



# The Power of Data

**EMMDS**  
Copenhagen, 2009

**Ricardo Baeza-Yates**  
*VP, Yahoo! Research*  
*Barcelona & Santiago*

## 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

-3-

# 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

-4-

# More complete information in one search

The screenshot shows a Yahoo search results page for the query "legal sea foods boston ma". The search results are categorized into "Shortcuts", "Deep Links", and "Enhanced Results".

- Shortcuts:** A map of Boston with a red box highlighting the "Legal Seafoods near Boston" section, which lists three nearby locations with their addresses and phone numbers.
- Deep Links:** A red box highlights several search results from various domains like "legalseafoods.com", "citysearch.com", and "yelp.com", providing detailed information about the restaurant chain.
- Enhanced Results:** A red box highlights two Yelp listings for "Legal Seafoods - Waterfront - Boston, MA 02109" and "Legal Seafoods - East Boston - Boston, MA", including user reviews, photos, and contact information.

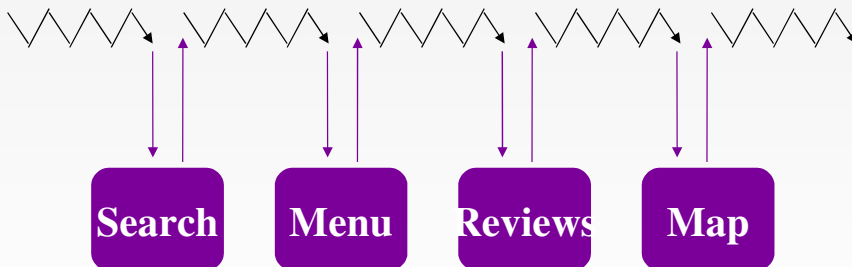
- 5 -

## 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**

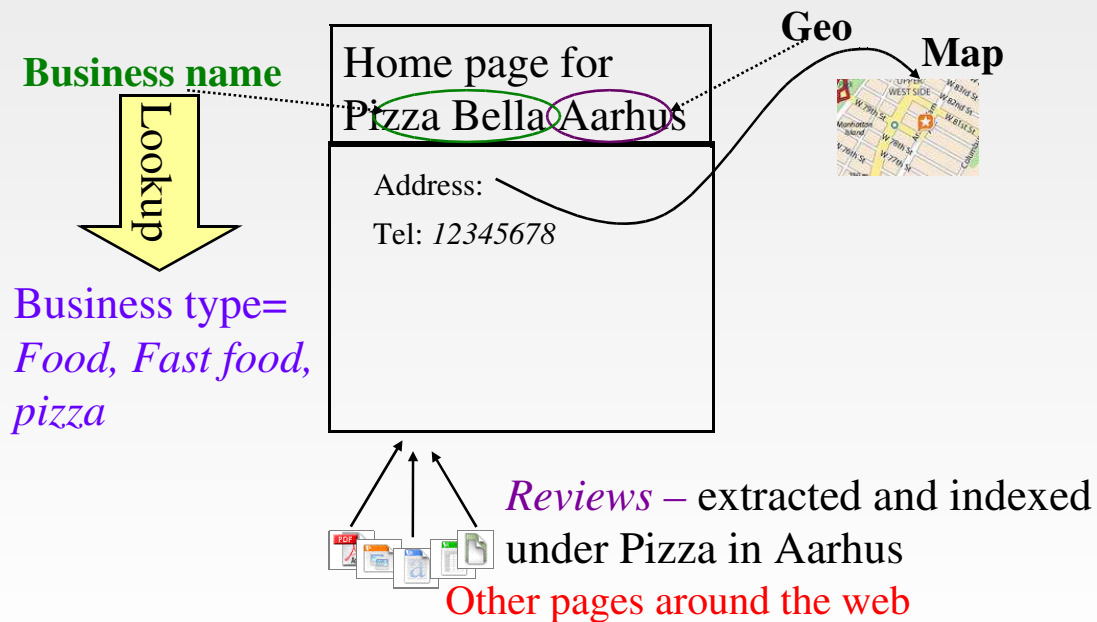


7

- 7 -

# How this might work – I

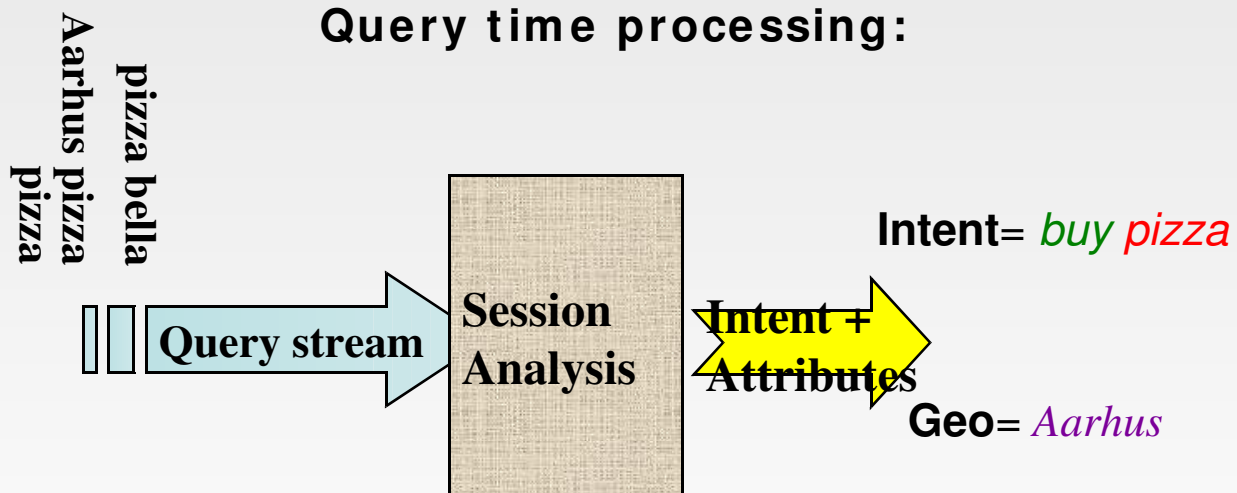
## Index time processing:



- 8 -

# How this might work – II

## Query time processing:



- 9 -

# Net

- 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

10

- 10 -

## 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
  - <sup>11</sup> – E.g. SearchMonkey

- 11 -

# Content and Metadata trends

Content type	Amount of content produced per day
Published content	3-4 GB
Professional web content	~ 2 GB
User generated content	8-10 GB
Private text content	~ 3 TB (300x more)
Upper bound on typed content	~700 TB (~200x more)

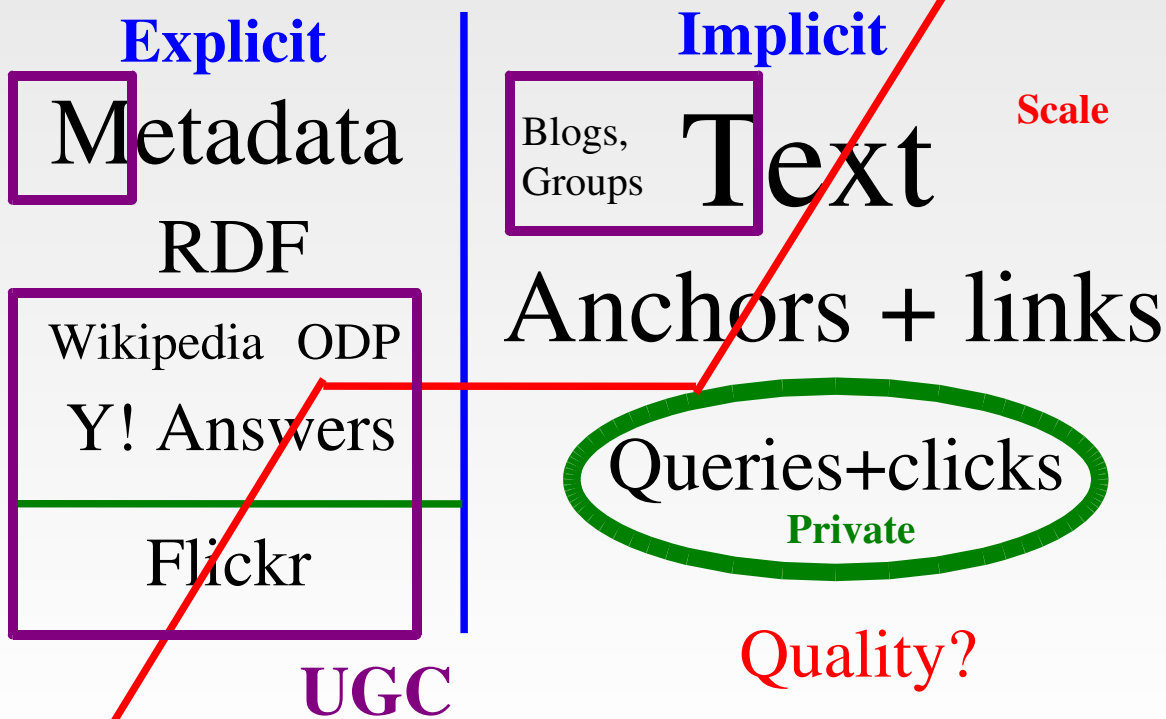
Metadata type	Amount of metadata produced per day
Anchortext	100 MB
Tags	40 MB
Pageviews	180 GB
Reviews	Around 10 MB

