Graph based techniques for tag cloud generation

Martin Leginus, Peter Dolog and Ricardo Lage

Department of Computer Science, Aalborg University

May, 2013

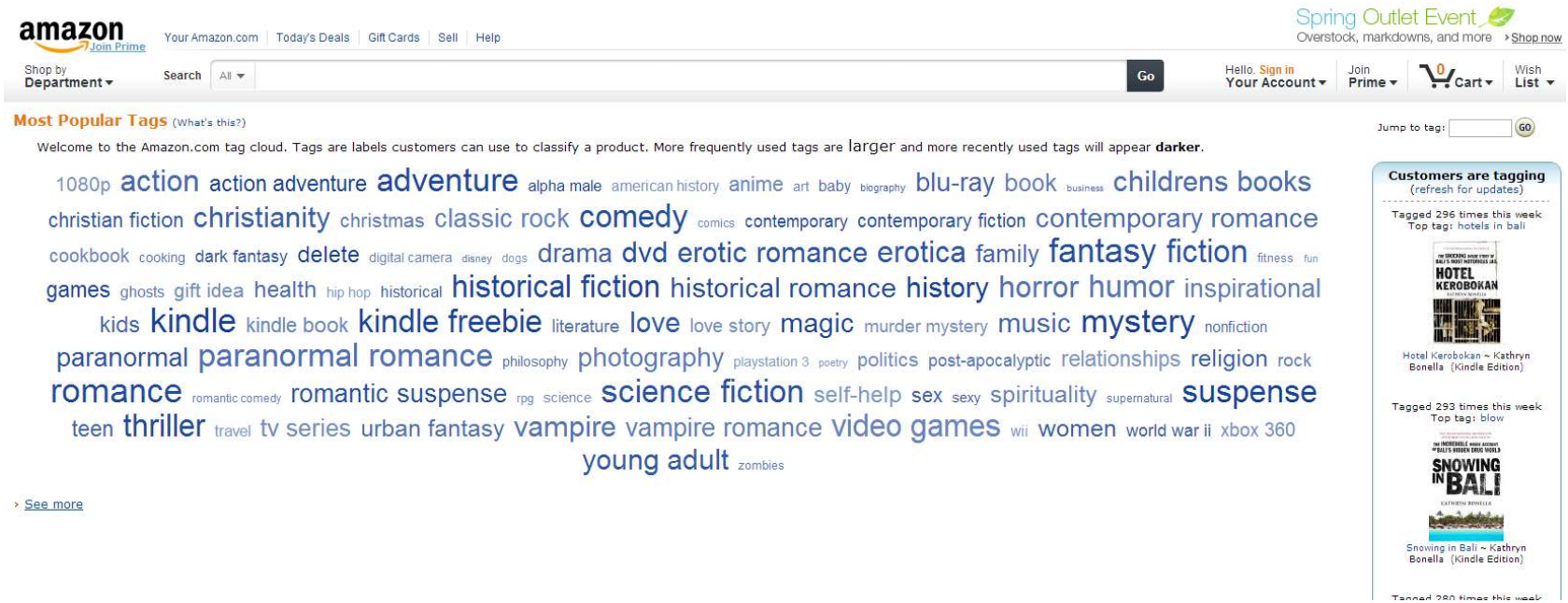
Agenda

- Motivation
- Graph representation
- Importance of tags with respect to a query tag
- External relevance score
- Results
- Discussion
- Future work

