# 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: "" <takworlld@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:

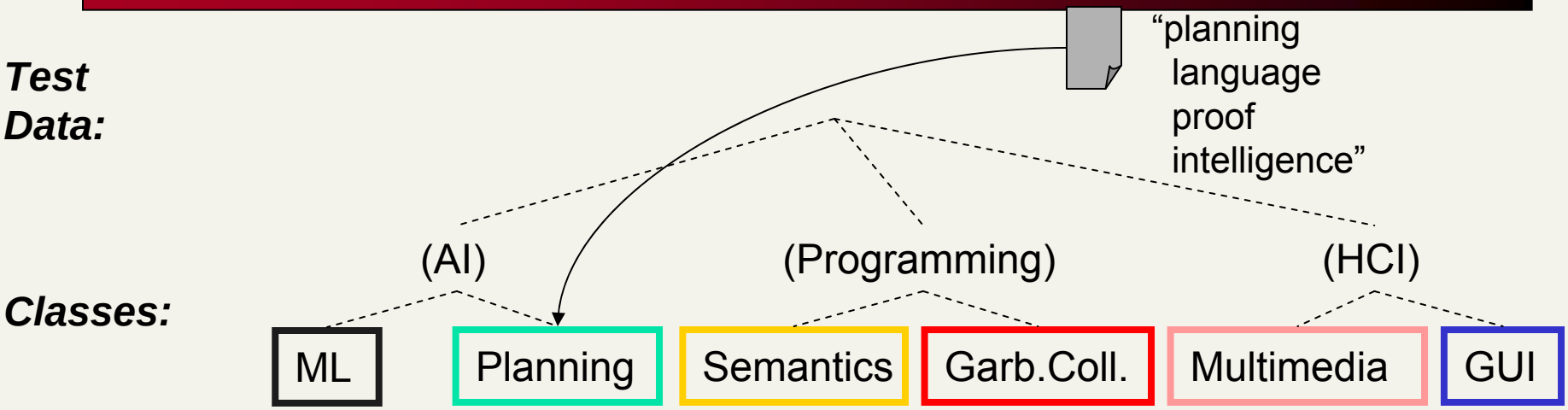
    $C = \{c_1, c_2, \ldots, c_J\}$
  - A training set D of labeled documents with each labeled document $(d,c) \in X \times C$
- Determine:
  - A learning method or algorithm which will enable us to learn a classifier $\gamma: X \rightarrow C$
  - For a test document $d$, we assign it the class $\gamma(d) \in C$

# 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:

    $C = \{c_1, c_2, \ldots, c_J\}$
- 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




(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
  evidence topic        ** 0.50 WORD
  topic definition modifier     /wordtext = ballet
  evidence topic        ** 0.50 STEM
  topic definition modifier     /wordtext = dance
  evidence topic        ** 0.50 WORD
  topic definition modifier     /wordtext = opera
  evidence topic        ** 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)

