Faceted search

Outline

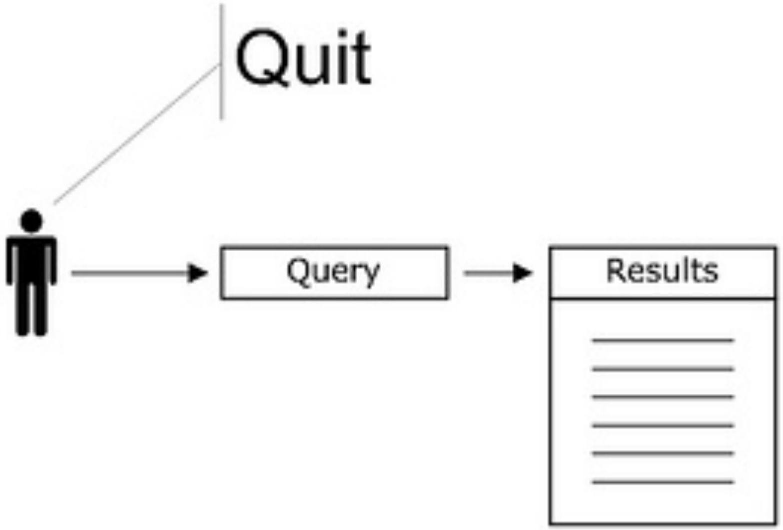
- Exploratory search and ways to support it
- Faceted search:
  - Interfaces
  - Interaction styles
- Faceted search solutions:
  - with structured metadata
  - with unstructured metadata
  - without ready-made metadata
- Future challenges

Users demand: explore

more **control** over search!

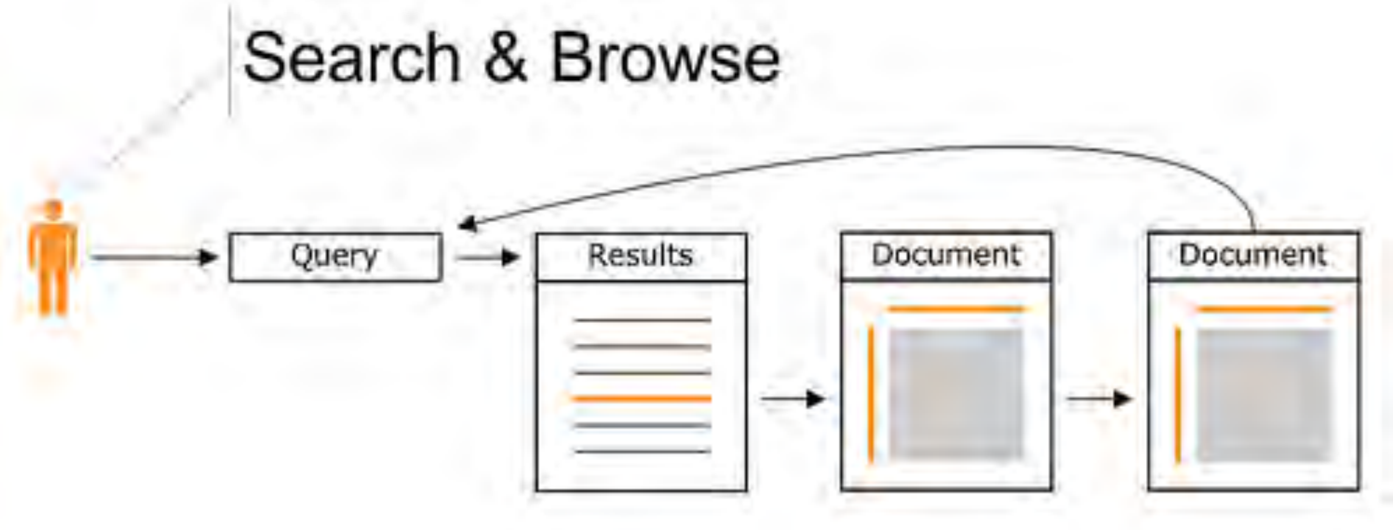
They want to **explore**!

Search is a look-up?



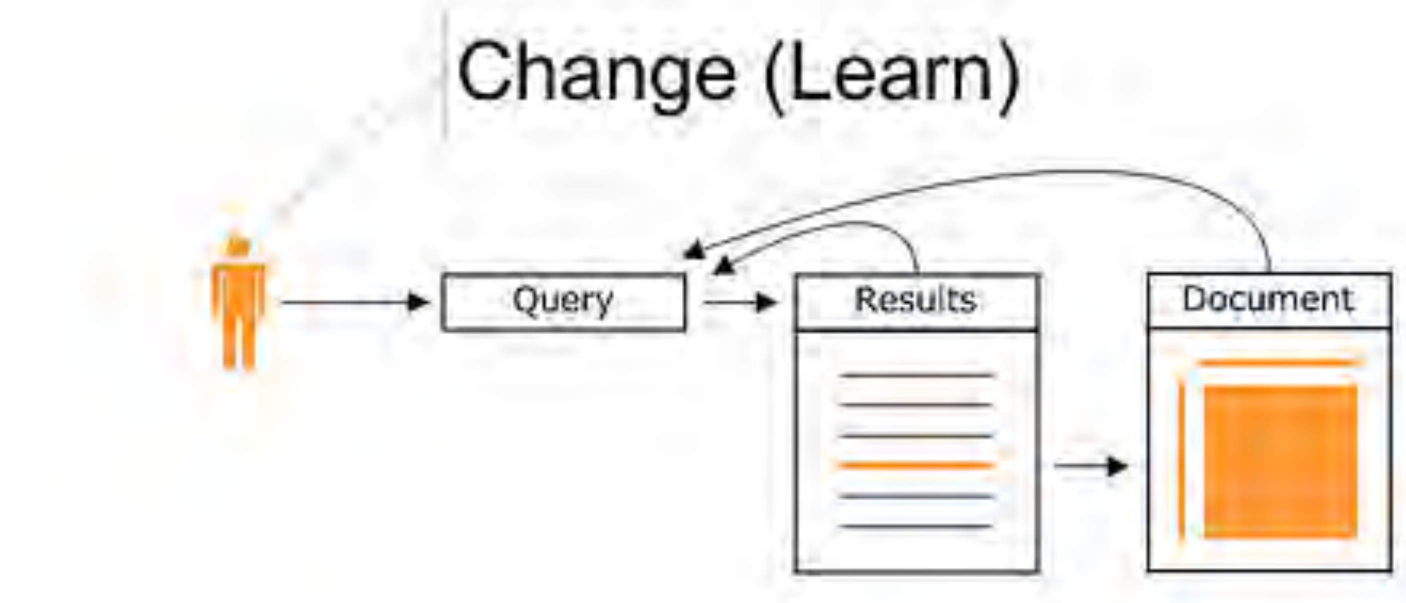
Is that all?

Search is a journey!



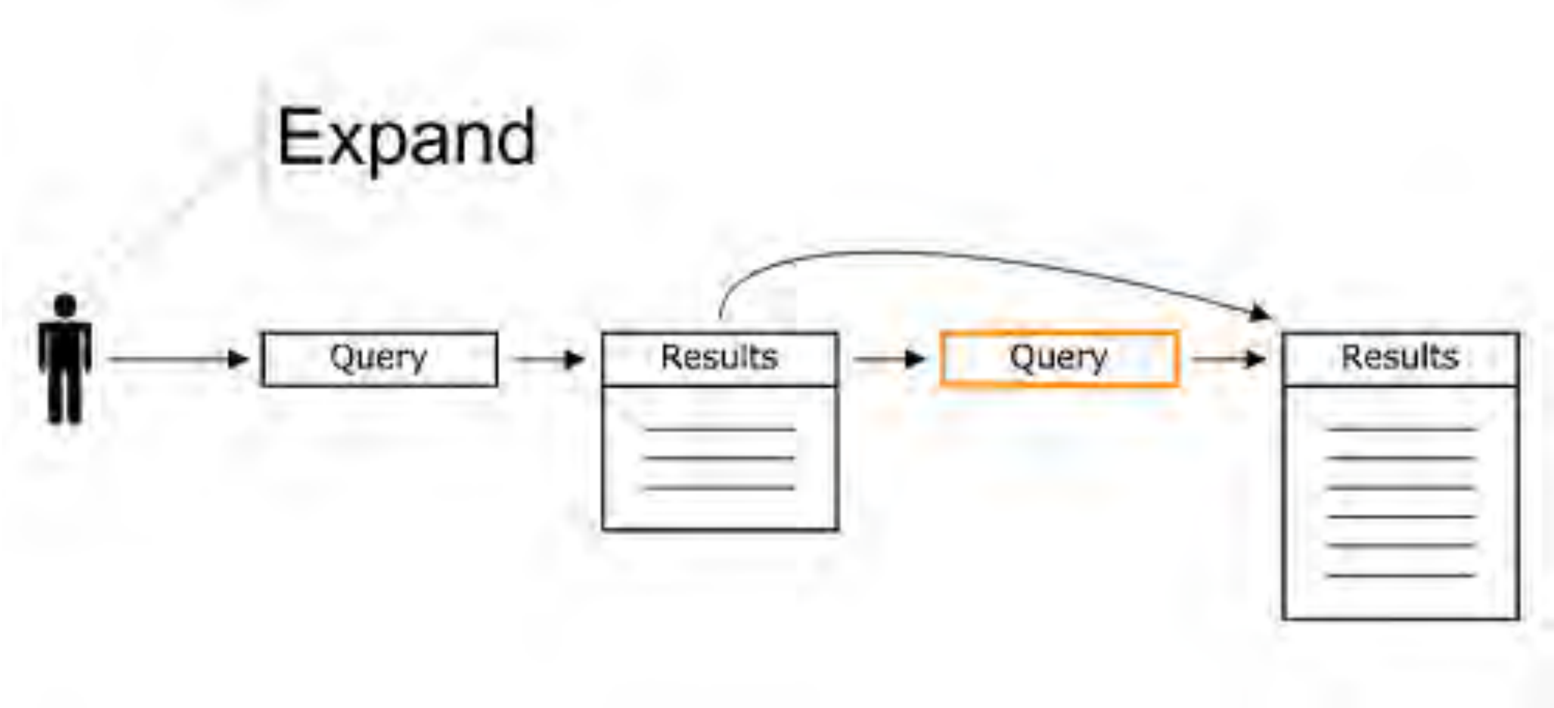
- Exploratory search involves:
  - – browsing the result
  - – analyzing returned documents
  - – coming back to the initial ranking again and again

Search is a journey!



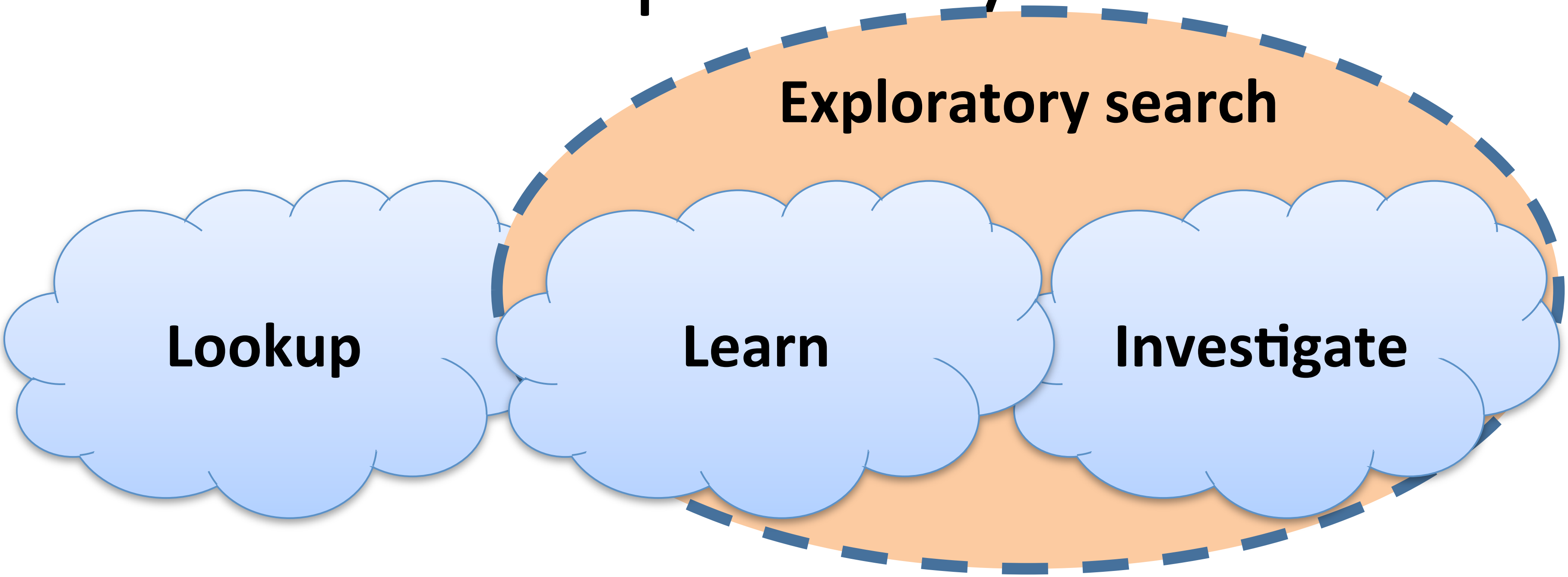
- Exploratory search involves:
  – Querying the last returned result set
  – Looking for similar documents (relevance feedback)

Search is a journey!



- Exploratory search is also about…
  - Query reformulation, same information need:
    - Specialization: **mp3 players** => **ipod**
    - Generalization: **ipod** => **mp3 players**

What is exploratory search

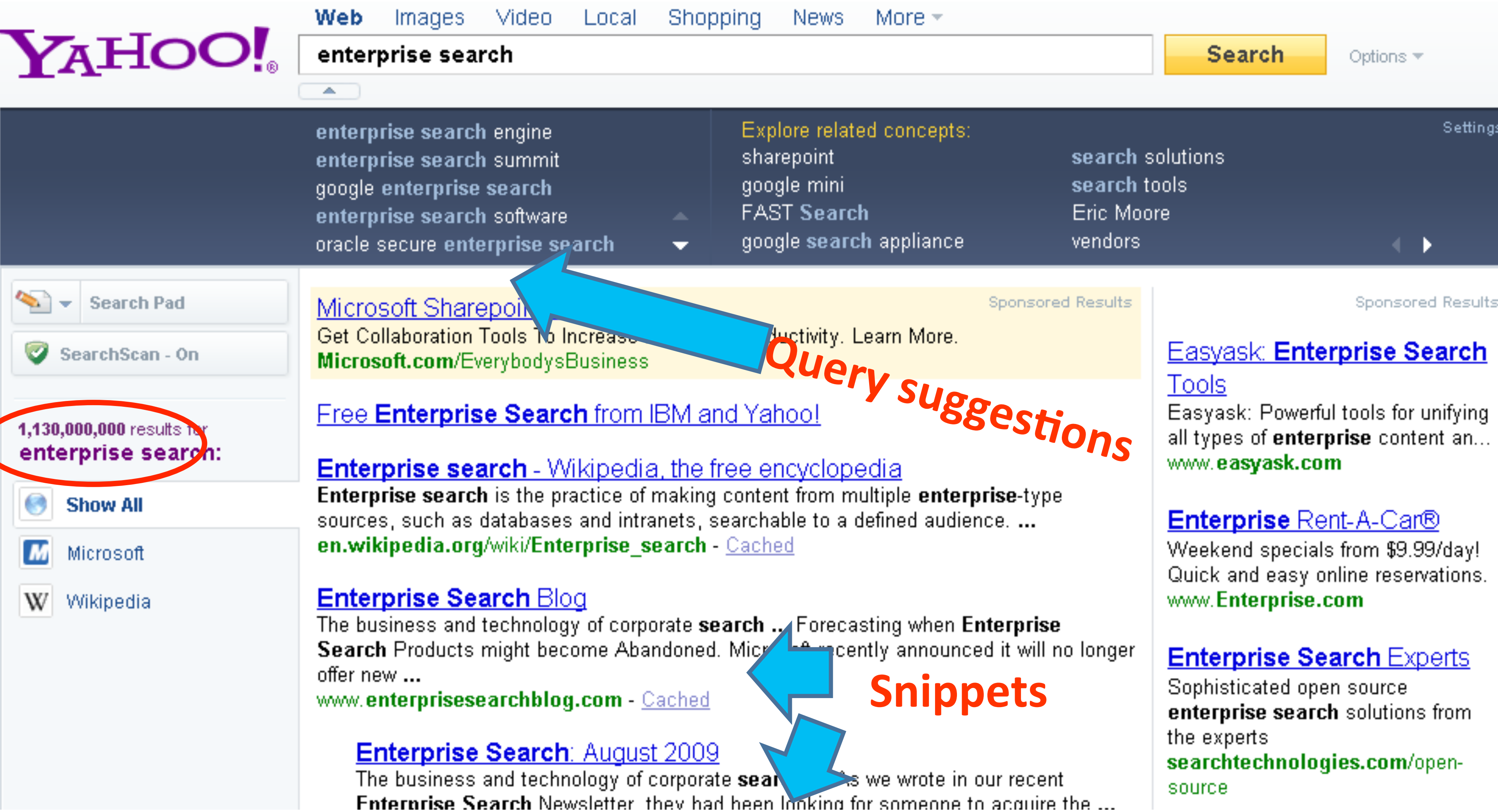


Question answering
Fact retrieval
Known-item search
Navigational search
Lasts for seconds

Knowledge acquisition
Comprehension
Comparison
Discovery
Serendipity

Incremental search
Driven by uncertainty
Non-linear behavior
Result analysis
Lasts for hours

What web search engines offer



enterprise search

enterprise search engine
enterprise search summit
google enterprise search
enterprise search software
oracle secure enterprise search

Explore related concepts:

sharepoint	search solutions
google mini	search tools
FAST Search	Eric Moore
google search appliance	vendors

1,130,000,000 results for
enterprise search:

Microsoft Sharepoi
Get Collaboration Tools To Increase ... uctivity. Learn More.
Microsoft.com/EverybodysBusiness

Free Enterprise Search from IBM and Yahoo!