Tag clouds

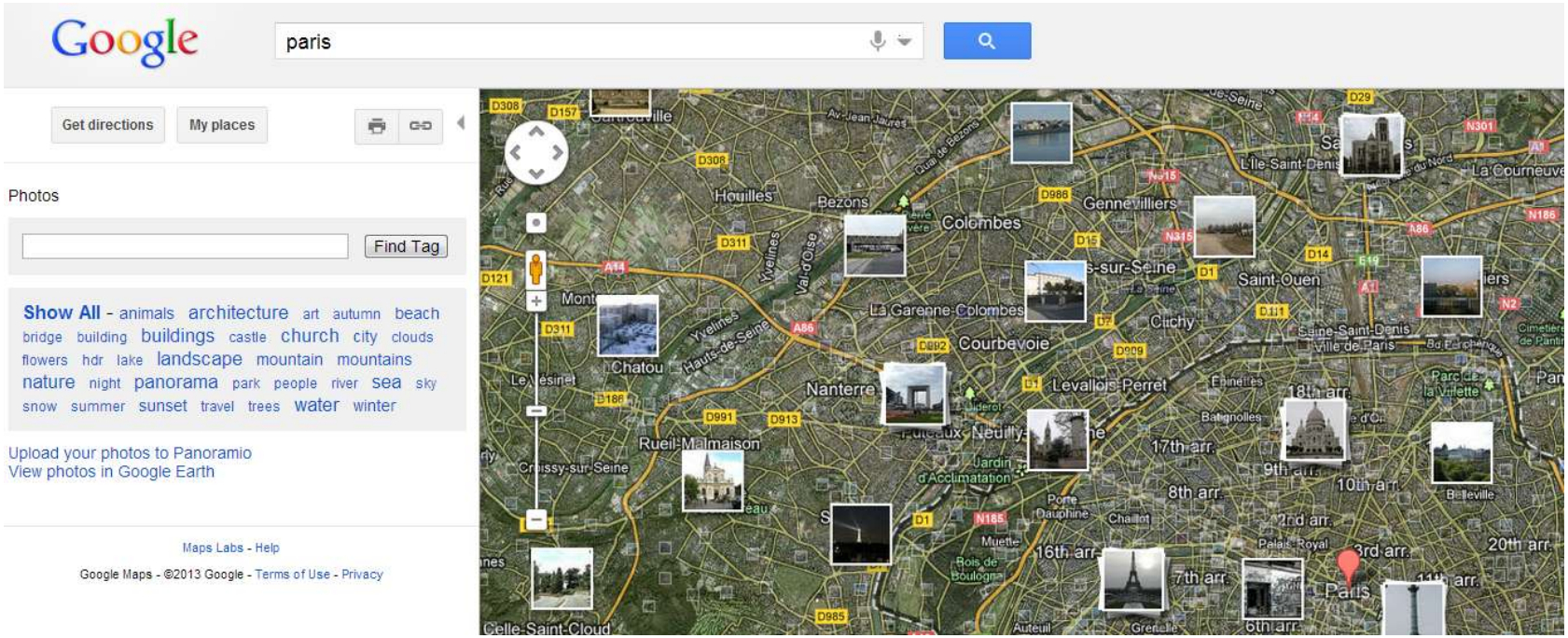
- Tag cloud is a visual retrieval interface depicting the most important terms of a dataset.
- Tag clouds build on top of the entire dataset.
- In this work, we focus on query based tag clouds.

Tag clouds



- Tag clouds build on top of the entire dataset.

Tag clouds



- In this work, we focus on query based tag clouds.

Coverage, Overlap or Relevance?

- Previous works focused on the cloud optimization of Coverage or Overlap (Venetis 2011, Leginus 2012)
- In some cases, this results into irrelevant terms in the cloud. See example for a query *Samuel L. Jackson:*
  divx, nudity(topless), tumeys DVD's, Eric's dvds, classic, bmf, can't remember, action, own, gfei own it, sci-fi, seen more than once, futuristmovies:com; based on book, dvd, DVD, violence, violent

Coverage, Overlap or Relevance?

- General tags about movie character but do not have high discriminative value e.g., *based on book*, *classic*.
- Self-organizing tags assigned by users e.g., *tumeys DVD's*, *Eric's dvds*, *own*, *seen more than once*.
- Other tags e.g., *divx*, *bmf*, *own*, *dvd*, *futuristmovies.com* - no relevance for the given information goal.

Coverage, Overlap or Relevance?

- We utilize relevance as the main synthetic metric.
  Tarantino, uma thurman, john travolta, watched 1994, Winnfield, Samuel L. Jackson as Jules, Quentin Tarantino, too violent
- The actor was performing in the movies of Quentin Tarantino, together with Uma Thurman and John Travolta and one of his famous roles was as Jules Winnfield in Pulp Fiction.
- Coverage can be misleading in measuring quality of a tag cloud.

