一种基于社会性标注的网页排序算法

刘凯鹏[1)] 方滨兴[1),2)]

[1)](哈尔滨工业大学计算机网络与信息安全技术研究中心 哈尔滨 150001)

[2)](中国科学院计算技术研究所网络重点实验室 北京 100190)

摘 要 社会性标注作为一种新的资源管理和共享方式,吸引为数众多的用户参与其中,由此产生的大量社会性标注数据成为网页质量评价的一个新维度.文中研究如何利用社会性标注改进网页检索性能,提出一种有机结合网页和用户的查询相关性与互增强关系的网页排序算法.首先利用统计主题模型,使用相关标签为网页和用户建模,并计算查询相关性.然后利用二部图模型刻画网页和用户间的互增强关系,并使用相关标签与用户兴趣和网页内容的匹配度为互增强关系赋予权重.最后结合查询相关性和互增强关系,以迭代方式同时计算网页和用户的评分.实验结果表明,文中提出的检索模型和互增强模型能够有效地提高排序算法的性能.与目前的代表性算法相比,该算法在检索性能上有明显提高.

关键词 社会性标注;网页检索;网页质量;排序算法;主题模型

中图法分类号 TP391 **DOI 号**: 10.3724/SP.J.1016.2010.01014

A Novel Page Ranking Algorithm Based on Social Annotations

LIU Kai-Peng[1)] FANG Bin-Xing[1),2)]

[1)](*Research Center of Computer Network and Information Security Technology*, *Harbin Institute of Technology*, *Harbin* 150001)

[2)](*Key Laboratory of Network Science and Technology*, *Institute of Computing Technology*, *Chinese Academy of Sciences*, *Beijing* 100190)

Abstract With the rapid development of social tagging systems, large amount of social annotations have been created by large crowd of collaborative users, forming a new dimension of accessing the quality of Web pages. This paper proposes a novel page ranking algorithm for improving Web search performance. The authors explored the social annotations by effectively combining the language model of pages and users with the mutual reinforcement between pages and users, developed a probabilistic generative model to demonstrate the tagging scheme of users and resources, and modeled the mutual reinforcement relation between pages and users with a bipartite graph. Moreover, the authors assigned each one of the mutual reinforcement relations with a weight representing the coherence between annotating tags and language model of pages and users, and computed the importance of pages and users simultaneously in an iterative fashion based on both query relevance and mutual reinforcement. Experiments on a dataset collected from a real-world social tagging system show that the query model and mutual reinforcement model developed in this paper can effectively improve the performance of the ranking algorithm, outperforms other state-of-the-art algorithms in retrieval performance measured by MAP and NDCG.

Keywords social annotations; page retrieval; page quality; ranking algorithm; topic models

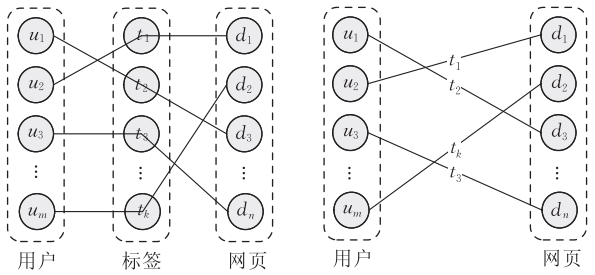
收稿日期:2010-01-09;最终修改稿收到日期:2010-05-13.本课题得到国家自然科学基金(60703014,60933005)、国家"九七三"重点基础研究发展规划项目基金(G2007CB311100)、国家"八六三"高技术研究发展计划项目基金(2006AA010105-02,2007AA01Z416,2007AA01Z442,2009AA01Z437)资助.**刘凯鹏**,男,1981 年生,博士研究生,研究方向为信息检索、数据挖掘和机器学习.E-mail:liukaipeng@gmail.com.**方滨兴**,男,1960 年生,博士,教授,中国工程院院士,研究领域为信息安全、信息检索和分布式系统.

1 引　言

随着 Web2.0 的兴起，基于社会性标注的内容共享系统，如共享网页的 Delicious①、共享图片的 Flickr② 和共享学术论文的 CiteULike③ 等，作为其中的典型应用，得到了迅速的发展. 社会性标注机制允许相互协作的用户通过一个开放的平台，对共享的资源赋予简短而富于个性化的标签，从而实现资源的有效管理和共享. 作为一种用户驱动的社会性协作机制，社会性标注系统吸引了大量的用户参与其中，并籍此形成了被称之为 Folksonomy 的体现大众智慧的大量社会性标注数据. 本文定义 Folksonomy 为四元组 $\mathcal{F}:=(U,T,D,A)$，其中 U、T 和 D 分别为用户、标签和资源的有限集合，定义在其上的三元关系 $A\subseteq U\times T\times D$ 称为标注集. 一个标注 $a=(u,t,d)\in A$ 表示用户 u 使用标签 t 标注了资源 d.

