# Harnessing Folksonomies for Resource Classification

PhD Thesis

Arkaitz Zubiaga

UNED

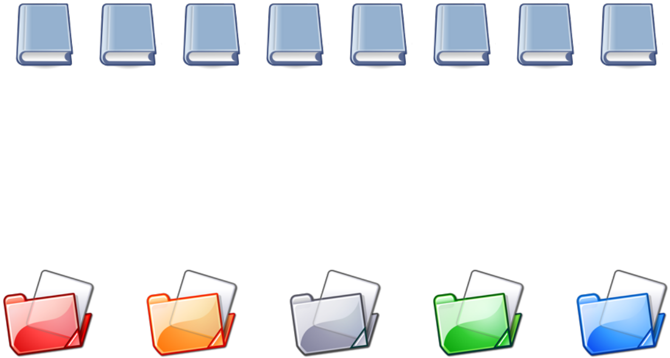
July 12th, 2011

Advisors:
Raquel Martínez Unanue
Víctor Fresno Fernández

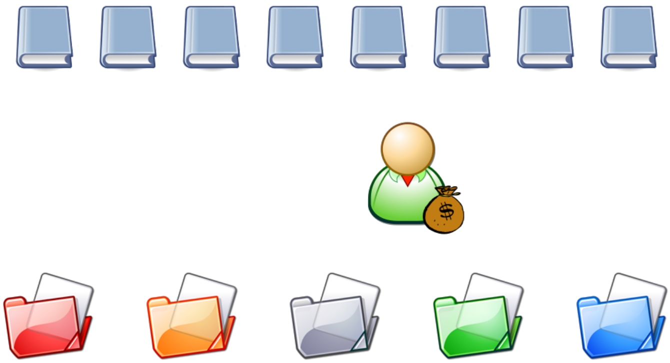
# Table of Contents

# Index

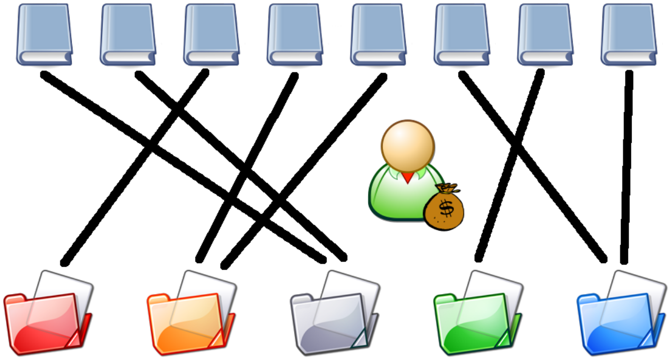
# Resource Classification



# Resource Classification



# Resource Classification



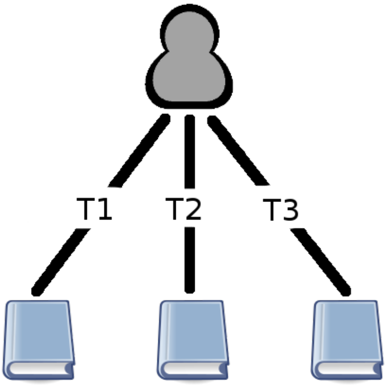
# Resource Classification

- **Classifying** resources is a **common task**.
  - Web pages, books, movies, files,...
- **Large collections** of resources $\rightarrow$ **expensive & effortful** to classify manually.
  - LoC reported an average **cost of $94.58 for cataloging each book** in 2002.
- Enormous costs and efforts $\rightarrow$ **automatic classification**.

# Resource Classification

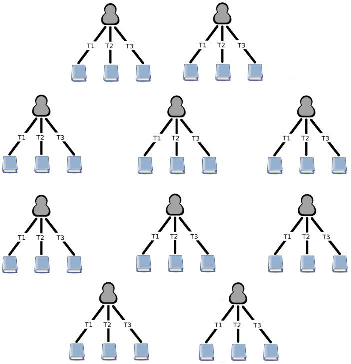
- **Representation** of resources → **self-content**.
- Use of **self-content** of resources presents some **issues**:
  - **Not** always **representative** enough.
  - **Not** always **accessible** (e.g., books).
- **Social tags** provided by users → **alternative** to solve the problem.

# Tagging




**T1**, **T2**, **T3** = sets of **tags**.

# Social Tagging




- **Aggregation** of user **annotations** → **folksonomy**.
- Folksonomy: **Folk** (People) + **Taxis** (Classification) + **Nomos** (Management).

# Organization of Resources

- User **annotations** → **own organization of resources**.

| **A user's tags** | |
|---|---|
| **Tag** | **# Resources** |
| `research` | 82 |
| `twitter` | 28 |
| `web2.0` | 35 |
| `language` | 42 |
| `english` | 64 |
| ... | ... |

# Example of Bookmarks

| | User | Resource | Tags |
|---|---|---|---|
| 1 | user1 | flickr.com | `photo`, `web2.0`, `social` |
| 2 | user2 | flickr.com | `photography`, `images` |
| 3 | user1 | google.com | `searchengine` |
| 4 | user3 | twitter.com | `microblogging`, `twitter` |

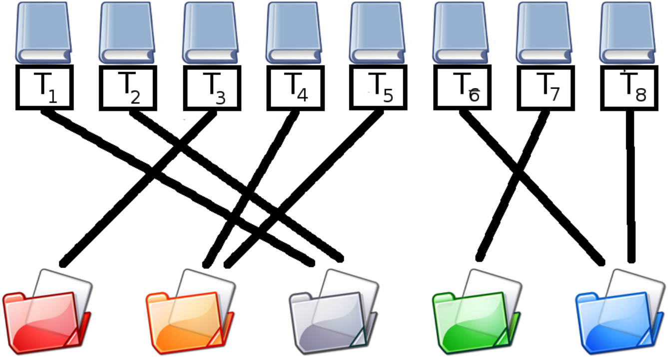
**Bookmark:** (1) **user** $u_i \in U$ who annotates
(2) **resource** $r_j \in R$ being annotated
(3) **tags** $T_{ij} = \{t_1, ..., t_n\} \in T$ utilized.

