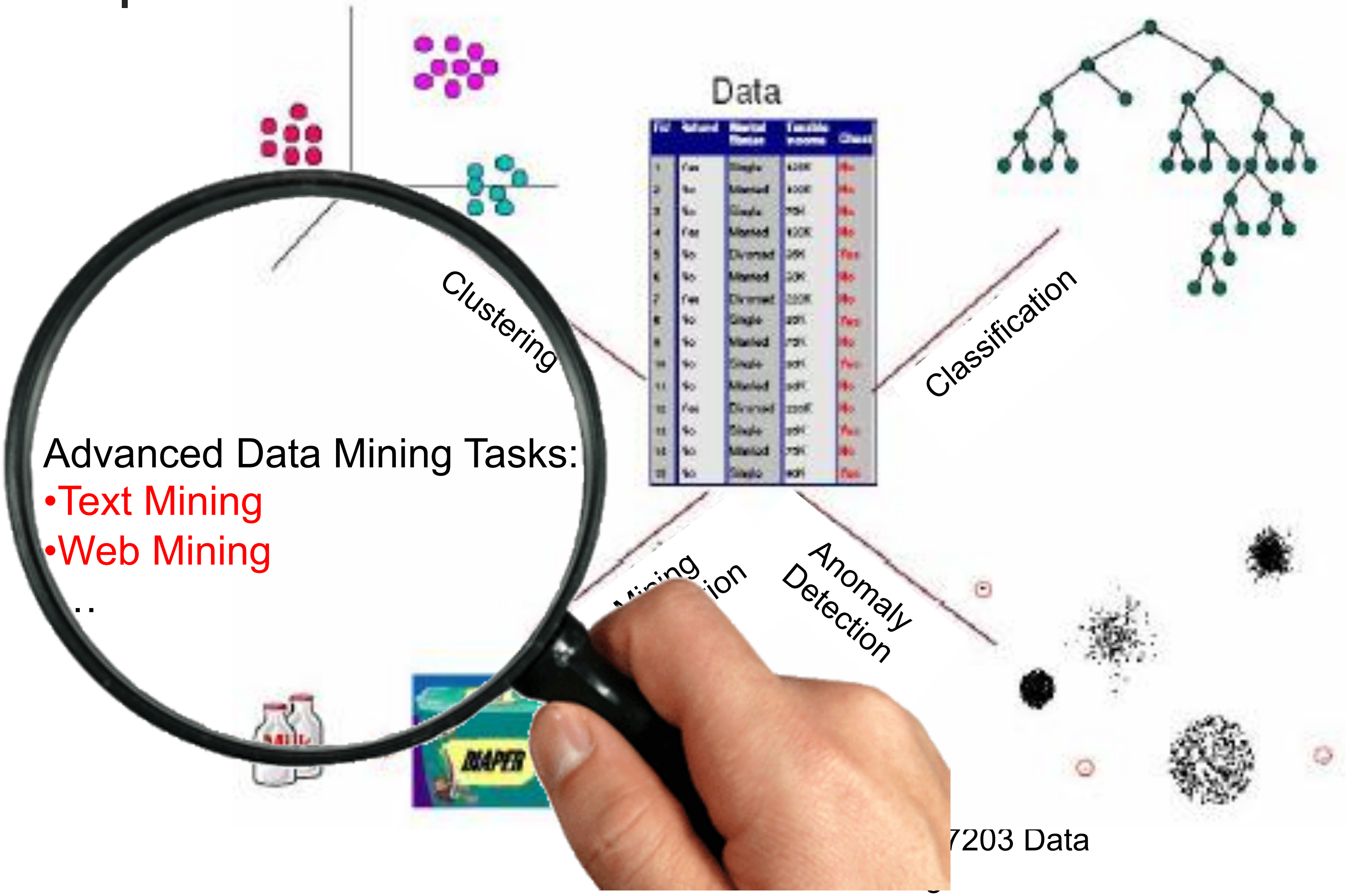
Data Mining Tasks



Text Mining

Introduction

- Text mining refers to data mining using text **documents** as data

- Text mining uses **Information Retrieval (IR)** methods to pre-process text documents

- IR methods are quite different from data pre-processing methods used for relational tables

