On the Selection of Tags for Tag Clouds

Petros Venetis\textsuperscript{1} Georgia Koutrika\textsuperscript{2} Hector Garcia-Molina\textsuperscript{1}

\textsuperscript{1}Computer Science Department
Stanford University

\textsuperscript{2}IBM Almaden

February 11, 2011
Search example

“,Hong Kong”

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

<table>
<thead>
<tr>
<th>System</th>
<th>Objects</th>
<th>Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search engines (e.g., quintura.com)</td>
<td>Webpages</td>
<td>Extracted keywords</td>
</tr>
<tr>
<td>CourseRank</td>
<td>Courses</td>
<td></td>
</tr>
<tr>
<td>Technorati</td>
<td>Blog posts</td>
<td></td>
</tr>
<tr>
<td>PubCloud</td>
<td>Medical publications</td>
<td></td>
</tr>
<tr>
<td>flickr.com</td>
<td>Photographs</td>
<td>User supplied words</td>
</tr>
<tr>
<td>del.icio.us</td>
<td>Webpages/Bookmarks</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Our goal

Questions
▶ What makes a tag cloud good?
▶ User model?
▶ Algorithms?

Our focus
▶ Exploration
▶ Actual tags (not: color, font size, etc.)
Outline

What is a tag cloud?

Metrics

User model

Algorithms

Experiments
Outline

What is a tag cloud?

Metrics

User model

Algorithms

Experiments
System model

Tag cloud

Objects (query results)

Tags (keywords)

Sim

Rank
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

![Diagram showing tag clouds and locations: Bay area, San Francisco, Milpitas]
Metrics: Relevance

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

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

San Francisco

Coit tower

Not similar

Very similar
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

---

![Diagram](image-url)

- Coverage
- Relevance
- Cohesiveness
- Overlap

**Synthetic user**

**Failure probability**
User model details: coverage only

Failure probability = 1 − coverage(T) = 1 − \frac{4}{5} = 0.2
User model details: coverage and relevance

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

Coverage
Relevance
Cohesiveness
Overlap

Synthetic user

Failure probability

Model trends

<table>
<thead>
<tr>
<th>Coverage ↑</th>
<th>Failure probability ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevance  ↑</td>
<td>Failure probability ↓</td>
</tr>
<tr>
<td>Cohesiveness ↑</td>
<td>Failure probability ↓</td>
</tr>
<tr>
<td>Overlap ↑</td>
<td>Failure probability ↑</td>
</tr>
</tbody>
</table>
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
- \( \sim 400 \text{K} \) tags applied in total

**CourseRank**

- \( \sim 18 \text{K} \) courses
- \( \sim 11.5 \text{M} \) keywords (excluding stop-words)
Algorithms ordering is stable

- 30 queries: various sizes of query results

![Graph showing failure probability vs. tag cloud size for different algorithms: TF, WTF, POP, COV.](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

![Failure probability vs. tag cloud size](image1)

**Figure:** CourseRank

![Failure probability vs. tag cloud size](image2)

**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

![Graph showing user disagreement and failure probability difference](image-url)
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

![Graph showing user agreement on prediction vs. failure probability difference. The x-axis represents failure probability difference ranging from 0 to 0.3. The y-axis represents the percentage of tag cloud pairs ranging from 0.8 to 1. The graph shows an increasing trend in user agreement as the failure probability difference increases.]
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?