

Enterprise and Desktop Search

Lecture 3: Exploratory search

Pavel Dmitriev, **Pavel Serdyukov**, Sergey Chernov

Yahoo! Labs,
Sunnyvale, US

Delft University
Of Technology
The Netherlands

L3S Research Center,
Hannover, Germany

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

Relevance in the Enterprise

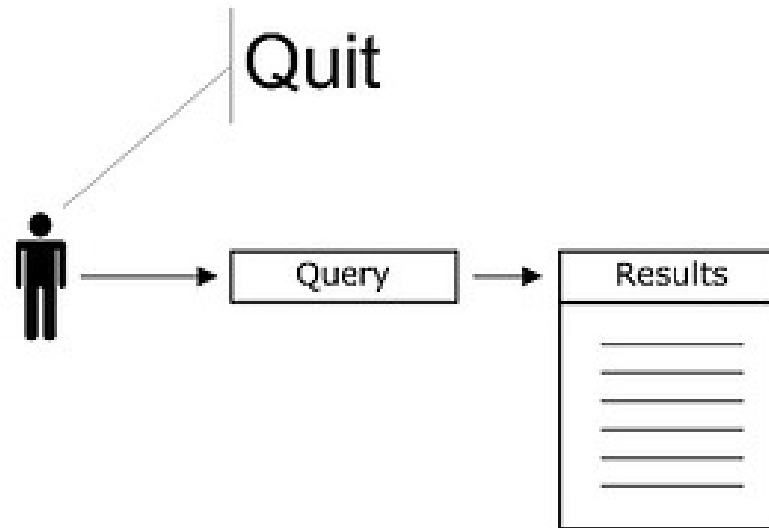
Search in enterprise is hard!
Initial guess is often wrong

Users want to be aware of
everything in the
Enterprise

Users demand
more **control** over search!
They want to **explore**!

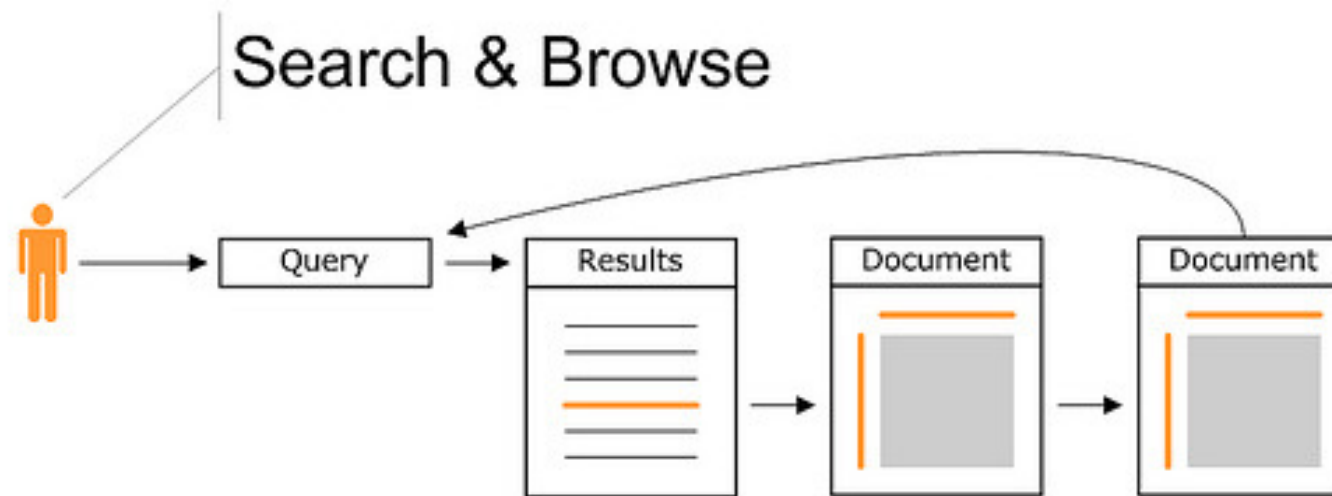


Search is a look-up?



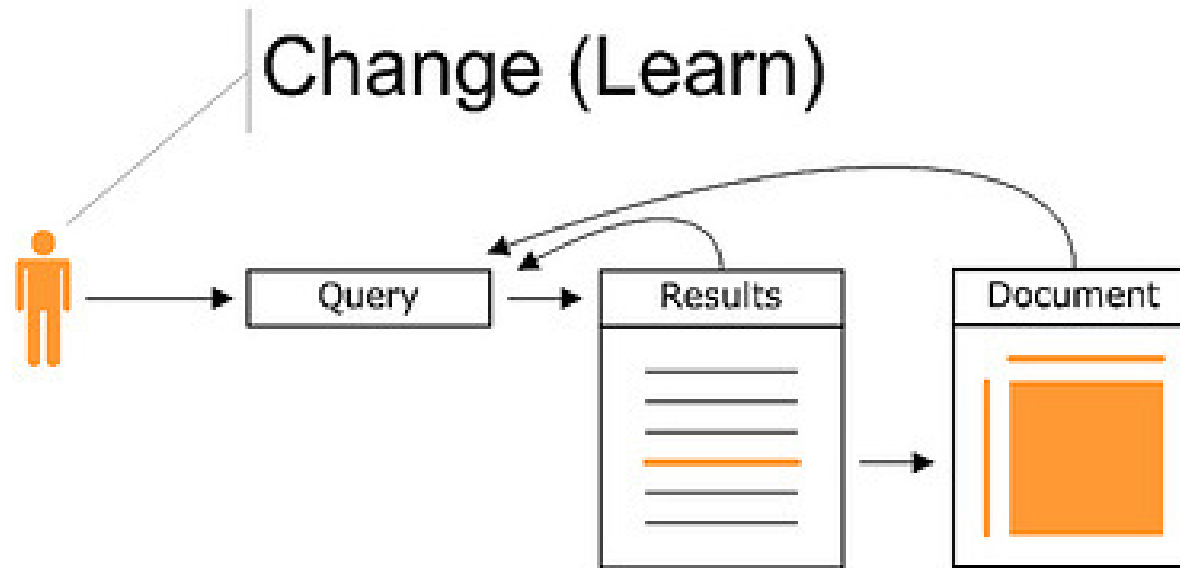
Is that all?
Certainly not in
enterprises

Search is a journey!



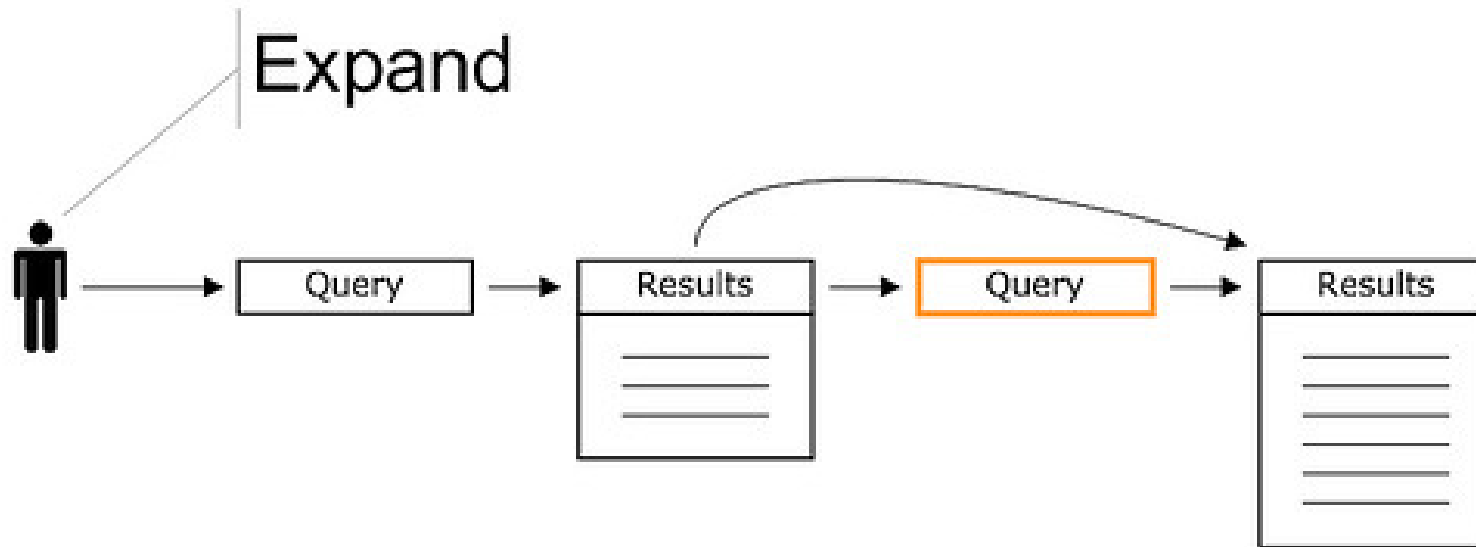
- Exploratory search involves:
 - browsing the result
 - analyzing returned documents
 - coming back to the initial ranking again and again

Search is a journey!



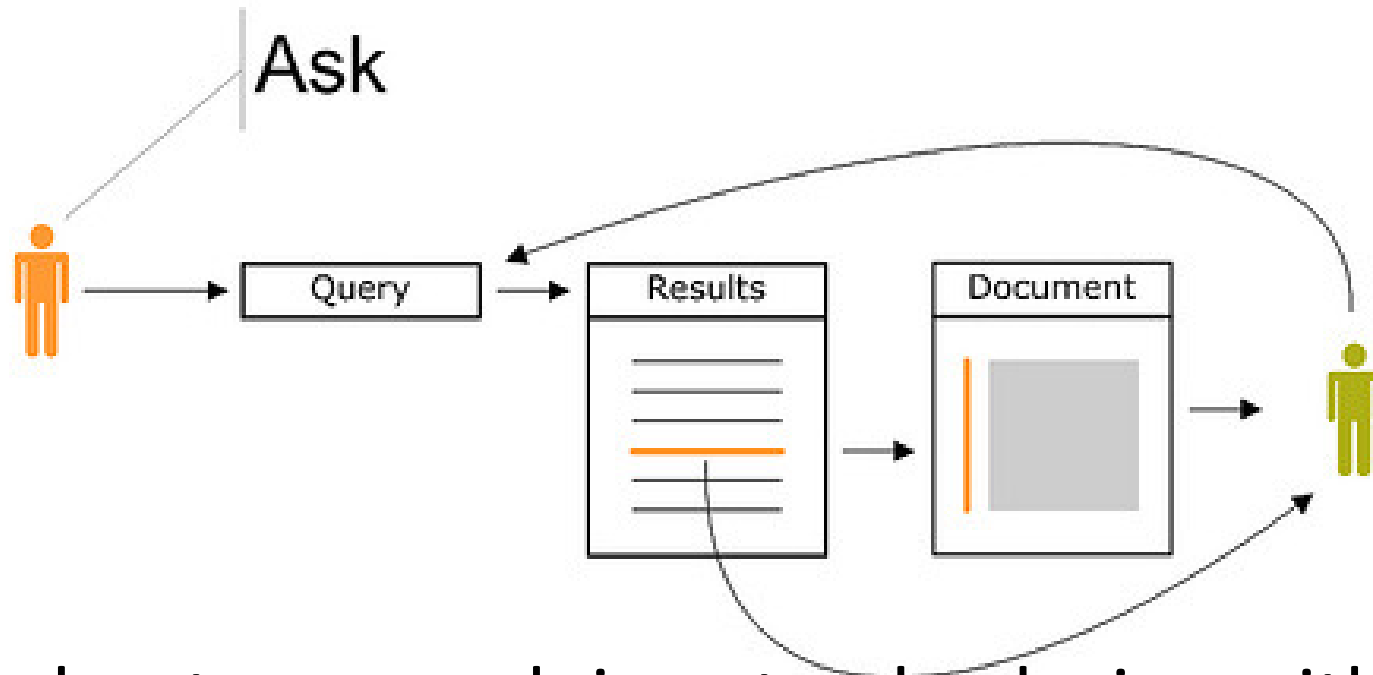
- Exploratory search involves:
 - Querying the last returned result set
 - Looking for similar documents (relevance feedback)

Search is a journey!



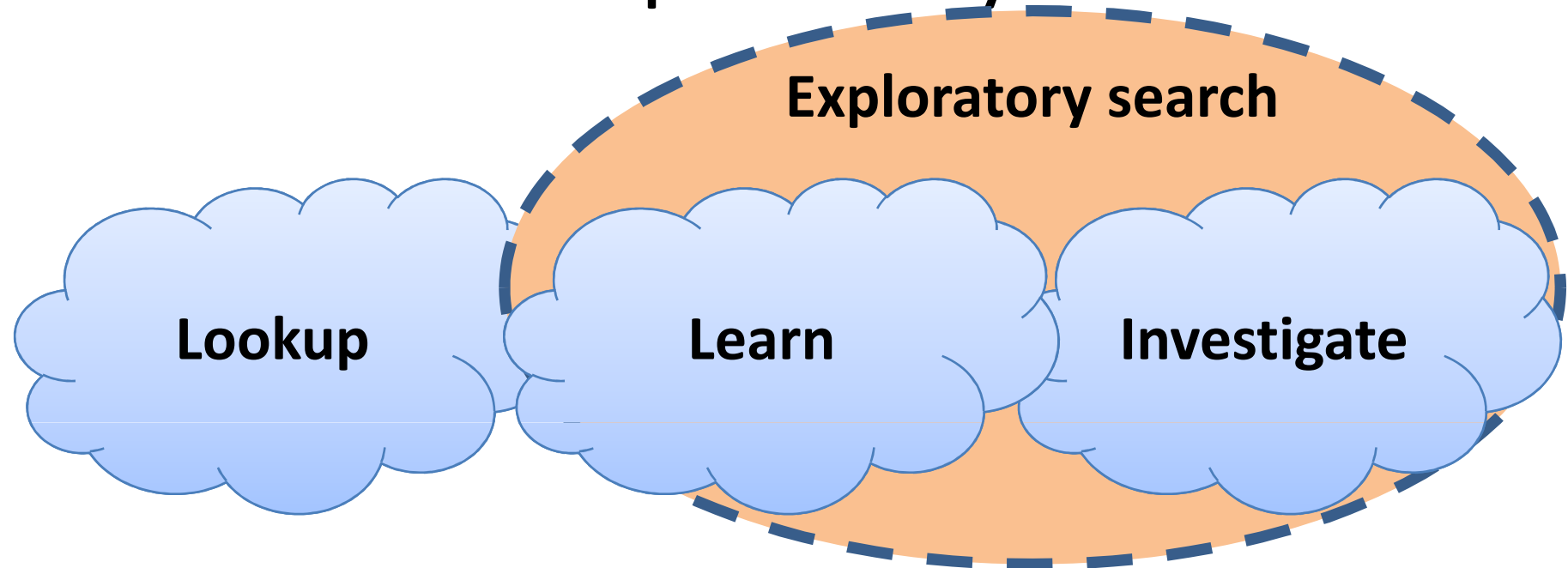
- Exploratory search is also about...
 - Query reformulation, same information need:
 - Specialization: **mp3 players** => **ipod**
 - Generalization: **ipod** => **mp3 players**

Search is a journey!