- Web search also has its root in IR

Information Retrieval

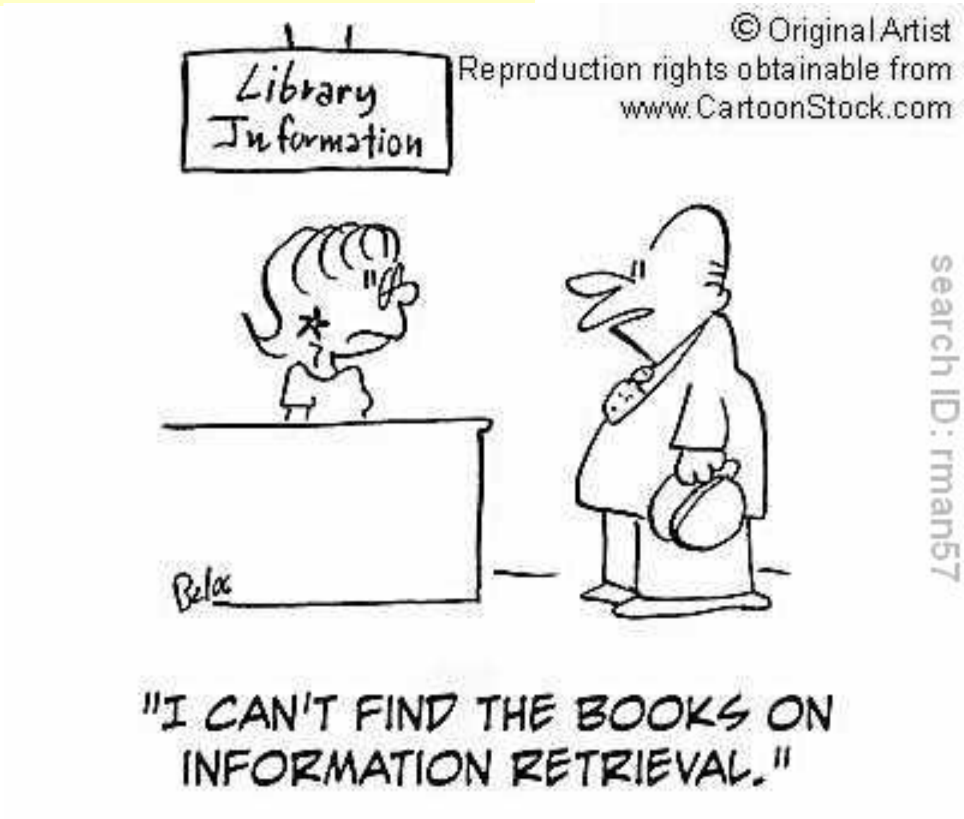
Information retrieval

From Wikipedia, the free encyclopedia

Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text indexing.

Automated information retrieval systems are used to reduce what has been called "information overload". Many universities and public libraries use IR systems to provide access to books, journals and other documents. Web search engines are the most visible IR applications.


SEARCH
INFORMATION
RETRIEVAL
GOOGLE
CORPUS
QUERY
RANK
USEFUL
DATA
DOCUMENTS
DIFFICULT
DOCUMENT
DATABASE
CRAWL
INDEX
WEB
LARGE
IMPORTANT
HOW
COMPLEX
COOL
AMAZING
CONFUSING
MATHEMATICAL
EFFICIENT
TEXT
LIST
FAST
FUN
SEARCHING
CRAWLER
DISTRIBUTED
GATHER
METADATA
RETRIEVE
SLOW
SCIENCE
PAGE
THE
FOR
FREQUENCY
PATTERSON
MULTI
INVERTED
PROGRESSED
WEIGHTING
ONE
QUERIES
WORD
BUSINESS
MONEY
QUEUING
ENGINES
MICROSOFT
PURPLE
REUTERS
KNOWING
SORT
EVALUATE
STEMMING
AGGREGATE
COLORFUL
STUDIED
DIGGING
INTERESTING
STATISTICS
INTRACTABLE
GROWING
DISCOVERING
COSINE
ENTHRALLING
MAGICAL
TECHNOLOGICAL
SPECIAL
RELATIONAL
BIG
UNSTRUCTURED
CHALLENGE
STOP
MODIFY
TECHNICAL
CONSUMING
INFO
ORGANIZATION
PARSING
DOCS
OF
NEED
AND
GET
MATH
HADOOP
SCORE
SPEEDY
BRIGHT
NEEDED
TO
IN
MAP
WAY
SCORES
EXPENSIVE
WWW
HARD
TIME
TRUST
INDEXING
STUFF
ABOUT
NOT
ALGORITHMIC
SAD
OVERLOAD
THINGS
QUIRKY
SORTING
EXCITING
CRAWLERS
SEXY
FUTURE
COMPLICATED
SCHOLARLY
EVOLVING
FRUSTRATING
SEARCHES
WIKIPEDIA
RETRIEVING
MOVE
CRAWLING
KEY
DIFFERENT
VARIED
A
ALL
VAST
TRICKY
EFFICIENTLY
DISTRIBUTEDLY
THROUGH
LINEAR
POSTING
HAPPY
RELEVANT
POLITE


© Original Artist
Reproduction rights obtainable from
www.CartoonStock.com
Library
Information
search ID: rman57
"I CAN'T FIND THE BOOKS ON
INFORMATION RETRIEVAL."

