

The Art of Tagging: Measuring the Quality of Tags

Ralf Krestel and Ling Chen

L3S Research Center, Hannover, Germany

February 4th, 2009

Motivation: Tagging Systems

flickr®

citeulike 

You  Tube

 last.fm



delicious

Motivation: Del.icio.us

Searching Everybody's bookmarks for:

aswc Search

Sign In to search your own bookmarks

See all bookmarks tagged aswc

Search all of Delicious for "aswc"

Search for this on the web

Web Search

Everybody's bookmarks 29 results - show all

3rd Asian Semantic Web Conference (ASWC 2008) SAVE 16

First saved by: chelana conference semanticweb 2008 asia web

1st Asian Semantic Web Conference SAVE 9

First saved by: lechatpito conference semanticweb 2006 semantic aswc

Alaska State Writing Consortium SAVE 4

First saved by: msdcavid writing apple arts language aswc

Lara Crome-Guthrie > Introduction > Who I Am SAVE

First saved by: laracguthrie imported aswc

Course: LEAD TECH GROUP SAVE

First saved by: SusanAlexis aswc

The 6th International Semantic Web Conference(ISWC 2007) SAVE 38

First saved by: castagna semanticweb conference 2007 iswc conferences

ASWC - Utah College, Private Liberal Arts College, Westminster College, Salt Lake City, Utah SAVE 2

First saved by: solpin film westminster movie documentary

conferences on the SemanticDesktop calendar SAVE

First saved by: wroble 2006 conference semanticdesitopws2006 escw aswc

1st Asian Semantic Web Conference SAVE

First saved by: ShengliangKu aswc

Semantic Web Applications and Tools Workshop SAVE 2

First saved by: ShengliangKu workshop research resources conference tools

Everyone's Related Tags

Outline

- 1 Introduction
- 2 Related Work
- 3 Measuring Tag Quality
 - Tagging System Model
 - Quality Propagation
 - Seed Selection Strategies
- 4 Evaluation
 - Dataset
 - Method
 - Results
- 5 Conclusions & Future Work

Outline

- 1 Introduction
- 2 Related Work
- 3 Measuring Tag Quality
 - Tagging System Model
 - Quality Propagation
 - Seed Selection Strategies
- 4 Evaluation
 - Dataset
 - Method
 - Results
- 5 Conclusions & Future Work

Outline

- 1 Introduction
- 2 Related Work
- 3 Measuring Tag Quality
 - Tagging System Model
 - Quality Propagation
 - Seed Selection Strategies
- 4 Evaluation
 - Dataset
 - Method
 - Results
- 5 Conclusions & Future Work

Outline

- 1 Introduction
- 2 Related Work
- 3 Measuring Tag Quality
 - Tagging System Model
 - Quality Propagation
 - Seed Selection Strategies
- 4 Evaluation
 - Dataset
 - Method
 - Results
- 5 Conclusions & Future Work

Outline

- 1 Introduction
- 2 Related Work
- 3 Measuring Tag Quality
 - Tagging System Model
 - Quality Propagation
 - Seed Selection Strategies
- 4 Evaluation
 - Dataset
 - Method
 - Results
- 5 Conclusions & Future Work

Introduction

Exploitation of Tags

Tags (user generated annotations) can facilitate search and browsing

Quality of Tags

Important in tagging systems for:

- High quality search
- Handling of spam
- Recommendation
- Browsing

Introduction

Exploitation of Tags

Tags (user generated annotations) can facilitate search and browsing

Quality of Tags

Important in tagging systems for:

- High quality search
- Handling of spam
- Recommendation
- Browsing

- 1 Introduction
- 2 Related Work
- 3 Measuring Tag Quality
 - Tagging System Model
 - Quality Propagation
 - Seed Selection Strategies
- 4 Evaluation
 - Dataset
 - Method
 - Results
- 5 Conclusions & Future Work

Related Work

Graph Model for Folksonomies

Andreas Hotho, Robert Jäschke,
Christoph Schmitz and Gerd Stumme:
Information Retrieval in Folksonomies: Search and Ranking

Quality Propagation to Fight Web Spam

Zoltán Gyöngyi, Hector Garcia-Molina and Jan O. Pedersen:
Combating Web Spam with TrustRank

- 1 Introduction
- 2 Related Work
- 3 Measuring Tag Quality**
 - Tagging System Model
 - Quality Propagation
 - Seed Selection Strategies
- 4 Evaluation
 - Dataset
 - Method
 - Results
- 5 Conclusions & Future Work

Measuring Tag Quality

Motivation

- hidden value of tag data
- application dependent usefulness
- no limitation on vocabulary
- varying tag quality
- tag quality not independent of resource