- Exploratory search is not only playing with a search box, but also... looking for people:
 - Who know the answer
 - Who know where to find answers
 - Who know much more than just an answer

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

Exploratory search: from finding to understanding.
Marchionini. Commun ACM. 2006

Support exploratory behavior

- Support learning
 - About the search topic
 - About the collection
- Support query reformulation
 - Broadening
 - Narrowing
 - Changing the focus
- Support socialization
 - Looking for experts
 - Collaborative search

What web search engines offer

The image is a screenshot of a Yahoo! search results page. At the top, the search bar contains the text "russian school". Below the search bar, a dark blue box displays "Query suggestions" with a large blue arrow pointing to it. The suggestions listed are: "russian school hostage", "russian school siege", "russian school massacre", "russian schoolroom", and "russian school of mathematics". Below the suggestions, the "Search In:" section shows "the Web" selected. A red circle highlights the text "1 - 10 of 195,000 for школа информационного поиска". Below this, the first search result is shown with a blue arrow pointing to its "Snippets" section. The snippet text is: "... где есть в СНГ центры которые работают в проблематике **информационного поиска**? ... Есть еще **школа** Яндекса, может там тоже найдутся желающие. sashchernuh ...". The second search result is also shown with a blue arrow pointing to its snippet. The snippet text is: "III Российская летняя **школа** по информационному поиску ... **информационного** ... текста и другое лингвистическое обеспечение **информационного поиска**; ...".

Web | Images | Video | Local | Shopping | more ▼

russian school Search Options Customize YAHOO!

Search Assist Se

Query suggestions

Search In: ☒ the Web ☐ pages from Netherlands

1 - 10 of 195,000 for школа информационного поиска (About) - 0.43 s | SearchScan

ru_ir: Центры **информационного поиска** - Translate

... где есть в СНГ центры которые работают в проблематике **информационного поиска**? ... Есть еще **школа** Яндекса, может там тоже найдутся желающие. sashchernuh ...

community.livejournal.com/ru_ir/67501.html - Cached

ru_ir: RuSSIR 2009 - Translate

III Российская летняя **школа** по информационному поиску ... **информационного** ... текста и другое лингвистическое обеспечение **информационного поиска**; ...

community.livejournal.com/ru_ir/76336.html - Cached

Snippets

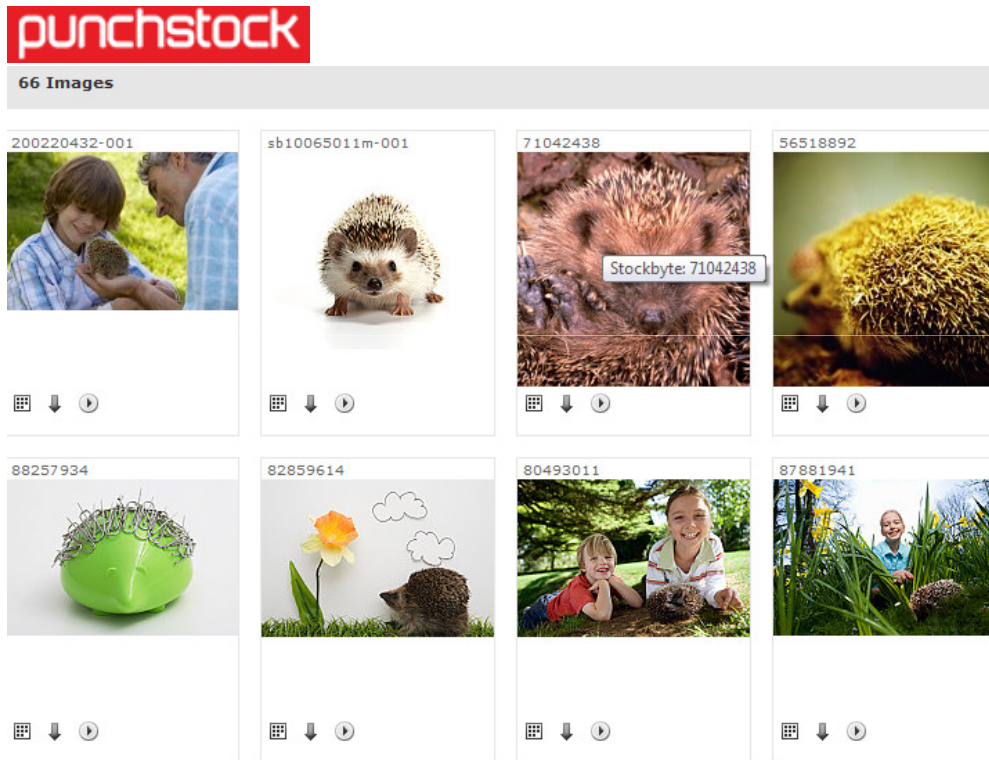
Does it really help to learn?

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

facet

facet values

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

What is faceted search?

punchstock

6 Images

You searched for:
"hedgehog" > One Woman Only
All results are visible on the page.

200245554-001

LS010977

LS010987


LS010978

LS010979

LS010988

ormulation!

What is faceted search?



Real People. Real Reviews.™

Search for (e.g. *taco, salon, Max's*)

Near (Address, [Neighborhood](#), City, State or Zip)

Search

Now in the UK!

[Welcome](#) [About Me](#) [Write a Review](#) [Find Reviews](#) [Invite Friends](#) [Messaging](#) [Talk](#) [Events](#) [Member Search](#) | [Account](#) | [Log In](#)


taco Boston

1 to 10 of 316 - Results per page:


[Hide Filters](#)

Sort By	Neighborhoods	Distance	Features	Price	Category
» Best Match Highest Rated Most Reviewed	<input type="checkbox"/> East Boston <input type="checkbox"/> Back Bay <input type="checkbox"/> Inman Square <input type="checkbox"/> Roslindale ... More Neighborhoods »	» Bird's-eye View Driving (5 mi.) Biking (2 mi.) Walking (1 mi.) Within 4 blocks	<input type="checkbox"/> Open Now (9:07pm) <input type="checkbox"/> Good for Dinner <input type="checkbox"/> Good for Kids <input type="checkbox"/> Take-out ... More features »	<input type="checkbox"/> \$\$\$\$ <input type="checkbox"/> \$\$\$ <input type="checkbox"/> \$\$ <input type="checkbox"/> \$	<input type="checkbox"/> Mexican <input type="checkbox"/> Fast Food <input type="checkbox"/> Latin American <input type="checkbox"/> Bars ... More categories »

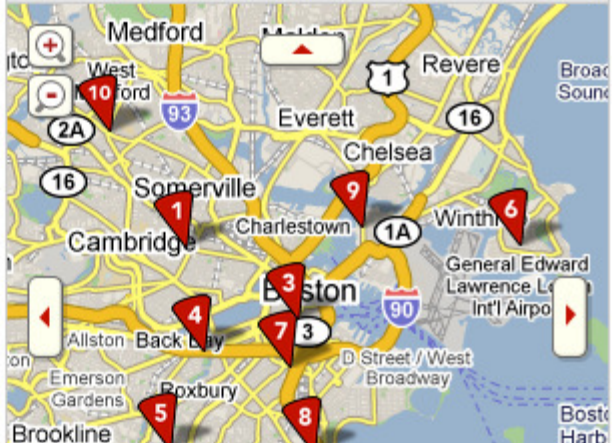
1. **Olecito**
Category: Mexican
Neighborhood: [Inman Square](#)

 ridiculously large **tacos** and a fountain Boylan's soda (yes, that's right, they have a Boylan's FOUNTAIN). But on with the food: These **tacos** were mind-numbingly good. I really believe that a good **taco**

2. **The Wapo Taco**
Category: Mexican
Neighborhoods: [Roslindale Village](#), [Roslindale](#)

 I particularly like this place because it has vegan options, including **tacos** and burritos made with vegetarian "meat," but it's equally popular with the meat-eaters in my family. My daughter likes

« Mo' Map ☐ Map, stay put! ☐ Redo search in map



What is faceted search?



Search:

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



News Media



Audio Video



Clinical Trials



Research Articles

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?



INTERNATIONAL CHILDREN'S DIGITAL LIBRARY

A Library for the World's Children



Rainbow Covers



Red Covers



Orange Covers



Yellow Covers



Green Covers



Blue Covers



Three to Five



Six to Nine



Ten to Thirteen



Make Believe Books



True Books



Kid Characters



Real Animal Characters



Imaginary Creature Characters



Picture Books



Chapter Books

Show books

Featured Books:
[\(About\)](#)



[The birds who flew beyond time](#)
English



[The giant mushroom](#)
English



[The alphabet in rhyme](#)
English

From Our Shelves:
[\(More books\)](#) 



[Why zippers have teeth and...](#)
Mongolian



[The epics of Amirarsalan](#)
Persian / Farsi



[A long, long way](#)
Persian / Farsi



Short Books



Medium Books



Long Books



Recently Added Books



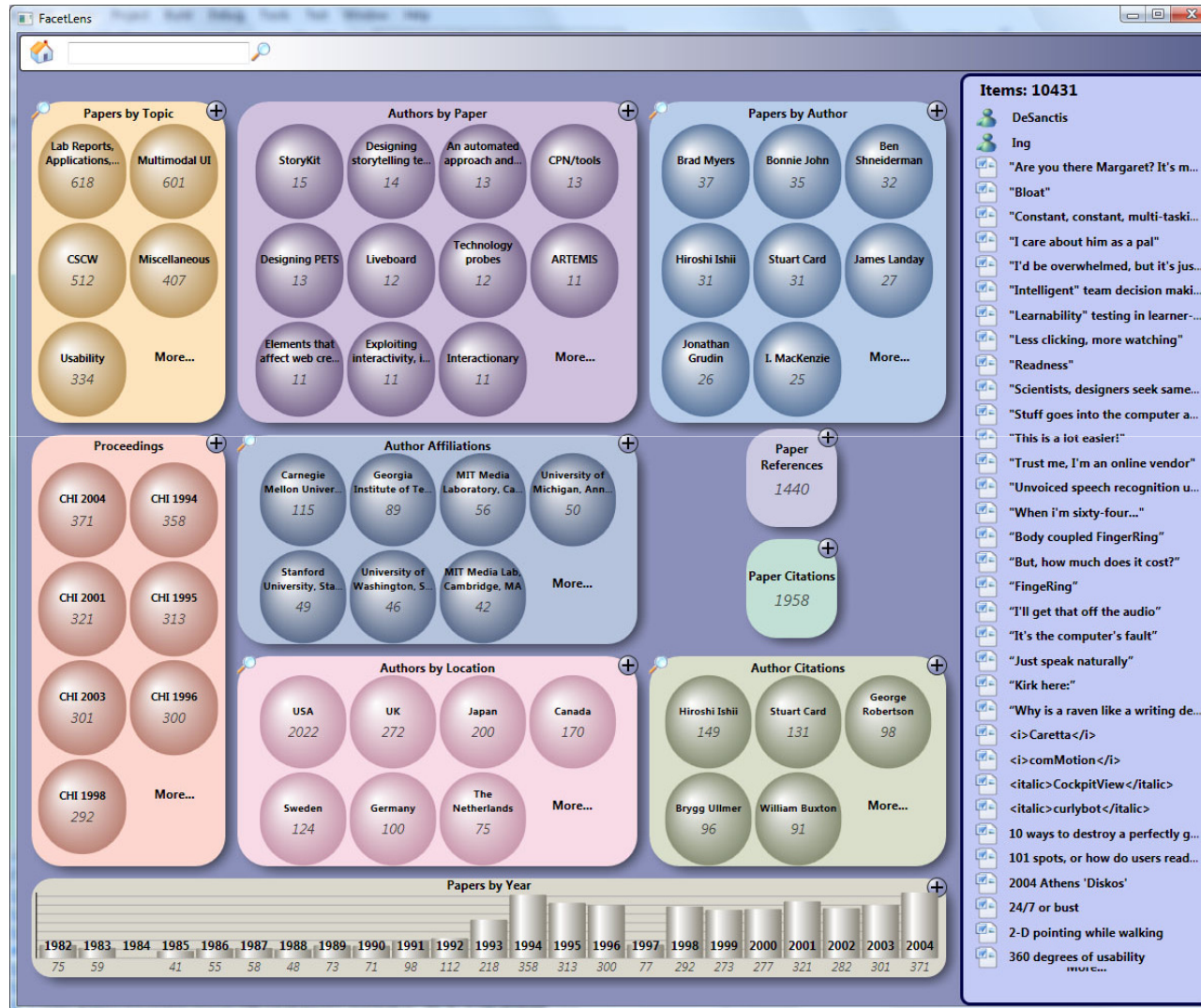
Award Winning Books



Fairy Tales and Folk Tales

Keywords in

What is faceted search?



FacetLens (Microsoft Research)

What is not faceted search?

The image shows a screenshot of the Snooth website, which is a wine rating and review platform. The top section features the Snooth logo and a search bar containing the text "khvanchkara". Below the search bar, the text "Khvanchkara Wine Ratings & Reviews" and "Results 1-10 of hundreds" is displayed. A message states: "Your search **khvanchkara** did not match any wines."

Below this, a "Refine Your Search" section is shown. It includes a checkbox for "Include out of stock items" (checked), a "Sort By" dropdown menu set to "Recommended", and a price range slider. The price range is currently set to "from us\$90 to us\$250+", which is circled in red. Other filters include "Vintage" (set to "any vintage"), "Show Wines Available In" (set to "All Countries"), and "Partner Search" (with checkboxes for "Winezap" and "Wine-Searcher"). A "Refine Search" button is located at the bottom right of this section.

Below the refined search section, another message states: "Your search **1998 khvanchkara** did not match any wines."

A second "Refine Your Search" section is shown below this. It includes the same "Include out of stock items" checkbox and "Sort By" dropdown. The price range slider is now set to "from us\$0 to us\$250+", which is also circled in red. The "Vintage" filter is now set to "1998", which is also circled in red. The "Show Wines Available In" and "Partner Search" options remain the same. A "Refine Search" button is located at the bottom right of this section.

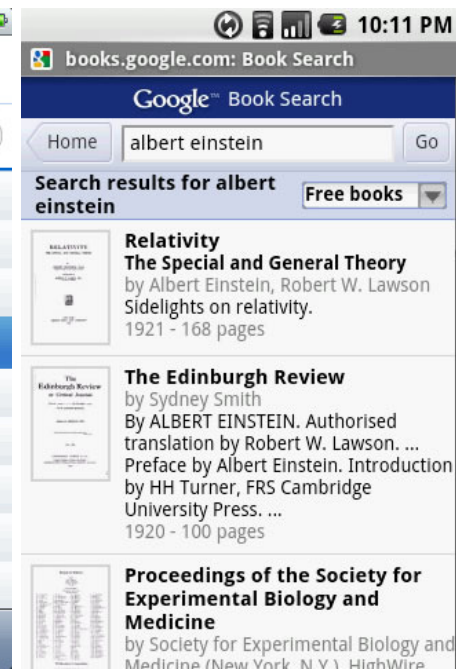
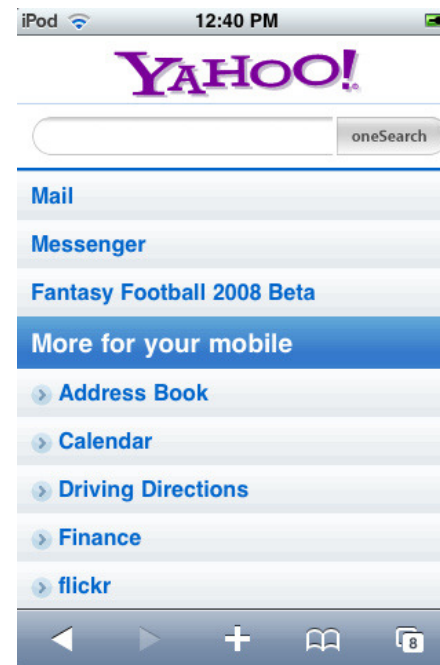
At the bottom of the page, there is a section for "User tags" with a list of wine descriptors: "fruits, cherries, sophisticated, jam, mouthful, color descriptors, unique, varietal". Below this list is a "Read reviews and more +" link. There is also a "Compare Prices and Buy" button.