Natural Language Processing

An example of part-of-speech tagging:

This	sentence	serves	as	an	example.
Det	Noun	Verb	P	Det	Noun

An example of Named Entity Recognition:

The University of Queensland,	St. Lucia	Brisbane
University	Suburb	City

Text Mining Tasks

- Text Classification
  - Assigning a document to one of several classes
- Text Clustering
  - Unsupervised learning
- Text Summarization
  - Extracting a summary from a document
- … …

Data Mining vs. Text Mining

- In traditional data mining:
  - Data is structured:
    - data stored in a database
    - Very clear structure: tables, records, attributes
  - Data is numeric:
    - Easy to measure similarity
    - Need to find a suitable way to transform data (text, images, videos, etc) into numbers

Text Mining Challenge

- In text mining, data is unstructured!
  - Given two documents
    - how to compute their similarity?
    - Based on what dimensions?
- Idea:
  - Unstructured => Structured

How to achieve a structured representation of an unstructured document?

Document Representation

- Document
- Word (term)
- “This is a data mining course. Data mining is important.”

term	This	is	a	data	mining	course	important
	1	1	1	1	1	1	
		1		1	1		1
frequency	**1**	**2**	**1**	**2**	**2**	**1**	**1**

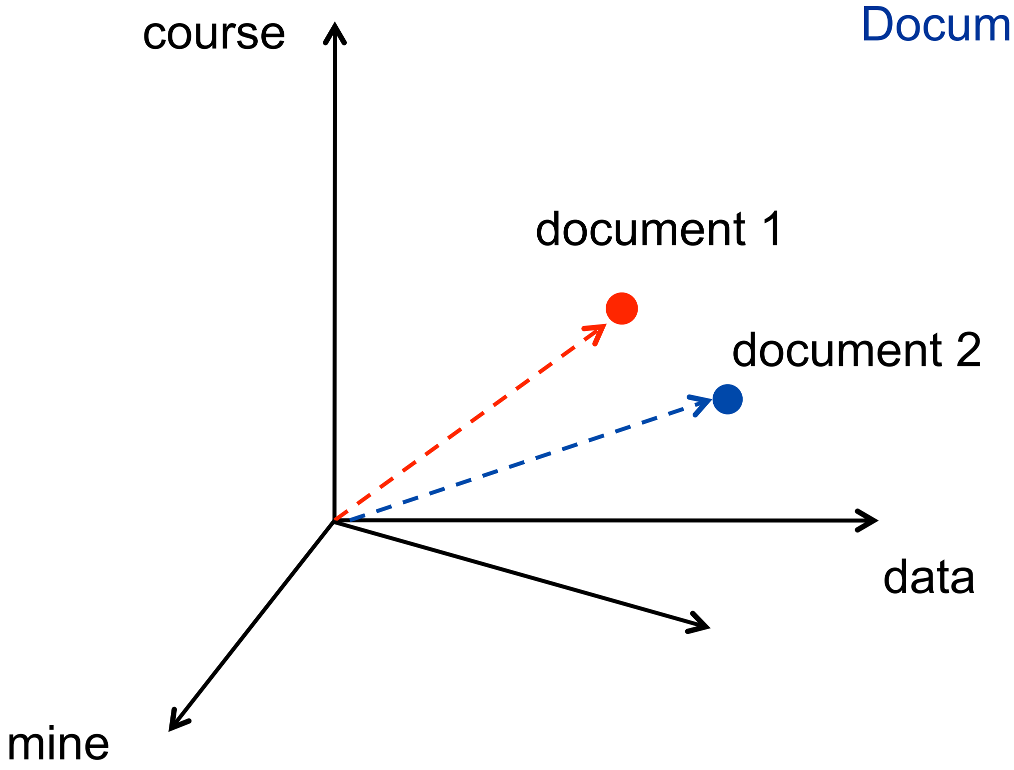
Vector Space Model (VSM)

