

# Dat5

## Introduction to Text Mining and Web Search

Gao Cong

[gaocong@cs.aau.dk](mailto:gaocong@cs.aau.dk)

Some slides are borrowed from Prof. Marti Hearst, Christopher Manning, Louis Eisenberg, Bing Liu, and Prabhakar Raghavan

# Objectives

- To have a rough idea of text mining and web search
  - Not about deep techniques
- Connections with StreamSpin
  - Mobile services as component of SS

# Corporate Knowledge “Ore”

Stuff not very accessible via standard data-mining

- Email
- Insurance claims
- News articles
- Web pages
- Patent portfolios
- IRC
- Scientific articles
- Customer complaint letters
- Contracts
- Transcripts of phone calls with customers
- Technical documents

# Definitions of Text Mining

- Text mining mainly is about somehow extracting the information and knowledge from text;
- 2 definitions:
  - Any operation related to gathering and analyzing text from external sources for business intelligence purposes;
  - Discovery of knowledge previously unknown to the user in text;
- Text mining is the process of compiling, organizing, and analyzing large **document collections** to support the delivery of targeted types of information to analysts and decision makers and to discover relationships between related facts that span wide domains of inquiry.

# Text Mining

- Text classification
- Text clustering
- Named entity recognition
- Information extraction
- Information retrieval engine
- Web spider/search
- Question answering
- Opinion mining
- Summarization
- Topic detection and tracking
- ...

# Text Classification (Categorization) Clustering

Text Classification and its application

Clustering and its application

# Is this spam?

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>

=====

# More Examples

Assign labels to each document or web-page:

- Labels are most often **topics** such as Yahoo-categories  
*e.g., "finance," "sports," "news>world>asia>business"*
- Labels may be **opinion**  
*e.g., "like", "hate", "neutral"*
- Labels may be **domain-specific binary**  
*e.g., "interesting-to-me" : "not-interesting-to-me"*  
*e.g., "spam" : "not-spam"*  
*e.g., "contains adult language" : "doesn't"*
- Labels may be **genres**  
*e.g., "editorials" "movie-reviews" "news"*



# Classification

- *Training data*
  - A description of an instance by a set of features.
    - Issue: how to represent text documents.
    - Feature selection
  - A fixed set of categories
  - *Build classification model*
- *Apply classification model to new data instance described by a set of features*

# Classification model

- K nearest neighbor
- Naïve Bayesian model
- Rule based model
  - Decision tree
  - Association rule based model
- Support Vector Machine
- Neural Network
- ...

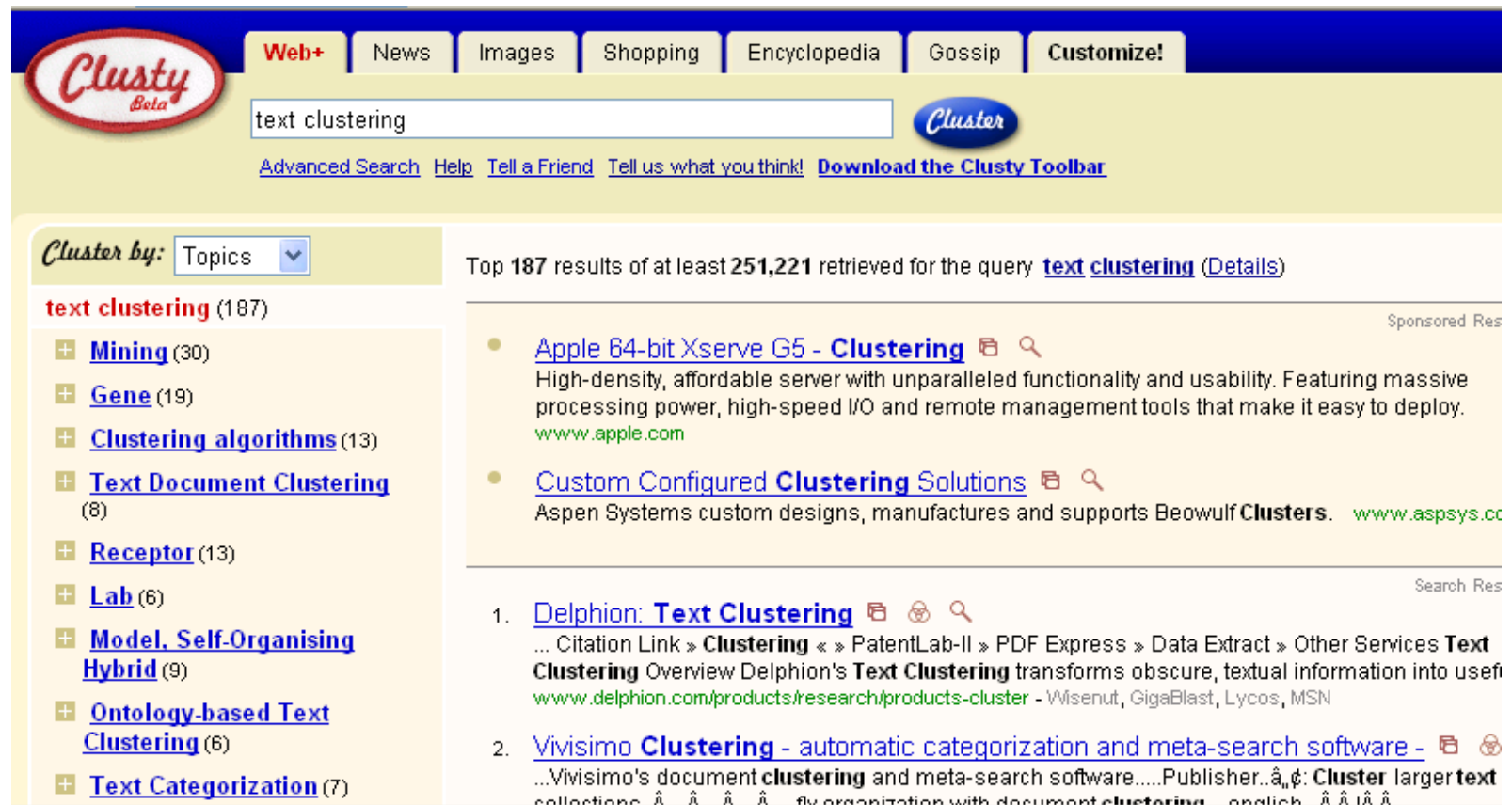
Yiming Yang & Xin Liu, A re-examination of text categorization methods.  
*Proceedings of SIGIR, 1999*

# Clustering

- **Clustering**: the process of grouping a set of objects into classes of similar objects
  - *unsupervised learning: no training data*
  - A common and important task that finds many applications in IR and other places
    - Whole corpus analysis/navigation
      - Better user interface
    - For improving recall in search applications
      - Better search results

# For better navigation of search results

- For grouping search results thematically
  - clusty.com / Vivisimo



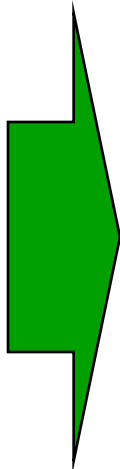
The screenshot displays the Clusty Beta search engine interface. At the top, there is a navigation bar with tabs for 'Web+', 'News', 'Images', 'Shopping', 'Encyclopedia', 'Gossip', and 'Customize!'. The search bar contains the text 'text clustering' and a 'Cluster' button. Below the search bar, there are links for 'Advanced Search', 'Help', 'Tell a Friend', 'Tell us what you think!', and 'Download the Clusty Toolbar'.

The main content area is divided into two columns. The left column, titled 'Cluster by: Topics', shows a list of thematic clusters for the search results:

- text clustering (187)
- + Mining (30)
- + Gene (19)
- + Clustering algorithms (13)
- + Text Document Clustering (8)
- + Receptor (13)
- + Lab (6)
- + Model, Self-Organising Hybrid (9)
- + Ontology-based Text Clustering (6)
- + Text Categorization (7)

The right column displays the search results. It starts with 'Top 187 results of at least 251,221 retrieved for the query **text clustering** (Details)'. The first result is a sponsored link for 'Apple 64-bit Xserve G5 - Clustering', described as a high-density server with massive processing power. The second result is 'Custom Configured Clustering Solutions' from Aspen Systems. Below these, there are two search results from Vivisimo:

1. **Delphion: Text Clustering** ... Citation Link » Clustering « » PatentLab-II » PDF Express » Data Extract » Other Services **Text Clustering** Overview Delphion's **Text Clustering** transforms obscure, textual information into usefu  
[www.delphion.com/products/research/products-cluster](http://www.delphion.com/products/research/products-cluster) - Wisenut, GigaBlast, Lycos, MSN
2. **Vivisimo Clustering - automatic categorization and meta-search software -** ...Vivisimo's document **clustering** and meta-search software.....Publisher..â„¢: **Cluster** larger**text** collections. â„¢ â„¢ â„¢ â„¢ fly organization with document **clustering**... english. â„¢ â„¢ â„¢