Too many facets ? Too many facet values?

Information overload



Mobile interfaces



Facet selection: interface-based approach



<http://mspace.fm>

Available Columns

Decade	Year	Subject	Theme	Story Title
2000s	2000	Animal Science	Crime, Law & Justice	First On 5: New Breed of Super Rat
1990s	1999	Animation	Disaster & Accident	
1980s	1998	Applied Science	Economy, Business & Finance	
1970s	1997	Apprentices	Education	
1960s	1996	Archaeology	Environmental Issue	
1950s	1995	Archery	Health	
1940s	1994	Architecture	Human Interest	
1930s	1993	Armed Conflict	Politics	
1920s	1992	Arts (General)	Science & Technology	

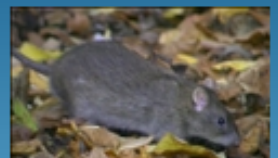
Available Columns: Contributor Role, Contributor, Segment, Month, Location, Type, Country, Publisher, Running Order, Person, Series Title, Keywords, Sub-Topic, Language, Day, Duration, Issue No.

You are currently browsing an online newsfilm archive

/ Animal Science / Environmental Issue (1 result)

Environmental Issue: All aspects of protection, damage, and condition of the ecosystem of the planet earth and its surroundings.

Prev (1-1 of 1) Next



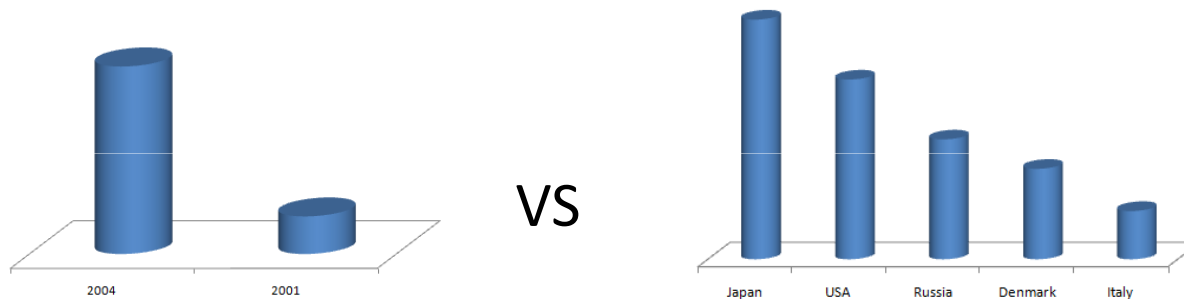
First On 5: New Breed of Super Rat
Growing fat on junk food, a new breed of rat is said to be on its way to our cities from Britain's countryside. The vermin carry lethal diseases, and experts say they could pose serious risks to health.

Redundancy-based facet selection

- Favor facets with high coverage in the result
 - Plenty of data formats in the enterprise
 - Metadata is not unified
 - There is no one classification scheme
 - **Select most frequent facets!**
- Avoid presenting highly correlating facets*
 - So, either **language** or **nationality**
- Consolidate similar facets:
 - author, editor, contributor => **people**

Interestingness-based facet selection

- Measure surprisingness of values distribution
- Favor facets with high-entropy distribution



$$Entropy = \sum_{i=1}^n P(w_i | R) \log P(w_i | R)$$

- Favor facets with query-specific distribution

$$Divergence = \sum_{i=1}^n (P(w_i | C) - P(w_i | R)) \log \frac{P(w_i | R)}{P(w_i | C)}$$

Facet values ranking

- Measure ***Relevance*** of facet value!
- Rank by frequency in result set
 - Most popular approach
- Rank by $\frac{P(f = v_i | R)}{P(f = v_i | C)}$
- Rank by aggregated document relevance:
 - Sum scores of all documents with value v_i

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

Collaborative facet values ranking (I)

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

$$\frac{\sum_{R \in User_k} P(f = v_i | R)}{P(f = v_i | C) \cdot |User_k|}$$

for result sets of all issued queries

Number of queries

Collaborative facet values ranking (II)

- Utilize collaborative filtering techniques*:

$$\alpha \frac{\sum_{R \in User_k} P(f = v_i | R)}{P(f = v_i | C) \cdot |User_k|} + (1 - \alpha) \underbrace{\frac{\sum_{User_j \in Users} \frac{\sum_{R \in User_j} P(f = v_i | R)}{|User_j|}}{P(f = v_i | C) \cdot |Users|}}_{\text{average preferences over all users}}$$

average preferences over all users

- Consider only users with similar tastes:

$$\alpha \frac{\sum_{R \in User_k} P(f = v_i | R)}{P(f = v_i | C) \cdot |User_k|} + (1 - \alpha) \frac{\sum_{User_j \in Users} \underbrace{sim(User_k, User_j)}_{\text{similarity}} \frac{\sum_{R \in User_j} P(f = v_i | R)}{|User_j|}}{P(f = v_i | C) \cdot |Users|}$$

For example, cosine similarity
or divergence of distributions

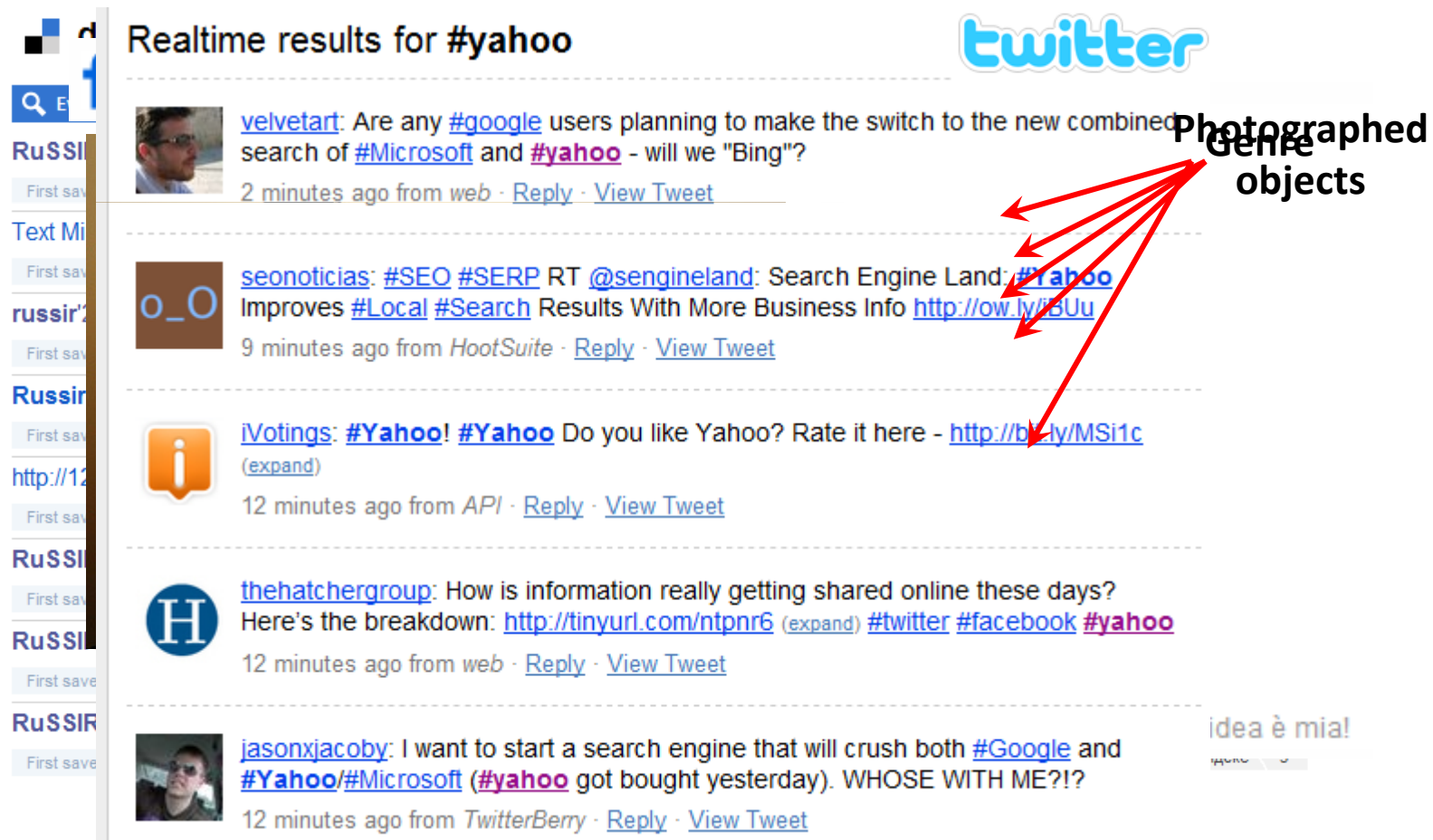
Summary

- Faceted search is must
 - 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
 - Giving answers right away
- 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



The image shows a screenshot of a Twitter search results page for the hashtag #yahoo. The page is titled "Realtime results for #yahoo" and features the Twitter logo. On the left, there is a sidebar with a search bar and a list of suggested filters. The main content area displays a list of tweets. Four red arrows originate from the text "Photographed objects" on the right and point to specific tweets, indicating that these tweets contain tagged objects. The tweets are as follows:

- 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](#)

On the right side of the image, the text "Photographed objects" is written in black, with four red arrows pointing to the tweets. Below this text, the phrase "idea è mia!" is visible.

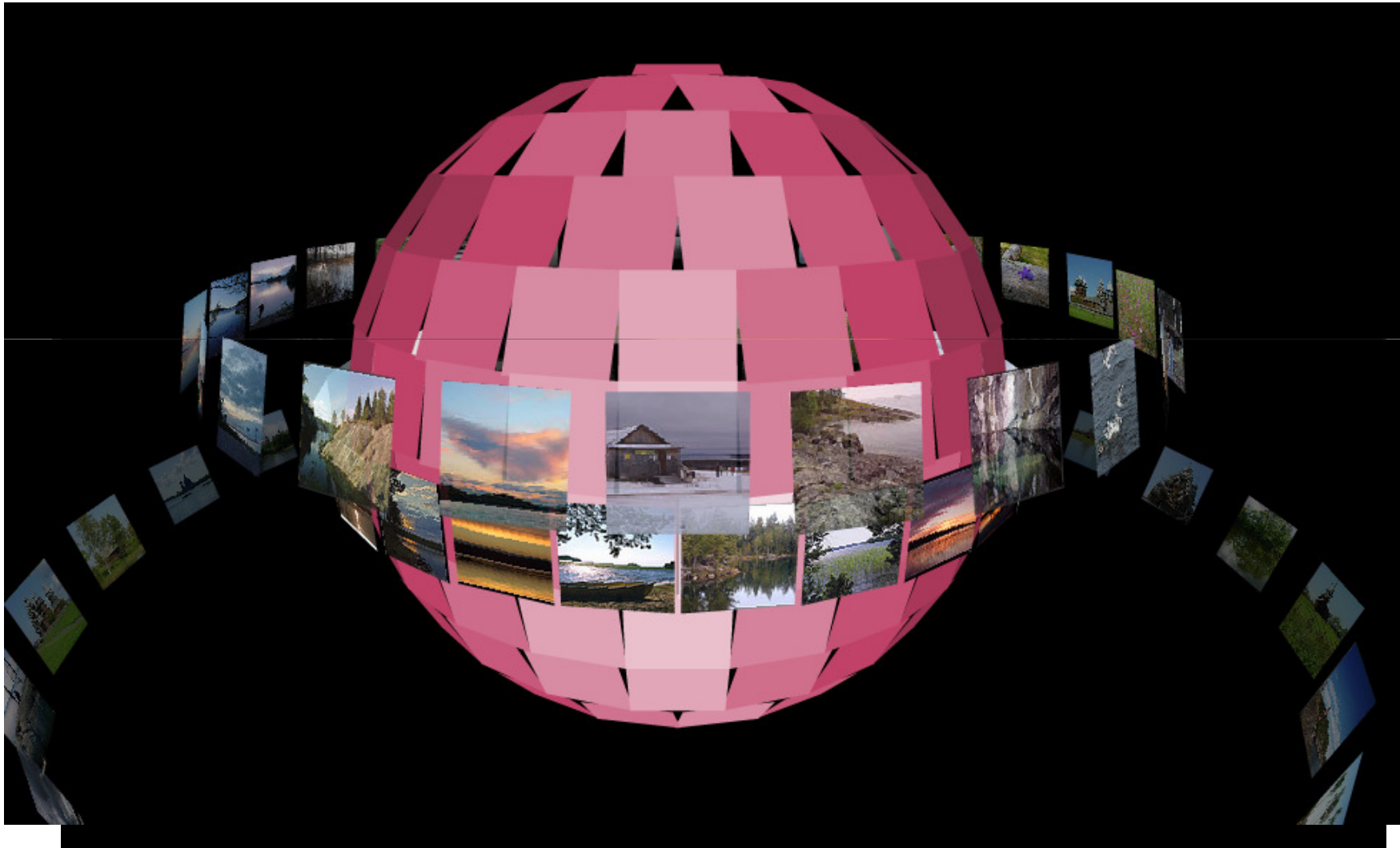
Tagging

- Disadvantages:
 - Not 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



Tag space



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

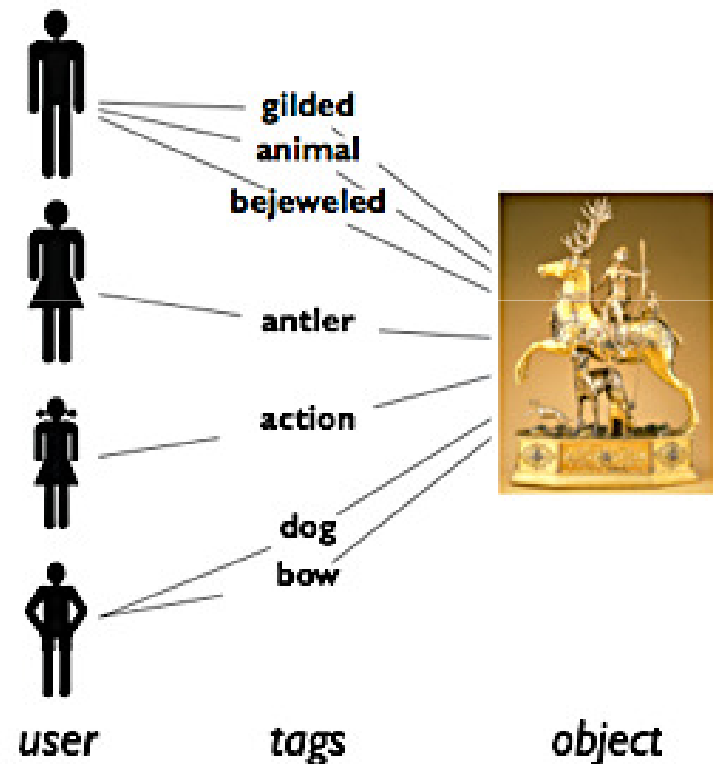
– inverted tag frequency in the collection

