

Data Mining

-- Text Mining

1

Text Mining

Synthesis of ...

- Information Retrieval
 - the science of searching for documents, for **information** within documents, and for **metadata** about documents, as well as that of searching **relational databases** and the **World Wide Web**.
- Natural Language Processing
 - Part of Speech Tagging
 - Phrase Chunking
 - Deep Parsing
 - Named Entity Recognition
 - Information Extraction

2

Information Retrieval

The screenshot shows a Google search for 'text mining'. The search bar contains 'text mining' and the search button is visible. Below the search bar, there are several search results. The first result is from 'www.clarabridge.com' titled 'Text Analytics Industry's First Self Service Offering, Sign Up Now'. The second result is from 'Wikipedia, the free encyclopedia' titled 'Text mining - Wikipedia, the free encyclopedia'. The third result is from 'MultiSearch' titled 'What is Text Mining?'. The fourth result is from 'Data Mining Software | SAS' titled 'Text mining applies the same analysis techniques to text-based documents. The knowledge gleaned from data and text mining can be used to fuel strategic ...'.

3

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 entity recognition:

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

4

Text Mining Tasks

- Text Classification
 - Assigning a document to one of several pre-specified classes
- Text Clustering
 - Unsupervised learning
- Text Summarization
 - Extracting a summary for a document
 - Based on syntax and semantics
-

5

Challenge of Text Mining

- In traditional data mining, all data are "structured".
 - We usually store the data into database.
 - Table structure. Very clear.
 - Every attribute is well defined.
 - We understand the record very well.
 - Each record is defined by a set of attributes
 - We can measure the similarity between any pair.

6

Challenge of Text Mining

- However, in text mining, data are "unstructured"!
 - Example:
 - Given two documents, how can you compute their similarity?
Base on what?
- So, what we need to do...
 - Unstructured => Structured
- In other words...
 - How to represent a document "structurally"???
 - Document representation problem.

7

Document Representation

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

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

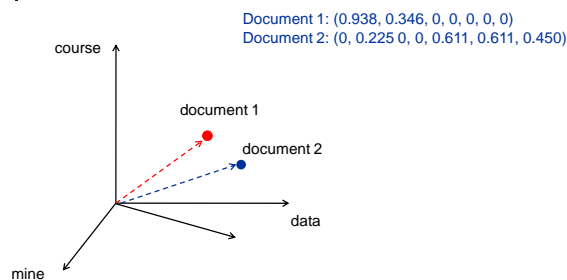
8

Vector Space Model

- Each word is a dimension
 - If we have M different words. Then, we have a M -dimensional vector space.
- Each document is regarded as a point in this vector space.
 - $d = \{w_1, w_2, \dots, w_m\}$
In term of geometry, w_i is the coordinate of dimension i in d .
Yet, conceptually, w_i denotes the importance of word i in d .

9

An Example of VSM



10

Vector Space Model

- Problems:
 1. There are soooooooooooooo many English words!
 2. How to determine the "importance of the words"?

11

Vector Space Model

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

12

Vector Space Model

- The second problem: how to determine the importance of the terms
- We solve the second problem by:
 - Using a weighting schema, the TF-IDF schema:

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

$TF(word_i)$ = number of times $word_i$ appears in the document

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

- Normalize the document into unit length

13

TF-IDF

- TF-IDF
 - Term frequency-inverse document frequency
 - Evaluate how important a **word** is to a **document** in a **collection**
 - 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**.

14

Why TF-IDF

- Can we simply use term frequency?
 - diminish the weight of terms that occur very frequently in the collection
 - increase the weight of terms that occur rarely in the collection
 - Example, the, a, ...
 - Example, UQ

15

TF-IDF Calculation

Term Importance:

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

Term Frequency :

$TF(word_i)$ = number of times $word_i$ appears in the document

Inverse Document Frequency :

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

16

A Running Example 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

17

A Running Example Step 2 – Remove stopwords

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

18

A Running Example 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. → 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.

19

A Running Example Step 4 – Stemming

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.

20

A Running Example Step 5 – Count the word frequencies

This is a data mining course. → 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. → We are studying text mining. Text mining is a subfield of data mining.
data:1, mine:3, study:1, subfield:1, text:2

Mining text is interesting, and I am interested in it. → Mining text is interesting and I am interested in it.
interest:2, mine:1, text:1

21

A Running Example Step 6 – Create an indexing file

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.