[Ramakrishnan and Tomkins 2007]

- 12 -

## Examples

Wordnet



- 13 -

# 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

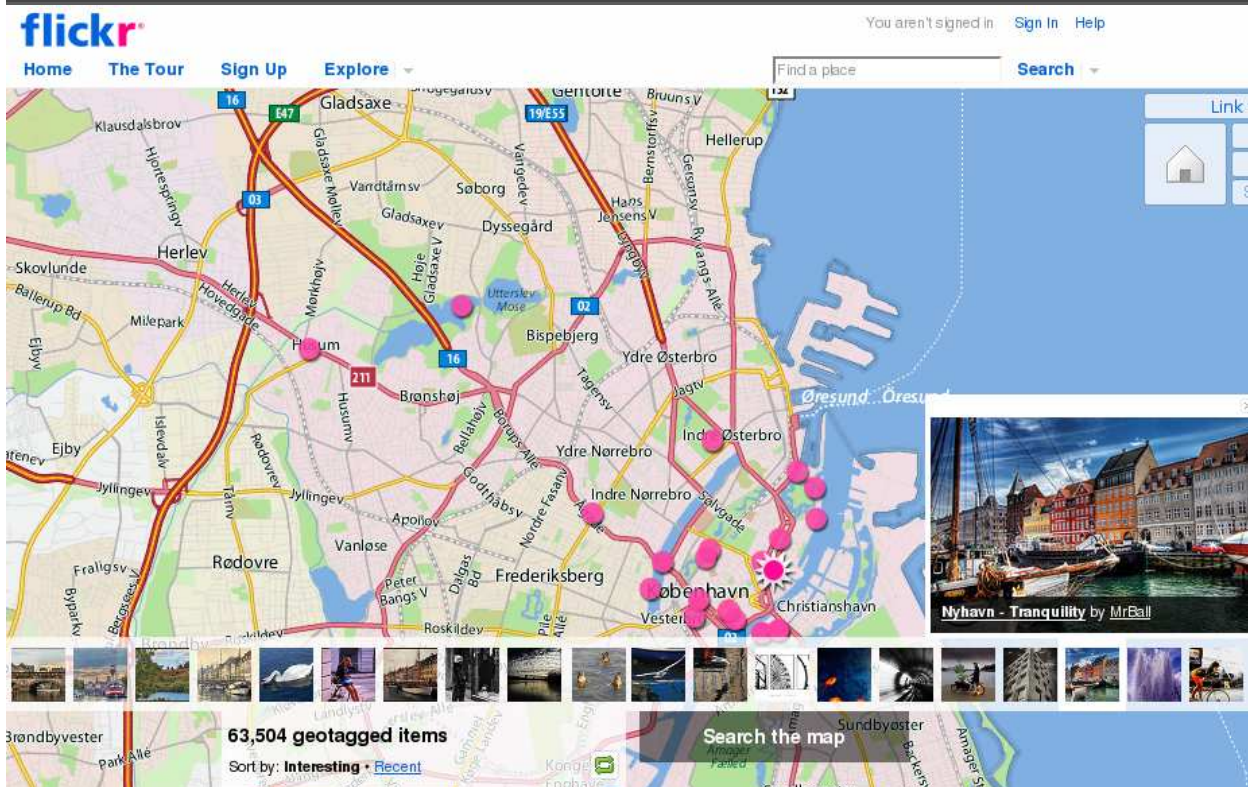
## Aggregating data

*“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”.*

- 14 -

The screenshot shows the Flickr website interface. At the top, there's a navigation bar with 'Home', 'The Tour', 'Sign Up', and 'Explore'. A search bar is visible with the text 'Search'. Below the navigation, the page title is 'Explore / Tags / danish / clusters'. A 'Jump to:' field contains the text 'danish'. The main content area displays four clusters of images, each with a grid of five thumbnails and a list of associated tags. The first cluster has tags: [denmark](#), [copenhagen](#), [danmark](#), [dansk](#), [design](#), [europe](#), [architecture](#), [girl](#), [art](#), [travel](#). The second cluster has tags: [modern](#), [vintage](#), [midcentury](#), [furniture](#), [teak](#), [retro](#), [chair](#), [lamp](#), [century](#), [mid](#). The third cluster has tags: [food](#), [pastry](#), [breakfast](#), [coffee](#), [bread](#), [dessert](#), [yummy](#), [baking](#). The fourth cluster has tags: [california](#), [solang](#). Each cluster includes a 'See more in this cluster...' link with a right-pointing arrow.

# The Wisdom of Crowds



# 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!)

- 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

- 20 -

## 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

- 23 -

# 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  $Vi_2(R)$  of an input  $R$
  - A view compatible to an input

- 33 -

## 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

- 35 -

- 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

- 36 -

```
// 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);
```

- 37 -

# 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**

- 38 -

# 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

- 39 -

# 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  $E(F) = E(G) \Delta E(H)$ .

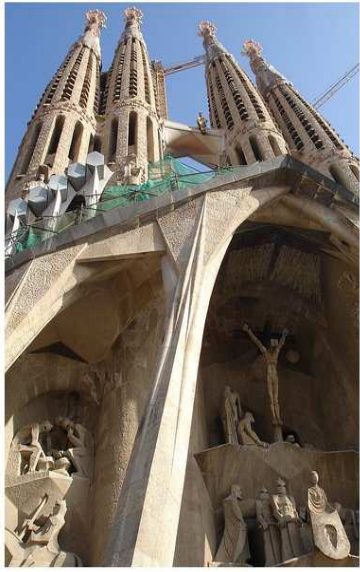
- 40 -

# Unary Operations

- Connected Components
- Biconnected Components
- Articulation Points
- Node Filtering
- Edge Filtering