Motivation and Contributions

- Tag cloud generation as ranking graph nodes and their relevance with respect to the root nodes.
- The underlying tag space can be transformed into a graph.

The main contributions:

- *Various graph based algorithms* for estimating tags relevance condition by a query tag.
- A *new notion of relevance* based on indicators from underlying data.

Co-occurrence calculation

- We compute a tag pair co-occurence using Jaccard similarity for all tags

$$\mathrm{JAC}(t_i, t_j) = \frac{\mathrm{cocr}(t_i, t_j)}{\mathrm{f}(t_i) + \mathrm{f}(t_j) - \mathrm{cocr}(t_i, t_j)} \tag{1}$$

- When the similarity is greater than a predefined threshold α, such tags are considered as similar.
- Each similar tag pair is transformed into two directed edges $t_1 \rightarrow t_2$ and $t_2 \rightarrow t_1$.

Co-occurrence calculation

- *Samuel L. Jackson* assigned to
  - Goodfellas (1990),Pulp Fiction (1994),Die Hard: With a Vengeance (1995),Kill Bill: Vol. 2 (2004)
- *Tarantino* assigned to
  - Reservoir Dogs (1992),Pulp Fiction (1994),Four Rooms (1995),Jackie Brown (1997),Kill Bill: Vol. 1 (2003),Kill Bill: Vol. 2 (2004)
- Cooccurring at Pulp Fiction and Kill Bill: Vol. 2

$$\mathrm{JAC}(\text{Samuel L. Jackson}, \text{Tarantino}) = \frac{2}{6+4-2} = \frac{1}{4} \quad (2)$$

Estimating relative importance of graph nodes

- Algorithms rank an importance of a tag t with respect to the query tag t_q where $\{t, t_q\} \in G$, denoted as:

$$I(t|t_q)$$

- Two distinct types of estimation:
  1. Distance based approaches
  2. **Stochastic approaches**

Stochastic approaches

Measuring importance of nodes in the graph through the simulation of a stochastic process i.e., random traversing of the graph.
The transition probability from a node t_1 to t_2 is defined as

$$p(t_2|t_1) = \frac{1}{d_{out}(t_1)}$$

for all nodes t_2 that have an ingoing edge from t_1.

Pagerank with priors

Pagerank models the behaviour of a random surfer.
Relative importance to a query tag is introduced through the vector of prior probabilities $p_R = \{p_1 \dots p_{|V|}\}$.
A random surfer is assured with a back probability β

$$\pi(v)^{(i+1)} = (1-\beta)\left(\sum_{u=1}^{d_{in}(v)} p(v|u)\pi^{(i)}(u)\right) + \beta p_v \qquad (3)$$

The resulting ranks biased towards t_q are considered as definition of importance after convergence i.e.;

$$I(t|t_q) = \pi(t) \qquad (4)$$

HITS with priors

Relative importance to a query tag is introduced through the vector of prior probabilities $p_R = \{p_1 \dots p_{|V|}\}$.

$$a(v)^{(i+1)} = (1-\beta)\left(\sum_{u=1}^{d_{in}(v)} \frac{h^{(t)}(u)}{H^{(i)}}\right) + \beta p_v \tag{5}$$

$$h(v)^{(i+1)} = (1-\beta)\left(\sum_{u=1}^{d_{out}(v)} \frac{a^{(t)}(u)}{A^{(i)}}\right) + \beta p_v \tag{6}$$

where

$$H^i = \sum_{v=1}^{|V|} \sum_{u=1}^{d_{in}(v)} h^i(u) \tag{7}$$

$$A^i = \sum_{v=1}^{|V|} \sum_{u=1}^{d_{out}(v)} a^i(u) \tag{8}$$

k-step Markov Chain

This method differs in the implementation of a random surfer model. It is assured with a path length limitation - determines how often we jump back to a root node.

$$I(v|R) = [A \cdot p_R + A^2 \cdot p_R \dots A^K \cdot p_R] \tag{9}$$

Relative importance to a root node is introduced through a vector p_R of prior probabilities.

Datasets

The Bibsonomy dataset contains 206589 items, 51565 tags and 466818 tagging posts.
The Movielens dataset contains 16518 tags, 7601 movies and 95580 tagging posts.

