

Web Information Retrieval

Lecture 13

Introduction to text classification and
clustering

Today's lecture

- Introduction to Text Classification
 - Also widely known as “text categorization”
- Introduction to Clustering

Text Classification

- Broadly the problem of **text classification** is to classify a set of documents as:
 - Spam / not spam
 - Topic (about art, health, etc.)
 - Language
 - Porn / not porn
 - ...
- The notion of **classification** is very general and has many applications within and beyond IR

Spam filtering: Another text classification task

From: "" <takworld@hotmail.com>

Subject: real estate is the only way... gem oalvgkay

Anyone can buy real estate with no money down

Stop paying rent TODAY !

There is no need to spend hundreds or even thousands for similar courses

I am 22 years old and I have already purchased 6 properties using the methods outlined in this truly INCREDIBLE ebook.

Change your life NOW !

=====

Click Below to order:

<http://www.wholesaledaily.com/sales/nmd.htm>

=====

Standing queries

- The path from IR to text classification:
 - You have an information need to monitor, say:
 - **Unrest in the Niger delta region**
 - You want to rerun an appropriate query periodically to find new news items on this topic
 - You will be sent new documents that are found
 - I.e., it's text classification, not ranking
- Such queries are called ***standing queries***
 - Long used by “information professionals”
 - A modern mass instantiation is **Google Alerts**
- Standing queries are (hand-written) text classifiers

Supervised Classification

- Given:
 - A description of an instance, $d \in X$
 - X is the *instance language* or *instance space*.
 - A fixed set of classes:
 $C = \{c_1, c_2, \dots, c_J\}$
 - A training set D of labeled documents with each labeled document $(d, c) \in X \times C$
- Determine:
 - A learning method or algorithm which will enable us to learn a classifier $\gamma: X \rightarrow C$
 - For a test document d , we assign it the class $\gamma(d) \in C$

Categorization/Classification

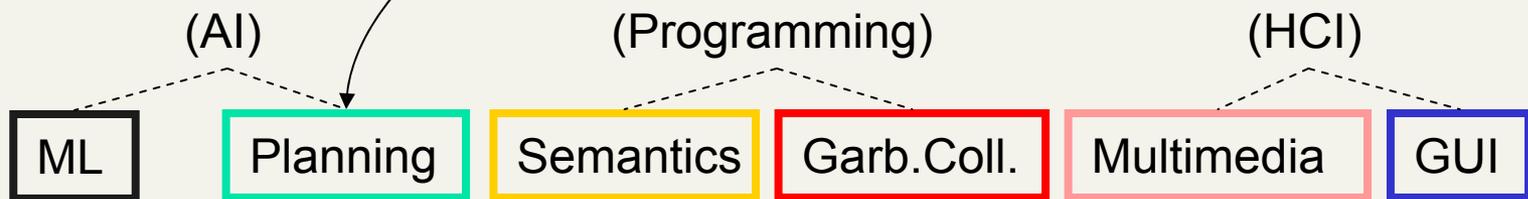
- Given:
 - A description of an instance, $d \in X$
 - X is the *instance language* or *instance space*.
 - Issue: how to represent text documents.
 - Usually some type of high-dimensional space
 - A fixed set of classes:
$$C = \{c_1, c_2, \dots, c_J\}$$
- Determine:
 - The category of d : $\gamma(d) \in C$, where $\gamma(d)$ is a *classification function* whose domain is X and whose range is C .
 - We want to know how to build classification functions (“classifiers”).

Document Classification

**Test
Data:**

“planning
language
proof
intelligence”

Classes:



**Training
Data:**

learning	<u>planning</u>	programming	garbage
<u>intelligence</u>	temporal	semantics	collection		
algorithm	reasoning	<u>language</u>	memory		
reinforcement	plan	<u>proof...</u>	optimization		
network...	<u>language...</u>		region...		