- 41 -

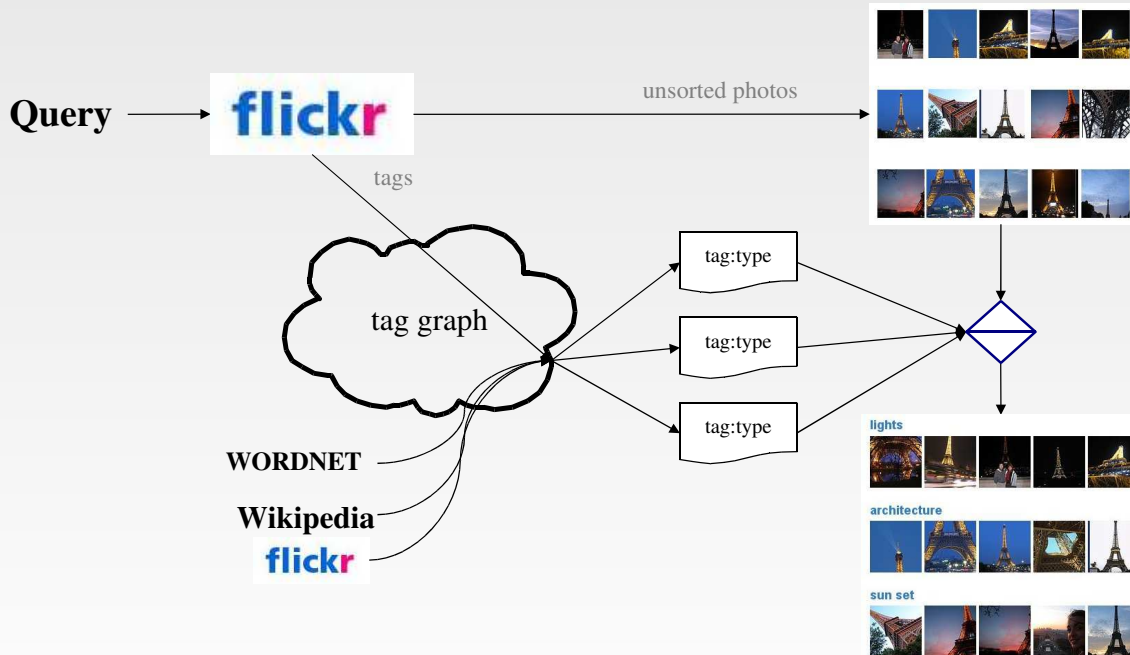
# Tag Mining - Collective Knowledge



- 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

- 42 -

## Relating Images



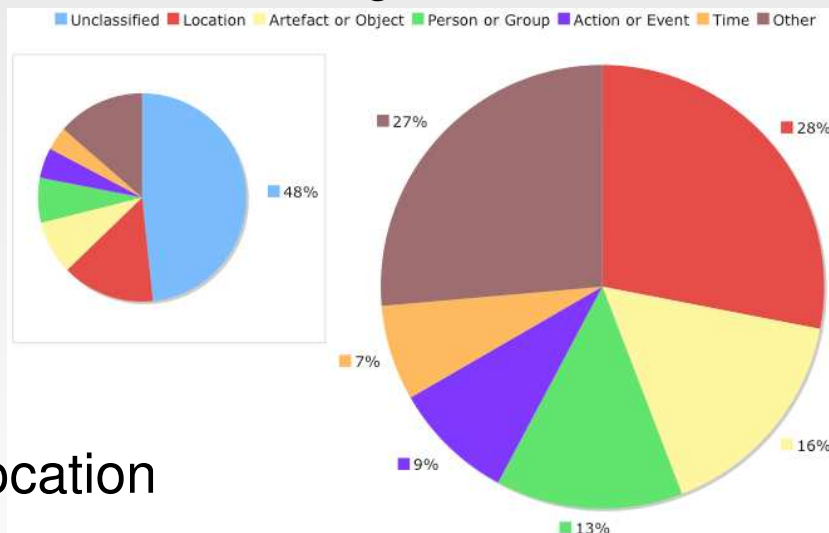
- 43 -

- <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

- 44 -

## Tag Mining - Classification

- Assign tag semantics using WordNet broad categories



- Paris :: location
- Eiffel Tower :: artefact
- Coverage: 52% of tag volume

- 45 -

# 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 et al, 2008

- 46 -

## TagExplorer - Example

TagExplorer  
Powered by Flickr

copenhagen

SEARCH

YAHOO!  
RESEARCH

Query: [copenhagen](#) ✕

locations

[amager](#) + [christiania](#) +  
[danmark](#) + [denmark](#) +  
[europe](#) + [frederiksberg](#) +  
[hvidovre](#) + [kobenhavn](#) +  
[kopenhagen](#) + [nyhavn](#) +  
[scandinavia](#) + [tivoli](#) +  
[ungdomshuset](#) +

subjects

[kastellet](#) +

activities

[travel](#) +

time

[2005](#) + [2006](#) + [2007](#) +

Help

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

- Use [tag](#) to post new query using tag
- Use + to add terms to query
- Use ✕ to remove terms from query

Photo Results



Photo Details



DSC00575, Amalienborg Palace, Copenhagen, Denmark  
Taken by: [jimg944](#)  
[View photo on Flickr](#)  
Tags: copenhagen denmark danish scandinavia amalienborgpalace danishroyal

## Could suggest tags: nice but ....

### London Eye



London Eye and Golden Jubilee Bridge seen from Westminster Bridge.

### Tag list

london eye, thames,

### Suggested tags

- london
- england
- uk
- river
- eye
- south bank
- big ben
- night
- bridge
- 2006

Update annotation

- 48 -

## 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

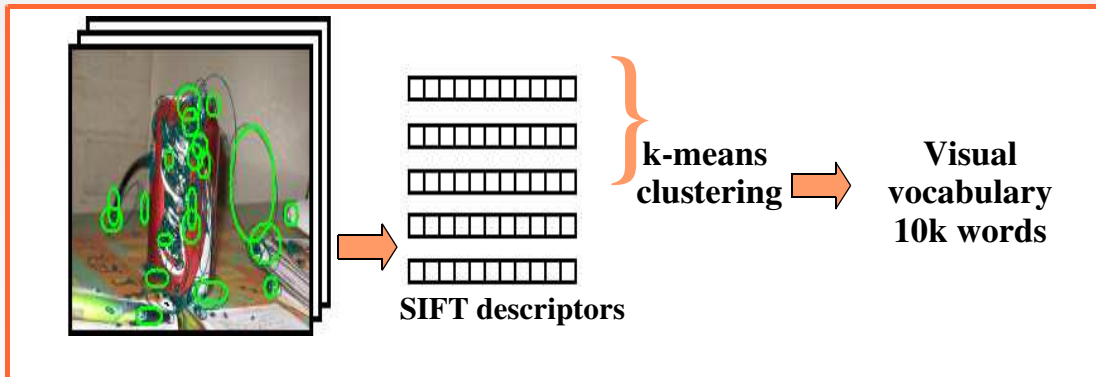


Olivares, Ciaramita, van Zwol. ACM Multimedia 2008

- 49 -

# Content-based Image Retrieval

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



- 50 -

## High-level search outline



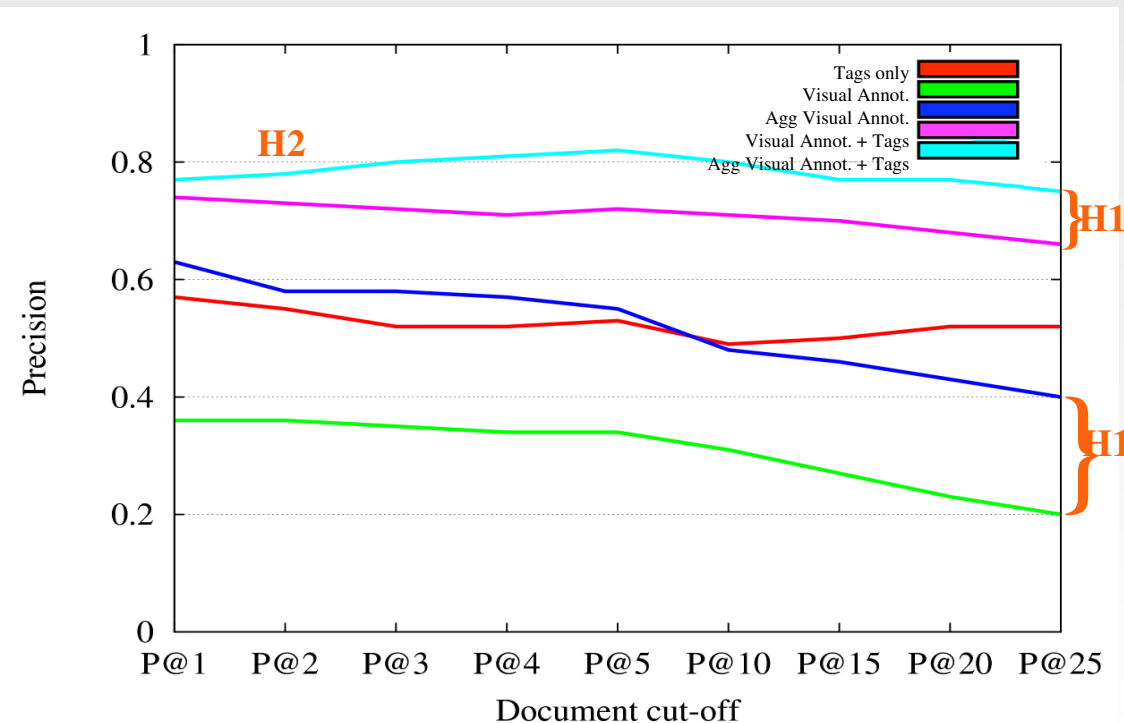
- 51 -

## 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

- 52 -

## Results: Systems comparison



- 54 -

# Bridging implicit and explicit metadata

- About Wikipedia
- Community portal
- Recent changes
- Contact Wikipedia
- Donate to Wikipedia
- Help

search

Go Search

toolbox

- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link
- Cite this article

languages

- Afrikaans
- العربية
- বাংলা
- Bân-lâm-gú
- Bosanski
- Brezhoneg
- Български
- Català
- Česky
- Cymraeg
- Dansk


**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 Clito Ruiz y Picasso**.<sup>[1]</sup> One of the most recognized figures in 20th century art, he is best known as the co-founder, along with Georges Braque, of cubism.

**Contents** [show]

## Biography [edit]


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.<sup>[2]</sup> 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,<sup>[3]</sup> his first word was "piz," a shortening of *lápiz*, the Spanish word for pencil.<sup>[4]</sup> It was from his father that Picasso had his first formal academic art 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


**Pablo Picasso**



Picasso (January 1962)

**Birth name** Pablo Diego José Francisco de Paula Juan Nepomuceno María de los Remedios Cipriano de la Santísima Trinidad Martyr Patricio Clito Ruiz y Picasso

**Born** October 25, 1881  Málaga, Spain

**Died** April 8, 1973 (aged 91)  Mougins, France

- 56 -

## Extending metadata

Pablo Picasso was born in Málaga, Spain.

