Exploring Social Tagging Graph for Web Object Classification

Zhijun Yin, Rui Li, Qiaozhu Mei, Jiawei Han
Department of Computer Science
University of Illinois at Urbana-Champaign
Outline

- Motivation: Web Object Classification
- Related Work
- Problem Formulation
- Classification Algorithm
- Experiments
- Conclusions
Web Object Classification

- *Web objects* become increasingly popular (>10⁶-10⁹)
  - *products* sold on *Amazon*
  - *videos* uploaded to *YouTube*
  - *research papers* referenced on *CiteULike*
  - *photos* uploaded to and collected by *Flickr* and *Facebook*

- Why classifying *web objects* into *semantic categories*?
  - Index and organize web objects efficiently
  - Browse and search of web objects conveniently
  - Discover interesting patterns from web objects
Subtlety on Web Object Classification

“Harry Potter” DVD
Class: “Movies & TV”

The fifth book of “Harry Potter”
Class: “Books”

“Harry Potter” Halloween costume
Class: “Apparel & Accessories”
Challenges for Web Obj. Classification

- Lack of features
  - Limited text description, e.g., title of a picture on Flickr
  - Inaccurate/difficult content features of images/videos
- Lack of interconnections
  - Often in isolate settings, w. limited interconnections
  - E.g. *Michael Jordan: a basketball star or a Berkeley professor?*
- Lack of labels
  - Impractical to obtain a huge number of labels
  - Without enough labels, how can one do effective classification?
Social Tagging I: Tagging Web Pages

Searching Everybody's www.google.com bookmarks for:

- google

Sign in to search your own bookmarks

See all bookmarks tagged google

Everybody's bookmarks

1572495 results - show all

Google

First saved by: Sajma

Google Guide Quick Reference: Google Advanced Operators (Cheat Sheet)

First saved by: TomSawyer

Google Trends

First saved by: atul

Google Reader

First saved by: rbns

Google Code - Google's Developer Network

First saved by: idealisms
Social Tagging II: Tagging Products

Harry Potter and the Order of the Phoenix (Book 5) (Paperback)
by J. K. Rowling (Author), Mary GrandPré (Author)

⭐⭐⭐⭐⭐ (5,879 customer reviews)

List Price: $42.00
Price: $10.18 & eligible for **FREE Super Saver Shipping** on orders over $25.
Details
You Save: $2.81 (22%)

**In Stock.**
Ships from and sold by Amazon.com. Gift-wrap available.

<table>
<thead>
<tr>
<th>Edition</th>
<th>List Price:</th>
<th>Our Price:</th>
<th>Other Offers:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hardcover (1st)</td>
<td>$29.99</td>
<td>$19.79</td>
<td>839 used &amp; new from $0.01</td>
</tr>
<tr>
<td>Paperback (Import)</td>
<td>$55.00</td>
<td>$47.25</td>
<td>19 used &amp; new from $1.25</td>
</tr>
<tr>
<td>Audio CD (Unabridged)</td>
<td>$75.00</td>
<td>$47.25</td>
<td>63 used &amp; new from $34.48</td>
</tr>
<tr>
<td>Mass Market Paperback</td>
<td>$34.99</td>
<td>$32.99</td>
<td>5 used &amp; new from $32.38</td>
</tr>
<tr>
<td>Library Binding (Library)</td>
<td>$34.99</td>
<td>$32.99</td>
<td>41 used &amp; new from $3.86</td>
</tr>
</tbody>
</table>

**See more editions and formats**

**Tags Customers Associate with This Product** *(What's this?)*
Click on a tag to find related items, discussions, and people.

- **harry potter** (159)
- **fantasy** (106)
- **jk rowling** (78)
- **book** (32)
- **fiction** (32)
- **fantasy series** (12)
- **london** (9)
- **for intelligent children** (7)
- **adventure** (6)
- **great juvenile fiction** (6)
- **another world** (5)

**See all 77 tags...**
Social Tagging Does Exist

- There exist many existing social tagging sites
  - Flickr (tagging pictures)
  - Digg (tagging news articles)
  - Technorati (tagging blogs)
  - Live search QnA (tagging questions)
  - ...
Intuition

- Social tagging can tackle the above challenges
  - Lack of features
    - Tags "art" and "architecture" are good features to characterize the book "ancient Greek art and architecture".
Intuition (Cont.)

- Social tagging can tackle the above challenges
  - Lack of interconnections
    - Although web page $P_1$ and web page $P_2$ do not have any tags in common, there is an implicit path from $P_1$ to $P_2$ via two tags and $P_2$. Class of $P_1$ can infer the class of $P_2$. 
Intuition (Cont.)