(Note: in real life there is often a hierarchy, not present in the above problem statement)

More Text Classification Examples

Many search engine functionalities use classification

Assigning labels to documents or web-pages:

- Labels are most often topics such as Yahoo-categories
 - *"finance," "sports," "news>world>asia>business"*
- Labels may be genres
 - *"editorials" "movie-reviews" "news"*
- Labels may be opinion on a person/product
 - *"like", "hate", "neutral"*
- Labels may be domain-specific
 - *"interesting-to-me" : "not-interesting-to-me"*
 - *"contains adult language" : "doesn't"*
 - *language identification: English, French, Chinese, ...*
 - *search vertical: about Linux versus not*
 - *"link spam" : "not link spam"*

Classification Methods (1)

- Manual classification
 - Used by the original Yahoo! Directory
 - ODP , Looksmart, about.com, PubMed
 - Very accurate when job is done by experts
 - Consistent when the problem size and team is small
 - Difficult and expensive to scale
 - Means we need automatic classification methods for big problems

Classification Methods (2)

- Automatic document classification
 - Hand-coded rule-based systems
 - One technique used by spam filters, Reuters, CIA, etc.
 - It's what Google Alerts is doing
 - Widely deployed in government and enterprise
 - E.g., assign category if document contains a given boolean combination of words
 - Standing queries: Commercial systems have complex query languages (everything in IR query languages +score accumulators)
 - Accuracy is often very high if a rule has been carefully refined over time by a subject expert
 - Building and maintaining these rules is expensive

A Verity topic

A complex classification rule

```

comment line      # Beginning of art topic definition
top-level topic  art ACCRUE
                  /author = "fsmith"
topic definition modifiers {
                  /date  = "30-Dec-01"
                  /annotation = "Topic created
                               by fsmith"
subtopic topic    * 0.70 performing-arts ACCRUE
evidencetopic    ** 0.50 WORD
topic definition modifier /wordtext = ballet
evidencetopic    ** 0.50 STEM
topic definition modifier /wordtext = dance
evidencetopic    ** 0.50 WORD
topic definition modifier /wordtext = opera
evidencetopic    ** 0.30 WORD
topic definition modifier /wordtext = symphony
subtopic         * 0.70 visual-arts ACCRUE
                  ** 0.50 WORD
                  /wordtext = painting
                  ** 0.50 WORD
                  /wordtext = sculpture
subtopic         * 0.70 film ACCRUE
                  ** 0.50 STEM
                  /wordtext = film
subtopic         ** 0.50 motion-picture PHRASE
                  *** 1.00 WORD
                  /wordtext = motion
                  *** 1.00 WORD
                  /wordtext = picture
                  ** 0.50 STEM
                  /wordtext = movie
subtopic         * 0.50 video ACCRUE
                  ** 0.50 STEM
                  /wordtext = video
                  ** 0.50 STEM
                  /wordtext = vcr
# End of art topic

```

■ Note:

- maintenance issues (author, etc.)
- Hand-weighting of terms

[Verity was bought by
Autonomy.]