Tag cloud's metrics

Synthetic metrics express a quality of tags selection process (Venetis 2011).

- A coverage for a particular tag t expresses how many of considered documents were annotated with a tag t

$$\text{Coverage}(t) = \frac{|D_t|}{|D_a|}, \tag{10}$$

- *Overlap of* T_c: Different tags in T_c may be assigned with the same item in D_{T_c}. The overlap metric captures the extent of such redundancy.

$$\text{Overlap}(T_c) = avg_{t_i \neq t_j} \frac{|D_{t_i} \cap D_{t_j}|}{\min\{|D_{t_i}|, |D_{t_j}|\}}, \tag{11}$$

- *Relevance of* T_c: Expresses how relevant the tags in T_c are to the query tag t_q. We compute a relevance of each tag t from T_c in the following fashion:

$$Relevance(T_c) = avg_{t \in T_c} \frac{|D_t \cap D_{t_q}|}{|D_t|}, \tag{12}$$

Evaluation methodology

We randomly select 30 distinct tag queries from each dataset. For each query tag t_q we perform the following evaluation:

1. Generate a tag cloud with respect to a given query tag t_q utilizing specific tags selection method such that tag cloud contains at most k-tags.
2. Measure *coverage*, *overlap* and *relevance* of the generated tag cloud (only on top of relevant resources).
3. Increase the size of tag cloud k. If maximum size is reached, change to the other selection method.

Baseline techniques

- Most Frequent Tags from Corpus (MFTC)
- Most Frequent Tags from query tag's documents (MFTQD)
- Most popular tags (POP)
- Term frequency - inverse document frequency selection (TFIDF)
- Weighted Term frequency - inverse document frequency selection (WTFIDF)
- Max coverage selection (COV)

Results

Dataset/Method	Overlap			
	25	50	75	100
Movielens - k-MarkovCh	0.65	0.43	0.27	0.18
Movielens - WTFIDF	0.51	0.37	0.29	0.24
Bibsonomy - k-MarkovCh	0.40	0.27	0.23	0.20
Bibsonomy - WTFIDF	0.49	0.35	0.35	0.31

Table: Mean values of overlap for the WTFIDF and k-step Markov Chain on BibSonomy and Movielens datasets.

Dataset/Method	Relevance			
	25	50	75	100
Movielens - k-MarkovCh	0.57	0.47	0.4	0.33
Movielens - WTFIDF	0.19	0.18	0.20	0.22
Bibsonomy - k-MarkovCh	0.37	0.28	0.24	0.23
Bibsonomy - WTFIDF	0.44	0.43	0.31	0.30

Table: Mean values of relevance for the WTFIDF and k-step Markov Chain on BibSonomy and Movielens datasets.