# Applications and Models of Clustering

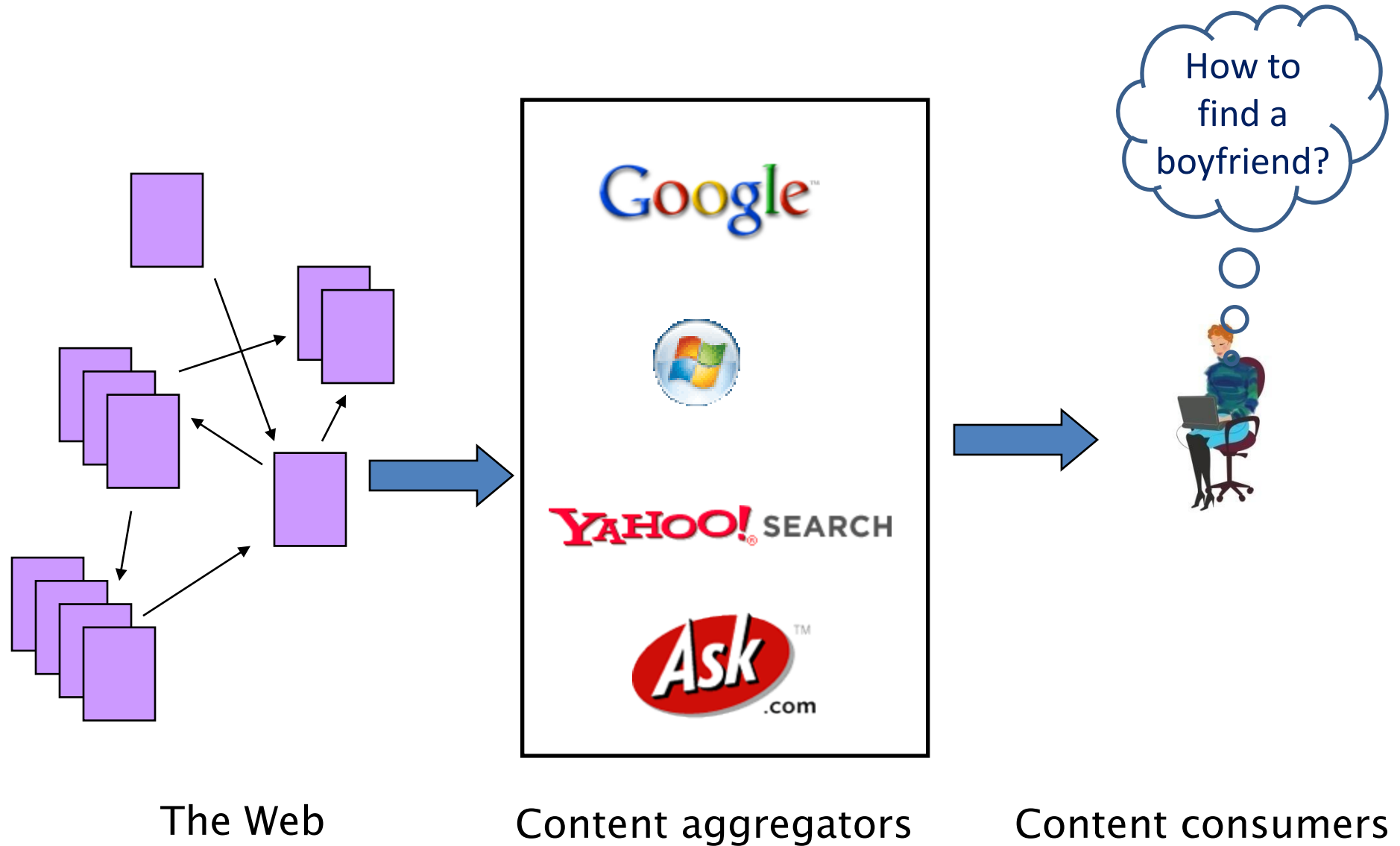
- Examples of Clustering Applications
  - Marketing: Help marketers discover **distinct groups** in their customer bases, and then use this knowledge to develop targeted marketing programs
  - Insurance: Identifying **groups of motor insurance policy holders** with a high average claim cost
  - City-planning: Identifying **groups of houses** according to their house type, value, and geographical location
- Clustering models
  - Partitioning Methods
  - Hierarchical Methods
  - Density-Based Methods
  - Grid-Based Methods
  - Model-Based Clustering Methods

# WEB SEARCH

Main components of search engines

Advertisement and search engineer

# Example1: Searching the Web



# Brief (non-technical) history

- Early keyword-based engines
  - Altavista, Excite, Infoseek, Inktomi, Lycos, ca. 1995-1997
- Paid placement ranking: Goto.com (morphed into Overture.com → Yahoo!)
  - Your search ranking depended on how much you paid
  - Auction for keywords: casino was expensive!



# Brief (non-technical) history

- 1998+: Link-based ranking pioneered by Google
  - Blew away all early engines save Inktomi
  - Great user experience in search of a business model
  - Meanwhile Goto/Overture's annual revenues were nearing **\$1 billion**
- Result: Google added paid-placement “ads” to the side, independent of search results
  - Internet providers to create content-rich [broadband](#) services ; 2003: Yahoo follows suit, acquiring Overture (for paid placement) and Inktomi (for search);

# Ads vs. search results

- Google has maintained that ads (based on vendors bidding for keywords) do not affect other rankings in search results

Sponsored Links

## [Fly to Aalborg](#)

**Aalborg** Flights with a flight site for the right flight.  
[www.cheapflightsite.co.uk](http://www.cheapflightsite.co.uk)

## [Studying In Denmark](#)

The Official Guide to studying in Denmark.  
[www.Denmark.dk/Study](http://www.Denmark.dk/Study)

Web

### [Aalborg University: Study Administration - Aalborg University](#)

**Aalborg** University The **Study** Administration (Studieforvaltningen) Fredrik Bajers Vej 5 DK - 9220 **Aalborg** East Denmark. Please note that The International ...

[en.aau.dk/About+Aalborg+University/University+Structure/Administration/Study+Administration](http://en.aau.dk/About+Aalborg+University/University+Structure/Administration/Study+Administration) - 8k - [Cached](#) - [Similar pages](#)

### [Study Abroad - International Students - Aalborg Universitet \(AAU\)](#)

If you are enrolled at **Aalborg** University as a student and are interested in a **study** period abroad, Ms. Mariann Simonsen at the International Office is ...

[studyguide.aau.dk/enrolledstudents/studyabroad](http://studyguide.aau.dk/enrolledstudents/studyabroad) - 7k - [Cached](#) - [Similar pages](#)

### [PAU Home: New in DK? - Study Cards - Aalborg University](#)

If not already received, you can get a PhD **study** card from the International Doctoral School. This will prove that you are a PhD student at **Aalborg** ...

[www.pau.aau.dk/newindk/1041744](http://www.pau.aau.dk/newindk/1041744) - 7k - [Cached](#) - [Similar pages](#)

### [\[PDF\] Aalborg brochure #2](#)

File Format: PDF/Adobe Acrobat - [View as HTML](#)

... lar program of **study** at **Aalborg**. Foreign language skills. are only required for students who wish to **study** in a. foreign language. ...

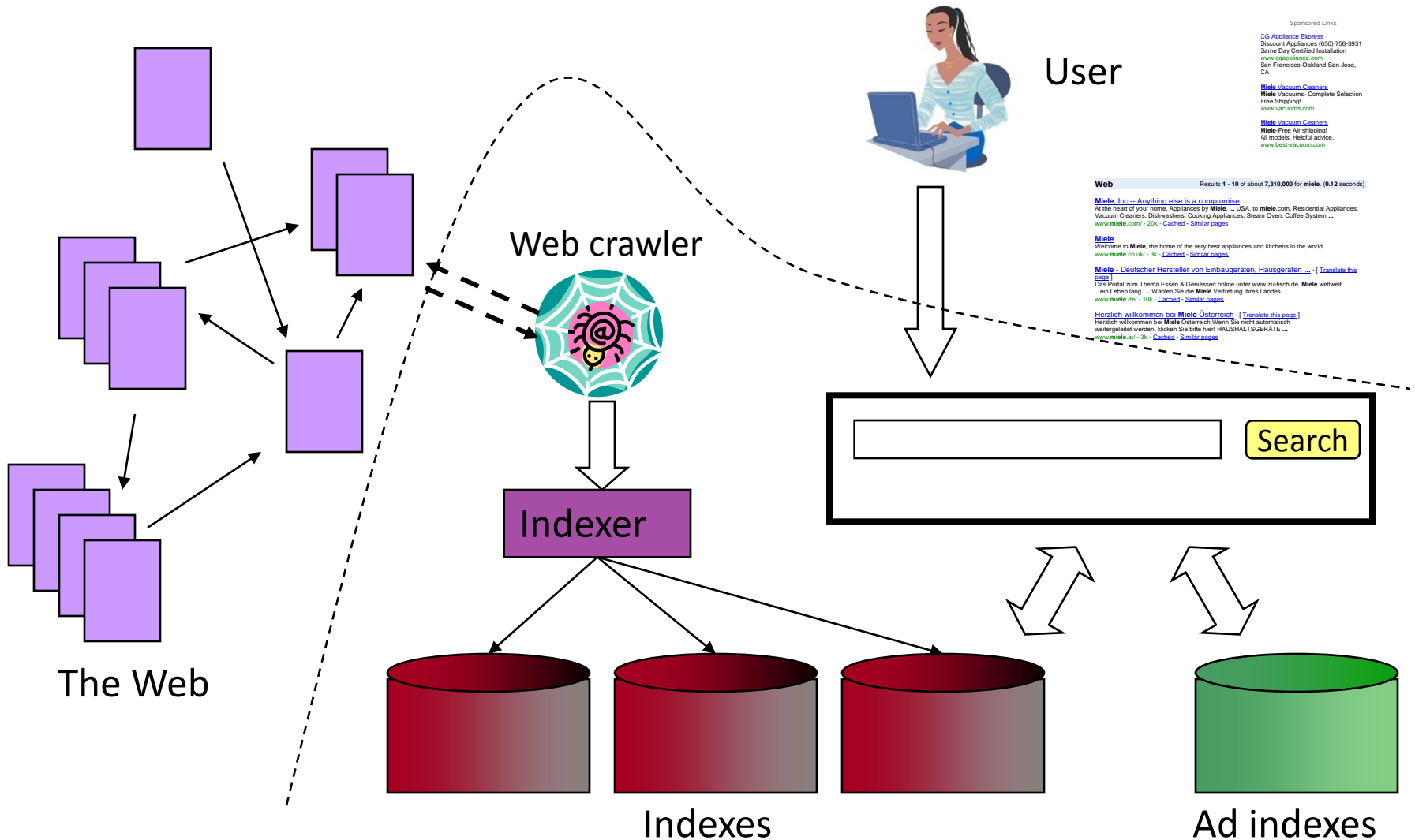
[studyabroad.uoregon.edu/brochure/aalborg.pdf](http://studyabroad.uoregon.edu/brochure/aalborg.pdf) - [Similar pages](#)

Search =  
*study in aalborg*

# Ads vs. search results

- Other vendors (Yahoo!, Live Search) have made similar statements from time to time
  - Any of them can change anytime
- We will focus primarily on search results independent of paid placement ads
  - Although the latter is a fascinating technical subject in itself
  - So, we'll look at it briefly later

# Web search basics



# How Search Engines Work

Three main parts:

**Crawler:** Gather the contents of all web pages  
(using a program called a **crawler** or **spider**)

**Indexer:** Organize the contents of the pages in a  
way that allows efficient retrieval (**indexing**)

**Ranker:** Take in a query, determine which pages  
match, and show the results (**ranking** and  
**display** of results)

Crawler

# Crawling Issues

- How to crawl?
  - *Quality*: “Best” pages first
  - *Efficiency*: Avoid duplication (or near duplication)
  - *Etiquette*: Robots.txt, Server load concerns
- How much to crawl? What to crawl?
  - *Coverage*: How big is the Web? How much do we cover?
  - *Relative Coverage*: How much do competitors have?
- How often to crawl?
  - *Freshness*: How much has changed?
  - Pages change (25%, 7% large changes)
    - At different frequencies

# Basic crawler algorithm

- Begin with known “seed” pages (on a queue)
- Fetch and parse them
  - Extract URLs they point to
  - Place the extracted URLs on a queue
- Fetch each URL on the queue and repeat



# Four Laws of Crawling

- A Crawler must show identification
- A Crawler must obey the robots exclusion standard
  - <http://www.robotstxt.org/wc/norobots.html>
  - For a URL, create a file `URL/robots.txt`
- A Crawler must not hog resources
  - Politeness – don't hit a server too often
- A Crawler must report errors

# Example robots.txt file

[www.whitehouse.gov/robots.txt](http://www.whitehouse.gov/robots.txt)  
(just the first few lines)

```
User-agent: *
Disallow: /cgi-bin
Disallow: /search
Disallow: /query.html
Disallow: /help
Disallow: /360pics/text
Disallow: /911/911day/text
Disallow: /911/heroes/text
Disallow: /911/messages/text
Disallow: /911/patriotism/text
Disallow: /911/patriotism2/text
Disallow: /911/progress/text
Disallow: /911/remembrance/text
Disallow: /911/response/text
Disallow: /911/sept112002/text
Disallow: /911/text
Disallow: /ConferenceAmericas/text
Disallow: /GOVERNMENT/text
Disallow: /QA-test/text
Disallow: /aci/text
Disallow: /afac/text
Disallow: /africanamerican/text
Disallow: /africanamericanhistory/text
Disallow: /agencycontact/text
Disallow: /americancompetitiveness/text
Disallow: /apec/2003/text
Disallow: /apec/2004-summit/text
Disallow: /apec/2004/text
```

# What really gets crawled?

- A small fraction of the Web that search engines know about; no search engine is exhaustive
- Not the “live” Web, but the search engine’s index
- Not the “Deep Web”, but start to do this
  - E.g., buy some stuff in Amazon, password protected
  - In 2000, a research project by Berkeley: **91,000 terabytes**. By contrast, the surface Web (which is easily reached by search engines) is only about **167** terabytes
- Mostly HTML pages but other file types too: PDF, Word, PPT, etc.