本文研究基于社会性标注的网页排序算法以改善网页检索性能. 传统的基于链接分析的网页排序算法，如 PageRank[1] 和 HITS[2] 等，利用网页间的链接关系来对网页排序. 这些链接是由网站作者出于不同的目的而加入到网页中，可以看作是对其他网页的间接评价. 与此相比，由网页读者通过收藏、标注和共享等行为而产生的社会性标注数据，则直接反映了用户对网页的质量评价和内容理解. 可以认为，社会性标注为我们提供了一个新的维度来评价网页的质量及其受欢迎程度. 因此，如何挖掘并利用蕴含在社会性标注数据中的社会性知识来提高信息检索的性能已经成为当前的研究热点. 虽然本文以资源类型为网页的社会性标注系统作为研究对象，但是本文提出的算法本身并不局限于网页排序，也可以应用在其他类型资源的检索中.

以往的研究工作大都基于社会性标注的三部图模型，将标签视为与网页和用户同等的对象进行排序（如图 1(a)所示）. 这类方法模型直观，算法简单，但存在一定不足. 由于标签只是用户对网页的描述，其本身并不具备独立的评价属性，认为使用重要标签的用户或被重要标签标注的网页也重要的观点缺乏依据. 因此，这类算法容易受到垃圾标注的影响：恶意用户可能大量使用广泛出现的标签来不正当地提升其影响力[3-6]. 实际上，作为依赖于网页和用户出现的描述性信息载体，标签的真正价值在于其语义信息可被用来描述用户兴趣和网页内容. 本文研究基于社会性标注的二部图模型（如图 1(b)所示）的网页排序算法，同时利用标签数据为网页和用户建立语言模型，用以计算查询相关性，优化检索性能.



(a) 三部图模型　　(b) 二部图模型

图 1　社会性标注的三部图与二部图模型的比较

不同对象间的互增强关系常被用来作为排序算法的基础. 例如在 HITS 算法中，网页被赋予两种属性：中心值（hub）和权威值（authority）. HITS 算法认为，被高中心值网页指向的网页具有高权威值，而指向高权威值网页的网页将获得高中心值. 通过迭代的计算，网页的中心值和权威值将收敛于合理判定网页排序的数值. 本文基于类似的思想，认为被高素质用户标注的网页具有高价值，而标注高价值网页的用户也具有高素质. 同时，我们也注意到，同一用户和不同网页以及不同用户和同一网页之间的互增强关系强弱并不相同. 通过网页和用户的语言模型以及用户标注网页时使用的标签，可以获知该标注与用户兴趣和网页内容之间的匹配度，从而量化互增强关系的强弱.

本文的主要贡献包括：(1) 提出一种社会性标注的主题模型刻画用户兴趣和网页内容；(2) 提出一种互增强模型刻画网页和用户之间的互增强关系；(3) 提出一种结合网页和用户的查询相关性与它们之间的互增强关系的网页排序算法. 本文在第 2 节介绍相关工作；在第 3 节描述网页排序算法；在第 4 节给出实验结果；在第 5 节作出总结.

2 相关工作

Hotho 等人最早提出了一种 FolkRank 算法[7] 来对社会性标注系统中的对象进行排序. 这种算法拓展了 PageRank 算法，将用户、资源和标签看作一个三部图中的 3 个顶点集合，不同集合中的顶点之间的权重为与它们共现的另一集合中的对象数. 所有对象基于 Folksonomy 的 PageRank 向量计算

① http://delicious.com
② http://www.flickr.com
③ http://www.citeulike.org

如下：

$$\boldsymbol{r}^{(i+1)}=\alpha\boldsymbol{r}^{(i)}+\beta\boldsymbol{W}\boldsymbol{r}^{(i)}+\gamma\boldsymbol{p},$$

