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

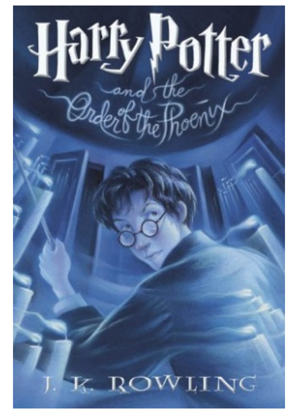
- *Web objects* become increasingly popular ($>10^6$-10^9)
  - *products* sold on *Amazon*
  - *videos* uploaded to *YouTube*
  - *research papers* referenced on *CiteULike*
  - *photos* uploaded to and collected by *Flickr* and *Facebook*
- Why classifying *web objects* into *semantic categories?*
  - 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**"

amazon.com



The fifth book of "Harry Potter"

Class: "**Books**"

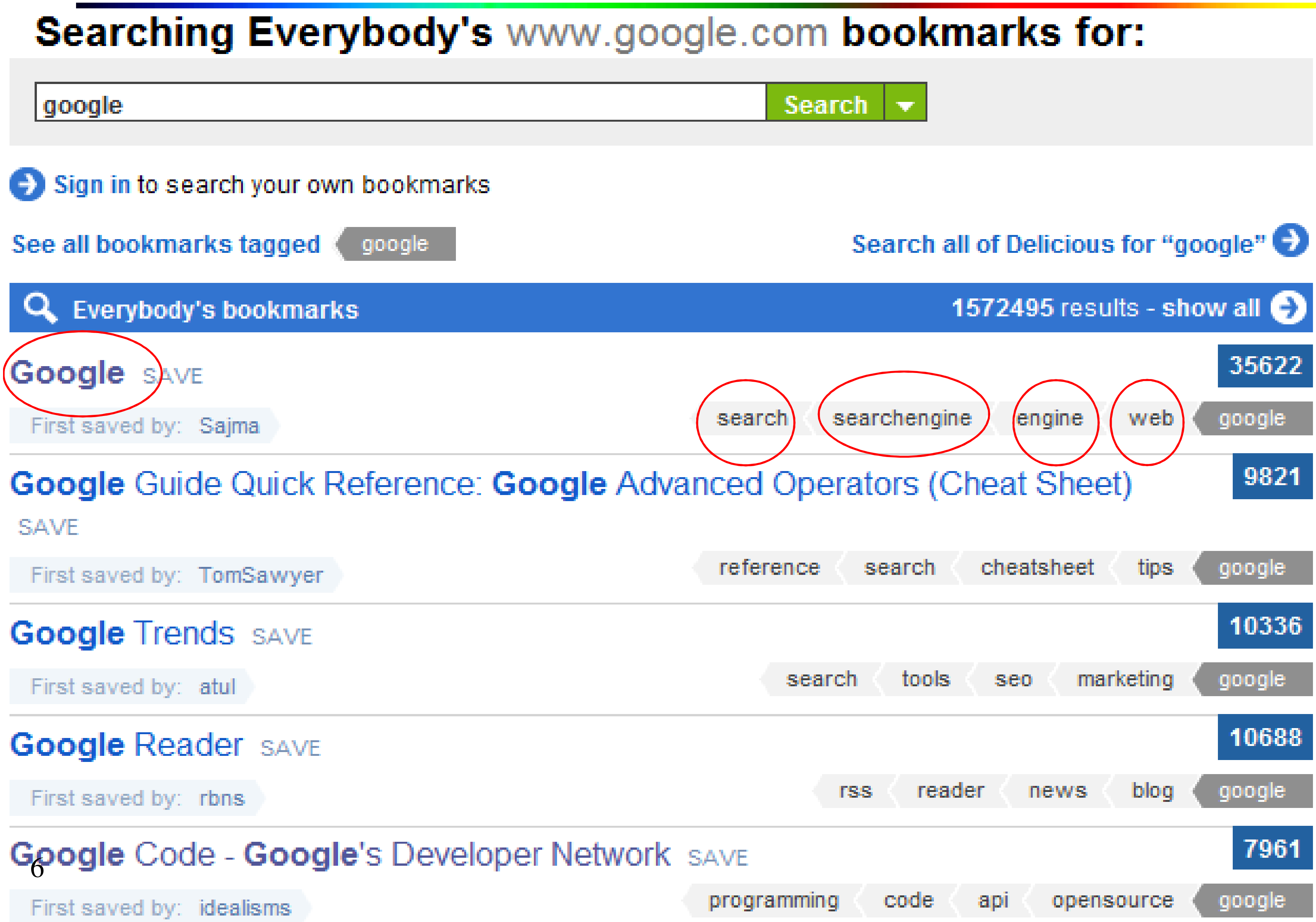
"Harry Potter" Halloween costume

Class: "**Apparel & Accessories**"

Challenges for Web Obj. Classification

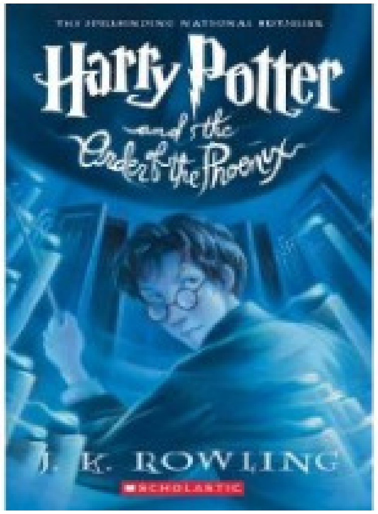
- Lack of features
  - Limited text description, e.g., title of a picture on Flickr
  - Inaccurate/difficult content features of images/videos
- Lack of interconnections
  - Often in isolate settings, w. limited interconnections
  - E.g. *Michael Jordan: a basketball star or a* Berkeley professor?
- 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



del.icio.us

Social Tagging II: Tagging Products



See larger image

See 1 customer image

Share your own customer images

Publisher: learn how customers can search inside this book.

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: ~~$12.99~~

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.

60 new from $6.00 **132 used** from $1.15 **15 collectible** from $10.00

Also Available in:	List Price:	Our Price:	Other Offers:
Hardcover (1st)	~~$29.99~~	$19.79	839 used & new from $0.01
Paperback (Import)			19 used & new from $1.25
Audio CD (Unabridged)	~~$75.00~~	$47.25	63 used & new from $34.48
Mass Market Paperback			5 used & new from $32.38
Library Binding (Library)	~~$34.99~~	$32.99	41 used & new from $3.86

Show more editions and formats

Tags Customers Associate with This Product (What's this?)

Click on a tag to find related items, discussions, and people.

Check the boxes next to the tags you consider relevant or enter your own tags in the field below.

harry potter (159)	fiction (32)	adventure (6)
fantasy (106)	fantasy series (12)	great juvenile fiction (6)
jk rowling (78)	london (9)	another world (5)
book (32)	for intelligent children (7)	**See all 77 tags...**