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

Explore / Tags / karelia / clusters

Jump to:



[russia](#), [lake](#), [kizhi](#), [nature](#), [church](#),
[landscape](#), [island](#), [sky](#), [petrozavodsk](#), [sunset](#)

➔ [See more in this cluster...](#)



[finland](#), [suomi](#), [wood](#), [clouds](#), [trees](#), [sun](#),
[forest](#), [helsinki](#), [karijala](#), [joensuu](#)

➔ [See more in this cluster...](#)



[water](#), [north](#), [summer](#), [rocks](#)

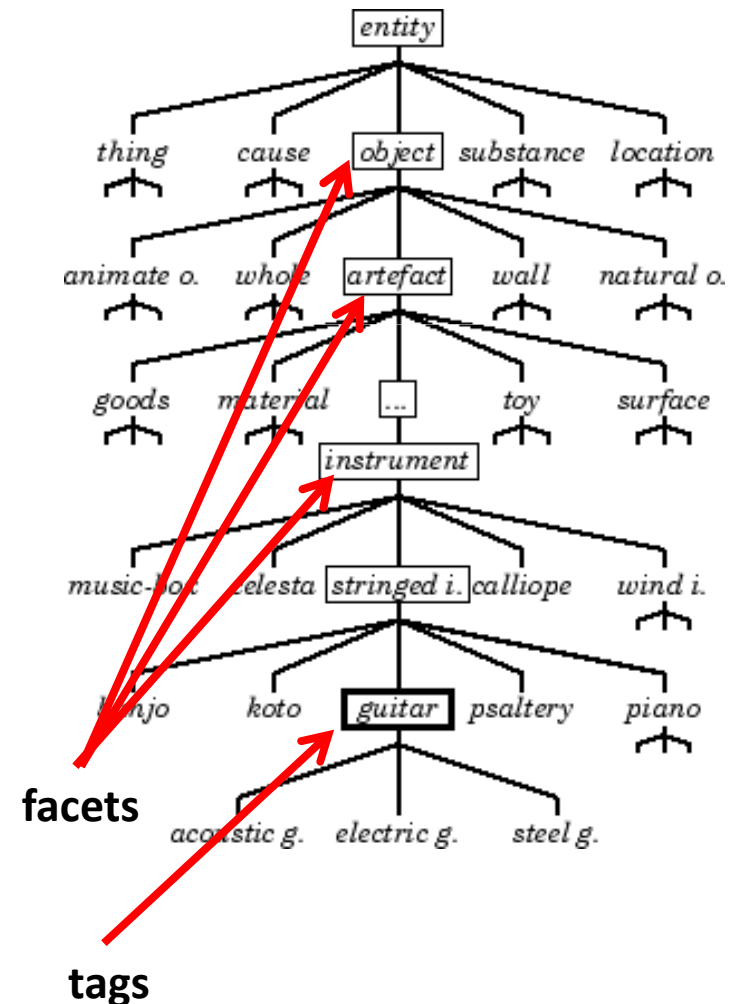
➔ [See more in this cluster...](#)

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 in this graph



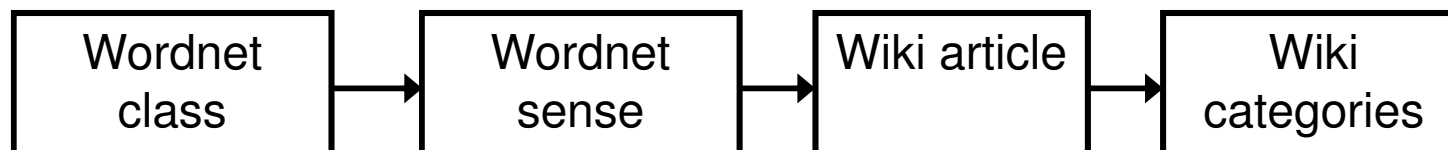
Tag classification with Wikipedia (I)

- Wordnet has nice selection of classes (facets)
- ... but not so many entities (facet values)
- 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 classification with Wikipedia (II)

- Direct Tag => Wiki matches may be too imprecise:
 - So, use only anchor text or titles
- Some Wikis are direct match with Wordnet senses!
 - “Guitar” => en.wikipedia.org/wiki/Guitar
 - Use these matches as training data
 - Build classifier for each Wordnet noun class (~25 classes)
- What features should describe Wordnet classes?
 - Using terms as features would introduce too much **noise** and problems with **dimensionality**
 - Categories of wiki-articles are better choice!



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

TagExplorer
Powered by Flickr

YAHOO!
RESEARCH

karelia

Query: [karelia](#) ✕

locations	subjects	activities
finland + kizhi + ladoga + onega +	lake + landscape + nature +	travel +
petrozavodsk + russia +	names	time
sortavala + suomi + valaam +	buddhism +	2006 + summer +

Photo Results



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

Interaction with faceted search system

- Traditional way:
 - Typing, typing, typing...
 - For the sake of query reformulation
- Faceted search?



Mousing & Browsing

Filtering – all search tags are made equal

Continue narrowing

Continue Start

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

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

Refine Search With Tags
No relevant tags

Refined Tag:

- garlic
- white wine, flour
- beef tenderloin
- tomato puree
- bay leaves
- tofu

Latest Entries(1)
Sorted by: Recent | Relevant | RSS

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

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

1 Tomato Puree Recipe at Epicurious.com

1 Steak Recipe at Epicurious.com

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

Tag weights

Tag feedback



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

Authentic Russian Recipes, Cuisine and Cooking
 recipes food russian cuisine cooking recipe russia reference kosthold europa
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 europa 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

$$P('food' | Q) = \frac{1}{5}$$

$$P('recipes' | Q) = \frac{1}{5}$$

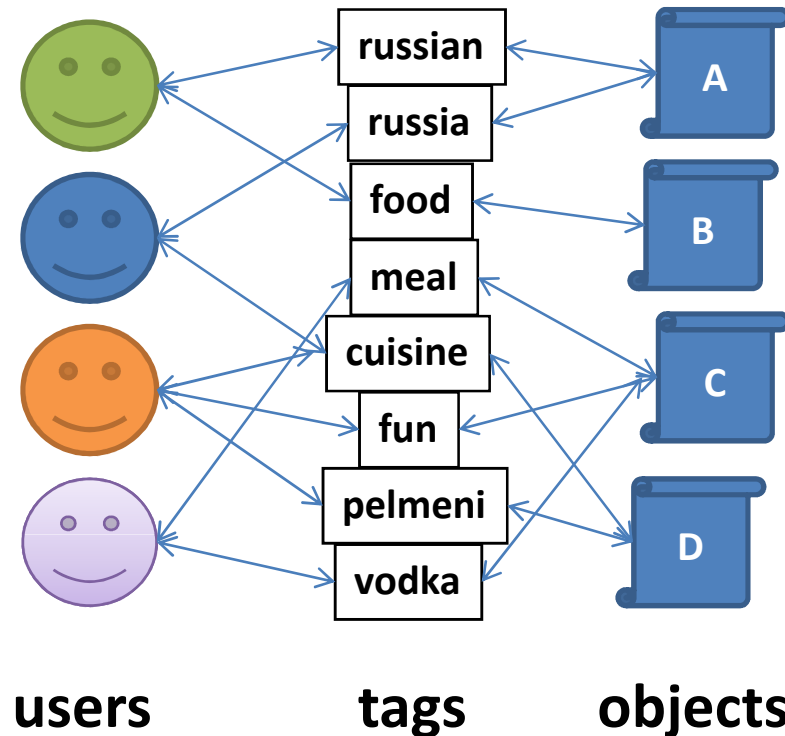
$$P('russia' | Q) = \frac{3}{5}$$

Irrelevance lang. model

-drinking -health -work -humor

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:

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

- How to define **distance** between nodes:

- Length of shortest path
- Number of shortest paths (of certain length)
- Distance-based similarity:

$$\sum_{\substack{\textit{path}(\textit{tag}, \textit{document}) \\ \in \textit{shortestpaths}}} c^{-\textit{length}(\textit{path})}$$

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)

The screenshot shows the DMOZ website interface. At the top, there's a green header with the DMOZ logo and the text 'open directory project'. To the right, it says 'In partnership with AOL search'. Below the header is a navigation bar with links: 'about dmoz', 'dmoz blog', 'suggest URL', 'help', 'link', and 'editor login'. A search bar is located below the navigation bar. The main content area is divided into several sections. On the left, there are category links: 'Arts' (with sub-links: Movies, Television, Music...), 'Business' (with sub-links: Jobs, Real Estate), 'Games' (with sub-links: Video Games, RPGs, Gambling...), 'Health' (with sub-links: Fitness, Medicine), 'Kids and Teens' (with sub-links: Arts, School Topics), 'Reference' (with sub-links: Maps, Educational Resources), and 'Shopping' (with sub-links: Clothing, Food, Gifts). The central part of the page features a large section titled 'Top: Science (110,319)'. Below this title is a list of sub-categories: 'Agriculture (3,874)', 'Environment (6,529)', 'Life Sciences (10,504)', 'Mathematics (4,528)', 'Physics (743)', 'Social Sciences (21,381)', 'Technology (11,372)', and 'Women@ (174)'. To the right of this list is a navigation bar with links: '[A | B | C | D | E | F | G | H | I | J]'. Below the 'Top: Science' section is another large section titled 'Top: Computers: Computer Science (2,111)'. This section contains two columns of sub-categories. The left column lists: 'Academic Departments (583)', 'Conferences (223)', 'Directories (8)', 'Organizations (75)', 'Artificial Intelligence@ (1,416)', 'Artificial Life@ (259)', 'Computational Geometry@ (66)', 'Computer Graphics (44)', and 'Database Theory (92)'. The right column lists: 'People (300)', 'Publications (81)', 'Reference (5)', 'Research Institutes (77)', 'Distributed Computing (245)', 'Parallel Computing@ (425)', 'Software Engineering@ (134)', and 'Theoretical (378)'.

dmaz open directory project

In partnership with
AOL search

[about dmoz](#) | [dmoz blog](#) | [suggest URL](#) | [help](#) | [link](#) | [editor login](#)

Search [advanced](#)

Top: Science (110,319)

[[A](#) | [B](#) | [C](#) | [D](#) | [E](#) | [F](#) | [G](#) | [H](#) | [I](#) | [J](#)]

- [Agriculture](#) (3,874)
- [Environment](#) (6,529)
- [Life Sciences](#) (10,504)
- [Mathematics](#) (4,528)
- [Physics](#) (743)
- [Social Sciences](#) (21,381)
- [Technology](#) (11,372)
- [Women@](#) (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)

Arts
[Movies](#), [Television](#), [Music](#)...

Business
[Jobs](#), [Real Estate](#)

Games
[Video Games](#), [RPGs](#), [Gambling](#)...

Health
[Fitness](#), [Medicine](#)

Kids and Teens
[Arts](#), [School Topics](#)

Reference
[Maps](#), [Educational Resources](#)

Shopping
[Clothing](#), [Food](#), [Gifts](#)

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

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

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

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

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

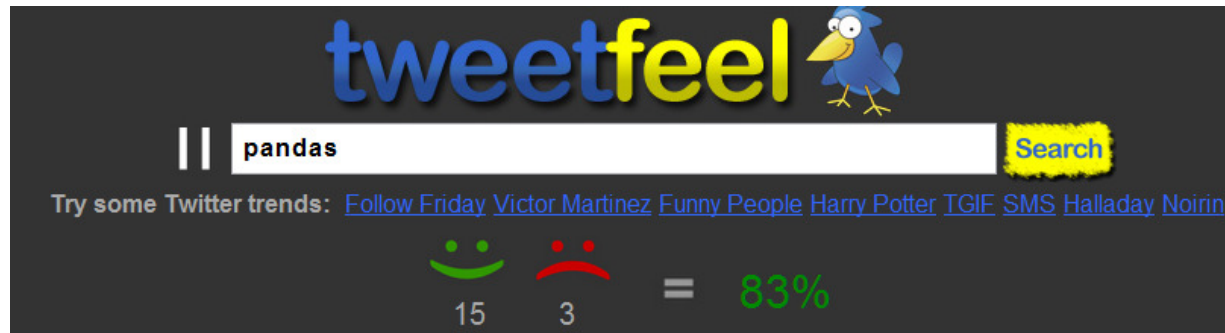
$$P(Class | D) \Rightarrow P(Class | D, Q) = P(Class | D)P(Class | Q)$$

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)



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**!!!

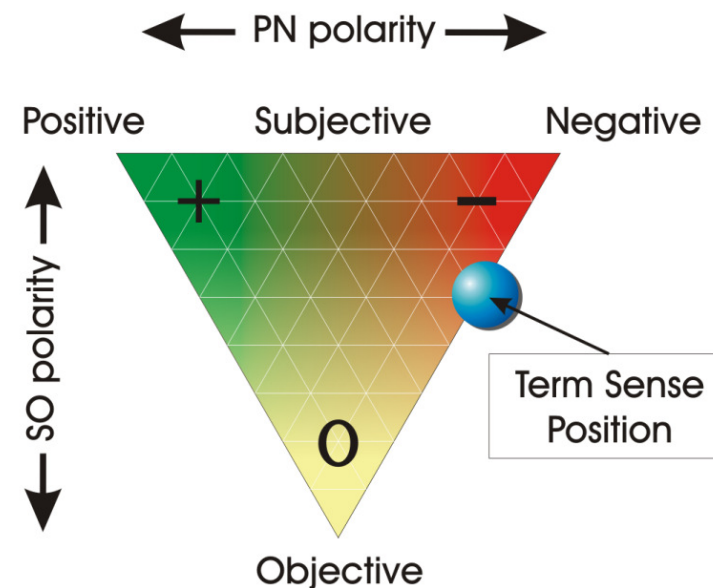
@Amber_Lily omg a panda!!! i love **pandas** and you know what! when i'm older i wanna be a panda :) well know!!!