式中 $\boldsymbol{r}^{(i)}$ 为在第 i 次迭代时的 PageRank 向量，$\boldsymbol{p}$ 为个性化向量，且 $\alpha+\beta+\gamma=1$. 分别用 $\boldsymbol{r}_0$ 和 $\boldsymbol{r}_1$ 表示上式在 $\beta=1$ 和 $\beta=0$ 时的解，则对象的评分向量 $\boldsymbol{r}$ 计算如下：

$$\boldsymbol{r}=\boldsymbol{r}_1-\boldsymbol{r}_0.$$

另一种与 FolkRank 类似的算法是 Bao 等人提出的 SocialPageRank 算法[8]. 他们采用与 FolkRank 算法相类似的模型，定义用户和资源、用户和标签以及资源和标签间的权重矩阵 $\boldsymbol{W}^{UD}$，$\boldsymbol{W}^{DU}$，$\boldsymbol{W}^{UT}$，$\boldsymbol{W}^{TU}$，$\boldsymbol{W}^{RT}$ 和 $\boldsymbol{W}^{TR}$，并用下式计算资源、用户和标签的评分 $\boldsymbol{r}$，$\boldsymbol{s}$ 和 $\boldsymbol{t}$：

$$\boldsymbol{s}^{(i)}=\boldsymbol{W}^{UD}\boldsymbol{r}^{(i)},\quad \boldsymbol{t}^{(i)}=\boldsymbol{W}^{TU}\boldsymbol{s}^{(i)},\quad \boldsymbol{r}^{(i)}=\boldsymbol{W}^{RT}\boldsymbol{t}^{(i)},$$

$$\boldsymbol{t}'^{(i)}=\boldsymbol{W}^{TR}\boldsymbol{r}^{(i)},\ \boldsymbol{s}'^{(i)}=\boldsymbol{W}^{UT}\boldsymbol{t}'^{(i)},\ \boldsymbol{r}^{(i+1)}=\boldsymbol{W}^{DU}\boldsymbol{s}'^{(i)}.$$

Noll 等人提出了一种 SPEAR 算法来支持在社会性标注系统中的专家搜索[9]. 他们同样利用用户和资源之间的互增强关系来对用户进行排序. 根据用户标注资源的时间先后，互增强关系被赋予不同的权重，以使较早进行标注的用户获得较高的评分. SPEAR 算法用下式计算用户和资源的评分 $\boldsymbol{r}$ 和 $\boldsymbol{s}$：

$$\boldsymbol{s}^{(i+1)}=\boldsymbol{W}\boldsymbol{r}^{(i)},\ \boldsymbol{r}^{(i+1)}=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{s}^{(i+1)},$$

式中 $\boldsymbol{W}$ 为权重矩阵，$\boldsymbol{W}^{\mathrm{T}}$ 为 $\boldsymbol{W}$ 的转置矩阵，$W_{u,d}=(C_{u,d}+1)^{0.5}$，而 $C_{u,d}$ 为在用户 u 之后标注网页 d 的用户数. 虽然 SPEAR 算法是被提出用来对用户进行排序的，但其计算的网页评分实际上可以被用来对网页进行排序.

3 网页排序算法

本节结合网页和用户的查询相关性与它们之间的互增强关系，提出一种基于社会性标注的网页排序算法. 首先利用用户使用的标签和网页被标注的标签数据来学习网页和用户的语言模型，计算查询相关性(第 3.1 节). 然后使用二部图模型刻画网页和用户之间的互增强关系，并结合语言模型和标注数据为这些关系赋予不同的权重(第 3.2 节). 最后结合查询相关性和互增强关系，利用迭代算法同时计算网页和用户的排序(第 3.3 节). 本节也给出该算法的收敛性和时间复杂性分析(第 3.4 节).

3.1 网页和用户的语言模型

从统计语言模型的观点来看，如果把用户使用的标签和网页被标注的标签看作它们的“语言”，那么通过学习它们的语言模型，可以更好地理解用户兴趣和网页内容. 在社会性标注系统中，同一网页往往有不同的内容属性，而同一用户也往往有不同的兴趣领域，因此，本文使用统计主题模型[10-14]来发现标签词汇中不同主题之间的隐含关系，并基于此模型计算查询相关性.

3.1.1 主题模型

利用相关标签，基于 LDA(Latent Dirichlet Allocation)[13]为网页和用户建模. 给定一个网页集合 D，其被用户集合 U 使用标签集合 T 标注. 每个网页 $d\in D$ 表示为其被标注的标签集合 $\{t_1,t_2,\cdots,t_{N_d}\}$. 假设主题数目固定为 K，则 d 由下面的过程产生：

1. 从 Dirichlet 分布 $Dir(\boldsymbol{\alpha})$ 中随机产生一个 K 维的向量 $\boldsymbol{\theta}_d$，表示网页被标注的标签 d 中的主题混合比例；

2. 对网页被标注的每个标签 t_i：

2.1. 从多项式分布 $Multinomial(\boldsymbol{\theta}_d)$ 中产生一个主题 z_j；

2.2. 从标签在所有主题上的分布 $\boldsymbol{\beta}$ 在主题 z_j 下的分布中产生 t_i.