$$b_{ij} : u_i \times r_j \times T_{ij}$$

# Sum of Annotations

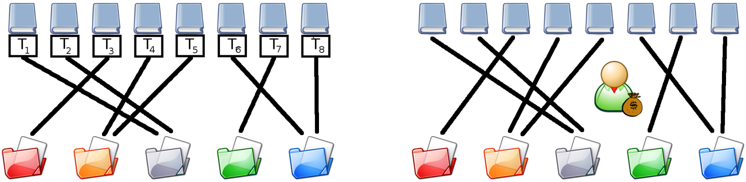
| Top tags (79,681 users) | | |
|---|---|---|
| Tag Rank | Tag | User Count |
| 1 | photos | 22,712 |
| 2 | flickr | 19,046 |
| 3 | photography | 15,968 |
| 4 | photo | 15,225 |
| 5 | sharing | 10,648 |
| 6 | images | 9,637 |
| 7 | web2.0 | 9,528 |
| 8 | community | 4,571 |
| 9 | social | 3,798 |
| 10 | pictures | 3,115 |

# Tag-based Resource Classification



# Problem Statement

How can the **annotations** provided by users on social tagging systems be **exploited** to **improve** the accuracy of a **resource classification** task?



# Related Work

**Social tags** for **information management**:

- **Search:** Bao et al. (2007) & Heymann et al. (2008).
- **Recommender Systems:** Shepitsen et al. (2008) & Li et al. (2008).
- **Enhanced Browsing:** Smith (2008).

# Related Work

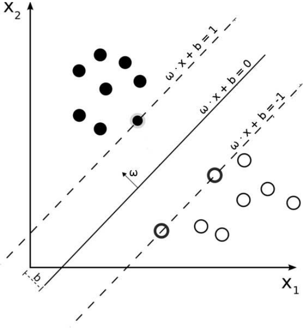
- **Classification:** Noll and Meinel (2008) → **statistical analysis** of matches between **tags** & **taxonomies**.
    - Tags are **useful** for **broad categorization**.
    - **Not** for **narrower categorization**.
- **Lack** of further **research** with:
    - **Actual classification** experiments.
    - **Other types** of **resources**.
    - **Different representations** of social tags.

# Index

# Characteristics of the task

- We have:
  - **Large set** of **resources**: some labeled + many unlabeled.
  - **Multiclass** taxonomy.
- Automated classifiers **learn a model** from **labeled resources**.
  - This **model** is used to **classify unlabeled resources** afterward.
- 2 learning settings:
  - **Supervised**: only labeled resources considered for learning.
  - **Semi-supervised**: unlabeled resources are also taken into account.

# Support Vector Machines (SVM)




- **Hyperplane** that separates with **largest margin**.
- Use of **kernels** → **redimensions** the space.
- Resource/Hyperplane **margin** → Classifier's **reliability**.

# Selection of a Classifier

- **SVMs** solve **binary** problems **by default**.
- To solve multiclass tasks:
  - **Native multiclass** classifier (**mSVM**).
  - **Combining binary** classifiers:
    - one-against-all (**oaaSVM**).
    - one-against-one (**oaoSVM**).
- Both **supervised (s)** and **semi-supervised (ss)**.

# Experiment Settings

- **3** benchmark **datasets** to analyze **suitability** of **classifiers**:

| Dataset | # web pages | # trainset | # categories |
|---|---|---|---|
| **BankSearch** | 10,000 | 3,000 | 10 |
| **WebKB** | 4,518 | 1,000 | 6 |
| **Y! Science** | 788 | 100 | 6 |

- We present **accuracy** to show performance.
- We perform **6 runs**, and show the **average** accuracy.

# Results

| | BankSearch | WebKB | Y! Science |
|---|---|---|---|
| **mSVM (s)** | **.925** | **.810** | .825 |
| **mSVM (ss)** | .923 | .778 | **.836** |
| **oaaSVM (s)** | .843 | .776 | .536 |
| **oaaSVM (ss)** | .842 | .773 | .565 |
| **oaoSVM (s)** | .826 | .775 | .483 |
| **oaoSVM (ss)** | .811 | .754 | .514 |

- **Native multiclass** classifier **performs best**, while **supervised** $\simeq$ **semi-supervised**.
- We used the **supervised** approach, as it is **computationally less expensive**.

# Index

# Requirements

- **Selected STS should** have:
  - **Large communities** involved.
  - Public **access to data**.
  - **Consolidated taxonomies** as a ground truth.
- We chose **Delicious**, **LibraryThing** & **GoodReads**.