Enterprise search - Wikipedia, the free encyclopedia
Enterprise search is the practice of making content from multiple enterprise-type sources, such as databases and intranets, searchable to a defined audience. ...
en.wikipedia.org/wiki/Enterprise_search - Cached

Enterprise Search Blog
The business and technology of corporate search ... Forecasting when Enterprise Search Products might become Abandoned. Mic ... ecently announced it will no longer offer new ...
www.enterprisesearchblog.com - Cached

Enterprise Search: August 2009
The business and technology of corporate sear ... s we wrote in our recent Enterprise Search Newsletter, they had been looking for someone to acquire the ...

Easyask: Enterprise Search Tools
Easyask: Powerful tools for unifying all types of enterprise content an...
www.easyask.com

Enterprise Rent-A-Car®
Weekend specials from $9.99/day! Quick and easy online reservations.
www.Enterprise.com

Enterprise Search Experts
Sophisticated open source enterprise search solutions from the experts
searchtechnologies.com/open-source

Query suggestions

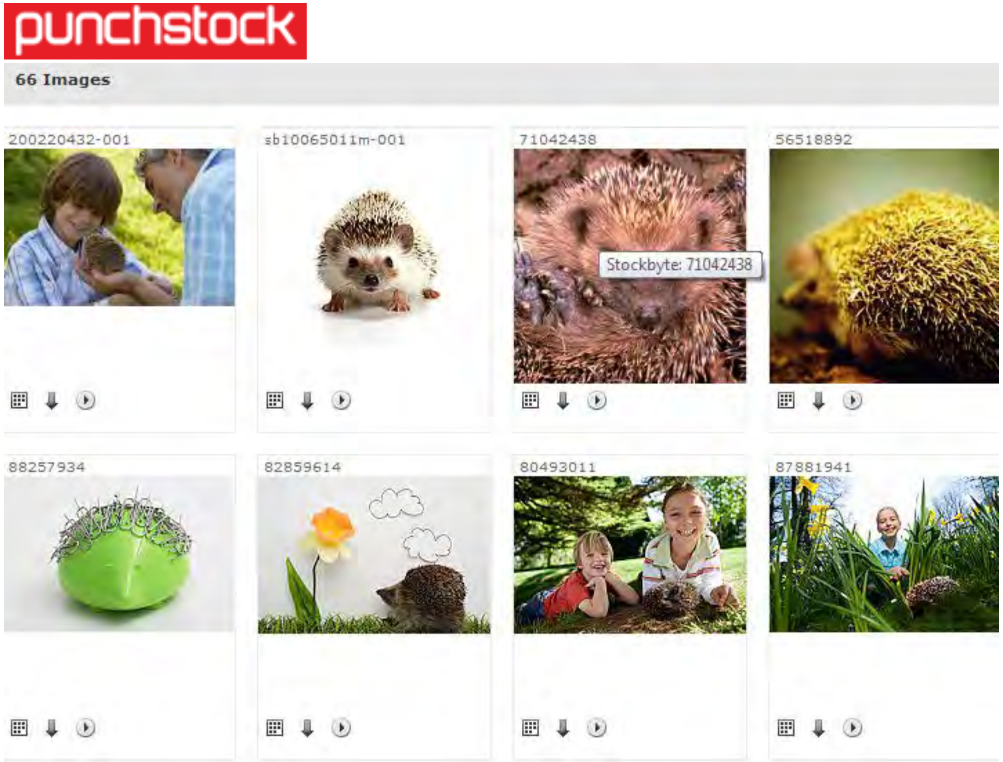
Snippets

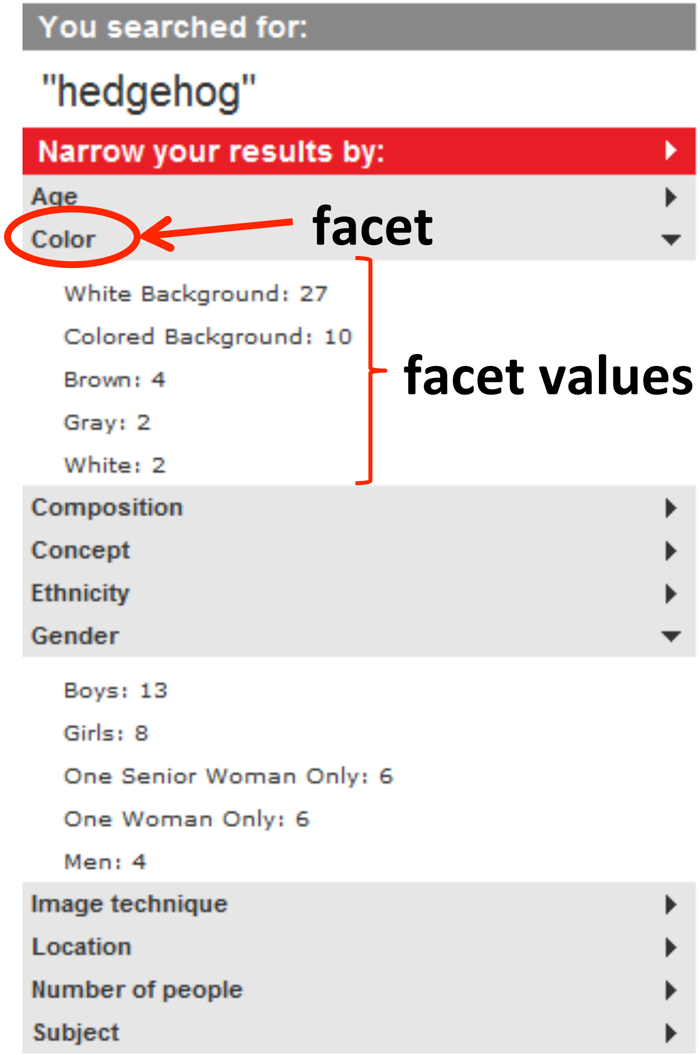
Can we do better?

- Certainly, when we have metadata for docs!
  - So, some summarization is done for us
- **Structured metadata:**
  - **Classic faceted search scenario**
- Unstructured metadata
  - Tag-based analysis and navigation
- No metadata?
  - Result clustering
  - More? Let' s see…

Faceted search:
with structured metadata

What is faceted search?




You searched for:

"hedgehog"

Narrow your results by:

Age

Color

- White Background: 27
- Colored Background: 10
- Brown: 4
- Gray: 2
- White: 2

Composition

Concept

Ethnicity

Gender

- Boys: 13
- Girls: 8
- One Senior Woman Only: 6
- One Woman Only: 6
- Men: 4

Image technique

Location

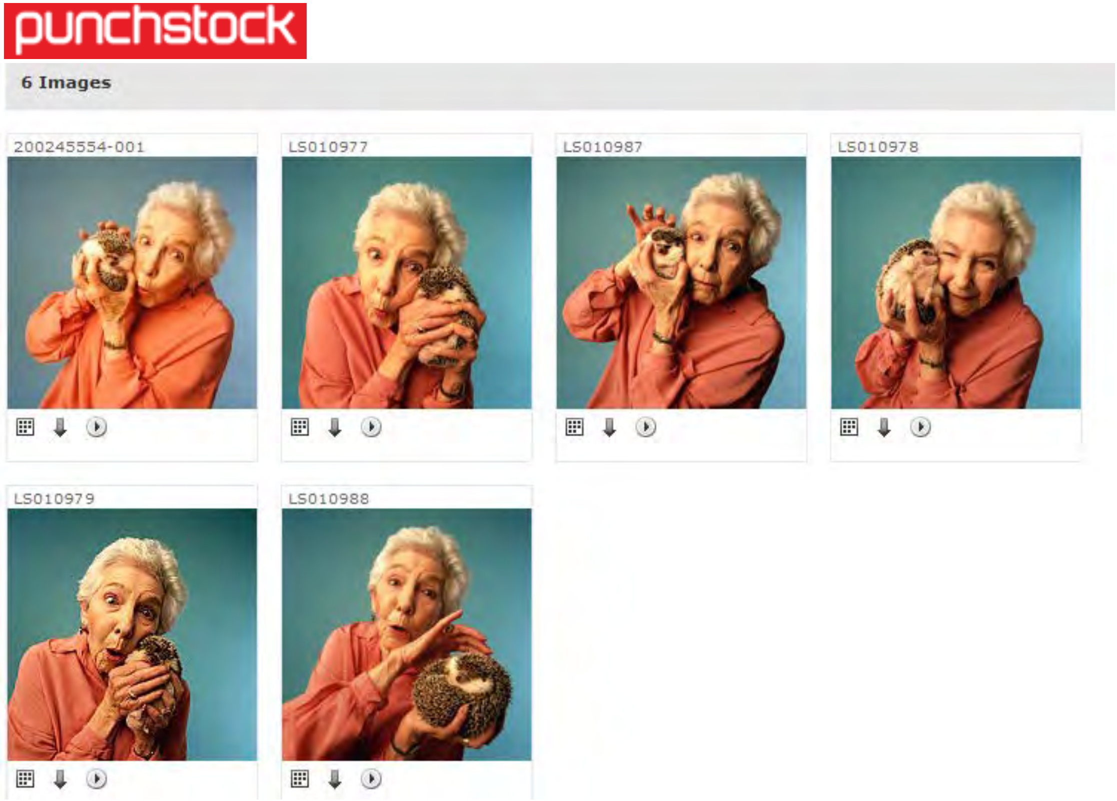
Number of people

Subject

facet

facet values

What is faceted search?



punchstock

6 Images

200245554-001

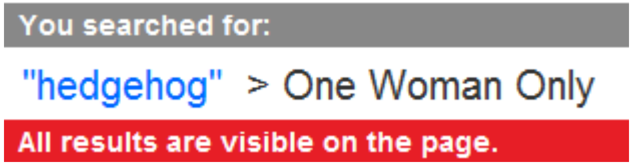
LS010977

LS010987

LS010978

LS010979

LS010988



You searched for:

"hedgehog" > One Woman Only

All results are visible on the page.

It’ s about

Query

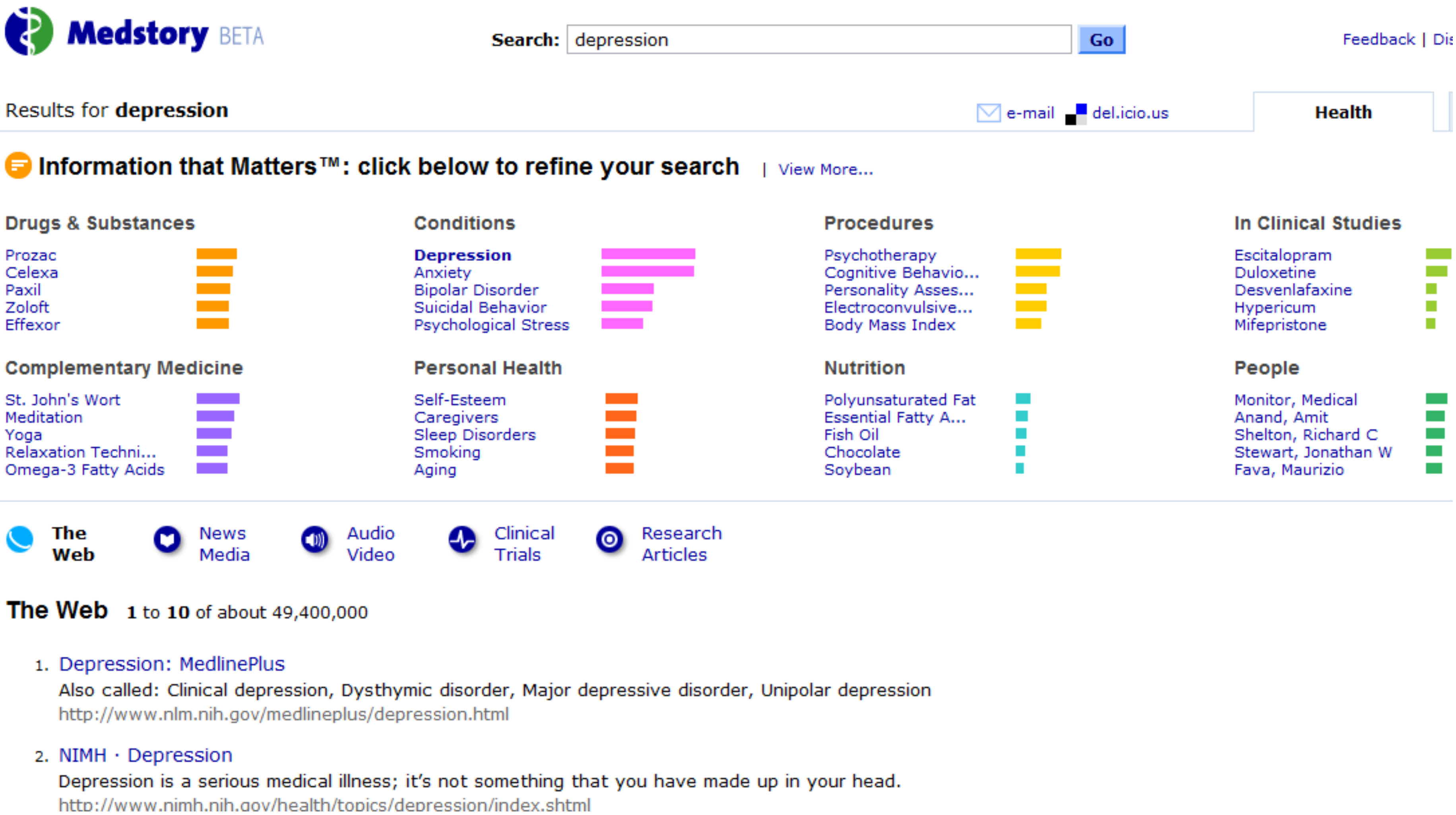
Reformulation!

Faceted search as **query reformulation**

- Traditional way:
  - Typing, typing, typing…
  - For the sake of query reformulation
- Faceted (exploratory) search?

Mousing & Browsing

What is faceted search?



Medstory BETA

Search: depression Go

Feedback | Dis

Results for **depression**

e-mail del.icio.us **Health**

Information that Matters™: click below to refine your search | View More...

