

Course Overview (II)

(Search and Data Mining)

Hui Fang
Department of Electrical and Computer Engineering
University of Delaware

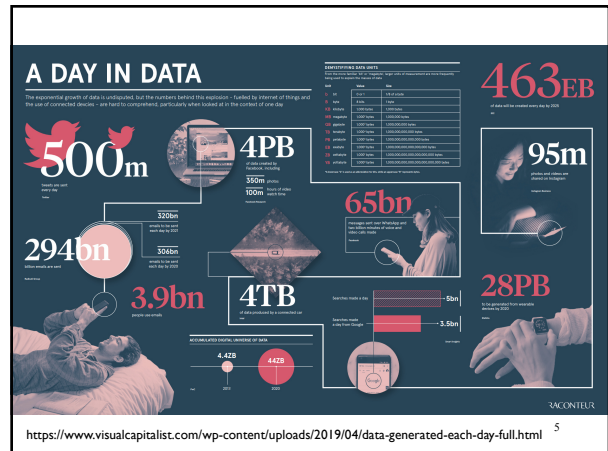
1

Why are you taking this course?

2

Information Overload

4



5

2019 This Is What Happens In An Internet Minute



6



We need new tools to help us organize, access and understand these huge amounts of information.

How to help users manage and exploit all the information?

Search engines help us access useful information effectively

Google™

Web Images Video News Maps more...

Google Search I'm Feeling Lucky



5 billion searches per day

What can a search engine do?

- Google's mission:
 - To organize the world's information and make it universally accessible and useful

9

Search + Mining

- Search is also known as Information Retrieval (IR)
- Challenges of Information Management
 - How to organize information automatically?
 - How to find useful information?
 - How to discover knowledge and extract patterns?



<https://trec.nist.gov>

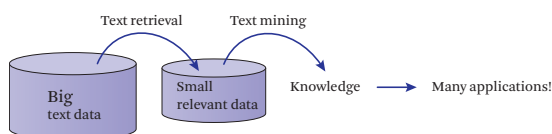
10

Text Information is crucial!

- The most natural way of encoding knowledge
- The most common type of information
- The most expressive form of information

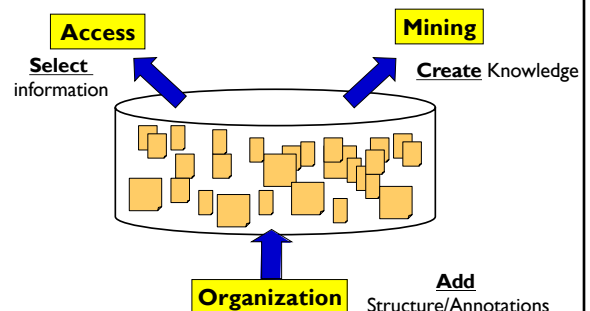
11

Text retrieval and text mining are two main techniques for analyzing big text data.



