On the Selection of Tags for Tag Clouds

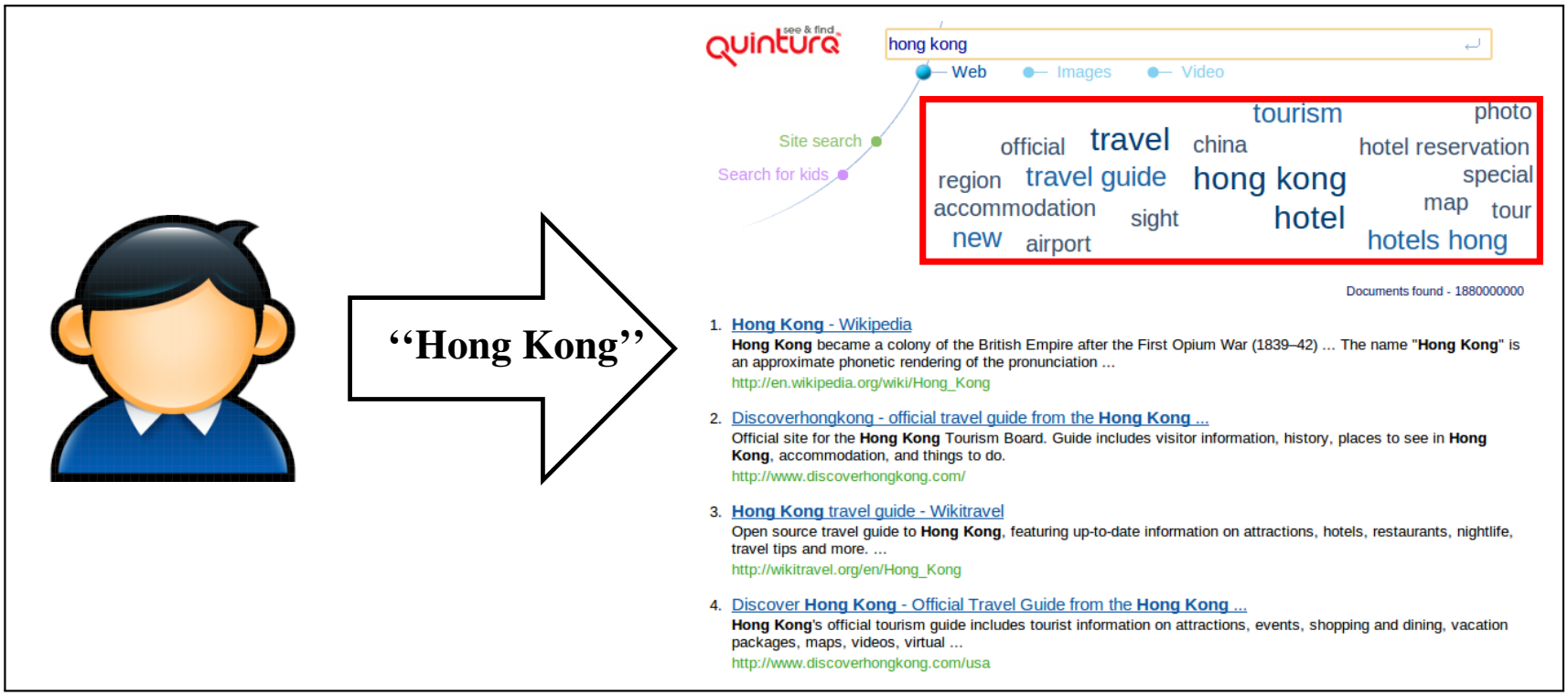
Petros Venetis[1] Georgia Koutrika[2] Hector Garcia-Molina[1]

[1]Computer Science Department
Stanford University

[2]IBM Almaden

February 11, 2011

Search example

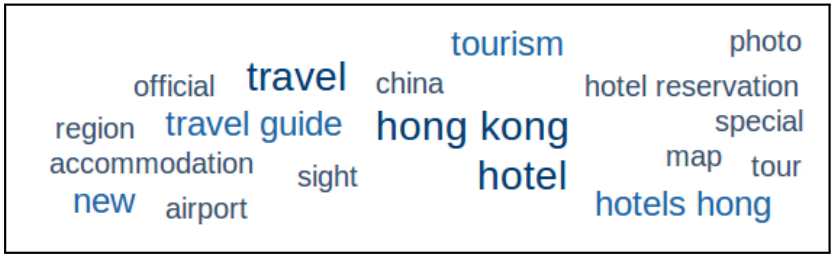


Tag cloud: set of keywords that describe a set of objects mostly for exploration

Other uses

System	Objects	Tags
Search engines (e.g., `quintura.com`)	Webpages	Extracted keywords
CourseRank	Courses	
Technorati	Blog posts	
PubCloud	Medical publications	
`flickr.com`	Photographs	User supplied words
`del.icio.us`	Webpages/Bookmarks	
⋮	⋮	⋮

Our goal



Questions

- What makes a tag cloud good?
- User model?
- Algorithms?

