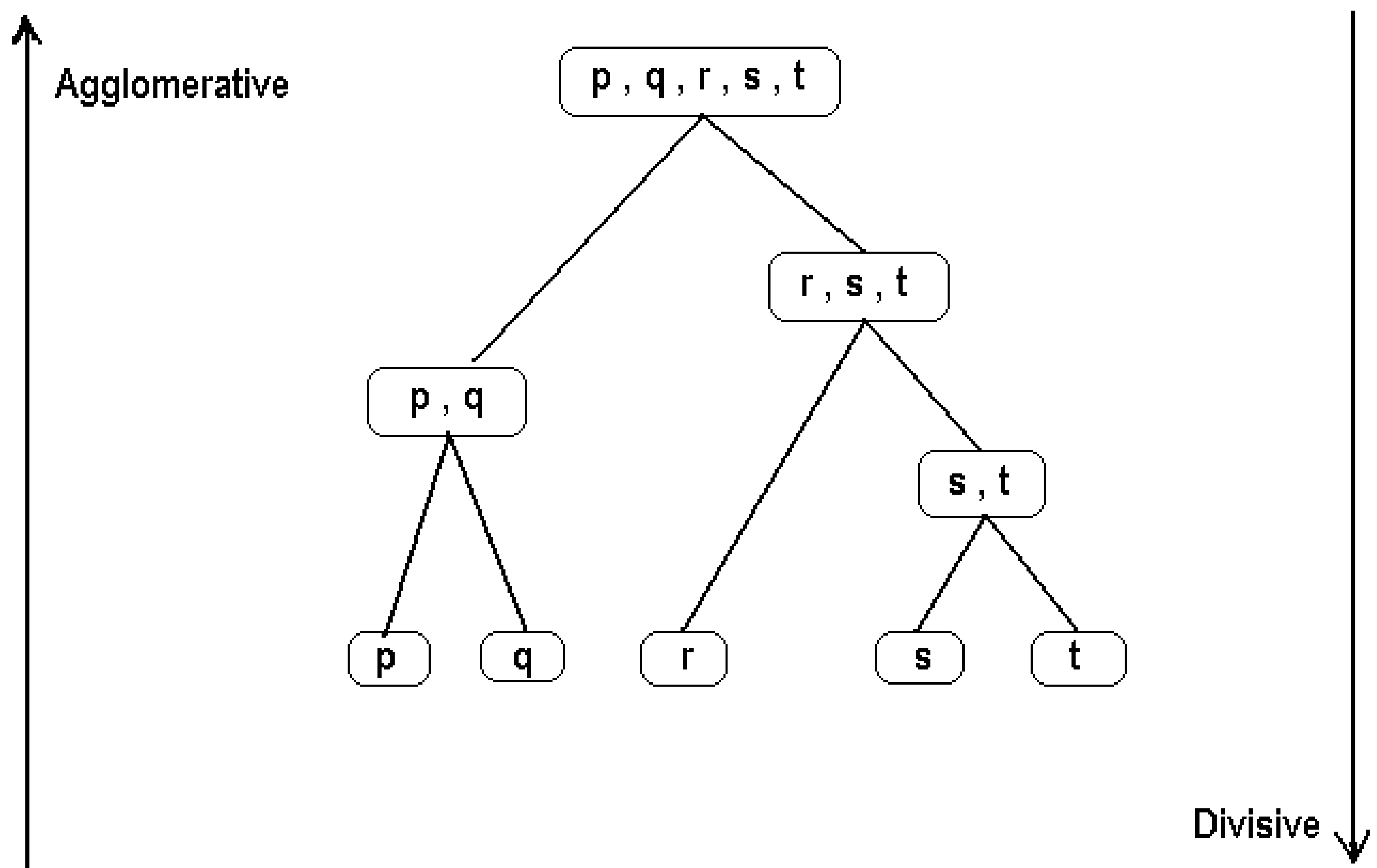
k-means Algorithm (1)

1. Partition documents into *k* nonempty clusters.
2. Compute seed points as the centroids of the clusters of the current partition. The centroid is the center (mean point) of the cluster.
3. Assign each document to the cluster with the nearest seed point.
4. If there is a change in the clusters, go to Step 2.

k-means Algorithm (2)



Hierarchal Clustering



Questions