CPSC 304
Introduction to Database Systems

Introduction to Information Retrieval

Textbook Reference
Database Management Systems: Sections 27.1 - 27.2

Hazra Imran

Based partly on Ramakrishnan & Gehrke, DB Management Systems

Recap : TF x IDF Calculation

$$w_{ik} = tf_{ik} * \log(N / n_k)$$

$T_k = \text{term } k \text{ in document } D_i$

$tf_{ik} = \text{frequency of term } T_k \text{ in document } D_i$

$idf_k = \text{inverse document frequency of term } T_k \text{ in } C$

$N = \text{total number of documents in the collection } C$

$n_k = \text{the number of documents in } C \text{ that contain } T_k$

$$idf_k = \log\left(\frac{N}{n_k}\right)$$

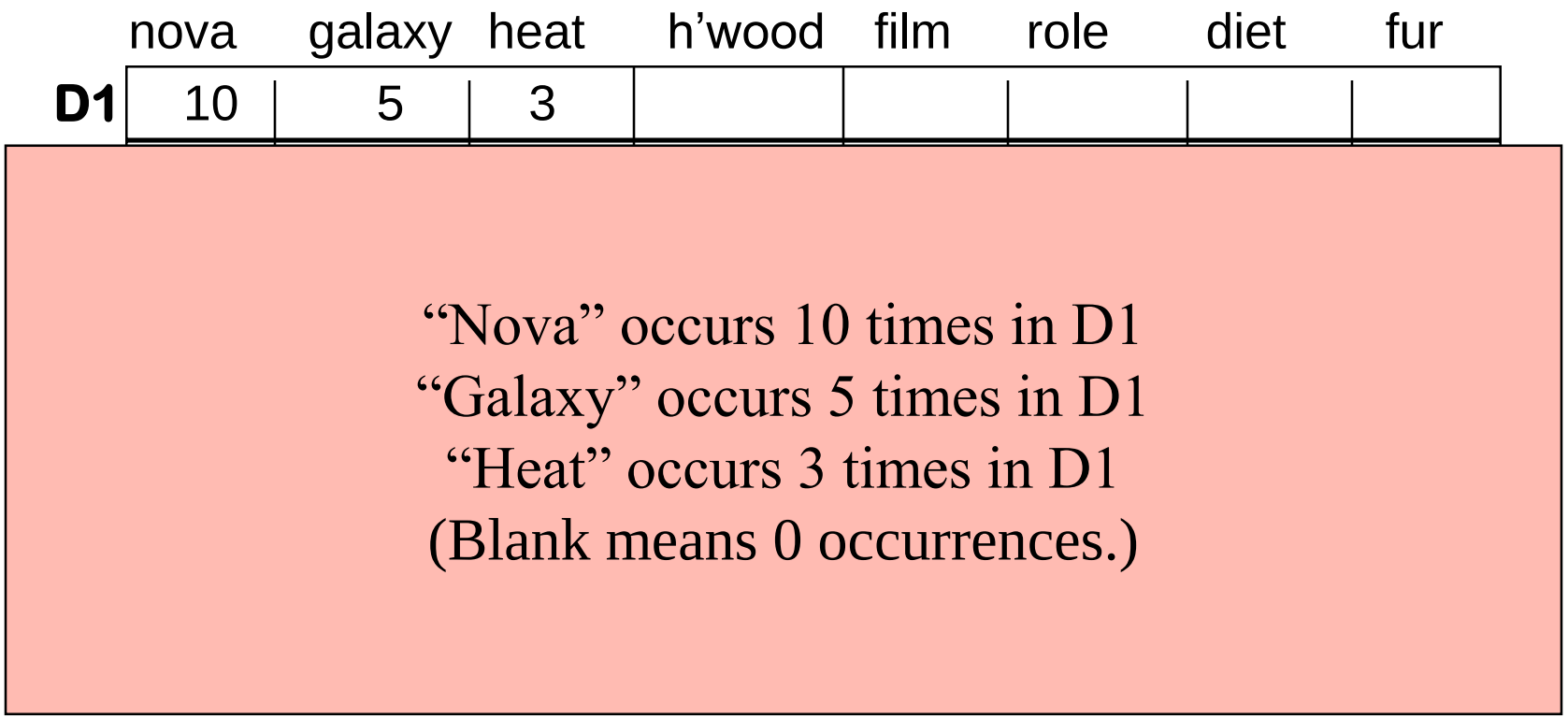
Vector Space Model

- Task:
  - Document collection
  - Query specifies information need: free text
  - Relevance judgments: 0/1 for all docs

Document Vectors

- Documents are represented as “bags of words”
- Represented as vectors when used computationally
  - A vector is like an array of floating point
  - Has direction and magnitude
  - Each vector holds a place for every term in the collection
  - Therefore, most vectors are sparse