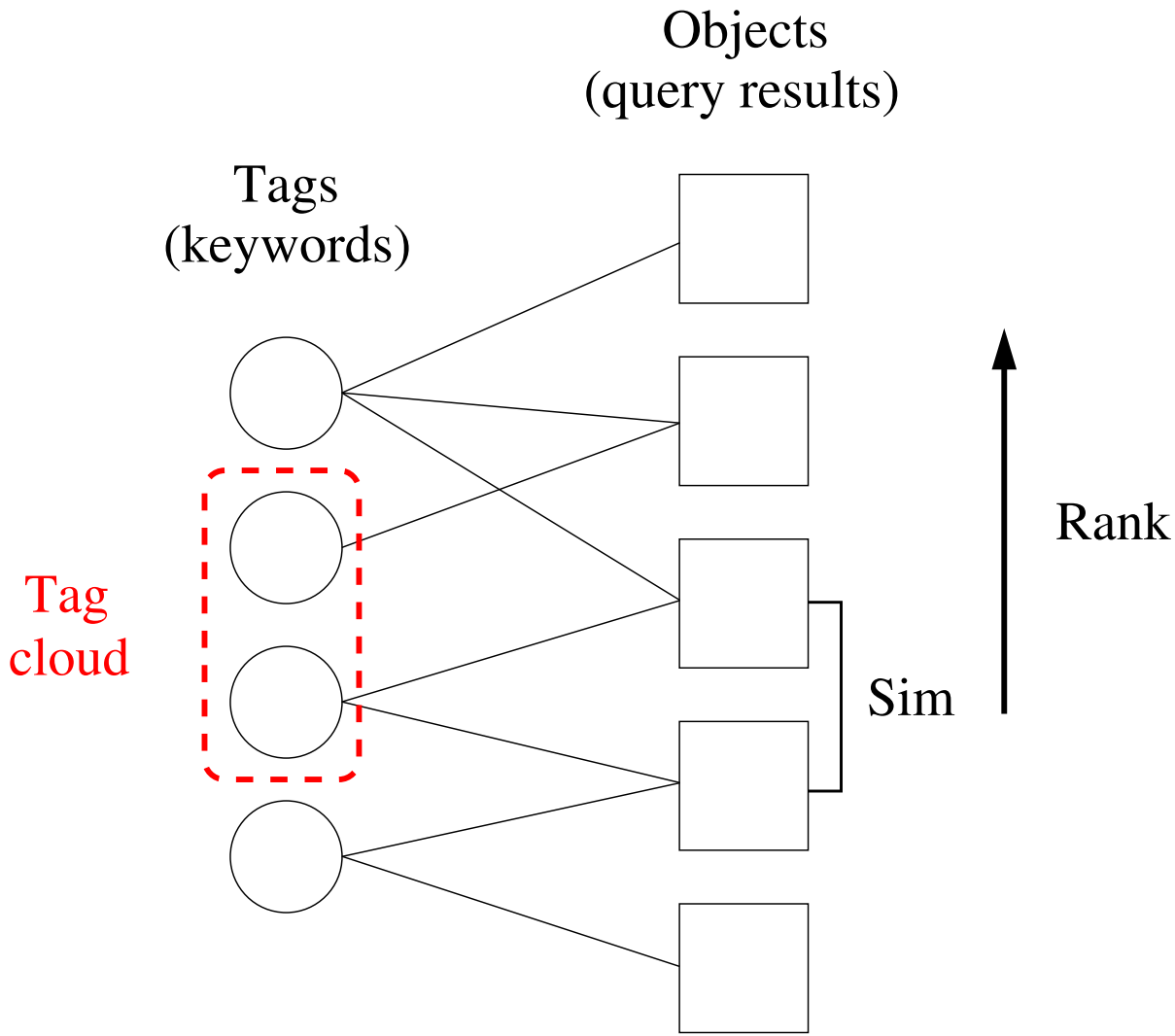
Our focus

- Exploration
- Actual tags (not: color, font size, etc.)

