Agenda

• Motivation: Search
• Web Mining
• Examples from the Web 2.0 & Usage
  – Flickr example
  – Wikipedia example
  – The Power of Queries
• Concluding Remarks
Motivation

• Web search is no longer about document retrieval
  – Means for web-mediated goals
• New breed of search experiences
  – Demands search ecosystem combining content with intent
  – Exploiting the Wisdom of Crowds behind the Web 2.0

Search is Evolving

• Already, more than a list of docs
• Moving towards identifying a user’s task
• Enabling means for task completion

• New experiences based on the Web 2.0

• Challenges: on-line, scalability
More complete information in one search

Search: Content vs. Intent

Premise:
– People don’t want to search
– People want to get tasks done and get straight to their answers

Start

I am craving for a good coffee in Copenhagen

Finish

Search Menu Reviews Map
How this might work – I

**Index time processing:**

- **Business name**: Lookup
- **Business type**: Food, Fast food, pizza
- **Home page for Pizza Bella Aarhus**
  - Address: Tel: 12345678
- **Geo**
- **Map**
- **Reviews** – extracted and indexed under Pizza in Aarhus
- Other pages around the web

How this might work – II

**Query time processing:**

- **Query stream**: Aarhus pizza bella
- **Session Analysis**
  - Intent: buy pizza
  - Geo: Aarhus
Ne t

• We move from a web of pages to a web of objects
• Objects are people, places, businesses, restaurants …
• Objects have attributes
  – Missing, noisy, etc.
• Intents are satisfied by presenting objects and attributes
• Attributes define faceted search

How do we get structured objects/attributes?

• Web Content
  – Metadata/Taxonomies/Folksonomies
  – ML/ Classification/Extraction/Semantic Web

• Web Usage
  – Implicit relations

• Building out an open ecosystem
  – Publishers have incentives to contribute
  – E.g. SearchMonkey
Content and Metadata trends

<table>
<thead>
<tr>
<th>Content type</th>
<th>Amount of content produced per day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Published content</td>
<td>3-4 GB</td>
</tr>
<tr>
<td>Professional web content</td>
<td>(\sim 2 ) GB</td>
</tr>
<tr>
<td>User generated content</td>
<td>8-10 GB</td>
</tr>
<tr>
<td>Private text content</td>
<td>(\sim 3 ) TB (300x more)</td>
</tr>
<tr>
<td>Upper bound on typed content</td>
<td>(\sim 700 ) TB (\sim 200x more)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metadata type</th>
<th>Amount of metadata produced per day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anchortext</td>
<td>100 MB</td>
</tr>
<tr>
<td>Tags</td>
<td>40 MB</td>
</tr>
<tr>
<td>Pageviews</td>
<td>180 GB</td>
</tr>
<tr>
<td>Reviews</td>
<td>Around 10 MB</td>
</tr>
</tbody>
</table>

[Ramakrishnan and Tomkins 2007]

Examples

Wordnet

Explicit

Metadata

RDF

Wikipedia ODP

Y! Answers

Flickr

UGC

Implicit

Text

Anchors + links

Queries+clicks

Private

Scale

Quality?
The Wisdom of Crowds

• James Surowiecki, a *New Yorker* columnist, published this book in 2004
  – “Under the right circumstances, groups are remarkably intelligent”
• Importance of diversity, independence and decentralization

   “large groups of people are smarter than an elite few, no matter how brilliant—they are better at solving problems, fostering innovation, coming to wise decisions, even predicting the future”.

---

Explore / Tags / danish / clusters
The Wisdom of Crowds

- Crucial for Search Ranking
- Text: Web Writers & Editors
  - not only for the Web!
- Links: Web Publishers
- Tags: Web Taggers
- Queries: All Web Users!
  - Queries and actions (or no action!)
Web Mining

– Fast Prototyping
– Quality vs. Performance
  • Bring more data!
– Graph Mining
– Parallel computing is not easy!
– Different sources of information

Fast Prototyping: WIM

WIM – Web Information Mining
(Pereira, Baeza-Yates, Ziviani; WSDM 2009)

• WIM goal: facilitate fast Web mining prototyping
• Main research challenges:
  – Data model
  – Algebra
  – Software prototype
  • Architecture and implementation issues
Data Model – Design Goals

• Feasibility
• Simplicity
• Extensibility
• Data representativity
• Uniformity among operators
• Applicability to many scenarios

Relation Type

The type of a relation is either node or link

• Node relations represent nodes of a graph
  – Such as documents of a Web dataset
• Link relations represent edges of a graph
  – Such as links between Web documents