# Characteristics of STS

| | Delicious | LibraryThing | GoodReads |
|---|---|---|---|
| **Resources** | web documents | books | books |
| **Tag suggestions** | based on earlier bookmarks on the resource | no | based on earlier tags utilized by the user |
| **Tag insertion** | space-separated | comma-separated | one by one text-box |
| **Saving a resource** | prompts user to add tags | prompts user to add tags at second step | user needs to click again to add tags |

# Characteristics of STS

| | Delicious | LibraryThing | GoodReads |
|---|---|---|---|
| **Resources** | web documents | books | books |
| **Tag suggestions** | based on earlier bookmarks on the resource | no | based on earlier tags utilized by the user |
| **Tag insertion** | space-separated | comma-separated | one by one text-box |
| **Saving a resource** | prompts user to add tags | prompts user to add tags at second step | user needs to click again to add tags |

# Characteristics of STS

| | Delicious | LibraryThing | GoodReads |
|---|---|---|---|
| **Resources** | web documents | books | books |
| **Tag suggestions** | based on earlier bookmarks on the resource | no | based on earlier tags utilized by the user |
| **Tag insertion** | space-separated | comma-separated | one by one text-box |
| **Saving a resource** | prompts user to add tags | prompts user to add tags at second step | user needs to click again to add tags |

# Characteristics of STS

| | Delicious | LibraryThing | GoodReads |
|---|---|---|---|
| **Resources** | web documents | books | books |
| **Tag suggestions** | based on earlier bookmarks on the resource | no | based on earlier tags utilized by the user |
| **Tag insertion** | space-separated | comma-separated | one by one text-box |
| **Saving a resource** | prompts user to add tags | prompts user to add tags at second step | user needs to click again to add tags |

# Characteristics of STS

| | Delicious | LibraryThing | GoodReads |
|---|---|---|---|
| **Resources** | web documents | books | books |
| **Tag suggestions** | based on earlier bookmarks on the resource | no | based on earlier tags utilized by the user |
| **Tag insertion** | space-separated | comma-separated | one by one text-box |
| **Saving a resource** | prompts user to add tags | prompts user to add tags at second step | user needs to click again to add tags |

# Retrieval of Categorized Resources

Retrieval of **popular annotated resources**, which were also **categorized** by experts.

| | | **Top level (L1)** | | **Second level (L2)** | |
|---|---|---|---|---|---|
| | | **Resources** | **Classes** | **Resources** | **Classes** |
| **Web** | **ODP** | 12,616 | 17 | 12,286 | 243 |
| **Books** | **DDC** | 27,299 | 10 | 27,040 | 99 |
| | **LCC** | 24,861 | 20 | 23,565 | 204 |

# Retrieval of Additional User Annotations

- **Delicious:** 300,571,231 bookmarks
  $\rightarrow$ 273,478,137 annotated (**91.00%**)

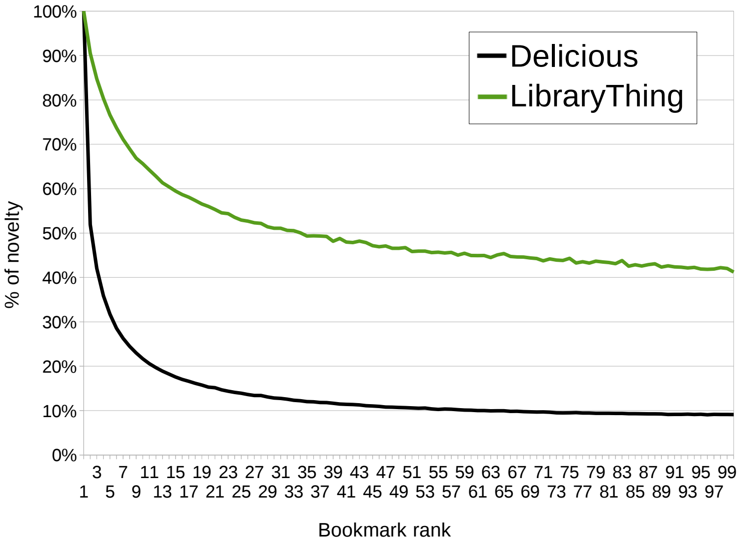
- **LibraryThing:** 44,612,784 bookmarks
  $\rightarrow$ 22,343,427 annotated (**50.08%**)

- **GoodReads:** 47,302,861 bookmarks
  $\rightarrow$ 9,323,539 annotated (**19.71%**)

**Importance** of system's **encouragement** to **tagging resources**.

# Tag Popularity on Resources

# Tag Novelty in Bookmarks by Rank



# Retrieval of Additional Data

- **URLs:**
  - **Self-content**, by crawling URLs.
  - **User reviews** (Delicious & StumbleUpon).
- **Books:**
  - ~~**Self-content (unavailable)**~~:
    - **Synopses** (Barnes&Noble).
    - **Editorial reviews** (Amazon).
  - **User reviews** (LibraryThing, GoodReads & Amazon).

# Summary of the Analysis of Datasets

- **Few users annotate** resources when the **system does not encourage** to do it.
- Resource-based **tag suggestions** → **Repeated** use of **popular tags**.

# Index

# Representing Resources Using Tags

- Different ways to **aggregate** user **annotations** on a **vectorial representation**.
- **2 major factors** to consider:
  - **What tags** to use?
  - **How to weigh** those tags?

# Representing Resources Using Tags

- Use of **all tags** (**FTA**), or just **top 10** tags for each resource.
- **4** different **weightings**.
- **Example** of a resource (**100 users**):
  $t_1$ (50), $t_2$ (30), $t_3$ (20), ..., $t_9$ (1), $t_{10}$ (1), ..., $t_n$ (1)

| | **FTA** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **Top 10** | | | | | | | |
| | $t_1$ | $t_2$ | $t_3$ | **...** | $t_9$ | $t_{10}$ | **...** | $t_n$ |
| **Ranks** | 1 | 0.9 | 0.8 | ... | 0.2 | 0.1 | ... | 0 |
| **Fractions** | 0.5 | 0.3 | 0.2 | ... | 0.02 | 0.01 | ... | 0.01 |
| **Binary** | 1 | 1 | 1 | ... | 1 | 1 | ... | 1 |
| **TF** | 50 | 30 | 20 | ... | 2 | 1 | ... | 1 |

# Representing Resources Using Other Data Sources

- To **represent** resources using **content** and **reviews**:
  1. **Removal** of **HTML** tags.
  2. **Removal** of **stopwords**.
  3. **Stem** of remaining words.
  4. **TF-IDF** weighting of words.