Outline

Outline

System model



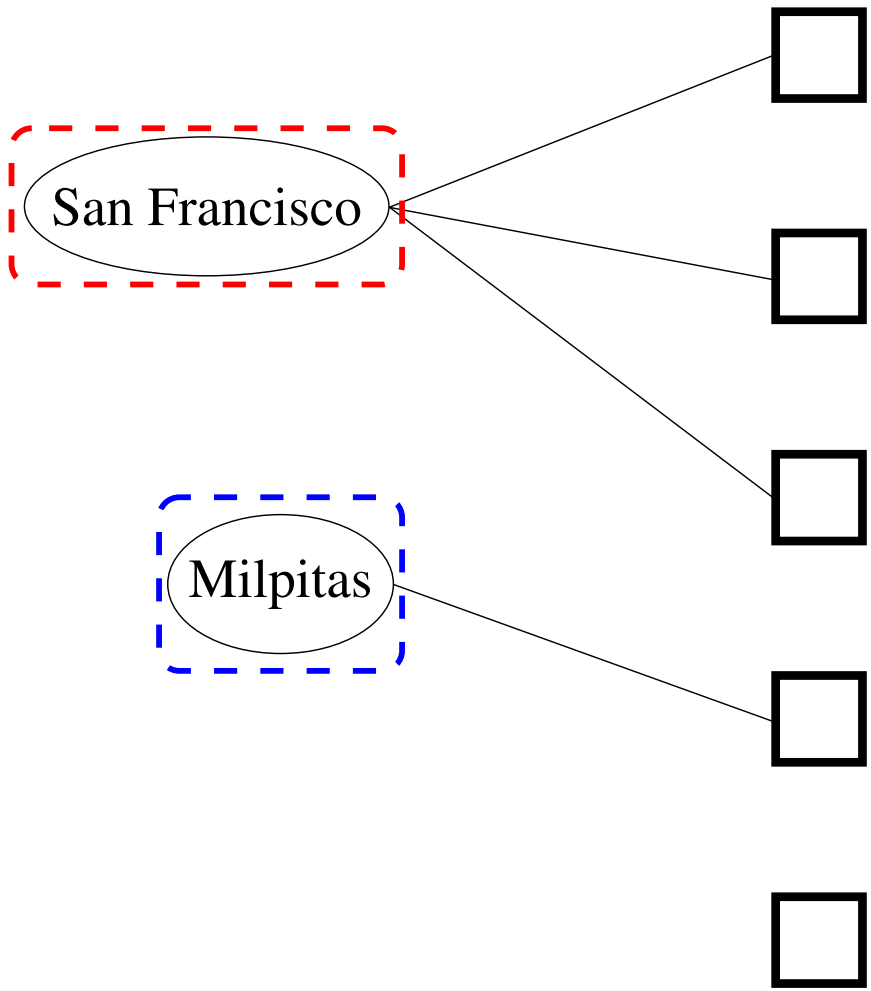
Examined metrics

Metrics examined in our paper

- Coverage
- Overlap
- Cohesiveness
- Relevance
- Extent
- Balance
- Independence
- Popularity