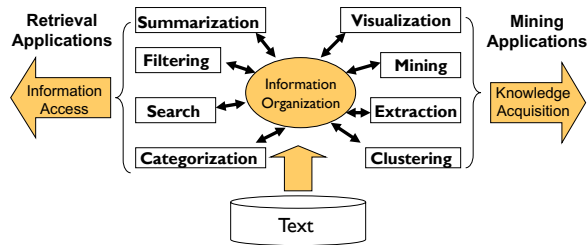
12

Text Management Applications



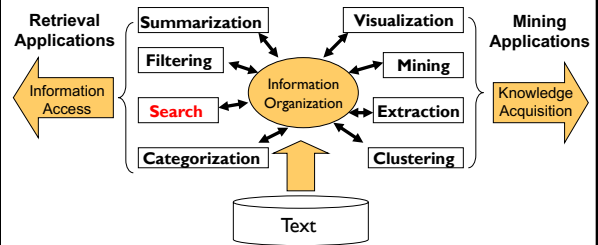
13

Elements of Text Info Management Technologies



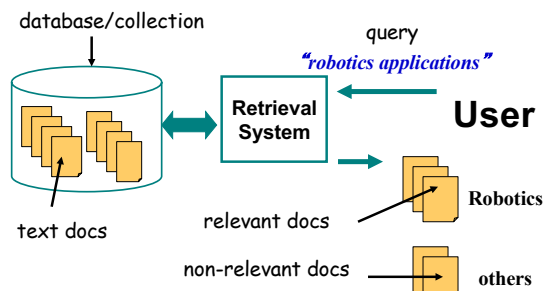
14

Elements of Text Info Management Technologies



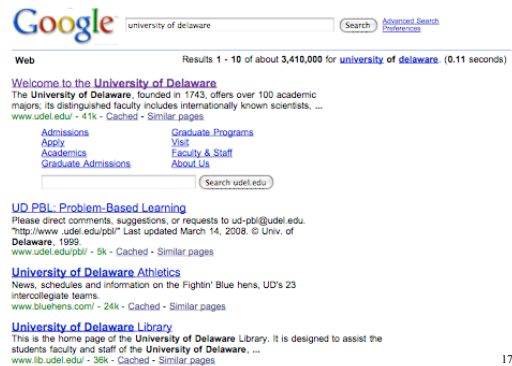
15

Search (Text Retrieval)



16

Web Search Engines



17

Which one is your favorite? Why?

Google™

Web Images Video News Maps more »

Google Search I'm Feeling Lucky



5 billion searches per day

18

How do search engines make money?

23

Web Search Engines

The screenshot shows a Google search results page for the query 'university of delaware'. The page includes the Google logo, a search bar with the query, and a search button. Below the search bar, it shows 'Results 1 - 10 of about 3,410,000 for university of delaware. (0.11 seconds)'. The results are divided into 'Web' and 'Sponsored Links'. The 'Web' section includes links to the University of Delaware website, Admissions, Graduate Programs, and UD PBL. The 'Sponsored Links' section includes links to the University of Delaware website, University of Delaware Gifts, and University of Delaware Library. A red box highlights the 'Sponsored search results' section.

Organic search results

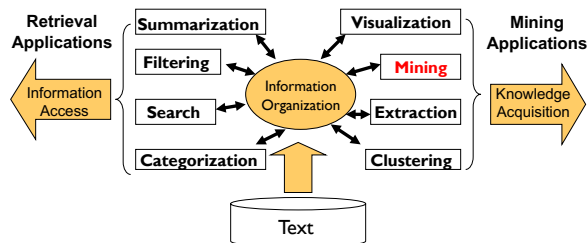
24

Is this a perfect solution?

Have you noticed any new features of search engines?

25

Elements of Text Info Management Technologies



The Database of Intentions

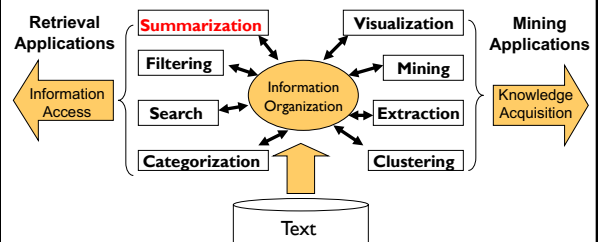
[John Battelle, 2004]

- The aggregate results of
 - every search ever entered
 - every result list ever tendered
 - ...
- This information represents, in aggregate form, a place holder for the intentions of humankind
- Tell us who we are and what we want as a culture

Top searches of 2019

- Disney Plus
- Cameron Boyce
- Nipsey Hussle
- Hurricane Dorain
- Antonio Brown

Elements of Text Info Management Technologies



Text Summarization

- **Motivation – Information overload**
 - 4 Billion URLs indexed by Google
 - 200 TB of data on the Web [Lyman and Varian 03]
 - Information is created every day in enormous amount
- **Goal of text summarization**
 - take an information source, extract the most important content from it and present it to the user in a condensed form and in a manner sensitive to the user's needs.