# Experiment Setup

- **Multiclass SVMs**.
- Show the **average accuracy** of **6 runs**.
- For clarity of presentation, we limit results to:
  - **LCC** taxonomy for books.
  - Training sets of **6,000 URLs** (6,616 (L1)/6,286 (L2) for test).
  - Training sets of **18,000 books** (8,861 (L1)/5,565 (L2) for test).

# Experiment Setup

## Compared Representations

- **Self-content** (baseline).
- **Reviews**.
- **Tags**:
  - **Ranks** (Top **10**).
  - **Fractions** (Top **10** & **FTA**).
  - **Binary** (Top **10** & **FTA**).
  - **TF** (Top **10** & **FTA**).

# Results of Tags vs Other Data Sources

| | | Delicious | | LThing | | GReads | |
|---|---|---|---|---|---|---|---|
| | | **L1 (17)** | **L2 (243)** | **L1 (20)** | **L2 (204)** | **L1 (20)** | **L2 (204)** |
| **Content** | | .610 | .470 | .807 | .673 | .807 | .673 |
| **Reviews** | | .646 | .524 | .828 | .705 | **.828** | **.705** |
| **Tags** | **Ranks** | .484 | .360 | .795 | .511 | .630 | .405 |
| | **Fractions (10)** | .464 | .349 | .738 | .411 | .663 | .427 |
| | **Fractions (FTA)** | .461 | .336 | .712 | .409 | .654 | .432 |
| | **Binary (10)** | .531 | .361 | .770 | .550 | .623 | .422 |
| | **Binary (FTA)** | .572 | .529 | .655 | .606 | .639 | .481 |
| | **TF (10)** | .654 | .545 | .855 | .722 | .713 | .491 |
| | **TF (FTA)** | **.680** | **.568** | **.857** | **.736** | .731 | .517 |

- Usually, **FTA** > **10**.

# Results of Tags vs Other Data Sources

| | | Delicious | | LThing | | GReads | |
|---|---|---|---|---|---|---|---|
| | | **L1 (17)** | **L2 (243)** | **L1 (20)** | **L2 (204)** | **L1 (20)** | **L2 (204)** |
| **Content** | | .610 | .470 | .807 | .673 | .807 | .673 |
| **Reviews** | | .646 | .524 | .828 | .705 | **.828** | **.705** |
| **Tags** | **Ranks** | .484 | .360 | .795 | .511 | .630 | .405 |
| | **Fractions (10)** | .464 | .349 | .738 | .411 | .663 | .427 |
| | **Fractions (FTA)** | .461 | .336 | .712 | .409 | .654 | .432 |
| | **Binary (10)** | .531 | .361 | .770 | .550 | .623 | .422 |
| | **Binary (FTA)** | .572 | .529 | .655 | .606 | .639 | .481 |
| | **TF (10)** | .654 | .545 | .855 | .722 | .713 | .491 |
| | **TF (FTA)** | **.680** | **.568** | **.857** | **.736** | .731 | .517 |

- **TF (FTA)** is the **best approach** for **tags**.

# Results of Tags vs Other Data Sources

| | Delicious | | LThing | | GReads | |
|---|---|---|---|---|---|---|
| | **L1 (17)** | **L2 (243)** | **L1 (20)** | **L2 (204)** | **L1 (20)** | **L2 (204)** |
| **Content** | .610 | .470 | .807 | .673 | .807 | .673 |
| **Reviews** | .646 | .524 | .828 | .705 | **.828** | **.705** |
| **Tags** | **.680** | **.568** | **.857** | **.736** | .731 | .517 |

- **Tags** clearly **outperform** content and reviews on **Delicious** and **LibraryThing**.

# Results of Tags vs Other Data Sources

| | Delicious | | LThing | | GReads | |
|---|---|---|---|---|---|---|
| | **L1 (17)** | **L2 (243)** | **L1 (20)** | **L2 (204)** | **L1 (20)** | **L2 (204)** |
| **Content** | .610 | .470 | .807 | .673 | .807 | .673 |
| **Reviews** | .646 | .524 | .828 | .705 | **.828** | **.705** |
| **Tags** | **.680** | **.568** | **.857** | **.736** | .731 | .517 |

- **GoodReads' disencouragement to tagging** makes it **insufficient** to outperform content and reviews.

# Results of Tags vs Other Data Sources

| | Delicious | | LThing | | GReads | |
|---|---|---|---|---|---|---|
| | **L1 (17)** | **L2 (243)** | **L1 (20)** | **L2 (204)** | **L1 (20)** | **L2 (204)** |
| **Content** | .610 | .470 | .807 | .673 | .807 | .673 |
| **Reviews** | .646 | .524 | .828 | .705 | **.828** | **.705** |
| **Tags** | **.680** | **.568** | **.857** | **.736** | .731 | .517 |

- **Tags** are also **useful** for **deeper categorization** (L2).