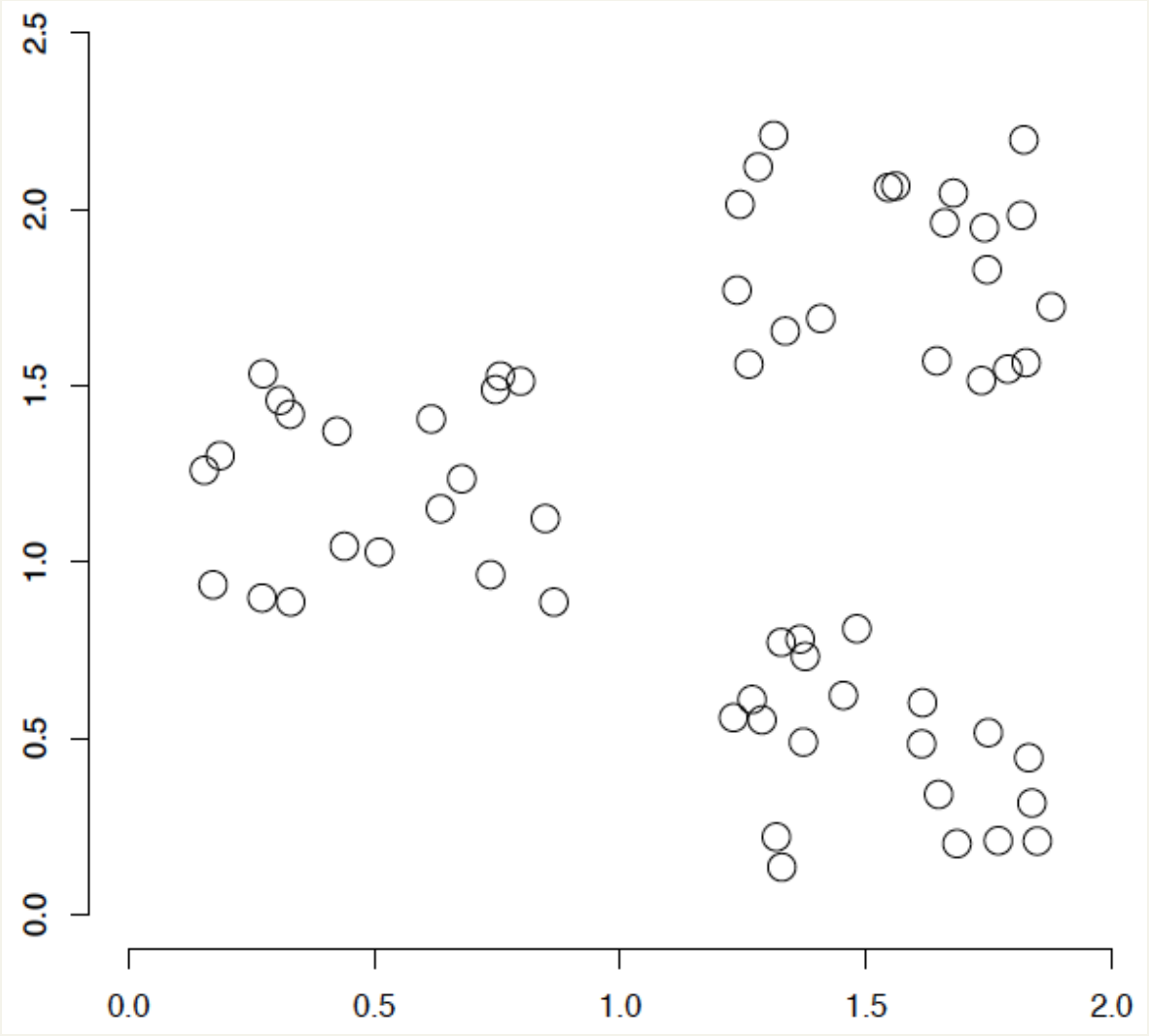
- Supervised learning of a document-label assignment function
  - Many systems partly rely on machine learning (Autonomy, Microsoft, Enkarta, Yahoo!, Google News, …)
    - k-Nearest Neighbors (simple, powerful)
    - Naive Bayes (simple, common method)
    - Support-vector machines (newer, more powerful)
    - … 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?

