


## More on Information Retrieval

Week 12


COS60009: Data Management for the Big Data Age



1

## Learning Objectives

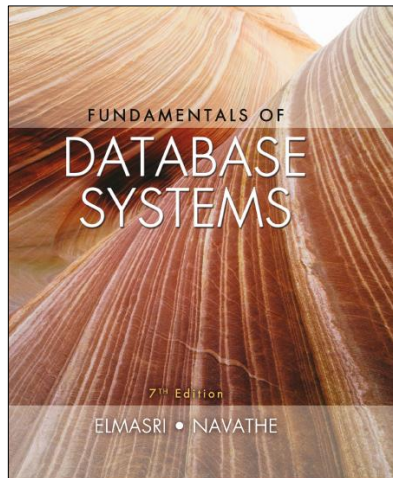
- Evaluation Measures of Search Relevance
- Web Search and Analysis
- Web ranking
- Trends in IR



2

# Fundamentals of Database Systems

Seventh Edition



## Chapter 27

Introduction to  
Information Retrieval and  
Web Search



Copyright © 2016, 2011, 2007 Pearson Education, Inc. All Rights Reserved

3

## Evaluation Measures of Search Relevance

(1 of 4)

- Topical relevance
  - Measures result topic match to query topic
- User relevance
  - Describes 'goodness' of retrieved result with regard to user's information need
- Web information retrieval
  - No binary classification made for relevance or nonrelevance
  - Ranking of documents



Copyright © 2016, 2011, 2007 Pearson Education, Inc. All Rights Reserved

4

## Evaluation Measures of Search Relevance

(2 of 4)

- Recall
  - Number of relevant documents retrieved by a search divided by the total number of actually relevant documents existing in the database
- Precision
  - Number of relevant documents retrieved by a search divided by total number of documents retrieved by that search

5

## Retrieved Versus Relevant Search Results

- TP: true positive
- FP: false positive
- TN: true negative
- FN: false negative





		Relevant?	
		Yes	No
Retrieved?	Yes	 Hits TP	 False Alarms FP
	No	Misses FN 	Correct Rejections TN 

Figure 27.5 Retrieved versus relevant search results

6

## Evaluation Measures of Search Relevance

(3 of 4)

- Recall can be increased by presenting more results to the user
  - May decrease the precision

Doc. No.	Rank Position $i$	Relevant	Precision( $i$ )	Recall( $i$ )
10	1	Yes	$1/1 = 100\%$	$1/10 = 10\%$
2	2	Yes	$2/2 = 100\%$	$2/10 = 20\%$
3	3	Yes	$3/3 = 100\%$	$3/10 = 30\%$
5	4	No	$3/4 = 75\%$	$3/10 = 30\%$
17	5	No	$3/5 = 60\%$	$3/10 = 30\%$
34	6	No	$3/6 = 50\%$	$3/10 = 30\%$
215	7	Yes	$4/7 = 57.1\%$	$4/10 = 40\%$
33	8	Yes	$5/8 = 62.5\%$	$5/10 = 50\%$
45	9	No	$5/9 = 55.5\%$	$5/10 = 50\%$
16	10	Yes	$6/10 = 60\%$	$6/10 = 60\%$

Table 27.2 Precision and recall for ranked retrieval



Copyright © 2016, 2011, 2007 Pearson Education, Inc. All Rights Reserved

7

## Evaluation Measures of Search Relevance

(4 of 4)

- Average precision
  - Computed based on the precision at each relevant document in the ranking
- Recall/precision curve
  - Based on the recall and precision values at each rank position
    - x-axis is recall and y-axis is precision
- F-score
  - Harmonic mean of the precision ( $p$ ) and recall ( $r$ ) values



Copyright © 2016, 2011, 2007 Pearson Education, Inc. All Rights Reserved

8

## Web Search and Analysis (1 of 6)

- Search engines must crawl and index Web sites and document collections
  - Regularly update indexes
  - Link analysis used to identify page importance
- Vertical search engines
  - Customized topic-specific search engines that crawl and index a specific collection of documents on the Web

