A k-Nearest-Neighbour Method for

Classifying Web Search Results

with Data in Folksonomies

by Ching-man Au Yeung, Nicholas Gibbins, Nigel Shadbolt

Intelligence, Agents, Multimedia Group
School of Electronics and Computer Science
University of Southampton
Introduction

The Problem of Ambiguity

- Queries by ambiguous terms return many irrelevant results
- Example: *bridge*
  1) a kind of card games;
  2) a form of architectural structure;
  3) a design pattern in software development;
  4) a device in computer networking
Introduction

Delicious

BibSonomy

WI’08 - Classifying Web Search Results with Data from Folksonomies
Collaborative Tagging Systems

- Aggregate user-contributed metadata of Web resources
- Provide rich information about the relations between different tags
- Sources for understanding how keywords are used on the Web
Multiple Meanings of Tags

- Tags have multiple meanings, or they are used in different contexts
- It is possible to extract related tags in different contexts
- E.g. sf:
  
  (california, bayarea, travel, ...) ➔ San Francisco
  (scifi, fantasy, fiction, ...) ➔ Science Fiction
Building Classifiers

Our Proposal

Building classifiers using data in folksonomies:

Wikipedia page of San Francisco
The City and County of San Francisco is the fourth most populous city in California and the 14th most populous city in the United States... Among the most densely populated cities in the country, San Francisco is part of the San Francisco Bay Area... The city is located at the tip of the San Francisco Peninsula, with the Pacific Ocean to the west, ...

Wikipedia page of Science Fiction
Science fiction (abbreviated SF or sci-fi with varying punctuation and capitalization) is a broad genre of fiction that often involves speculations based on current or future science or technology. Science fiction is found in books, art, television, films, games, theatre, and other media... this includes fantasy, horror, and related genres.
Building Classifiers

Clustering of Folksonomy Networks

- Construct a document network from a folksonomy
- Cluster documents based on the users who have used the tag on the documents
  (the community-discovery algorithm described in [Newman 2004] is used in this paper)
- Extract frequently co-occurred tags as representations of the different classes (meanings)
Clustering of Folksonomy Networks
Clustering of Folksonomy Network
## Clustering of Folksonomy Nework

<table>
<thead>
<tr>
<th>Category</th>
<th>Relevant Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design pattern</td>
<td>bridge, programming, development, library, code, ruby, tools, software, adobe, dev</td>
</tr>
<tr>
<td>Card game</td>
<td>bridge, games, cards, game, imported, howto, conventions, card, bidding, online</td>
</tr>
<tr>
<td>Computer networking</td>
<td>bridge, networking, linux, network, howto, software, sysadmin, firewall, virtualization, security</td>
</tr>
<tr>
<td>Architecture</td>
<td>bridge, bridges, structures, engineering, science, physics, school, education, building, reference</td>
</tr>
</tbody>
</table>
Web Search Result Classification

1. Submit query term
2. Get documents associated with the query term
3. Get tags
4. Construct sets of keywords
5. Search Result Classification
6. Output Result

**Google**

Example: sf

**Delicious**

Example:
- D1: Wikipedia - Science Fiction
- D2: Wikipedia - San Francisco
- D3: ...

Example:
- C1: {scifi, fiction, books, ...}
- C2: {california, bayarea, ...}

Example:
- D1 \(\rightarrow\) C1
- D2 \(\rightarrow\) C2
Web Search Result Classification

**k-Nearest-Neighbour Classifier**

A document returned by the search engine

Labelled Documents from Delicious

Similarity Calculation

Example:
- D9 (C1): 0.432
- D5 (C1): 0.401
- D1 (C1): 0.310
- D7 (C2): 0.230

Ranked List of Documents

Example:
- D1 \(\rightarrow\) C1
- D2 \(\rightarrow\) C2
- D3 \(\rightarrow\) C1

Assign document to class according to the majority vote and a threshold \(\beta\)
Evaluation

Data Preparation

♦ Ten tags which are used in multiple contexts in Delicious are chosen

♦ Documents associated with the tags as well as the users who assigned the tags are retrieved

♦ Testing dataset obtained by submitting query to Google and obtaining the top 50 pages returned
### Classifiers Built

<table>
<thead>
<tr>
<th>Tag</th>
<th>Context</th>
<th>Class Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>sf</td>
<td>San Francisco</td>
<td>sanfrancisco, bayarea, san, francisco, california, travel, events</td>
</tr>
<tr>
<td></td>
<td>Science fiction</td>
<td>scifi, fiction, books, sci-fi, literature, writing, science, fantasy</td>
</tr>
<tr>
<td>soap</td>
<td>Cleaning agent</td>
<td>soapmaking, diy, recipes, crafts, shopping, making, howto</td>
</tr>
<tr>
<td></td>
<td>Web services</td>
<td>webservices, webservice, programming, web, xml, soa, java</td>
</tr>
<tr>
<td>wine</td>
<td>Software application</td>
<td>linux, ubuntu, howto, windows, software, tutorial, emulation</td>
</tr>
<tr>
<td></td>
<td>Beverage</td>
<td>food, shopping, drink, vino, cooking, alcohol, blog, news</td>
</tr>
<tr>
<td>xp</td>
<td>Windows XP</td>
<td>windows, software, tools, pc, computer, tech, winxp, microsoft</td>
</tr>
<tr>
<td></td>
<td>Extreme programming</td>
<td>software, programming, process, methodology, development</td>
</tr>
</tbody>
</table>
Evaluation

Performance Measures

- **Precision**
  Measures the percentage of documents which are classified correctly.

- **Recall**
  Measures the percentage of classifiable documents which are classified correctly.

- **Coverage**
  Measures the proportion of documents the classifiers are able to classify.
**Evaluation**

**An Example**

Documents returned by Google

- \(C_1\): Doc1, Doc2
- \(C_2\): Doc3, Doc4
- \(C_0\): Doc5

Results of manual classification

- \(C_1\): Doc1, Doc2
- \(C_2\): Doc3, Doc4
- \(C_0\): Doc5

Results returned by classifier

- \(C_1\)
- \(C_2\)
- \(C_0\)

**Evaluation Metrics**

- Precision = 2/3
- Recall = 2/4
- Coverage = 3/4
Evaluation

Figure 1. Precision, recall and coverage against different values of $\beta$. 
**Evaluation**

![Graph of Precision, Recall, and Coverage](image)

*Figure 2. Precision, recall and coverage for different tags. (k = 11, β = 0.15)*
Discussions

Precision

- Precision is generally quite high (67-100%)
- Clustering process provides good basis for the kNN classifier
- Low precision cases: different keywords even for same context

WI’08 - Classifying Web Search Results with Data from Folksonomies
Recall

- Recall ranges from 65% to 93%
- Some documents cannot be classified (recognised)
- Mainly because keywords do not match well
**Discussions**

**Coverage**

- Has the largest range: 28-81%
- Due partly to low recall
- Some contexts not discovered by the clustering process (e.g. tube)
- Also, there are irrelevant results (e.g. bridge)
Conclusions

♦ Folksonomies offer rich information on the relations and semantics of tags, and can be used to enhance Web search

♦ Advantages over using of dictionaries or thesauruses (able to keep up with new meanings)

♦ Future research directions:
  1. Building more comprehensive classifiers
  2. Use of other clustering methods
  3. Larger scale of evaluation
Thank You!

Albert Au Yeung
cmay06r@ecs.soton.ac.uk
http://users.ecs.soton.ac.uk/cmay06r/