Drugs & Substances
- Prozac
- Celexa
- Paxil
- Zoloft
- Effexor

Conditions
- **Depression**
- Anxiety
- Bipolar Disorder
- Suicidal Behavior
- Psychological Stress

Procedures
- Psychotherapy
- Cognitive Behavio...
- Personality Asses...
- Electroconvulsive...
- Body Mass Index

In Clinical Studies
- Escitalopram
- Duloxetine
- Desvenlafaxine
- Hypericum
- Mifepristone

Complementary Medicine
- St. John's Wort
- Meditation
- Yoga
- Relaxation Techni...
- Omega-3 Fatty Acids

Personal Health
- Self-Esteem
- Caregivers
- Sleep Disorders
- Smoking
- Aging

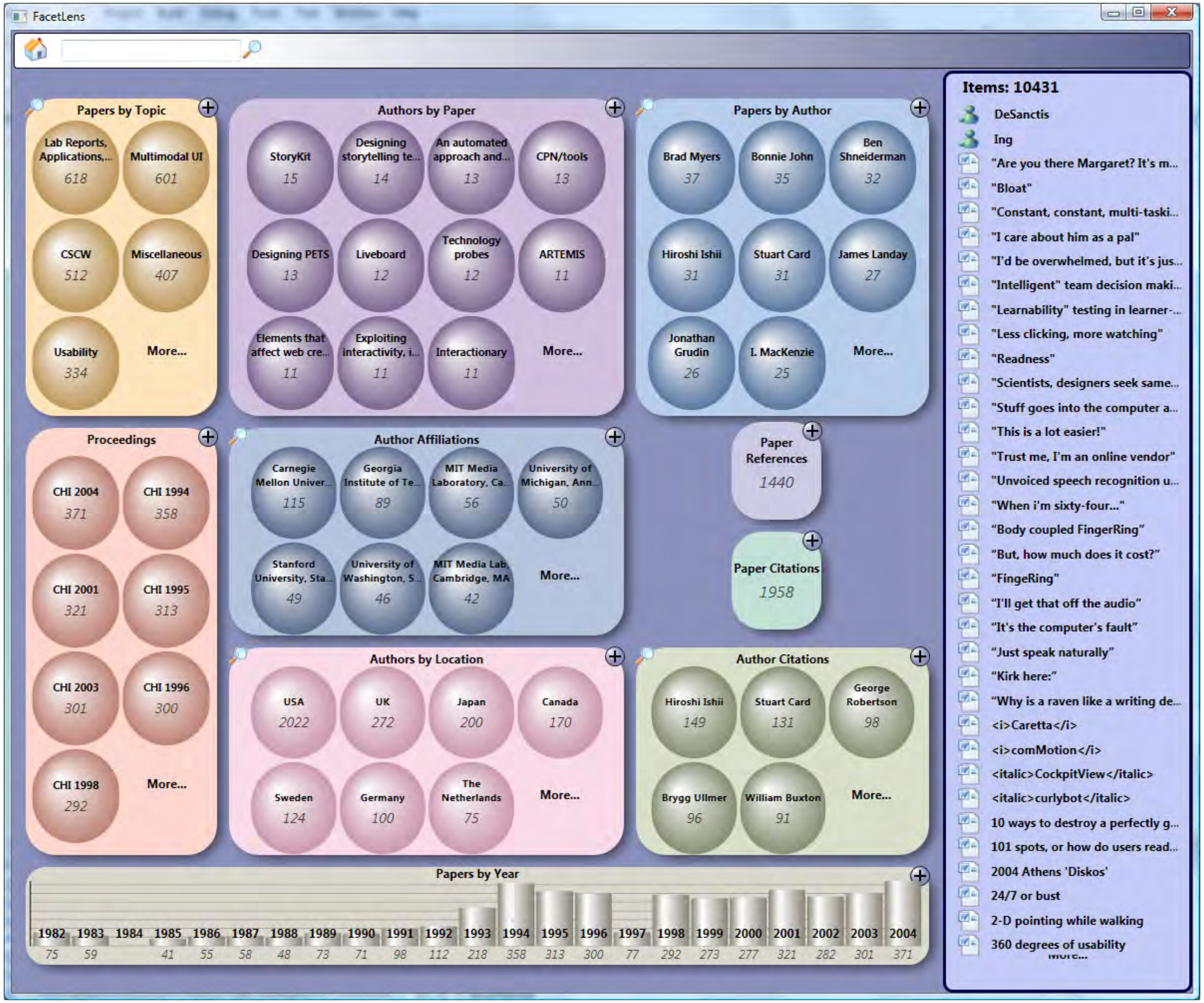
Nutrition
- Polyunsaturated Fat
- Essential Fatty A...
- Fish Oil
- Chocolate
- Soybean

People
- Monitor, Medical
- Anand, Amit
- Shelton, Richard C
- Stewart, Jonathan W
- Fava, Maurizio

The Web **1** to **10** of about 49,400,000

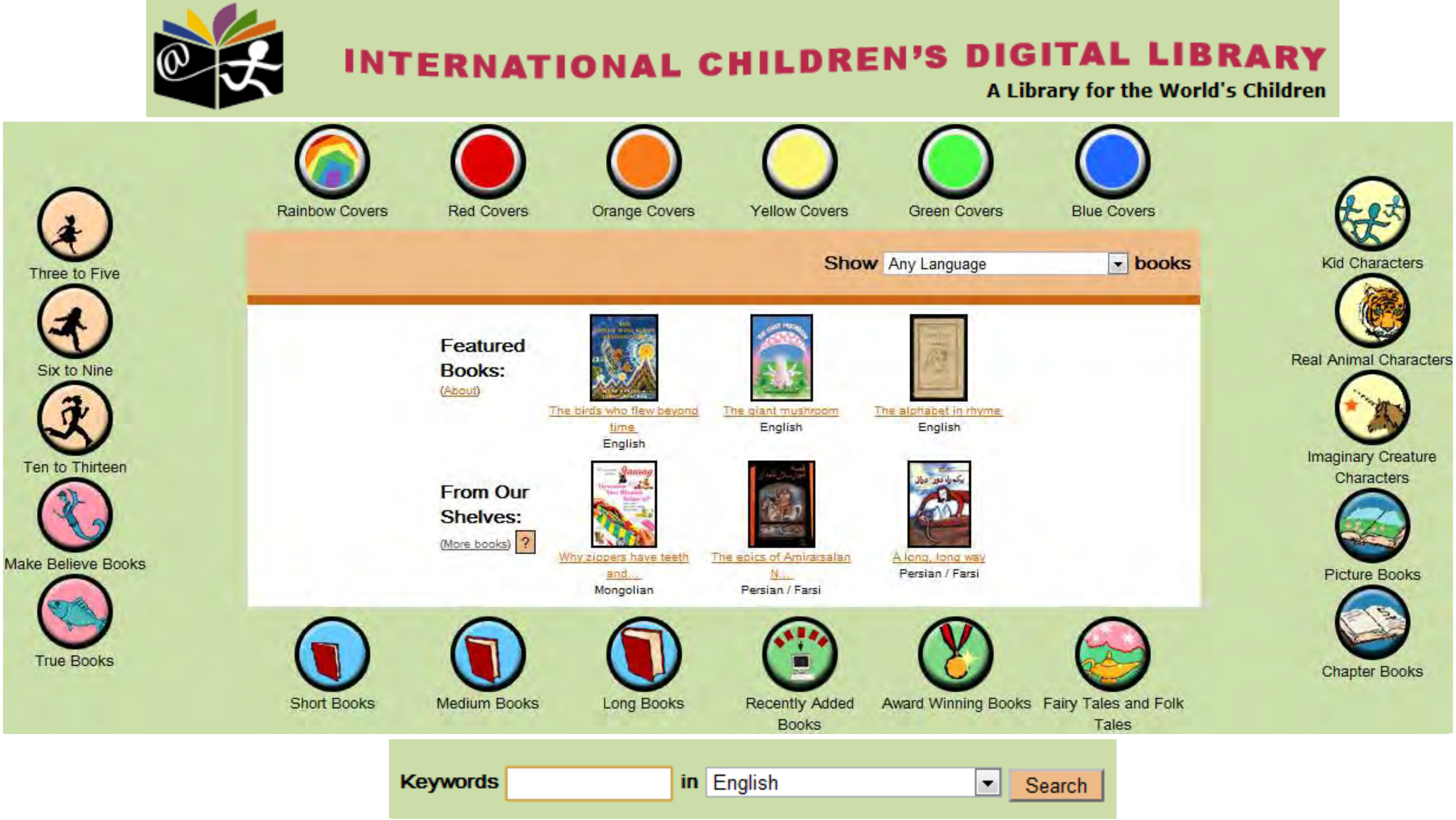
1. Depression: MedlinePlus
   Also called: Clinical depression, Dysthymic disorder, Major depressive disorder, Unipolar depression
   http://www.nlm.nih.gov/medlineplus/depression.html
2. NIMH · Depression
   Depression is a serious medical illness; it's not something that you have made up in your head.
   http://www.nimh.nih.gov/health/topics/depression/index.shtml

What is faceted search?



FacetLens (Microsoft Research)

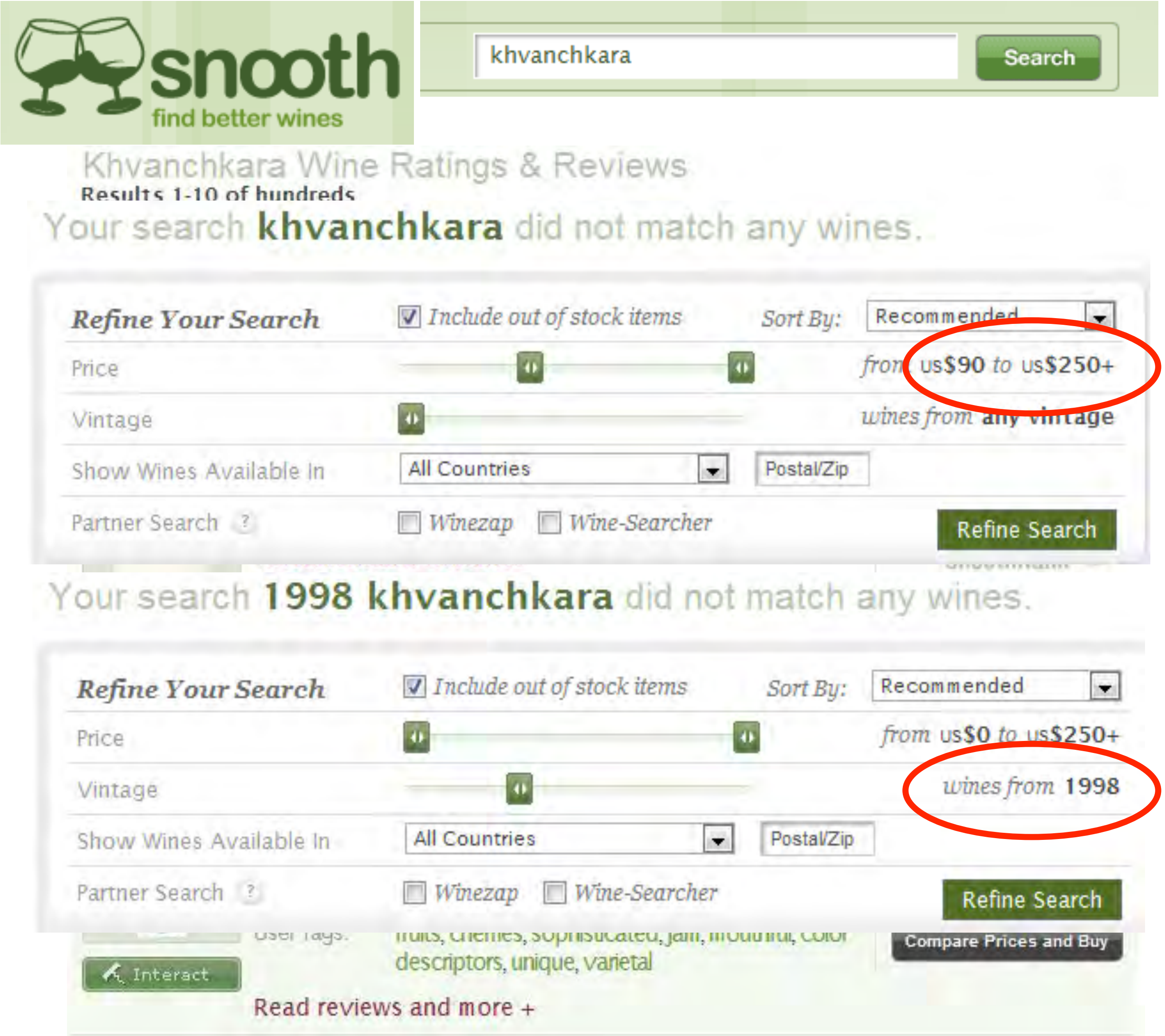
What is faceted search?



Show Any Language books

Keywords in English Search

What is not faceted search?



khvanchkara

Khvanchkara Wine Ratings & Reviews

Results 1-10 of hundreds

Your search **khvanchkara** did not match any wines.

Refine Your Search

☑ *Include out of stock items*

Sort By: Recommended

Price: *from* us$90 *to* us$250+

Vintage: *wines from* **any vintage**

Show Wines Available In: All Countries | Postal/Zip

Partner Search: □ *Winezap* □ *Wine-Searcher*

Your search **1998 khvanchkara** did not match any wines.

Refine Your Search

☑ *Include out of stock items*

Sort By: Recommended

Price: *from* us$0 *to* us$250+

Vintage: *wines from* 1998

Show Wines Available In: All Countries | Postal/Zip

Partner Search: □ *Winezap* □ *Wine-Searcher*

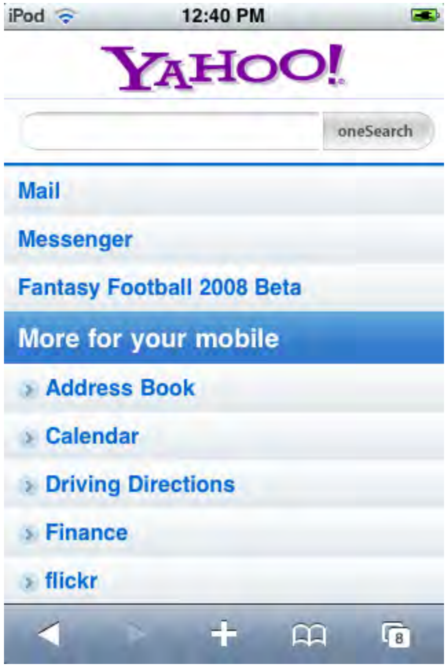
User Tags: fruits, cherries, sophisticated, jam, mouthful, color descriptors, unique, varietal

Read reviews and more +

Too many facets ?
Too many facet values?

Information overload

Mobile interfaces



iPod 12:40 PM

YAHOO!

oneSearch

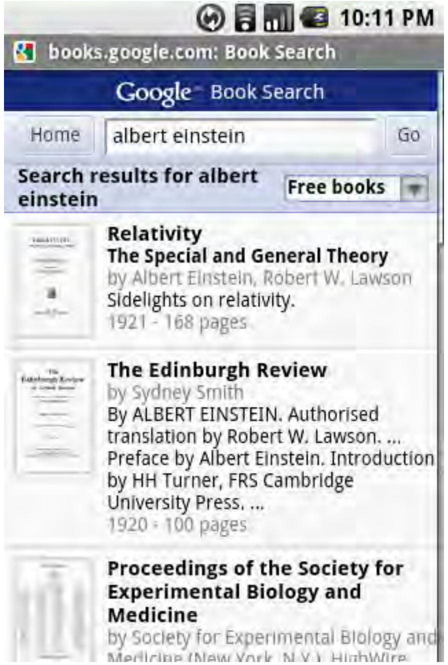
Mail

Messenger

Fantasy Football 2008 Beta

More for your mobile

- Address Book
- Calendar
- Driving Directions
- Finance
- flickr



10:11 PM

books.google.com: Book Search

Google Book Search

Home | albert einstein | Go

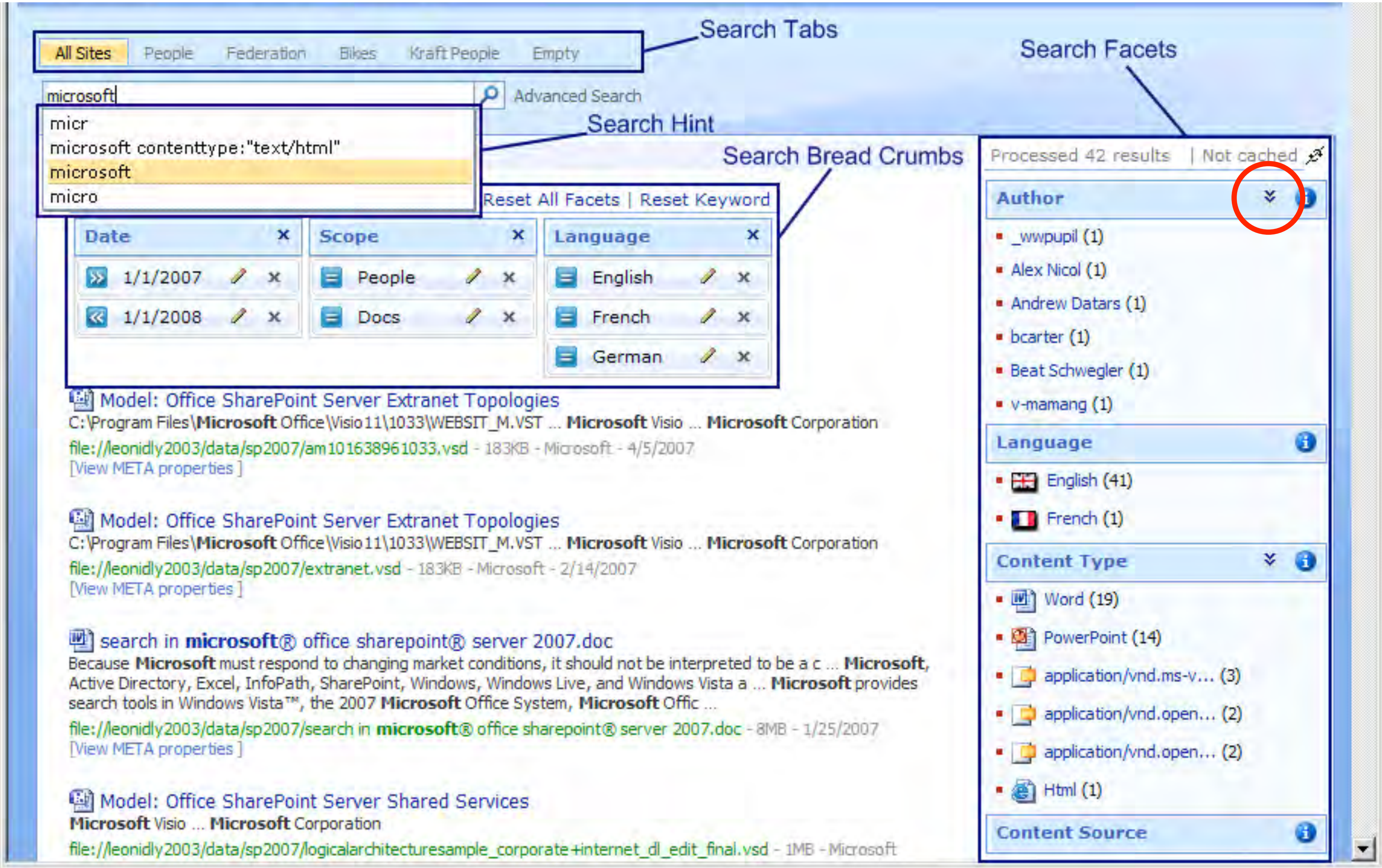
Search results for albert einstein | Free books