PER

LOC

LOC

E:PERSON

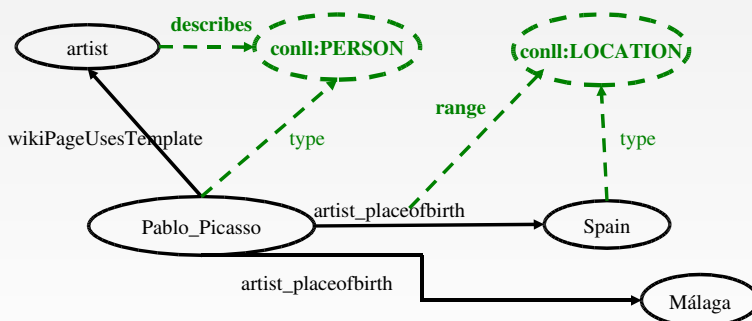
GPE:CITY GPE:COUNTRY

artist:name

artist:placeofbirth artist:placeofbirth

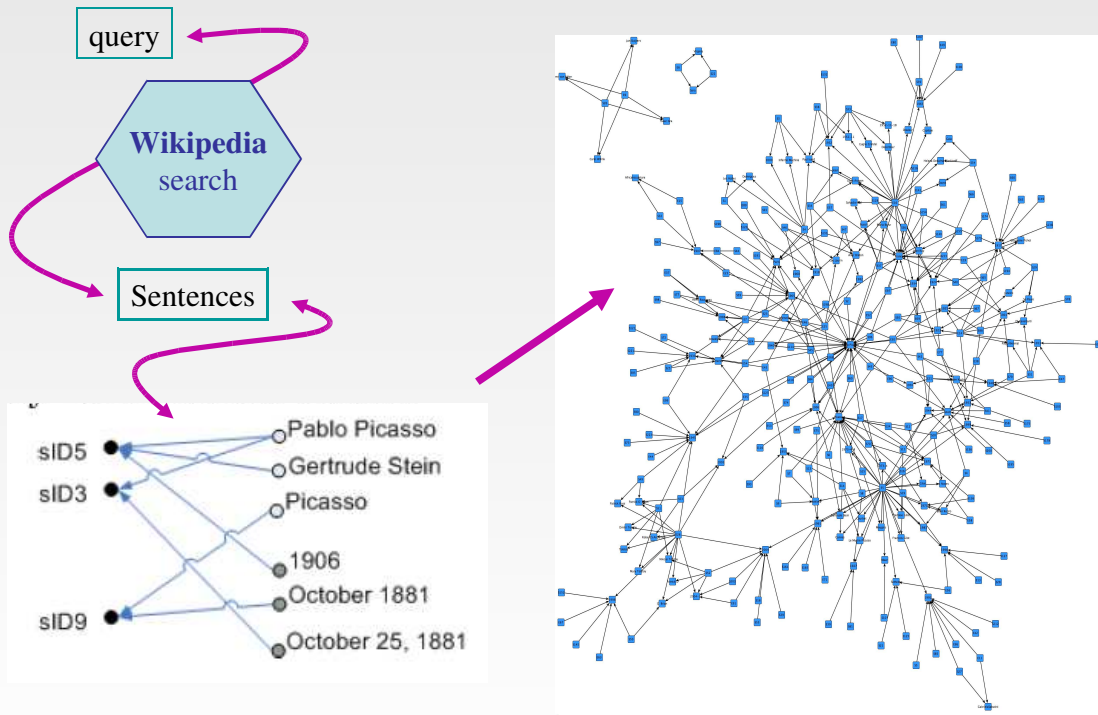


If most artists are persons, then let's assume all artists are persons.  
If most places of birth are locations, then let's assume all are.

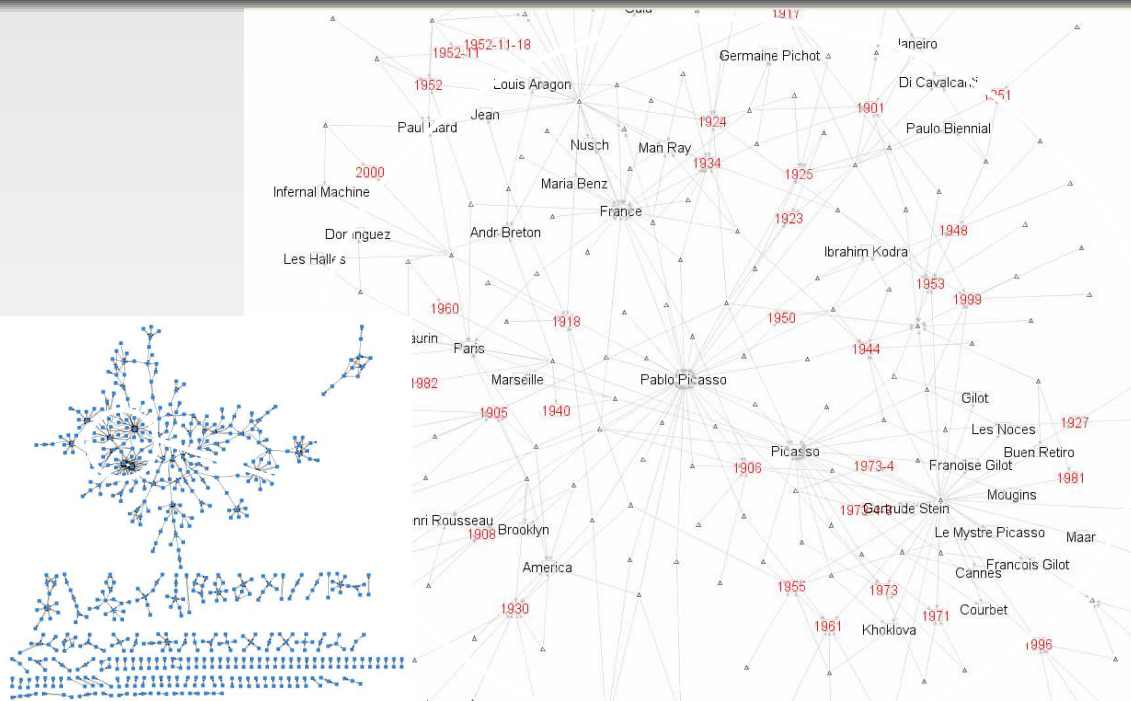


- 58 -

# Entity Containment Graph



## Example: Picasso



# Correlator

- URL: [correlator.sandbox.yahoo.com](http://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

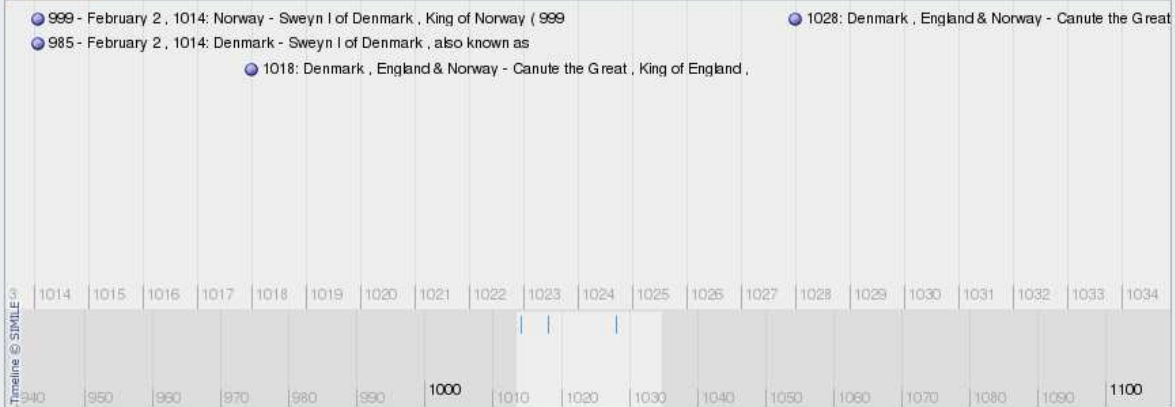
Zaragoza, Attardi, Ciaramita, Atserias, Castillo, Mika, Surdeanu, .....

- 62 -

## Correlator - Examples

### Events related to "denmark"

#### Timeline



#### Events in the timeline

##### 985 - February 2, 1014

(From [W List of state leaders in 1008](#)) \*Denmark - Sweyn I of Denmark , also known as Sweyn Forkbeard , King of Denmark ( 985 - February 2 , 1014 ) ; also King of Norway\*

(From [W List of state leaders in 1004](#)) \*Denmark - Sweyn I of Denmark , also known as Sweyn Forkbeard , King of Denmark ( 985 - February 2 , 1014 ) ; also King of Norway\*

(From [W List of state leaders in 1002](#)) \*Denmark - Sweyn I of Denmark , also known as Sweyn Forkbeard , King of Denmark ( 985 - February 2 , 1014 ) ; also King of Norway\*

##### 999 - February 2, 1014

(From [W List of state leaders in 1008](#)) \*Norway - Sweyn I of Denmark , King of Norway ( 999 - February 2 , 1014 ) ; also King of Denmark\*

(From [W List of state leaders in 1004](#)) \*Norway - Sweyn I of Denmark , King of Norway ( 999 - February 2 , 1014 ) ; also King of Denmark\*

(From [W List of state leaders in 1002](#)) \*Norway - Sweyn I of Denmark , King of Norway ( 999 - February 2 , 1014 ) ; also King of Denmark\*

1018

# Overview page

- For topics without a Wikipedia page, Correlator creates a “synthetic page” with an overview of the topic
- Query:
  - art deco chicao
- Synthetic page:
  - Defines Art Deco
  - Defines Chicago
  - Shows relations between Art Deco and Chicago

Search

### Art Deco

**Art Deco** was a popular international design movement from 1922 until 1939, affecting the decorative arts such as [architecture](#), [interior design](#), and [industrial design](#), as well as the [visual arts](#) such as [fashion](#), [painting](#), the [graphic arts](#), and [film](#). This movement was, in a sense, an amalgam of many different styles and movements of the early 20th century, including [Constructivism](#), [Cubism](#), [Modernism](#), [Bauhaus](#), [Art Nouveau](#), and [Futurism](#). Its popularity peaked in Europe during the [Roaring Twenties](#) and continued strongly in the [United States](#) through the 1930s. Although many [design movements](#) have political or philosophical roots or intentions, [Art Deco](#) was purely decorative. At the time, this style was seen as elegant, functional, and ultra modern.