其中，$\boldsymbol{\alpha}$ 是一个 K 维的 Dirichlet 参数，概率密度

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha})=\frac{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}{\prod_{i=1}^{K}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\theta_2^{\alpha_2-1}\cdots\theta_K^{\alpha_K-1} \tag{1}$$

$\boldsymbol{\beta}$ 是一个 $|T|\times K$ 的二维矩阵，其元素值为 $\beta_{i,j}=P(t_i=1|z_j=1)$. 图 2 是网页主题模型的图模型表示.

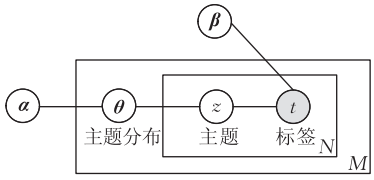


图 2 网页主题模型的图模型表示

本文使用基于 Variational Bayesian 推理的 EM 算法[13]来训练主题模型，估计模型参数 $\hat{\boldsymbol{\theta}}$ 和 $\hat{\boldsymbol{\beta}}$，并可以计算

$$\begin{aligned}P(t|d)&=P_{\mathrm{LDA}}(t|d)=P(t|\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\beta}})\\&=\sum_{i=1}^{K}P(t|z_i,\hat{\boldsymbol{\beta}})P(z_i|\hat{\boldsymbol{\theta}})\end{aligned} \tag{2}$$

对于用户，可以同样地利用用户使用的标签定义和训练用户的主题模型.

3.1.2 检索模型

在查询时，假设查询者从一个能够满足信息需求的理想网页中抽取词汇形成查询 Q，它可以看作是能够表征该理想网页的词序列. 由此检索任务被转化为查找网页集合中与理想网页最接近的网页模型 d. 为此，需要估计后验概率 $P(d|Q)$. 假设先验概

率 $P(d)$均匀分布，根据贝叶斯公式，有

$$P(d|Q)=\frac{P(Q|d)P(d)}{P(Q)}\propto P(Q|d)P(d)\propto P(Q|d) \tag{3}$$

所以只需计算查询似然 $P(Q|d)$来确定查询和网页间的相关性. 一般假设对给定的网页模型，查询中的查询词相互独立，故有

$$P(Q|d)=\prod_{t\in Q}P(t|d) \tag{4}$$

这样最终把对查询似然 $P(Q|d)$的计算转换为对 $P(t|d)$的估计.

估计 $P(t|d)$的最简单方法是极大似然估计(Maximum Likelihood Estimation，MLE). 令 $c(t,d)$表示网页 d 被标签 t 标注的次数，有

$$P(t|d)=P_{\mathrm{ML}}(t|d)=\frac{c(t,d)}{\sum_{t\in d}c(t,d)} \tag{5}$$

由于数据稀疏性，使用 MLE 会导致 $P(t|d)=0$ 的情况发生，因此需要对模型进行平滑. 根据 Jelinek-Mercer 平滑方法[15]，有

$$P(t|d)=(1-\lambda_{\mathrm{JM}})P_{\mathrm{ML}}(t|d)+\lambda_{\mathrm{JM}}P(t|D) \tag{6}$$

式中 λ_{JM}为独立于网页的平滑参数，$P(t|D)$为网页集 D 的语言模型，可以用 MLE 估计.

实际上，使用网页和用户的主题模型，可以使用式(2)直接估计 $P(t|d)$，即

$$P(t|d)=P_{\mathrm{LDA}}(t|d) \tag{7}$$

同时，Wei 和 Croft 指出，结合主题模型对语言模型进行平滑可以有效地提高信息检索的性能[16]. 本文采用与之类似的方法进行平滑：

$$\begin{aligned}P(t|d)=&(1-\lambda_{\mathrm{TM}})((1-\lambda_{\mathrm{JM}})P_{\mathrm{ML}}(t|d)+\\&\lambda_{\mathrm{JM}}P(t|D))+\lambda_{\mathrm{TM}}P_{\mathrm{LDA}}(t|d)\end{aligned} \tag{8}$$

式中 λ_{TM}为平滑参数，$P_{\mathrm{LDA}}(t|d)$和 $P_{\mathrm{ML}}(t|d)$分别由式(2)和式(5)给出. 实际上，该式有机地融合了查询词在单独网页、网页集合以及隐含主题中的似然分布，因此能够获得很好的检索性能.

对于用户，可以使用同样的方法计算依赖于用户主题模型的用户模型 u 的查询似然 $P(Q|u)$.