Problem Specification

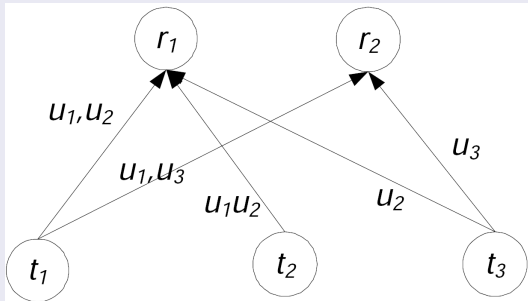
- Let \mathcal{T} be a set of tags, \mathcal{R} be a set of resources, and \mathcal{U} be a set of users.
- We denote a tag assignment of a tag $t \in \mathcal{T}$ to a resource $r \in \mathcal{R}$ as a tag-resource pair tr .
- $\mathcal{TR} = \{tr | t \in \mathcal{T}, r \in \mathcal{R}\}$ denotes all tag assignments.
- The function $getU(tr)$ retrieves the set of users who assigned t to r .
- Given the complete set of tag-resource pairs $\mathcal{TR} = \{tr_1, \dots, tr_n\}$, and associated users of each tag-resource pair $getU(tr_i) \subseteq \mathcal{U}$, our goal is to find a function $Q(tr_i)$ which assigns a score to each tag-resource pair tr_i .

Tagging System Model

Tagging Example

User u_1 annotates Resource r_1 with Tag t_1

Scenario



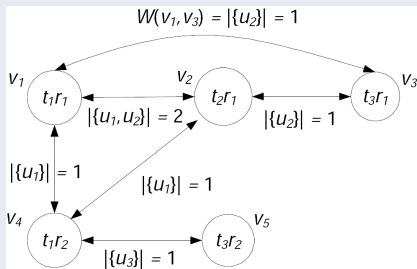
Tagging System Model

Tagging Example

Each tag-resource pair becomes a node.

The weight between the node is determined by the number of users who have annotated the same resource with the same tag

Data Model



Tagging System Model

Right Stochastic Transition Matrix

$$T(i, j) = \begin{cases} 0 & \text{if } (v_i, v_j) \notin \mathcal{E} \\ \frac{W(v_i, v_j)}{\sum_{v_k \in \mathcal{V}} W(v_i, v_k)} & \text{if } (v_i, v_j) \in \mathcal{E} \end{cases}$$

Adjacency and Transition Matrix

	v_1	v_2	v_3	v_4	v_5
v_1		2	1	1	
v_2	2		1	1	
v_3	1	1			
v_4	1	1			1
v_5				1	

$$T = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\ \frac{1}{2} & 0 & \frac{1}{4} & \frac{1}{4} & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Quality Propagation

PageRank

$$\text{p-rank}_{i+1} = \alpha \cdot T \cdot \text{p-rank}_i + (1 - \alpha) \cdot \frac{1}{N} \cdot 1_N. \quad (1)$$

TrustRank

$$\text{t-rank}_{i+1} = \alpha \cdot T \cdot \text{t-rank}_i + (1 - \alpha) \cdot d. \quad (2)$$

The set of seed sites is selected using an inverse PageRank algorithm. Particularly, nodes from where lots of other nodes can be reached are identified and ranked accordingly, similar to the idea of HITS. Then, the top-k nodes are manually assigned values 1, or 0 in case of a spam web site, and these initial values are stored in d .

Quality Propagation

PageRank

$$\text{p-rank}_{i+1} = \alpha \cdot T \cdot \text{p-rank}_i + (1 - \alpha) \cdot \frac{1}{N} \cdot 1_N. \quad (1)$$

TrustRank

$$\text{t-rank}_{i+1} = \alpha \cdot T \cdot \text{t-rank}_i + (1 - \alpha) \cdot d. \quad (2)$$

The set of seed sites is selected using an inverse PageRank algorithm. Particularly, nodes from where lots of other nodes can be reached are identified and ranked accordingly, similar to the idea of HITS. Then, the top-k nodes are manually assigned values 1, or 0 in case of a spam web site, and these initial values are stored in d .

Quality Propagation

Example Continued

$$\text{trp-rank}_{i+1} = 0.85 \cdot \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\ \frac{1}{2} & 0 & \frac{1}{4} & \frac{1}{4} & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \cdot \text{trp-rank}_i + (1-0.85) \cdot \begin{pmatrix} 0 \\ 0 \\ -1 \\ 1 \\ 0 \end{pmatrix}$$

Example Results

$i = 10$	v_1	v_2	v_3	v_4	v_5
trp-rank(10)	-0.03341879	-0.03341879	-0.16368952	0.180295	0.05023218