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**
- Machine learning based approaches:
 - Build classification models for **texts** and **terms**:
 - Objective vs. Subjective
 - Positive vs. Negative
 - Better for each domain
 - Better use 2,3-grams
 - “long battery life”
 - “long execution time”



Categorization by location (I)

- Some documents, photos, videos, tweets...
 - are location agnostic and **some are not!**



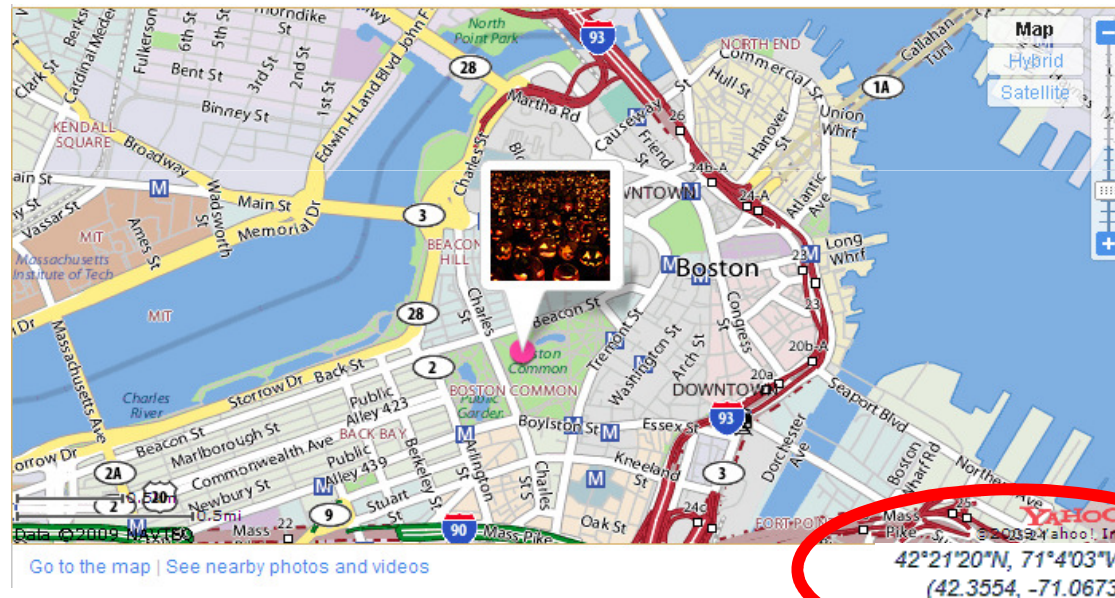
kitchen cats dogs



russia river brownbear

Categorization by location (II)

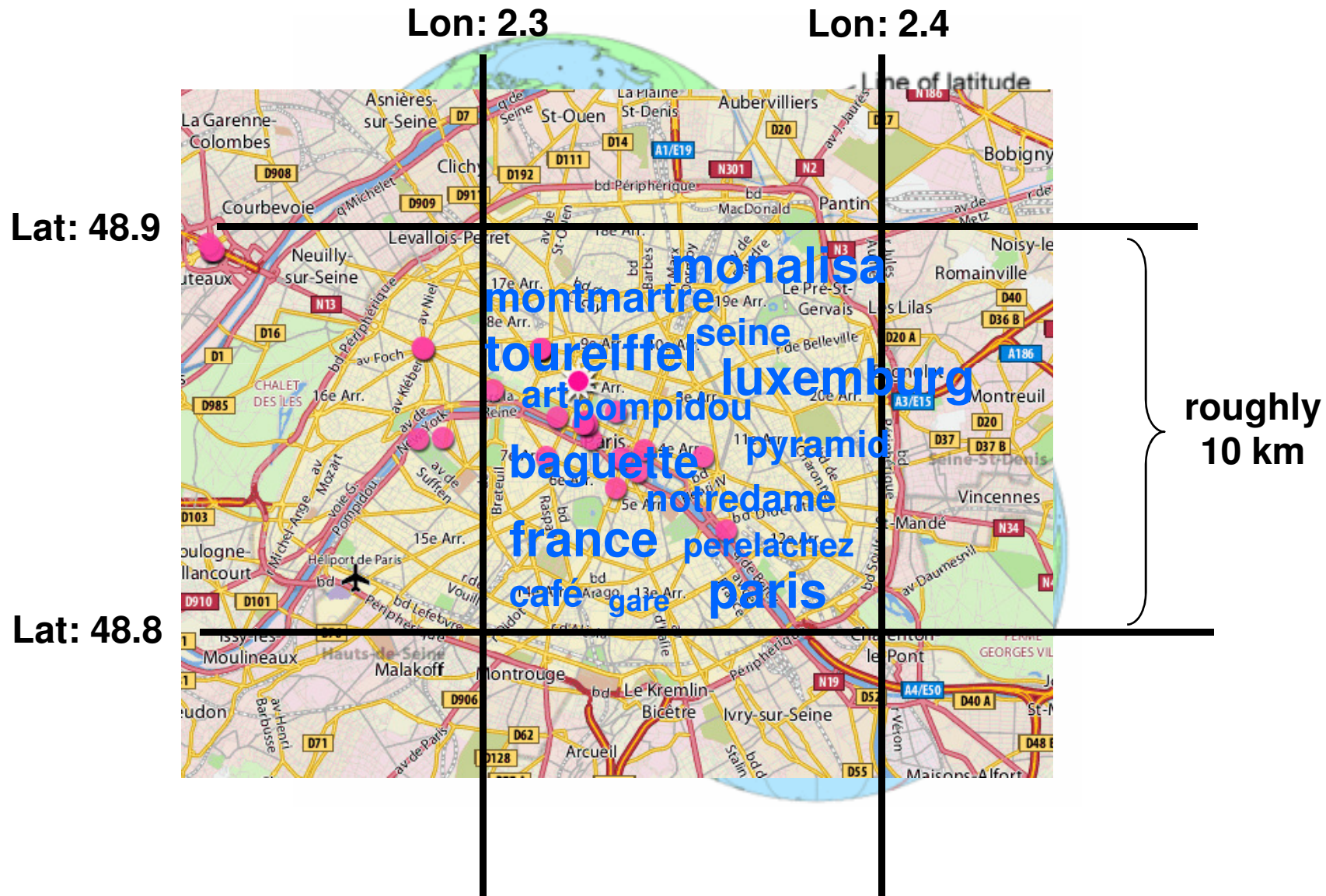
- **Some documents are geo-tagged**
 - There are more than 100 millions of them at Flickr!
 - Are we done?



geo-tags: latitude, longitude

Around 96% of Flickr photos are not geo-tagged!

Categorization by location (III)



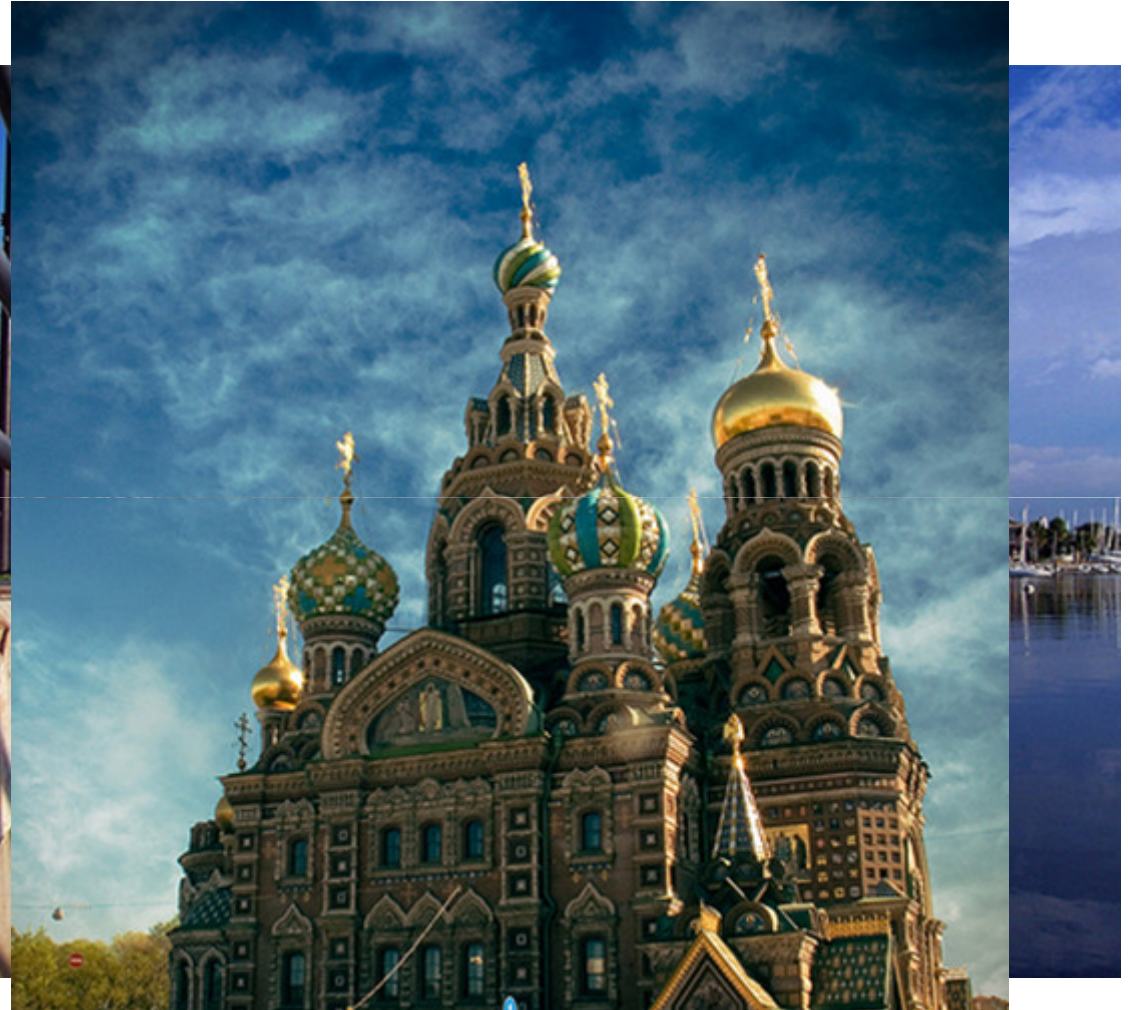
Categorization by location (IV)

St. Petersburg



Popular tags

russia, church, bridge, cathedral,
hermitage, russian, winter, baltic,
warpedtour, bird, petersburg, bay,
light, neva, petersburg, water,
tampabay, vnoypark, pelican, water,
hermitage, russian, winter, baltic,
warpedtour, bird, petersburg, bay



Categorization by location (V)

- ▶ 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 :

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

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

- ▶ But there is much more we can do*:
 - ▶ Consider spatial ambiguity of tags?
 - ▶ Consider neighboring locations?
 - ▶ Consider that some of them are toponyms?
- ▶ Apply for place non-tagged photos? Not only photos?

***Placing Flickr Photos on a Map.**

Serdyukov P., Murdock V., van Zwol R. SIGIR 2009

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*...
- Relevance of phrase $l = w_1 w_2 w_3$ to document:

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

- Relevance of document to phrase. **Minimize:**

$$Dist(l, D) = - \sum_w P(w | l) \frac{P(w | l)}{P(w | D)}$$

Over all docs where l occurs

- Uniqueness on document level. **Maximize:**

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

- Uniqueness on collection level. **Maximize:**

$$\frac{1}{|C| - 1} \sum_{D' \neq D} Dist(l, D')$$

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

Metadata extraction (III)

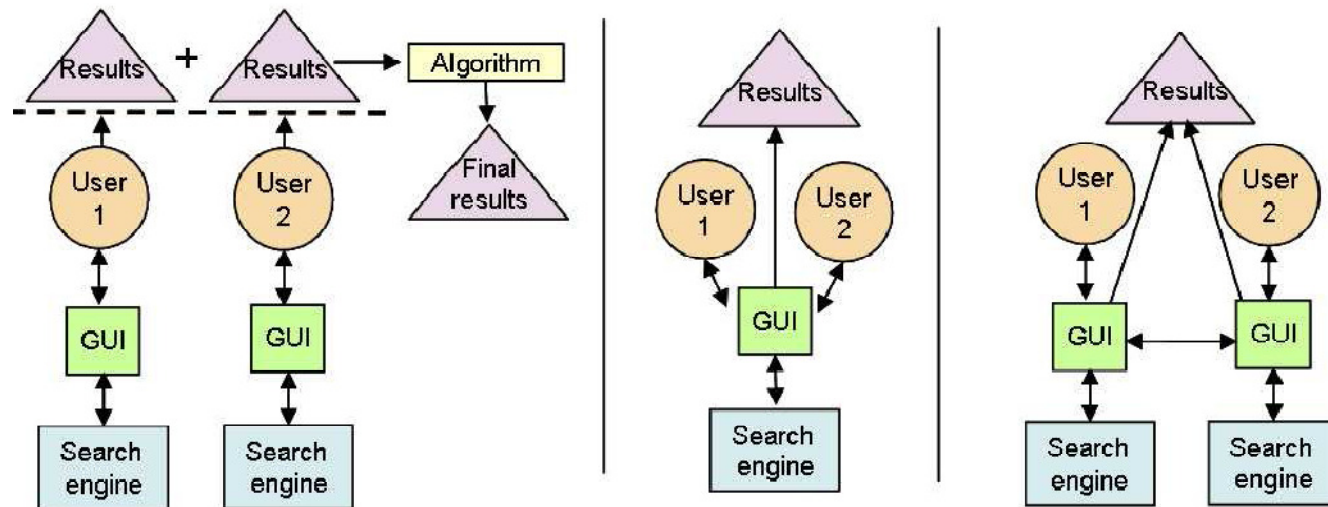
- So far not query-driven, right?
- Let's move away from bag-of-words
- Possible algorithm:
 - Cluster sentences in a document
 - Select keywords for each cluster (as shown)
 - Find cluster(s) most relevant to a query
 - Represent document by keywords from relevant cluster(s)
- Just consider text windows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

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...

What about the Future?

Collaborative exploratory search



- Collaborative search*:
 - Many queries, many people, one information goal
 - How to suggest and route queries?
 - How to route documents for evaluation?
 - How to aggregate opinions on documents?