Usage data can be represented as both node or link relations
Operations

• The act of applying an operator to a view or relation
• An operator is a function defined in the WIM algebra
  – Unary or binary
• Operators' output is one of these:
  – A totally new relation \( R' \)
  – A view \( V_i R \) of an input \( R \)
  – A view compatible to an input

Two Classes of Operators

• Seven data manipulation operators
  – Select, Calculate, CalcGraph, Aggregate, Set, Join, Materialize
• Eight data mining operators
  – Search, Compare, CompGraph, Cluster, Disconnect, Associate, Analyze, Relink

• Operators:
  – Have options and sub-options
  – Are often applied to one or a few attributes
WIM Program

- Sequence of operations applied to relations
  - Result of users' interaction through the WIM language
  - The WIM language:
    - Is built upon the WIM algebra
    - Is declarative
    - Is a dataflow programming language
    - Facilitates parallelism

```c
// Clustering duplicates for both old and new collections:
relDupOld = Compare(relOld, sparse, total, at.text);
relClusterOld = Disconnect(relDupOld, connected, newat.clus);
relDupNew = Compare(relNew, sparse, total, at.text);
relClusterNew = Disconnect(relDupNew, connected, newat.clus);

// Comparing the collections:
relSearch = Search(relClusterOld, relClusterNew, shingles, 20%,
                   at.text, at.text);

// Eliminating children with the same URL of parents:
relSearchUrl = CompGraph(relSearch, total, at.url, at.url, newat.sim);
relSeDifUrl = Select(relSearchUrl, value, ==, 0, at.sim);

// Translating start and end nodes into instance nodes:
relStart = Set(relClusterOld, relSeDifUrl, intersection, at.id, at.start);
relStartInst = Aggregate(relStart, grouping, count, at.clus);
relEnd = Set(relClusterNew, relSeDifUrl, intersection, at.id, at.end);
relEndInst = Aggregate(relEnd, grouping, count, at.clus);

// Merging instance nodes with the similarity graph:
relGenEnd = Set(relSeDifUrl, relEndInst, intersection, at.end, at.id);
relGenSt = Set(relGenEnd, relStartInst, intersection, at.start, at.id);

// Selecting only one parent per child:
relGenFinal = Aggregate(relGenSt, grouping, count, at.end);
```
Multi-Graph Mining

• Performing a joint analysis of multi-graphs to capture different semantic aspects of the same knowledge domain.
  
  – General framework
    • set of operations and graph algorithms
  – Efficient and scalable implementation
  – Applications

Bordino, Donato & Baeza-Yates, Scalable analysis of query logs through multiple graph projections, submitted

Algebra

• Data Model:
  - \( G = \{ V, E, w_V, w_E \} \)
    
    - \( w_V : V \rightarrow N \)
    - \( w_E : E \rightarrow R \)

• Operations:
  • Binary operations
  • Unary operations
**Binary Operations**

- **Union.** Given two query log graphs $G$ and $H$, their union is represented by a graph $F = G \cup H$ such that $V(F) = V(G) \cup V(H)$ and $E(F) = E(G) \cup E(H)$.

- **Intersection.** The intersection of two graphs $G$ and $H$ is a graph $F = G \cap H$ such that $V(F) = V(G) \cap V(H)$ and $E(F) = E(G) \cap E(H)$.

- **Difference.** The set difference of two graphs $G$ and $H$ is a graph $F = G \setminus H$ such that $V(F) = V(G) \setminus V(H)$ and $E(F) = E(G) \setminus E(H)$.

- **Symmetric difference.** The symmetric difference of two graphs $G$ and $H$ is a graph $F = G \Delta H$ such that $V(F) = V(G) \Delta V(H)$ and $E(F) = E(G) \Delta E(H)$.

**Unary Operations**

- Connected Components
- Biconnected Components
- Articulation Points
- Node Filtering
- Edge Filtering
Tag Mining - Collective Knowledge

- Many users annotate photos of “La Sagrada Familia”:
  - Sagrada Familia, Barcelona
  - Sagrada Familia, Gaudi, architecture, church
  - church, Sagrada Familia
  - Sagrada Familia, Barcelona, Spain

- Derived collective knowledge:
  - Barcelona, Gaudi, church, architecture

Relating Images

Query → unsorted photos

WORDNET
Wikipedia
flickr

unsorted photos

architecture
sun set
lights

tag graph

tag:type

tag:type

tag:type
TagExplorer

- http://sandbox.yahoo.com/TagExplorer
- A prototype for browsing Flickr photos
- Provides query refinement for …
  - … drilling in to more specific topics
  - … zooming out to more general topics
  - … side-track to a related topic
- Organizes refinement terms …
  - … in a tag-cloud
  - … groups together semantically similar terms