- Social tagging can tackle the above challenges
  - Lack of labels
  - Assume that the labels of web pages are easy to obtain, the class of web page "www.art.com" can infer the class of an art picture in Flickr via tag "art".
Motivation: Web Object Classification

Related Work

Problem Formulation

Classification Algorithm

Experiments

Conclusions
Related Work

- Web object classification
  - Web page classification
  - Multimedia classification
- Social tag usage
  - Web search
  - Information retrieval
  - Semantic web
  - Web page clustering
  - User interest mining
Outline

- Motivation: Web Object Classification
- Related Work
- Problem Formulation
- Classification Algorithm
- Experiments
- Conclusions
Given: Social Tagging Graph

- Objects of type $T$ are the target objects to be assigned category labels
- Objects of type $S$ are labeled objects from another domain
Notations: Social Tagging Graph

- **$C$:** a category set, {$c_1, c_2, \ldots, c_k$}
- **$G = (V,E)$:** a social tagging graph. Every object, $u$, and every tag, $v$, is a vertex in the graph $G$. If an object $u$ is associated with a tag $v$, there will be an edge between $u$ and $v$
- **$V_S$:** a set of objects of type $S$
- **$V_T^l$:** a set of labeled objects of type $T$
- **$V_T^u$:** a set of unlabeled objects of type $T$
- **$V_{tag}$:** a set of tags
Web Object Classification Problem

- Achieve consistency on social tagging graph
  - Category assignment of a vertex in *should not deviate* much from its original label
  - Category assignment of the vertex in *should remain* the same with its original label if it is fully trustable
  - Category of the vertex in $V$ *should take the prior knowledge* into consideration if there is any
  - Category assignment of any vertex in graph $G$ *should be* as consistent as possible to the categories of its neighbors
The Optimization Framework

\[ O(f) = \alpha \sum_{u \in V_S} \| f_u - \hat{f}_u \|^2 + \beta \sum_{u \in V_T^L} \| f_u - \hat{f}_u \|^2 + \gamma \sum_{u \in V_T^u} \| f_u - \hat{f}_u \|^2 + \sum_{(u,v) \in E} w_{uv} \| f_u - f_v \|^2 \]

- \( f_u \): a \( k \)-dimension vector that represents the class distribution of vertex \( u \in V \), where \( k \) is the number of categories. \( f_u[i] \) represents the possibility that \( u \) belongs to category \( i \), s.t. \( \sum_{i=1}^{k} f_u[i] = 1 \). We denote \( \{f_u\}_{u \in V} \) as \( f \).

- \( \hat{f}_u \): for \( u \in V_S \cup V_T^L \), \( \hat{f}_u \) is the class distribution estimated from the original category labels of vertex \( u \). For \( u \in V_T^u \), \( \hat{f}_u \) is the class distribution estimated from some prior knowledge of the unlabeled object \( u \) (e.g., the label assignments by a domain classifier).

- \( w_{uv} \): a weight of the importance of edge \((u,v)\). Given an object \( u \) and its associated tag \( v \), \( w_{uv} \) is the frequency that \( v \) is used to tag \( u \).
The Optimization Framework

$$O(f) = \alpha \sum_{u \in V_S} \| f_u - \hat{f}_u \|^2 + \beta \sum_{u \in V_T} \| f_u - \hat{f}_u \|^2 + \gamma \sum_{u \in V_T} \| f_u - \hat{f}_u \|^2 + \sum_{(u,v) \in E} w_{uv} \| f_u - f_v \|^2$$

1. $\sum_{u \in V_S} \| f_u - \hat{f}_u \|^2$ means that the category of a vertex in $V_S$ should not deviate much from its original label(s).

2. $\sum_{u \in V_T} \| f_u - \hat{f}_u \|^2$ means that the category of a vertex in $V_T$ should keep close to its initial label(s).

3. $\sum_{u \in V_T} \| f_u - \hat{f}_u \|^2$ means that the category of a vertex in $V_T$ should keep close to the prior knowledge if any.