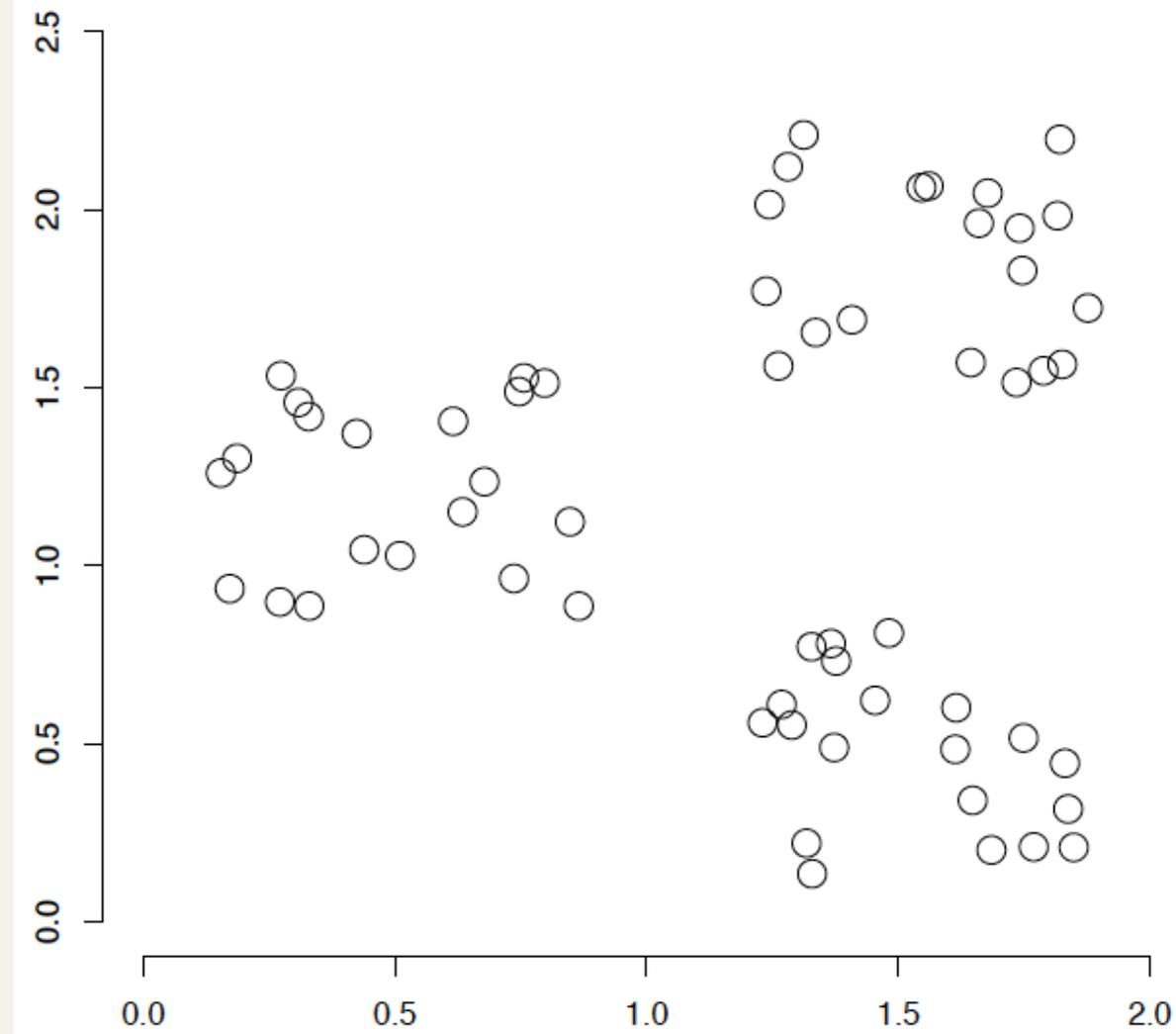
Classification Methods (3)

- Supervised learning of a document-label assignment function
 - Many systems partly rely on machine learning (Autonomy, Microsoft, Enkarta, Yahoo!, Google News, ...)
 - k-Nearest Neighbors (simple, powerful)
 - Naive Bayes (simple, common method)
 - Support-vector machines (newer, more powerful)
 - ... plus many other methods
 - No free lunch: requires hand-classified training data
- Many commercial systems use a mixture of methods
- More next lecture

What is clustering?

- **Clustering**: the process of grouping a set of objects into classes of similar objects
 - Documents within a cluster should be similar
 - Documents from different clusters should be dissimilar
- The commonest form of **unsupervised learning**
 - Unsupervised learning = learning from raw data, as opposed to supervised data where a classification of examples is given
- A common and important task that finds many applications in IR and other places

A data set with clear cluster structure



- How would you design an algorithm for finding the three clusters in this case?