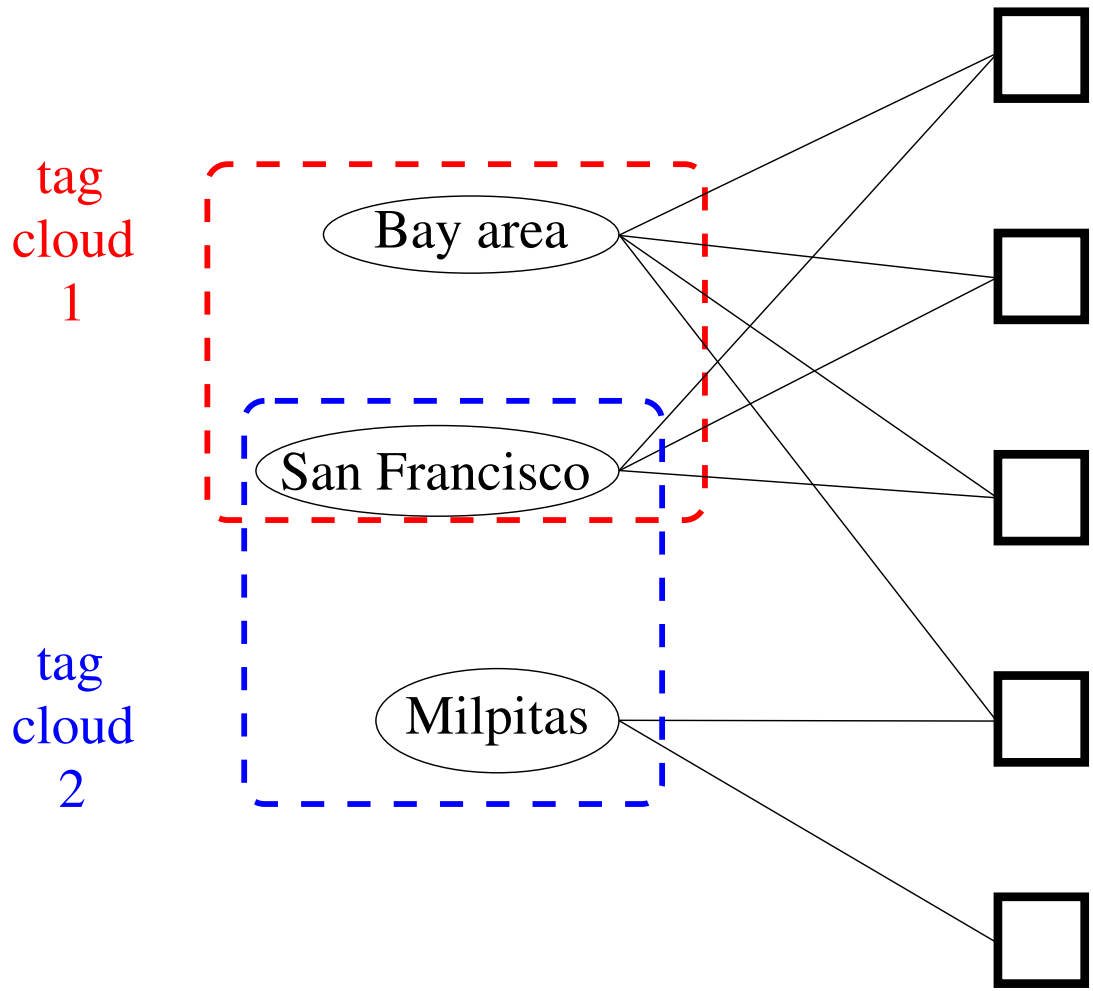
Metrics: Coverage

- Query: “california”
- Result: 5 photographs
- Tag cloud size = 1



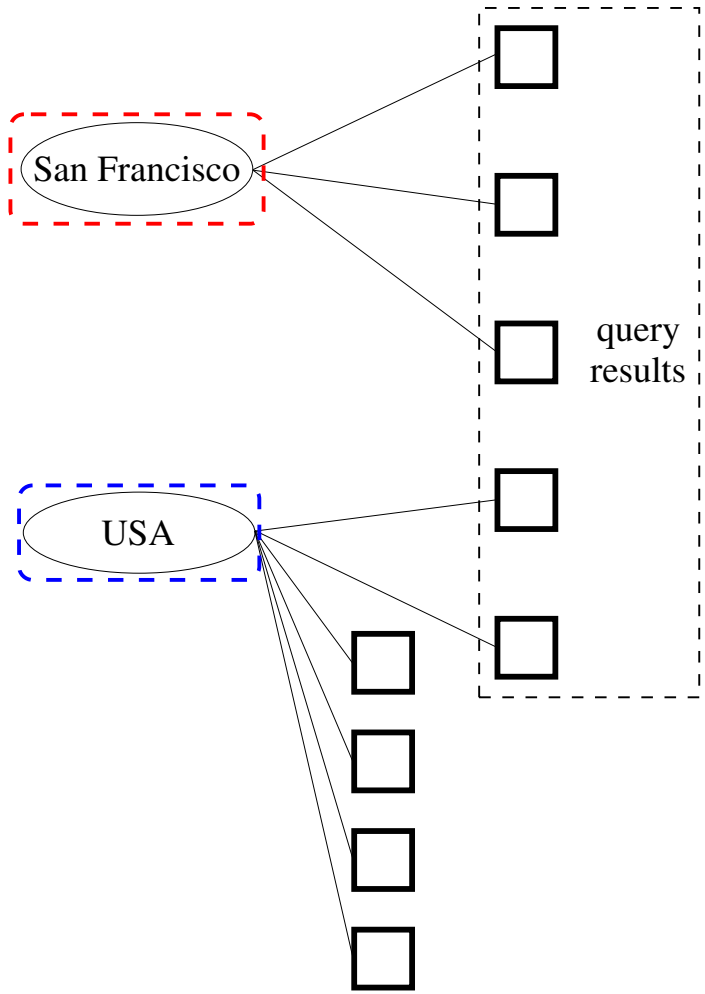
Metrics: Overlap

- Query: “california”
- Result: 5 photographs
- Tag cloud size = 2



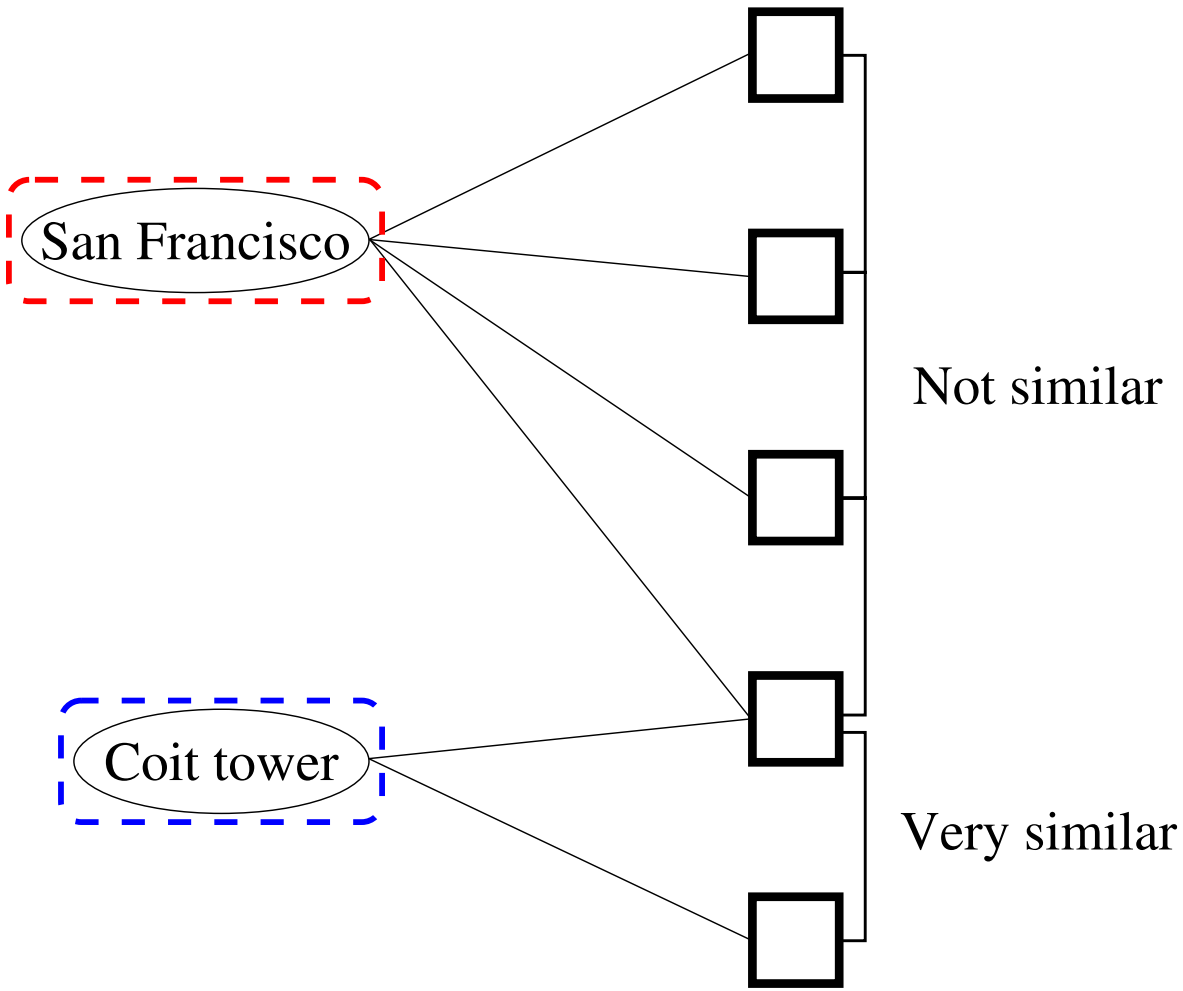
Metrics: Relevance

- Query: “california”
- Result: 5 photographs
- Tag cloud size = 1



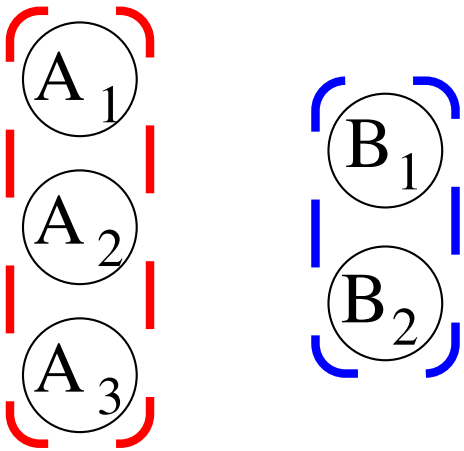