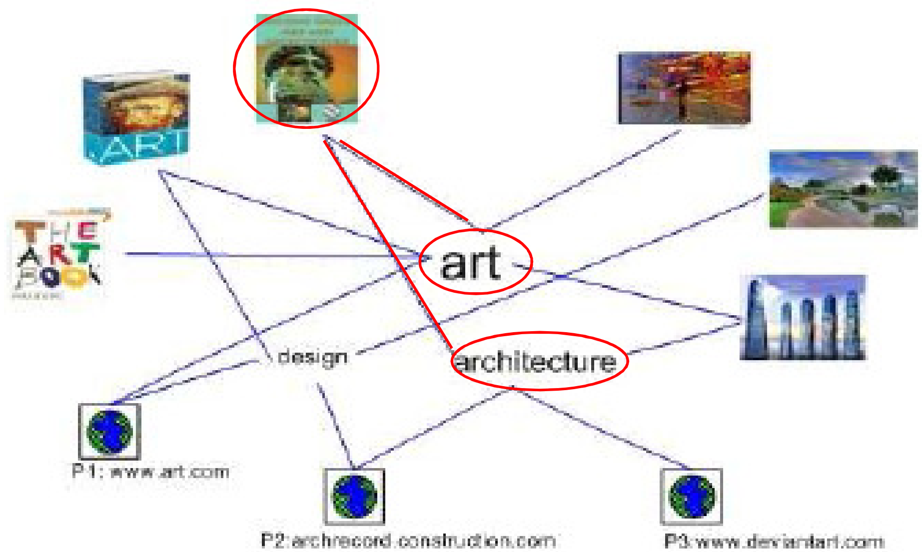
amazon.com

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)
  - …

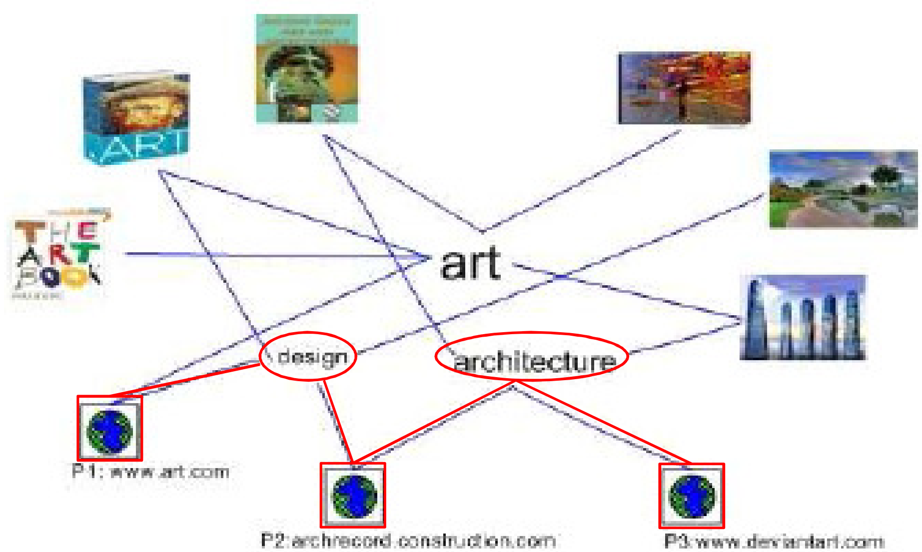
Intuition

- 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".*



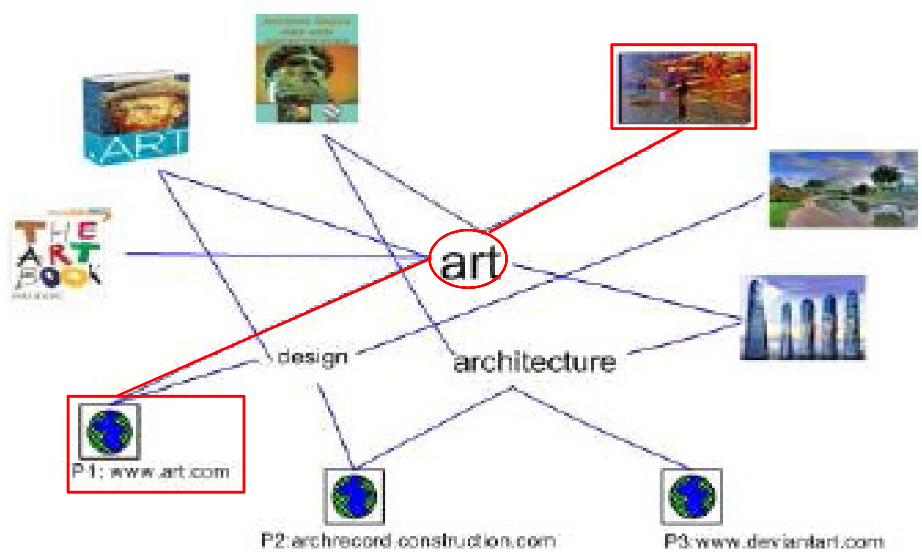
Intuition (Cont.)

- 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"*.