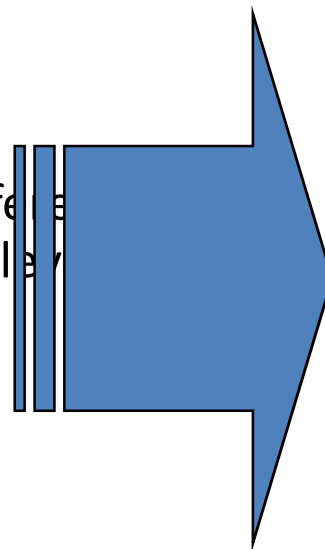
* **Algorithmic mediation for collaborative exploratory search.** J. Pickens et. al. SIGIR 08

Aggregated exploratory search

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

Query “hairspray”

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



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

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

References: advanced exploratory search

- Collaborative search:
 - http://en.wikipedia.org/wiki/Collaborative_search_engine
 - **Algorithmic mediation for collaborative exploratory search.** J. Pickens et. al. SIGIR 2008
 - **Discovering and Using Groups to Improve Personalized Search.** J. Teevan. WSDM 2009
 - Download and play:
<http://research.microsoft.com/en-us/um/redmond/projects/searchtogether/>
- Aggregated search:
 - **Integration of News Content into Web Results.** F. Diaz. WSDM 2009. (Best paper award)
 - **Sources of evidence for vertical selection.** J. Arguello et. al. SIGIR 2009. (Best paper award)

- 4-years
- EU-proj
- Summ



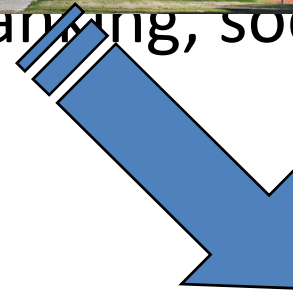
ft

entity ranking, social search

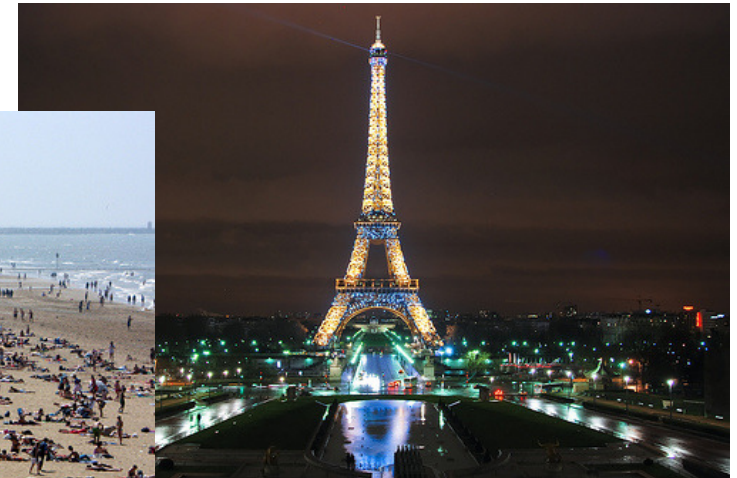
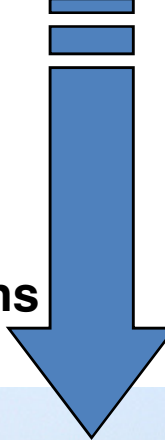
60 mins



3.5 hours



40 mins



Enterprise and Desktop Search

Lecture 4: Expert finding

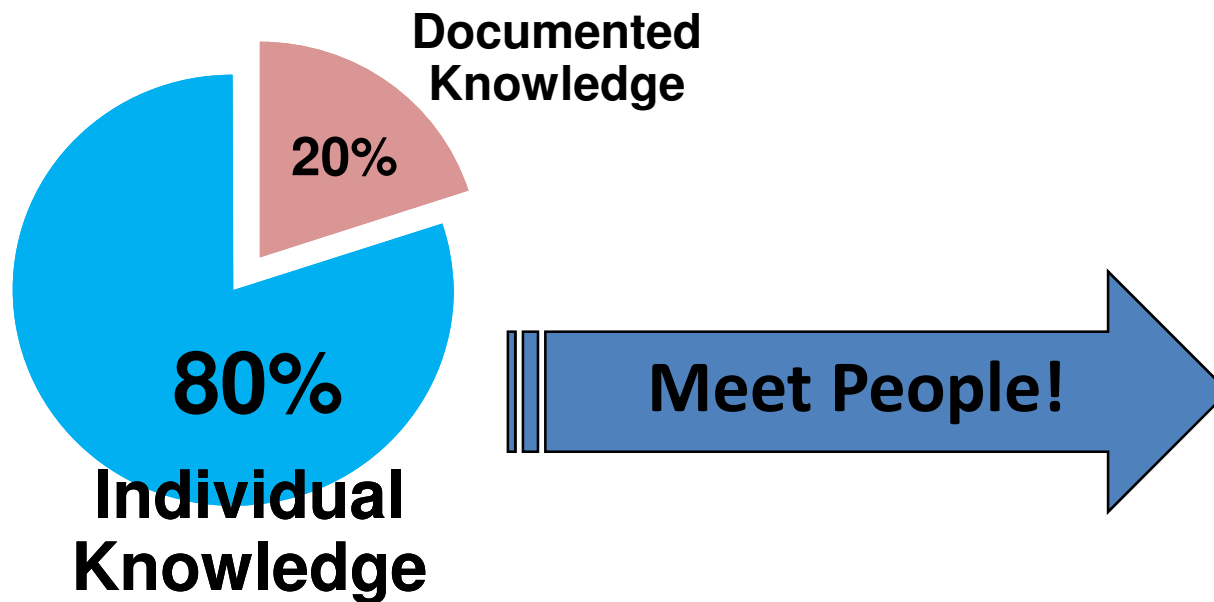
Pavel Dmitriev, Pavel Serdyukov, Sergey Chernov

Outline

- The need for expert finding
- State-of-the-art approaches
- Advanced techniques:
 - Mining for personal language models
 - Proximity-aware expert finding
 - Looking for additional evidence in the enterprise
 - Looking for additional evidence on the Web
- Future challenges

Search for experts

- Some knowledge is not easy to find
 - Not stored in documents
 - Not stored in databases
 - **It is stored in peoples' minds!**



Search for experts

- Let's search for ~~documents~~ people
- Who is ~~relevant~~ expert on topic X?
- Basically, a special case of **faceted search**
 - Facets “people”, “employees”
- Try some expert search right now:



Search in personal profiles

Search for experts in retrieval

Search only among known people

Working in Europe

Ever worked at Yahoo!

Faceted search for experts!

The screenshot shows a LinkedIn search interface. The search bar contains the word 'retrieval'. To the right, the 'Refine By' section is expanded, showing several filters. The 'Relationship' filter is set to '1st Connections (6)'. The 'Location' filter is set to 'Geneva Area, Switzerland (1)'. The 'Past Company' filter is set to 'Yahoo! (6)'. The search results list five profiles: Gleb Skobeltsyn (Post-Doc Engineer at Google), Vanessa Murdock (Researcher at Yahoo! Research Barcelona), Vassilis Plachouras (Researcher in Information Retrieval), Paul - Alexandru Chirita (Engineering Manager at Adobe Systems Inc.), and Maarten Clements (Ph.D. Researcher at Delft University of Technology). Red arrows point from the text annotations to the corresponding filters in the 'Refine By' section.

Search

retrieval

Refine By

Current Company

Relationship

☐ All LinkedIn Members

☒ 1st Connections (6)

☐ 2nd Connections (0)

☐ Group Members (4)

☐ 3rd + Everyone Else (0)

Industry

Location

☐ All Locations

☐ Montreal, Canada Area (1)

☒ Geneva Area, Switzerland (1)

☒ Barcelona Area, Spain (1)

☒ Greece (1)

☒ Amsterdam Area, Netherlands (1)

☒ The Hague Area, Netherlands (1)

☒ Romania (1)

☐ Greater Atlanta Area (1)

Enter location name

show less...

Past Company

☐ All Companies

☒ Yahoo! (6)

☐ University of Amsterdam (4)

☐ Universitat Pompeu Fabra (2)

☐ CWI (2)

☐ Delft University of Technology

Gleb Skobeltsyn 1st

Post-Doc Engineer at Google

Geneva Area, Switzerland | Information Technology and Services

In Common: ▶ 29 shared connections ▶ 1 shared group

Vanessa Murdock 1st

Researcher at Yahoo! Research Barcelona

Barcelona Area, Spain | Research

In Common: ▶ 29 shared connections

Vassilis Plachouras 1st

Researcher in Information Retrieval

Greece | Research

In Common: ▶ 26 shared connections

Paul - Alexandru Chirita 1st

Engineering Manager at Adobe Systems Inc.

Romania | Internet

In Common: ▶ 19 shared connections ▶ 1 shared group

Maarten Clements 1st

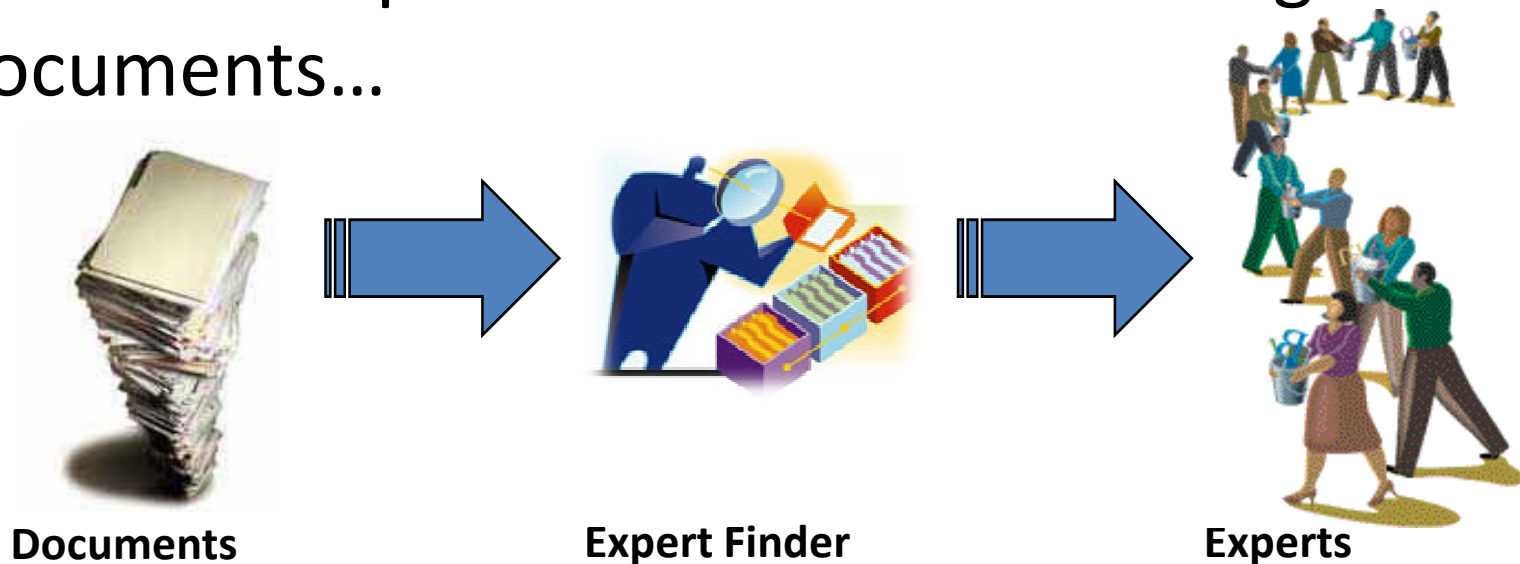
Ph.D. Researcher at Delft University of Technology

The Hague Area, Netherlands | Information Technology and Services


In Common: ▶ 31 shared connections ▶ 1 shared group

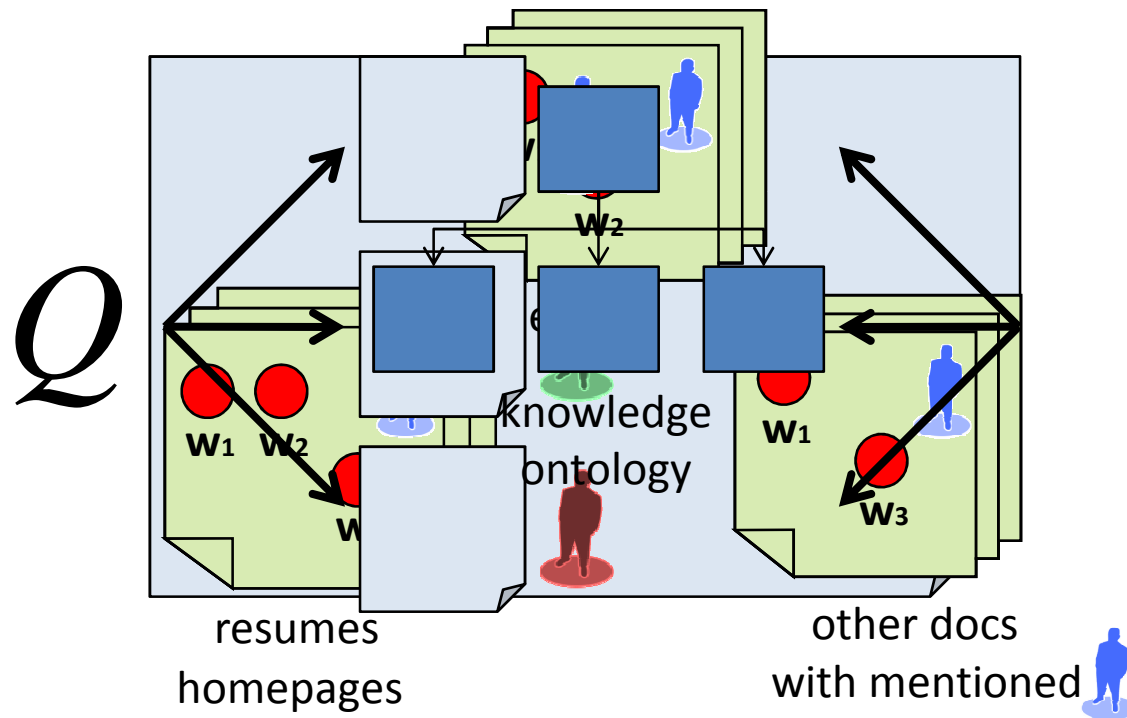
Expert finding via document analysis

- Analyze self-made profiles?
 - Need some enthusiasm to maintain
 - Subjective due to over/under-estimation
- Sleuth for expertise evidence in existing documents...




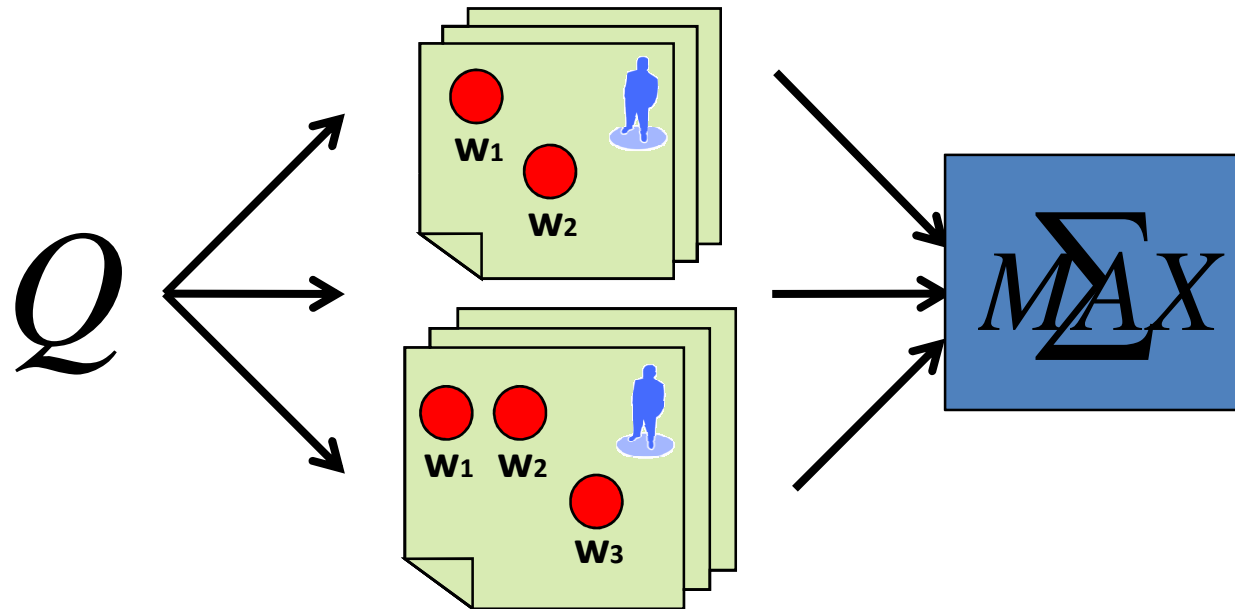
Profile-based expert finding

- 1st step: Build a personal profile for 
- 2nd step: Match it to a query as a document



Document-centric expert finding

- 1st step: Rank all documents with 
- 2nd step: Aggregate document scores



- Remember facet values ranking?

