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

Exploratory search

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

Learn
- Knowledge acquisition
- Comprehension
- Comparison
- Discovery
- Serendipity

Investigate
- Incremental search
- Driven by uncertainty
- Non-linear behavior
- Result analysis
- Lasts for hours
What web search engines offer

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?
What is faceted search?

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?

Information that Matters™: click below to refine your search

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

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

Procedures
- psychotherapy
- Cognitive Behavioral Therapy
- Personality Assessment
- Electroconvulsive Therapy
- Body Mass Index

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

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

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

Nutrition
- Polyunsaturated Fatty Acids
- Essential Fatty Acids
- Fish Oil
- Chocolate
- Soybean

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

The Web
1. Depression: MedlinePlus
   Also called: Clinical depression, Dysthymic disorder, Major depressive disorder, Unipolar depression

2. NIMH - Depression
   Depression is a serious medical illness; it's not something that you have made up in your head.
What is faceted search?

FacetLens (Microsoft Research)
What is faceted search?
What is not faceted search?
Too many facets?
Too many facet values?

Information overload

Mobile interfaces
Facet selection: interface-based approach
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

\[ \left| \text{docs} \in \text{Facet}_1 \cup \text{docs} \in \text{Facet}_2 \cup \ldots \cup \text{docs} \in \text{Facet}_K \right| \]
Redundancy-based selection

• Avoid presenting both of correlating facets:
  – Language
  – Nationality

• Consolidate similar facets:
  – Author
  – Editor
  – Contributor

*Beyond Basic Faceted Search. Ben-Yitzhak et. al. WSDM 2008
Interestingness-based facet selection

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

  \[
  \text{Entropy} = \sum_{i=1, value_i \in \text{Facet}}^{n} P(value_i | R) \log P(value_i | R)
  \]

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

  \[
  \text{Divergence}(\text{Facet}, \text{Query}) = \sum_{value_i \in \text{Facet}}^{n} (P(value_i | C) - P(value_i | R)) \log \frac{P(value_i | R)}{P(value_i | C)}
  \]
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:

\[
Relevance(v_i) = \sum_{Doc \in Result, \ f \in Doc \ f = v_i} Score(Doc)
\]

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:

\[
P(f \mid User_k) = \frac{\sum I(f = clicked, Query)}{|Queries \in User_k|}
\]
Collaboratively recommended selection

- Utilize collaborative filtering techniques*:

\[ \alpha P(f \mid User_k) + (1 - \alpha) \sum_{User_j \in Users} \frac{P(f \mid User_j)}{|Users|} \]

- Consider only users with similar tastes:

\[ \alpha P(f \mid User_j) + (1 - \alpha) \sum_{User_j \in Users} P(User_j \mid User_k)P(f \mid User_j) \]

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
Tags in the Enterprise

Operation systems, Genre, Location

What Is Web 2.0 - O'Reilly Media
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 o’reilly, o’reilly media, next generation web
21 minutes ago by John Taylor

The Connectbeam Social Computing Blog: Social Networking
... 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)

Twitter for Enterprise - Eclipsys - Confluence
Tags: social software yesterday by John Taylor into Enterprise 2.0

Sales Performance International :: Solution Selling
Solution Selling provides comprehensive sales skills training for sales professionals, managers...
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 space

http://taggalaxy.de/
How to measure tag size?

\[ \text{fontsize}_i = \frac{\text{fontsize}_{\text{max}} \times (\text{tfidf}_i - \text{tfidf}_{\text{min}})}{(\text{tfidf}_{\text{max}} - \text{tfidf}_{\text{min}})} \]

- \text{tf} – tag frequency in the result set
- \text{idf} – inverted tag frequency in the collection
- \text{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
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 120 000 nouns
- Match tags to nouns
- Disambiguate!
  - Find senses with minimum distance to each other on graph
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 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 classes
   – Use categories as features would introduce too much noise
   – Categories of wiki-articles are better choice
• 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

Continue narrowing

Start
How to incorporate feedback (I)

\[ \text{Score}(Q, D) = -D(\theta_Q \parallel \theta_D) + \beta \cdot D(\theta_N \parallel \theta_D) \]

Relevance lang. model

- \textit{food} +++ \textit{russia} \textit{recipes}

Irrelevance lang. model

- drinking
- health
- work
- humor

\[
\begin{align*}
P('food'|Q) &= \frac{1}{5} \\
P('recipes'|Q) &= \frac{1}{5} \\
P('russia'|Q) &= \frac{3}{5}
\end{align*}
\]

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:

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

• How to define distance between nodes:
  – Length of shortest path
  – Number of shortest paths (of certain length)
  – Distance-based similarity:
    \[
    \sum_{\text{path}(tag, document) \in \text{shortestpaths}} c^{-\text{length}(path)}
    \]
    \(c - \text{parameter}\)

• 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)
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

  \[
P(Class \mid D) = P(Class) \prod_{w=1}^{\mid D \mid} P(w \mid Class)
  \]

  \[P_{\text{smoothed}}(w\mid"\text{Databases}\") = \]

  \[= \lambda_1 P(w\mid"\text{Databases}\") + \lambda_2 P(w\mid"\text{ComputerScience}\") + \lambda_3 P(w\mid"\text{Science}\") , \sum \lambda_i = 1\]

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

  \[
P(Class \mid D) \Rightarrow P(Class \mid D, Q) = P(Class \mid D)P(Class \mid Q)
  \]
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](http://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](http://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 :) we'll know!!!

didn't play very well at gig tonight. that makes me a mad panda. why panda? i like pandas, that's why!"

search: pandas

try some twitter trends: follow friday victor martinez funny people harry potter tgif sms halladay noerin

15

3

= 83%
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:
  - 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 \(T\):
  \[
P(T \mid L) = \prod_{i=1}^{\mid T \mid} P(t_i \mid L)
  \]
  \[
P(t \mid L) = \frac{|L|}{|L| + \lambda} P(t \mid L)_{ML} + \frac{\lambda}{|L| + \lambda} P(t \mid G)_{ML}
  \]
- 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?
  \[
  \text{Score}(l, D) = \alpha \frac{P(l \mid D)}{P(l \mid C)} + (1 - \alpha) \sum_w \frac{P(w \mid D)}{P(w \mid C)}
  \]

• Is document on the topic of \( l \)?
  \[
  \text{Dist}(l, D) = -\sum_w P(w \mid l) \frac{P(l \mid l)}{P(l \mid D)}
  \]

• Select top tags using the rule:
  – At each step choose tag that maximizes:
  \[
  \max_{l' \in \text{selected}} \text{Dist}(l, l')
  \]

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”

• 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
• Conferences: SIGIR, ECIR, WWW, WSDM, KDD, HCIR