3.2　网页和用户的互增强模型

社会性标注系统是一种用户驱动的资源共享平台，用户基于备忘存储、知识管理或社会激励等不同原因而标注其兴趣领域内的网页. 内容质量较高的网页会吸引更多的用户对其进行标注；而资深的系统用户则会标注更多的网页. 因此，有理由认为，网页和用户之间存在着一种互增强关系：被高素质用户标注的网页具有高价值，而标注高价值网页的用户也具有高素质. 利用这种关系，可以设计出高效的排序算法.

3.2.1　互增强关系的模型

用一个如图 1(b)所示的二部图模型来刻画网页和用户之间的互增强关系. 令 $G=(U\cup D,\boldsymbol{W})$为表示网页和用户之间的互增强关系的边带权二部图. 对于任意用户 u 和网页 d，如果用户 u 曾经标注过网页 d，则 $W^{UD}_{u,d}\neq 0, W^{DU}_{d,u}\neq 0$；否则 $W^{UD}_{u,d}=0, W^{DU}_{d,u}=0$. 可以认为，$W^{UD}_{u,d}$和 $W^{DU}_{d,u}$分别表示从 u 到 d 和从 d 到 u 的转移概率，由于从一个顶点转移到其他顶点的概率和为 1，故需满足归一化条件

$$\sum_{d}W^{UD}_{u,d}=1,\quad \sum_{u}W^{DU}_{d,u}=1 \tag{9}$$

实际上，在该模型中如果只考虑两个顶点集中的一个，而将另一个视为隐含状态集合，那么该模型也可转化为在单一顶点集上的互增强关系模型. 例如，对于顶点集D，可以计算转移概率$W^{D}_{d_i,d_j}=\sum_{u}W^{DU}_{d_i,u}W^{UD}_{u,d_j}$，同时有

$$\begin{aligned}\sum_{d_j}W^{D}_{d_i,d_j}&=\sum_{d_j}\sum_{u}W^{DU}_{d_i,u}W^{UD}_{u,d_j}=\sum_{u}(W^{DU}_{d_i,u}\sum_{d_j}W^{UD}_{u,d_j})\\&=\sum_{u}W^{DU}_{d_i,u}=1\end{aligned} \tag{10}$$

故转移概率符合归一化条件.

3.2.2　互增强关系的权重

必须注意到，同一用户和不同网页以及不同用户和同一网页之间的互增强关系强弱并不相同. 例如，某数学家对有关微积分的网页的标注具有较强的质量评价意义，但其对音乐网站的标注则不尽然. 对于一个给定的网页 d，可以认为所有用户对其标注的标签是其内容在大众智慧中的抽象. 这一抽象在网页主题模型中表示为标签词汇 T 在该网页上的条件分布. 对标签词汇 T 中的标签 t，可以使用式(2)来得到 $P_{\mathrm{LDA}}(t|d)$，并用同样的方法得到 $P_{\mathrm{LDA}}(t|u)$. 设所有用户对网页 d 使用的标签集合为 $\{t_1,t_2,\cdots,t_{N_d}\}$，而用户 u 对网页 d 使用的标签集合为$\{t_{i_1},t_{i_2},\cdots,t_{i_M}\}(M\leq N_d)$. 受到信息检索评价指标滑动系数[17]的启发，本文用标注滑动系数来计算该标注与用户兴趣和网页内容的匹配度 $S^{UD}_{u,d}$和 $S^{DU}_{d,u}$：

$$S^{UD}_{u,d}=\frac{\sum_{k=1}^{M}P_{\mathrm{LDA}}(t_{i_k}|u)}{\sum_{k=1}^{M}P_{\mathrm{LDA}}(t_{s_k}|u)},\quad S^{DU}_{d,u}=\frac{\sum_{k=1}^{M}P_{\mathrm{LDA}}(t_{i_k}|d)}{\sum_{k=1}^{M}P_{\mathrm{LDA}}(t_{s_k}|d)} \tag{11}$$

式中$[s_1,s_2,\cdots,s_{N_d}]$为在$\{1,2,\cdots,N_d\}$上的一个特定排列，满足 $P_{\mathrm{LDA}}(t_{s_1}|\cdot)\geq P_{\mathrm{LDA}}(t_{s_2}|\cdot)\geq\cdots\geq P_{\mathrm{LDA}}(t_{s_{N_d}}|\cdot)$. 本文采用标注滑动系数作为评价匹配