Tag Mining - Classification

- Assign tag semantics using WordNet broad categories

- Paris :: location
- Eiffel Tower :: artefact
- Coverage: 52% of tag volume
Tag Mining – Classification

- Extend this mapping using patterns found in Wikipedia
  - Upper bound for coverage: 78.6% of the tag volume
  - Based on SVM approach
    - Features: Wikipedia templates and categories
    - Training data: Wikipedia entries found in WordNet
  - Extended coverage: 68% of the tag volume
  - Mapping from Wikipedia pages to tags
    - Reduces ambiguity in the classification

Van Zwol at al, 2008
Could suggest tags: nice but ....

Use Visual Annotations

Flickr allows another kind of annotations (notes)

- Associate text with visual area
- Highly relevant to content
  → Visual Annotation
- Valuable to learn different visual representations of an object
- Tagging untagged images

Olivares, Ciaramita, van Zwol. ACM Multimedia 2008
Content-based Image Retrieval

1. Extract visual features and describe them
2. Build visual vocabulary

- SIFT descriptors

{k-means clustering} → Visual vocabulary 10k words

High-level search outline

(1) "coke can"

(2) Search

(3) Visual Annotations

(4) Aggregate
Evaluation

Hypotheses:

- **H1**: Rank aggregation using visual annotations will significantly improve the retrieval performance in terms of precision

- **H2**: Tag-based search combined with CBIR using visual annotations will improve retrieval in terms of precision

Results: Systems comparison
Bridging implicit and explicit metadata

Pablo Ruiz Picasso (October 25, 1881 – April 8, 1973), often referred to simply as Picasso, was a Spanish painter and sculptor. His full name is Pablo Diego José Francisco de Paula Juan Nepomuceno María de los Remedios Cipriano de la Santísima Trinidad Ruiz y Picasso. One of the most recognized figures in 20th century art, he is best known as the co-founder, along with Georges Braque, of cubism.

**Biography**

Pablo Picasso was born in Málaga, Spain, the first child of José Ruiz y Blasco and María Picasso y López. He was christened with the names Pablo, Diego, José, Francisco de Paula, Juan Nepomuceno, María de los Remedios, and Cipriano de la Santísima Trinidad. Picasso's father was a painter whose specialty was the naturalistic depiction of birds and who for most of his life was also a professor of art at the School of Crafts and a curator of a local museum. The young Picasso showed a passion and a skill for drawing from an early age, according to his mother; his first word was 'piz,' a shortening of 'Pizarrón,' the Spanish word for pencil. It is from his father that Picasso had his first formal academic training, such as figure drawing and painting in oil. Although Picasso attended art schools throughout his childhood, often those where his father taught, he never finished his college-level course of study at the Academy of Arts.

---

Extending metadata

**Pablo Picasso** was born in Málaga, Spain.

<table>
<thead>
<tr>
<th>PER</th>
<th>LOC</th>
<th>LOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>artist:name</td>
<td>artist:placeofbirth</td>
<td>artist:placeofbirth</td>
</tr>
</tbody>
</table>

|  E:PERSON | GPE:CITY | GPE:COUNTRY |

If most artists are persons, than let’s assume all artists are persons. If most places of birth are locations, then let’s assume all are.

---
Entity Containment Graph

Example: Picasso

[Diagram of a graph displaying connections between entities, with nodes labeled such as 'Pablo Picasso', 'Gertrude Stein', 'Picasso', '1906', 'October 1881', and 'October 25, 1881', connected by lines indicating relationships.]
Correlator

- URL: correlator.sandbox.yahoo.com
- Find relations in the Wikipedia
  - Relate entities: names, places, dates
  - Change the result interface
- If the query is not an entry in the wikipedia
  - Synthetic page is created
- Based on linear time entity detection with competitive quality

Zaragoza, Attardi, Ciaramita, Atserias, Castillo, Mika, Surdeanu, ....

---

**Correlator - Examples**

Events related to "denmark"

**Timeline**

985 - February 2, 1014
- Norway: Sveyn Forkbeard, King of Norway
- Denmark: Sveyn Forkbeard, also known as Sveyn Forkbeard, King of Denmark

999 - February 2, 1014
- Norway: Sveyn Forkbeard, King of Norway
- Denmark: Sveyn Forkbeard, also known as Sveyn Forkbeard, King of Denmark

---

**Events in the timeline**

985 - February 2, 1014
- Norway: Sveyn Forkbeard, King of Norway
- Denmark: Sveyn Forkbeard, also known as Sveyn Forkbeard, King of Denmark

999 - February 2, 1014
- Norway: Sveyn Forkbeard, King of Norway
- Denmark: Sveyn Forkbeard, also known as Sveyn Forkbeard, King of Denmark
For topics without a Wikipedia page, Correlator creates a “synthetic page” with an overview of the topic.

