

Data Mining - Text Mining

By Dr Heng Tao SHEN
School of Information Technology and Electrical Engineering
The University Of Queensland
<http://www.itee.uq.edu.au/~shenht>

Outline

- Motivation
- Text Mining Techniques
- Performance Evaluations

P. 2

Motivation

- Information filtering & retrieval
- Q&A
- Information reasoning
- Document classification
- Document management
 - Routing, Summarizing, Indexing, Labelling, categorization, etc.

P. 3

Question

- **How Do We Mine Text Data?**

P. 4

Challenge of Text Mining

- In traditional data mining, all data are “structural”.
 - 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.
- However, in text mining, data are “unstructural”!!!
 - Example:
 - Given two documents, how can you compute their similarity? Base on what?

P. 5

Challenge of Text Mining

- So, what we need to do...
 - Unstructural => Structural
- In other words...
 - How to represent a document “structurally”???
 - Document representation problem.

P. 6

Information Retrieval Techniques

- Approaches
 - VSM (Vector Space Model)
 - TF/IDF (Term Frequency/Inverse Document Frequency)
 - CF (Collaborative Filtering)
- Principles & Laws in IR
 - Similarity Computation
 - IR evaluation: Recall and Precision
 - Power Law Distribution

P.7

Document Representation

- Vector Space Model
 - Each word is a dimension
 - Hence, in text mining, we may deal with a few hundred thousand dimensions!
 - 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 .
- Problems:
 1. There are soooooooooooooo many English words!
 2. How to determine the "importance of the words"?

P.8

Document Representation

- We solve the first problem by:
 - Remove stopword
 - Stemming
- We solve the second problem by:
 - Using a weighting schema, the tf-idf schema:

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

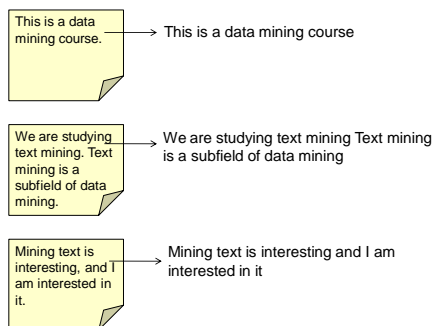
$$TF(word_i) = \text{number of times } word_i \text{ appears in the document}$$

$$IDF(word_i) = \log \frac{\text{total documents}}{\text{document frequency}}$$
 - Normalize the document into unit length

P.9

A Running Example

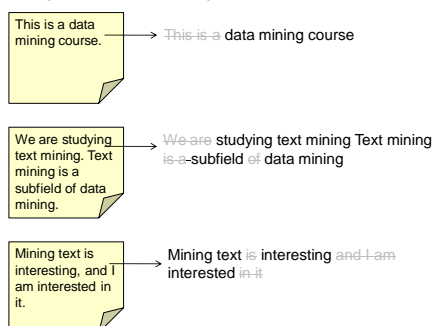
- Step 1 – Extract text (i.e. no preposition)



P.10

A Running Example

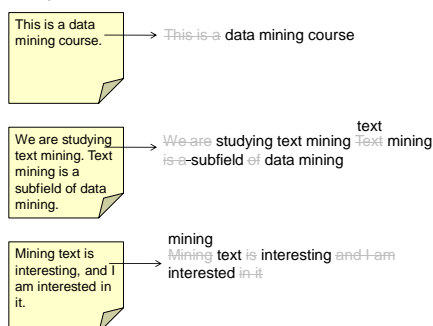
- Step 2 – Remove stopword



P.11

A Running Example

- Step 3 – Convert all words to lower case



P.12

A Running Example

• Step 4 – Stemming

This is a data mining course. → This is a data ^{mine} ~~mining~~ course

We are studying text mining. Text mining is a subfield of data mining. → We are studying ^{study} ~~text~~ ^{mine} ~~mining~~ ^{text} ~~mining~~ ^{mine} ~~mining~~
is a subfield of ^{data} ~~mining~~ ^{mine} ~~mining~~

Mining text is interesting, and I am interested in it. → ^{mine} ~~mining~~ ^{interest} ~~text~~ is interesting and I am interested in ^{interest} ~~it~~

P. 13

A Running Example

• Step 5 – Count the word frequencies

This is a data mining course. → This is a ^{data} ~~mining~~ ^{course} ~~course~~ ^{coursex1, datax1, minex1}

We are studying text mining. Text mining is a subfield of data mining. → We are studying ^{study} ~~text~~ ^{mine} ~~mining~~ ^{text} ~~mining~~ ^{mine} ~~mining~~
is a subfield of ^{data} ~~mining~~ ^{mine} ~~mining~~ ^{datax1, minex2, studyx1, subfieldx1, textx2}

Mining text is interesting, and I am interested in it. → ^{mine} ~~mining~~ ^{interest} ~~text~~ is interesting and I am interested in ^{interest} ~~it~~ ^{interestx2, minex1, textx1}

P. 14

A Running Example

• Step 6 – Create an indexing file

This is a data mining course. → This is a ^{data} ~~mining~~ ^{course} ~~course~~ ^{coursex1, datax1, minex1}