4. $\sum_{(u,v) \in E} w_{uv} \| f_u - f_v \|^2$ makes sure that the class distribution of the vertices are smooth over the whole graph, i.e., the class distribution of a vertex is consistent with its neighbors.
Target: Minimizing $O(f)$

- Our target is to find $f^*$ to minimize the $O(f)$

- The class label $c$ of object $o$

\[
f^* = \arg \min \ O(f)
\]

\[
c = \arg \max \frac{P(o|c)}{P(o)} = \arg \max \frac{P(c|o)}{P(c)}
\]

\[
c = \arg \max_{1 \leq i \leq k} \sum_{u' \in V_T} f_{u'}^{*}[i]
\]
Outline

- Motivation: Web Object Classification
- Related Work
- Problem Formulation
- Classification Algorithm
- Experiments
- Conclusions
Classification Algorithm

- Finding the close solution of the above optimization problem requires the computation of the inverse of a matrix with the size of all web objects and tags.
- In reality, this is usually not feasible due to the complexity of computation.
- An efficient iterative algorithm to solve the optimization problem.
Classification Algorithm

Overall, it takes $O(k(|V| + \text{iter}|E|))$ time

The initialization steps (i.e., steps 1-4) take $O(k|V|)$ time

The iteration steps (i.e., steps 5-12) take $O(2k|E|)$ time

At step 7, the class distributions of objects of type $S$ are updated from the class distributions of the associated tags

At step 8, the class distributions of the labeled objects of type $T$ are updated from the class distributions of the associated tags

At step 9, the class distributions of the unlabeled objects of type $T$ are updated from the class distributions of the associated tags

At step 10, the class distributions of the tags are updated from the class distributions of the connected object

It takes $O(k|V_t^u|)$ time to get the class labels (i.e., steps 13-14).
Parameter Setting (Semi-Supervised Learning)

\[ \alpha = 0, \beta \neq 0, \gamma = 0 \]
Parameter Setting (Transfer Learning)

\[ \alpha \neq 0, \beta = 0, \gamma = 0 \]
Parameter Setting (Prior Integration)

\[ \alpha \neq 0, \beta \neq 0, \gamma \neq 0 \]
Classification Algorithm

- Convergence proof
  - Equivalent to absorption random walk on a new graph
  - Details in the paper
Outline

- Motivation: Web Object Classification
- Related Work
- Problem Formulation
- Classification Algorithm
- Experiments
- Conclusions
Experiments: Data Collections

- 6123 *products* from *Amazon*
- 5536 *web pages* (under *ODP Shopping* category)
- Tags of web pages are collected from *Delicious*

<table>
<thead>
<tr>
<th>ODP:Shopping</th>
<th>Count</th>
<th>Amazon</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Publications/Books</td>
<td>558</td>
<td>Books</td>
<td>937</td>
</tr>
<tr>
<td>Consumer_Electronics</td>
<td>494</td>
<td>Electronics</td>
<td>945</td>
</tr>
<tr>
<td>Health</td>
<td>1009</td>
<td>HealthPersonCare</td>
<td>747</td>
</tr>
<tr>
<td>Home_and_Garden</td>
<td>1976</td>
<td>HomeGarden</td>
<td>841</td>
</tr>
<tr>
<td>Jewelry</td>
<td>452</td>
<td>Jewelry</td>
<td>386</td>
</tr>
<tr>
<td>Music</td>
<td>527</td>
<td>Music</td>
<td>944</td>
</tr>
<tr>
<td>Office</td>
<td>77</td>
<td>OfficeProducts</td>
<td>695</td>
</tr>
<tr>
<td>Pet</td>
<td>443</td>
<td>PetSupplies</td>
<td>628</td>
</tr>
</tbody>
</table>
Experiments: Measurement

- **Measurement**
  \[
  F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
  \]

  - Micro-averaged scores (MicroF1) tend to be dominated by the performance on common categories.
  - Macro-averaged scores (MacroF1) are influenced by the performance in rare categories.

- **Baseline**
  - **SVM+TITLE**: SVM using product titles as feature.
  - **SVM+TAG**: SVM using tags as feature.
  - **HG+TITLE**: Harmonic Gaussian field method using titles. Use cosine similarity of the titles as edge weight.
  - **HG+TAG**: Harmonic Gaussian field method using tags. Use cosine similarity of the tags as edge weight.
Experiment: Overall Performance

- Overall performance comparison
  - $TM$ (Tag-based classification Model) to refer to our method.

<table>
<thead>
<tr>
<th>Label Ratio</th>
<th>Measure</th>
<th>1% MicroF1</th>
<th>1% MacroF1</th>
<th>5% MicroF1</th>
<th>5% MacroF1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SVM+TITLE</td>
<td>0.4233</td>
<td>0.3812</td>
<td>0.5967</td>
<td>0.6091</td>
</tr>
<tr>
<td></td>
<td>SVM+TAG</td>
<td>0.4045</td>
<td>0.4059</td>
<td>0.6397</td>
<td>0.6435</td>
</tr>
<tr>
<td></td>
<td>HG+TITLE</td>
<td>0.6251</td>
<td>0.6038</td>
<td>0.6778</td>
<td>0.6689</td>
</tr>
<tr>
<td></td>
<td>HG+TAG</td>
<td>0.7174</td>
<td>0.7127</td>
<td>0.7856</td>
<td>0.7859</td>
</tr>
<tr>
<td></td>
<td>TM *</td>
<td>0.7870</td>
<td>0.7872</td>
<td>0.8027</td>
<td>0.8030</td>
</tr>
</tbody>
</table>

*$\alpha = 1000, \beta = \infty, \gamma = 0.1$
Experiment (Cont.)