Supervised and Unsupervised Learning

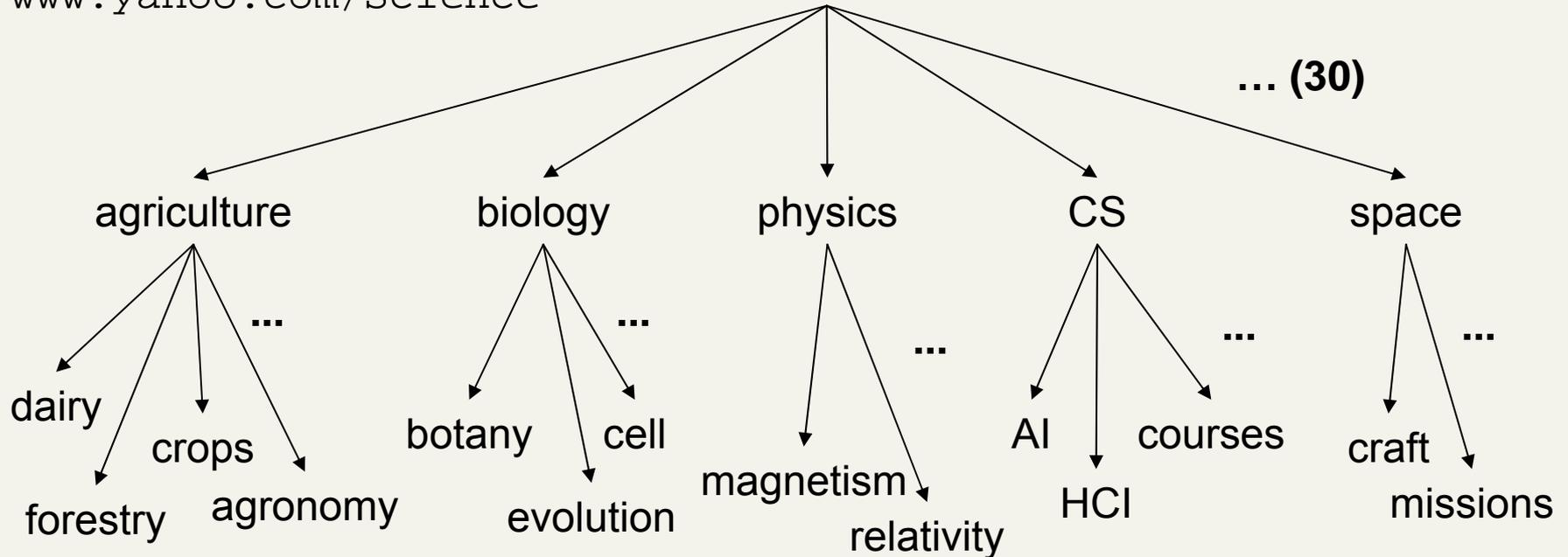
- **Learning:** We are given a collection X , and we want to learn some function γ over X
 - Classification: $\gamma(d)$ tells you the class of document d
 - Clustering: $\gamma(d)$ tells you the cluster in which d belongs
- **Supervised learning:** We have examples of some $\gamma(d)$ and we want to compute γ for the rest
 - Classification
- **Unsupervised learning:** We only work on raw data. We don't know the value of any $\gamma(d)$.
 - Clustering

Applications of clustering in IR

- Whole corpus analysis/navigation
 - Better user interface: search without typing
- For improving recall in search applications
 - Better search results (like pseudo RF)
- For better navigation of search results
 - Effective “user recall” will be higher
- For speeding up vector space retrieval
 - Cluster-based retrieval gives faster search