# Lots of tricky aspects

- Servers are often down or slow
- Hyperlinks can get the crawler into cycles
- Some websites have spam in the web pages
- Now many pages have dynamic content
  - E.g. javascript, the deep web

# Lots of tricky aspects

- The web is **HUGE**
  - Distributed crawling
  - Crawl order
    - Most sites, stay at  $\leq 5$  levels of URL hierarchy
    - Which URLs are most promising for building a high-quality corpus
  - Filtering duplicates, and Mirror detection
    - Fetch from the fastest, inlink and outlink
  - Malicious pages, Spider traps
  -

Indexer

# Index (the database)

Record information about each page

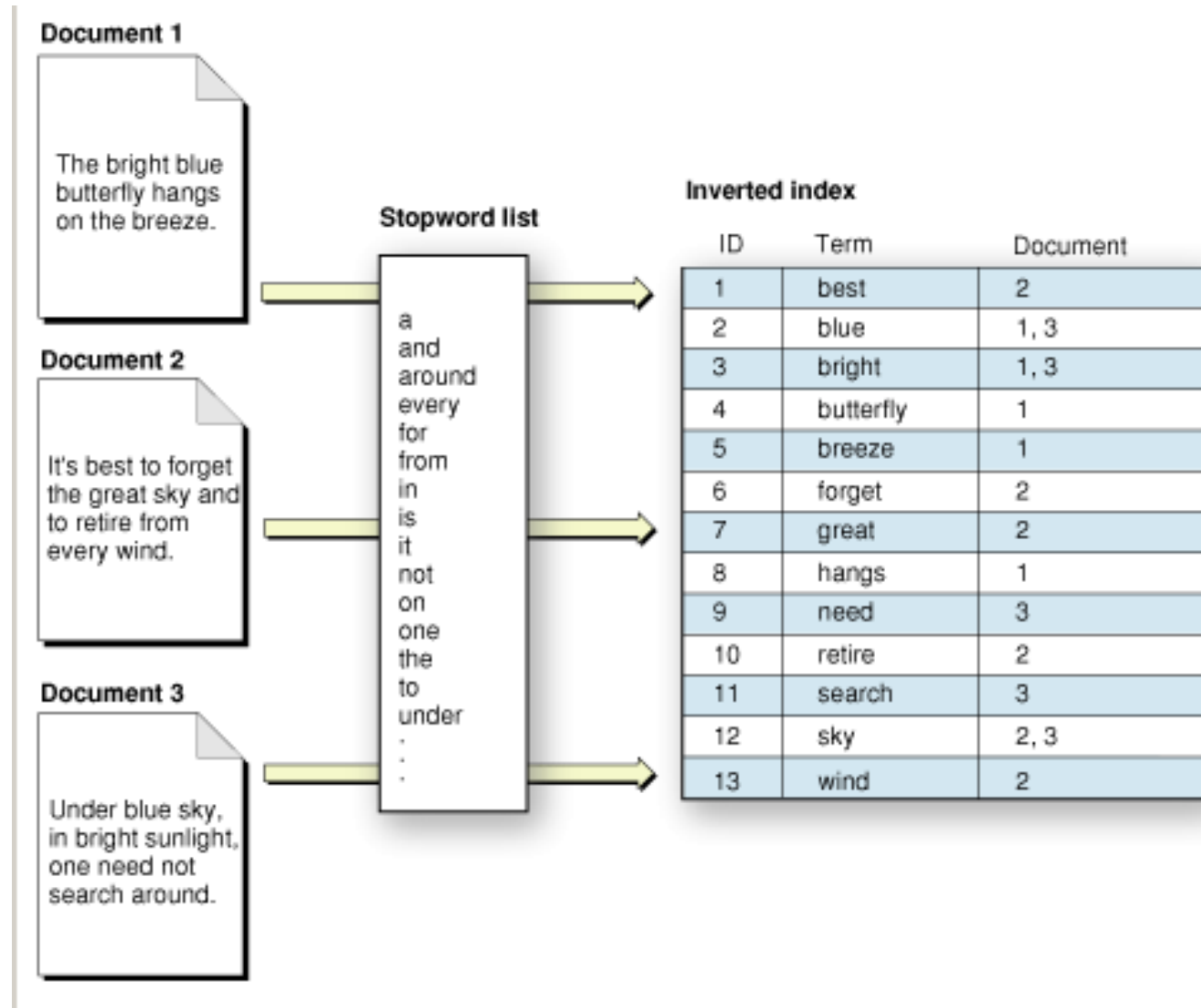
- List of words
  - In the title?
  - How far down in the page?
  - Was the word in boldface?
- URLs of pages pointing to this one
- Anchor text on pages pointing to this one
  - [AAU DB](#)

# Inverted Index

- How to store the words for fast lookup
- Basic steps:
  - Make a “dictionary” of all the words in all of the web pages
  - For each word, list all the documents it occurs in.
  - Often omit very common words
    - “**stop words**”
  - Sometimes **stem** the words
    - (also called **morphological analysis**)
    - cats -> cat
    - running -> run



# Inverted Index Example



# Query processing: AND

- Consider processing the query:

## *Computer AND science*

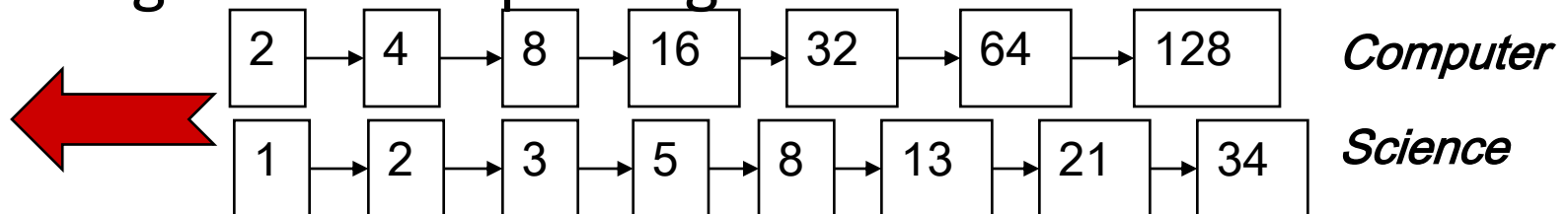
- Locate *computer* in the Dictionary;

- Retrieve its postings.

- Locate science in the Dictionary;

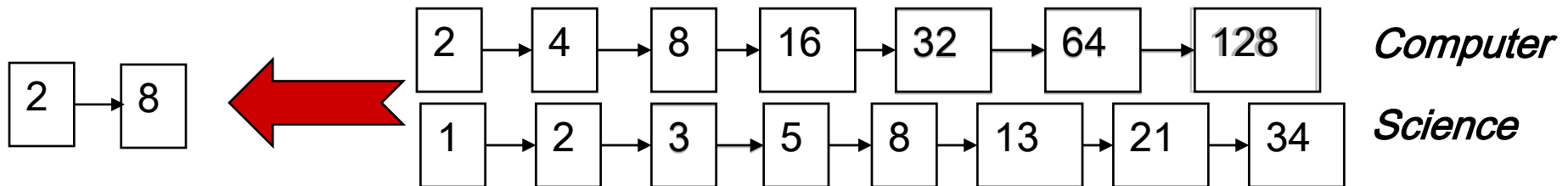
- Retrieve its postings.

- “Merge” the two postings:



# The merge

- Walk through the two postings simultaneously, in time linear in the total number of postings entries



If the list lengths are  $x$  and  $y$ , the merge takes  $O(x+y)$  operations.

Crucial: postings sorted by docID.

# Inverted Index

- In reality, this index is HUGE
- Need to store the contents across many machines
- Need to do optimization tricks to make lookup fast.

# Discussion: what information is missing from the simple index?

- Frequency
  - Party in Aalborg
- Words are the same important?
  - Introduction to XML
- Location, in title, in anchor text
- Proximity
- ....

# term frequency: $tf$

- $tf$  the frequency of a term in a page
- Weighting  $tf$  is the relative importance of
  - 0 vs. 1 occurrence of a term in a doc
  - 1 vs. 2 occurrences
  - 2 vs. 3 occurrences ...
- Unclear: while it seems that more is better, a lot isn't proportionally better than a few
  - Can just use raw  $tf$
  - Another option commonly used in practice:

$$wf_{t,d} = 0 \text{ if } tf_{t,d} = 0, \quad 1 + \log tf_{t,d} \text{ otherwise}$$

# Score computation

- Score for a query  $q$  = sum over terms  $t$  in  $q$ :

$$= \sum_{t \in q} tf_{t,d}$$

- [Note: 0 if no query terms in document]
- This score can be zone-combined
- Can use  $wf$  instead of  $tf$  in the above
- Still doesn't consider term scarcity in collection (*ides* is rarer than *of*)

# Weighting should depend on the term overall

- Which of these tells you more about a doc?
  - 10 occurrences of *house*?
  - 10 occurrences of *the*?
- Would like to attenuate the weight of a common term
  - But what is “common”?



# Document frequency

- $df$  = number of docs in the corpus containing the term,  $cf$  = number of occurrences in a collection

Word	$cf$	$df$
<i>ferrari</i>	10422	17
<i>insurance</i>	10440	3997

- Document/collection frequency weighting is only possible in known (static) collection.
- So how do we make use of  $df$  ?