- Each word/term is a dimension
  - *M* different words ➔ *M*-dimensional vector space
- Each document is treated as a **"bag"** of words or terms
- Given a collection of documents *D*, let $V = \{t_1, t_2, ..., t_V\}$ be the set of distinctive words/terms in the collection
  - *V* is called the vocabulary
- A **weight** $w_{ij} > 0$ is associated with each term t_i of a document $\mathbf{d}_j \in D$
  - For a term that does not appear in document $\mathbf{d}_j$, $w_{ij} = 0$

$$\mathbf{d}_j = (w_{1j}, w_{2j}, ..., w_{Vj}),$$

An Example of VSM

Document 1: (0.938, 0.346, 0, 0, 0, 0, 0)
Document 2: (0, 0.225 0, 0, 0.611, 0.611, 0.450)



Each document is regarded as a **point** in the m-dimensional vector space

conceptually, w_{ij} denotes the **importance** of the word i in d_j

Vector Space Model

- Problems:
  1. There are soooooooooooooo many words in the English language!
  2. How to determine the “importance” of each words?

Vector Space Model

- The first problem: too many words
- We solve this problem by:
  1. Removing stop words
     - A, the, this, that …
  2. Stemming
     - study
     - study, studying, studied

Vector Space Model

- The second problem: how to determine the importance of each term
- We solve this problem by:
  - Using a weighting scheme; the **TF-IDF** scheme:

$$w(word_i) = TF(word_i) \times IDF(word_i)$$

$$TF(word_i) = \text{number of times } word_i \text{ appears in the document}$$

$$IDF(word_i) = \log \frac{\text{total documents}}{\text{document frequency}}$$

TF-IDF

- TF-IDF
  - Term frequency-inverse document frequency
  - Given a collection of documents, it estimates how important is a term is to a document
  - the number of times a term occurs in a document is called its term frequency
  - the number of documents a term occurs in is called its document frequency

Why TF-IDF

- Why the IDF component?
  - Can we simply use term frequency?
  - IDF allows to:
    - **increase** the weight of terms that occur rarely in the collection
    - **decrease** the weight of terms that occur very frequently in the collection
      - Example: the, a, … (if stop words are not removed)
      - Example: UQ

TF-IDF Calculation

Term Importance:

$$w(word_i) = TF(word_i) \times IDF(word_i)$$

Term Frequency:

$$TF(word_i) = \text{number of times } word_i \text{ appears in the document}$$

Inverse Document Frequency:

$$IDF(word_i) = \log \frac{\text{total documents}}{\text{document frequency}}$$

Running Example

This is a data mining course.

We are studying text mining. Text mining is a subfield of data mining.

Mining text is interesting, and I am interested in it.

Step 1 – Extract text

This is a data mining course. → This is a data mining course

We are studying text mining. Text mining is a subfield of data mining. → We are studying text mining Text mining is a subfield of data mining

Mining text is interesting, and I am interested in it. → Mining text is interesting and I am interested in it

Step 2 – Remove stop words

This is a data mining course. → ~~This is a~~ data mining course

We are studying text mining. Text mining is a subfield of data mining. → ~~We are~~ studying text mining Text mining ~~is a~~ subfield ~~of~~ data mining

Mining text is interesting, and I am interested in it. → Mining text ~~is~~ interesting ~~and I am~~ interested ~~in it~~

Step 3 – Convert all words to lowercase

This is a data mining course.

~~This is a~~ data mining course

We are studying text mining. Text mining is a subfield of data mining.

text
~~We are~~ studying text mining ~~Text~~ mining
~~is a~~-subfield ~~of~~ data mining

Mining text is interesting, and I am interested in it.

mining
~~Mining~~ text ~~is~~ interesting ~~and I am~~
interested ~~in it~~

Step 4 – Stemming

This is a data mining course.

→

mine
~~This is a~~ data ~~mining~~ course

We are studying text mining. Text mining is a subfield of data mining.

→

study mine text mine
~~We are studying~~ text ~~mining~~ ~~Text~~ ~~mining~~
~~is a~~-subfield ~~of~~ data ~~mining~~
mine

Mining text is interesting, and I am interested in it.