Relativity
The Special and General Theory
by Albert Einstein, Robert W. Lawson
Sidelights on relativity.
1921 - 168 pages

The Edinburgh Review
by Sydney Smith
By ALBERT EINSTEIN. Authorised translation by Robert W. Lawson. ... Preface by Albert Einstein. Introduction by HH Turner, FRS Cambridge University Press. ...
1920 - 100 pages

Proceedings of the Society for Experimental Biology and Medicine
by Society for Experimental Biology and Medicine (New York, N.Y.), HighWire

Facet selection: interface-based approach



microsoft

micr
microsoft contenttype:"text/html"
microsoft
micro

Date	Scope	Language
1/1/2007	People	English
1/1/2008	Docs	French
		German

Model: Office SharePoint Server Extranet Topologies
C:\Program Files\Microsoft Office\Visio11\1033\WEBSIT_M.VST ... Microsoft Visio ... Microsoft Corporation
file://leonidly2003/data/sp2007/am101638961033.vsd - 183KB - Microsoft - 4/5/2007
[View META properties]

Model: Office SharePoint Server Extranet Topologies
C:\Program Files\Microsoft Office\Visio11\1033\WEBSIT_M.VST ... Microsoft Visio ... Microsoft Corporation
file://leonidly2003/data/sp2007/extranet.vsd - 183KB - Microsoft - 2/14/2007
[View META properties]

search in microsoft® office sharepoint® server 2007.doc
Because Microsoft must respond to changing market conditions, it should not be interpreted to be a c ... Microsoft, Active Directory, Excel, InfoPath, SharePoint, Windows, Windows Live, and Windows Vista a ... Microsoft provides search tools in Windows Vista™, the 2007 Microsoft Office System, Microsoft Offic ...
file://leonidly2003/data/sp2007/search in microsoft® office sharepoint® server 2007.doc - 8MB - 1/25/2007
[View META properties]

Model: Office SharePoint Server Shared Services
Microsoft Visio ... Microsoft Corporation
file://leonidly2003/data/sp2007/logicalarchitecturesample_corporate+internet_dl_edit_final.vsd - 1MB - Microsoft

Processed 42 results | Not cached

Author
- _wwpupil (1)
- Alex Nicol (1)
- Andrew Datars (1)
- bcarter (1)
- Beat Schwegler (1)
- v-mamang (1)

Language
- English (41)
- French (1)

Content Type
- Word (19)
- PowerPoint (14)
- application/vnd.ms-v... (3)
- application/vnd.open... (2)
- application/vnd.open... (2)
- Html (1)

Content Source

SharePoint 2010

health.

Redundancy-based selection

- Favor facets with high coverage in the result
- Most popular strategy:
  - Select most frequent facets with best cover!
- Let' s reach more documents in one click:
  - **Greedy solution:** at each step select the facet with the maximum number of **unseen documents**

$$\left| docs \in Facet_1 \cup docs \in Facet_2 \cup \ldots \cup docs \in Facet_K \right|$$

W. Dakka et. al. Automatic Construction of Multifaceted Browsing Interfaces. **CIKM 2005**

Redundancy-based selection

- Avoid presenting both of correlating facets:
  - Language
  - Nationality

Language

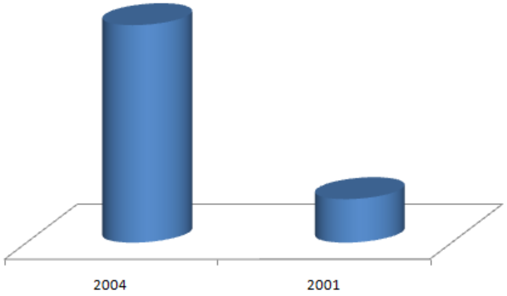
- Consolidate similar facets:
  - Author
  - Editor
  - Contributor

People

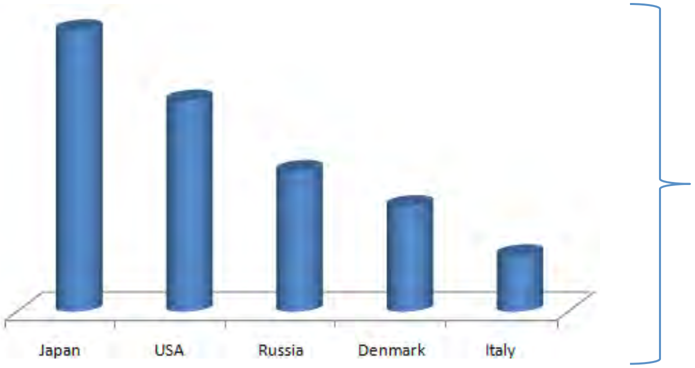
***Beyond Basic Faceted Search.** Ben-Yitzhak et. al. WSDM 2008

Interestingness-based facet selection

- Favor facets with **high-entropy distribution** of facet values:



VS



Documents in result list ***R***

$$Entropy = \sum_{i=1, value_i \in Facet}^{n} P(value_i \mid R) \log P(value_i \mid R)$$

- Favor facets with **query-specific distribution** of facet values:

$$Divergence(Facet, Query) = \sum_{\substack{i=1 \\ value_i \in Facet}}^{n} (P(value_i \mid C) - P(value_i \mid R)) \log \frac{P(value_i \mid R)}{P(value_i \mid C)}$$

Relevance based selection

- Rank **facets** by relevance of their documents
  - Consider all documents with the facet
- Rank **facet values** within a facet
  - Consider all documents with certain facet values
- Aggregate scores of documents:

$$Relevance(v_i) = \sum_{\substack{Doc \in Result, \\ f \in Doc \\ f = v_i}} Score(Doc)$$

To rank facets

To rank facet values

Preference based selection

- Suppose we have long history of interactions
  - Queries + returned documents
  - Maybe even clicks
  - Or just personal/bookmarked documents
- So, let's build a user model!
- User preferences over all ever issued queries:

$$P(f \mid User_k) = \frac{\sum_{Query \in User_k} I(f = clicked, Query)}{\left| Queries \in User_k \right|}$$

Collaboratively recommended selection

- Utilize collaborative filtering techniques*:

$$\alpha P(f \mid User_k) + (1-\alpha)\underbrace{\frac{\sum_{User_j \in Users} P(f \mid User_j)}{\mid Users \mid}}_{\textbf{average preferences over all users}}$$

- Consider only users with similar tastes:

$$\alpha P(f \mid User_j) + (1-\alpha) \sum_{User_j \in Users} P\left(User_j \mid User_k\right) P(f \mid User_j)$$

For example, based on cosine similarity or divergence of prob. distributions over facets

***Personalized Interactive Faceted Search.** Koren et. al. WWW 2008

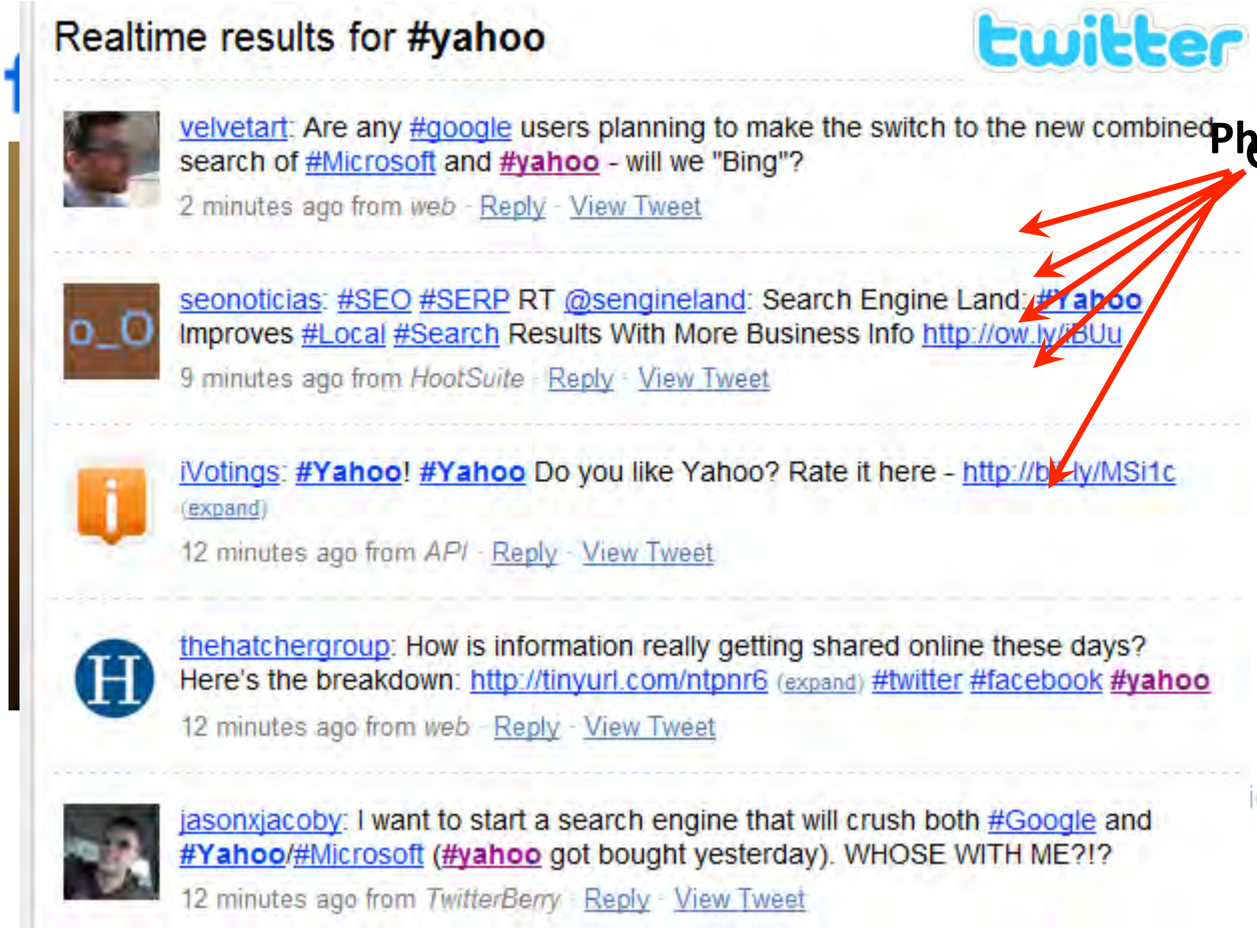
Summary

- Faceted search is a must
  - Especially, when metadata is structured
- Interfaces are crucially important to satisfy the user and help to learn
  - Need to be simple, but customizable
  - Allow to **navigate** the result
- Summarization should be
  - Result-set oriented, query specific
  - Giving answers right away, helping to learn
- Facets/values should be selectively presented!

Faceted search with unstructured metadata: Tags!

Tagging

- Make the way to annotate as easy as possible
- Get metadata for free



Realtime results for #yahoo

velvetart: Are any #google users planning to make the switch to the new combined search of #Microsoft and #yahoo - will we "Bing"?
2 minutes ago from web · Reply · View Tweet

seonoticias: #SEO #SERP RT @sengineland: Search Engine Land: #Yahoo Improves #Local #Search Results With More Business Info http://ow.ly/iBUu
9 minutes ago from HootSuite · Reply · View Tweet

iVotings: #Yahoo! #Yahoo Do you like Yahoo? Rate it here - http://bit.ly/MSi1c (expand)
12 minutes ago from API · Reply · View Tweet

thehatchergroup: How is information really getting shared online these days? Here's the breakdown: http://tinyurl.com/ntpnr6 (expand) #twitter #facebook #yahoo
12 minutes ago from web · Reply · View Tweet

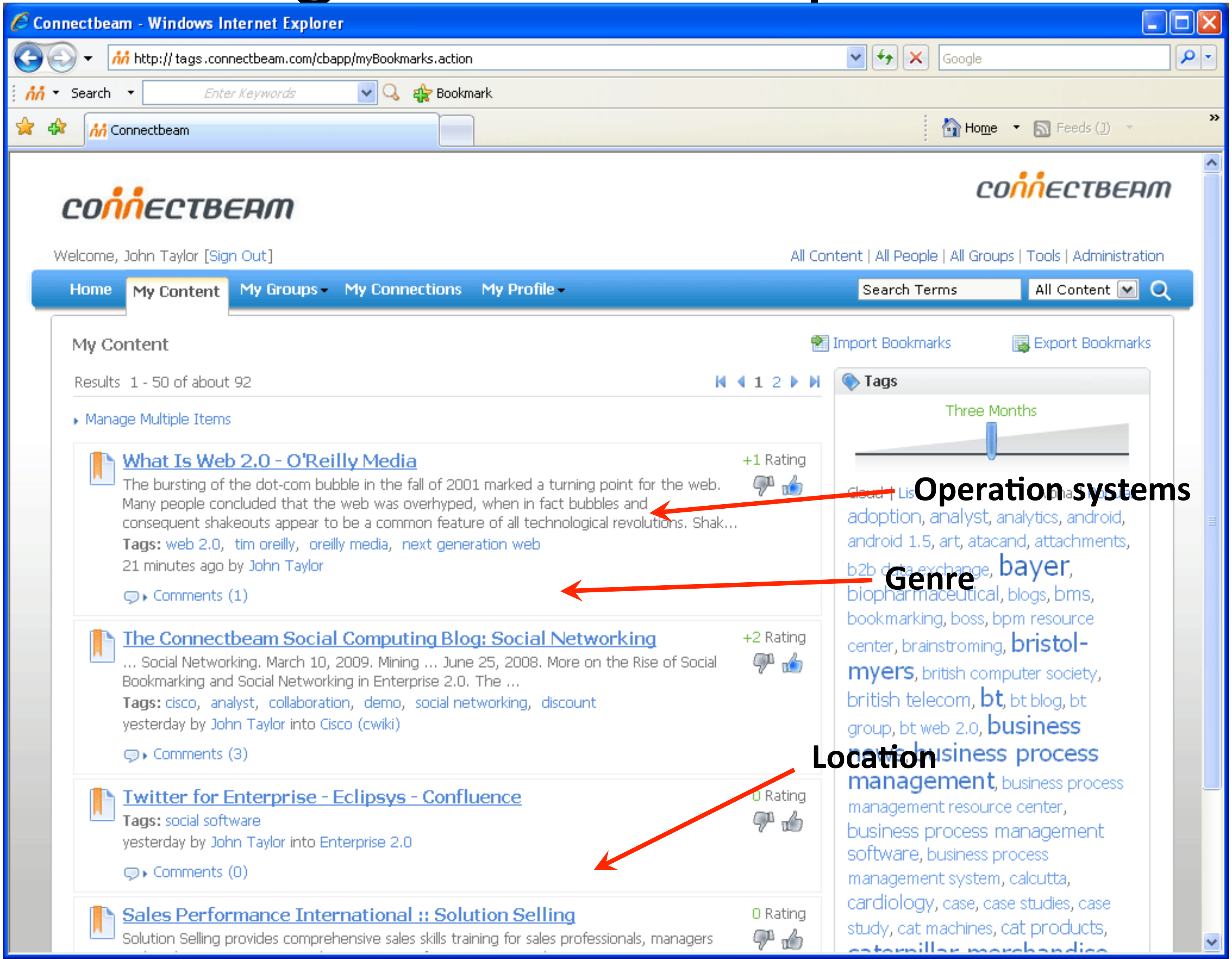
jasonxjacoby: I want to start a search engine that will crush both #Google and #Yahoo/#Microsoft (#yahoo got bought yesterday). WHOSE WITH ME?!?
12 minutes ago from TwitterBerry · Reply · View Tweet

Photographed objects

Genre

idea è mia!