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

22

A Running Example Step 7 – Create the vector space model

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

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

23

A Running Example Step 8 – Compute the 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

24

A Running Example Step 9 – Compute the weights of the words

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)

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

$w(\text{word}_i) = TF(\text{word}_i) \times IDF(\text{word}_i)$
 $TF(\text{word}_i) = \text{number of times word}_i \text{ appears in the document}$

25

A Running Example Step 10 – Normalize all documents to unit length

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)

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

$w(\text{word}_i) = \frac{w(\text{word}_i)}{\sqrt{w^2(\text{word}_1) + w^2(\text{word}_2) + \dots + w^2(\text{word}_n)}}$

26

Normalization

This is a data mining course. → (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

Everything become structural!

- 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)

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

28

Query A Document

- How can we query the document?
 - Simple! Just similar to the previous steps:
 - Remove stopwords.
 - Stem every word of the query string.
 - Transform the query string into a vector space model (VSM) by using TD-IDF schema.
 - Normalize the VSM into unit length.

29

A Running Example Q = {interested in interesting data and text}

Original Query: (interested in interesting data and text)

Step 1: Remove stop word: (interested interesting data text)

Step 2: Stemming: (interest interest data text)

Step 3: Remove duplication: (interest data text)

Step 4: Construct a vector space model: (0, 1, 1, 0, 0, 0, 1)

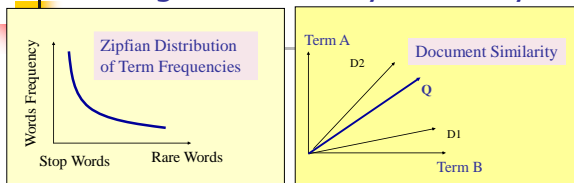
Step 5: Compute the weight of each word: (0, 0, 0.477, 0, 0, 0, 0.176)

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

30

Ranking Document by Similarity



Vector similarity (dot product):

$$sim(Q, D) = \sum_{k=1}^t w_{qk} \cdot w_{dk}$$

Cosine vector similarity:

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

31

A Running Example – The Result

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

Document 1: (0.938, 0.346, 0, 0, 0, 0)
 Document 2: (0, 0.225, 0, 0, 0.611, 0.450)
 Document 3: (0, 0, 0.983, 0, 0, 0.181)

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

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

$$\cosine(D1, Q) = 0$$

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

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

Conclusion: Return Document 3

32

QUIZ!

- 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.

33

IDF

Word list	DF	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

34

VSM

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

35

Normalization

- D1= [0.6634 0.6634 0.2448 0 0.2448 0 0 0 0 0]
- D2= [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]

36

Query

- $Q=(0,0,0,0,0.176,0,0,0,0,0)$
 $(0,0,0,0,1,0,0,0,0)$
- $\text{Cosine_sim}(Q,D1)=0.2448$
- $\text{Cosine_sim}(Q,D2)=0.2525$
- $\text{Cosine_sim}(Q,D3)=0$

37

Simple? Well...

- What we have discussed so far is a general framework only.
- There are still a lot of issues:
 - How to define stopwords?
 - "A" is usually regarded as a stopwords. However, "Vitamin A" may be an important term in an article.
 - How to perform stemming?
 - What to stem and what not to stem?
 - Should "booking" be converted to "book"?
 - How to stem? There are many new words everyday!

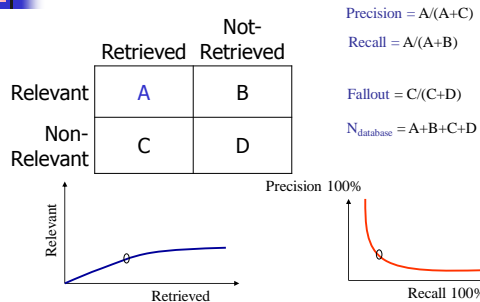
38

Spelling error?

- Spelling error always appears in documents! Should we consider two similar word as a same word?
 - Are they the same: "classification" and "classificatiam"?
 - But then, how about "Information" and "informatics"?

39

IR Evaluation: Recall and Precision



40

Summary

- Information Retrieval Concepts
 - VSM Model
 - Similarity Measure
 - IR Evaluations

41