

COMP3420: Advanced Databases and Data Mining

Text data mining

Lecture outline

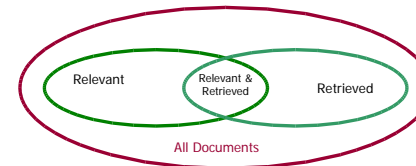
- Text data analysis and text/information retrieval
 - Basic measures for text/information retrieval
 - Information retrieval techniques
 - Boolean and vector space model
 - Similarity-based retrieval in text data
 - TF-IDF weighting
 - Vector space model
- Types of text data mining
 - Keyword based association analysis
 - Text classification and categorisation
 - Document clustering

Text data analysis and information retrieval

- Typical information retrieval systems
 - Online library catalogs
 - Online document management systems
 - Internet search engines
- Information retrieval (IR) versus database (DB) systems
 - Some DB problems are not present in IR, such as: updates, transaction management, complex structured objects
 - Some IR problems are not addressed well in DBMS, for example: unstructured documents, approximate search using keywords and relevance

3

Basic measures for text retrieval



- Precision: the percentage of retrieved documents that are in fact relevant to the query (i.e., the “correct” responses)

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

- Recall: the percentage of documents that are relevant to the query and were, in fact, retrieved

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

4

Information retrieval techniques

- Basic concepts
 - A document can be described by a set of representative keywords called *index terms*
 - Different index terms have varying relevance when used to describe document contents
 - This effect is captured through the *assignment of numerical weights* to each index term of a document (for example, frequency, or TF-IDF)
- DBMS analogy
 - Index terms → Attributes
 - Weights → Attribute values

5

Information retrieval techniques (2)

- Index terms (attribute) selection
 - Stop word list
 - Word stem
 - Index terms weighting methods
- Term and document frequency matrices
- Information retrieval models
 - Boolean model
 - Vector model
 - Probabilistic model (categories modeled by probability distributions, find likelihood a document belongs to a certain category, similar to Bayesian classification)

6

Boolean model

- Consider that index terms are either present or absent in a document
 - For example: $1=$ present, $0=$ absent
 - As a result, the index term weights are assumed to be all binaries
- A query is composed of index terms linked by three connectives: *not*, *and*, and *or*
 - For example: “*car and repair*”, “*plane or airplane*”
- The Boolean model predicts that each document is either relevant or non-relevant based on the match of a document to the query

7

Similarity-based retrieval in text data

- Finds similar documents based on a set of common keywords
- Answer should be based on the degree of relevance based on the nearness of the keywords, relative frequency of the keywords, etc.
- Use of stop lists
 - Set of words that are deemed “irrelevant”, even though they may appear frequently
 - For example: *a*, *the*, *of*, *for*, *to*, *with*, etc.
 - Stop lists may vary when document set varies

8

Similarity-based retrieval in text data (2)

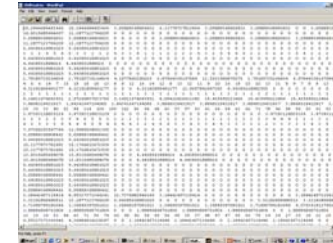
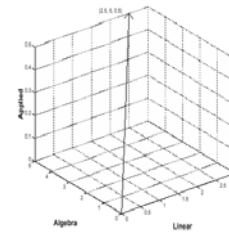
- Apply word stemming
 - Several words are small syntactic variants of each other since they share a common word stem
 - For example, *drug, drugs, drugged* → *drug*
- A term and document frequency matrix (or table)
 - Each entry $frequent_table(i, j) =$ number of occurrences of the word t_i in document d_j
 - Usually, the *ratio* instead of the absolute number of occurrences is used
- Similarity metrics: measure the closeness of a document to a query (a set of keywords)
 - Relative term occurrences
 - Cosine distance:

$$sim(v_1, v_2) = \frac{v_1 \cdot v_2}{|v_1| |v_2|}$$

9

Vector space model

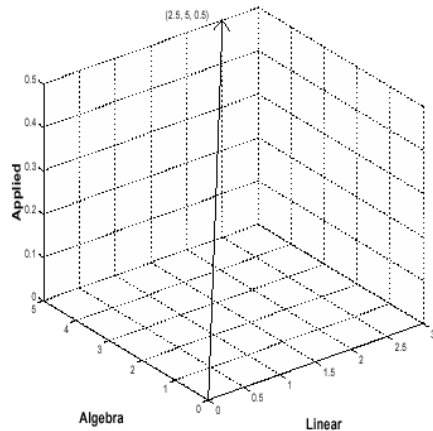
- Documents and user queries are represented as m -dimensional vectors, where m is the total number of index terms in the document collection
- The degree of similarity of the document d with regard to the query q is calculated as the correlation between the vectors that represent them, using measures such as the *Euclidean distance* or the *cosine* of the angle between these two vectors



Source: Han and Kamber, DM Book, 2nd Ed. (Copyright © 2006 Elsevier Inc.)

10

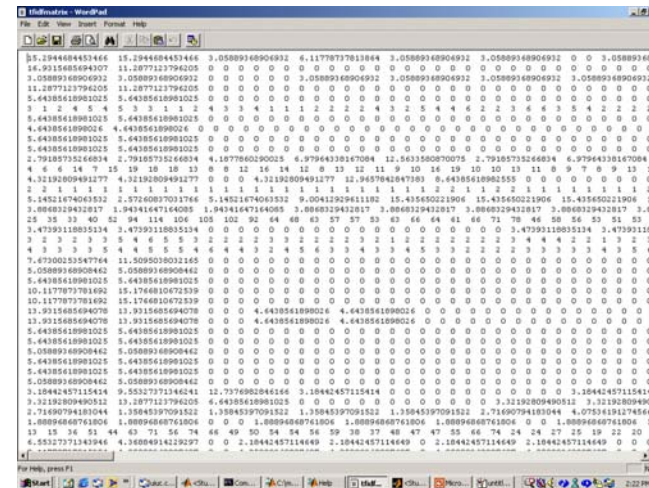
Vector space model (2)



Source: Han and Kamber, DM Book, 2nd Ed. (Copyright © 2006 Elsevier Inc.)