# tf x idf term weights

- tf x idf measure combines:
  - term frequency ( $tf$ )
    - or  $wf$ , some measure of **term density** in a doc
  - inverse document frequency ( $idf$ )
    - measure of **informativeness** of a term: its rarity across the whole corpus
    - could just be raw count of number of documents the term occurs in ( $idf_i = 1/df_i$ )
    - but by far the most commonly used version is:

$$idf_i = \log \left( \frac{n}{df_i} \right)$$

- See Kishore Papineni, NAACL 2, 2002 for theoretical justification

Ranker

# Results ranking

- Search engine receives a query, then
- Looks up the words in the index, retrieves many documents, then
- Rank orders the pages and extracts “snippets” or summaries containing query words.
  - Most web search engines assume the user wants all of the words (Boolean AND, not OR).
- These are complex and highly guarded algorithms unique to each search engine.

# Some ranking criteria

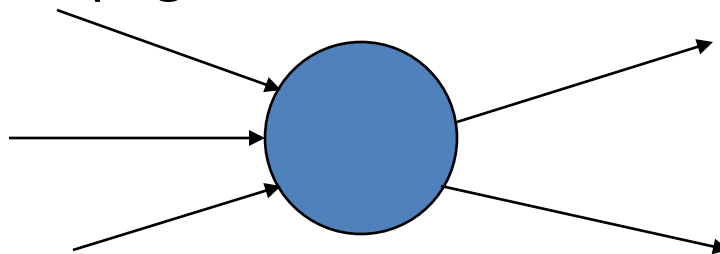
- For a given candidate result page, use:
  - Number of matching query words in the page
  - Proximity of matching words to one another
  - Location of terms within the page
  - Location of terms within tags e.g. <title>, <h1>, link text, body text
  - Anchor text on pages pointing to this one
  - Frequency of terms on the page and in general
  - Link analysis of which pages point to this one
  - (Sometimes) Click-through analysis: how often the page is clicked on
  - How “fresh” is the page
- Complex formulae combine these together. How ?

# One possible ranking

- First retrieve all pages meeting the text query (say *venture capital*).
- Order these by their link popularity (either variant on the previous page).

# Link analysis

- First generation: using link counts as simple measures of popularity.
- Two basic suggestions:
  - Undirected popularity:
    - Each page gets a score = the number of in-links plus the number of out-links ( $3+2=5$ ).
  - Directed popularity:
    - Score of a page = number of its in-links (3).



# Spamming simple popularity

- *Discussion:* How do you spam each of the following heuristics so your page gets a high score?
- Each page gets a score = the number of in-links plus the number of out-links.
- Score of a page = number of its in-links.



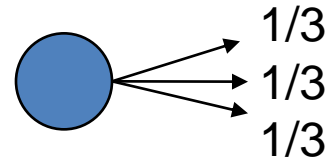
# Measuring Importance of Linking

- PageRank Algorithm
  - Idea: important pages are pointed to by other important pages
  - Method:
    - Each link from one page to another is counted as a “vote” for the destination page
    - But the importance of the starting page also influences the importance of the destination page.
    - And those pages scores, in turn, depend on those linking to them.

# Pagerank scoring

- Imagine a browser doing a random walk on web pages:

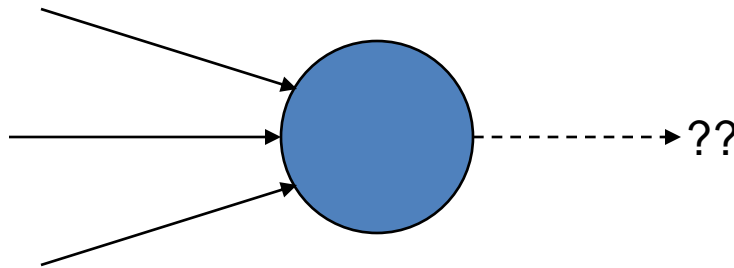
- Start at a random page



- At each step, go out of the current page along one of the links on that page, equiprobably
- “In the steady state” each page has a long-term visit rate - use this as the page’s score.

# Not quite enough

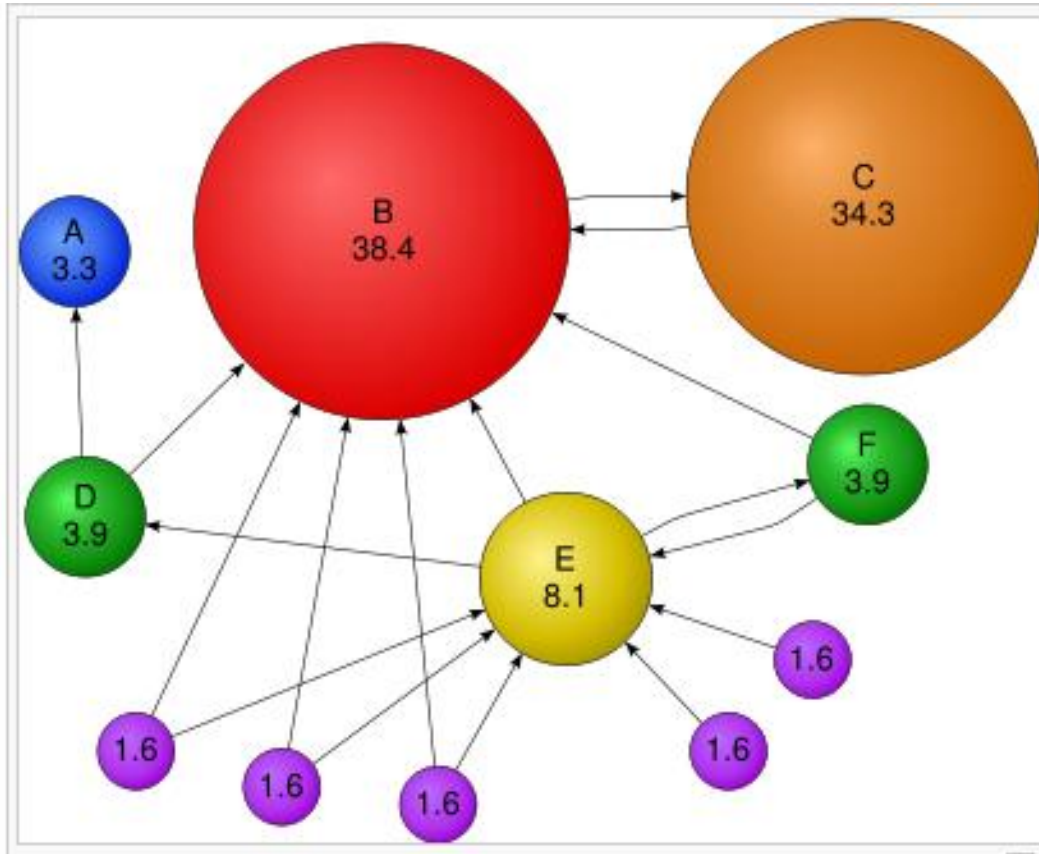
- The web is full of dead-ends.
  - Random walk can get stuck in dead-ends.
  - Makes no sense to talk about long-term visit rates.



# Teleporting

- At a dead end, jump to a random web page.
- At any non-dead end, with probability 10%, jump to a random web page.
  - With remaining probability (90%), go out on a random link.
  - 10% - a parameter.

# Pagerank



# Machine Learned Ranking

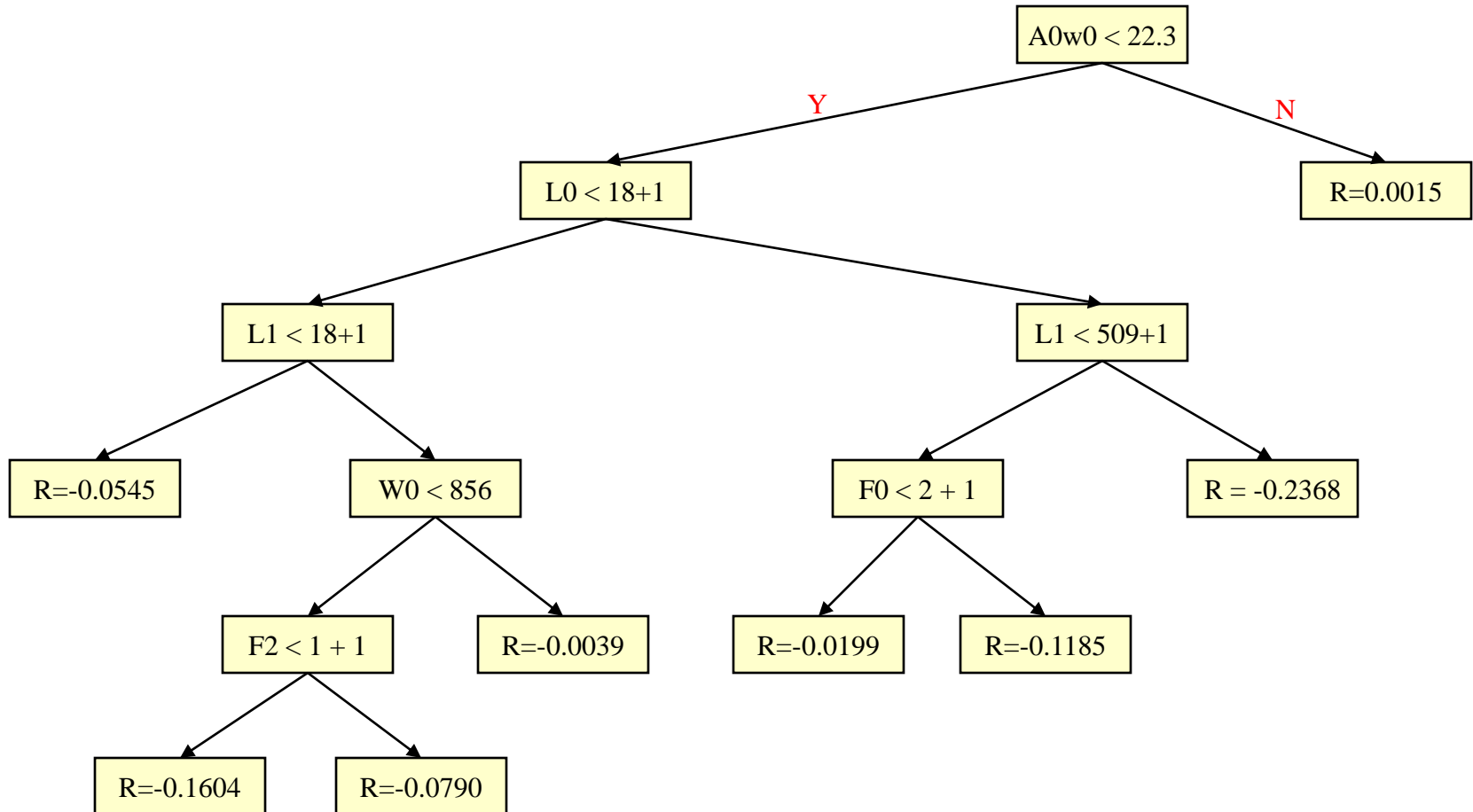
- Goal: Automatically construct a ranking function
  - Input:
    - Large number training examples
    - Features that predict relevance
    - Relevance metrics
  - Output:
    - Ranking function
- Enables rapid experimental cycle
  - Scientific investigation of
    - Modifications to existing features
    - New feature

