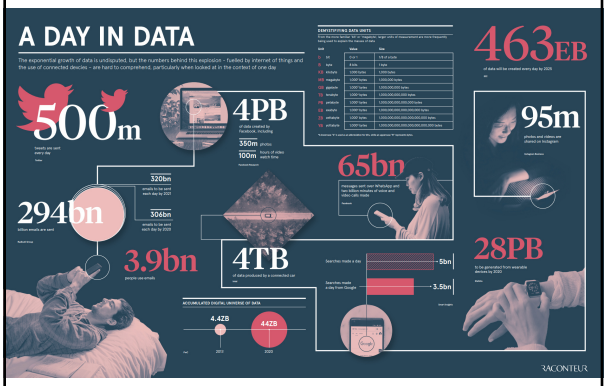
Course Overview (II)

(Search and Data Mining)

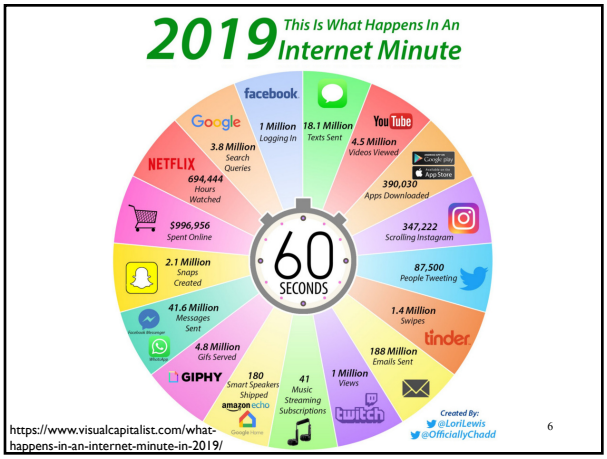
Hui Fang
Department of Electrical and Computer Engineering
University of Delaware

Why are you taking this course?

Information Overload



https://www.visualcapitalist.com/wp-content/uploads/2019/04/data-generated-each-day-full.html



https://www.visualcapitalist.com/what-happens-in-an-internet-minute-in-2019/

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

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

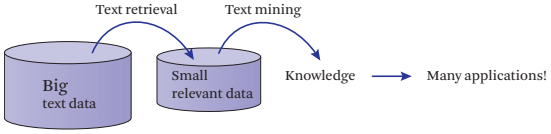
10

Text Information is crucial!