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

We are studying text mining. Text mining is a subfield of data mining. → We are studying ^{study} ~~text~~ ^{mine} ~~mining~~ ^{text} ~~mining~~ ^{mine} ~~mining~~
is a subfield of ^{data} ~~mining~~ ^{mine} ~~mining~~ ^{datax1, minex2, studyx1, subfieldx1, textx2}

Mining text is interesting, and I am interested in it. → ^{mine} ~~mining~~ ^{interest} ~~text~~ is interesting and I am interested in ^{interest} ~~it~~ ^{interestx2, minex1, textx1}

P. 15

A Running Example

• Step 7 – Create the vector space model

This is a data mining course. → This is a ^{data} ~~mining~~ ^{course} ~~course~~ ^{coursex1, datax1, minex1}
 $(1, 1, 0, 1, 0, 0, 0)$

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

We are studying text mining. Text mining is a subfield of data mining. → We are studying ^{study} ~~text~~ ^{mine} ~~mining~~ ^{text} ~~mining~~ ^{mine} ~~mining~~
is a subfield of ^{data} ~~mining~~ ^{mine} ~~mining~~ ^{datax1, minex2, studyx1, subfieldx1, textx2}
 $(0, 1, 0, 2, 1, 1, 2)$

Mining text is interesting, and I am interested in it. → ^{mine} ~~mining~~ ^{interest} ~~text~~ is interesting and I am interested in ^{interest} ~~it~~ ^{interestx2, minex1, textx1}
 $(0, 0, 2, 1, 0, 0, 1)$

P. 16

A Running Example

• Step 8 – Compute the inverse document frequency

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

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

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

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

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

P. 17

A Running Example

• Step 9 – Compute the weights of the words

This is a data mining course. → $(1, 1, 0, 1, 0, 0, 0)$
 $(0.477, 0.176, 0, 0, 0, 0, 0)$

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

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

We are studying text mining. Text mining is a subfield of data mining. → $(0, 1, 0, 2, 1, 1, 2)$
 $(0, 0.176, 0, 0, 0.477, 0.477, 0.352)$

Mining text is interesting, and I am interested in it. → $(0, 0, 2, 1, 0, 0, 1)$
 $(0, 0, 0.954, 0, 0, 0, 0.176)$

P. 18

A Running Example

- Step 10 – Normalize all documents to unit length

This is a data mining course.

(1, 1, 0, 1, 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, 1, 0, 2, 1, 1, 2)
(0, 0.225, 0, 0, 0.611, 0.611, 0.450)

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

(0, 0, 2, 1, 0, 0, 1)
(0, 0, 0.983, 0, 0, 0, 0.181)

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

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

P. 19

A Running Example

- Finally, we obtain the following:
 - 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

P. 20

Query A Document

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

P. 21

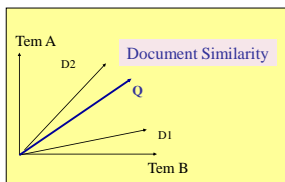
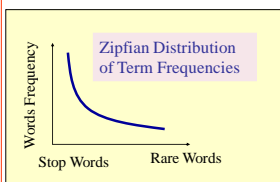
A Running Example

- 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

P. 22

Ranking Document by Similarity



Vector similarity (dot product):

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

Cosine vector similarity:

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

Question: What is the difference between cosine vector similarity and the Euclidean similarity?

A Running Example – The Result

- Q = {interested in interesting data and text}

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

Document 1: (0.938, 0.346, 0, 0, 0, 0, 0)

Document 2: (0, 0.225, 0, 0, 0.611, 0.611, 0.450)

Document 3: (0, 0, 0.983, 0, 0, 0, 0.181)

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

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

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

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

Conclusion: Return Document 3

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

P. 24

Simple? Well...

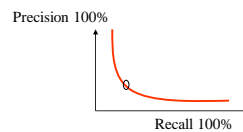
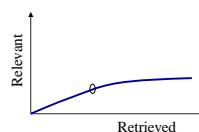
- What we have discussed so far is a general framework only.
- There are still a lot of issues:
 - How to define stopword?
 - "A" is usually regarded as a stopword. However, "Vitamin A" may be an important term in an article.
 - How to perform stemming?
 - What to stem and what not to stem?
 - Should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

P. 25

IR Evaluation: Recall and Precision

	Retrieved	Not- Retrieved
Relevant	A	B
Non- Relevant	C	D

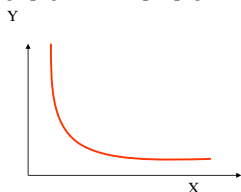
$Precision = A/(A+C)$
 $Recall = A/(A+B)$
 $Fallout = C/(C+D)$
 $N_{database} = A+B+C+D$
 $Generality = A / N_{database}$



Power Law on Internet

$$x \cdot y = C$$

e.g., page# • link# per page



"The Laws of the Web: Patterns in the Ecology of Information", by Huberman, B. A.; Huberman, Bernardo A.
ISBN: 0262083035

P. 27

Summary

- Information Retrieval Concepts
 - VSM Model
 - Similarity Measure
 - IR Evaluations
 - Power Law on the Internet
- Next Week:
 - Web Mining

P. 28