Metrics: Cohesiveness

- Query: “california”
- Result: 5 photographs
- Tag cloud size = 1



Problem

How to compare tag clouds



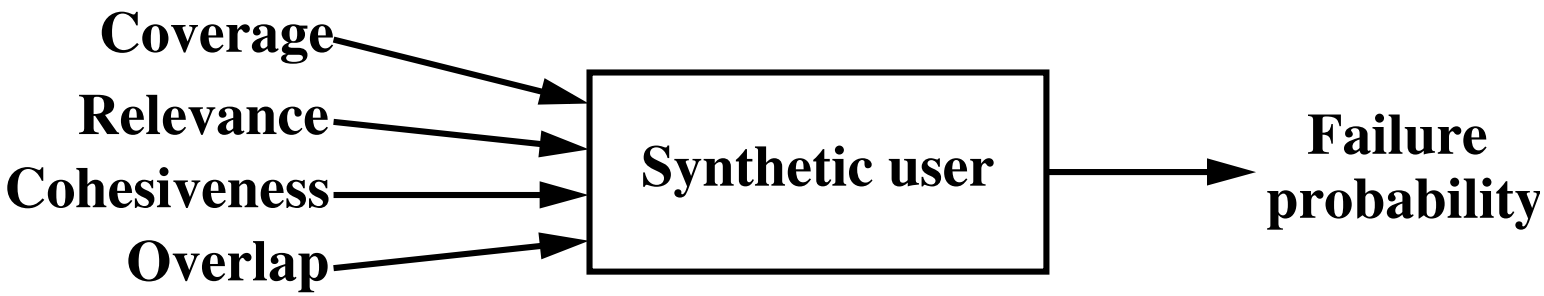
1. Humans
2. Synthetic User

User model

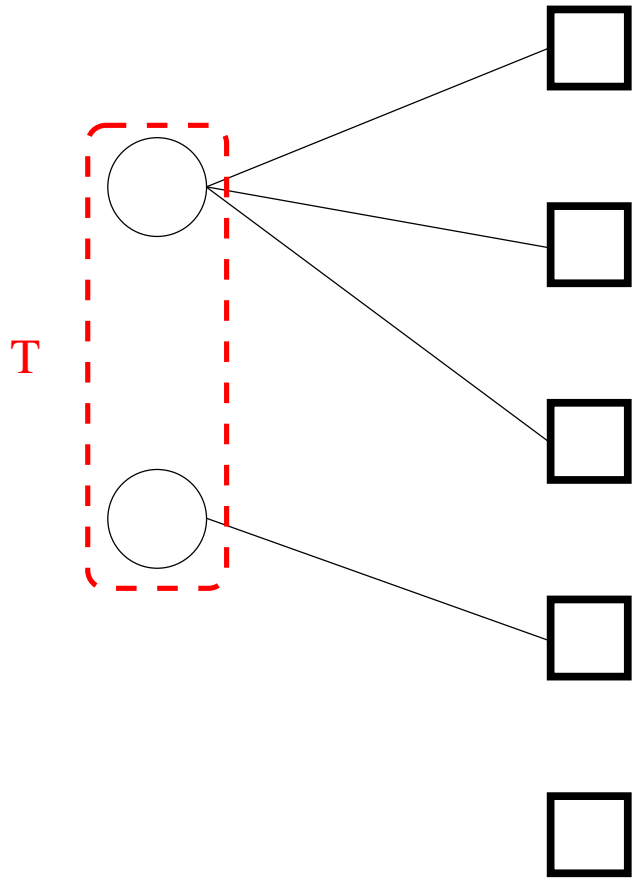
Synthetic user:

1. Searches for a particular object
2. Queries the system
3. Abundance of results
4. Unable to refine query
5. Has to use tag cloud