... California [Art Deco Society](#) [Chicago landmarks](#) , [Art Deco architecture](#) [tour](#) France -- The Bibliographical Society of the University of Virginia presents Illustrations for The [Art Deco Book](#) in France , by Gordon N. Ray , Edited by G. Thomas Tanselle ...

[W Art Deco](#): View full article

### Chicago

**Chicago** ([IPA](#): /jɪˈkɑːɡoo/ ), is a city in the state of [Illinois](#) and the largest in the [Midwest](#). With its population of over 2.8 million people located almost entirely in [Cook County](#) (a portion of the city's [O'Hare International Airport](#) overlaps into [DuPage County](#)), Chicago is the [third largest city](#) in the [United States](#). The population of Chicago's [metropolitan area](#), which covers several counties, contains over 9.7 million people in [Illinois](#), [Wisconsin](#) and [Indiana](#), making it the [third largest metropolitan area](#) in the U.S. [Chicago](#) has been [classified as an alpha world city](#) for its worldwide economic influence.

[Chicago](#) was incorporated as a city in 1837. Its location at the site of a portage between the [Great Lakes](#) and the [Mississippi River watershed](#), aided the city's rapid growth. Today, Chicago is a major transportation hub, as well as the business, financial, and cultural capital of the [Midwest](#).

[W Chicago](#): View full article

**Category: 1930 architecture**

**W Merchandise Mart**: Massive in its construction , and serving as a monument to early 20th century merchandising and architecture , the [art deco](#) landmark anchors the daytime skyline at the junction of the [Chicago River](#) branches . ... Second only to [Holabird & Root](#) in [Chicago art deco architecture](#) , the firm had a long-standing relationship with the Field family . Started in 1928 , completed in 1931 , and built in the same [art deco](#) style as the [Chicago Board of Trade Building](#) , its cost was reported as both \$ 32 million and \$ 38 million .

**W Chicago Board of Trade Building**: The current structure is known for its [art deco](#) architecture , sculptures and large-scale stone carving , as well as large trading floors . A three-story [art deco](#) statue of Ceres , goddess of grain , caps the building . ... The project included restoration of the main lobby to emphasize the design features of the [art deco](#) era , elevator modernization , facade renovation and cleaning , and the continued renovation of upper floor corridors and hallways .

**W Grace Building (Sydney)**: Inspired by the Gothic revival-modernist [Tribune Tower](#) in [Chicago](#) -- the headquarters of the [Chicago Tribune](#) -- the building was of the [art deco](#) architectural style and had stat-of-the-art innovations and facilities for the time .

[1930 architecture](#): View more entries from this category

---

**Category: Skyscrapers in Chicago**

**W Chicago Board of Trade Building**: The current structure is known for its [art deco](#) architecture , sculptures and large-scale stone carving , as well as large trading floors . A three-story [art deco](#) statue of Ceres , goddess of grain , caps the building . ... The project included restoration of the main lobby to emphasize the design features of the [art deco](#) era , elevator modernization , facade renovation and cleaning , and the continued renovation of upper floor corridors and hallways .

**W LaSalle National Bank Building**: LaSalle National Bank Building ( formerly known as the [Field Building](#) ) is an [art deco](#) building in the LaSalle Street corridor in the Loop community area of [Chicago](#) , [Illinois](#) , [USA](#). The construction of LaSalle National Bank Building was completed 1934 as a 335 feet ( 163 m ) 45-story skyscraper on S. Clark Street in [Chicago](#) , [U.S.A.](#). The architect was Graham , Anderson , Probst & White .

**W Four Seasons Hotel Chicago**: [Four Seasons Hotel Chicago](#) will soon undergo its first renovation . The renovation will provide a [French Art Deco](#) design to the structure , patterned after a 1930s style .

[Skyscrapers in Chicago](#): View more entries from this category

- 64 -

## Step 1: Definitions of query concepts

- Parse query using Wikipedia titles and redirects
  - nyc parks => “New York City” parks
  - art deco chicao => “Art Deco” Chicago
- Display first paragraphs of each from each concept’s Wikipedia page and sentences connecting the concepts

### Art Deco

**Art Deco** was a popular international design movement from 1922 until 1939, affecting the decorative arts such as [architecture](#), [interior design](#), and [industrial design](#), as well as the [visual arts](#) such as [fashion](#), [painting](#), the [graphic arts](#), and [film](#). This movement was, in a sense, an amalgam of many different styles and movements of the early 20th century, including [Constructivism](#), [Cubism](#), [Modernism](#), [Bauhaus](#), [Art Nouveau](#), and [Futurism](#). Its popularity peaked in Europe during the [Roaring Twenties](#) and continued strongly in the [United States](#) through the 1930s. Although many [design movements](#) have political or philosophical roots or intentions, [Art Deco](#) was purely decorative. At the time, this style was seen as elegant, functional, and ultra modern.

... California [Art Deco Society](#) [Chicago landmarks](#) , [Art Deco architecture](#) [tour](#) France -- The Bibliographical Society of the University of Virginia presents Illustrations for The [Art Deco Book](#) in France , by Gordon N. Ray , Edited by G. Thomas Tanselle ...

[W Art Deco](#): View full article

### Chicago

**Chicago** ([IPA](#): /jɪˈkɑːɡoo/ ), is a city in the state of [Illinois](#) and the largest in the [Midwest](#). With its population of over 2.8 million people located almost entirely in [Cook County](#) (a portion of the city's [O'Hare International Airport](#) overlaps into [DuPage County](#)), Chicago is the [third largest city](#) in the [United States](#). The population of Chicago's [metropolitan area](#), which covers several counties, contains over 9.7 million people in [Illinois](#), [Wisconsin](#) and [Indiana](#), making it the [third largest metropolitan area](#) in the U.S. [Chicago](#) has been [classified as an alpha world city](#) for its worldwide economic influence.

[Chicago](#) was incorporated as a city in 1837. Its location at the site of a portage between the [Great Lakes](#) and the [Mississippi River watershed](#), aided the city's rapid growth. Today, Chicago is a major transportation hub, as well as the business, financial, and cultural capital of the [Midwest](#).