# Classifier Committees

- **Despite** the **superiority** of social **tags**, **all** data sources **perform well**.
- Their outputs can be **combined** by using **classifier committees**.
- Classifier committees **add up margins** (i.e., reliability values) outputted by several classifiers, and provide a single combined prediction.

| | **Cat. #1** | **Cat. #2** | **Cat. #3** |
|---|---|---|---|
| **Classif. A** | **1.2** | 1.1 | 0.6 |
| **Classif. B** | 0.5 | 1.0 | **1.2** |
| **Classif. committees** | 1.7 | **2.1** | 1.8 |

# Results of Classifier Committees

| | | Delicious | | LThing | | GReads | |
|---|---|---|---|---|---|---|---|
| | | **L1 (17)** | **L2 (243)** | **L1 (20)** | **L2 (204)** | **L1 (20)** | **L2 (204)** |
| **Content (C)** | | .610 | .470 | .807 | .673 | .807 | .673 |
| **Reviews (R)** | | .646 | .524 | .828 | .705 | .828 | .705 |
| **Tags (T)** | | .680 | .568 | .857 | .736 | .731 | .517 |
| **Commit.** | **C + R** | .670 | .547 | .817 | .704 | .817 | .704 |
| | **C + T** | .696 | .587 | .821 | .720 | .832 | .696 |
| | **R + T** | .694 | .584 | **.859** | **.755** | **.857** | **.730** |
| | **C + R + T** | **.699** | **.588** | .827 | .732 | .843 | .727 |

- Classifier **committees successfully improve** performance.
  - **Even on GoodReads**, where tags were not good enough on their own.

# Results of Classifier Committees

| | | Delicious | | LThing | | GReads | |
|---|---|---|---|---|---|---|---|
| | | **L1 (17)** | **L2 (243)** | **L1 (20)** | **L2 (204)** | **L1 (20)** | **L2 (204)** |
| **Content (C)** | | .610 | .470 | .807 | .673 | .807 | .673 |
| **Reviews (R)** | | .646 | .524 | .828 | .705 | .828 | .705 |
| **Tags (T)** | | .680 | .568 | .857 | .736 | .731 | .517 |
| **Commit.** | **C + R** | .670 | .547 | .817 | .704 | .817 | .704 |
| | **C + T** | .696 | .587 | .821 | .720 | .832 | .696 |
| | **R + T** | .694 | .584 | **.859** | **.755** | **.857** | **.730** |
| | **C + R + T** | **.699** | **.588** | .827 | .732 | .843 | .727 |

- Data sources must be chosen with care:
  - **All** 3 are **helpful** on **Delicious**.
  - **Content** is **harmful** for **books**. Inappropriate considering synopses and ed. reviews as a summary of content?

# Summary of Results

- Better represent using **all tags** with **TF weighting**.
- **Tags** perform **accurately** even for **deeper levels**.
  - The **system** must **encourage** the user to tag **to make it useful** enough.
- **Tags** can be **combined** with other data to **improve performance**.
  - Combined **data sources** must be **chosen** with **care**.

# Index

# Tag Distributions

- So far, we have considered that **tags annotated by the same number of users** are **equally representative** to the resource.
- **Distributions of tags** in a collection **could help determine representativity** of tags.

# TF-IDF

**TF-IDF** is an **inverse weighting function** (**IWF**) that computes:

- the **term frequency** (TF).
- the **inverse document frequency** (IDF).

$$tf\text{-}idf_{ij} = tf_{ij} \times \log \frac{|D|}{|\{d : t_i \in d\}|}$$

- **High IDF** value for **terms** appearing in **few documents**.
- **Low IDF** value for **terms** appearing in **many documents**.

# Tag Weighting Functions

- **Analogous** to TF-IDF on **folksonomies**:
  - **TF-IRF** $\rightarrow$ distributions across **resources**.
  - **TF-IUF** $\rightarrow$ distributions across **users**.
  - **TF-IBF** $\rightarrow$ distributions across **bookmarks**.
- **TF-IRF** and **TF-IUF** had been **barely used**, and their **suitability** was yet **unexplored**.
- **TF-IBF** had not been used.

# Results Using IWFs

| | | Delicious | | LThing | | GReads | |
|---|---|---|---|---|---|---|---|
| | | **L1** | **L2** | **L1** | **L2** | **L1** | **L2** |
| **TF** | | **.680** | **.568** | .857 | .736 | .731 | .517 |
| **IWFs** | **TF-IRF** | .639 | .529 | .894 | .809 | .799 | .622 |
| | **TF-IBF** | .641 | .532 | **.895** | **.811** | **.800** | **.628** |
| | **TF-IUF** | .661 | .555 | .892 | .803 | .794 | .623 |

- **All 3 IWFs** clearly **outperform TF** for **LibraryThing** and **GoodReads**.
  - **Similar** performance of **IWFs**.

# Results Using IWFs

| | | Delicious | | LThing | | GReads | |
|---|---|---|---|---|---|---|---|
| | | **L1** | **L2** | **L1** | **L2** | **L1** | **L2** |
| **TF** | | **.680** | **.568** | .857 | .736 | .731 | .517 |
| **IWFs** | **TF-IRF** | .639 | .529 | .894 | .809 | .799 | .622 |
| | **TF-IBF** | .641 | .532 | **.895** | **.811** | **.800** | **.628** |
| | **TF-IUF** | .661 | .555 | .892 | .803 | .794 | .623 |

- **IWFs underperform** on **Delicious**, due to **tag suggestions** that make top tags utmost popular.
  - **IUF superior to IBF and IRF**. **Users** who make their **own choices** make the difference.

# Using Classifier Committees with IWFs

How about using **tags** represented with **IWFs** on **classifier committees**?

# Results Using IWF with Committees

| | | Delicious | | LThing | | GReads | |
|---|---|---|---|---|---|---|---|
| | | **L1** | **L2** | **L1** | **L2** | **L1** | **L2** |
| **TF** | | .699 | .588 | .859 | .755 | .857 | .730 |
| **IWFs** | **TF-IRF** | .697 | .592 | .885 | .793 | .864 | .748 |
| | **TF-IBF** | .698 | .592 | **.887** | **.797** | **.866** | **.751** |
| | **TF-IUF** | **.700** | **.595** | .885 | .792 | .864 | .749 |

- **IWF-based committes** are even **better than TF-based ones**.
  - Even on **Delicious**, where **IWFs were not appropriate**, **committees** perform slightly **better**.