→

mine interest
~~Mining~~ text ~~is interesting and I am~~
~~interested in it~~
interest

Step 5 – Count term frequencies

This is a data mining course.

mine

~~This is a~~ data ~~mining~~ course

course:1, data:1, mine:1

We are studying text mining. Text mining is a subfield of data mining.

study mine text mine

~~We are studying~~ text ~~mining Text mining~~

~~is a~~-subfield ~~of~~ data ~~mining~~

mine

data:1, mine:3, study:1, subfield:1, text:2

Mining text is interesting, and I am interested in it.

mine interest

~~Mining~~ text ~~is interesting and I am~~

~~interested in it~~

interest

Interest:2, mine:1, text:1

Step 6 – Create an indexing file

在几个docment出现过

ID	word	document frequency
1	course	1
2	data	2
3	interest	1
4	mine	3
5	study	1
6	subfield	1
7	text	2

low IDF 因为三个文件都有，无法区别

This is a data mining course.

mine

~~This is a~~ data ~~mining~~ course

course:1, data:1, mine:1

We are studying text mining. Text mining is a subfield of data mining.

study mine text mine

~~We are studying~~ text ~~mining~~ ~~Text~~ ~~mining~~ ~~is a~~ subfield ~~of~~ data ~~mining~~

mine

data:1, mine:3, study:1, subfield:1, text:2

Mining text is interesting, and I am interested in it.

mine interest

~~Mining~~ text ~~is interesting and I am interested in it~~

interest

interest:2, mine:1, text:1

Step 7 – Create the vector space model

This is a data mining course.

mine
~~This is a~~ data ~~mining~~ course

course:1, data:1, mine:1

(1, 1, 0, 1, 0, 0, 0)

右表中word在这个document出现的次数

We are studying text mining. Text mining is a subfield of data mining.

study mine text mine
~~We are studying~~ text ~~mining Text mining~~
~~is a~~-subfield ~~of~~ data ~~mining~~
mine

data:1, mine:3, study:1, subfield:1, text:2

(0, 1, 0, 3, 1, 1, 2)

Mining text is interesting, and I am interested in it.

mine interest
~~Mining~~ text ~~is interesting and I am~~
~~interested in it~~
interest

interest:2, mine:1, text:1

(0, 0, 2, 1, 0, 0, 1)

ID	word	document frequency
1	course	1
2	data	2
3	interest	1
4	mine	3
5	study	1
6	subfield	1
7	text	2

Step 8 – Compute inverse document frequency

This is a data mining course. → (1, 1, 0, 1, 0, 0, 0)

We are studying text mining. Text mining is a subfield of data mining. → (0, 1, 0, 3, 1, 1, 2)

Mining text is interesting, and I am interested in it. → (0, 0, 2, 1, 0, 0, 1)

$$IDF(word) = \log \frac{\text{total documents}}{\text{document frequency}}$$

ID	word	document frequency	IDF
1	course	1	0.477
2	data	2	0.176
3	interest	1	0.477
4	**mine**	**3**	**0**
5	study	1	0.477
6	subfield	1	0.477
7	text	2	0.176

Step 9 – Compute term weights

This is a data mining course.

→ (1, 1, 0, 1, 0, 0, 0)

(0.477, 0.176, 0, 0, 0, 0, 0)

We are studying text mining. Text mining is a subfield of data mining.

→ (0, 1, 0, 3, 1, 1, 2)

(0, 0.176, 0, 0, 0.477, 0.477, 0.352)

Mining text is interesting, and I am interested in it.

→ (0, 0, 2, 1, 0, 0, 1)

(0, 0, 0.954, 0, 0, 0, 0.176)

ID	word	document frequency	IDF
1	course	1	0.477
2	data	2	0.176
3	interest	1	0.477
4	mine	3	0
5	study	1	0.477
6	subfield	1	0.477
7	text	2	0.176

$$w(word_i) = TF(word_i) \times IDF(word_i)$$

$$TF(word_i) = \text{number of times } word_i \text{ appears in the document}$$

Step 10 – Normalize all documents to unit length

This is a data mining course.

(0.477, 0.176, 0, 0, 0, 0, 0)

(0.938, 0.346, 0, 0, 0, 0, 0)

We are studying text mining. Text mining is a subfield of data mining.

(0, 0.176, 0, 0, 0.477, 0.477, 0.352)

(0, 0.225 0, 0, 0.611, 0.611, 0.450)

$$w(word_i) = \frac{w(word_i)}{\sqrt{w^2(word_1) + w^2(word_2) + \cdots + w^2(word_n)}}$$

Mining text is interesting, and I am interested in it.

(0, 0, 0.954, 0, 0, 0, 0.176)

(0, 0, 0.983, 0, 0, 0, 0.181)

Normalization

This is a data mining course.