Query:
- art deco chicago

Synthetic page:
- Defines Art Deco
- Defines Chicago
- Shows relations between Art Deco and Chicago

Step 1: Definitions of query concepts

- Parse query using Wikipedia titles and redirects
  - nyc parks => “New York City” parks
  - art deco chicago => “Art Deco” Chicago
- Display first paragraphs of each from each concept’s Wikipedia page and sentences connecting the concepts
Step 2: Relations between query concepts (1/2)

- Retrieve related sentences
  - **Output:** Ranked list of sentences
- Aggregate sentences over Wikipedia pages
  - **Page score is the sum of the score of its sentences**
  - **Output:** Ranked list of pages
- Aggregate pages over Wikipedia categories
  - **Each relevant page votes for its categories**
  - **Category score is the sum of its votes**
  - **Output:** Ranked list of categories containing relevant pages

Web Usage

- **Clicks** – follow hyperlinks
- **Queries** – user interest
- **Sequence of actions** – time

- **Strong Assumption:**
  
  When you use the Web you are thinking
  
  - **Users** – **Actions** – **Objects**
Relating All (Baeza-Yates, 2007)

Qualitative Analysis

<table>
<thead>
<tr>
<th>Graph</th>
<th>Strength</th>
<th>Sparsity</th>
<th>Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>Medium</td>
<td>High</td>
<td>Polysemy</td>
</tr>
<tr>
<td>Session</td>
<td>Medium</td>
<td>High</td>
<td>Physical sessions</td>
</tr>
<tr>
<td>Click</td>
<td>High</td>
<td>Medium</td>
<td>Click spam</td>
</tr>
<tr>
<td>Link</td>
<td>Weak</td>
<td>Medium</td>
<td>Link spam</td>
</tr>
<tr>
<td>Term</td>
<td>Medium</td>
<td>Low</td>
<td>Term spam</td>
</tr>
</tbody>
</table>
Session (Query-Flow) Graph

[Diagram of session graph with nodes like 'ebay', 'autotrader', 'used fox vw', 'barcelona hotel', 'barcelona rent', 'soccer', 'barcelona soccer', 'barcelona', 'barcelona fc', and arrows connecting them.]


Click Graph

[Diagam of click graph with nodes like 'Map of France', 'France flag', 'France map', 'Montpellier France map', and connections between them.]
Click Distribution

Data per user is a power law

Connected Components
Implicit Folksonomy?

Set Relations

- Identical sets: **equivalence**
- Subsets: **specificity**
  - directed edges
- Non empty intersections (with threshold)
  - degree of relation
- Dual graph: URLs related by queries
  - High degree: multi-topical URLs
  - Queries relate content

Baeza-Yates & Tiberi
ACM KDD 2007
Implicit Knowledge? Web slang!

Evaluation: ODP Similarity

- A simple measure of similarity among queries using ODP categories
  - Define the similarity between two categories as the length of the longest shared path over the length of the longest path
  - Let $c_1,.., c_k$ and $c'_1,.., c'_k$ be the top $k$ categories for two queries. Define the similarity ($@k$) between the two queries as $\max\{\text{sim}(c_i,c'_j) \mid i,j=1,..,K\}$
Experimental Evaluation

• We evaluated a 1000 thousand edges sample for each kind of relation
• We also evaluated a sample of random pairs of not adjacent queries (baseline)
• We studied the similarity as a function of $k$ (the number of categories used)
Open Issues

• Data Volume versus Better Algorithms

• Explicit versus implicit social networks
  − Any fundamental similarities?

• How to evaluate with (small) partial knowledge?
  − Data volume amplifies the problem

• User aggregation versus personalization
  − Optimize common tasks
  − Move away from privacy issues
The Virtuous Cycle

Explicit

Metadata
RDF
Wikipedia
ODP
Y! Answers
Flickr

Text
Anchors + links
Queries + clicks

Wordnet

Second edition coming soon

Questions?

Contact: rbaeza@acm.org

Thanks to Carlos Castillo, Debora Donato, Aris Gionis, Alexandre Francisco, Peter Mika, Prabhakar Raghavan, Borkur Sigurbjornsson, Roelof van Zwol, Hugo Zaragoza