Outline

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

Related Work

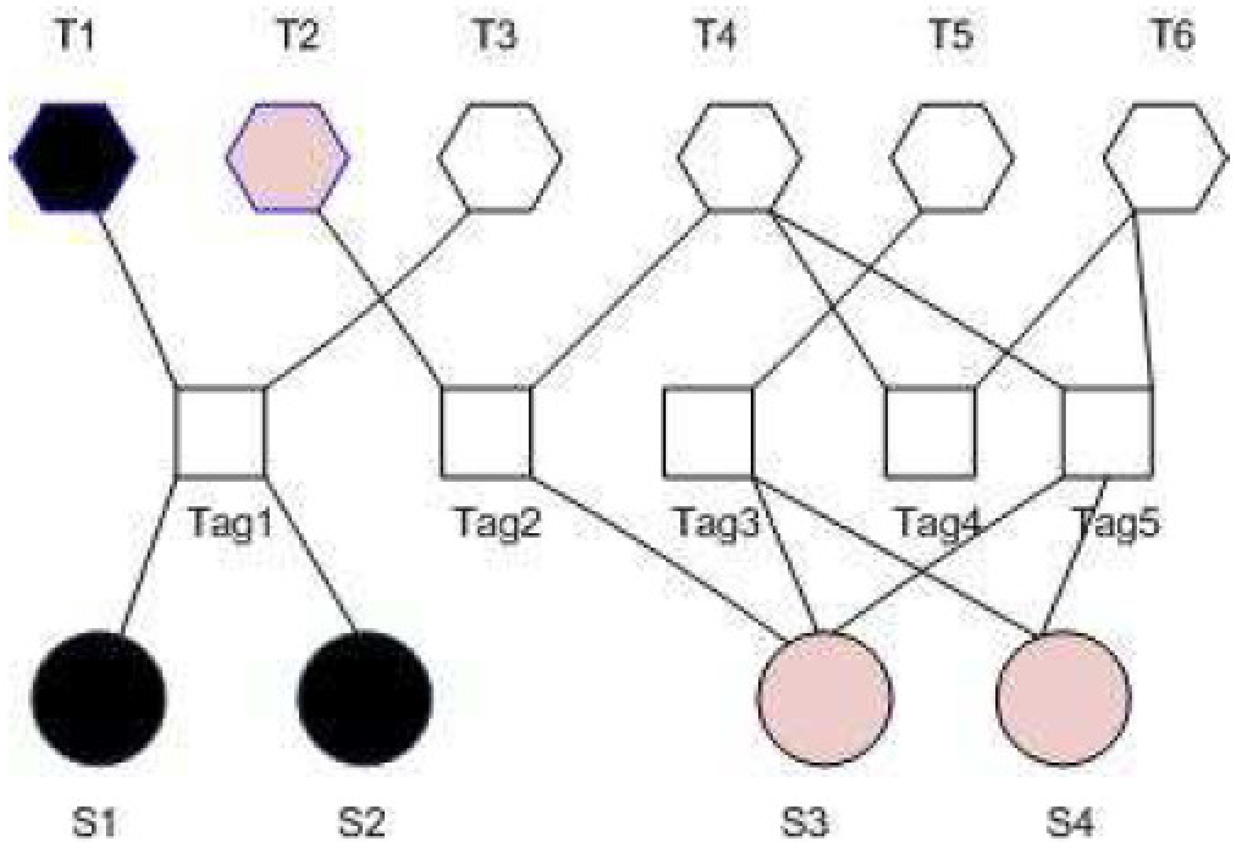
- Web object classification
  - Web page classification
  - Multimedia classification
- Social tag usage
  - Web search
  - Information retrieval
  - Semantic web
  - Web page clustering
  - User interest mining

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 knowl*edge 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

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

- 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

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

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

DAIS The Database and Information Systems Laboratory at The University of Illinois at Urbana-Champaign

Target: Minimizing *O(f)*

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

- The class label c of object o

$$f^* = \arg\min O(f)$$

$$c = \arg\max \frac{P(o|c)}{P(o)} = \arg\max \frac{\boxed{P(c|o)}}{\boxed{P(c)}}$$

$$c = \arg\max_{1 \le i \le k} \frac{\boxed{f_u^*[i]}}{\boxed{\sum_{u' \in V_T^l \cup V_T^u} f_{u'}^*[i]}}$$

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

Algorithm 1: Iterative Algorithm

Input: category size k, class labels $C(x)$ for $x \in V_S \cup V_T^l \cup V_T^u$