# Results Using IWF with Committees

| | | Delicious | | LThing | | GReads | |
|---|---|---|---|---|---|---|---|
| | | **L1** | **L2** | **L1** | **L2** | **L1** | **L2** |
| **TF** | | .699 | .588 | .859 | .755 | .857 | .730 |
| **IWFs** | **TF-IRF** | .697 | .592 | .885 | .793 | .864 | .748 |
| | **TF-IBF** | .698 | .592 | **.887** | **.797** | **.866** | **.751** |
| | **TF-IUF** | **.700** | **.595** | .885 | .792 | .864 | .749 |

- Despite this outperformance of IWFs using committees, **IWFs on their own perform better on LibraryThing** (**.895 & .811**).

# Summary of Results Using IWFs

- **IWFs** are an **appropriate** way to weight tags **when used on classifier committees**.
  - The **exception** is **LibraryThing**, where **tags on their own perform better**.
- Combined **data** sources **must be appropriately chosen** (e.g., synopses & ed. reviews are harmful with books).

# Index

# User Behavior: Categorizers and Describers

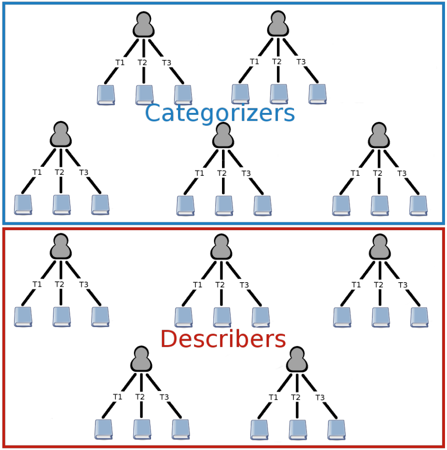
- Körner[1] suggested **2 kinds of user behavior**:

| | **Categorizer** | **Describer** |
|---|---|---|
| Goal of Tagging | later browsing | later retrieval |
| Change of Tag Vocabulary | costly | cheap |
| Size of Tag Vocabulary | limited | open |
| Tags | subjective | objective |

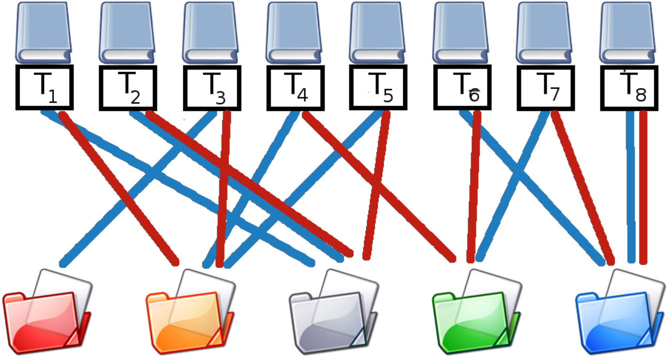
- They found that **Describers** help infer **semantic relations** among tags.
- Do these **tagging behaviors** affect the usefulness of tags for **resource classification**?

[1] C. Körner. Understanding the Motivation behind Tagging. Hypertext 2009.

# Categorizers and Describers



# Categorizers and Describers



# Weighting Measures

We use **3 measures to weight users**, based on Koerner et al. (2010).

- 2 factors are considered: **verbosity** & **diversity**.

# Weighting Measures: TPP

- **Tags per Post** (TPP) – **Verbosity**

$$TPP(u) = \frac{\sum^{r} |T_{ur}|}{|R_u|}$$

# Weighting Measures: ORPHAN

- **Orphan Ratio** (ORPHAN) – **Diversity**

$$n = \left\lceil \frac{|R(t_{max})|}{100} \right\rceil$$

$$ORPHAN(u) = \frac{|T_u^o|}{|T_u|}, T_u^o = \{t||R(t)| \leq n\}$$

# Weighting Measures: TRR

- **Tag Resource Ratio** (TRR) – **Verbosity + Diversity**

$$TRR(u) = \frac{|T_u|}{|R_u|}$$

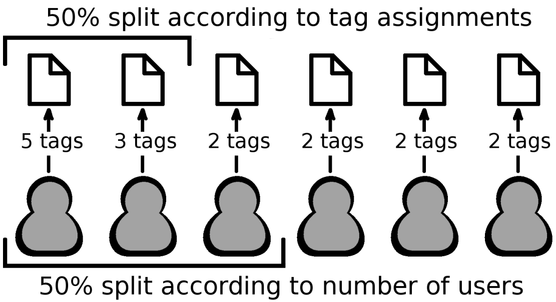
# Use of Weighting Measures

- These **3 measures** provide:
  - A **weight** for each user.
  - **Ranking of users** according to each measure.
- From rankings → subsets of users as **extreme Categorizers** (highest-ranked) and **extreme Describers** (lowest-ranked).
- **Subsets** range from **10% to 100%** (step size = 10%).

# Experiment Setup

- We select **subsets of users** according to **number of tag assignments**.

- Selecting by **percents of users** would be **unfair** → different **amounts of data**.



# Experiments

## Classification

- We use a **multiclass SVM**, with **TF** weighting of tags.