- Challenge in web object classification
  - Lack of features
  - Lack of interconnections
  - Lack of labels
Experiment (Cont.)

- Lack of features?
  - Effectiveness of tag feature
- Lack of interconnection?
  - Exploring the interconnections of objects

<table>
<thead>
<tr>
<th></th>
<th>SVM+TITLE</th>
<th>SVM+TAG</th>
<th>HG+TITLE</th>
<th>HG+TAG</th>
<th>TM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p%</td>
<td>MicroF1</td>
<td>MacroF1</td>
<td>MicroF1</td>
<td>MacroF1</td>
</tr>
<tr>
<td>5%</td>
<td></td>
<td>0.5967</td>
<td>0.6091</td>
<td>0.6397</td>
<td>0.6435</td>
</tr>
<tr>
<td>10%</td>
<td></td>
<td>0.6700</td>
<td>0.6789</td>
<td>0.7168</td>
<td>0.7334</td>
</tr>
<tr>
<td>15%</td>
<td></td>
<td>0.7181</td>
<td>0.7218</td>
<td>0.7417</td>
<td>0.7366</td>
</tr>
<tr>
<td>20%</td>
<td></td>
<td>0.7343</td>
<td>0.7399</td>
<td>0.7674</td>
<td>0.7722</td>
</tr>
<tr>
<td>25%</td>
<td></td>
<td>0.7545</td>
<td>0.7597</td>
<td>0.7763</td>
<td>0.7780</td>
</tr>
</tbody>
</table>

*α = 0, β = ∞, γ = 0
Experiment (Cont.)

- Lack of labels?
  - Handling lack of labeling issue

<table>
<thead>
<tr>
<th>p%</th>
<th>HG+TITLE MicroF1</th>
<th>HG+TAG MicroF1</th>
<th>HG+TAG MacroF1</th>
<th>HG+TAG MacroF1</th>
<th>α = 1000 MicroF1*</th>
<th>MacroF1*</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>0.7594</td>
<td>0.7606</td>
</tr>
<tr>
<td>1%</td>
<td>0.6251</td>
<td>0.7174</td>
<td>0.7127</td>
<td>0.7708</td>
<td>0.7719</td>
<td></td>
</tr>
<tr>
<td>2%</td>
<td>0.6499</td>
<td>0.7510</td>
<td>0.7434</td>
<td>0.7771</td>
<td>0.7766</td>
<td></td>
</tr>
<tr>
<td>3%</td>
<td>0.6368</td>
<td>0.7695</td>
<td>0.7666</td>
<td>0.7774</td>
<td>0.7769</td>
<td></td>
</tr>
<tr>
<td>4%</td>
<td>0.6503</td>
<td>0.7566</td>
<td>0.7513</td>
<td>0.7885</td>
<td>0.7891</td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>0.6778</td>
<td>0.7856</td>
<td>0.7859</td>
<td>0.7872</td>
<td>0.7866</td>
<td></td>
</tr>
</tbody>
</table>

* $\beta = \infty, \gamma = 0$
Experiment (Cont.)

- Sensitivity of parameter $\alpha$
Experiment (Cont.)

- Prior Knowledge
  - With prior > Without prior
  - $\textit{SVM+TAG}$ with prior > $\textit{SVM+TAG}$
  - $\textit{HG+TAG}$ with prior > $\textit{HG+TAG}$
Outline

- Motivation: Web Object Classification
- Related Work
- Problem Formulation
- Classification Algorithm
- Experiments
- Conclusions
Conclusions

- Web object classification: An emerging task and increasingly important
- Web object classification problem can take advantage from social tags in three aspects
  - represent web objects in a meaningful feature space
  - interconnect objects to indicate implicit relationship
  - bridging heterogeneous objects so that category information can be propagated from one domain to another
- We propose a general framework to model the problem as an optimization problem on a social tagging graph, which covers different scenarios of web object classification problem
- In our model, we only consider the setting of two types of web objects
  - It is interesting to generalize our model to manage multi-types of objects
THANKS!
Q&A