$\rightarrow$ (1, 1, 0, 1, 0, 0, 0)

(0.477, 0.176, 0, 0, 0, 0, 0)

$$w(\text{course}) = \frac{0.477}{\sqrt{0.477^2 + 0.176^2 + 0 + 0 + 0 + 0 + 0}} = 0.938$$

(0.938, 0.346, 0, 0, 0, 0, 0)

A Running Example

- Documents become structured!
  - We can perform classification, clustering, etc

This is a data mining course.	→ (0.938, 0.346, 0, 0, 0, 0, 0)
We are studying text mining. Text mining is a subfield of data mining.	→ (0, 0.225 0, 0, 0.611, 0.611, 0.450)
Mining text is interesting, and I am interested in it.	→ (0, 0, 0.983, 0, 0, 0, 0.181)

Querying Documents

- How can we query the document collection?
  - Similar to the previous steps:
    1. Remove stop words
    2. Stem every word in the query string
    3. Transform the query string into a vector space model (VSM) by using the TD-IDF scheme
    4. Normalize the VSM into unit length

Querying Documents - Example

Query ***Q = {interesting data and text}***

Step 1: Remove stop words
(interesting data text)

Step 2: Stemming
(interest data text)

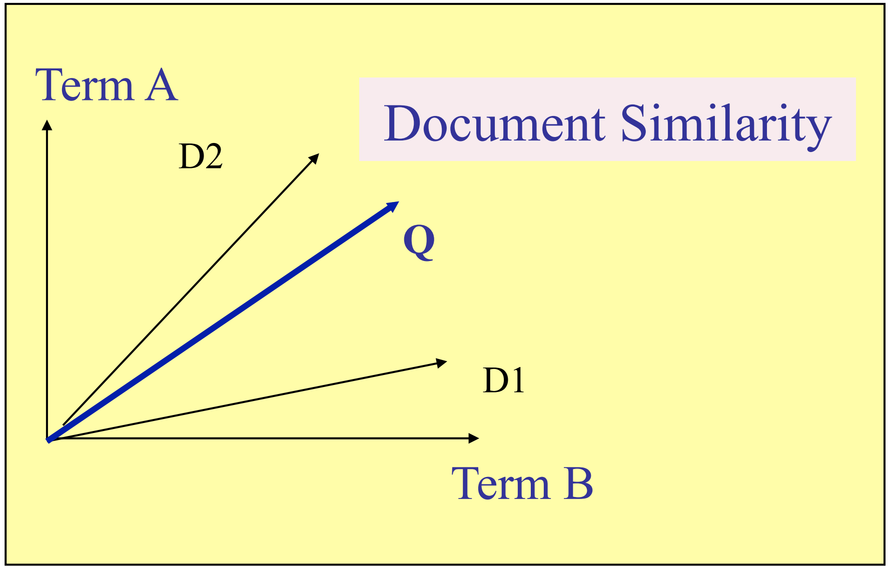
Step 3: Construct a vector space model
(0, 1, 1, 0, 0, 0, 1)

Step 4: Compute the weight of each word
(0, 0, 0.477, 0, 0, 0, 0.176)

Step 5: Normalize the vector space model
(0, 0, 0.938, 0, 0, 0, 0.346)

ID	word	document frequency	IDF
1	course	1	0.477
2	data	2	0.176
3	interest	1	0.477
4	mine	3	0
5	study	1	0.477
6	subfield	1	0.477
7	text	2	0.176