Results

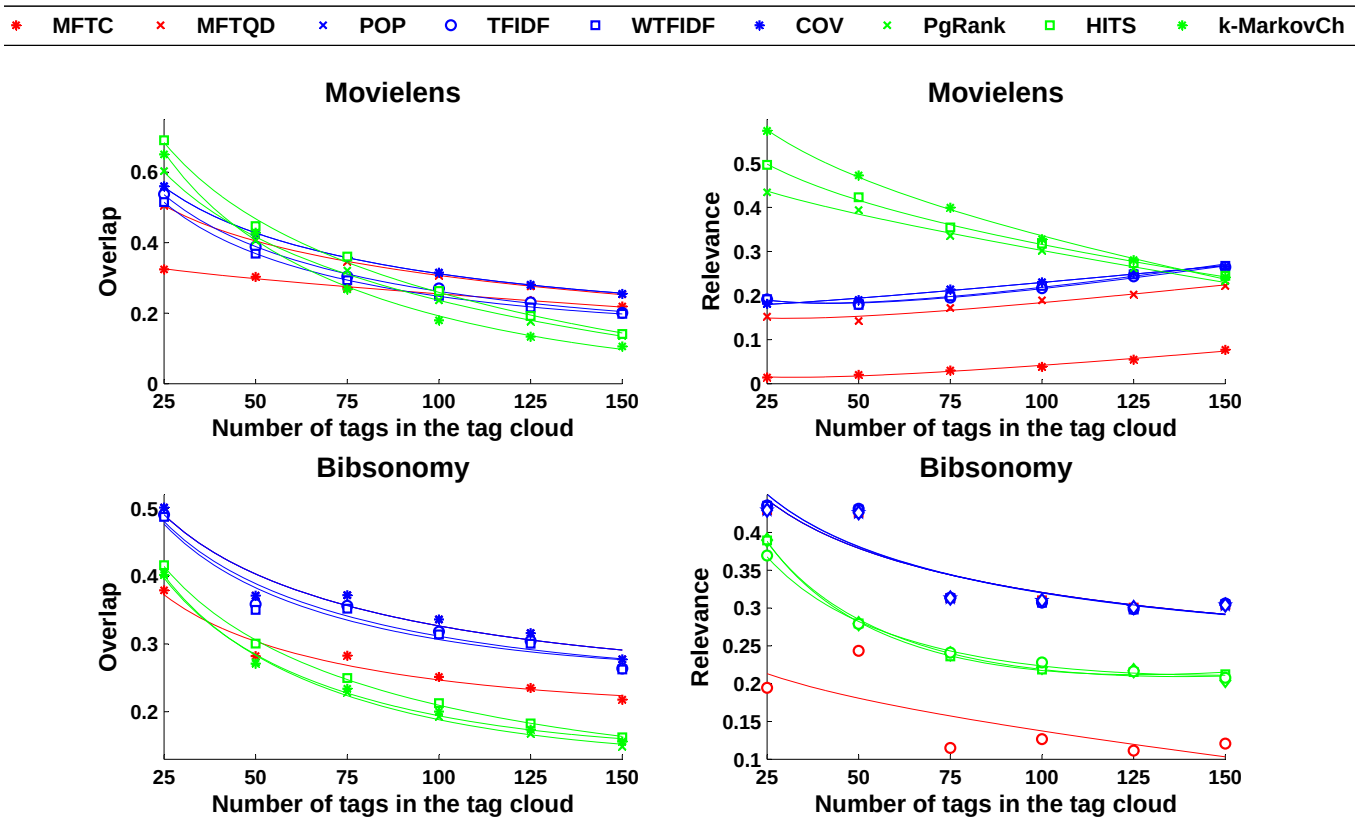


Figure: Overlap and Relevance with normal selection of resources on Bibsonomy and Movielens datasets with different selection techniques and their corresponding logarithmic fit.

External indicators of relevance

We introduce indicators of relevance based on corresponding sources such are citation count and average movies ratings.
For each research publication from Bibsonomy, we downloaded a citation number using Microsoft Academic Search.
We calculated an average rating assigned by users for a movie from Movielens.

Relevance based on indicators

Dataset/Method	Overlap			
	25	50	75	100
Movielens - k-MarkovCh	0.66	0.43	0.27	0.18
Movielens - WTFIDF	0.52	0.37	0.29	0.25
Bibsonomy - k-MarkovCh	0.43	0.29	0.25	0.21
Bibsonomy - WTFIDF	0.41	0.30	0.32	0.27

Table: Mean values of overlap for the WTFIDF and k-step Markov Chain on BibSonomy and Movielens datasets.

Dataset/Method	Relevance			
	25	50	75	100
Movielens - k-MarkovCh	0.6	0.5	0.42	0.35
Movielens - WTFIDF	0.19	0.17	0.20	0.22
Bibsonomy - k-MarkovCh	0.43	0.33	0.29	0.27
Bibsonomy - WTFIDF	0.32	0.28	0.30	0.29

Table: Mean values of relevance for the WTFIDF and k-step Markov Chain on BibSonomy and Movielens datasets.