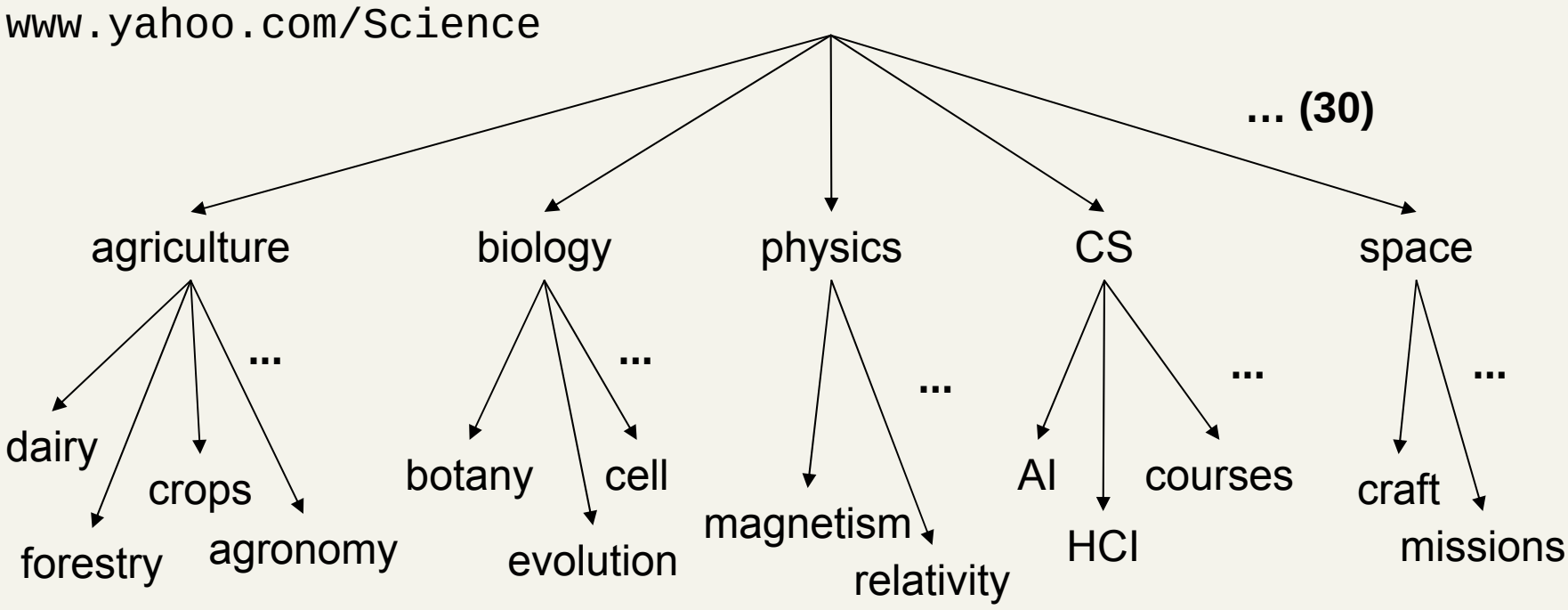
# Supervised and Unsupervised Learning

- **Learning:** We are given a collection X, and we want to learn some function γ over X
  - Classification: γ(d) tells you the class of document d
  - Clustering: γ(d) tells you the cluster in which d belongs
- **Supervised learning:** We have examples of some γ(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 γ(d).
  - Clustering