Document Vectors:
One location for each word

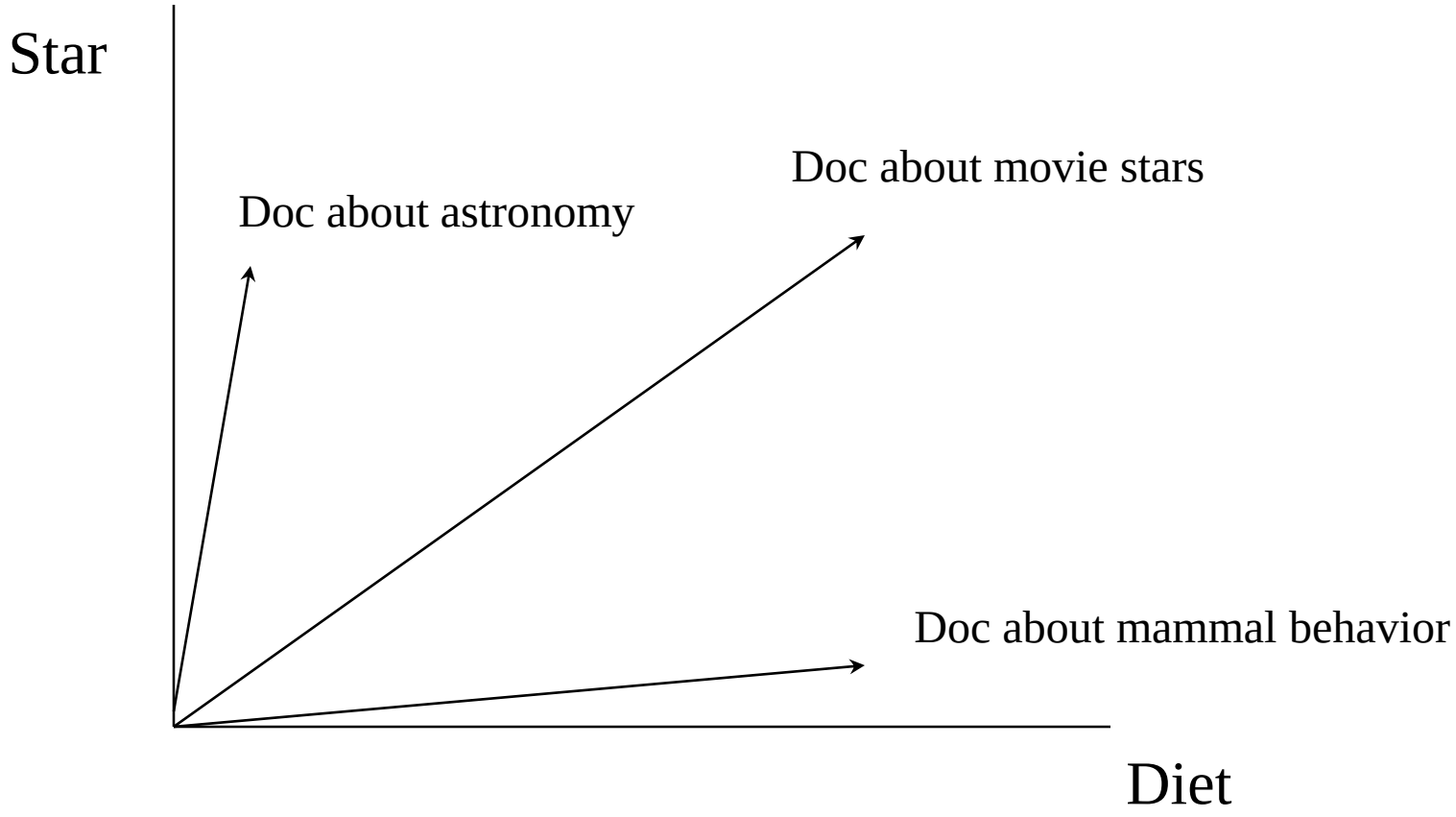


Document Vectors

Document ids

	nova	galaxy	heat	h'wood	film	role	diet	fur
D1	10	5	3					
D2	5	10						
D3				10	8	7		
D4				9	10	5		
D5							10	10
D6							9	10
D7	5		7			9		
D8		6	10	2	8			
D9				7	5		1	3

We Can Plot the Vectors



Assumption: Documents that are “close” in space are similar.

Vector Space Model

- Documents are represented as *vectors* in term space
  - Terms are usually stems
  - Documents represented by binary vectors of terms
- Queries represented the same as documents
- A vector distance measure between the query and documents is used to rank retrieved documents

Vector space model

- Vector space = contain all the keywords

  $<t_1, \ t_2, \ t_3, \ldots, \ t_n>$

- Document

  $D = \ < a_1, a_2, a_3, \ldots, a_n>$

  a_i = weight of t_i in D

- Query

  $Q = \ < b_1, b_2, b_3, \ldots, b_n>$

  b_i = weight of t_i in Q

- R(D,Q) = Sim(D,Q)