## Descriptivity

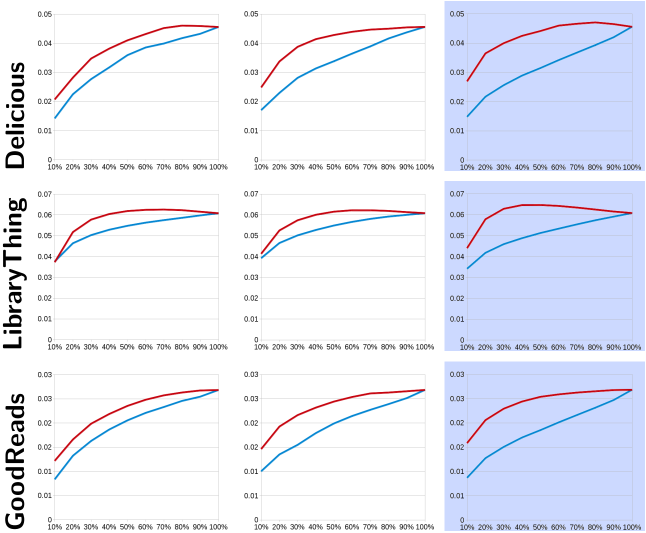
- **Vectorial** representations of **resources**:
  - $T_r \rightarrow$ **tag frequencies**.
  - $R_r \rightarrow$ **term frequencies on descriptive data** (self-content).

- **Cosine similarity** between $T_r$ and $R_r$:

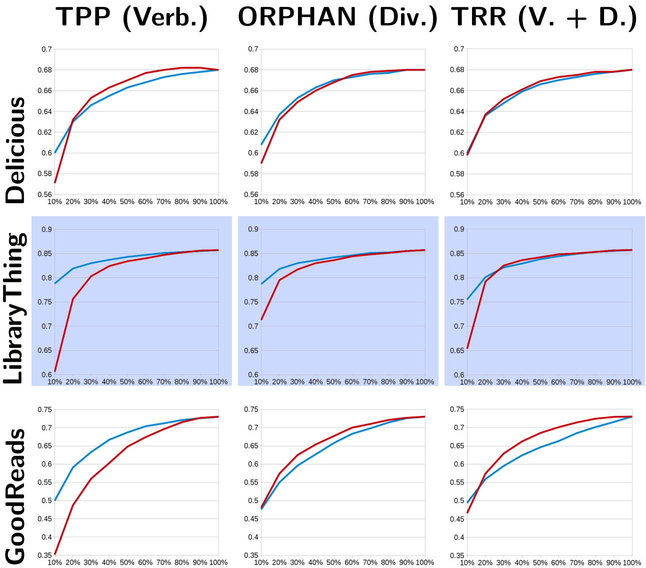
$$\cos(\theta_r) = \sum_{i=1}^{n} \frac{T_{ri} \times R_{ri}}{\sqrt{\sum_{i=1}^{n} (T_{ri})^2} \times \sqrt{\sum_{i=1}^{n} (R_{ri})^2}}$$

# Descriptivity Results

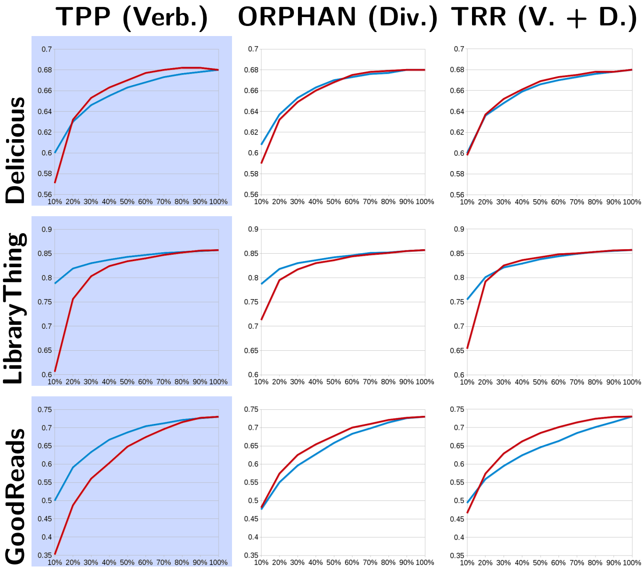
**TPP (Verb.)** **ORPHAN (Div.)** **TRR (V. + D.)**



# Classification Results



# Classification Results



# Overall Categorizers/Describers Results

- Discriminating by **verbosity (TPP)** does **best for finding extreme Categorizers**.
- The use of **non-descriptive tags** provide more **accurate classification**.

PhD Thesis

Arkaitz Zubiaga

Motivation

Selection of a Classifier

STS & Datasets

Representing the Aggregation of Tags

Tag Distributions on STS

User Behavior on STS

Conclusions & Outlook

Publications

# Index

# Contributions

- Generation & analysis of **3 large-scale social tagging datasets**.
- **Release** of some **tagging datasets**, used by Godoy and Amandi (2010), Strohmaier et al. (2010), Li et al. (2011), and Ares et al. (2011).

# Contributions

- **First research work** performing **actual classification** experiments using social tags.
  - Analysis of **different representations** of social tags.
  - Analysis of effect of **tag distributions**.
  - Study of **user behavior**.
- It **paves the way** to future **researchers** interested in the **task** & in the exploration of **STS**.

# Research Questions

- Apart from the **Problem Statement**:
  - How can the **annotations** provided by users on social tagging systems be **exploited** to **improve** the accuracy of a **resource classification** task?
- We set forth **10 research questions**.

# Research Questions (1)

**RQ 1** What is a **suitable SVM** classifier for the **task**?

- Native **multiclass** SVM $>>$ Combinations of **binary** SVMs.

# Research Questions (2)

**RQ 2** What is a **suitable learning method** for the **task**?

- **Supervised** $\simeq$ **Semi-supervised**.
- **Unlike** for **binary** tasks, where Semi-supervised >> Supervised (Joachims, 1999).