# Ranking Features

- A0 - A4 anchor text score per term
- W0 - W4 term weights
- L0 - L4 first occurrence location  
(encodes hostname and title match)
- SP spam index: logistic regression of 85 spam filter variables  
(against relevance scores)
- F0 - F4 term occurrence frequency within document
- DCLN document length (tokens)
- ER Eigenrank
- HB Extra-host unique inlink count
- ERHB  $ER * HB$
- A0W0 etc.  $A0 * W0$
- QA Site factor –  
logistic regression of 5 site link and url count ratios
- SPN Proximity
- FF family friendly rating
- UD url depth

(from Jan Pedersen's lecture)

# Ranking Decision Tree





# Advertisement and Manipulating ranking

# advertisement

- How does SE sell keyword to advertisers?
  - One keyword can be sold to multiple advertisers
  - The more you bid, the more chance that it appears
    - It is not true. It is secret to rank ads.
    - An optimization problem to maximize its revenue
    - Bid, relevance to query, click rate

# advertisement

- Where to place advertisement?
  - Adword: Appear in the sidebar
  - Adsense: Appear in the other web pages
    - Apply to google; a good website, e.g. blog; is related to a keyword?
    - Google then sends a script to insert ad to the webpages
    - Google shares revenue with host providers
  - How to decide if a page is related to a search keyword?
  - How to detect spam click to advertisement?
    - The quality

# Trademarks and paid placement

- Consider searching Google for ***geico***
  - Geico is a large insurance company that offers car insurance
- Sponsored Links

## [Car Insurance Quotes](#)

Compare rates and get quotes from top car insurance providers.

[www.dmv.org](#)

## [It's Only Me, Dave Pell](#)

I'm taking advantage of a popular case instead of earning my traffic.

[www.davenetics.com](#)

## [Fast Car Insurance Quote](#)

21st covers you immediately. Get fast online quote now!

[www.21st.com](#)

# Who has the rights to your name?

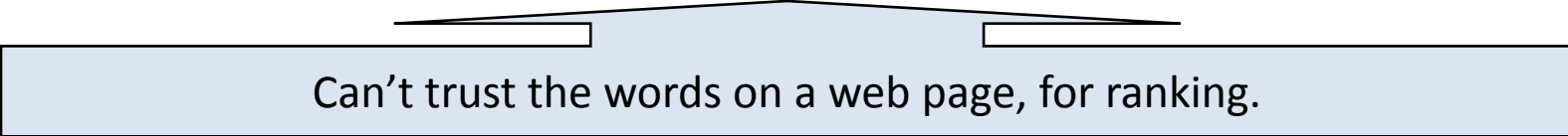
- Geico sued Google, contending that it owned the trademark “Geico” – thus ads for the keyword ***geico*** couldn’t be sold to others
  - Unlikely the writers of the constitution contemplated this issue
- Courts recently ruled: search engines can sell keywords including trademarks
  - [Personal names](#), too
- No court ruling yet: whether the ad itself can use the trademarked word(s) e.g., ***geico***

# The trouble with paid placement

- It costs money. What's the alternative?
- Search Engine Optimization:
  - “Tuning” your web page to rank highly in the search results for select keywords
  - Alternative to paying for placement
  - Thus, intrinsically a marketing function
  - Also known as Search Engine Marketing
- Performed by companies, webmasters and consultants (“Search engine optimizers”) for their clients

# Simplest forms

- Early engines relied on the density of terms
  - The top-ranked pages for the query ***aalborg resort*** were the ones containing the most ***aalborg's*** and ***resort's***
- SEOs responded with dense repetitions of chosen terms
  - e.g., ***aalborg resort aalborg resort aalborg resort***
  - Often, the repetitions would be in the same color as the background of the web page
    - Repeated terms got indexed by crawlers
    - But not visible to humans on browsers



Can't trust the words on a web page, for ranking.

# Variants of keyword stuffing

- Misleading meta-tags, excessive repetition

**Meta-Tags =**

"... London hotels, hotel, holiday inn, hilton, discount, booking, reservation, sex, mp3, britney spears, viagra, ..."



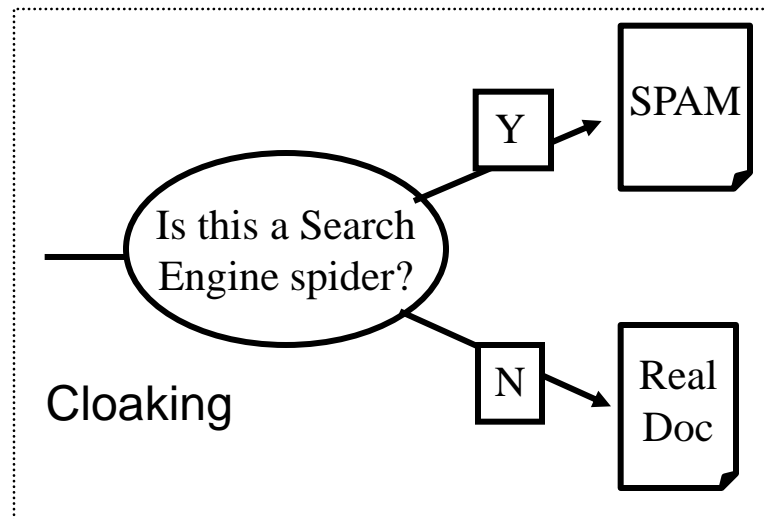
# Search engine optimization (Spam)

- Motives
  - Commercial, political, religious, lobbies
  - Promotion funded by advertising budget
- Operators
  - Contractors (Search Engine Optimizers) for lobbies, companies
  - Web masters
  - Hosting services
- Forum
  - Web master world ( [www.webmasterworld.com](http://www.webmasterworld.com) )
    - Search engine specific tricks
    - Discussions about academic papers 😊
    - More pointers in the Resources

# More spam techniques

- **Cloaking**

- Serve fake content to search engine spider
- *DNS cloaking*: Switch IP address. Impersonate





## Tutorial: Cloaking and Stealth Technology

Featured as an ongoing multi part section in our newsletter, we are offering you all the stuff you need to know, straight from the horse's mouth. Learn the secrets of the pros – subscription terminated anytime you wish.

### “Stealth, Cloaking, Phantom Tech

#### FAQ

- [What are Ghost Pages?](#)
- [What are Doorway Pages, then?](#)
- [And Hallway Pages?](#)
- [How are cloaked pages submitted?](#)
- [How about changing stealth pages?](#)
- [What are the mechanics of cloaking?](#)
- [What's a key switch?](#)
- [Isn't this really simple redirection technique?](#)
- [What about penalization?](#)

fanto  
**spide**

The

### Don't risk nasty surprises from spiders sneaking on your site under wraps!

Sure, they tend to add and switch engines, IPs and User Agents almost all the time, and keeping up with their antics is a grueling task at best. But it's also a fact that professional traffic evaluation, stealthing technology and even page submission management depend on reliable search engine reference data, if you don't want to waste your valuable resources on inventing the wheel over and over.

And consider the risks: **one single unrecognized spider crawling your doorways or debunking your stealth pages, and your top ranking with that engine may be gone for keeps!** If they don't bar you from page submissions

Tutorial on Cloaking & Stealth Technology

# More spam techniques

- **Doorway pages**
  - Pages optimized for a single keyword that re-direct to the real target page
- **Link spamming**
  - Mutual admiration societies, hidden links, awards
    - more on these later
  - *Domain flooding*: numerous domains that point or re-direct to a target page
- .....

# Acid test

- Which SEO's rank highly on the query *seo*?
- Web search engines have policies on SEO practices they tolerate/block
  - See pointers in Resources
- Adversarial IR: the unending (technical) battle between SEO's and web search engines
- See for instance <http://airweb.cse.lehigh.edu/>

# Question Answering

Transitional QA

Community based QA

# Can google find answers?

- Question: *How much money did IBM spend on advertising in 2002?*
- Answer: *I dunno, but I'd like to ...* 😞



[Advanced Search](#) [Preferences](#) [Language Tools](#) [Search Tips](#)

How much money did IBM spend on a

The following words are very common and were not included in your search: **How on in.** [\[details\]](#)

[Web](#) [Images](#) [Groups](#) [Directory](#) [News](#)

Searched the web for **How much money did IBM spend on advertising in 2002?**. Results 1 - 10 of about 13,000. Search tools

Asking a question? Try out [Google Answers](#).

**Money 2002 - Software on eBay - Buy or Sell Here!**

[www.eBay.com](http://www.eBay.com) Software for PCs, Macs, & More

Lot of ads on Google these days!

**Money 2002 - Find Prices, Reviews, Pictures, and more!**

[www.pricegrabber.com](http://www.pricegrabber.com) Comparison Shopping Beyond Compare on Money 2002

**Advertising metrics**

... How **much money** does it cost us to get a sales lead or a new customer? ... How **did** you ... more can you get by spending that **money** again ...

[www.perrymarshall.com/marketing/10.htm](http://www.perrymarshall.com/marketing/10.htm) - 22k - Cached - Similar pages

No relevant info (Marketing firm page)

**Business 2.0 - Magazine Article - Shelly Lazarus**

... don't think you can **spend** too **much money** on marketing ... **much** ... simulate real-life interaction ... the value of a brand is so **much** ...

[www.business2.com/articles/mag/0,1640,17509,00.html](http://www.business2.com/articles/mag/0,1640,17509,00.html) - 34k - Mar 1, 2002

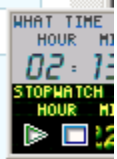
No relevant info (Mag page on ad exec)

[Similar pages](#)

**B2B requires simple marketing materials - 2002-07-29**

... historic first sale to **IBM**, he didn't have a product, **much** less ... B2B companies with a direct sales force **spend** tons of **money** on mark

No relevant info (Mag page on MS-IBM)





ask.

answer.

discover.

Search for questions:

Search

Advanced

My Profile

Home > Travel > Asia Pacific > Singapore



Ready to Participate?  
Get Started!