Tags in the Enterprise



Connectbeam - Windows Internet Explorer

http:// tags.connectbeam.com/cbapp/myBookmarks.action

Welcome, John Taylor [Sign Out]

All Content | All People | All Groups | Tools | Administration

Home | My Content | My Groups | My Connections | My Profile

Search Terms | All Content

My Content

Import Bookmarks | Export Bookmarks

Results 1 - 50 of about 92

1 2

Manage Multiple Items

What Is Web 2.0 - O'Reilly Media +1 Rating

The bursting of the dot-com bubble in the fall of 2001 marked a turning point for the web. Many people concluded that the web was overhyped, when in fact bubbles and consequent shakeouts appear to be a common feature of all technological revolutions. Shak...

Tags: web 2.0, tim oreilly, oreilly media, next generation web

21 minutes ago by John Taylor

Comments (1)

The Connectbeam Social Computing Blog: Social Networking +2 Rating

... Social Networking. March 10, 2009. Mining ... June 25, 2008. More on the Rise of Social Bookmarking and Social Networking in Enterprise 2.0. The ...

Tags: cisco, analyst, collaboration, demo, social networking, discount

yesterday by John Taylor into Cisco (cwiki)

Comments (3)

Twitter for Enterprise - Eclipsys - Confluence 0 Rating

Tags: social software

yesterday by John Taylor into Enterprise 2.0

Comments (0)

Sales Performance International :: Solution Selling 0 Rating

Solution Selling provides comprehensive sales skills training for sales professionals, managers

Tags

Three Months

Cloud | List

adoption, analyst, analytics, android, android 1.5, art, atacand, attachments, b2b data exchange, bayer, biopharmaceutical, blogs, bms, bookmarking, boss, bpm resource center, brainstroming, bristol-myers, british computer society, british telecom, bt, bt blog, bt group, bt web 2.0, business news, business process management, business process management resource center, business process management software, business process management system, calcutta, cardiology, case, case studies, case study, cat machines, cat products, caterpillar merchandise

Operation systems

Genre

Location

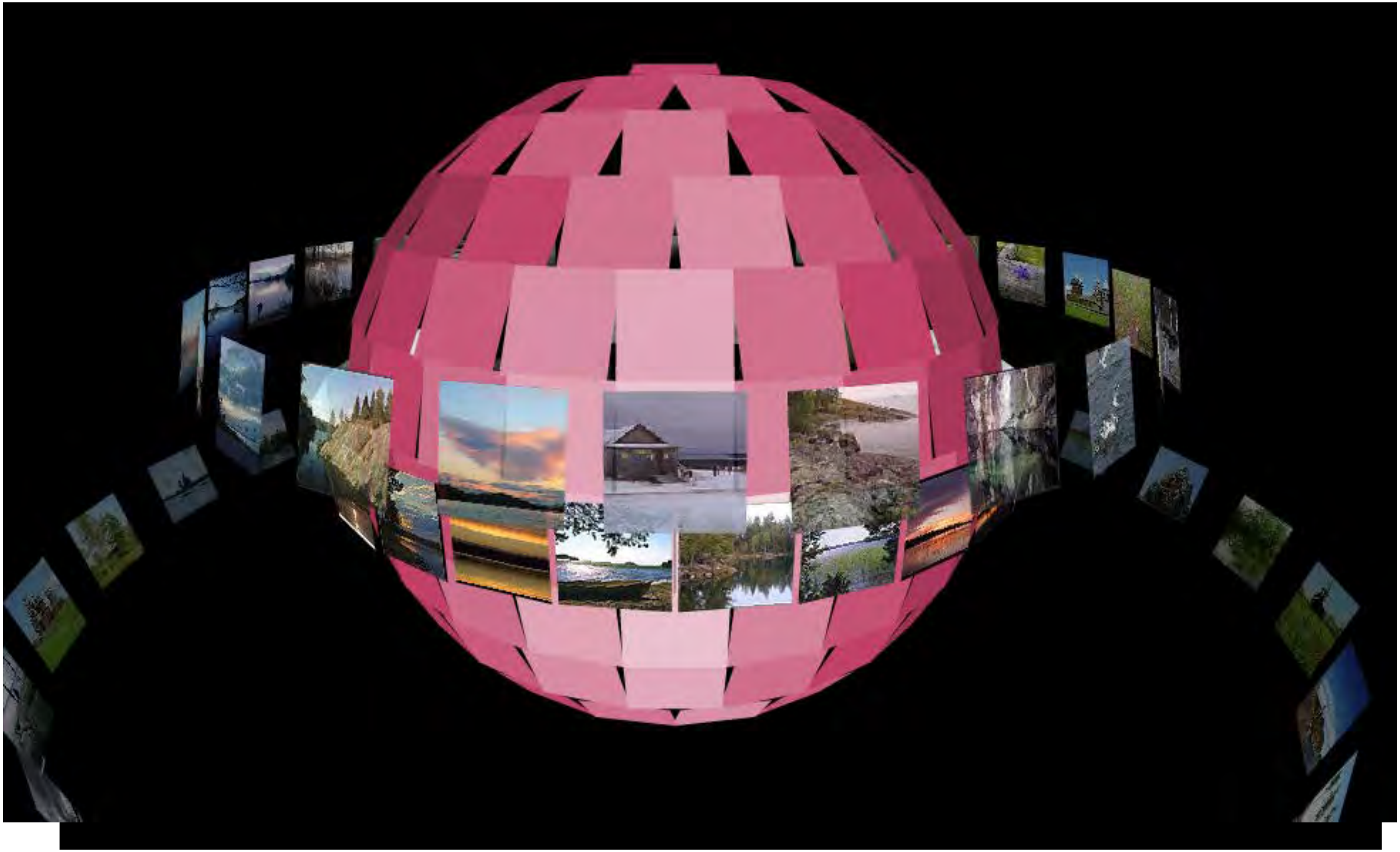
Tagging

- Disadvantages:
  - Nor ranked by relevance to the tagged resource
  - Not organized
  - Not categorized
- But still plenty of ways to summarize!
  - Find “relevant” tags
  - Demonstrate their importance to the user
  - Guess the tag purpose
  - Guess the tag meaning

Tag cloud

http://www.wordle.net/

Tag space



http://taggalaxy.de/

How to measure tag size?

$$fontsize_i = \frac{fontsize_{\max}(tfidf_i - tfidf_{\min})}{(tfidf_{\max} - tfidf_{\min})}$$

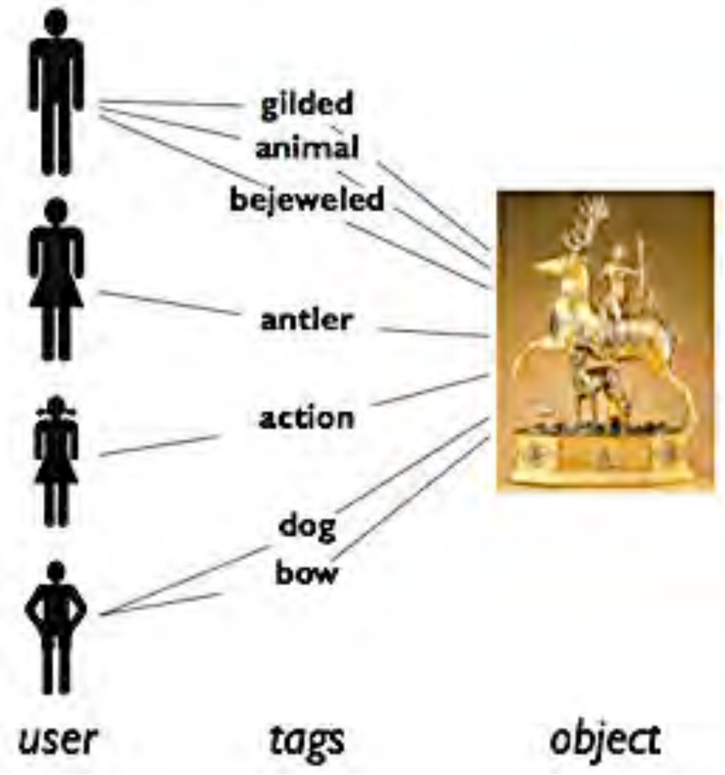
tf – tag frequency in the result set

idf – inverted tag frequency in the collection

tfidf – non-normalized tag importance

Cloud or clouds?

- Group tags by topic!
- Cluster them*!
- Similarity function?
- Tags as vectors of objects
  - But tagging can be non-collaborative
- Tags as vectors of users
  - But co-occurrence less meaningful



***Personalization in folksonomies based on tag clustering.** Gemmel et. al. AAAI 2008

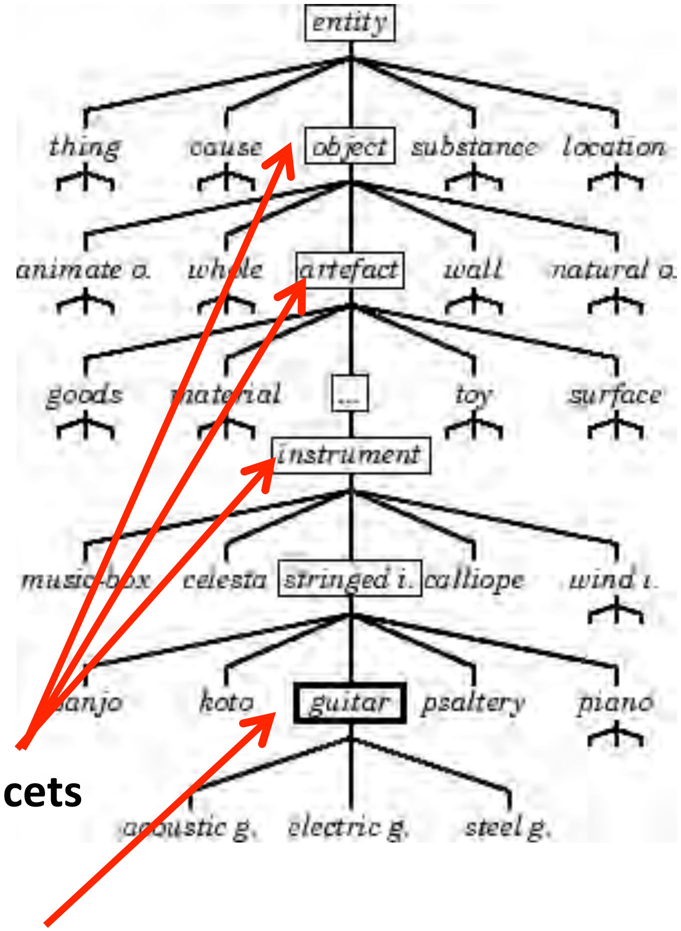
Flickr example

Tag classification for faceted search

- Clusters are nice, but…
  - Random
  - Not always of high quality
- We need some knowledge-based classification
  - To discover more meaningful structure
  - To represent tags as values of facets (classes)
  - To provide the feeling of control for users
- Who knows everything about a word (tag)?
  - Lexical databases: **Wordnet**
  - Encyclopedias: **Wikipedia**

Tag classification with Wordnet

- Contains various semantic relations between word senses
  - guitar is a type of instrument
  - string is part of guitar
  - java is a type of island OR coffee OR language
- About 150 000 senses
  - of 120 00 nouns
- Match tags to nouns
- Disambiguate!
  - Find senses with minimum distance to each other on graph

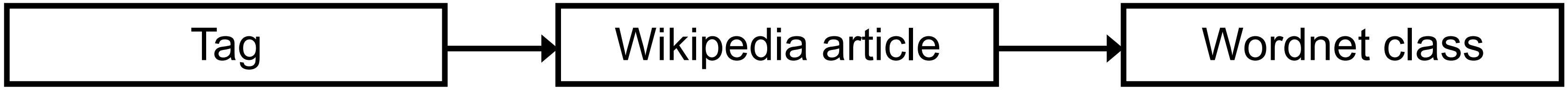


facets

Tags (facet values)

Tag classification with Wikipedia (I)

- Wordnet has nice selection of classes (facets)
- … but no so many entities (facet values)
  - And is not growing as fast as other resources
- Let’ s use larger knowledge repository… **Wikipedia** - more than 3 million articles!
- But it has too many classes (categories)
  - ~ 400,000, their hierarchy is very fuzzy
- Use Wikipedia **just** as a middle layer!



Tag

Wikipedia article

Wordnet class

Tag classification with Wikipedia (II)

1) Match **Tags** => **Wiki articles**

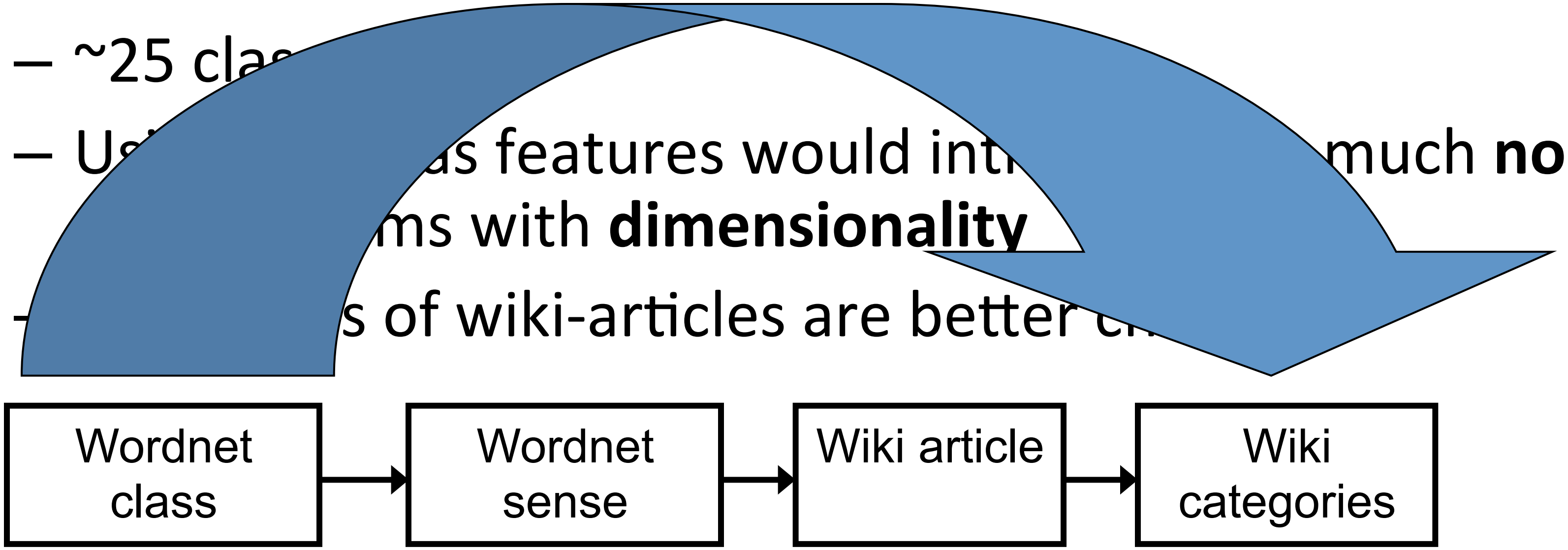
- Match to Wiki titles, anchor text or first sentences

2) Match **Wiki articles => Wordnet senses**

- Some Wikis are direct match with Wordnet senses!
- "Guitar" => en.wikipedia.org/wiki/Guitar
- Use these matching Wikis as training data