Matrix representation

Document space →

	t_1	t_2	t_3	…	t_n
D_1	a_{11}	a_{12}	a_{13}	…	a_{1n}
D_2	a_{21}	a_{22}	a_{23}	…	a_{2n}
D_3	a_{31}	a_{32}	a_{33}	…	a_{3n}
…					
D_m	a_{m1}	a_{m2}	a_{m3}	…	a_{mn}
Q	b_1	b_2	b_3	…	b_n

← Term vector space

Some formulas for Sim

Dot product $$Sim(D,Q) = \sum (a_i * b_i)$$

Cosine $$Sim(D,Q) = \frac{\sum_i (a_i * b_i)}{\sqrt{\sum_i a_i^2 * \sum_i b_i^2}}$$

Pair-wise Document Similarity

	nova	galaxy	heat	h'wood	film	role	diet	fur
A	1	3	1					
B	5	2						
C				2	1	5		
D				4	1			

How to compute document similarity?

Pair-wise Document Similarity

$$D_1 = w_{11}, w_{12,}, \ldots, w_{1t}$$

$$D_2 = w_{21}, w_{22,}, \ldots, w_{2t}$$

$$sim(D_1, D_2) = \sum_{i=1}^{t} w_{1i} * w_{2i}$$

$$sim(A, B) = (1*5) + (3*2) = 11$$

$$sim(A, C) = 0$$

$$sim(A, D) = 0$$

$$sim(B, C) = 0$$

$$sim(B, D) = 0$$

$$sim(C, D) = (2*4) + (1*1) = 9$$

	nova	galaxy	heat	h'wood	film	role	diet	fur
A	1	3	1					
B	5	2						
C				2	1	5		
D				4	1			