Failure probability: probability the synthetic user did not find desired object

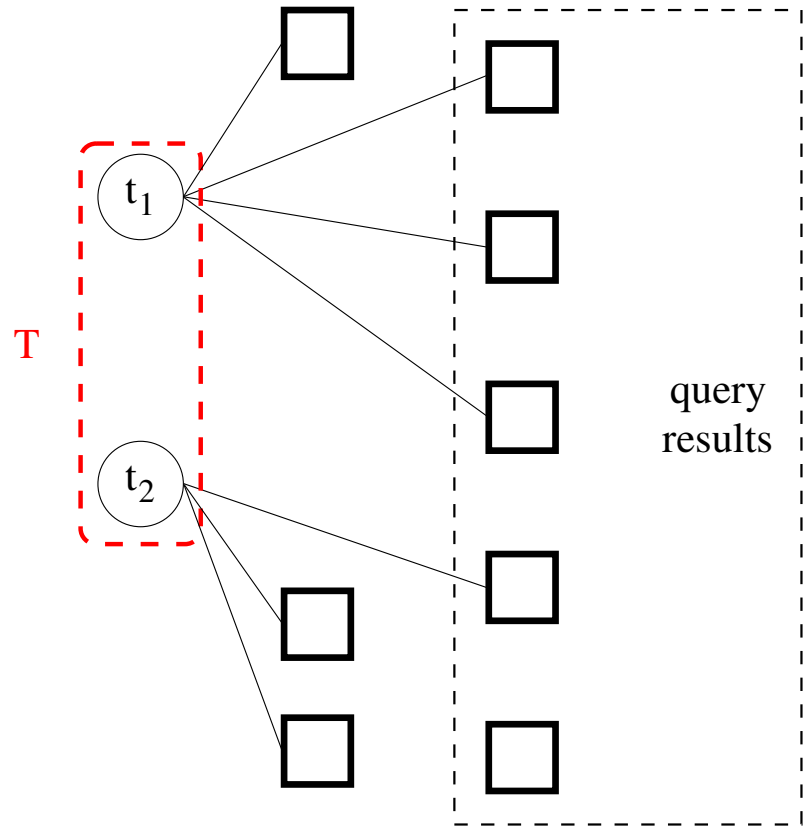


User model details: coverage only



$$\text{Failure probability} = 1 - \text{coverage}(T) = 1 - \frac{4}{5} = 0.2$$

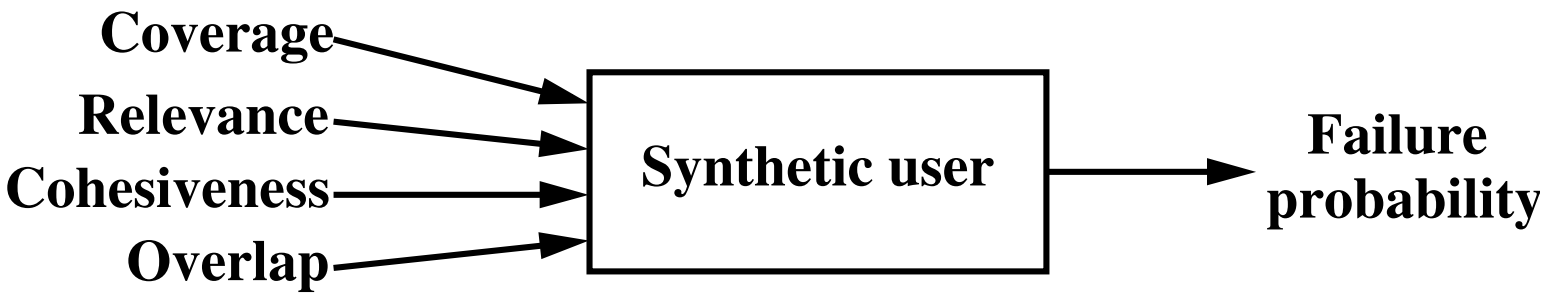
User model details: coverage and relevance



$$\text{Failure probability} = 1 - \sum_{t \in T} \text{coverage}(\{t\}) \cdot \text{relevance}(t, q) =$$

$$= 1 - \frac{3}{5} \cdot \frac{3}{4} - \frac{1}{5} \cdot \frac{1}{3} = 0.48$$

User model details



Model trends

Coverage ↑	Failure probability ↓
Relevance ↑	Failure probability ↓
Cohesiveness ↑	Failure probability ↓
Overlap ↑	Failure probability ↑

Existing algorithms

Algorithms' interface

Input:
- Query results and associated tags
- Budget of tags

Output: Subset of the tags of the bipartite graph

Explored algorithms

- Maximum coverage algorithm (`COV`)
- Popularity-based algorithm (`POP`)
- Tf-idf algorithms (2 versions: `TF` and `WTF`)

Experiments: focus on user model

- Which algorithms work best in real data?
- Can humans agree on the best tag cloud?
- Does our model predict what real users prefer?

Experiments: datasets

`del.icio.us` (thanks Paul Heymann!)

- 100K urls
- ∼400K tags applied in total

CourseRank

- ∼18K courses
- ∼11.5M keywords (excluding stop-words)