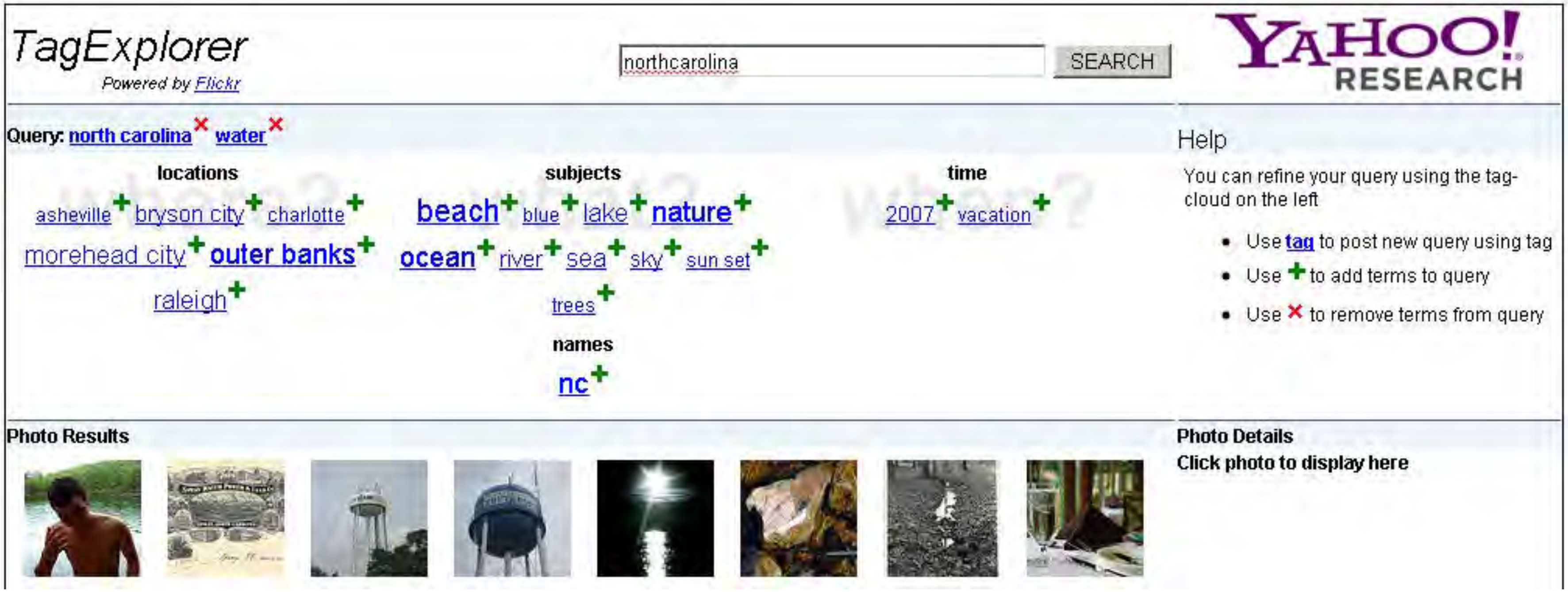
3) Build classifier for each Wordnet noun class

- ~25 clas
- Us as features would intr much **noise** ms with **dimensionality**
- s of wiki-articles are better c



Wordnet class → Wordnet sense → Wiki article → Wiki categories

http://tagexplorer.sandbox.yahoo.com/


TagExplorer
Powered by Flickr
northcarolina
SEARCH
YAHOO! RESEARCH
Query: north carolina ✕ water ✕
locations
asheville + bryson city + charlotte +
morehead city + outer banks +
raleigh +
subjects
beach + blue + lake + nature +
ocean + river + sea + sky + sun set +
trees +
names
nc +
time
2007 + vacation +
Photo Results
Photo Details

Help

You can refine your query using the tag-cloud on the left

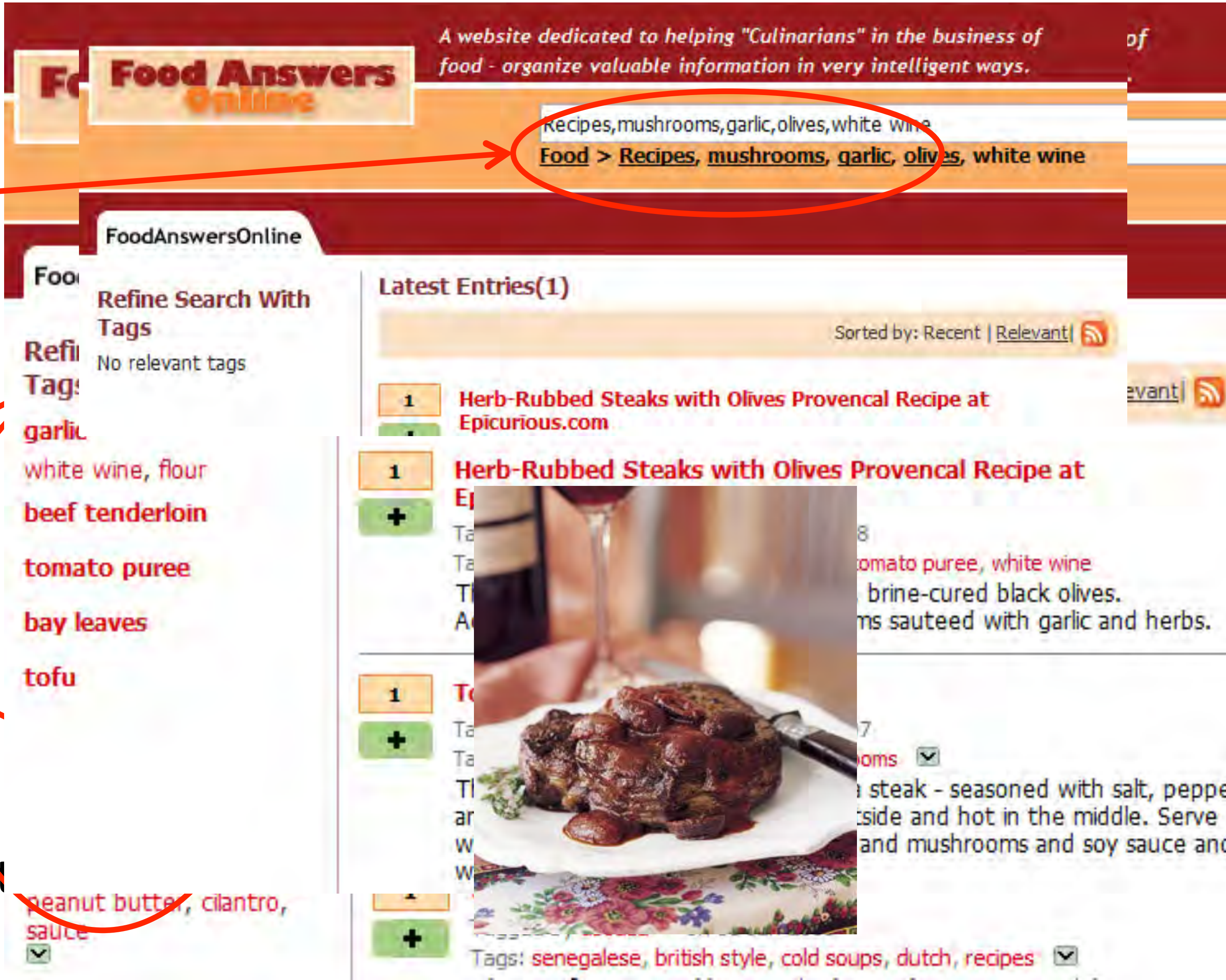
- Use tag to post new query using tag
- Use + to add terms to query
- Use ✕ to remove terms from query

Click photo to display here

- Classified 22% of Flickr tags with Wordnet
- Classified 70% of Flickr tags with Wikipedia

Classifying Tags using Open Content Resources. Overell et. al. WSDM 2008

Filtering – all search tags are made equal



A website dedicated to helping "Culinarians" in the business of food - organize valuable information in very intelligent ways.

Recipes,mushrooms,garlic,olives,white wine

Food > Recipes, mushrooms, garlic, olives, white wine

FoodAnswersOnline

Refine Search With Tags

No relevant tags

Latest Entries(1)

Sorted by: Recent | Relevant|

1 Herb-Rubbed Steaks with Olives Provencal Recipe at Epicurious.com

Refi Tags

garlic

white wine, flour

beef tenderloin

tomato puree

bay leaves

tofu

peanut butter, cilantro, sauce

1 Herb-Rubbed Steaks with Olives Provencal Recipe at Ep

tomato puree, white wine

brine-cured black olives.

ms sauteed with garlic and herbs.

steak - seasoned with salt, peppe

tside and hot in the middle. Serve

and mushrooms and soy sauce and

Tags: senegalese, british style, cold soups, dutch, recipes

Continue narrowing

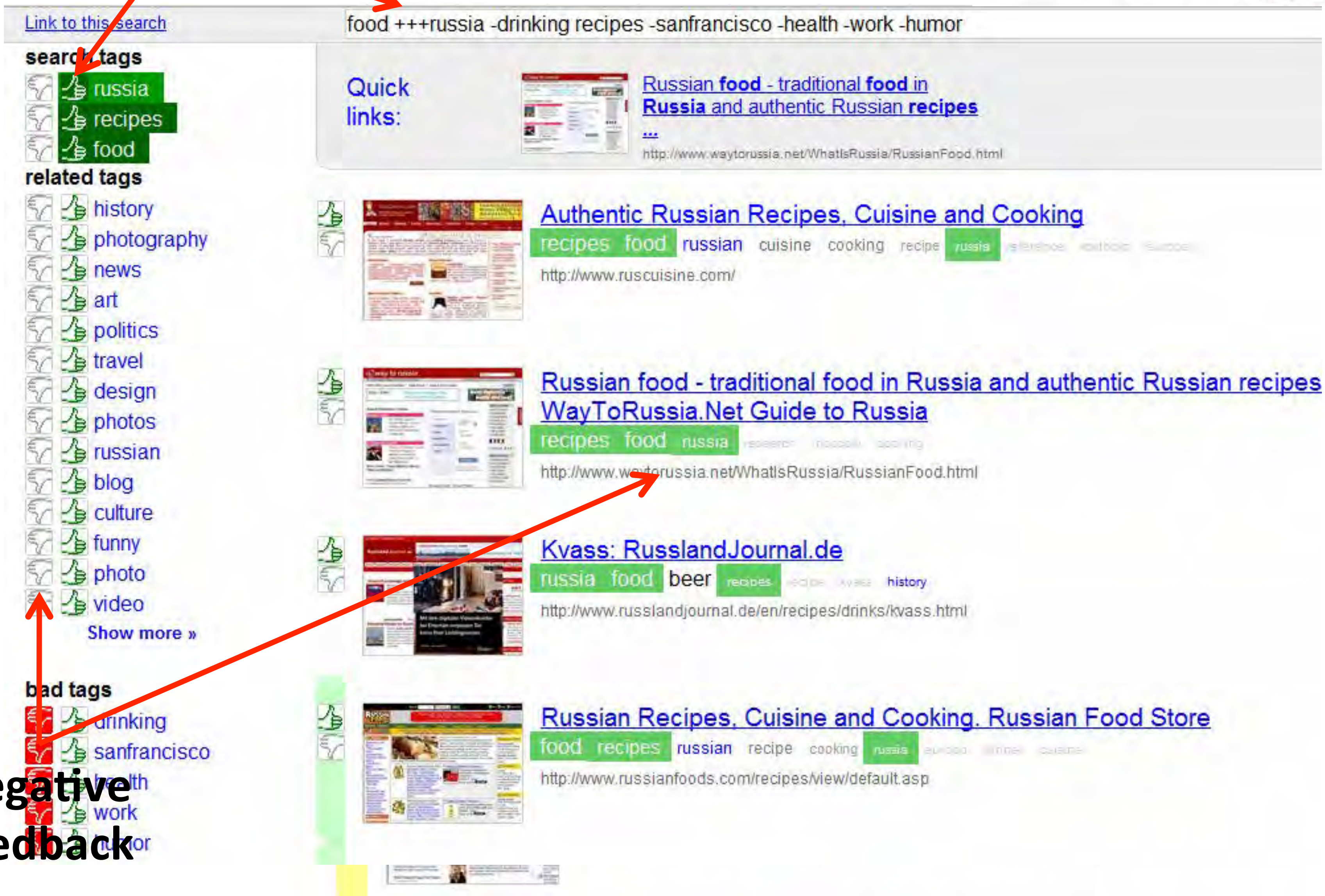
Continue

Start

Tag feedback

MrTaggy PROTOTYPE

Tag weights



Link to this search

food +++russia -drinking recipes -sanfrancisco -health -work -humor

search tags

- russia
- recipes
- food

related tags

- history
- photography
- news
- art
- politics
- travel
- design
- photos
- russian
- blog
- culture
- funny
- photo
- video

Show more »

bad tags

- drinking
- sanfrancisco
- health
- work
- humor

Quick links:

[Russian **food** - traditional **food** in **Russia** and authentic Russian **recipes** ...]

http://www.waytorussia.net/WhatIsRussia/RussianFood.html

[Authentic Russian Recipes, Cuisine and Cooking]

recipes food russian cuisine cooking recipe russia reference cookbook europe

http://www.ruscuisine.com/

[Russian food - traditional food in Russia and authentic Russian recipes WayToRussia.Net Guide to Russia]

recipes food russia research moscow cooking

http://www.waytorussia.net/WhatIsRussia/RussianFood.html

[Kvass: RusslandJournal.de]

russia food beer recipes recipe kvass history

http://www.russlandjournal.de/en/recipes/drinks/kvass.html

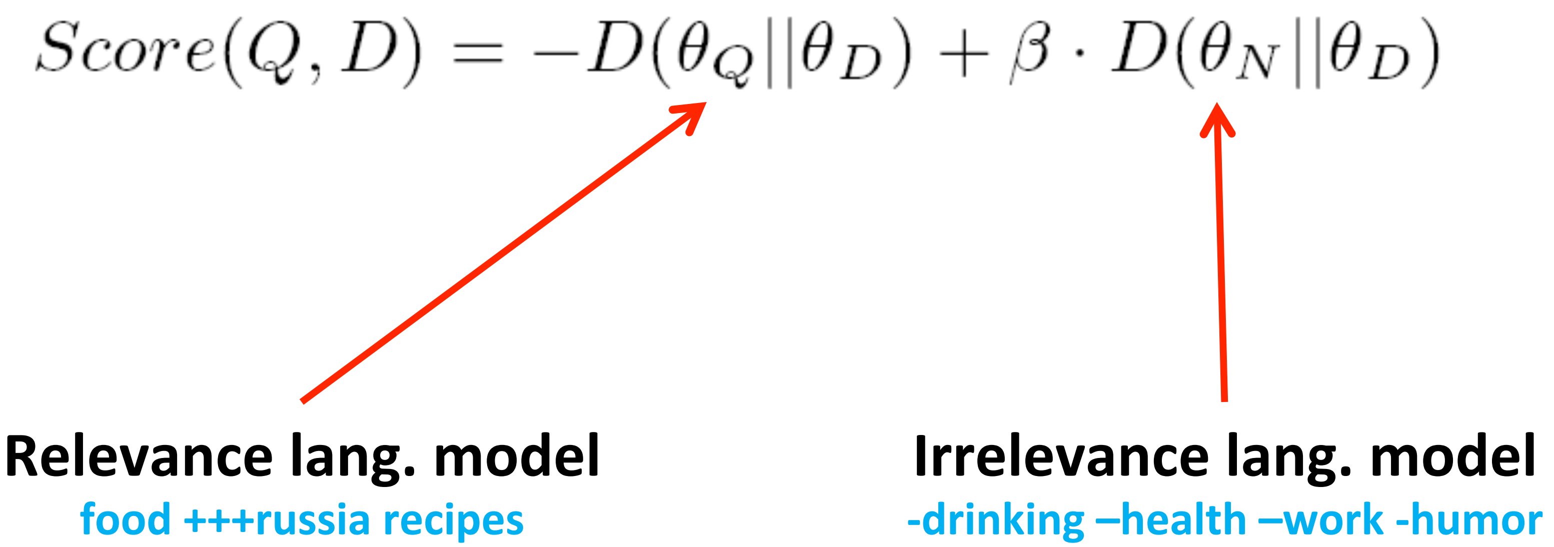
[Russian Recipes, Cuisine and Cooking. Russian Food Store]

food recipes russian recipe cooking russia eurofood dinner cuisine

http://www.russianfoods.com/recipes/view/default.asp

Negative feedback

How to incorporate feedback (I)



$$Score(Q, D) = -D(\theta_Q||\theta_D) + \beta \cdot D(\theta_N||\theta_D)$$

Relevance lang. model

food +++russia recipes

Irrelevance lang. model

-drinking –health –work -humor

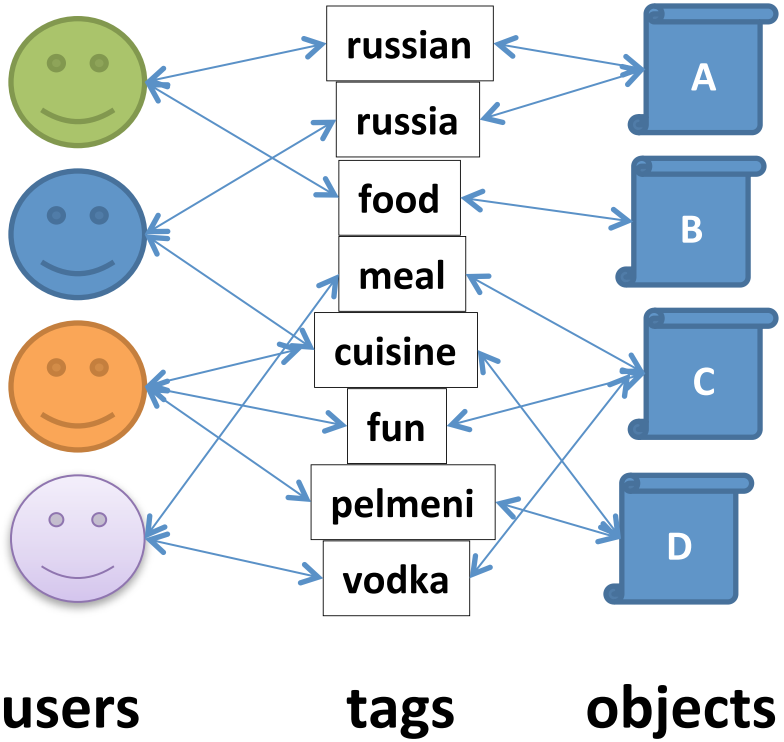
$$P('food'|Q) = 1/5$$

$$P('recipes'|Q) = 1/5$$

$$P('russia'|Q) = 3/5$$

A study of methods for negative relevance feedback Wang et. al. SIGIR 2008

How to incorporate feedback (II)