Output: class labels $\tilde{C}(x)$ for $x \in V_T^u$

```
   // Initialization
1  foreach x ∈ V_S ∪ V_T^l ∪ V_T^u do
2      f̂_x[C(x)] ← 1
3  foreach x ∈ V do
4      foreach i ← 1 to k do  f_x[i] ← 1/k
   // Iteration
5  repeat
6      foreach x ∈ V_S do
7          f'_x ← α/(α+Σ_{v∈V_tag} w_xv) f̂_x + (Σ_{v∈V_tag} w_xv f_v)/(α+Σ_{v∈V_tag} w_xv)
       foreach x ∈ V_T^l do
8          f'_x ← β/(β+Σ_{v∈V_tag} w_xv) f̂_x + (Σ_{v∈V_tag} w_xv f_v)/(β+Σ_{v∈V_tag} w_xv)
       foreach x ∈ V_T^u do
9          f'_x ← γ/(γ+Σ_{v∈V_tag} w_xv) f̂_x + (Σ_{v∈V_tag} w_xv f_v)/(γ+Σ_{v∈V_tag} w_xv)
       foreach x ∈ V_tag do
10         f'_x ← (Σ_{s∈V_S} w_sx f_s + Σ_{l∈V_T^l} w_lx f_l + Σ_{u∈V_T^u} w_ux f_u)/(Σ_{s∈V_S} w_sx + Σ_{l∈V_T^l} w_lx + Σ_{u∈V_T^u} w_ux)
       foreach x ∈ V do
11         f_x ← f'_x
12 until converged ;
   // Get Class Label
13 foreach x ∈ V_T^u do
14     C̃(x) = arg max_{1≤i≤k} f_x[i] / Σ_{u∈V_T^l ∪ V_T^u} f_u[i]
```

Overall, it takes *O(k(|V | + 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

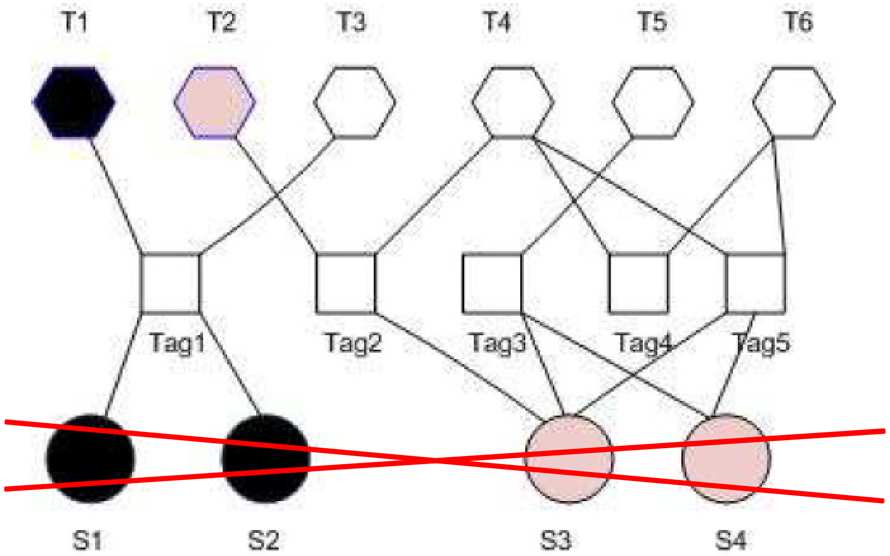
At step 10, the class distributions of the tags are updated from the class distributions of the connected object

It takes $O(k|V_T^u|)$ time to get the class labels (i.e., steps 13-14).

DAIS The Database and Information Systems Laboratory at The University of Illinois at Urbana-Champaign