Algorithms ordering is stable

- 30 queries: various sizes of query results

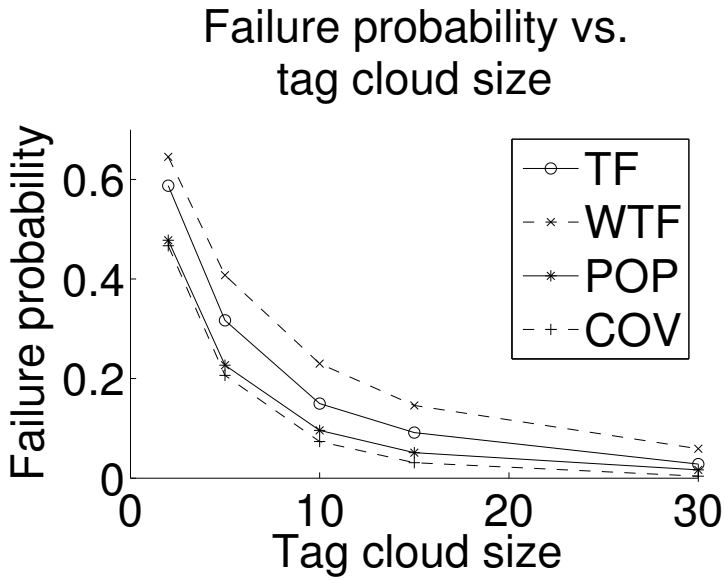


Figure: CourseRank

Algorithms ordering is stable

- 30 queries: various sizes of query results

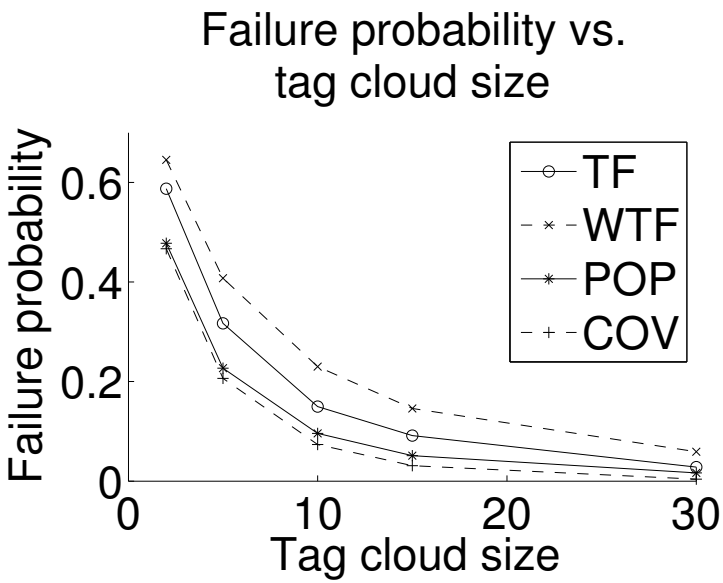


Figure: CourseRank

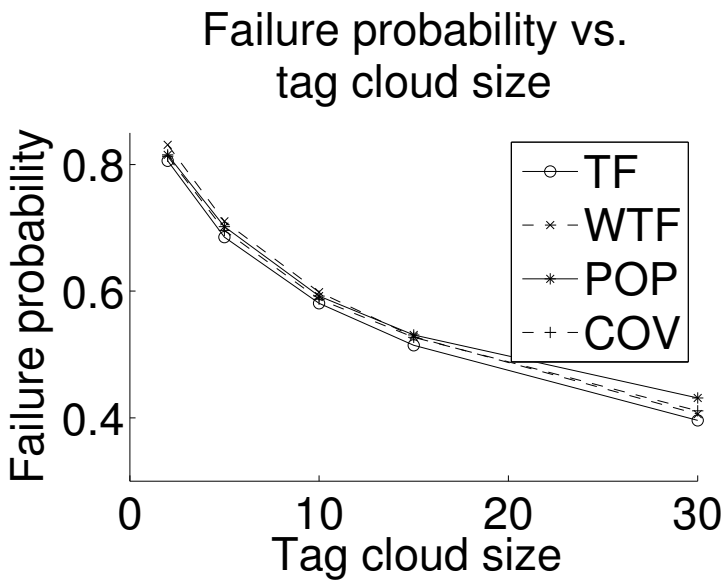


Figure: `del.icio.us`

Algorithms ordering is stable

- 30 queries: various sizes of query results

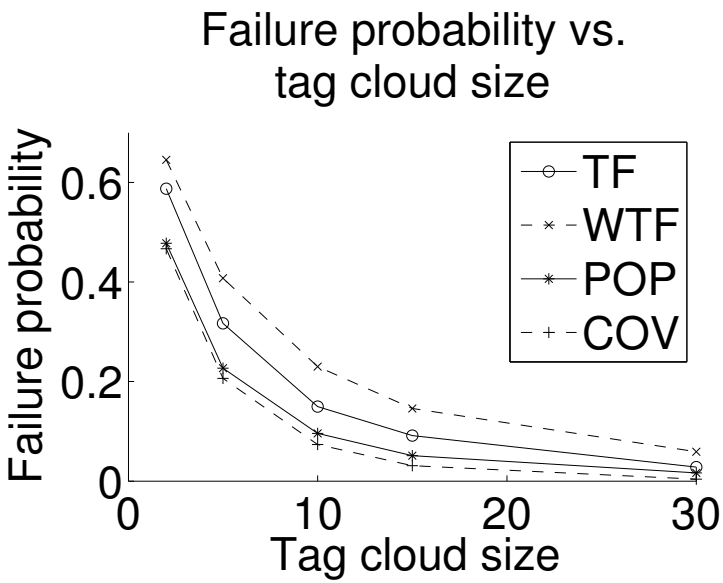


Figure: CourseRank

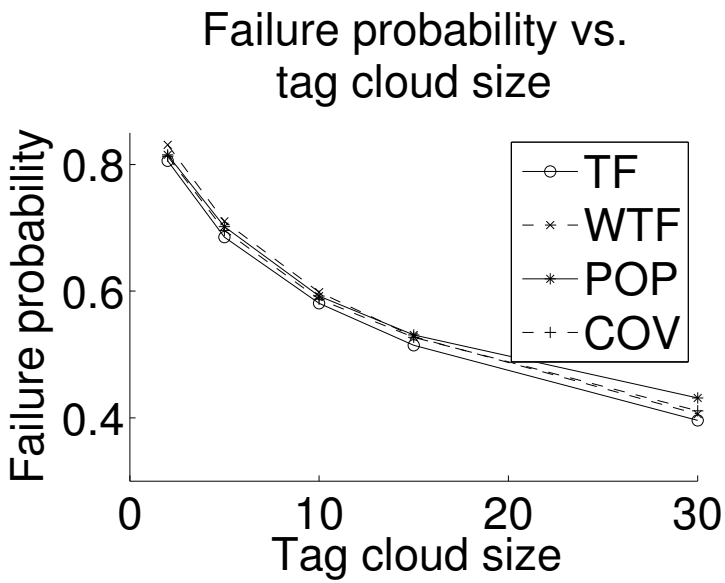
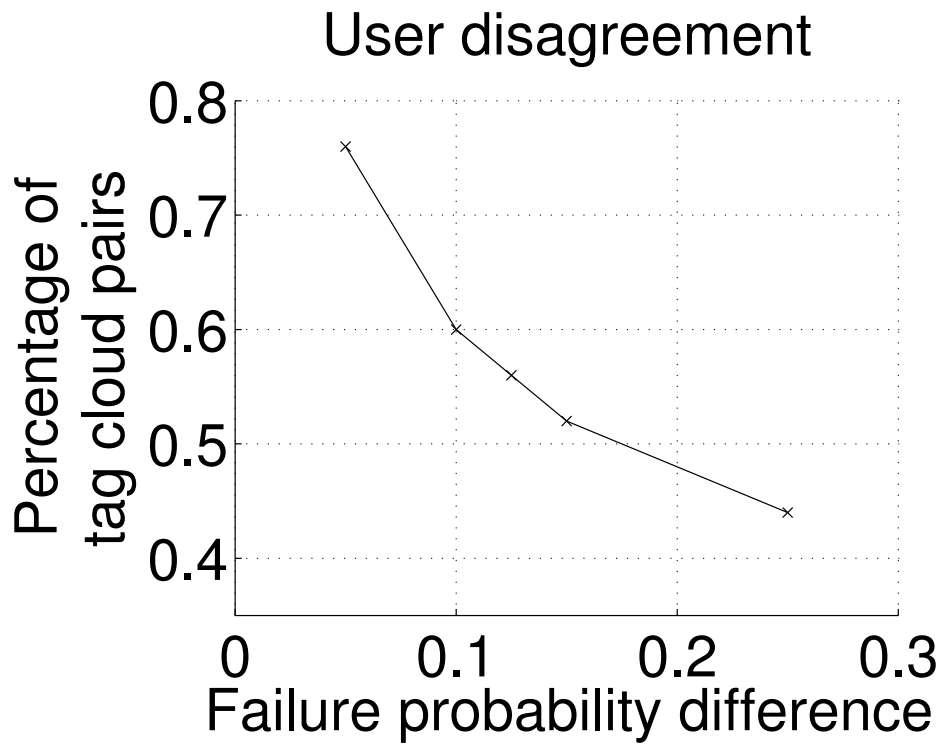


Figure: `del.icio.us`

- Different ordering! (`COV` better in CourseRank, `TF` better in `del.icio.us`)