- We have a tripartite graph
  – Many tags are related, but not used in our query
- It’s good **to be close to positive** tags
- It’s good **to be far from negative** tags

How to incorporate feedback (III)

- Express language models in graph terms:

$$P(tag \mid Document) = \frac{Distance(tag, Document)^{-1}}{\sum_{tag \in alltags} Distance(tag, Document)^{-1}}$$

- How to define **distance** between nodes:
  - Length of shortest path
  - Number of shortest paths (of certain length)
  - Distance-based similarity: $$\sum_{\substack{path(tag,document) \\ \in shortestpaths}} c^{-length(path)}$$

$$c - parameter$$

- What else to consider?
  - Downweight paths with nodes of high indegree/outdegree

Summary

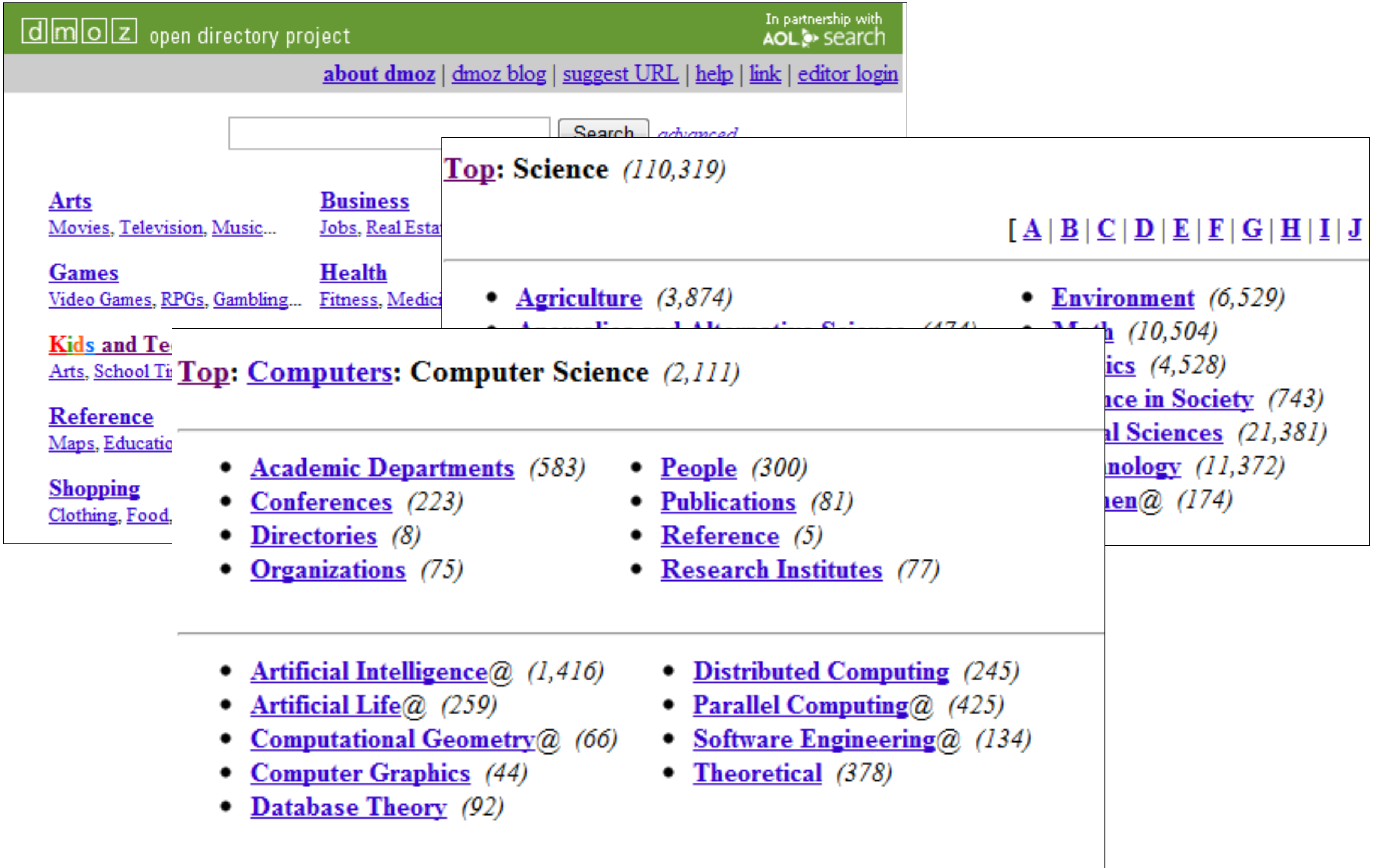
- Faceted search is possible with unstructured metadata…
  - But we need to make some effort **to structure** it!
- Visualization is always important
  - But not enough to understand the summary
- So, it's better to explain the result
  - By clustering tags/objects
  - By classifying tags/objects into semantic categories
- And, finally, it's about navigation and click-based query reformulation
  - Provide ways to react for the user
  - Provide ways to give different kinds of feedback

Faceted search:
No metadata!

No metadata? No panic!

- Facet-value pairs are manual classification
- Tags are basically important terms
- Why not classify automatically?
  - Categorize into known topics
  - Cluster and label clusters
- Why not automatically discover tags?
  - Extract important keywords from documents
- Well, some metadata always exists
  - Time, source….

Categorize by topic (I)



dmoz open directory project

In partnership with AOL search

about dmoz | dmoz blog | suggest URL | help | link | editor login

Search advanced

Arts
Movies, Television, Music...

Business
Jobs, Real Esta

Games
Video Games, RPGs, Gambling...

Health
Fitness, Medici

Kids and Te
Arts, School Ti

Reference
Maps, Educatio

Shopping
Clothing, Food,

Top: Science (110,319)

[A | B | C | D | E | F | G | H | I | J

- Agriculture (3,874)
- Environment (6,529)
- Math (10,504)
- ics (4,528)
- nce in Society (743)
- l Sciences (21,381)
- nology (11,372)
- nen@ (174)

Top: Computers: Computer Science (2,111)

- Academic Departments (583)
- Conferences (223)
- Directories (8)
- Organizations (75)
- People (300)
- Publications (81)
- Reference (5)
- Research Institutes (77)

- Artificial Intelligence@ (1,416)
- Artificial Life@ (259)
- Computational Geometry@ (66)
- Computer Graphics (44)
- Database Theory (92)
- Distributed Computing (245)
- Parallel Computing@ (425)
- Software Engineering@ (134)
- Theoretical (378)

Categorize by topic (II)

- Document categorization
  - – Shallow (Flat) vs. Deep (Hierarchical)
- Shallow classification: only top level
  - – Makes no sense for very focused queries:

    java vs. **biology**
- Deep classification*:
  - – Lack of training examples (labeled documents) with each next level of hierarchy
  - – Documents can be assigned to **too many classes**

Deep Classifier: Automatically Categorizing Search Results into Large-Scale Hierarchies. Xing et. al. WSDM 2008

Categorize by topic (III)

- Solution for sparsity:
  - Suppose, we use Bayesian classification

$$P(Class \mid D) = P(Class)\prod_{w=1}^{|D|} P(w \mid Class)$$

$$P^{smoothed}(w \mid "Databases") =$$

$$= \lambda_1 P(w \mid "Databases") + \lambda_2 P(w \mid "ComputerScience") + \lambda_3 P(w \mid "Science"), \sum \lambda_i = 1$$

- Solution for “too many classes” problem
  - Many documents focus on several topics
  - Let’s care only about those that user cares about:

$$P(Class \mid D) \Rightarrow P(Class \mid D, Q) = P(Class \mid D) P(Class \mid Q)$$

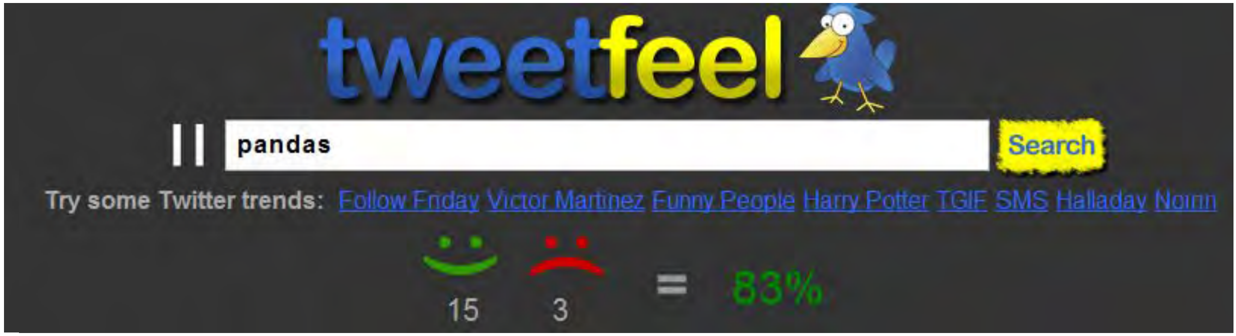
Robust Classification of Rare Queries Using Web Knowledge. Broder et. Al. SIGIR 2007

Non-topical categorization

- Classification by genre
  - patent, news article, meeting report, discussion, resume , tutorial, presentation, source code, blog post?
  - Not only words are features:
    - Average sentence length, layout structure (number of tables, lists), file format, classes of words (dates, times, phone numbers), sentence types (declarative, imperative, question), number of images, links…
- Classification by reading difficulty*
  - Compare definitions of **sugar:**
  - **Sugar** is something that is part of food or can be added to food. It gives a sweet taste © **simple.wikipedia.org/wiki/Sugar**
  - **Sugar** is a class of edible crystalline substances, mainly sucrose, lactose, and fructose. Human taste buds interpret its flavor as sweet © **wikipedia.org/wiki/Sugar**

***A Language Modeling Approach to Predicting Reading Difficulty.** Collins-Thompson et. al. 2004

Categorization by sentiment (I)



Try some Twitter trends: Follow Friday Victor Martinez Funny People Harry Potter TGIF SMS Halladay Norm

15 3 = 83%

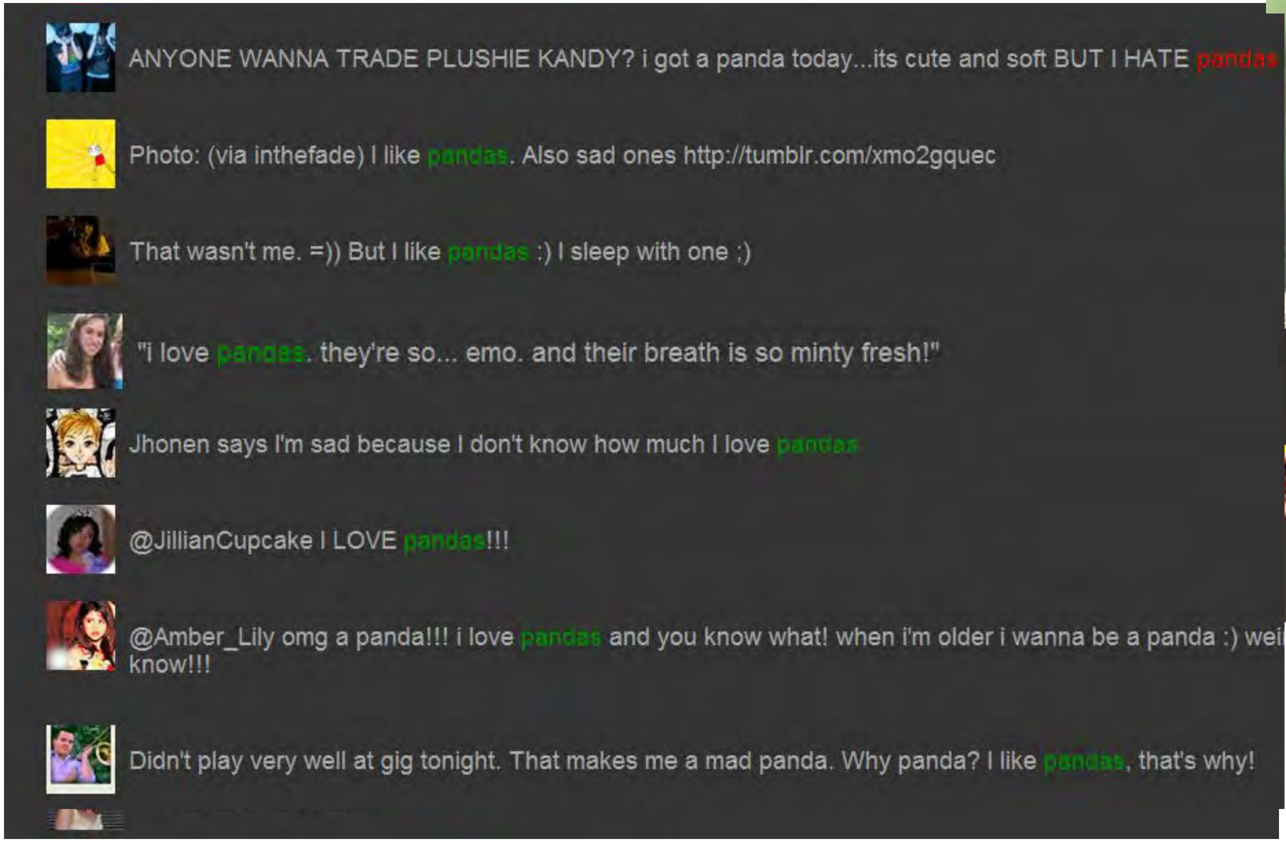

ANYONE WANNA TRADE PLUSHIE KANDY? i got a panda today...its cute and soft BUT I HATE pandas

Photo: (via inthefade) I like pandas. Also sad ones http://tumblr.com/xmo2gquec

That wasn't me. =)) But I like pandas :) I sleep with one ;)

"i love pandas. they're so... emo. and their breath is so minty fresh!"

Jhonen says I'm sad because I don't know how much I love pandas

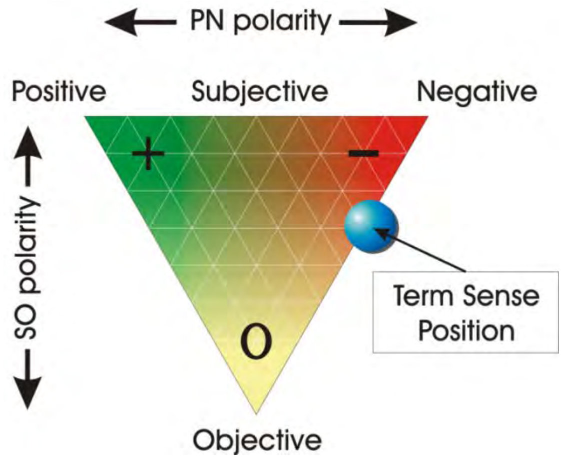
@JillianCupcake I LOVE pandas!!!

@Amber_Lily omg a panda!!! i love pandas and you know what! when i'm older i wanna be a panda :) wei know!!!

Didn't play very well at gig tonight. That makes me a mad panda. Why panda? I like pandas, that's why!

Categorization by sentiment (II)

- Lexicon-based approaches:
  - Calculate ratio of negative/positive words/smileys
  - Weight contribution of every subjective term by its **inverse distance to query terms**
- Build classification models:
  - Objective vs. Subjective
  - Positive vs. Negative

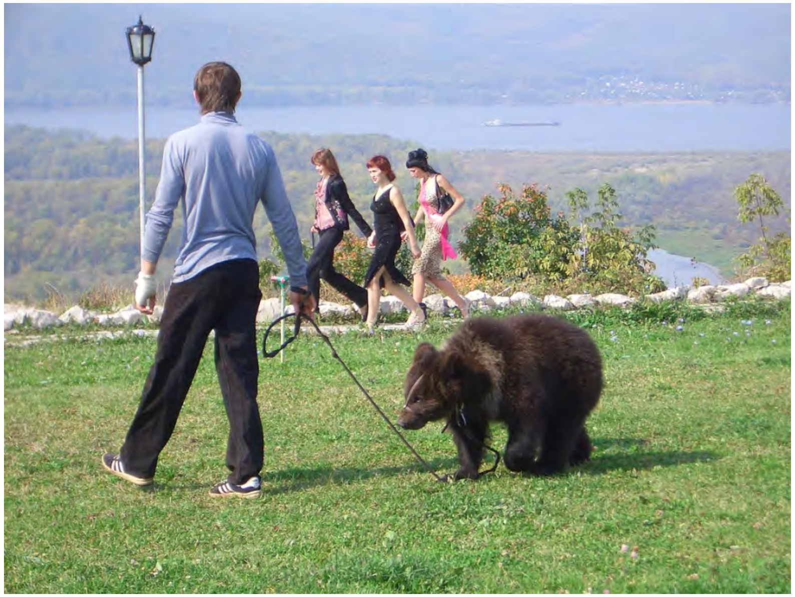


- Enterprises?
  - **Harder**: people try to avoid really strong language
  - **Easier**: domain-specific models can be trained, feedback from users is available, etc.