Search Result Summarization

Google computer engineering

Web Results 1 - 10 of about 68,500,000 for computer engineering

Computer engineering - Wikipedia, the free encyclopedia
 Computer engineering (also called Electronic and Computer engineering or Computer Systems Engineering) is a discipline that combines elements of both ...
[en.wikipedia.org/wiki/Computer_engineering](#) - 59k - Cached - Similar pages -

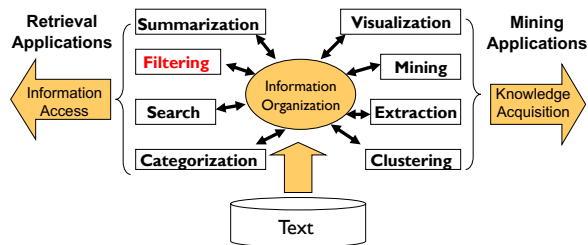
Computer Software Engineers
 Excellent job prospects are expected for applicants with at least bachelor's degree in computer engineering or computer science and with practical work ...
[www.bls.gov/ocoocoos267.htm](#) - 55k - Cached - Similar pages -

Electrical and Computer Engineering at Carnegie Mellon University
 The Department of Electrical and Computer Engineering at Carnegie Mellon University is recognized worldwide for its undergraduate and graduate programs. ...
[www.ece.cmu.edu/](#) - 12k - Cached - Similar pages -

Computer Engineering
 Various test prep options available, pick one based on your own learning style: online, classroom, private tutoring.
[www.priestonreview.com/Maple.aspx?mpid=70](#) - 96k - Cached - Similar pages -

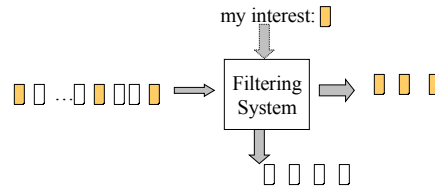
Summaries of search results

Elements of Text Info Management Technologies



Information Filtering

- Stable & long term interest, dynamic info source
- System must make a delivery decision immediately as a document “arrives”



Information Filtering

Google Alerts

Monitor the web for interesting new content

Create an alert about...

Search vs. Filtering

- **Short-term information need (Search)**
 - “Temporary need”, e.g., info about used cars
 - Information source is relatively static
 - User “pulls” information
 - Application example: library search, Web search
- **Long-term information need (Filtering)**
 - “Stable need”, e.g., new data mining algorithms
 - Information source is dynamic
 - System “pushes” information to user
 - Applications: news filter

Collaborative Filtering

The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture (BARGAIN PRICE) (Hardcover)
by John Battelle (Author) "By the fall of 2003, the Internet industry was in full retreat..."

Key Feature: Bargain Price: \$7.99 (49% off) (List Price: \$15.99)
You Save: \$8.00 (50%)

In Stock.
Ship from and sold by Amazon.com. Gift-wrap available.

Want it delivered Thursday, January 24? Order it in the next 1 hour and 12 minutes, and choose **One-Day Shipping** at checkout. [Details](#)

Customers Who Bought This Item Also Bought

Page 1 of 17

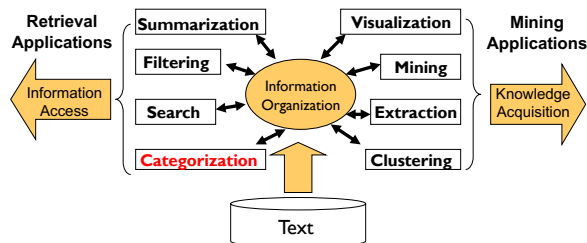
• Netflix Prize

- ☹
- Content began 2006
- Qualify for the \$1M grand prize if new algorithms can be at least 10% better than the baseline method.
- ☹
- The prize was awarded in 2009.

• Many competitions

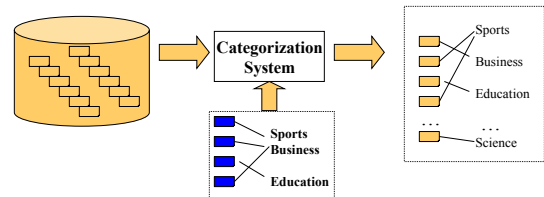
- Kaggle
- TREC
- WSDM/KDD cup (spotify, etc.)

Elements of Text Info Management Technologies



What is Text Categorization?

- Pre-given categories and labeled document examples (Categories may form hierarchy)
- Classify new documents
- A standard supervised learning problem



Text Categorization

Google News

Search and browse 4,500 news sources updated continuously.

[News archive search](#) | [Advanced news search](#) | [Blog search](#)

Top Stories

- World
- U.S.
- Business
- Sci/Tech
- Entertainment
- Sports
- Health
- Most Popular

News categories

Sci/Tech

DTV delay...delayed? House votes down DTV postponement
Ars Technica - 38 minutes ago
The DTV transition may go ahead as scheduled on February 17 after the House unexpectedly rejected a Senate bill to delay it until June 12.