Pair-wise Document Similarity (cosine normalization)

$$D_1 = w_{11}, w_{12,}, \ldots, w_{1t}$$

$$D_2 = w_{21}, w_{22,}, \ldots, w_{2t}$$

$$sim(D_1, D_2) = \sum_{i=1}^{t} w_{1i} * w_{2i} \quad \text{unnormalized}$$

$$sim(D_1, D_2) = \frac{\sum_{i=1}^{t} w_{1i} * w_{2i}}{\sqrt{\sum_{i=1}^{t} (w_{1i})^2 * \sum_{i=1}^{t} (w_{2i})^2}} \quad \text{cosine normalized}$$

Vector Space "Relevance" Measure

$$D_i = w_{d_{i1}}, w_{d_{i2}}, ..., w_{d_{it}}$$

$$Q = w_{q1}, w_{q2,}, ..., w_{qt} \qquad w = 0 \text{ if a term is absent}$$

$$\text{if term weights normalized:} \quad sim(Q, D_i) = \sum_{j=1}^{t} w_{qj} * w_{d_{ij}}$$

otherwise normalize in the similarity comparison:

$$sim(Q, D_i) = \frac{\sum_{j=1}^{t} w_{qj} * w_{d_{ij}}}{\sqrt{\sum_{j=1}^{t} (w_{qj})^2 * \sum_{j=1}^{t} (w_{d_{ij}})^2}}$$

Computing Relevance Scores

Say we have query vector $Q = (0.4, 0.8)$

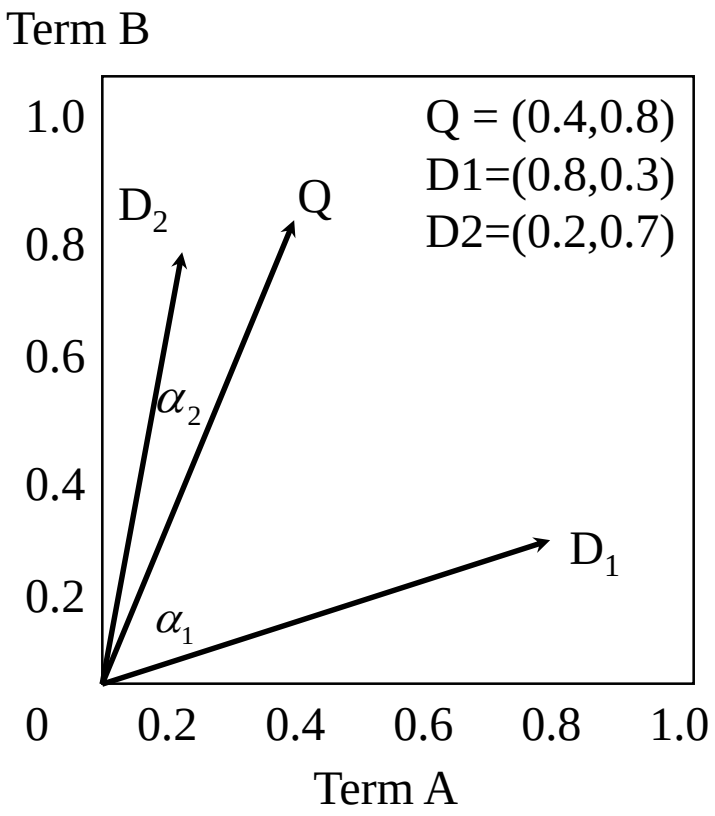
Also, document $D_2 = (0.2, 0.7)$

What does their similarity comparison yield?