[W Chicago](#): View full article

65 -

## 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**

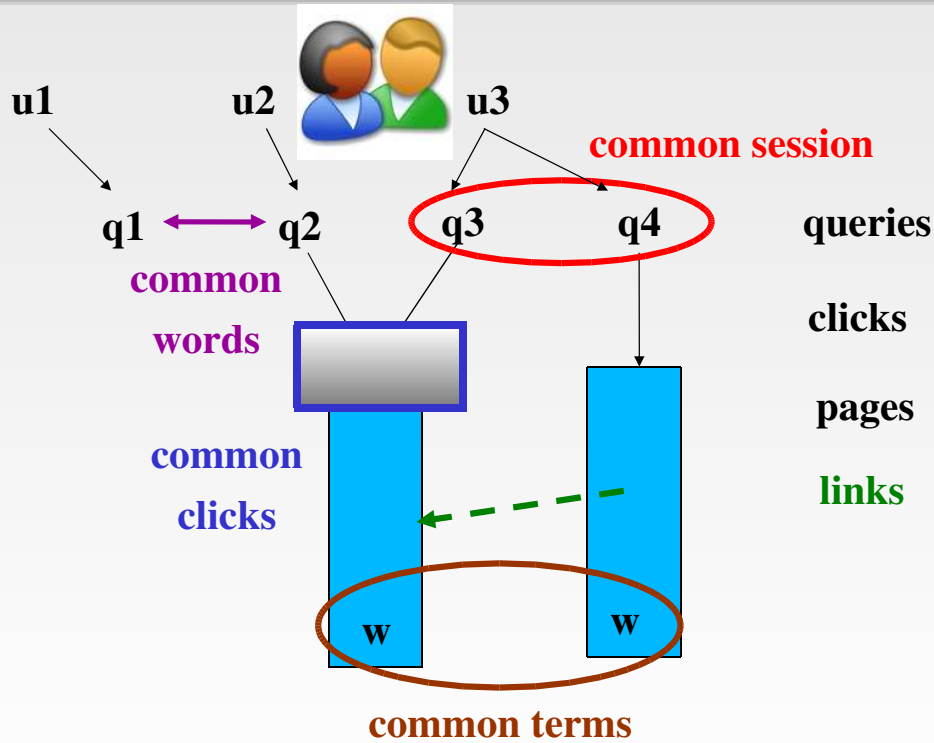
- 66 -

## 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**

- 71 -

# Relating All (Baeza-Yates, 2007)



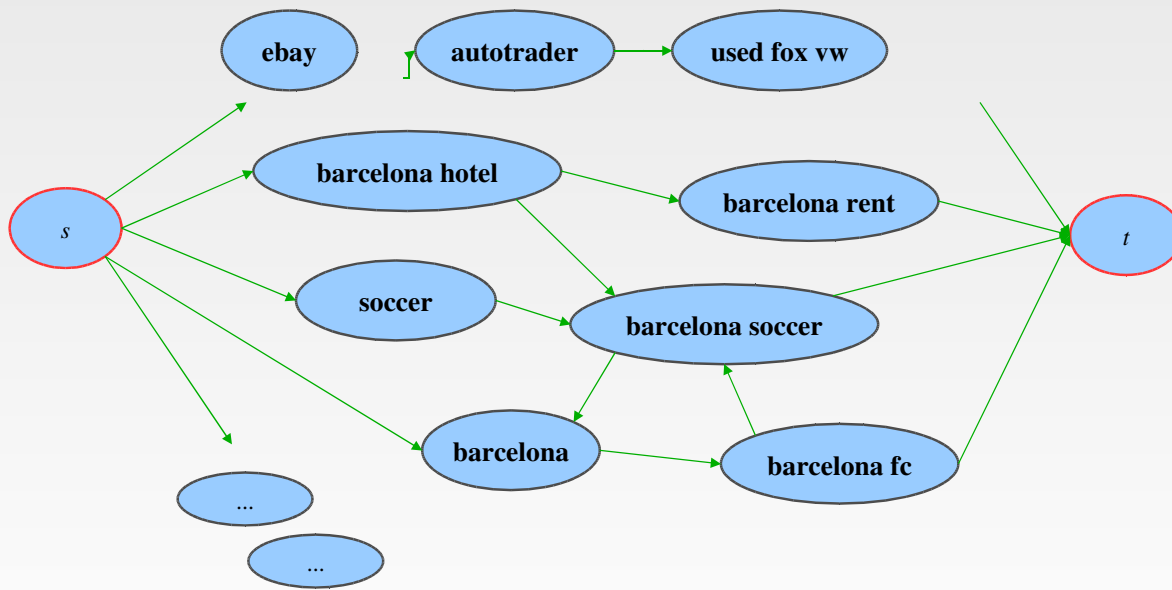
- 72 -

## Qualitative Analysis

Graph	Strength	Sparsity	Noise
Word	Medium	High	Polysemy
Session	Medium	High	Physical sessions
Click	High	Medium	Click spam
Link	Weak	Medium	Link spam
Term	Medium	Low	Term spam

- 73 -

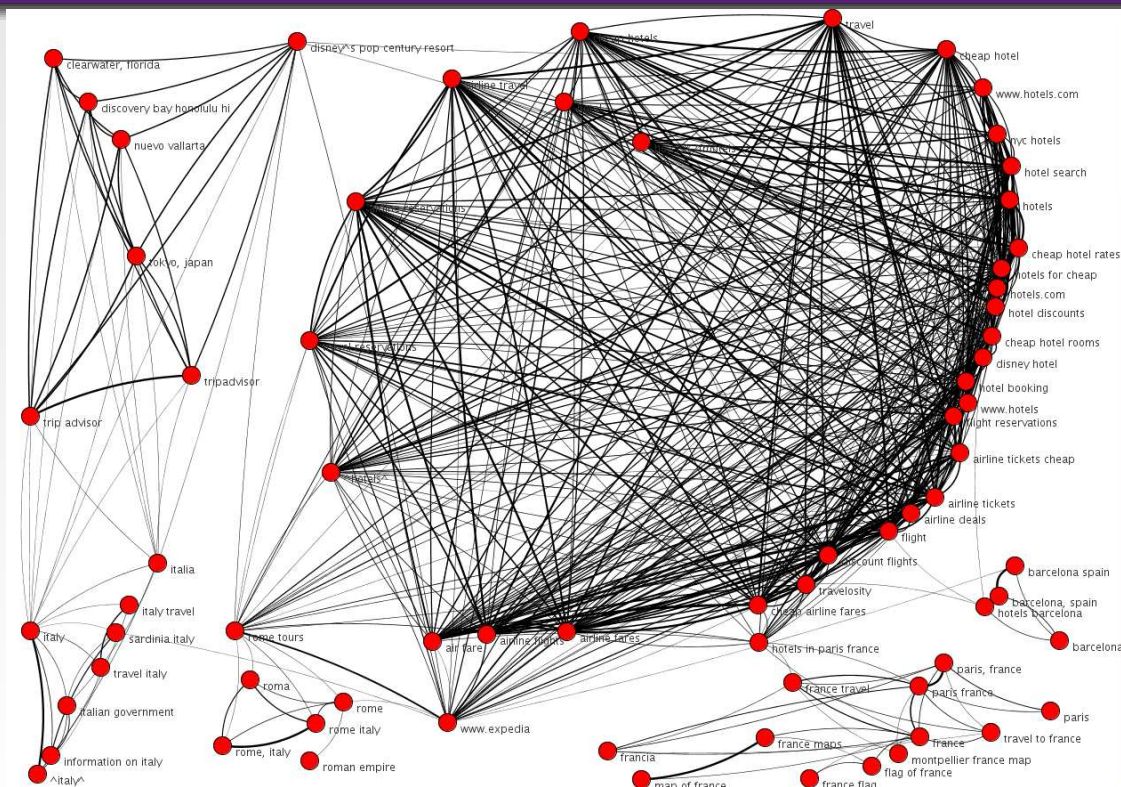
# Session (Query-Flow) Graph



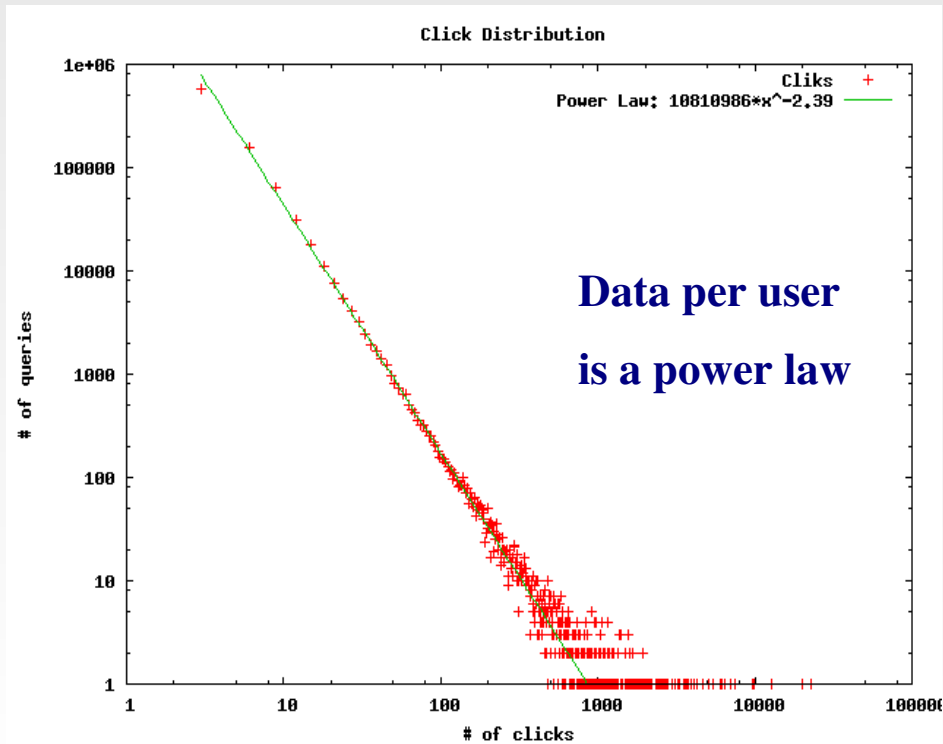
Boldi, Bonchi, Castillo, Donato, Gionis, Vigna. CIKM 2008.