Video: Delay Coming to Digital TV Switch
Associated Press
House fails to pass DTV delay bill Reuters

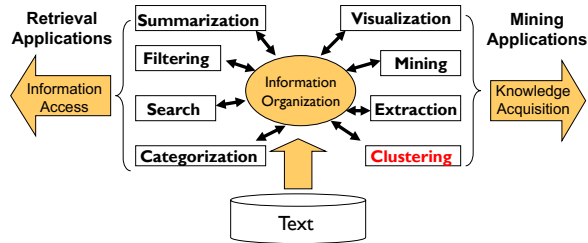
The Associated Press - B3journalism.com - TG Daily - New York Times
all 1,853 news articles x

AT&T Revenues Influenced By Strong iPhone Sales
InformationWeek - 2 hours ago

Text Categorization (Cont.)

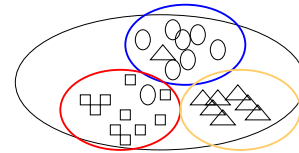
- Email Spam Filters
 - Classify emails to spam and non-spam
- Sentiment Analysis
 - Classify user reviews to positive and negative

Elements of Text Info Management Technologies



The Clustering Problem

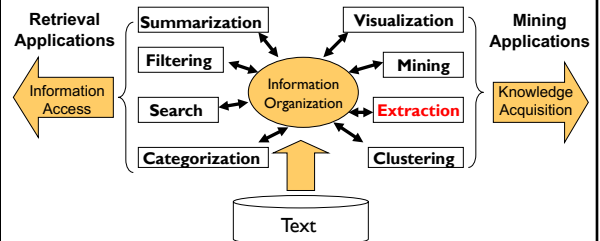
- Discover “natural structure”
- Group similar objects together
- Object can be document, term, passages
- Example



Clustering News Articles

The screenshot shows Google News search results. The top result is 'Obama Says Not a Moment to Spare' on Stimulus Plan, with a link to 'all 8,224 news articles'. Below it is a result about 'Distraught Wilmington dad Ervin Lupoe likely shot wife Ana and...', with a link to 'all 1,539 news articles'. The results are from various sources like New York Times, Guardian, and BBC News.

Elements of Text Info Management Technologies



What is Information Extraction?

- Recovering structured data from formatted text

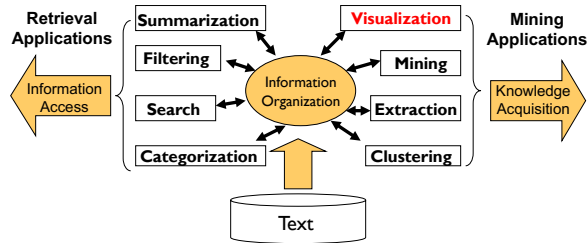
Subject: faculty meeting
Date: January 15, 2020
To: Hui Fang

Event: faculty mtg
Date: Jan-16-2020
Start: 10:00am
End: 11:30am
Where: Evans 204

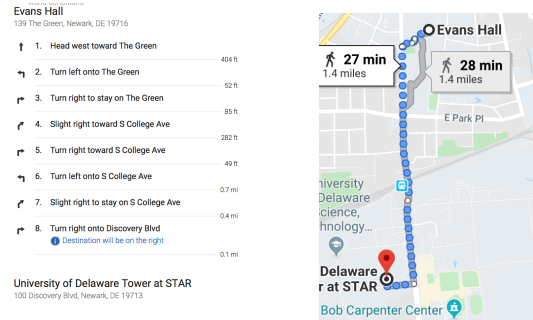
Hi Hui, we've now scheduled the faculty meeting.
It will be in Evans 204 tomorrow from 10:00-11:30.

Create new Calendar entry

Elements of Text Info Management Technologies



Information Visualization



Many Unsolved Challenges

- New types of textual information
 - Messages, EHR, ...
- New contextual information
 - Social networks, instagram, ...
- New Information needs
 - People search, ...
- New Concerns
 - Privacy
 - Security,
 - ...

SOCIAL MEDIA EXPLAINED