11

Vector space model (3)



12

Vector space model (4)

- Represent a document by a *term* or *feature vector*
 - Term: basic concept, for example, *word* or *phrase* (like “data mining”)
 - Each term defines one dimension (large number of dimensions!)
 - N terms define a N -dimensional space
 - Element of vector corresponds to term weight
 - For example, $d = (x_1, \dots, x_N)$, x_i is “importance” of term i
 - These term vectors are *sparse* (most weights are 0)
- New document is assigned to the most likely category based on vector similarity

13

How to assign weights

- Two-fold heuristics based on frequency
- TF (Term Frequency)

- More frequent *within* a document \rightarrow more relevant to semantics
- For example, “classification” versus “SVM”
- Raw TF = $f(t, d)$ (how many times term t appears in doc d)
- Document length varies \Rightarrow relative frequency preferred
- Perform normalisation (for example, *maximum frequency normalisation*)

$$TF(t, d) = 0.5 + \frac{0.5 * f(t, d)}{MaxFreq(d)}$$

- IDF (Inverse Document Frequency)

- Less frequent *among* documents \rightarrow more discriminative
- For example “algebra” versus “science”
- Formula:
 - n = total number of documents
 - k = number of documents with term t appearing

$$IDF(t) = 1 + \log\left(\frac{n}{k}\right)$$

14

TF-IDF weighting

- TF-IDF weighting: $weight(t, d) = TF(t, d) * IDF(t)$
 - Frequent within doc \rightarrow high TF \rightarrow high weight
 - Selective among docs \rightarrow high IDF \rightarrow high weight
- Recall vector space model
 - Each selected term represents one dimension
 - Each document is represented by a *term* or *feature vector*
 - Its t -term coordinate of document d is the TF-IDF weight
- Just for illustration ...
 - Many complex and more effective weighting variants exist in practice

15

How to measure similarity?

- Given two document

$$D_i = (w_{i1}, w_{i2}, \dots, w_{iN}) \quad D_j = (w_{j1}, w_{j2}, \dots, w_{jN})$$

- Similarity definition

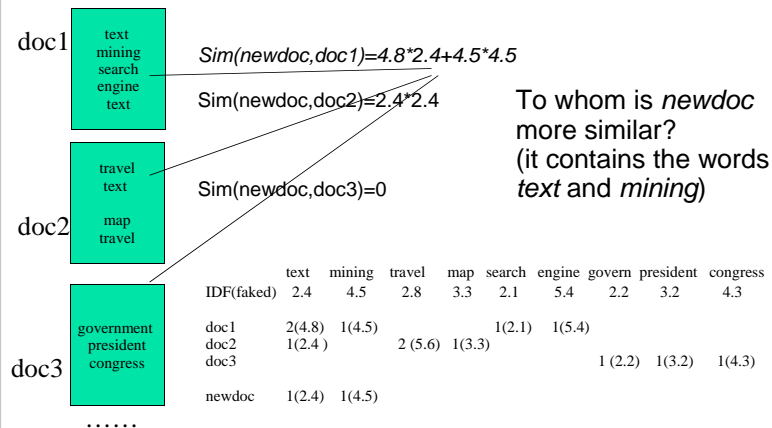
- Dot product $Sim(D_i, D_j) = \sum_{t=1}^N w_{it} * w_{jt}$

Normalised dot product (or *cosine similarity*)

$$Sim(D_i, D_j) = \frac{\sum_{t=1}^N w_{it} * w_{jt}}{\sqrt{\sum_{t=1}^N (w_{it})^2 * \sum_{t=1}^N (w_{jt})^2}}$$

16

Illustrative example



17

Vector space model-based classifiers

- What do we have so far?
 - A feature space with similarity measure
 - This is a classic supervised learning problem
 - Search for an approximation to classification hyper plane
- Vector space model based classifiers
 - Decision tree based
 - Neural networks
 - Support vector machine
 - ...

18

Types of text data mining


- Keyword-based association analysis
- Automatic document classification
- Similarity detection
 - Cluster documents by a common author
 - Cluster documents containing information from a common source
- Link analysis: unusual correlation between entities
- Sequence analysis: predicting a recurring event
- Anomaly detection: find information that violates usual patterns
- Hypertext analysis
 - Patterns in anchors/links (for example, anchor text correlations with linked objects)
- Applications: news article classification, automatic e-mail filtering, Web page classification, etc.

19

Keyword-based association analysis

- Motivation
 - Collect sets of keywords or terms that occur frequently together and then find the *association* or *correlation* relationships among them
- Association analysis process
 - Pre-process the text data by parsing, stemming, removing stop words, etc.
 - Evoke association mining algorithms
 - Consider each document as a transaction
 - View a set of keywords in the document as a set of items in the transaction
 - Term level association mining
 - No need for human effort in tagging documents
 - The number of meaningless results and the execution time is greatly reduced

20



Document clustering

- **Motivation**

- Automatically group related documents based on their contents
- No predetermined training sets or taxonomies
- Generate a taxonomy at runtime

- **Clustering process**

- Data preprocessing: remove stop words, stem, feature extraction, lexical analysis, etc.
- Hierarchical clustering: compute similarities applying clustering algorithms
- Model-based clustering (neural network approach): clusters are represented by "exemplars" (for example Self-Organising Maps, SOM)