Categorization by location (I)

- Some documents, photos, videos, tweets…
  - are location agnostic and **some are not!**
  - Where to take location metadata for them?

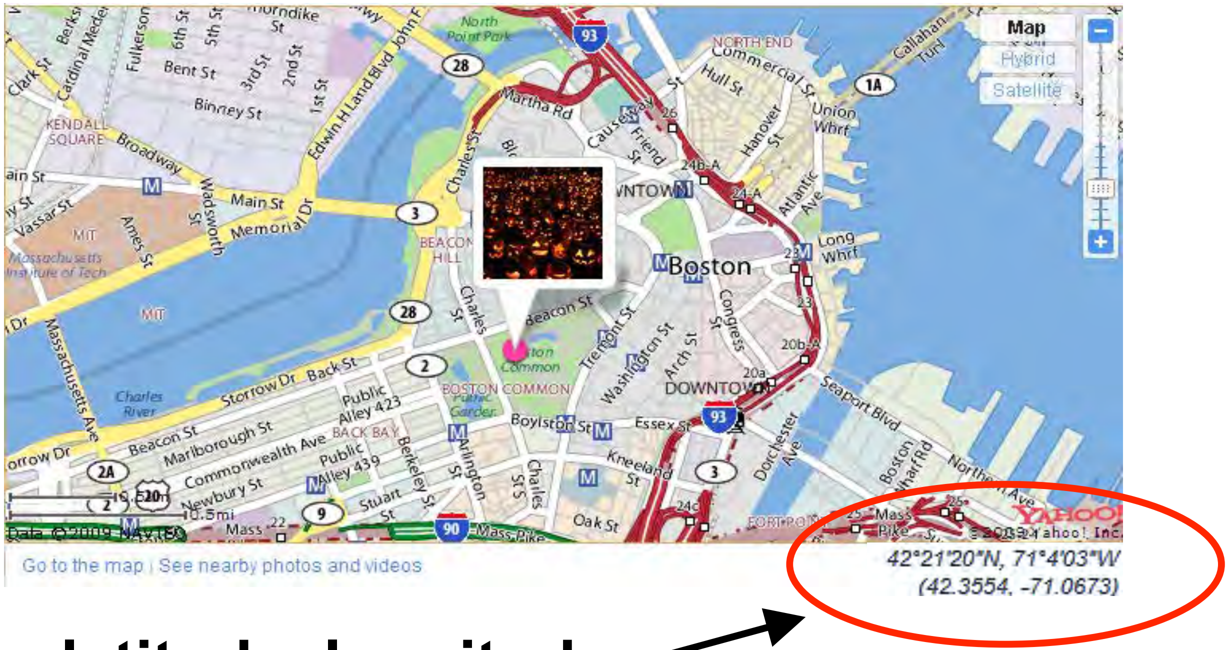
kitchen cats dogs



russia river brownbear

Categorization by location (II)

- Some documents are geo-tagged:

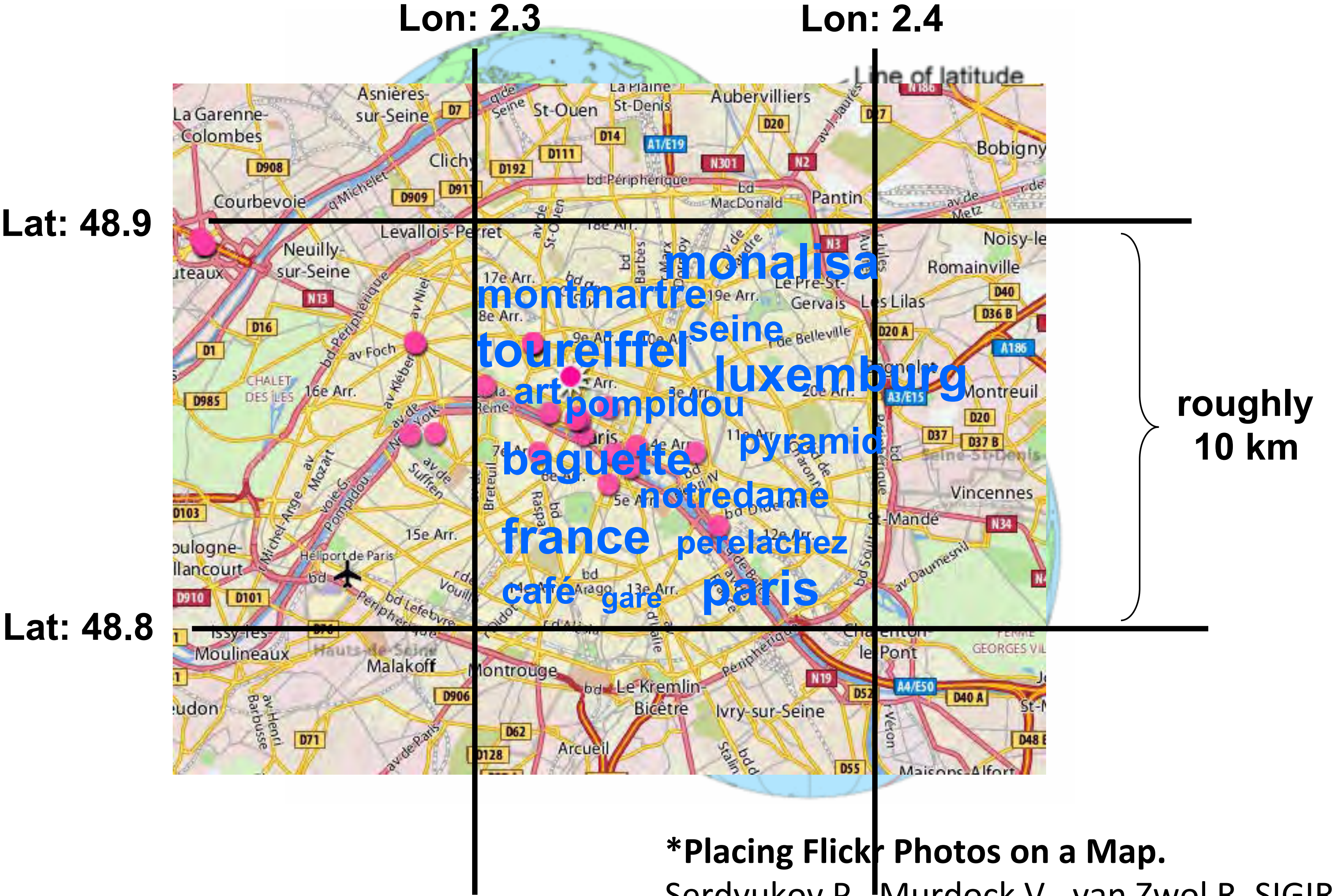


42°21'20"N, 71°4'03"W
(42.3554, -71.0673)

geo-tags: latitude, longitude

- Some documents contain location metadata
- Some users/departments generate only location-specific data

Categorization by location (III)



***Placing Flickr Photos on a Map.**
Serdyukov P., Murdock V., van Zwol R. SIGIR 2009

Categorization by location (IV)

- Locations – documents (L), tagsets – queries (T)
- Tags of photos are query terms (t_i)
- How likely that location L produced the image with a tagset T :

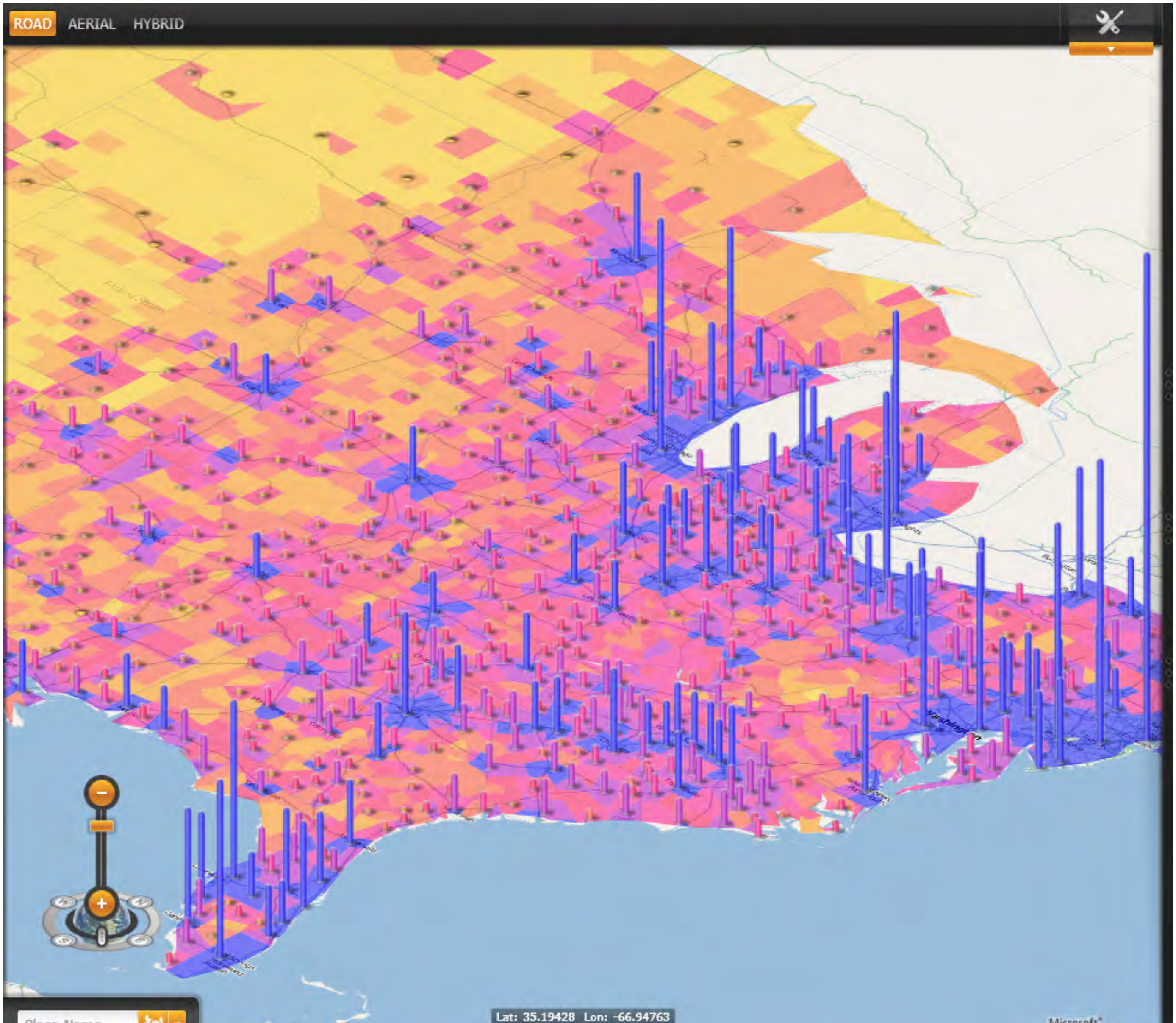
$$P(T \mid L) = \prod_{i=1}^{|T|} P(t_i \mid L)$$

$$P(t \mid L) = \frac{|L|}{|L| + \lambda} P(t \mid L)_{ML} + \frac{\lambda}{|L| + \lambda} P(t \mid G)_{ML}$$

- But there is much more we can do*:
  - Consider spatial ambiguity of tags?
  - Consider neighboring locations?
  - Consider that some of them are toponyms?

***Placing Flickr Photos on a Map.**
Serdyukov P., Murdock V., van Zwol R. SIGIR 2009

Location in Enterprises (SharePoint Example)



Metadata extraction (I)

- Tags provide intuitive description
- Allow not only summarize, but aggregate
- Natural query terms suggestions
- Let' s generate tags (***topic labels***)
  - For each document
  - For clusters of documents
  - For documents grouped by some (boring) facet
    - e.g. Year or Department
- Technically , we can build classification model for **each tag assigned to sufficient number of docs***
  - But let' s do that in an unsupervised way

***Social Tag Prediction.** Heyman et. al. SIGIR 08

Metadata extraction (II)

- Plenty of ways to extract keyphrases…
  - What to consider? Several dimensions*…
- Does phrase $l = w_1 w_2 w_3$ represent document well?

$$Score(l,D) = \alpha \frac{P(l \mid D)}{P(l \mid C)} + (1-\alpha) \sum_{w} \frac{P(w \mid D)}{P(w \mid C)}$$

- Is document on the topic of ***l*** ?

$$Dist(l,D) = -\sum_{w} P(w \mid l) \frac{\boxed{P(w \mid l)}}{P(w \mid D)} \longleftarrow \textbf{Over all docs where } \boldsymbol{l} \textbf{ occurs}$$

- Select top tags using the rule:
  - At each step choose tag that maximizes:

$$\max_{l' \in selected} Dist(l,l')$$

***Automatic Labeling of Multinomial Topic Models.** Mei et. al. KDD 2007

Metadata extraction (III)

- So far not query-driven, right?
- Let's move away from bag-of-words
- Possible algorithm:
  - Cluster sentences in a document
  - Select keywords for each cluster (as shown)
  - Find cluster(s) most relevant to a query
  - Represent document by keywords from relevant cluster(s)
- Just consider text windows around query terms
- So, we can also **just add another constraint**

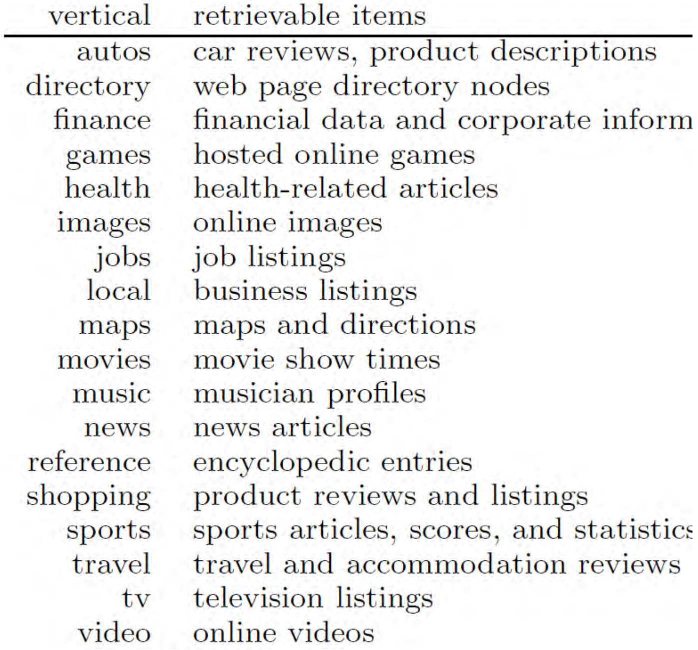
Summary

- No metadata?
- Categorize, categorize, categorize…
  - Semantic classes
  - Genres
  - Reading difficulty levels
  - Sentiments
  - Locations
  - **What else?**
- Or extract metadata from text to summarize!
  - Find tags, entities, etc...

Aggregated exploratory search

- Find not only relevant facets/values, but…
- Find relevant domains (verticals) !

Query "hairspray"



vertical	retrievable items
autos	car reviews, product descriptions
directory	web page directory nodes
finance	financial data and corporate inform
games	hosted online games
health	health-related articles
images	online images
jobs	job listings
local	business listings
maps	maps and directions
movies	movie show times
music	musician profiles
news	news articles
reference	encyclopedic entries
shopping	product reviews and listings
sports	sports articles, scores, and statistics
travel	travel and accommodation reviews
tv	television listings
video	online videos

- Present result sets from different verticals in the order of their total relevance!

References: Exploratory search

- http://en.wikipedia.org/wiki/Exploratory_search
- http://en.wikipedia.org/wiki/Faceted_search
- **Exploratory search: Beyond the Query-Response Paradigm.** R. White and R. Roth. 2009
- **Faceted search**. D. Tunkelang. 2009
- **Search User Interfaces.** M. Hearst. 2009.
  free at: http://searchuserinterfaces.com/
- **Opinion Mining and Sentiment Analysis.** B. Pang and L. Lee. 2008

  free at: http://www.cs.cornell.edu/home/llee/
- **A Survey on Automatic Text Summarization.** D. Das, A. Martins. 2007

  free at: http://www.cs.cmu.edu/~afm/
- **Conferences:** SIGIR, ECIR, WWW, WSDM, KDD, HCIR