Yahoo! Hierarchy *isn't* clustering but *is* the kind of output you want from clustering

www.yahoo.com/Science



Google News: automatic clustering gives an effective news presentation metaphor

The screenshot shows the Google News interface in a browser window. The address bar displays 'http://news.google.com/'. The page is organized into a grid of news stories. On the left, there are three main sections under the 'World' tab: 'Pirates Demand \$25 Million Ransom for Hijacked Tanker (Update1)', 'Pakistan protests over US missile strikes', and 'Nighttime attack on Thai antigovernment protesters wounds at least 20'. On the right, there are two main sections under the 'U.S.' tab: 'Top Court in California Will Review Proposition 8' and 'Drop That Cigarette, Today Is The Great American Smokeout'. Each article includes a headline, a byline with the source and time ago, a short summary, and a list of related links. Small thumbnail images are placed next to each article. At the bottom of the page, there are buttons for 'Show more stories' and 'Show fewer stories' for each section, and a URL: 'http://www.google.com/hostednews/ap/article/ALeqM5hGjNxBi6O23C8QzqZMY0pGPAik--AD94INLTG1'.

World » [edit](#)

Pirates Demand \$25 Million Ransom for Hijacked Tanker (Update1) 
Bloomberg - 36 minutes ago
By Caroline Alexander and Hamsa Omar Nov. 20 (Bloomberg) -- Somali pirates are demanding \$25 million in ransom to release an oil-laden Saudi supertanker seized off the East African coast, and called on the ship's owners to pay up "soon."
[Somali pirates demand \\$25M for Saudi ship](#) United Press International
[African Union says Somali politicians fuel piracy](#) Washington Post
[BBC News](#) - [guardian.co.uk](#) - [Aljazeera.net](#) - [RIA Novosti](#)
[all 4,015 news articles »](#)

Pakistan protests over US missile strikes 
Reuters - 2 hours ago
By Simon Cameron-Moore ISLAMABAD (Reuters) - Pakistan summoned US ambassador Anne Patterson on Thursday to protest over missile strikes launched by pilotless drone aircraft against militant targets in Pakistan.
[Pakistan protests US drone attacks, Taliban warns of reprisals](#) AFP
[Pakistan warns US over missile strike](#) CNN International
[Telegraph.co.uk](#) - [China Daily](#) - [Xinhua](#) - [PRESS TV](#)
[all 560 news articles »](#)

Nighttime attack on Thai antigovernment protesters wounds at least 20 
Christian Science Monitor - 30 minutes ago
The government denied attacking demonstrators, who have called for the ouster of the prime minister. By Huma Yusuf One person has been killed and 23 others wounded in a grenade attack Thursday against antigovernment protesters occupying the Thai prime ...
[Blast Kills 1, Wounds 23 at Thai Prime Minister's Office](#) Washington Post
[Anti-government protestor in Thailand dies in grenade attack](#) International Herald Tribune
[Xinhua](#) - [United Press International](#) - [The Associated Press](#) - [AsiaOne](#)
[all 688 news articles »](#)

[Show more stories](#) [Show fewer stories](#)

U.S. » [edit](#)

Top Court in California Will Review Proposition 8 
New York Times - 1 hour ago
By JESSE MCKINLEY SAN FRANCISCO - Responding to pleas for legal clarity from those on both sides of the issue, the California Supreme Court said Wednesday that it would take up the case of whether a voter-approved ban on same-sex unions was ...
[California Supreme Court to decide fate of Prop. 8 same-sex ...](#)
San Jose Mercury News
[Prop. 8 gay marriage ban goes to Supreme Court](#) Los Angeles Times
[The Miami Herald](#) - [San Diego Union Tribune](#) - [Indiana Daily Student](#) - [San Francisco Chronicle](#)
[all 1,241 news articles »](#)