Quality Propagation

Example Continued

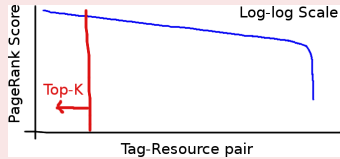
$$\text{trp-rank}_{i+1} = 0.85 \cdot \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\ \frac{1}{2} & 0 & \frac{1}{4} & \frac{1}{4} & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \cdot \text{trp-rank}_i + (1-0.85) \cdot \begin{pmatrix} 0 \\ 0 \\ -1 \\ 1 \\ 0 \end{pmatrix}$$

Example Results

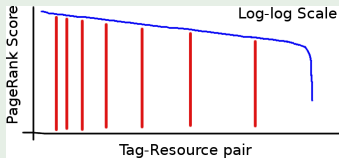
$i = 10$	v_1	v_2	v_3	v_4	v_5
trp-rank(10)	-0.03341879	-0.03341879	-0.16368952	0.180295	0.05023218

Seed Selection Strategies

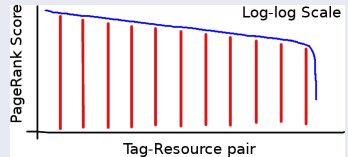
Top-k Seed Set



Exponential Base Seed Set



Constant Base Seed Set



- 1 Introduction
- 2 Related Work
- 3 Measuring Tag Quality
 - Tagging System Model
 - Quality Propagation
 - Seed Selection Strategies
- 4 **Evaluation**
 - Dataset
 - Method
 - Results
- 5 Conclusions & Future Work

Evaluation

Setup

- No corpus with annotated tag quality scores
- There exist corpora to identify spam users
- *Hypothesis*: Quality of tags assigned by spam users is poor
- *Problem*: Mapping spam scores of users to quality scores of tags

Data Provenance

- ECML PKDD Discovery Challenge 2008
- <http://www.kde.cs.uni-kassel.de/ws/rsdc08/>
- <http://www.bibsonomy.org>

Dataset

Some Figures

- 221,354 tag assignments by
- 1,328 users
- 118 spammer
- 1,210 non-spammer

Preprocessing

- Discarding *trs* from users with only one *tr*
- Stemming and ignoring of capital letters
e.g. “Book”, “books”, and “Books” get one ID

Method

Oracle Function

$$O(tr) = \begin{cases} 1 & \text{if } \frac{1}{|getU(tr)|} \sum_{u \in getU(tr)} \text{notSpammer}(u) > 0 \\ -1 & \text{if } \frac{1}{|getU(tr)|} \sum_{u \in getU(tr)} \text{notSpammer}(u) < 0 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Result Mapping

$$isSpammer(u) = \begin{cases} 1 & \text{if } \frac{1}{|getTR(u)|} \sum_{tr_i \in getTR(u)} Q(tr_i) < 0 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Results 1

Theoretical Maximum using complete Data as Seeds

Positive and Negative spread information

True Positives:	1210	True Negatives	89
False Positives:	29	False Neagatives	0

Only positive spread information

True Positives:	1079	True Negatives	114
False Positives:	4	False Neagatives	131

Only negative spread information

True Positives:	1210	True Negatives	91
False Positives:	27	False Neagatives	0

Results 2

Different Seed Set Selection Strategies

Strategy	Accuracy	
	Seed Set Size	
	10000	20000
Top-k	91.11 %	91.11 %
ExponentialBase	94.58 %	96.39 %
ConstantBase	94.88 %	96.31 %

Results 3

Different Seed Set Selection Strategies

Seed Set Size	TP	FP	TN	FN	Accuracy
132520	1210	29	89	0	97.82 %
50000	1210	29	89	0	97.82 %
20000	1210	49	69	0	96.31 %
10000	1210	68	50	0	94.88 %
5000	1210	87	31	0	93,45 %

Results 4

Data Reduction

- Minimum 10 Users → Accuracy 94.80 %
- Minimum 3 Users → Accuracy 95.63 %
- Minimum 1 Users → Accuracy 97.67 %

We observe that the performance drops by only 2.94 % when considering only tags that were used by at least 10 users (compared with the performance where $x = 1$), while the transition matrix size is reduced by more than 50%.

- 1 Introduction
- 2 Related Work
- 3 Measuring Tag Quality
 - Tagging System Model
 - Quality Propagation
 - Seed Selection Strategies
- 4 Evaluation
 - Dataset
 - Method
 - Results
- 5 Conclusions & Future Work

Conclusions & Future Work

Conclusions

- Measure for quality of tags
- Different seed set selection approaches
- Evaluation on manually labeled data set

Future Work

- More fine-grained seed values
- Automatic/user-generated assessment of seed scores
- Using a similar model to represent users instead of tag-resources to identify spammers

Conclusions & Future Work

Conclusions

- Measure for quality of tags
- Different seed set selection approaches
- Evaluation on manually labeled data set

Future Work

- More fine-grained seed values
- Automatic/user-generated assessment of seed scores
- Using a similar model to represent users instead of tag-resources to identify spammers