## Web Search and Analysis (2 of 6)

- Metasearch engines
  - Query different search engines simultaneously and aggregate information
- Digital libraries
  - Collections of electronic resources and services for the delivery of materials in a variety of formats
- Web analysis
  - Applies data analysis techniques to discover and analyze useful information from the Web

## Web Search and Analysis (3 of 6)

- Goals of Web analysis
  - Finding relevant information
  - Personalization of the information
  - Finding information of social value
- Categories of Web analysis
  - Web structure analysis
  - Web content analysis
  - Web usage analysis



Copyright © 2016, 2011, 2007 Pearson Education, Inc. All Rights Reserved

11

## Web Search and Analysis (4 of 6)

- Web structure analysis
  - Hyperlink
  - Destination page
  - Anchor text
  - Hub
  - Authority
- PageRank ranking algorithm
  - Used by Google
  - Analyzes forward/backward links: highly linked pages are more important



Copyright © 2016, 2011, 2007 Pearson Education, Inc. All Rights Reserved

12

## Web Search and Analysis (5 of 6)

- Web content analysis tasks
  - Structured data extraction
    - Wrapper
  - Web information integration
    - Web query interface integration
    - Schema matching
    - Ontology-based information integration
  - Building concept hierarchies
  - Segmenting web pages and detecting noise



Copyright © 2016, 2011, 2007 Pearson Education, Inc. All Rights Reserved

13

## Web Search and Analysis (6 of 6)

- Web usage analysis attempts to discover usage patterns from Web data
  - Preprocessing
    - Usage, content, structure
  - Pattern discovery
    - Statistical analysis, association rules, clustering, classification, sequential patterns, dependency modeling
  - Pattern analysis
    - Filter out patterns not of interest



Copyright © 2016, 2011, 2007 Pearson Education, Inc. All Rights Reserved

14

## Trends in Information Retrieval (1 of 3)

- Faceted search
  - Classifying content
- Social search
  - Collaborative social search
- Conversational information access
  - Intelligent agents perform intent extraction to provide information relevant to a conversation



Copyright © 2016, 2011, 2007 Pearson Education, Inc. All Rights Reserved

15

## Trends in Information Retrieval (2 of 3)

- Probabilistic topic modeling
  - Automatically organize large collections of documents into relevant themes
- Question-answering systems
  - Factoid questions
  - List questions
  - Definition questions
  - Opinion questions
  - Composed of question analysis, query generation, search, candidate answer generation, and answer scoring



Copyright © 2016, 2011, 2007 Pearson Education, Inc. All Rights Reserved

16



## Trends in Information Retrieval (3 of 3)

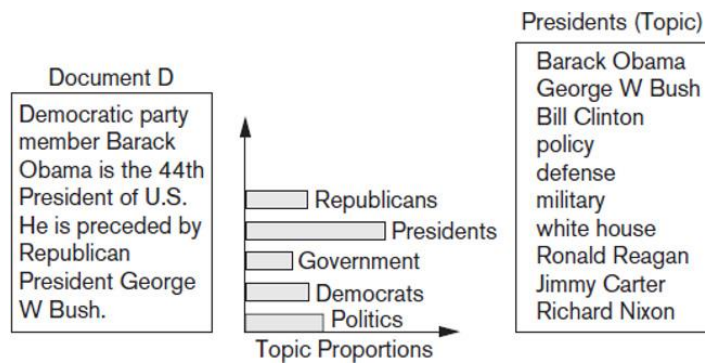


Figure 27.6 A document D and its topic proportions

## Summary

- Information retrieval mainly targeted at unstructured data
- Query and browsing modes of interaction
- Retrieval models
  - Boolean, vector space, probabilistic, and semantic
- Text preprocessing
- Web search
- Web ranking
- Trends

## Copyright



This work is protected by United States copyright laws and is provided solely for the use of instructors in teaching their courses and assessing student learning. Dissemination or sale of any part of this work (including on the World Wide Web) will destroy the integrity of the work and is not permitted. The work and materials from it should never be made available to students except by instructors using the accompanying text in their classes. All recipients of this work are expected to abide by these restrictions and to honor the intended pedagogical purposes and the needs of other instructors who rely on these materials.