# Click Graph

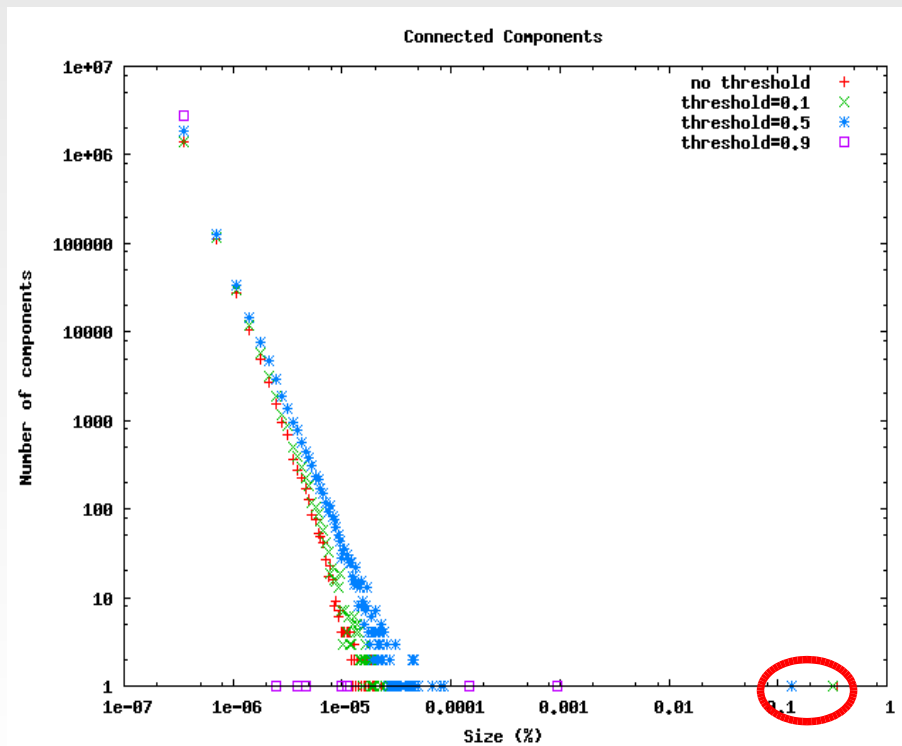


# Click Distribution



- 79 -

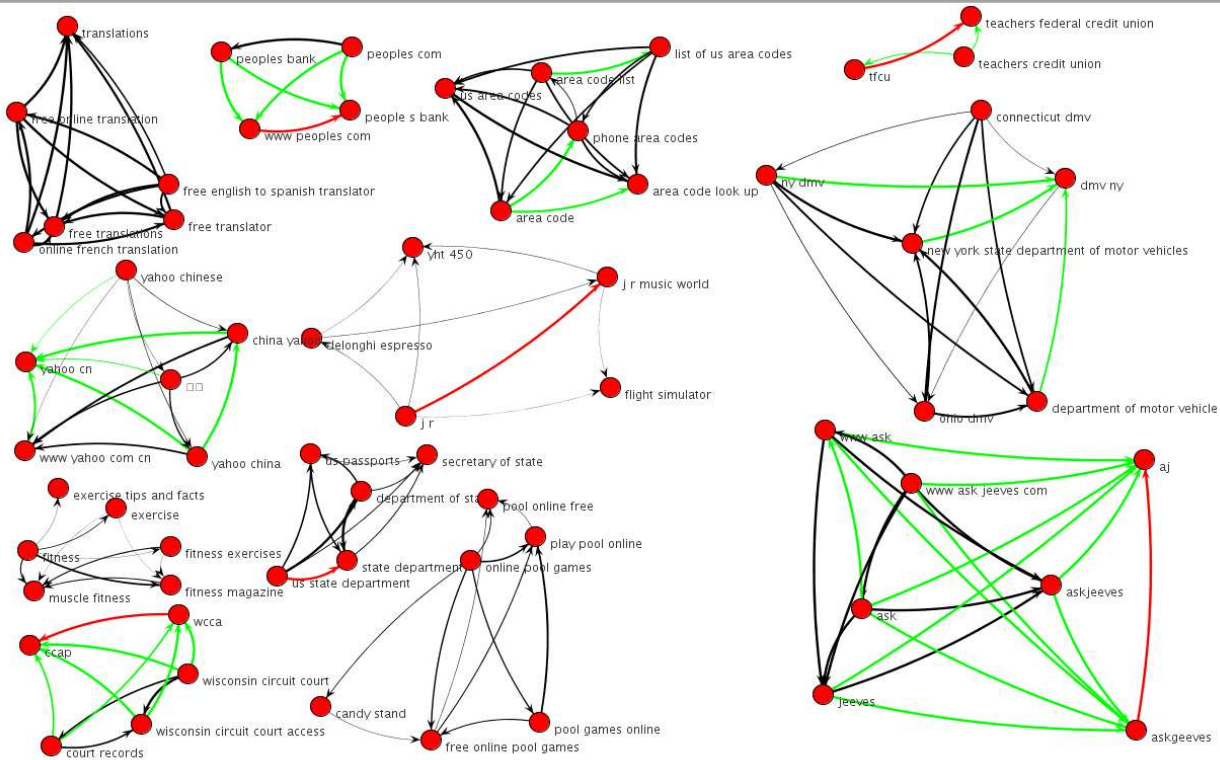
# Connected Components



- 80 -



# Implicit Knowledge? Web slang!



## Evaluation: ODP Similarity

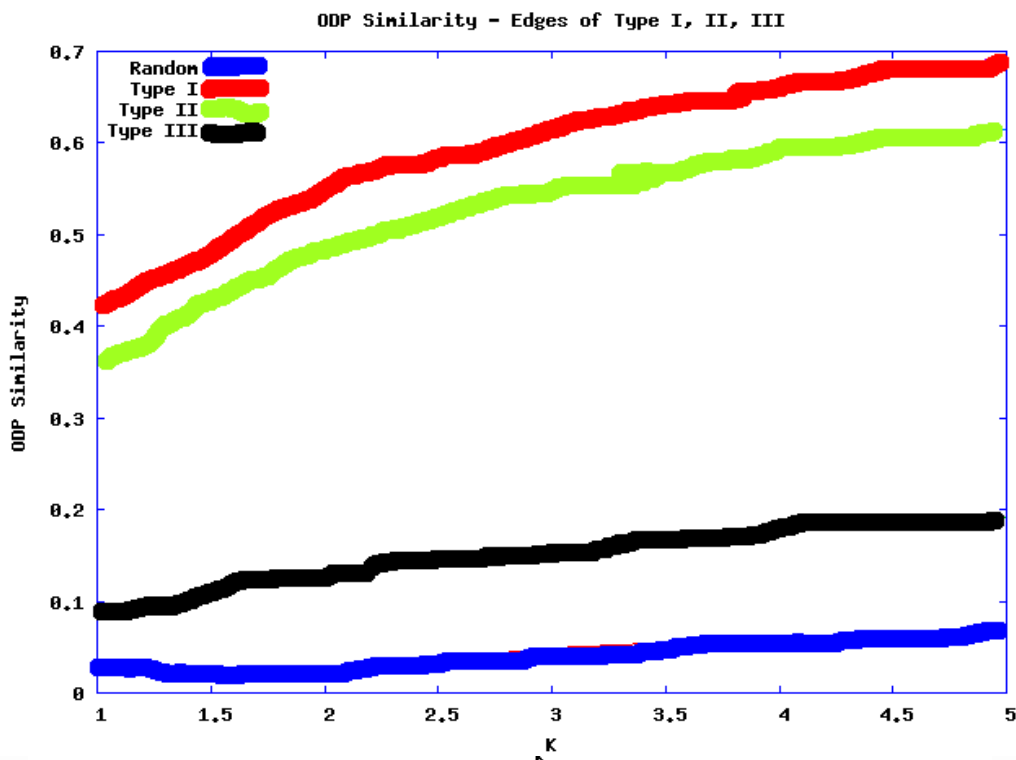
- A simple measure of similarity among queries using ODP categories
  - Define the similarity between two categories as the length of the longest shared path over the length of the longest path
  - Let  $c_1, \dots, c_k$  and  $c'_1, \dots, c'_k$  be the top  $k$  categories for two queries. Define the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