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

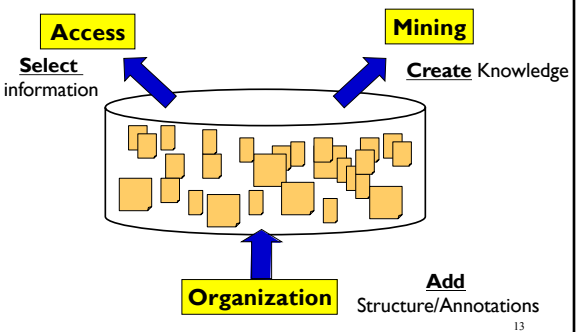
11

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



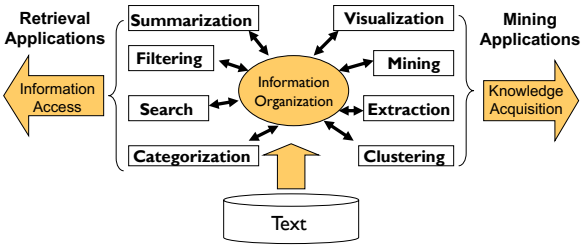
12

Text Management Applications



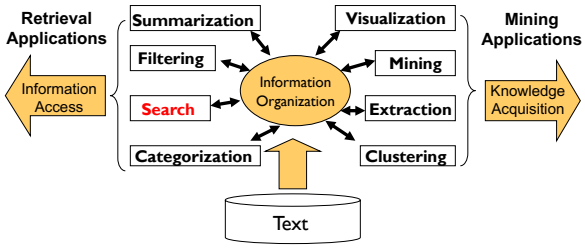
13

Elements of Text Info Management Technologies



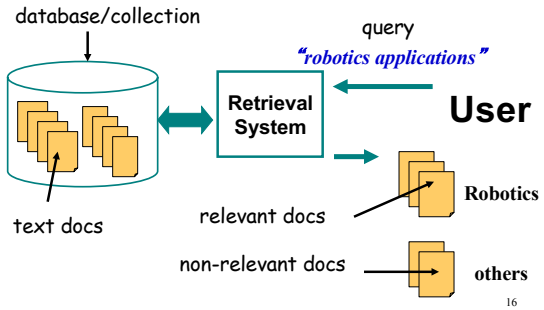
14

Elements of Text Info Management Technologies



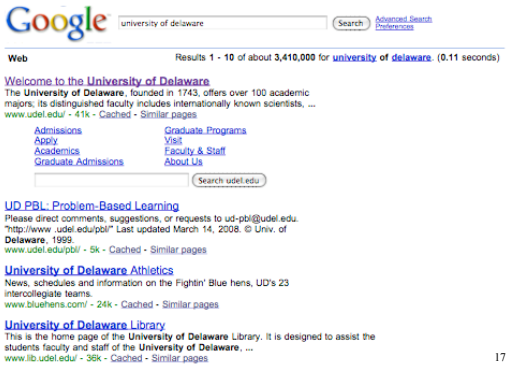
15

Search (Text Retrieval)



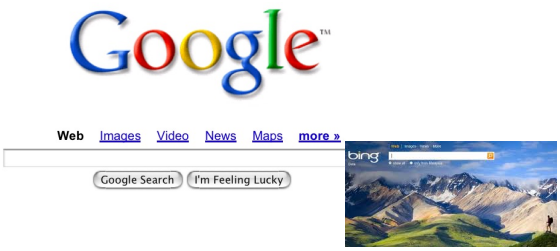
16

Web Search Engines



17

Which one is your favorite? Why?



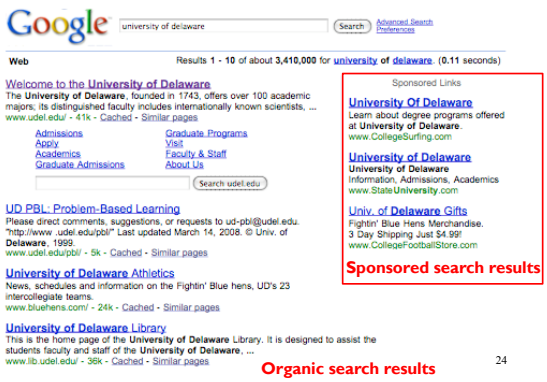
5 billion searches per day

18

How do search engines make money?

23

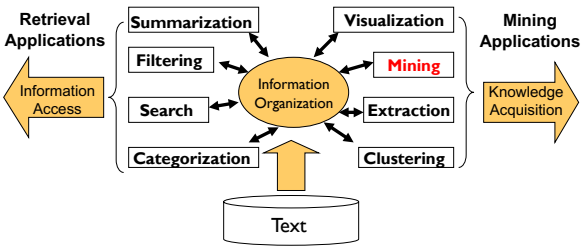
Web Search Engines



Is this a perfect solution?

Have you noticed any new features of search engines?

Elements of Text Info Management Technologies



The Database of Intentions

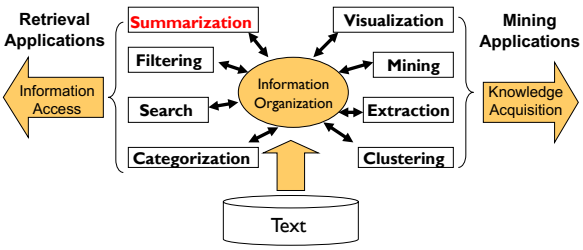
[John Battelle,2004]

- 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

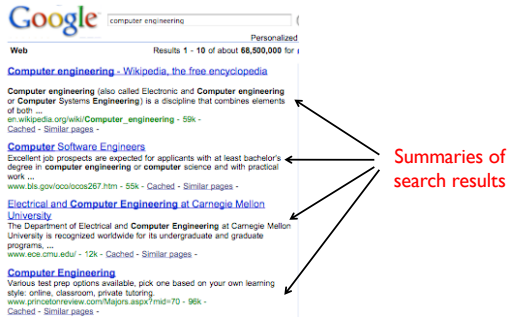
Elements of Text Info Management Technologies