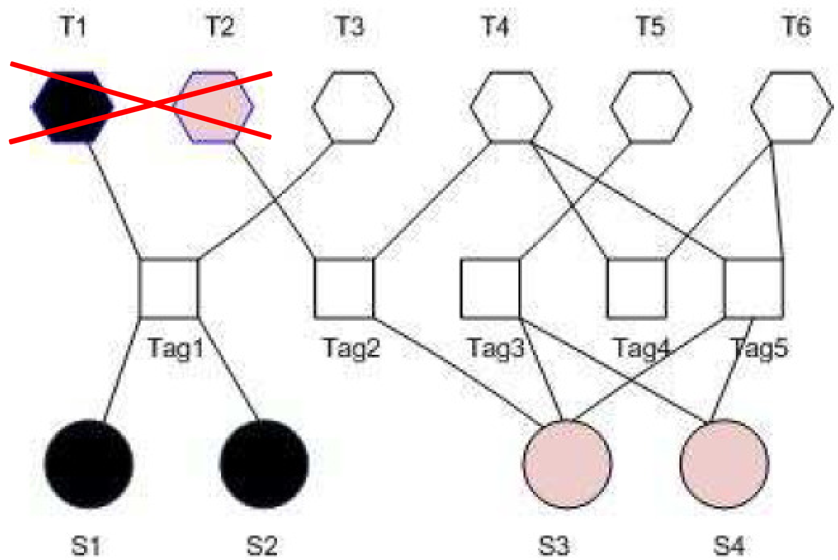
Parameter Setting (Semi-Supervised Learning)

$$\alpha = 0, \beta \neq 0, \gamma = 0$$



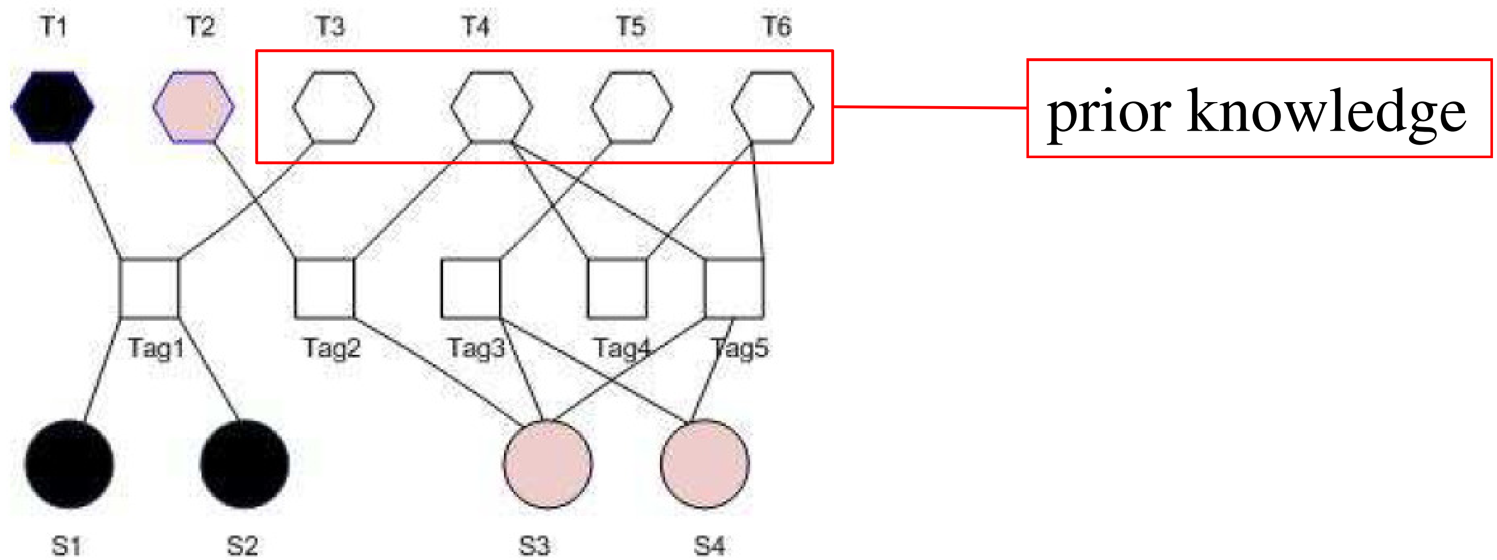
Parameter Setting (Transfer Learning)

$$\alpha \neq 0, \beta = 0, \gamma = 0$$



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*

ODP:Shopping		Amazon	
Name	Count	Name	Count
Publications/Books	558	Books	937
Consumer_Electronics	494	Electronics	945
Health	1009	HealthPersonCare	747
Home_and_Garden	1976	HomeGarden	841
Jewelry	452	Jewelry	386
Music	527	Music	944
Office	77	OfficeProducts	695
Pet	443	PetSupplies	628

Experiments: Measurement

- Measurement $$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$
  - 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* (*T*ag-based classification *M*odel) to refer to our method.