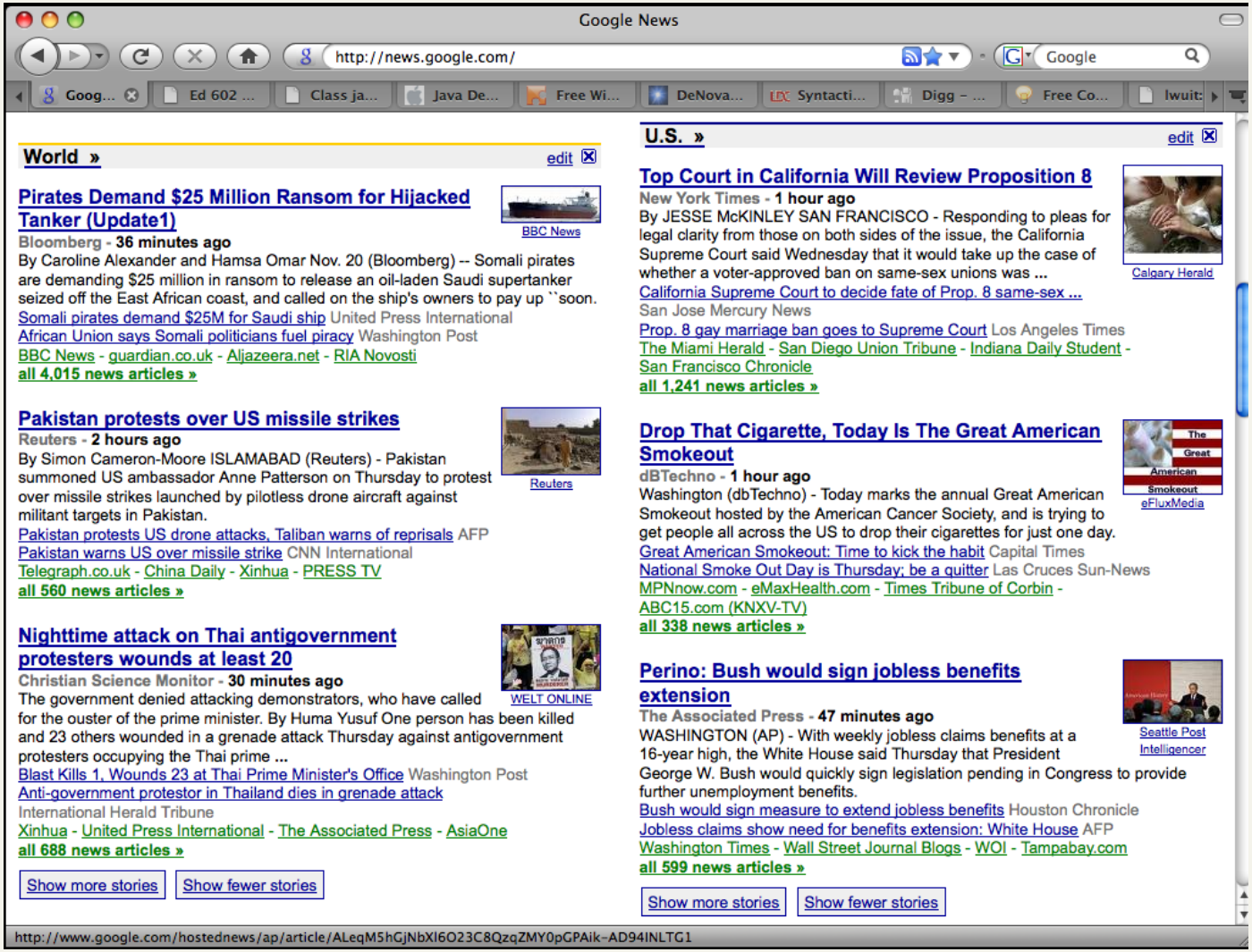
# Applications of clustering in IR

- 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

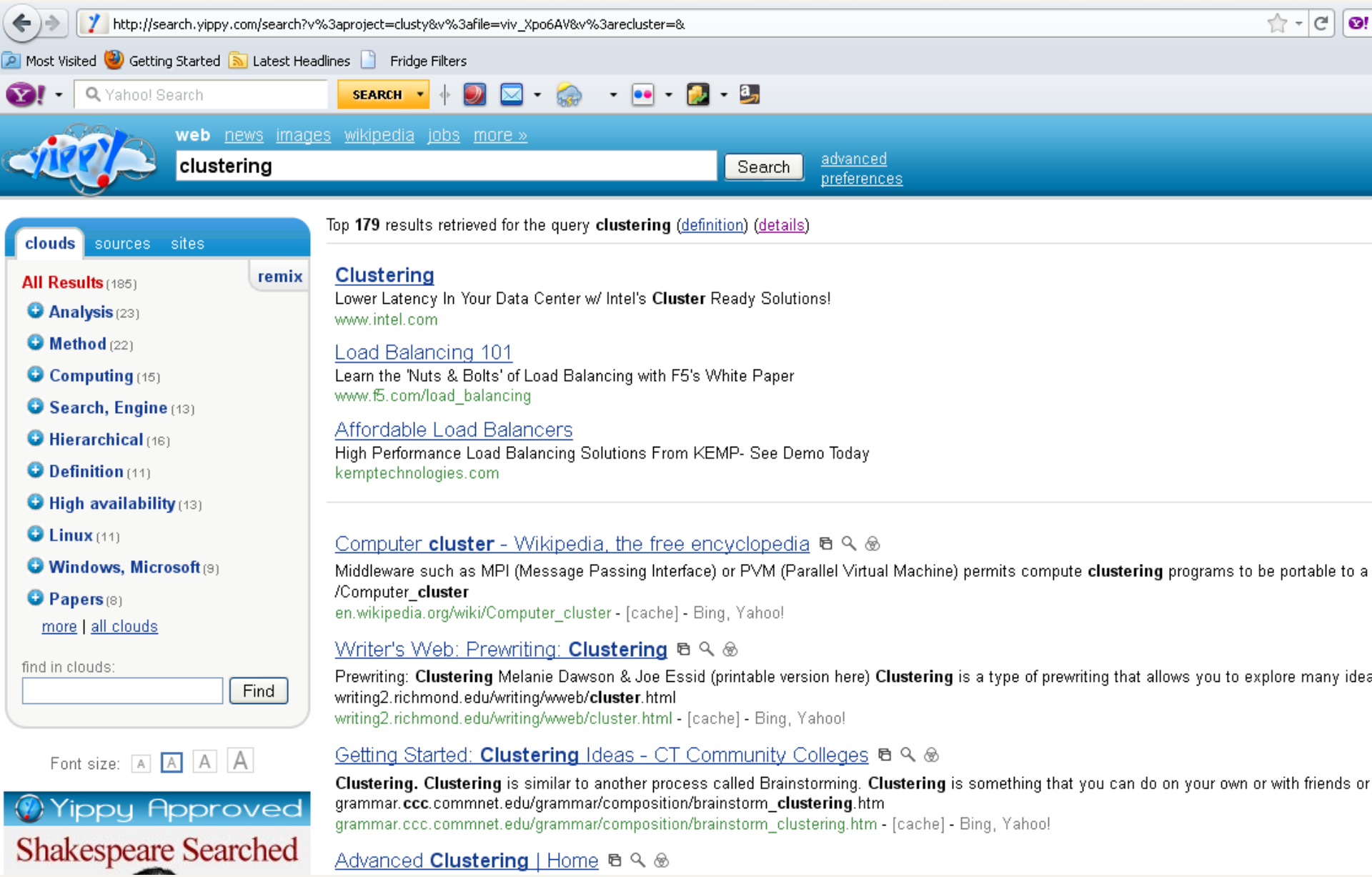


# Google News: automatic clustering gives an effective news presentation metaphor



# 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.*




yippy.com – grouping search results