Drop That Cigarette, Today Is The Great American Smokeout 
dBTechno - 1 hour ago
Washington (dbTechno) - Today marks the annual Great American Smokeout hosted by the American Cancer Society, and is trying to get people all across the US to drop their cigarettes for just one day.
[Great American Smokeout: Time to kick the habit](#) Capital Times
[National Smoke Out Day is Thursday, be a quitter](#) Las Cruces Sun-News
[MPNnow.com](#) - [eMaxHealth.com](#) - [Times Tribune of Corbin](#) - [ABC15.com \(KNXV-TV\)](#)
[all 338 news articles »](#)

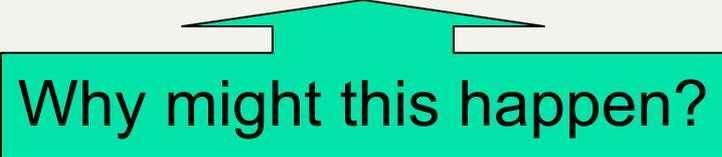
Perino: Bush would sign jobless benefits extension 
The Associated Press - 47 minutes ago
WASHINGTON (AP) - With weekly jobless claims benefits at a 16-year high, the White House said Thursday that President George W. Bush would quickly sign legislation pending in Congress to provide further unemployment benefits.
[Bush would sign measure to extend jobless benefits](#) Houston Chronicle
[Jobless claims show need for benefits extension: White House](#) AFP
[Washington Times](#) - [Wall Street Journal Blogs](#) - [WOI](#) - [Tampabay.com](#)
[all 599 news articles »](#)

[Show more stories](#) [Show fewer stories](#)

[http://www.google.com/hostednews/ap/article/ALeqM5hGjNxBi6O23C8QzqZMY0pGPAik--AD94INLTG1](#)

For improving search recall

- **Cluster hypothesis** - Documents in the same cluster behave similarly with respect to relevance to information needs
- Therefore, to improve search recall:
 - Cluster docs in corpus a priori
 - When a query matches a doc D , also return other docs in the cluster containing D
- Hope if we do this: The query “car” will also return docs containing *automobile*
 - Because clustering grouped together docs containing *car* with those containing *automobile*.



Why might this happen?

clouds sources sites remix

All Results (185)

- + Analysis (23)
- + Method (22)
- + Computing (15)
- + Search, Engine (13)
- + Hierarchical (16)
- + Definition (11)
- + High availability (13)
- + Linux (11)
- + Windows, Microsoft (9)
- + Papers (8)

more | all clouds

find in clouds: Find

Top 179 results retrieved for the query **clustering** (definition) (details)

Clustering

Lower Latency In Your Data Center w/ Intel's **Cluster** Ready Solutions!
www.intel.com

Load Balancing 101

Learn the 'Nuts & Bolts' of Load Balancing with F5's White Paper
www.f5.com/load_balancing

Affordable Load Balancers

High Performance Load Balancing Solutions From KEMP- See Demo Today
kemptechnologies.com

Computer **cluster** - Wikipedia, the free encyclopedia

Middleware such as MPI (Message Passing Interface) or PVM (Parallel Virtual Machine) permits compute **clustering** programs to be portable to a /Computer_ **cluster**
en.wikipedia.org/wiki/Computer_cluster - [cache] - Bing, Yahoo!

Writer's Web: Prewriting: **Clustering**

Prewriting: **Clustering** Melanie Dawson & Joe Essid (printable version here) **Clustering** is a type of prewriting that allows you to explore many ideas
writing2.richmond.edu/writing/wwweb/cluster.html
writing2.richmond.edu/writing/wwweb/cluster.html - [cache] - Bing, Yahoo!

Getting Started: **Clustering** Ideas - CT Community Colleges

Clustering. **Clustering** is similar to another process called Brainstorming. **Clustering** is something that you can do on your own or with friends or
grammar.ccc.commnet.edu/grammar/composition/brainstorm_cluster.htm
grammar.ccc.commnet.edu/grammar/composition/brainstorm_clustering.htm - [cache] - Bing, Yahoo!

Advanced **Clustering** | Home