ADVERTISEMENT



## Categories

All Categories

Travel

Asia Pacific

- China
- Indonesia
- Japan
- Korea
- Malaysia
- Maldives
- Philippines

» Singapore

Taiwan

## Singapore



Top Singapore Answerer

floozy\_niki - Level 6 - 150 Best Answers

Top 10 Answerers

Answer

Open Questions

Discover

Resolved Questions

Vote

Undecided Questions

View by: Date  | No. of Answers



**I go to Singapore in October. I'm from europe. Is it safe to eat in Singapore restaurants?**

1 ☆ In [Singapore](#) - Asked by [mabp](#) - 15 answers - 2 days ago



**Closest MRT stop near ikea?**

☆ In [Singapore](#) - Asked by [meredithk93](#) - 4 answers - 2 days ago



**Singapore??????????????????????...**

☆ In [Singapore](#) - Asked by [s0ck-puPPeT](#) wants YA fixed - 3 answers - 2 days ago



**What can we do in changi airport?**

☆ In [Singapore](#) - Asked by [Tsz Shan](#) - 7 answers - 3 days ago



**Isn't it a Disastrous Olympic Silver Medal Celebration?**

1 ☆ In [Singapore](#) - Asked by [Lin Dan](#) - 1 answer - 3 days ago



**How will you grade the life of a Lecturer/ Professor at National Univ of Singapore in terms of stress, etc?**

☆ In [Singapore](#) - Asked by [priyanka b](#) - 2 answers - 4 days ago



**Is the national animal/icon of Singapore a lion or a MERlion?**

☆ In [Singapore](#) - Asked by [Kush W](#) - 2 answers - 3 days ago

- Yah
- Bai
- Na
- Go
- Ma

ug,

# ELECTION 2008 & Yahoo! Answers



Hillary Clinton

OFFICIAL

Based on your own family's



John McCain

OFFICIAL

What would you do to stop wasteful government spending in Washington?

## Additional Details

6 months ago

Yahoo! Answers Staff Note: Watch a video of John McCain discussing government spending.

<http://video.yahoo.com/video/play?vid=f3...>

6 months ago

Yahoo! Answers Staff note: Yahoo! Answers is a forum for people from all over the world to engage with one another and to find information on topics that interest them. This is not an endorsement. We are not siding with any candidate or party -- in general or for the 2008 US elections. We're hopeful that people from all perspectives will realize the great insights that the Answers community can have, and will turn to us for future discussions.

5 months ago

Yahoo! Answers staff note: Read Sen. John McCain's own response to the Yahoo! Answers community.

<http://blog.360.yahoo.com/blog-d8ph0dcor...>

6 months ago **16713 answers** Report Abuse



Barack Obama

How can we engage more people in the democratic process?

I will be asking questions to help create dialogue around this and many other important topics so please add me to

Network so that we can begin exchanging hopefully make changes that will benefit the

## Details

ers Staff note: Yahoo! Answers is a forum for all over the world to engage with one another on topics that interest them. This is not an endorsement. We are not siding with any candidate in general or for the 2008 US elections. We're hopeful that people from all perspectives will realize the great insights that the Answers community can have, and will turn to us for future discussions.

I was deeply impressed by the thoughtfulness of the answers to my question. There is, of course, no single answer to how we all need to do more work on engaging more people to participate in the process and ensuring every community is included in the discussion. Hundreds of worthy and thought-provoking answers, in fact, some were so intriguing that I arranged to meet personally with their creators. To hear my conversation with one user, who wrote about the parallels between the Internet and democracy, click below.

<http://video.yahoo.com/video/play?vid=74...>

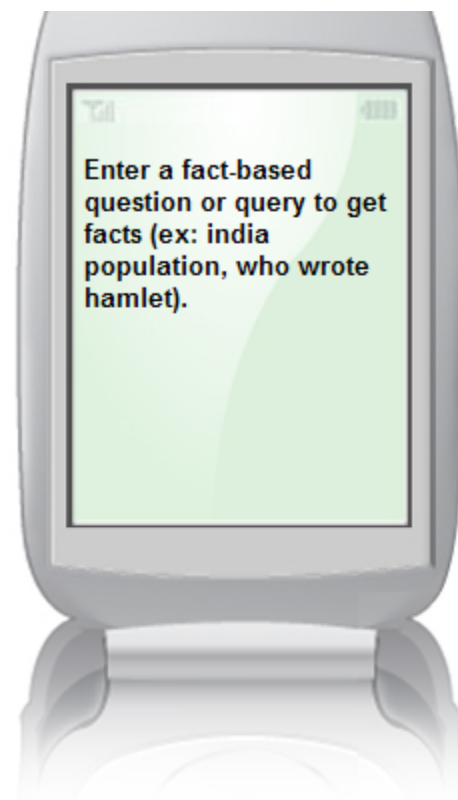
17382 answers Report Abuse

Enter a search term and click on the links under "Search Feature" in the table below to find specific information about those features, or click on the links in the "Sample Query" column to view sample search results.

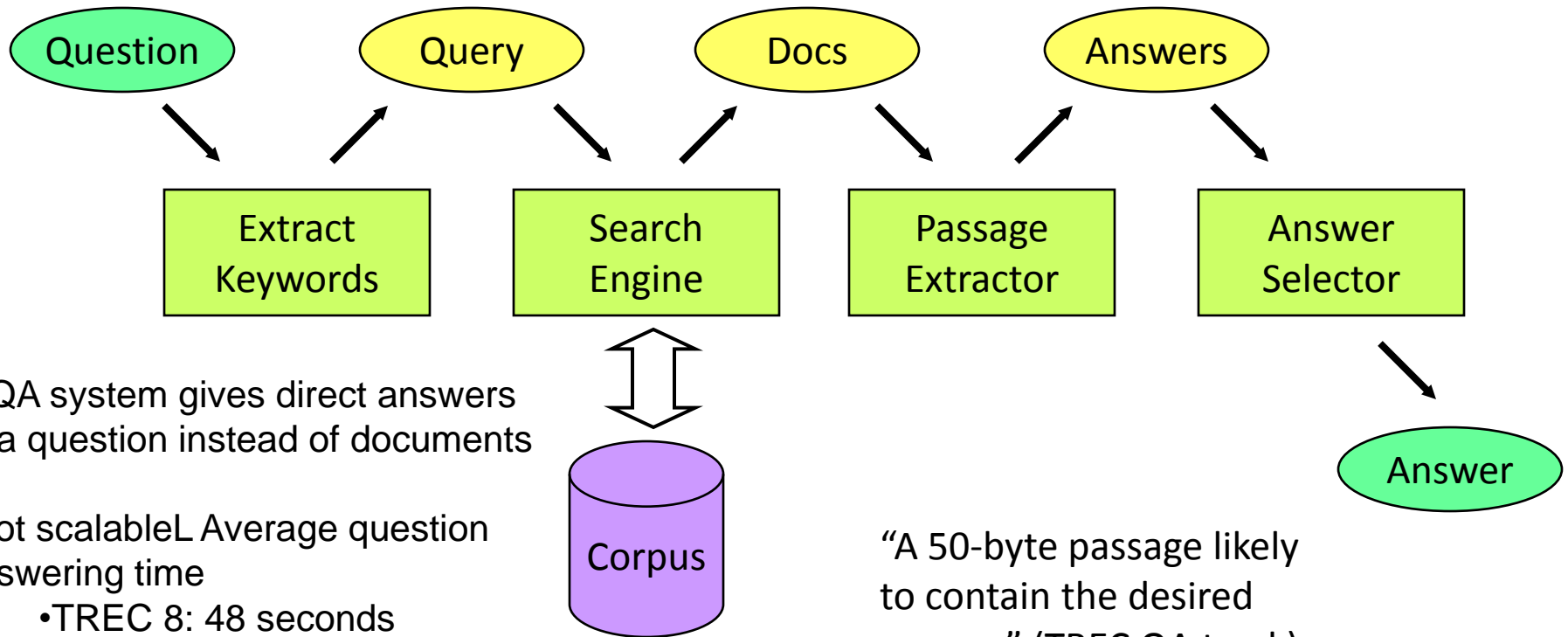
<b>Personalization</b>	
<a href="#">Change personalization setting</a>	<a href="#">save on</a> / <a href="#">save off</a> / <a href="#">clear</a>
<a href="#">Set / view / clear your location</a>	<a href="#">set location 10012</a> / <a href="#">view location</a> / <a href="#">clear location</a>
Search Queries Personalized*	<a href="#">pizza</a> / <a href="#">weather</a> / <a href="#">movies</a>

Search Feature	Sample Query
<a href="#">Local</a>	<a href="#">sushi 94040</a>
<a href="#">Weather</a>	<a href="#">weather boston</a>
<a href="#">Glossary</a>	<a href="#">define zenith</a>
<a href="#">Sports</a> **	<a href="#">score red sox</a>
<a href="#">Movies</a>	<a href="#">movies 94110</a>
<a href="#">Stocks</a>	<a href="#">stock tqf</a>
<a href="#">Zip Codes</a>	<a href="#">zip code 72202</a>
<a href="#">Directions</a>	<a href="#">directions pasadena ca to 94043</a>
<a href="#">Maps</a>	<a href="#">map 5th avenue new york</a>
<a href="#">Flights</a> ***	<a href="#">flight aa 2111</a>
<a href="#">Area Codes</a>	<a href="#">area code 650</a>
<a href="#">Products</a>	<a href="#">price ipod player 40gb</a>
<a href="#">Q&amp;A</a>	<a href="#">abraham lincoln birthday</a>
<a href="#">Airlines</a> ***	<a href="#">united airlines</a>
<a href="#">Translation</a>	<a href="#">translate hello in french</a>
<a href="#">Web Snippets</a>	<a href="#">web hubble telescope</a>
<a href="#">Calculator</a>	<a href="#">1 us pint in liters</a>
<a href="#">Currency Conversion</a>	<a href="#">8 usd in yen</a>



# Typical TREC QA Pipeline

“A simple factoid question”