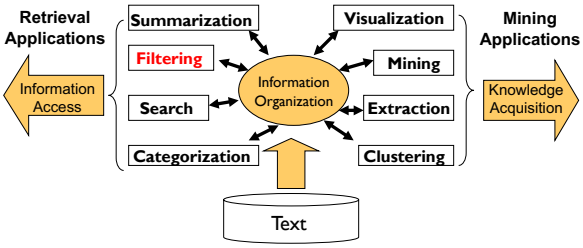
Text Summarization

- 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

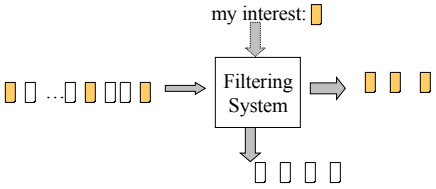


Elements of Text Info Management Technologies

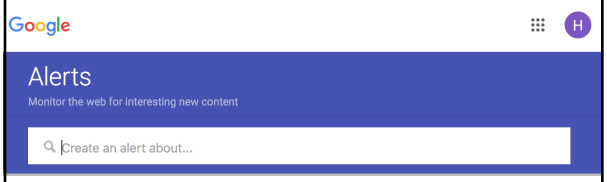


Information Filtering

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



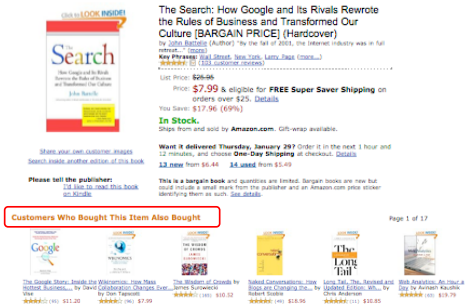
Information Filtering



Search vs. Filtering

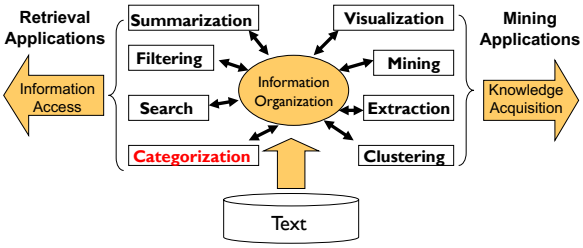
- Short-term information need (Search)
  - "Temporary need", e.g., info about used cars
  - Information source is relatively static
  - User "pulls" information
  - Application example: library search, Web search
- Long-term information need (Filtering)
  - "Stable need", e.g., new data mining algorithms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

Collaborative Filtering



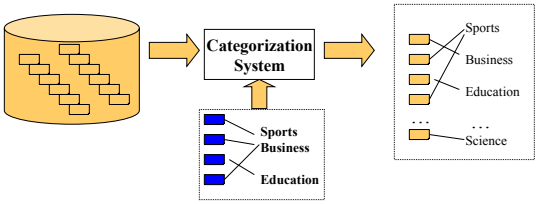
- **Netflix Prize**
  - ☺
    - **Content began 2006**
    - **Qualify for the $1M grand prize if new algorithms can be at least 10% better than the baseline method.**
  - ☹
    - **The prize was awarded in 2009.**
- **Many competitions**
  - **Kaggle**
  - **TREC**
  - **WSDM/KDD cup (spotify, etc.)**

Elements of Text Info Management Technologies

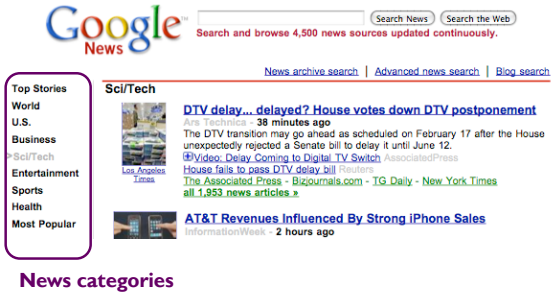


What is Text Categorization?