度的指标，主要基于以下 3 点原因：(1) 使用排名靠前(即与用户兴趣和网页内容较一致)的标签能够获得较大的匹配度；(2) 使用较多的标签并不能一定获得较大的匹配度，因为排序算法并不鼓励滥用标签(或制造垃圾标注)；(3) 使用标签的顺序不应该影响匹配度. 对 $S^{UD}_{u,d}$ 和 $S^{DU}_{d,u}$ 归一化，得到二部图 G 中边的权重 $\boldsymbol{W}$，有

$$W^{UD}_{u,d}=\frac{S^{UD}_{u,d}}{\sum_{d}S^{UD}_{u,d}},\quad W^{DU}_{d,u}=\frac{S^{DU}_{d,u}}{\sum_{u}S^{DU}_{d,u}} \tag{12}$$

这样，通过网页和用户的语言模型以及用户标注网页时使用的标签，可以获知该标注与用户兴趣和网页内容之间的匹配度，从而量化互增强关系.

3.3 网页排序算法描述

结合上述的查询模型和互增强关系，通过在二部图上迭代传播网页和用户的评分来计算排序. 首先根据查询似然为网页和用户赋予一个初始评分，然后基于互增强关系迭代的传播网页和用户的评分. 对于一个给定的查询 Q，归一化每个网页 d 和用户 u 的查询似然得到其初始评分 p_d 和 q_u 来体现查询相关性在评分计算中所起到的作用：

$$p_d=\frac{P(Q|d)}{\sum_{d}P(Q|d)},\quad q_u=\frac{P(Q|u)}{\sum_{u}P(Q|u)} \tag{13}$$

分别使用两个系数 $\lambda^{UD},\lambda^{DU}\in(0,1)$ 来控制在计算评分时网页和用户互增强关系所占的比重，结果互增强关系，用下式计算网页 d 和用户 u 的评分：

$$r_d^{(i+1)}=\lambda^{UD}\sum_{u}s_u^{(i)}W^{UD}_{u,d}+(1-\lambda^{UD})p_d \tag{14}$$

$$s_u^{(i+1)}=\lambda^{DU}\sum_{d}r_d^{(i+1)}W^{DU}_{d,u}+(1-\lambda^{DU})q_u \tag{15}$$

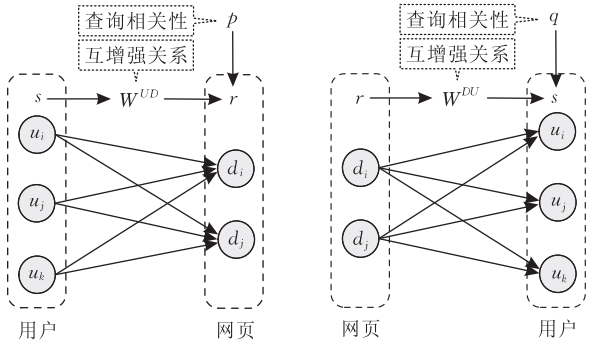
式中 $r_d^{(i)}$ 和 $s_u^{(i)}$ 分别表示网页 d 和用户 u 在第 i 次迭代时的评分. 式(14)和式(15)也可以写成矩阵形式：

$$\boldsymbol{r}^{(i+1)}=\lambda^{UD}\boldsymbol{s}^{(i)}\boldsymbol{W}^{UD}+(1-\lambda^{UD})\boldsymbol{p} \tag{16}$$

$$\boldsymbol{s}^{(i+1)}=\lambda^{DU}\boldsymbol{r}^{(i+1)}\boldsymbol{W}^{DU}+(1-\lambda^{DU})\boldsymbol{q} \tag{17}$$

式中 $\boldsymbol{r}\equiv[r_d]_{1\times|D|}$，$\boldsymbol{s}\equiv[s_u]_{1\times|U|}$，$\boldsymbol{p}\equiv[p_d]_{1\times|D|}$，$\boldsymbol{q}\equiv[q_u]_{1\times|U|}$. 网页和用户的评分计算过程如图 3 所示. 在算法实现中，该迭代过程终止于：(1) 评分向量在两次迭代中的值 $\boldsymbol{r}^{(i+1)}$ 和 $\boldsymbol{r}^{(i)}$ 满足 $\|\boldsymbol{r}^{(i+1)}-\boldsymbol{r}^{(i)}\|_2/\|\boldsymbol{r}^{(i)}\|_2\leq\theta$，其中 θ 为预先指定的阈值(在实验中取 $\theta=0.001$)；(2) 或迭代次数大于预先指定的阈值 $k_{\max}$(在实验中取 $k_{\max}=100$).