Relevance based on indicators

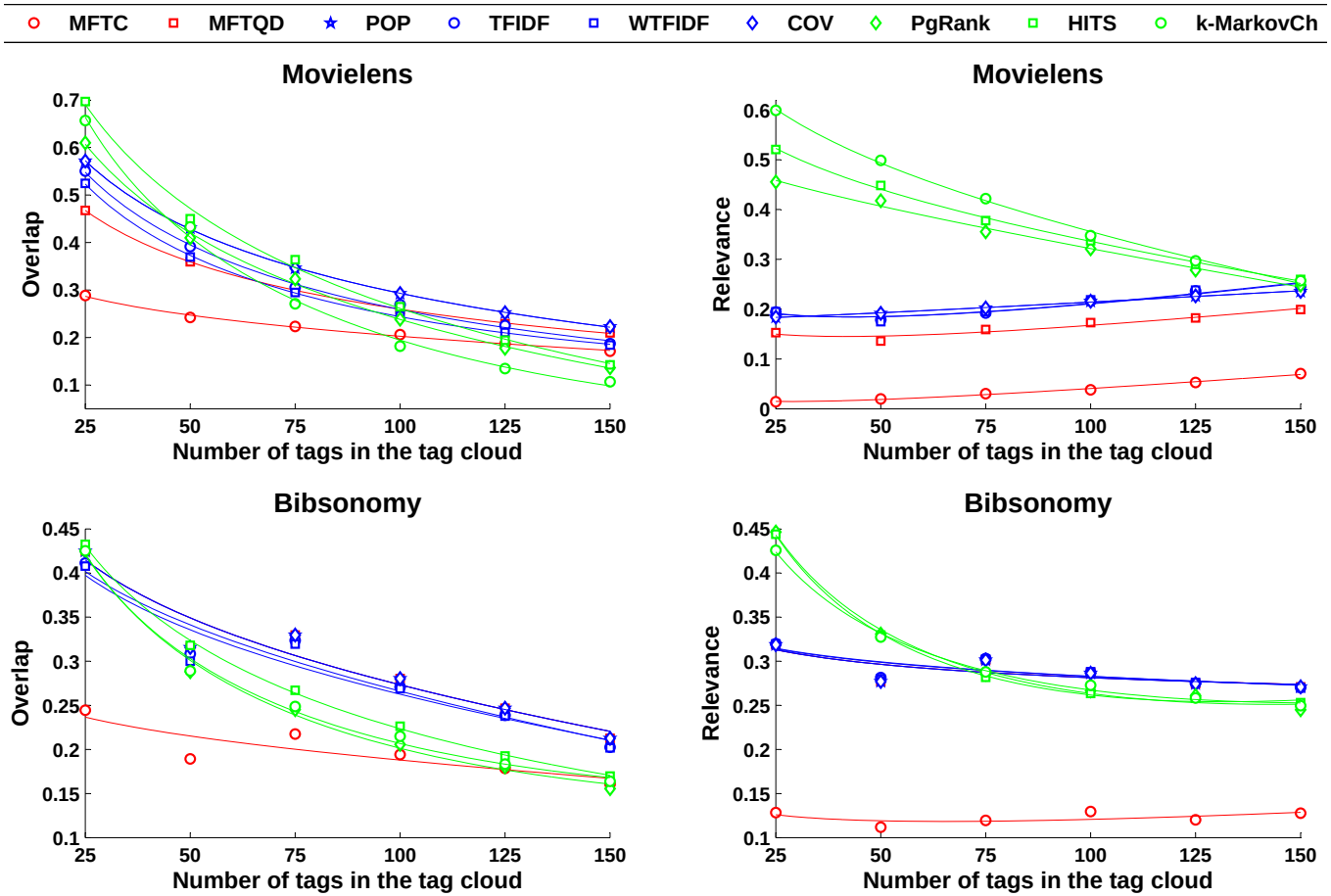


Figure: Overlap and Relevance with relevant selection of resources on Bibsonomy and Movielens datasets with different selection techniques and their corresponding logarithmic fit.

Limitations

- Graph methods are depedent on various parameters.
- Only binary relevance based on external indicators.

Contributions

The main contributions:

- *Various graph based algorithms* for estimating tags relevance condition by a query tag.
- A *new notion of relevance* based on indicators from underlying data.
- Improved relevance of tag clouds *with 41 % on Movielens* dataset and *11 % on Bibsonomy* in comparison to the state-of-the-art tag selection techniques.

Future work

- Explore various relevance scoring functions.
- Extend co-occurrence through relational information.
- Propose new methods for deriving relavance from underlying data.

Questions


Only Five Questions
Great, I can't see who has the prearranged questions!