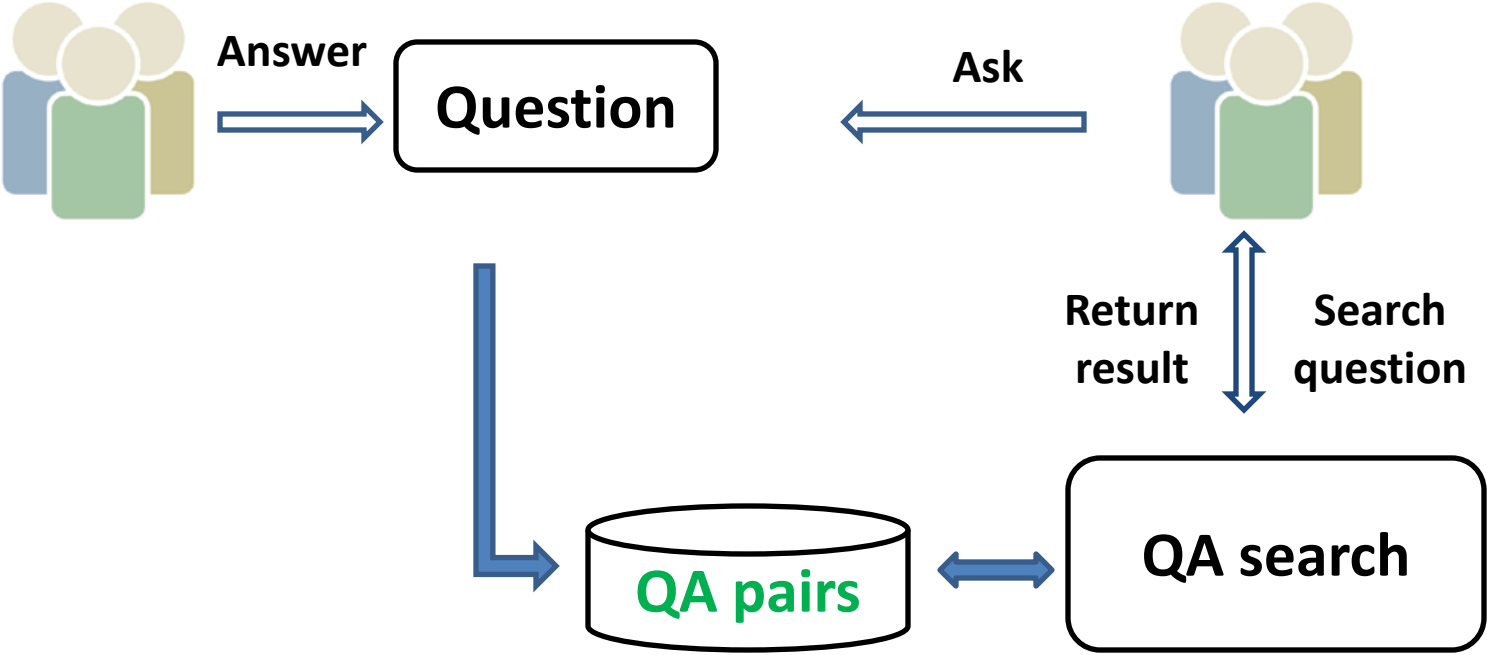


A QA system gives direct answers to a question instead of documents

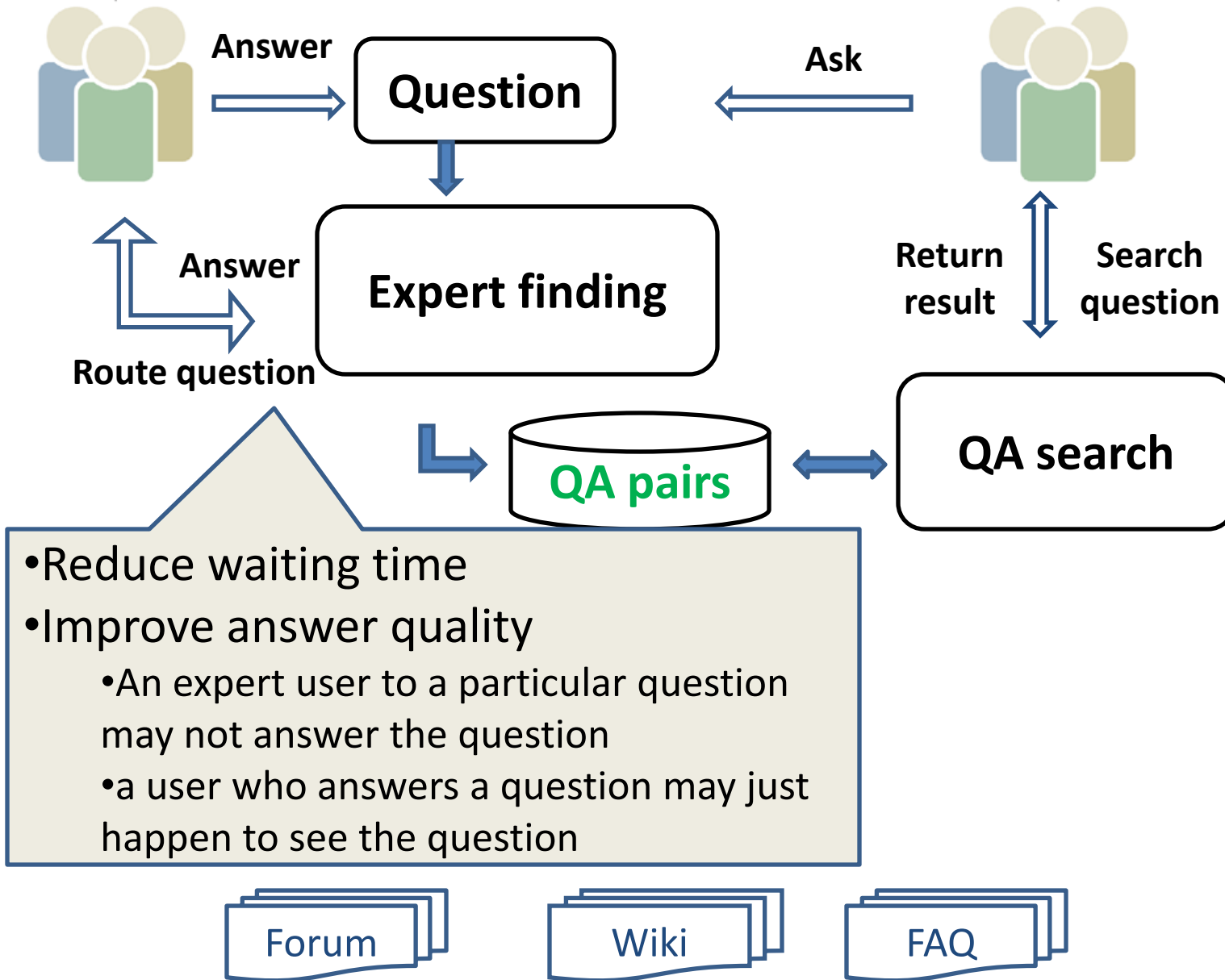
- Not scalable
- Average question answering time
  - TREC 8: 48 seconds
  - TREC 9: 94 seconds
- Process limited types of questions

“A 50-byte passage likely to contain the desired answer” (TREC QA track)

# CURRENT Community QA ARCHITECTURE



# PROPOSED CQA ARCHITECTURE




## weather in July and olympic games souvenirs

[el.di.](#)

Greece

Joined: Aug 2006

Forum posts: 137

Travel map pins: 69 


[More about el.di....](#)



Travel IQ: ?

REPLY TO THIS POST

Posted on: 8:39 am, July 08, 2008

 Save

Hi,

Q1

I know that weather in beijing in july is warm, humid and rainy, is that true? What clothes should we pack?

Q2

Q3

Also where can i find olympic games souvenirs to bring home?

Thank you

[Report inappropriate post](#)



[travellevret](#)

Joined: Jan 2005

Forum posts: 2,228

Travel map pins: 345

*destination expert*


for Beijing



Travel IQ: 131

[More about travellevret...](#)

Posted on: 9:57 am, July 08, 2008

 Save

A1,2

Shorts, T-shirts--clothes good for hot weather. July is the rainiest month, but not all that wet. You should pack raingear.

A3

You will no doubt come across Olympics stores without trying. The airport has stores, there must be one on Wangfujing. Some museum shops have them. Not to worry--their goal is to make sure you have plenty of opportunity to buy! Selection and prices are similar at all of the stores.

[Report inappropriate post](#)

Problem: find questions and answers in a forum thread

# PROBLEM AND MOTIVATIONS

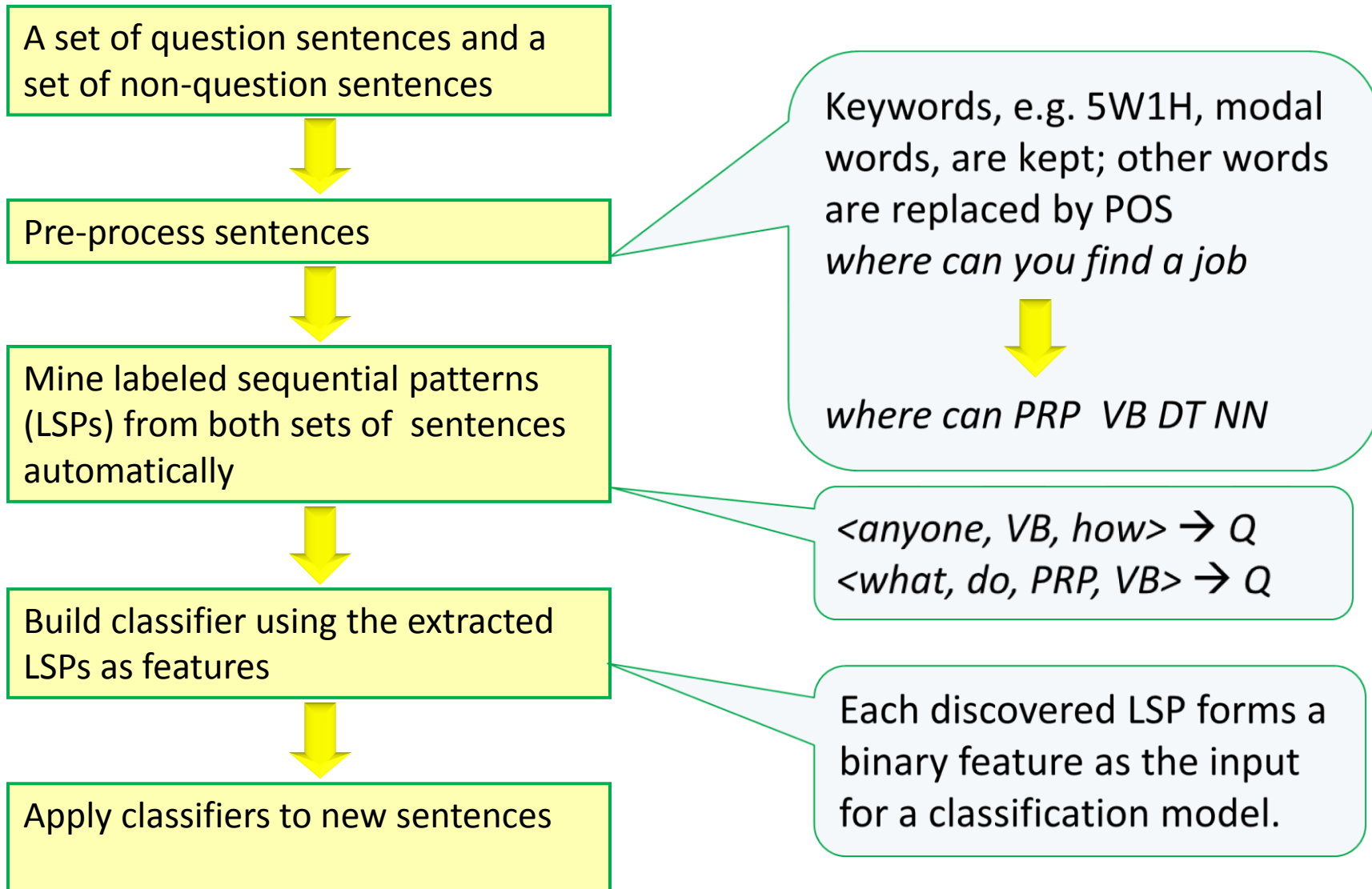
- Problem: Find questions, context and answers in forums
- Extract knowledge in QA form from forums
  - Large amount of existing QA pairs in forums
    - TripAdvisor: ~3M QA pairs (Jan 2008)
    - Yahoo! Answer Travel: ~700K (Jan 2008)
- Provide effective and efficient search & browsing UI
- Organize and manage forum information
  - Highlight the main content of a thread
  - Index question and answer pairs



# MAIN CHALLENGES: QUESTION DETECTION

- Simple rules are not adequate.
  - 30% questions have no ‘?’ (out of 1,000).
  - 9% sentences with ‘?’ are not questions.
- Questions can be expressed in imperative form.
  - *“I am wondering where I can buy cheap and good clothing in beijing.”*
- Question mark is often omitted in forums.
- Short informal expressions, such as “really?”, are not questions.