从图 3 中可以看出，排序算法的计算评分的过程中，有机地融合了网页和用户之间的互增强关系($\boldsymbol{W}^{UD}$ 和 $\boldsymbol{W}^{DU}$)和查询相关性($\boldsymbol{p}$ 和 $\boldsymbol{q}$)这两个影响网页排序的因素，从而使得具有较高质量且和当前查询相关的网页获得较高评分.



(a) 用户到网页的评分传播　　(b) 网页到用户的评分传播

图 3 网页和用户评分的计算过程

3.4 网页排序算法分析

本小节分析本文提出的网页排序算法收敛性和时间复杂性等特性.

3.4.1 收敛性分析

考虑网页的评分向量 $\boldsymbol{r}$. 将式(17)代入式(16)中，有

$$\begin{aligned}\boldsymbol{r}^{(i+1)}&=\lambda^{UD}(\lambda^{DU}\boldsymbol{r}^{(i)}\boldsymbol{W}^{DU}+(1-\lambda^{DU})\boldsymbol{q})\boldsymbol{W}^{UD}+(1-\lambda^{UD})\boldsymbol{p},\\&=\lambda^{DU}\lambda^{UD}\boldsymbol{r}^{(i)}\boldsymbol{W}^{DU}\boldsymbol{W}^{UD}+\lambda^{UD}(1-\lambda^{DU})\boldsymbol{q}\boldsymbol{W}^{UD}+(1-\lambda^{UD})\boldsymbol{p}\end{aligned} \tag{18}$$

令 $\lambda^D=\lambda^{DU}\lambda^{UD}\in(0,1)$，$\boldsymbol{W}^D=\boldsymbol{W}^{DU}\boldsymbol{W}^{UD}$，$\boldsymbol{p}^D=\lambda^{UD}(1-\lambda^{DU})\boldsymbol{q}\boldsymbol{W}^{UD}+(1-\lambda^{UD})\boldsymbol{p}$，则可将式(18)化为

$$\boldsymbol{r}^{(i+1)}=\lambda^D\boldsymbol{r}^{(i)}\boldsymbol{W}^D+\boldsymbol{p}^D \tag{19}$$

迭代式(19)，并取极限，有

$$\boldsymbol{r}^{(\infty)}=\boldsymbol{r}^{(0)}\lim_{n\to\infty}(\lambda^D\boldsymbol{W}^D)^n+\boldsymbol{p}^D\lim_{n\to\infty}\sum_{i=1}^{n}(\lambda^D\boldsymbol{W}^D)^{i-1} \tag{20}$$

对 $\boldsymbol{r}^{(\infty)}$ 的第一个分量，考虑

$$\begin{aligned}\sum_{d_j}(\lambda^D\boldsymbol{W}^D)^n_{d_i,d_j}&=\sum_{d_j}\sum_{d_k}(\lambda^D\boldsymbol{W}^D)^{n-1}_{d_i,d_k}(\lambda^D\boldsymbol{W}^D)_{d_k,d_j}\\&=\sum_{d_k}(\lambda^D\boldsymbol{W}^D)^{n-1}_{d_i,d_k}(\lambda^D\sum_{d_j}\boldsymbol{W}^D_{d_k,d_j})\end{aligned} \tag{21}$$

由式(10)可知 $\sum_{d_j}\boldsymbol{W}^D_{d_k,d_j}=1$，故

$$\sum_{d_j}(\lambda^D\boldsymbol{W}^D)^n_{d_i,d_j}=\lambda^D\sum_{d_k}(\lambda^D\boldsymbol{W}^D)^{n-1}_{d_i,d_k} \tag{22}$$

对于 $\lambda^D\in(0,1)$，存在 $\gamma\in(0,1)$，满足 $\lambda^D\leq\gamma$，则有

$$\sum_{d_j}(\lambda^D\boldsymbol{W}^D)^n_{d_i,d_j}\leq\gamma\sum_{d_k}(\lambda^D\boldsymbol{W}^D)^{n-1}_{d_i,d_k} \tag{23}$$

迭代式(23)，有

$$\sum_{d_j}(\lambda^D\boldsymbol{W}^D)^n_{d_i,d_j}\leq\gamma^n \tag{24}$$

故

$$\lim_{n\to\infty}\sum_{d_j}(\lambda^D\boldsymbol{W}^D)^n_{d_i,d_j}=0 \tag{25}$$

可知

$$\lim_{n\to\infty}(\lambda^D \boldsymbol{W}^D)^n=0 \tag{26}$$

即式(20)中 $\boldsymbol{r}^{(\infty)}$ 的第 1 个分量为 $\boldsymbol{0}$(零矩阵). 根据式(20),计算 $\boldsymbol{r}^{(\infty)}$ 的第 2 个分量,有

$$\boldsymbol{r}^{(\infty)}=\boldsymbol{p}^D\lim_{n\to\infty}\sum_{i=1}^{n}(\lambda^D\boldsymbol{W}^D)^{i-1}=\boldsymbol{p}^D(\boldsymbol{I}-\lambda^D\boldsymbol{W}^D)^{-1} \tag{27}$$

式中 $\boldsymbol{I}$ 为单位矩阵. 式(27)即是式(16)的收敛解. 式(17)的收敛性同理可证.