Querying Document by Similarity

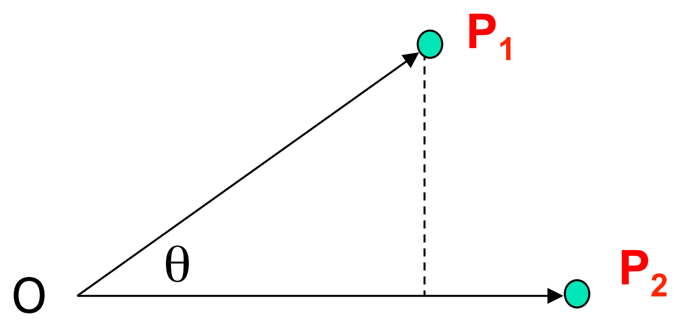


Cosine Distance

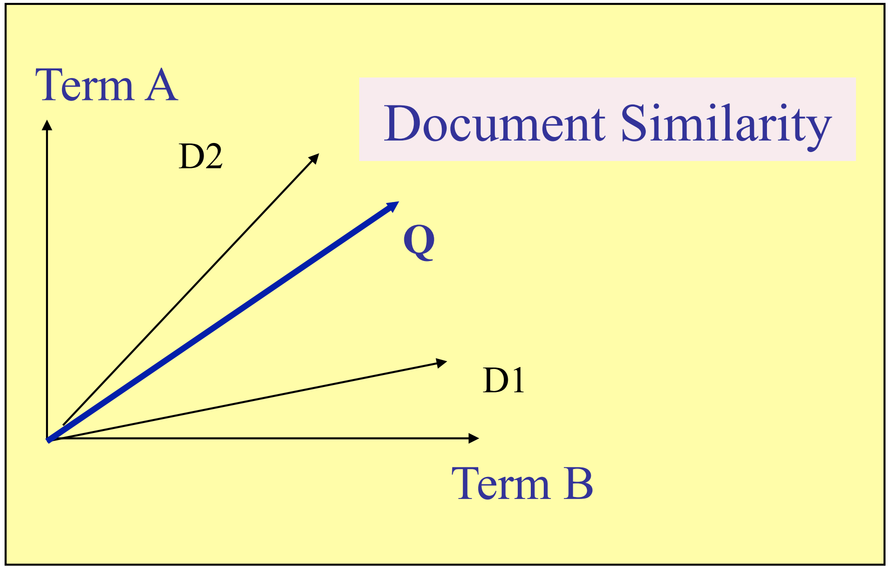
- Measures the distance between two vectors
  - Think of a point as:
    - a vector from the origin (0,0,…,0) to its location
  - Two point-vectors make an angle

angle cosine is $\cos(p_1, p_2) = (p_1 \bullet p_2) / ||p_1||\ ||p_2||$,

where • indicates vector dot product and $||\ d\ ||$ is the length of vector d.



Querying Document by Similarity



$$sim(Q,D) = \frac{\sum_{k=1}^{t} w_{qk} \bullet w_{dk}}{\sqrt{\sum_{k=1}^{t} (w_{qk})^2 \bullet \sum_{k=1}^{t} (w_{dk})^2}}$$

Example – Result

Q: (0, 0, 0.938, 0, 0, 0, 0.346)

Document D1: (0.938, 0.346, 0, 0, 0, 0, 0)

Document D2: (0, 0.225 0, 0, 0.611, 0.611, 0.450)

Document D3: (0, 0, 0.983, 0, 0, 0, 0.181)

$$cosine(P,Q) = \frac{\sum p_i \cdot q_i}{\sqrt{\sum p_i^2 \times \sum q_i^2}}$$

Example – Result

Q: (0, 0, 0.938, 0, 0, 0, 0.346)

Document D1: (0.938, 0.346, 0, 0, 0, 0, 0)

Document D2: (0, 0.225 0, 0, 0.611, 0.611, 0.450)

Document D3: (0, 0, 0.983, 0, 0, 0, 0.181)

$$\text{cosine}(P,Q) = \frac{\sum p_i \cdot q_i}{\sqrt{\sum p_i^2 \times \sum q_i^2}}$$

$$\text{cosine}(D1,Q) = 0$$

$$\text{cosine}(D2,Q) = \frac{0.346 \times 0.450}{\sqrt{(0.938^2 + 0.346^2) \times (0.225^2 + 0.611^2 + 0.611^2 + 0.450^2)}} = 0.156$$

$$\text{cosine}(D3,Q) = \frac{0.938 \times 0.983 + 0.346 \times 0.181}{\sqrt{(0.938^2 + 0.346^2) \times (0.983^2 + 0.181^2)}} = 0.985$$

Return Document 3

Example!

- Given a query of "W4 W5" and a collection of the following three documents:

- Document 1: "W1 W2 W3 W4 W5"
- Document 2: "W6 W7 W4 W5"
- Document 3: "W8 W3 W9 W4 W10"

- Use the Vector Space Model, TF/IDF weighting scheme, and Cosine vector similarity measure to find the most relevant document(s) to the query.

TF-IDF

Term list	TF	IDF
W1	1	0.477
W2	1	0.477
W3	2	0.176
W4	3	0
W5	2	0.176
W6	1	0.477
W7	1	0.477
W8	1	0.477
W9	1	0.477
W10	1	0.477

VSM

- D1=(1,1,1,1,1,0,0,0,0,0)
  (0.477,0.477,0.176,0,0.176,0,0,0,0,0)
- D2=(0,0,0,1,1,1,1,0,0,0)
  (0,0,0,0,0.176,0.477,0.477, 0,0,0)
- D3=(0,0,1,1,0,0,0,1,1,1)
  (0,0,0.176,0,0,0,0,0.477,0.477, 0.477)