# QUESTION DETECTION



# LABELED SEQUENTIAL PATTERN(LSP)

- LSP: sequence  $\rightarrow$  label
  - E.g. question: *i want to buy an office software and wonder which software company is best.*
  - *wonder which...is*  $\rightarrow$  question
- Two measures of LSP: a labeled sequence database  $D$  and a LSP  $p$ 
  - **Support** of  $p$ : the percentage of labeled sequences in  $D$  that contain  $p$ .

$$\text{sup}(p) = \frac{|\{s \mid s \in D, p \subseteq s\}|}{|D|}$$

- **Confidence** of  $p$ : the probability of the sequence of  $p$  being true

$$\text{conf}(p) = \frac{\text{sup}(p)}{\text{sup}(p.\text{sequence})}$$

- Mine LSP given minsup and minconf threshold

Labeled sequence database:

S1(Q): a d e f

S2(Q): a f e f

S3(NQ): d a f

LSP1:  $\langle a, e, f \rangle \rightarrow Q$

S1,S2 contain  $\langle a,e,f \rangle \rightarrow Q$

S1,S2 contain  $\langle a,e,f \rangle$

Sup(LSP1)=2/3 Conf(LSP1)=2/2

LSP2:  $\langle a, f \rangle \rightarrow Q$

S1,S2 contain  $\langle a,f \rangle \rightarrow Q$

S1,S2,S3 contain  $\langle a,f \rangle$

Sup(LSP2)=2/3 Conf(LSP2)=2/3

# QUESTION DETECTION

Data	Method	Prec(%)	Rec(%)	F <sub>1</sub> (%)
Q-TUnion	5W-1H words	69.0	14.8	24.4
	Question Mark	96.8	78.4	86.6
	SM [18]	81.9	87.8	84.6
	Our	96.5	98.5	97.5
Q-TInter	5W-1H words	69.0	15.3	25.0
	Question Mark	98.7	77.6	86.9
	SM [18]	92.7	86.8	89.7
	Our	97.8	97.0	97.4

- Ours and SM use Ripper classifier but different set of features
- 2,316 LSPs for questions (1,074 contains “?”) and 2,789 for non-questions

# MAIN CHALLENGES

- Answer detection
  - Multiple QA threads overlap and exist in parallel.
  - Reply relationship between posts is not explicit.
  - A post may answer many questions.
  - A question may have multiple answers.

# ANSWER DETECTION

- Problem: Given a question and a set of candidate answers, rank
- Benchmark methods
  - Cosine similarity
  - KL divergence language model
  - Query likelihood language model
  - Classification based re-ranking

# UNSUPERVISED APPROACH

- Our improvements: Make use of the **dependency** relationships between answers to improve the benchmark methods
  - if a candidate answer is related to an authoritative candidate answer with high score, the candidate answer is likely to be an answer
  - Build a graph on candidate answers of a question, rerank based on the graph
    - Build graph: using asymmetric generation probability
      - Assign weight:
      - Compute reranking score:

# ANSWER DETECTION ON A-TUNION DATA

Method	Question with answer		
	P@1	MRR	MAP
NA	0.644	0.718	0.618
Lex	0.649	0.756	0.721
Cla	0.722	0.818	0.774
CS	0.686	0.789	0.737
QL	0.697	0.791	0.719
KL	0.709	0.809	0.762
G+CS	0.739	0.830	0.784
G+QL	0.761	0.843	0.775
G+KL	<b>0.816</b>	<b>0.882</b>	<b>0.842</b>

- G+KL performs 15% better than KL



# OPINION MINING

- Problems & Application
- Approaches

# Examples

- **Product review mining:** What features of the ThinkPad T43 do customers like and which do they dislike?
- **Review classification:** Is a review positive or negative toward the movie?
- **Tracking sentiments toward topics over time:** Is anger ratcheting up or cooling down?
- Search engines do not search for opinions
  - Opinions are hard to express with a few keywords
  - How do people think of Motorola Cell phones?
  - Current search ranking strategy is not appropriate for
  - opinion retrieval/search

# Application

- **Businesses and organizations:** product and service benchmarking. Market intelligence.
  - Business spends a huge amount of money to find consumer sentiments and opinions.
    - Consultants, surveys and focused groups, etc
- **Individuals:** interested in other's opinions when
  - Purchasing a product or using a service,
  - Finding opinions on political topics,
- **Ads placements:** Placing ads in the user-generated content
  - Place an ad when one praises a product.
  - Place an ad from a competitor if one criticizes a product.
- **Opinion retrieval/search:** providing general search for opinions.



## Products

See also: [Web](#), [Images](#), [Video](#), [News](#), [Maps](#), [More](#) ▼

### Apple iPhone Smart Phone



\$499 - \$849 [Compare prices](#) (6)

★★★★☆ [User reviews](#) (1233)

★★★★☆ [Expert reviews](#) (1)

Quad Band - GSM 800, GSM 900, GSM 1800, GSM 1900 - Bluetooth, Wi-Fi - EDGE, GPRS - Polyphonic - 16.7 Million Colors - 4GB - Bar - Smartphone

[User reviews](#) | [Product details](#) | [Expert reviews](#) | [Compare prices](#)

[All user reviews](#)

General Comments (497 comments)

85% positive

[Ease Of Use](#) (217 comments)

93% positive

[Features](#) (132 comments)

93% positive

[Affordability](#) (126 comments)

52% positive

[Screen](#) (70 comments)

80% positive

[Appearance](#) (64 comments)

95% positive

### General Comments

View by: [Positive comments](#) (85%) | [Negative comments](#) (15%)

The Iphone is the best cell phone I had so far and i think that if you buy it you will love it. [More...](#)

ruben\_93 [catalog.ebay.com](#) 8/23/2008

I have decided to buy it because I like it very much and good product the apple iphone 16GB, has very good quality [More...](#)

davidxerach [catalog.ebay.com](#) 5/30/2008

We are very happy with the item, it is really a great product. [More...](#)

3703guns [catalog.ebay.com](#) 5/7/2008

Overall this is great item and I highly recommend this device to travelers. [More...](#)

clink1381 [catalog.ebay.com](#) 4/26/2008

Sponsored sites

[iPhone 3G - Official Site](#)

Watch the new iPhone TV ads featuring the App store today.

[www.apple.com/iphone](http://www.apple.com/iphone)

[Save on iPhone](#)

Shop for iPhone and save at Live Search cashback!

[Search.Live.com/cashback](http://Search.Live.com/cashback)

Live Search cashback

[Iphone' at Amazon](#)

Low prices on iPhone'. Qualified orders over \$25 ship free

[Amazon.com](http://Amazon.com)

[iphone](#)

Find Low Prices and Multiple Offers On iPhone

[shopping.yahoo.com](http://shopping.yahoo.com)

[By Iphone](#)

Explore 5,000+ Cell Phones & Gear. Save On By iPhone!

[CellPhones.Shopzilla.com](http://CellPhones.Shopzilla.com)

[See your message here...](#)

# General approach

- Identify features
  - Useful to identify opinion (positive, negative, neutral)
- Use classification approach or unsupervised approach to identify polarity (sentiment)
  - document level
  - Sentence level
  - Feature level

# Opinion Words or Phrases

- **Find** relevant words, phrases, patterns that can be used to express subjectivity
- **Determine** the polarity of subjective expressions
  - also called polar words, opinion bearing words, etc, e.g.,
  - Positive: beautiful, wonderful, good, amazing, *Ron Paul is the only **honest** man in Washington*
  - Negative: bad, poor, terrible.
  - **curious, peculiar, odd, likely, probable**, *The two species are **likely** to flower at different times*

# Three main ways to compile a list of opinion words

- Manual approach: not a bad idea, only an one-time effort
- Corpus-based approaches
  - This car is *beautiful and spacious*
- Dictionary-based approaches
  - WordNet, e.g. synonyms and antonyms

# Domain and context dependent

- **Domain** dependent
  - “go read the book” most likely indicates positive sentiment; for book reviews, but negative sentiment for movie reviews
  - Unpredictable for movie plot, for car’s steering abilities
- **Contextual** Polarity :
  - Some opinion words are context independent (e.g., good).
  - Some are context dependent (e.g. long)



# Example

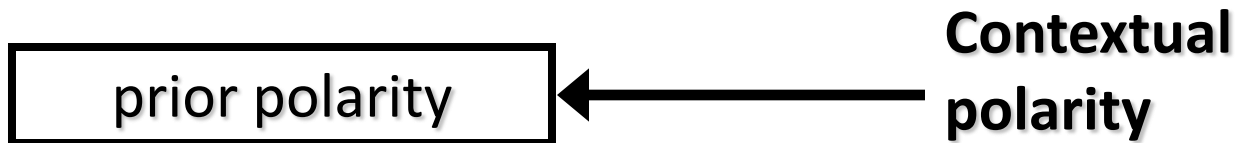
Philip Clap, President of the National Environment Trust, sums up well the general thrust of the reaction of environmental movements: there is no reason at all to believe that the polluters are suddenly going to become reasonable.

# Example

Philip Clap, President of the National Environment **Trust**, sums up **well** the general thrust of the reaction of environmental movements: there is no **reason** at all to believe that the **polluters** are suddenly going to become **reasonable**.

# Example

Philip Clap, President of the National Environment ~~Trust~~, sums up well the general thrust of the reaction of environmental movements: there is no reason at all to believe that the ~~polluters~~ are suddenly going to become reasonable.



# SUMMARY

- Background on classification and clustering
- Basic knowledge on search engines
  - Crawler, indexer, ranker
- Question Answering and QA extraction
  - QA detection in forum threads
- Opinion Mining