# Research Questions (3)

**RQ 3** **How** do the **settings** of **STS affect folksonomies**?

- Great **impact** of **tag suggestions**.
- **Importance** of **encouraging** users to **annotate**.

# Research Questions (4)

**RQ 4** How to **amalgamate annotations** to get a **representation** of a resource?

- Considering **all the tags rather than** only **those in the top**.
- **Weighting tags** according to **number of users annotating** them.

# Research Questions (5)

**RQ 5** Is it **worthwhile combining tags with other data sources**?

- **Combining** different **data sources** helps **improve performance**.
- **Data sources** must be **appropriately chosen**.

# Research Questions (6)

**RQ 6** Are **social tags specific** enough to **classify into narrower categories**?

- **Tags** are **as useful as** for **top level**.
- **Noll and Meinel** (2008) → **tags** were probably **not useful for deeper levels**.

# Research Questions (7)

**RQ 7** Can we consider **tag distributions** to get the **representativity of each tag**?

- **LibraryThing & GoodReads**: really **useful**.
- **Delicious**: **not useful**, because of **tag suggestions**
  $\rightarrow$ need of **committees** to **make them useful**.

# Research Questions (8)

**RQ 8** What **approach** to use **to weigh the representativity of tags**?

- **LibraryThing & GoodReads**: **IBF**, **IRF** & **IUF** are very **similar**.
- **Delicious**: **IUF** clearly **superior**, because of **users that get rid of suggestions**.

# Research Questions (9)

**RQ 9** Can we discriminate **users** who further **resemble an expert classification**?

- **Categorizers** > **Describers** for **classification**.
- **Need** of **appropriate measure** for **discriminating**.

# Research Questions (10)

**RQ 10** What **features identify** a **Categorizer**?

- **Categorizers** can be **found** when **discriminating by verbosity**.
- **Non-descriptive tags** produce **more accurate classification**.

# Future Directions

- **Increase of interest** in the **field**, still **much work** to do.
- We have considered **each tag** as a **diferent token**.
  → Considering **semantic meanings** of social tags could help.
- **Tag suggestions** leverage several **issues** in **folksonomies**.
  → Looking for a **weighting function** that fits the characteristics of systems with **tag suggestions**, e.g., Delicious.

# Index

# Publications

## Peer-Reviewed Conferences (I)

- Arkaitz Zubiaga, Christian Körner, Markus Strohmaier. 2011. *Tags vs Shelves: From Social Tagging to Social Classification*. In Proceedings of **Hypertext 2011**, the 22nd ACM Conference on Hypertext and Hypermedia, Eindhoven, Netherlands. (acceptance rate: 35/104, 34%)

- Arkaitz Zubiaga, Raquel Martínez, Víctor Fresno. 2009. *Getting the Most Out of Social Annotations for Web Page Classification*. In Proceedings of **DocEng 2009**, the 9th ACM Symposium on Document Engineering, pp. 74-83, Munich, Germany. (acceptance rate: 16/54, 29.6%) [15 citations]

# Publications

## Peer-Reviewed Conferences (II)

- Arkaitz Zubiaga. 2009. *Enhancing Navigation on Wikipedia with Social Tags*. **Wikimania 2009**, Buenos Aires, Argentina.
  [6 citations]

- Arkaitz Zubiaga, Alberto P. García-Plaza, Víctor Fresno, Raquel Martínez. 2009. *Content-based Clustering for Tag Cloud Visualization*. In Proceedings of **ASONAM 2009**, International Conference on Advances in Social Networks Analysis and Mining, pp. 316-319, Athens, Greece.
  [3 citations]

# Publications

## Journals (I)

- Arkaitz Zubiaga, Raquel Martínez, Víctor Fresno. 2011. *Augmenting Web Page Classifiers with Social Annotations*. **Procesamiento del Lenguaje Natural**. (acceptance rate: 33/60, 55%)

- Arkaitz Zubiaga, Raquel Martínez, Víctor Fresno. 2009. *Clasificación de Páginas Web con Anotaciones Sociales*. **Procesamiento del Lenguaje Natural**, vol. 43, pp. 225-233. (acceptance rate: 36/72, 50%)

- Arkaitz Zubiaga, Víctor Fresno, Raquel Martínez. 2009. *Comparativa de Aproximaciones a SVM Semisupervisado Multiclase para Clasificación de Páginas Web*. **Procesamiento del Lenguaje Natural**, vol. 42, pp. 63-70.

# Publications

**Journals (II)**

- Arkaitz Zubiaga, Víctor Fresno, Raquel Martínez. *Harnessing Folksonomies to Produce a Social Classification of Resources*. **IEEE Transactions on Knowledge and Data Engineering**. (**pending notification**)

# Publications

## Book Chapters

- Arkaitz Zubiaga, Víctor Fresno, Raquel Martínez. 2011. *Exploiting Social Annotations for Resource Classification*. **Social Network Mining, Analysis and Research Trends: Techniques and Applications.** IGI Global.

## Workshops

- Arkaitz Zubiaga, Víctor Fresno, Raquel Martínez. 2009. *Is Unlabeled Data Suitable for Multiclass SVM-based Web Page Classification?*. In Proceedings of the **NAACL-HLT 2009 Workshop on Semi-supervised Learning for Natural Language Processing**, pp. 28-36, Boulder, CO, United States.

# Thank You

Achiu Arigato Danke Dhannvaad Dua Netjer en ek Efcharisto Gracias Gràcies Gratia Grazie Guishepeli Hvala Kiitos Köszönöm Mercé Merci Mila esker Obrigado Shukran Shukriya Tack Tak Takk Tänan Tapadh leat Tesekkür ederim Thank you Toda



http://thesis.zubiaga.org/