Label Ratio	1%		5%	
Measure	MicroF1	MacroF1	MicroF1	MacroF1
SVM+TITLE	0.4233	0.3812	0.5967	0.6091
SVM+TAG	0.4045	0.4059	0.6397	0.6435
HG+TITLE	0.6251	0.6038	0.6778	0.6689
HG+TAG	0.7174	0.7127	0.7856	0.7859
TM *	**0.7870**	**0.7872**	**0.8027**	**0.8030**

$$^*\alpha = 1000, \beta = \infty, \gamma = 0.1$$

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

	SVM+TITLE		SVM+TAG		HG+TITLE		HG+TAG		TM	
$p\%$	MicroF1	MacroF1	MicroF1	MacroF1	MicroF1	MacroF1	MicroF1	MacroF1	MicroF1	MacroF1
5%	0.5967	0.6091	0.6397	0.6435	0.6778	0.6689	0.7856	0.7859	**0.7918***	**0.7919**
10%	0.6700	0.6789	0.7168	0.7334	0.6937	0.6802	0.7915	0.7864	**0.8005**	**0.7996**
15%	0.7181	0.7218	0.7417	0.7366	0.7139	0.7049	0.7921	0.7908	**0.8187**	**0.8199**
20%	0.7343	0.7399	0.7674	0.7722	0.7152	0.7059	0.8025	0.8004	**0.8217**	**0.8231**
25%	0.7545	0.7597	0.7763	0.7780	0.7131	0.7038	0.8109	0.8079	**0.8259**	**0.8273**

$*\alpha = 0, \beta = \infty, \gamma = 0$

Experiment (Cont.)

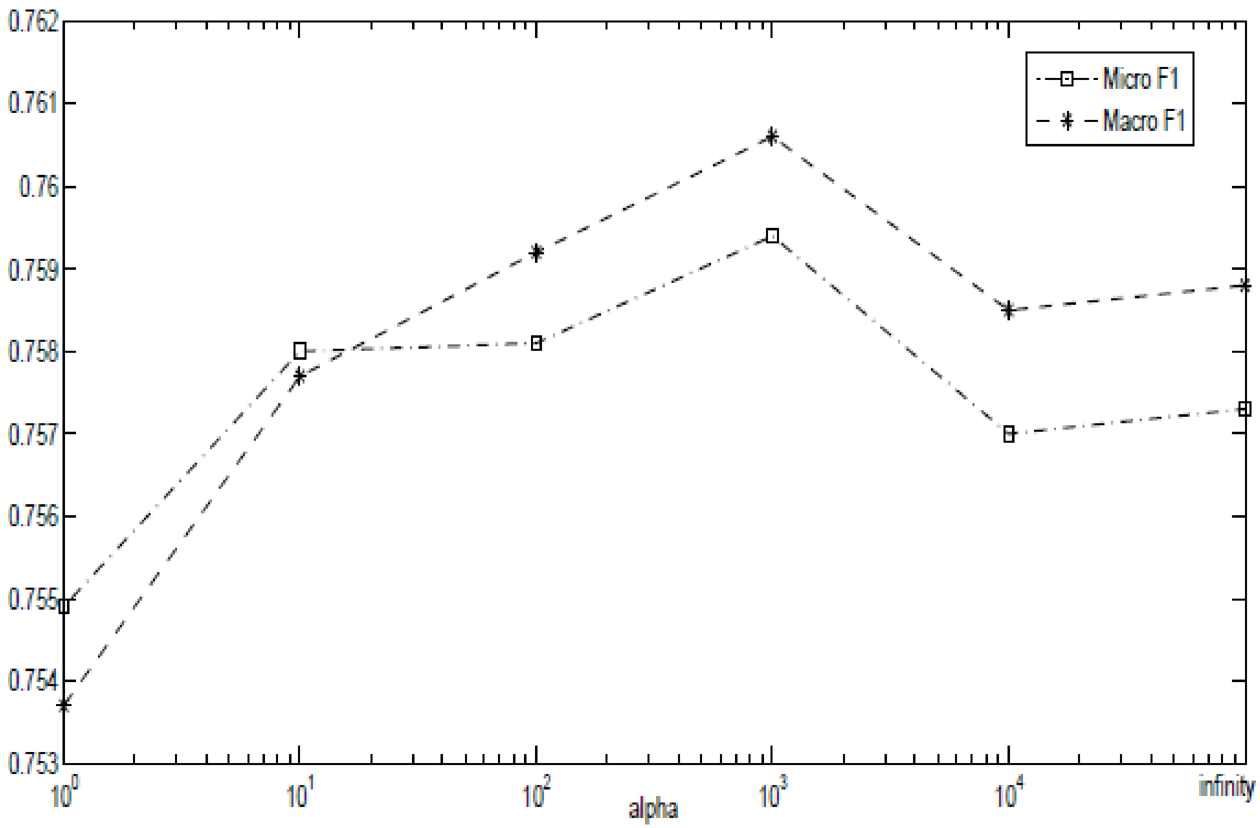
- Lack of labels?
  - Handling lack of labeling issue