Popular datasets

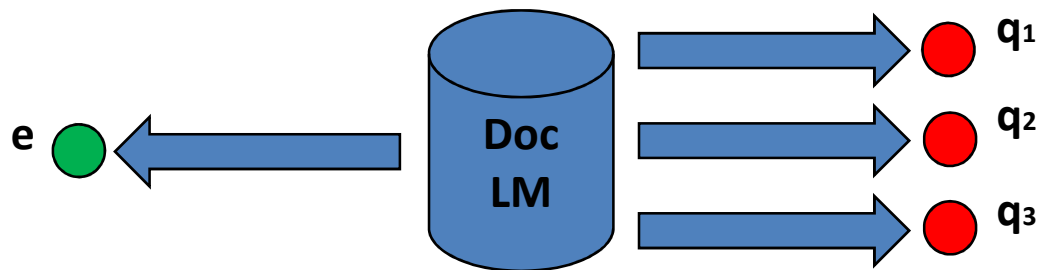
- **TREC 2005-2006: W3C data**
 - The largest part consists of mailing lists
 - About 1000 candidates provided
 - Judgments made by participants (50 queries)
 - Really many “experts” per query
- **TREC 2007-2008: CSIRO data**
 - www.csiro.au crawl
 - About 3500 candidates (just all persons mentioned)
 - Judgments made by the organization itself (49 queries)
 - Very few “experts” (key persons) per query
- **Three measures are analyzed**
 - MAP (Mean Average Precision) and P@5
 - MRR (Mean Reciprocal Rank)

Going beyond bag-of-words (I)

- **Popular Intuition:**

Expertise is proportional to the degree
of query terms and the person's
co-occurrence

- **Classic document-centric approach*:**

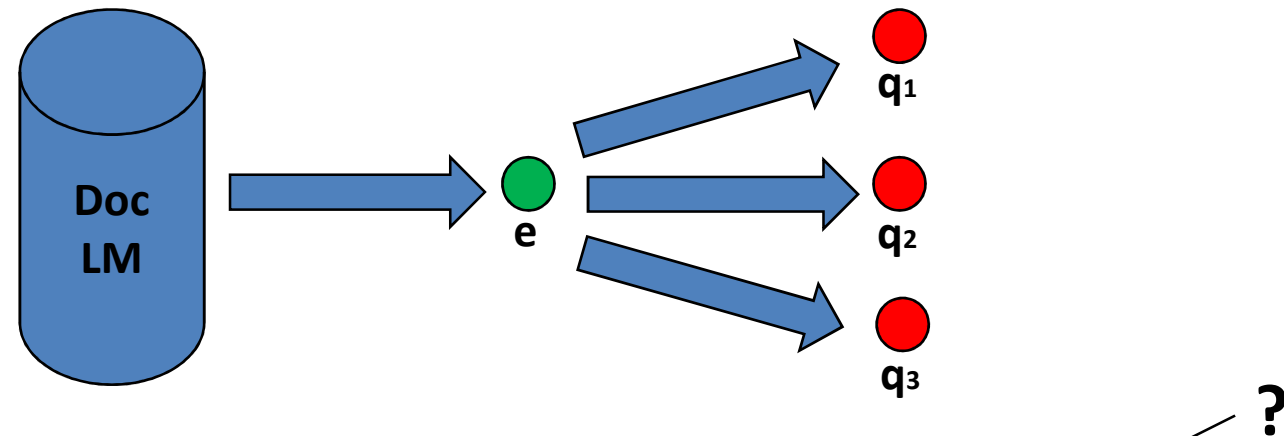


$$P(e, Q) = \sum_D P(e, Q | D) P(D) = \sum_D P(e | D) \underbrace{P(Q | D) P(D)}_{\approx P(\text{Relevance} | D)}$$

*A language modeling framework for expert finding. Balog et. al. SIGIR 06

Going beyond bag-of-words (II)

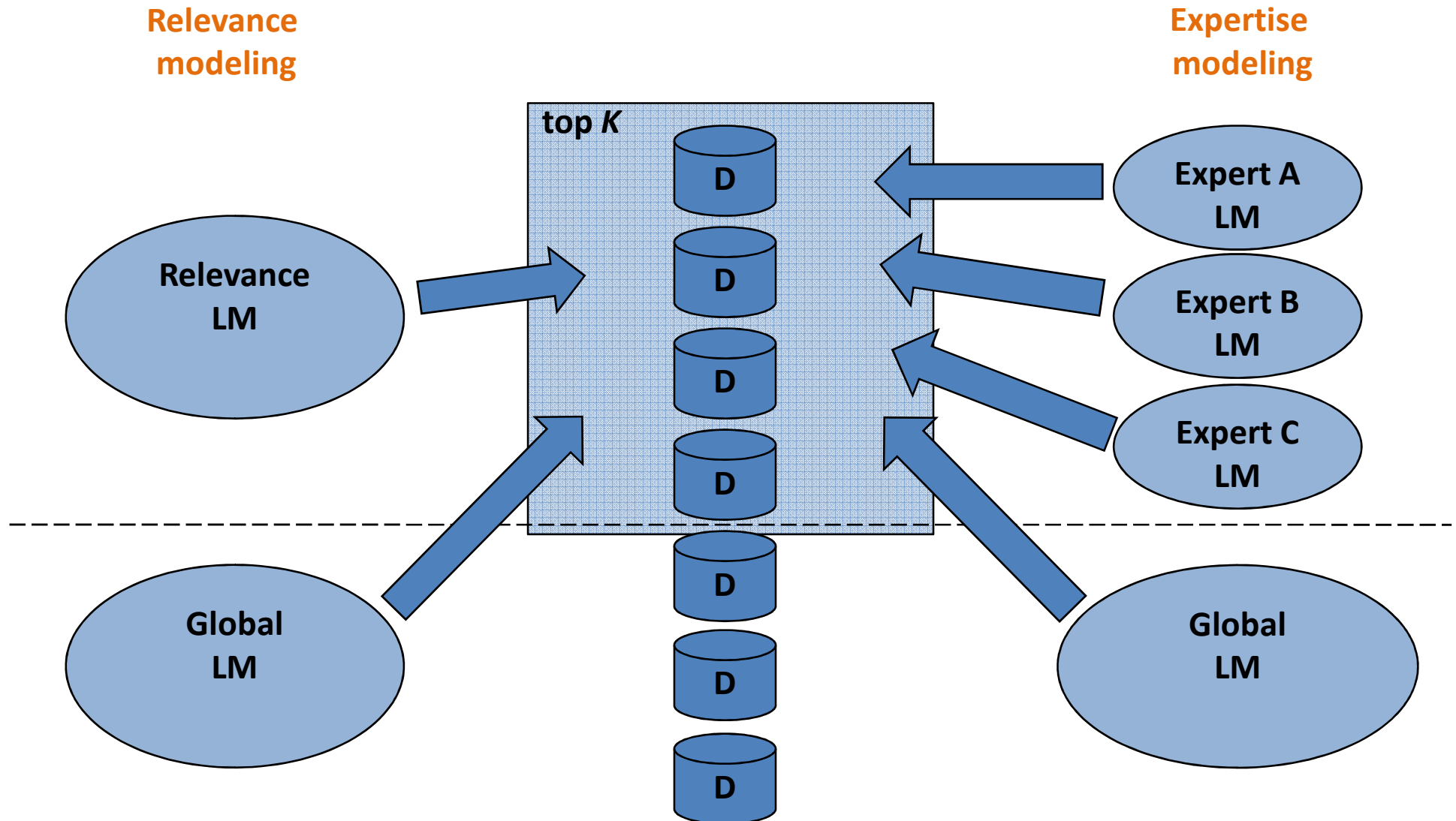
- Full Independence is not realistic
- Persons are responsible for terms!



$$\sum_D P(e, Q | D) P(D) = \sum_D P(Q | e) P(e | D) P(D)$$

**Modeling documents as mixtures of persons
for expert finding.** Serdyukov and Hiemstra. ECIR 2008

Mining personal language models (I)



Mining personal language models (II)

- Likelihood of Top K retrieved documents

$$\prod_D \prod_{w \in D} ((1 - \lambda_G) (\sum_{i=1}^m P(w | e_i) P(e_i | D)) + \lambda_G P(w | G))^{c(w, D)}$$

unknown params

- $c(w, D)$ - count of terms w in document D
- λ_G - probability of term generation from the Global LM
- $P(e | D)$? Previously, was inferred from:
 - Importance of a document's field
 - Number of candidates in a document

Mining personal language models (III)

- Steps for EM iterations:

E – step :

$$P(e | w, D) = \frac{(1 - \lambda_G) P(e | D) P(w | e)}{(1 - \lambda_G) (\sum_{i=1}^m P(e_i | D) P(w | e_i)) + \lambda_G P(w | G)}$$

M – step :

$$P(w | e) = \frac{\sum_{D \in TopK} c(w, D) P(e | w, D)}{\sum_w \sum_{D \in TopK} c(w, D) P(e | w, D)}$$

unfixed

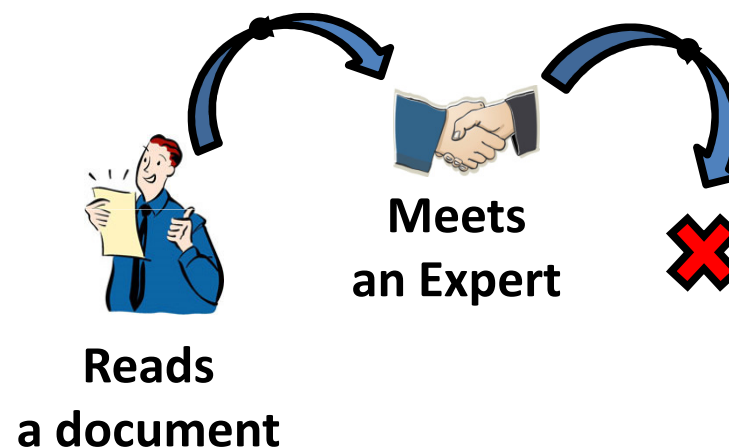
$$P(e | D) = \frac{1 + \sum_{w \in D} c(w, D) P(e | w, D)}{m + \sum_{i=1}^m \sum_{w \in D} c(w, D) P(e_i | w, D)}$$

Going beyond “personal” documents

- Look at the classic approach again:

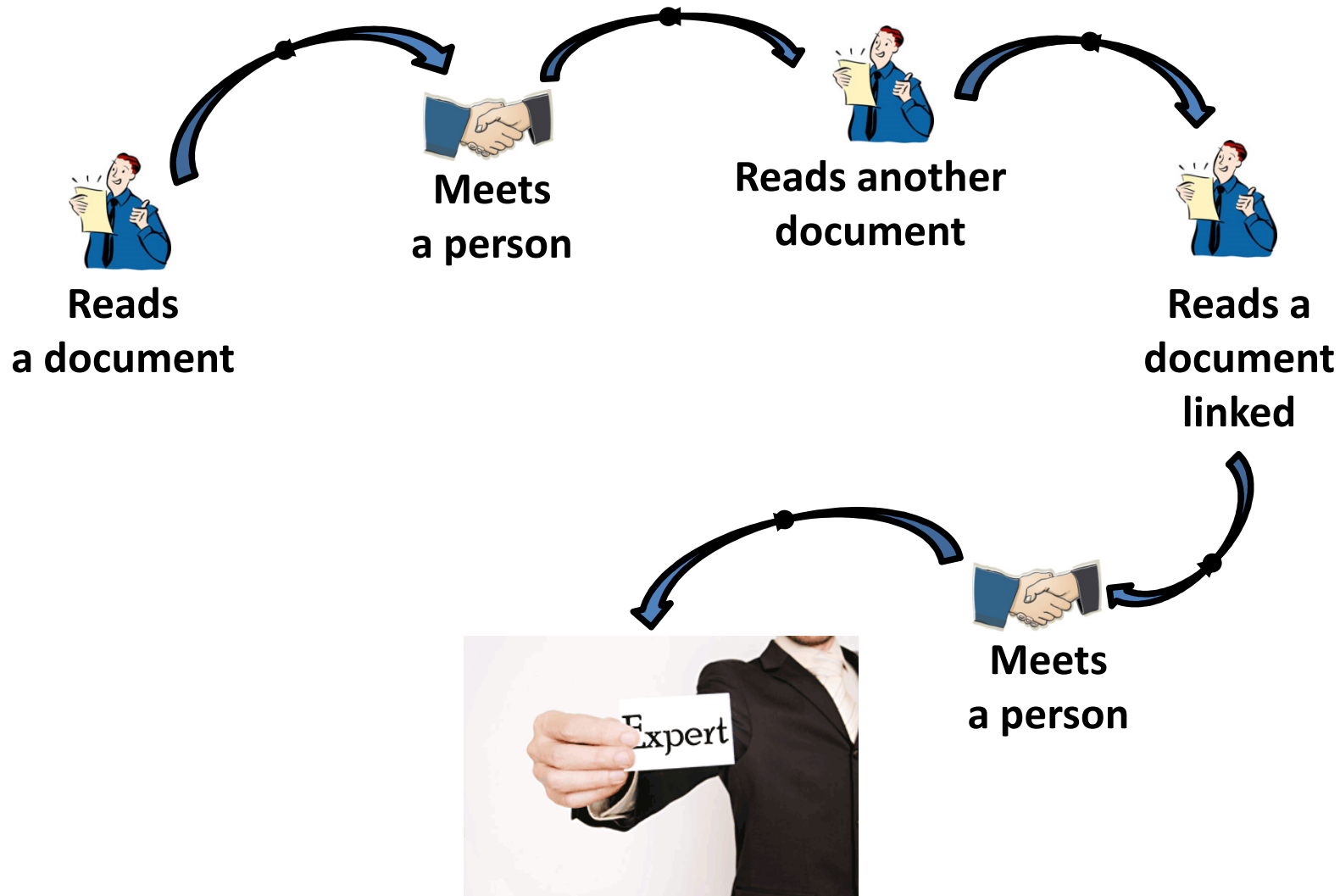
$$Expertise(e) = \sum_{D \in TopK} P(e|D)P(Q|D)P(D)$$

1. User selects a document from the top
2. User selects a person from the document
3. Finished? Well, not in exploratory mood