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



Text Categorization

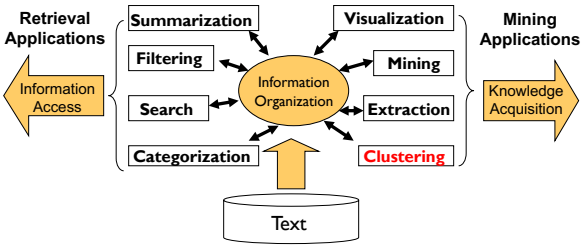


News categories

Text Categorization (Cont.)

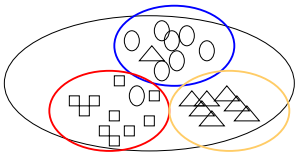
- **Email Spam Filters**
  - **Classify emails to spam and non-spam**
- **Sentiment Analysis**
  - **Classify user reviews to positive and negative**

Elements of Text Info Management Technologies

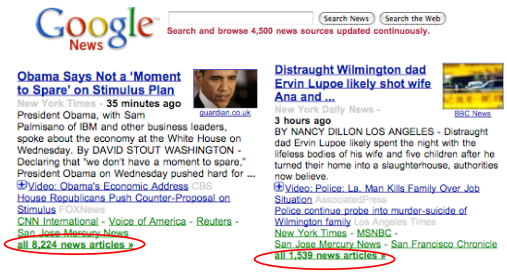


The Clustering Problem

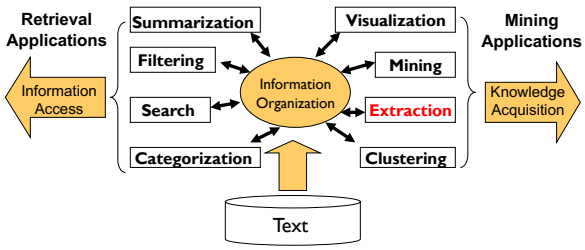
- Discover "natural structure"
- Group similar objects together
- Object can be document, term, passages
- Example



Clustering News Articles



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**

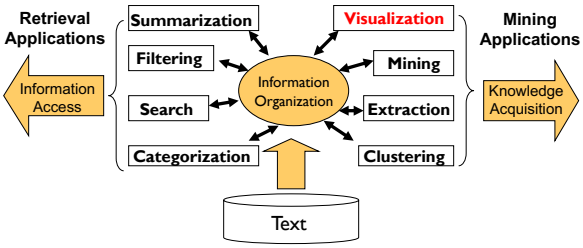
Event: faculty mtg
Date: Jan-16-2020
Start: 10:00am
End: 11:30am
Where: Evans 204

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

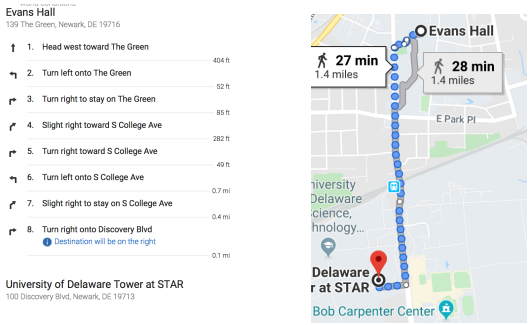
Create new Calendar entry

47

Elements of Text Info Management Technologies

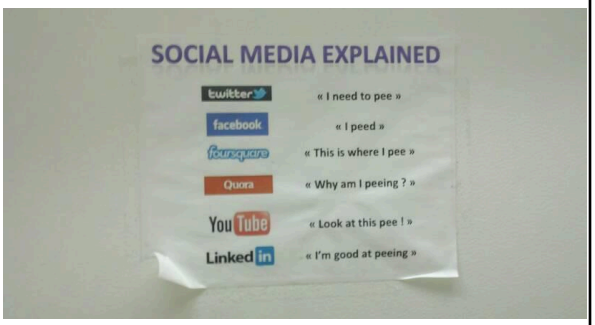


Information Visualization



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,
  - ….



51