	HG+TITLE		HG+TAG		$\alpha = 1000$	
$p\%$	MicroF1	MacroF1	MicroF1	MacroF1	MicroF1*	MacroF1
0	NA	NA	NA	NA	0.7594	0.7606
1%	0.6251	0.6038	0.7174	0.7127	**0.7708**	**0.7719**
2%	0.6499	0.6334	0.7510	0.7434	**0.7771**	**0.7766**
3%	0.6368	0.6368	0.7695	0.7666	**0.7774**	**0.7769**
4%	0.6503	0.6360	0.7566	0.7513	**0.7885**	**0.7891**
5%	0.6778	0.6689	0.7856	0.7859	**0.7872**	**0.7866**

$$^{*}\beta = \infty, \gamma = 0$$

Experiment (Cont.)

- Sensitivity of parameter α



DAIS The Database and Information Systems Laboratory at The University of Illinois at Urbana-Champaign

Experiment (Cont.)

- Prior Knowledge
  - With prior > Without prior
  - *SVM+TAG* with prior > *SVM+TAG*
  - *HG+TAG* with prior > *HG+TAG*

$p\%$	5%		10%		15%		20%		25%	
Measure	MicroF1	MacroF1	MicroF1	MacroF1	MicroF1	MacroF1	MicroF1	MacroF1	MicroF1	MacroF1
γ=0	0.7918	0.7919	0.8005	0.7996	0.8187	0.8199	0.8217	0.8231	0.8259	0.8273
SVM+TAG	0.6397	0.6435	0.7168	0.7334	0.7417	0.7366	0.7674	0.7722	0.7763	0.7780
(γ=0.001)+(SVM+TAG)	0.7938	0.7914	0.8000	0.7987	0.8214	0.8198	0.8229	0.8238	0.8281	0.8295
(γ=0.01)+(SVM+TAG)	**0.7964**	**0.7932**	0.8013	0.8005	0.8199	0.8184	0.8223	0.8231	0.8292	0.8306
(γ=0.1)+(SVM+TAG)	0.7796	0.7673	**0.8096**	**0.8109**	**0.8251**	**0.8201**	**0.8272**	**0.8277**	**0.8355**	**0.8364**
(γ=1)+(SVM+TAG)	0.6878	0.6846	0.7704	0.7803	0.7913	0.7843	0.8033	0.8051	0.8165	0.8163
HG+TAG	0.7856	0.7859	0.7915	0.7864	0.7921	0.7908	0.8025	0.8004	0.8109	0.8079
(γ=0.001)+(HG+TAG)	0.7968	0.7973	0.8038	0.8026	0.8214	0.8228	0.8251	0.8263	0.8300	0.8316
(γ=0.01)+(HG+TAG)	0.8012	0.8028	0.8056	0.8040	0.8222	0.8233	0.8249	0.8261	0.8313	0.8329
(γ=0.1)+(HG+TAG)	**0.8038**	**0.8043**	**0.8174**	**0.8151**	**0.8233**	**0.8238**	**0.8296**	**0.8301**	**0.8381**	**0.8387**
(γ=1)+(HG+TAG)	0.7950	0.7951	0.8036	0.7982	0.8082	0.8065	0.8206	0.8192	0.8339	0.8308

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