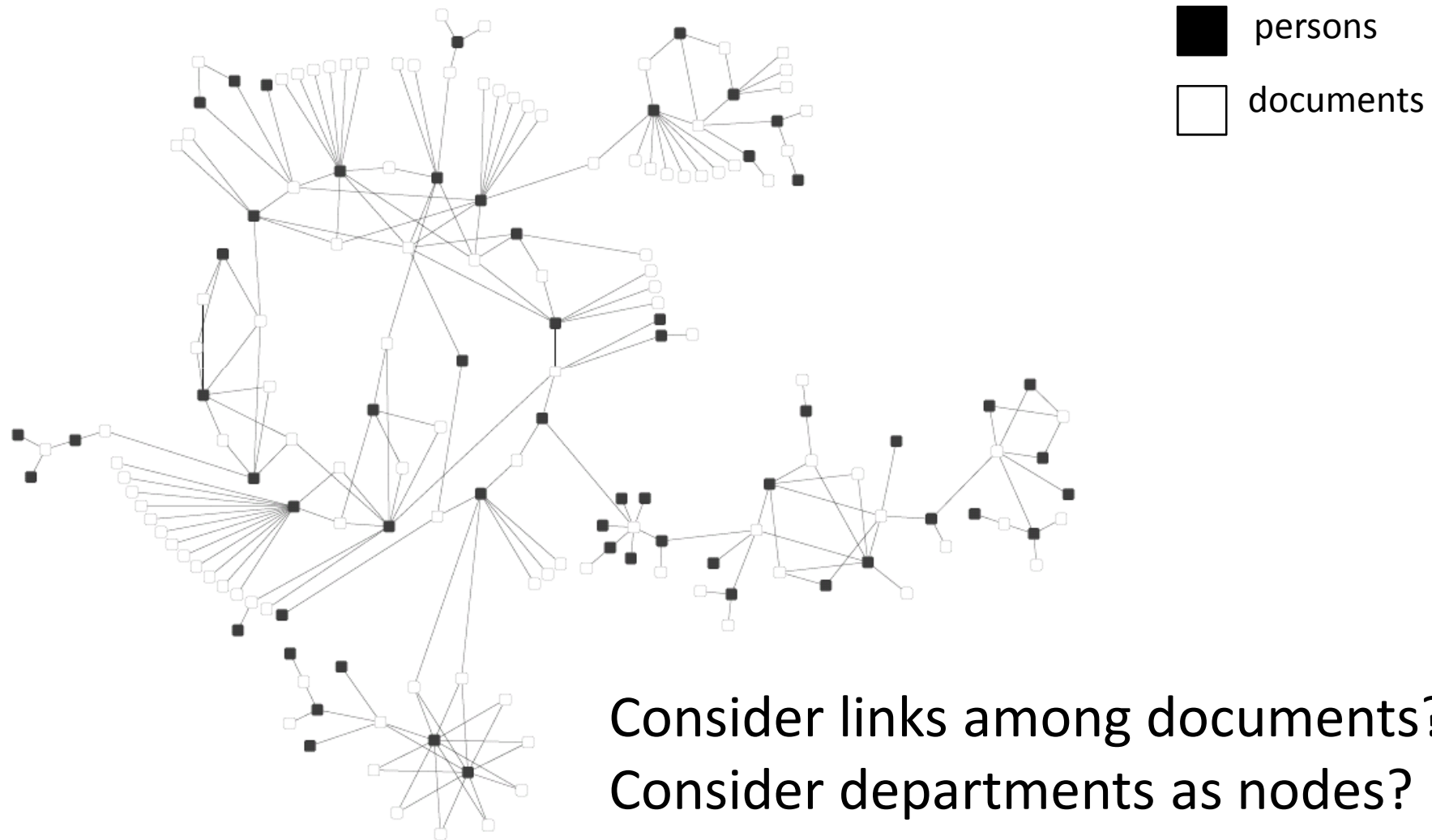


- Expertise evidence is never propagated further than to **mentioned persons**

Exploratory search for experts



Expertise graph



Consider links among documents?
Consider departments as nodes?
Consider social relationships?

Multi-step relevance propagation

- How to model this walk for expertise?
 - Although, considering that experts should be close to relevant documents
- How to propagate expertise evidence (relevance) further after the first step?
- **Answer:** Multi-step relevance propagation with random walk models
 - Finite-random walk (FRW)
 - Infinite random walk (IRW)
 - Absorbing random walk (ARW)

In **P. Serdyukov, H. Rode, and D. Hiemstra**. Modeling Multi-step Relevance Propagation for Expert Finding. In **CIKM 2008**.

Finite random walk

- Model the user as a lazy seeker:
 - So, who is the most probable expert to end up with after some ***K number of steps***?
- How to model laziness in a smart way?

$$P_0(D) = P(Q | D), P_0(e) = 0$$

$$P_i(D) = \underbrace{P(Q | D)}_{\text{Prob. to stay at } D} P_{i-1}(D) + \sum_{e \rightarrow D} P(D | e) P_{i-1}(e),$$

$$P_i(e) = \sum_{D \rightarrow e} \underbrace{(1 - P(Q | D)) P(e | D)}_{\text{Prob. to move on from } D} P_{i-1}(D)$$

Prob. to stay at D

Prob. to move on from D

$$Expertise(e) = P_K(e)$$

Infinite random walk

- Model the user as a tireless seeker:
 - So, who is the most probable expert to end up with after *infinite number of steps*?
- How to model tirelessness smartly?

$$P_i(e) = \sum_{D \rightarrow e} P(e | D) P_{i-1}(D)$$

$$P_i(D) = \underbrace{\lambda P(Q | D)}_{\text{Prob. of walk restart from } D} + (1 - \lambda) \sum_{e \rightarrow D} P(D | e) P_{i-1}(e),$$

Prob. of walk
restart from D

$$Expertise(e) = P_{\infty}(e)$$

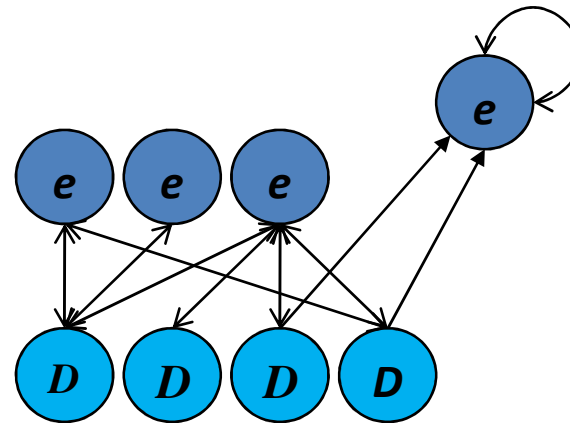
Absorbing random walk

- Absorbing walk:

$$P_0(D) = P(Q | D), P_0(e) = 0$$

$$P_i(D) = \sum_{e \rightarrow D} P(D | e) P_{i-1}(e),$$

$$P_i(e) = \sum_{D \rightarrow e} P(e | D) P_{i-1}(D) + P_{i-1}(e) P^{self}(e | e)$$



- What is the **generalization** of the classic one-step propagation:

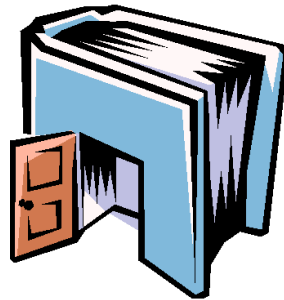
$$Expertise(e) = \sum_{D \in TopK} P^{mult}(e | D) P(Q | D) P(D)$$

Prob. to reach e from D in
minimum number of steps

In **P. Serdyukov, H. Rode, and D. Hiemstra**. Modeling Expert Finding as An Absorbing Random Walk. In **SIGIR 2008**.

Looking for better expertise evidence

- So far considered:
 - Documents are **black boxes (black bags of words)**
 - **There is no world outside the enterprise**
- Can we do better? Look at two extremes...
- Go deeper into the document on a word-level



- Escape the enterprise.... in search for better evidence



Proximity-aware expert finding (I)

- Remember document-centric model?

$$P(e, Q) = \sum_D P(e | D) P(Q | D) P(D)$$

- Why consider independence?

$$P(Q | D) \Rightarrow P(Q | e, D) = \prod_{q \in Q} P(q | e, D)$$

For every occurrence of a query
term and an expert mention

Proximity function

$$P(q | e, D) = \frac{\sum_{q \in D} \text{count}(q, D) \sum_{e \in D} \text{count}(e, D) k(q, e)}{Z}$$

Normalization constant

Proximity-aware expert finding (II)

- Linear function:

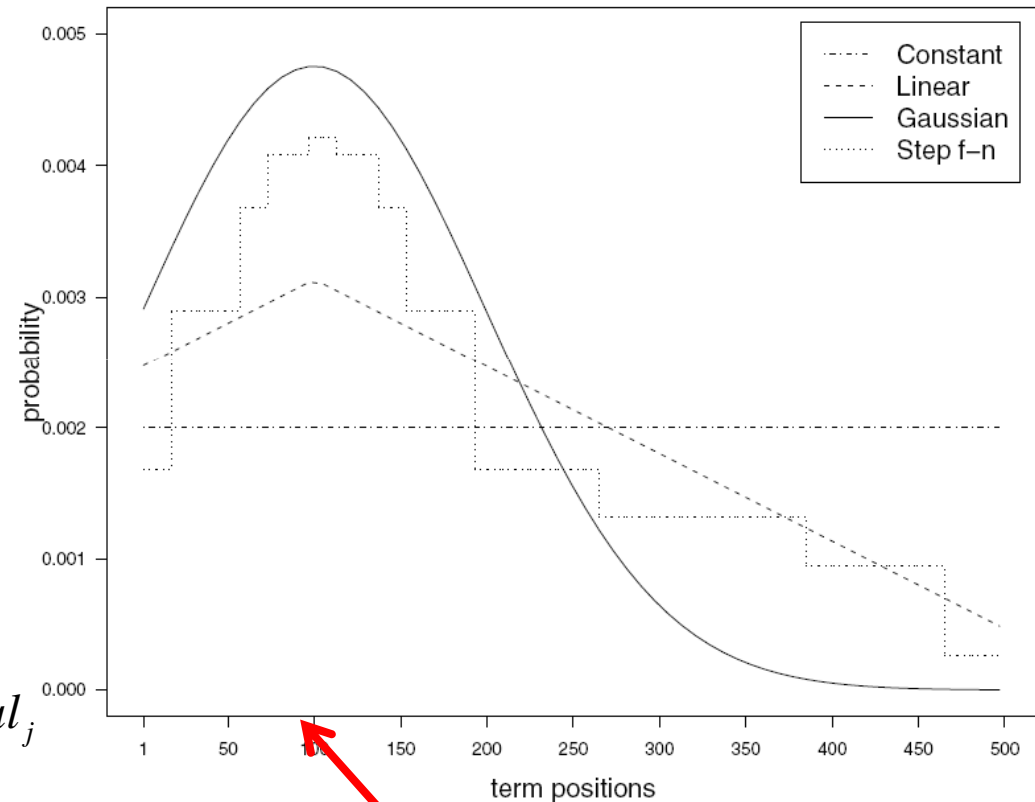
$$k(q, e) = 1 - |pos(q) - pos(e)|$$

- Gaussian function:

$$k(q, e) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(pos(q) - pos(e))^2}{2\sigma^2}\right]$$

- Step function:

$$k(q, e) = \alpha_j, \text{ if } |pos(q) - pos(e)| \in Interval_j$$



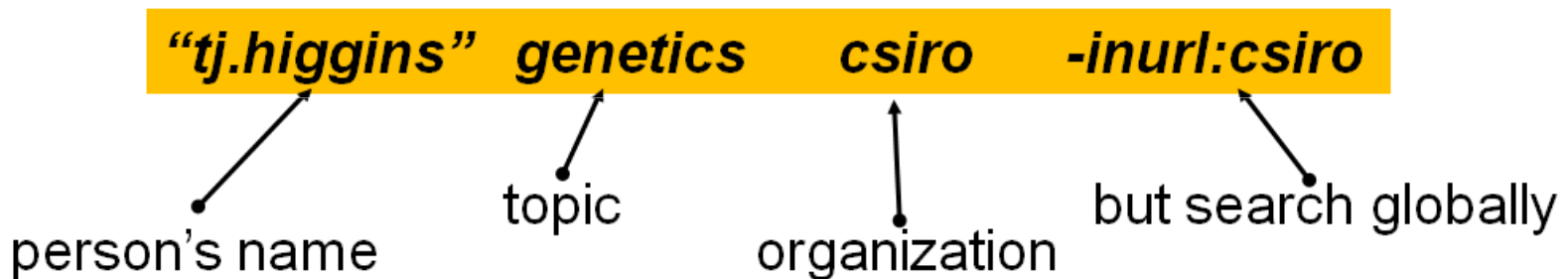
Going beyond the enterprise

- Why to search only in the enterprise?



Acquiring data via Search APIs

- Retrieve all pages with person name?
 - But APIs return at most 1000 results
- Build a query consisting of:



- The number of returned results is a **rough estimate of expertise**

In **P. Serdyukov and D. Hiemstra**. Being Omnipresent to Be Almighty: The Importance of the Global Web Evidence for Organizational Expert Finding. In **FCHER 2008 (SIGIR 2008 Workshop)**.

Where to start?

- Issue 3500 queries to APIs for each topic?
 - Takes about 30 minutes
- Some pre-selection stage for candidates?
 - Experts should be within some **Top-K**
- We are making Enterprise run anyway
 - And it is very fast
- We have full access to the Enterprise data
 - It should be the primary evidence

Web Search evidence

- We need precise estimates for the number of results:
 - Estimates of “total results” are very imprecise
 - Their precision depends on starting position

1st Yahoo! page: 1 - 10 of 273 for genetic modification"tj. higgins" csiro- inurl: csiro

Last Yahoo! page: 71 - 73 of 111 for genetic modification"tj. higgins" csiro- inurl: csiro

- Worst estimate
- Better estimate
- The best estimate

- Google API returns only 32 search items
 - And its estimates are less reliable

Results 21 - 30 of about 1,020 for genetic modification "tj.higgins" csiro -inurl:csiro

News evidence

- Good experts are often news-makers
 - Make discoveries
 - Receive awards
- Every engine has a News Search API !
 - But all of them allow to search **only in the news from the past month**
 - Google News Archives allows to search even in 19th century news, **but has no API**
- But, let's simulate it
 - By adding ***inurl:news*** clause

Blog evidence

- Blogs are knowledge marketplaces
- Even most corporate blogs are public
- Quoting is a social recommendation

Kevin Rose [writes](#) that Digg is launching a recommendation engine that "uses your past digging activity to identify what we call Diggers

Amit Singhal, the head of the Core Ranking team at Google has a post on Google's [philosophy of ranking](#).

John Langford just posted a list of [seven ICML '08 papers that he found interesting](#). I appreciate his taste in papers, and I particularly

- Two blog search engines have the best coverage:
 - Technorati API: almost not supported
 - Google Blog Search API: returns only 8 results

Academic search evidence

- Strong academic record is a must
 - Especially for R&D companies
- Big academic search engines have no API
 - Live Search Academic
 - Google Scholar (recommends experts itself!)

Results 1 - 2 of about 528,000 for [web retrieval](#).

Key authors: [G Salton](#) - [D Hawking](#) - [N Craswell](#) - [P Bailey](#) - [W Grosky](#)

- But Google Book Search API is available!
 - It's like a crippled Google Scholar with only books indexed

Combining evidences

- Why we need so many sources?
- Good expert is not only a local winner
 - Should be “omnipresent”
- Normalization of absolute values is hard
 - Vary a lot over queries and search engines
- Rank aggregation is a convenient solution

$$Expertise(e) = \sum_{Rankings} -Rank(e)$$

Considering URL quality

- What about result set quality?
 - Considering only its **size** is too naive
- We should measure the quality of each result item(*URL*, *Title*, *Summary*):

$$Expertise(e) = \sum_{Item \in WebResultSet} Quality(Item)$$

- Two types of quality measures:
 - **Query-independent**
 - **Query-dependent**

Future challenges for expert finding

- Modeling dependencies within a document
 - More complex topic models?
- Relevance propagation
 - Introduce new entities? Relevance sources?
Search for organizational units?
- Utilize more web sources



References: expert finding

- Expert finding in industry:
 - **Expert finding systems. Survey.** M. Maybury. 2006
free at: http://www.mitre.org/work/tech_papers/tech_papers_06/
- Expert finding in academia:
 - **People Search in the Enterprise.**
K. Balog. PhD Thesis. 2008
 - **The Voting Model for People Search.**
C. Macdonald. PhD Thesis . 2009
 - **Search for expertise: going beyond direct evidence.**
P. Serdyukov. PhD Thesis. 2009