$$sim(Q, D_2) = \frac{(0.4 * 0.2) + (0.8 * 0.7)}{\sqrt{[(0.4)^2 + (0.8)^2] * [(0.2)^2 + (0.7)^2]}}$$

$$= \frac{0.64}{\sqrt{0.42}} = 0.98$$

Vector Space with Term Weights and Cosine Matching



$$D_i=(d_{i1},w_{di1};d_{i2},\ w_{di2};\ldots;d_{it},\ w_{dit})$$
$$Q=(q_{i1},w_{qi1};q_{i2},\ w_{qi2};\ldots;q_{it},\ w_{qit})$$

$$sim(Q,D_i)=\frac{\sum_{j=1}^{t} w_{q_j} w_{d_{ij}}}{\sqrt{\sum_{j=1}^{t}(w_{q_j})^2 \sum_{j=1}^{t}(w_{d_{ij}})^2}}$$

$$sim(Q,D2)=\frac{(0.4\cdot 0.2)+(0.8\cdot 0.7)}{\sqrt{[(0.4)^2+(0.8)^2]\cdot[(0.2)^2+(0.7)^2]}}$$

$$=\frac{0.64}{\sqrt{0.42}}=0.98$$

$$sim(Q,D_1)=\frac{.56}{\sqrt{0.58}}=0.74$$

Advantage of VSM

- Simplicity
  - Ability to incorporate term weights

- Can measure similarities between almost anything:
  - documents and queries
  - documents and documents
  - queries and queries

Disadvantages of VSM

- Assumed independence relationship among terms
  - Though this is a very common retrieval model assumption

- Lack of justification for some vector operations
  - Selection of similarity function, term weights

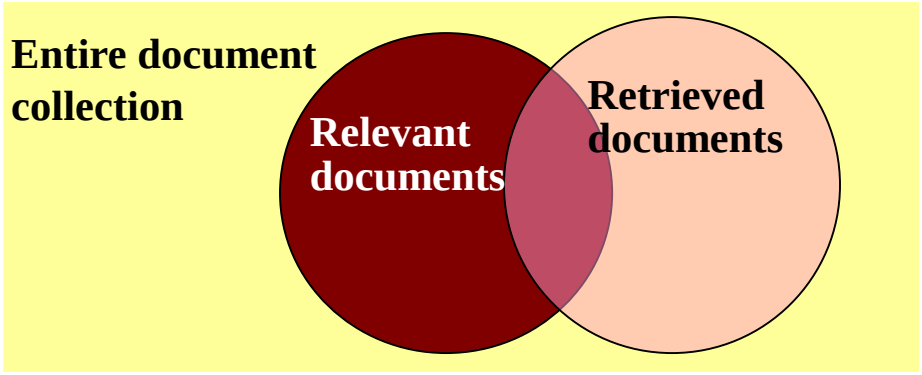
- Assumes a query and a document can be treated the same (symmetric)

In Class activity

System evaluation criteria

- Efficiency: time, space
- Effectiveness:
  - How is a system capable of retrieving relevant documents?
  - Is a system better than another one?
- Metrics often used (together):
  - Precision = retrieved relevant docs / retrieved docs
  - Recall = retrieved relevant docs / relevant docs

Evaluation Measures -Precision and Recall



	retrieved	not retrieved
irrelevant	retrieved & irrelevant	Not retrieved & irrelevant
relevant	retrieved & relevant	not retrieved but relevant

$$recall = \frac{Number\ of\ relevant documents retrieved}{Total\ number of\ relevant documents}$$

$$precision = \frac{Number\ of\ relevant documents\ retrieved}{Total\ number of\ documents\ retrieved}$$

Mean Average Precision?

Some techniques to improve IR effectiveness

- Interaction with user (relevance feedback)

  - Keywords only cover part of the contents

  - User can help by indicating relevant/irrelevant document

- Users usually do not cooperate (e.g. AltaVista in early years)
  - Pseudo-relevance feedback (Blind RF)
  - Using the top-ranked documents as if they are relevant.

Revisited Learning Objectives

- Be able to define the information retrieval
- Be able to understand the difference between IR and DBMS
- Be able to differentiate between document selection and document ranking
- Be able to represent the document and query using vectors.
- Be able to compute the similarity between documents and the query
- Be able to comment on the efficiency if the IRS based in recall and precision values.
- Be able to briefly explain the techniques to improve IR efficiency