# 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)

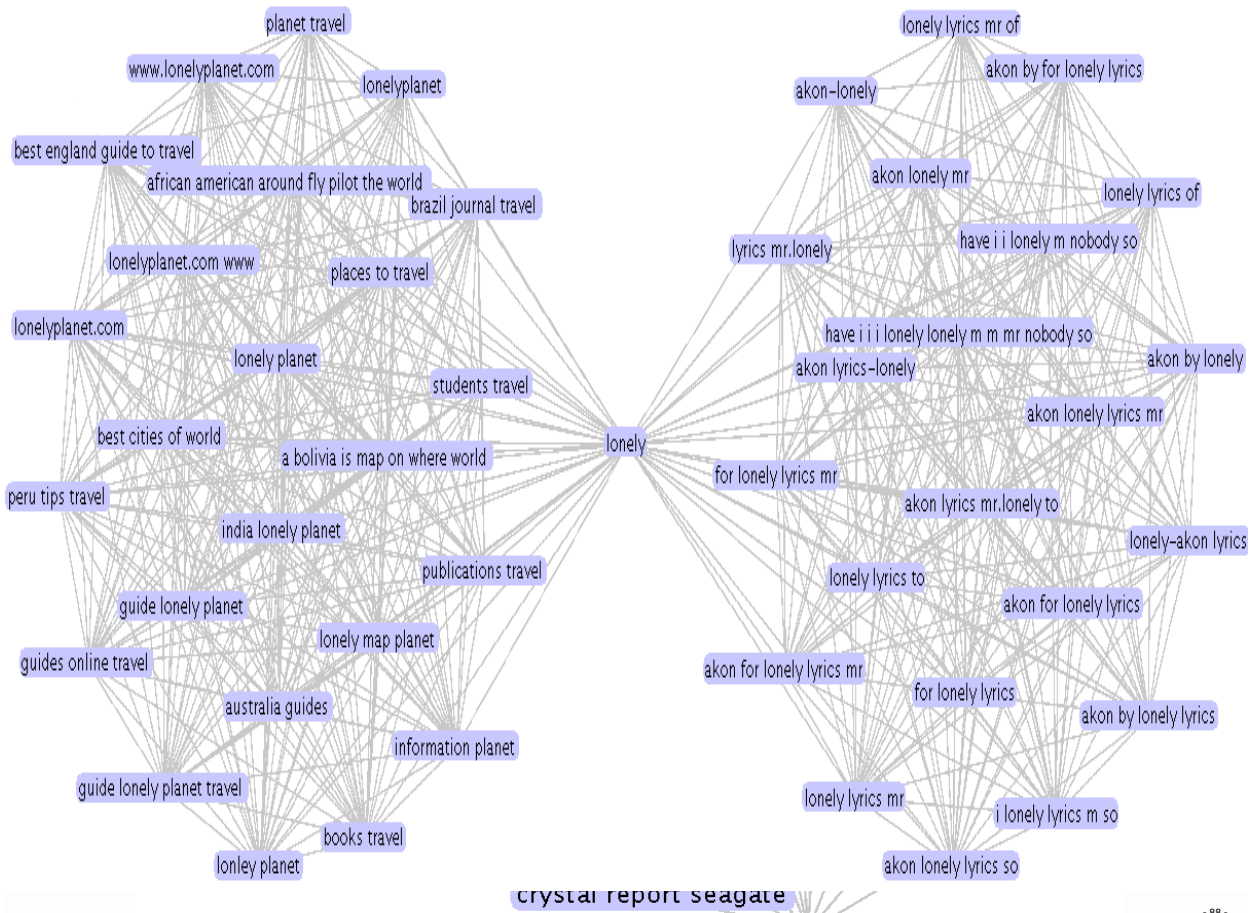
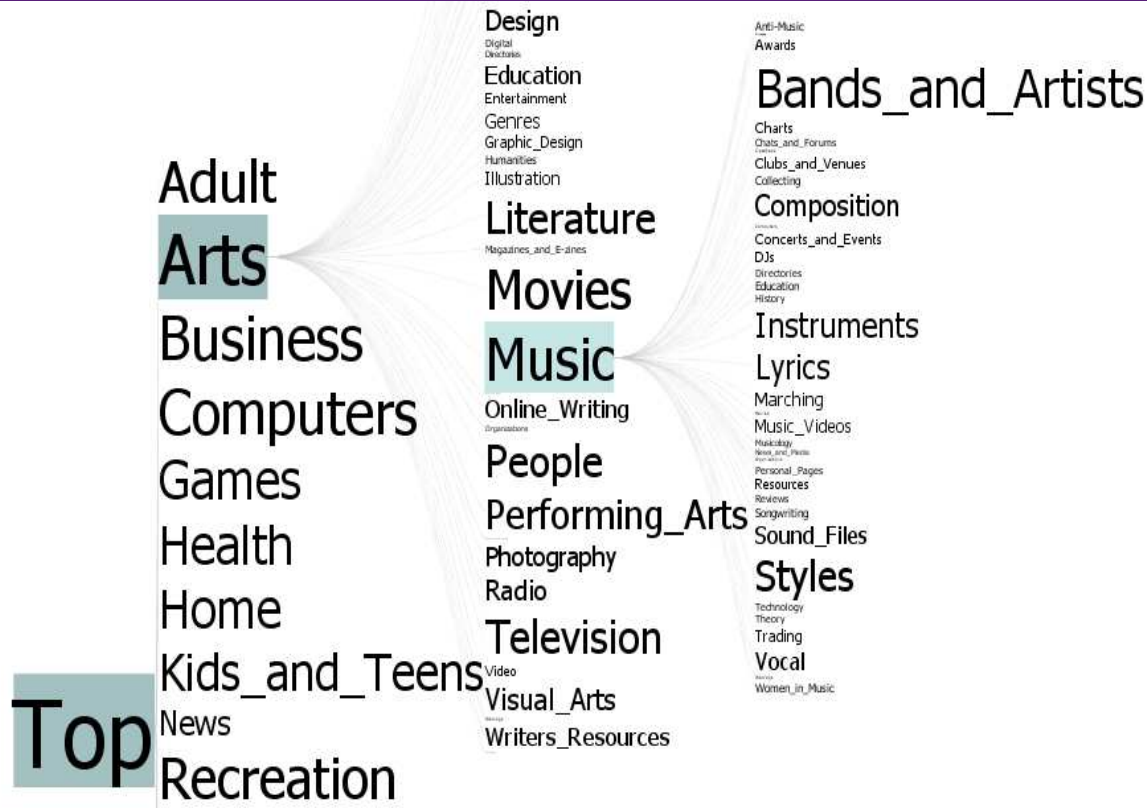
- 85 -

# Experimental Evaluation



- 86 -

# Mapping Queries to ODP



# Hierarchical Clustering

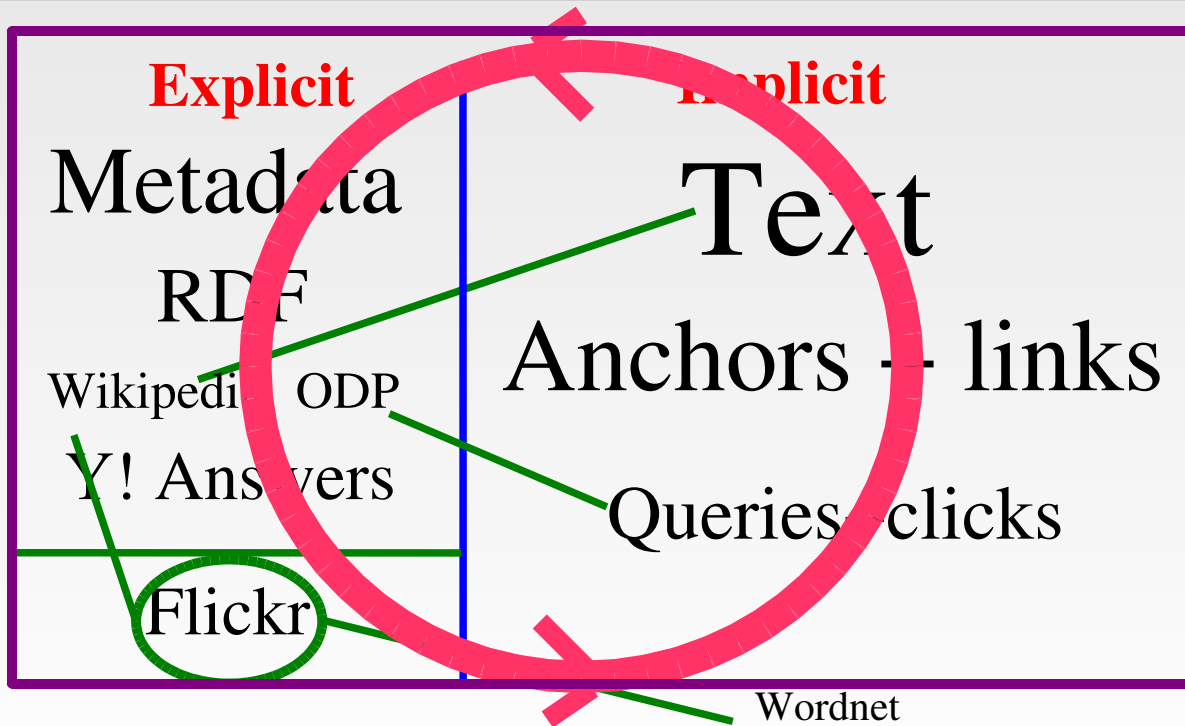
Francisco, Baeza-Yates & Oliveira, submitted



## Open Issues

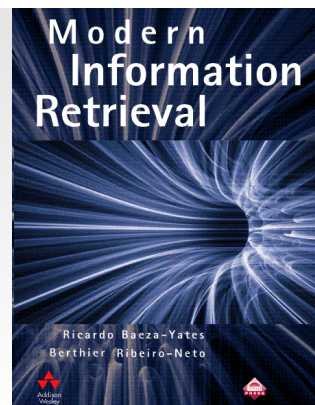
- Data Volume versus Better Algorithms
- Explicit versus implicit social networks
  - Any fundamental similarities?
- How to evaluate with (small) partial knowledge?
  - Data volume amplifies the problem
- User aggregation versus personalization
  - Optimize common tasks
  - Move away from privacy issues

# The Virtuous Cycle



- 91 -

**Second edition  
coming soon**



**Questions?**

**Contact: [rbaeza@acm.org](mailto:rbaeza@acm.org)**

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