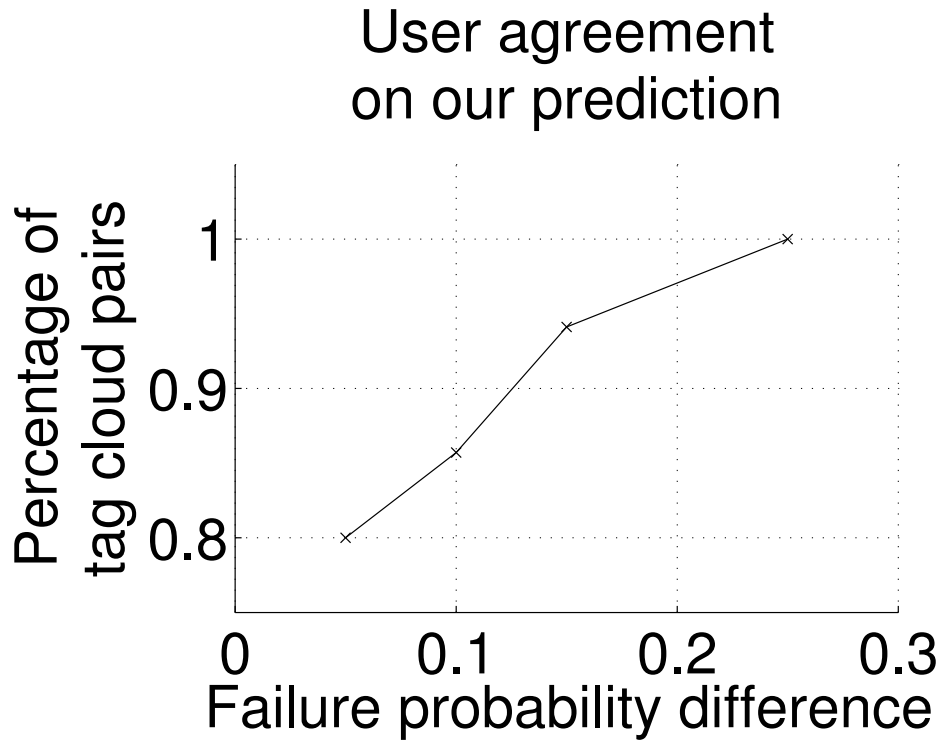
Do users agree on one tag cloud?

- `del.icio.us` dataset
- 450 "random" pairs of tag clouds
- 5 evaluators for each pair
- Agreement: 4 or more evaluators



Does our user model predict the best tag cloud?

- `del.icio.us` dataset
- 450 "random" pairs of tag clouds
- 5 evaluators for each pair
- Agreement: 4 or more evaluators



Conclusions

Summary

- Problem: tag cloud comparison
- Described metrics
- Proposed synthetic user model built on top of the metrics
- Experimentally justified user model
- Provided intuition about algorithms

Future work

- Construction of optimal algorithm
- Items with no assigned tags or spam tags

Thank you!
Questions?