Normalization

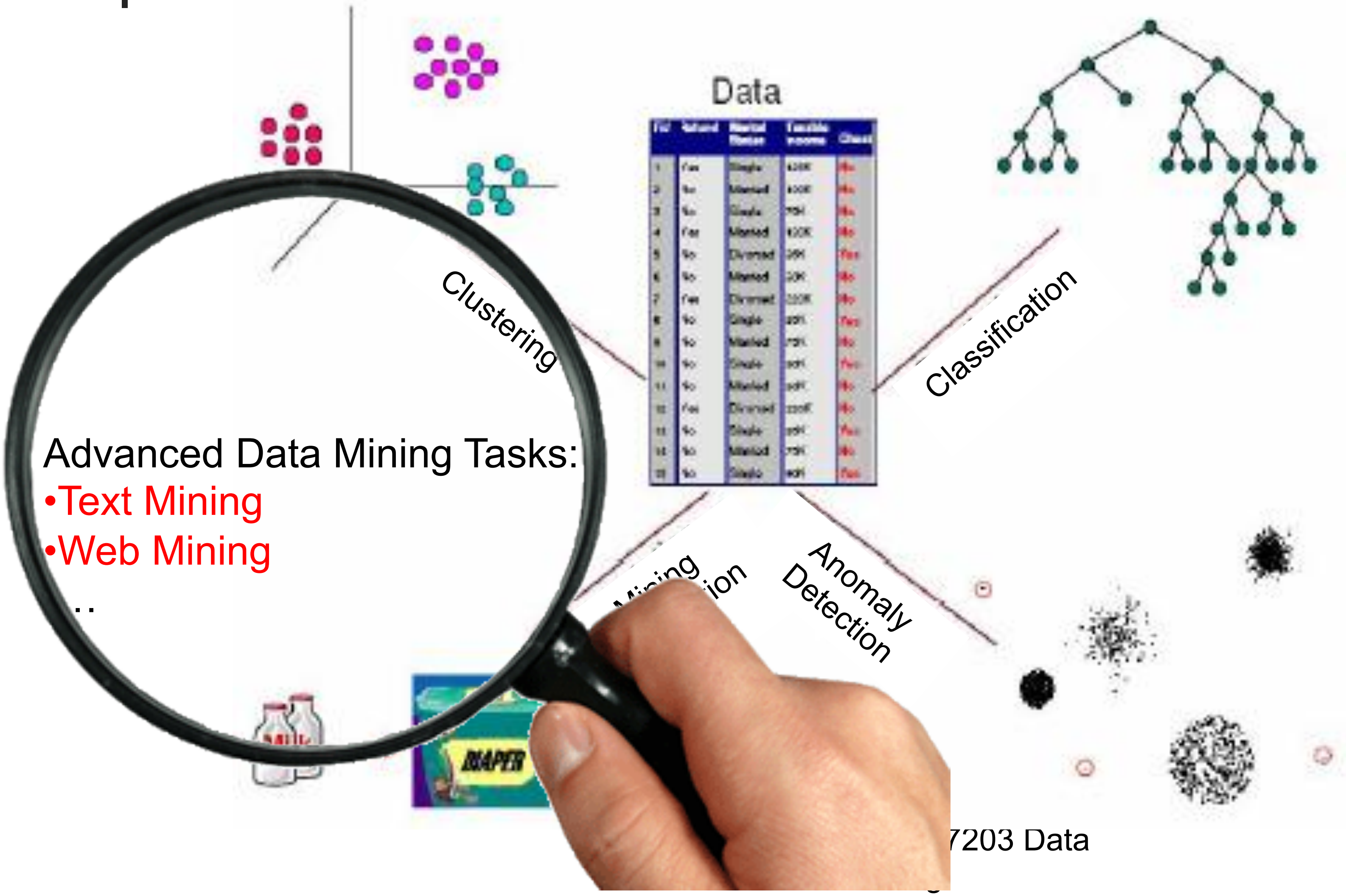
- D1= [0.6634 0.6634 0.2448 0 0.2448 0 0 0 0 0]
- D2= [0 0 0 0 0.2525 0.6842 0.6842 0 0 0]
- D3= [0 0 0.2084 0 0 0 0 0.5647 0.5647 0.5647]

Query

- Q=(0,0,0,0,0.176,0,0,0,0,0)
  (0,0,0,0,1,0,0,0,0,0)

- Cosine_sim(Q,D1)=0.2448
- Cosine_sim(Q,D2)=0.2525
- Cosine_sim(Q,D3)=0

Data Mining Tasks



Advanced Data Mining Tasks:
- Text Mining
- Web Mining
- ..