3.4.2　时间复杂性分析

该算法的迭代计算过程的时间复杂度为 $O(n(|U|L^U+|D|L^D))$. 其中 n 为迭代次数,L^U 和 L^D 分别为用户标注网页次数的均值和网页被标注次数的均值. 在实际计算时,算法收敛的迭代次数一般较小,故算法的运行时间主要依赖于网页和用户数量. 一种有效的提高运行效率的方法是,在计算时首先将网页和用户按照其查询相关性排序,然后按照性能需求,选择排序靠前的若干网页和用户继续进行迭代计算.

4　实　验

为了评价算法性能,在实际的社会性标注系统 Delicious 中采集标注数据进行实验. 在 6 个月的时间内,通过采集网页,抽取相关信息,得到包含有 367782 个网页,54707 个用户,86937 个标签以及 13876825 个标注的标注数据集.

4.1　主题模型选择

在 LDA 中,主题服从 Dirichlet 分布,该分布假设一个主题的出现与其它主题的出现无关. 在真实数据中,很多主题之间是存在关联的. 这种独立假设与真实数据的矛盾使得 LDA 对于主题数目 K 非常敏感. 本文通过分析不同主题数目对困惑度(Perplexity)的影响来确定最优的主题数目[13,18-20]:

$$Perplexity(D_{\text{test}})=\exp\left(-\frac{\sum_{d\in D_{\text{test}}}\ln P_{\text{LDA}}(d)}{\sum_{d\in D_{\text{test}}}N_d}\right).$$

式中 D_{test} 为测试网页标注集,N_d 为网页 d 被标注的标签数,$P_{\text{LDA}}(d)$ 是待测试的模型产生测试网页标注 d 的概率,

$$P_{\text{LDA}}(d)=\prod_{t\in d}P(t|\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\beta}}),$$

式中 $\hat{\boldsymbol{\theta}}$ 和 $\hat{\boldsymbol{\beta}}$ 为待测试模型的后验参数估计. 对于给定的模型,困惑度越小表明模型越具有推广性.

在网页标注数据集中随机选择 90% 作为训练数据,10% 作为测试数据来测试不同主题数目对困惑度的影响. 从图 4 中的实验结果可以看出,当主题数目 $K>30$ 时,继续增加主题数目对困惑度的改善效果已经很小. 对用户标注数据也有类似的结果. 在下面的实验中,采用主题数目 $K=30$.

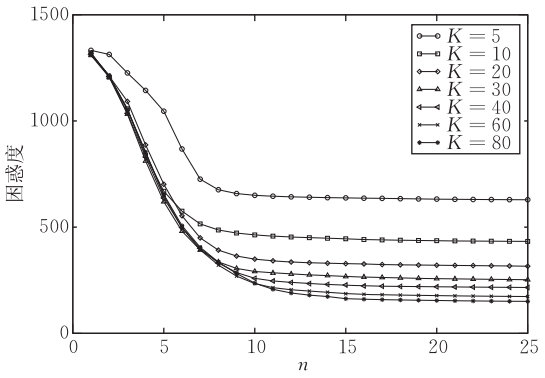


图 4　在不同迭代次数 n 时不同主题模型的困惑度

4.2　排序结果评价方法

由于人工标记标准检索结果十分困难,本文采用 Open Directory Project①(ODP)来自动生成检索结果评测集. ODP 以分类树的形式组织网页,树中每个节点均有一个分类标签及其相应的网页. 本文采用下面的方法来生成查询及其相应的检索结果. 首先随机选择一个分类路径,将路径中的每个部分作为一个查询词(除了"Top")来生成查询. 例如,对于分类路径"Computers /Artificial_Intelligence /Machine_Learning"将生成查询"computers+artificial+intelligence+machine+learning". 然后将该路径下的网页集合与 Delicious 中的网页集合的交集作为标准检索结果集合. 按照此方法,共生成了 1000 个查询以及 6474 个相关网页,平均查询长度为 6.976.

本文采用两个广泛使用的评价指标 MAP(Mean Average Precision)和 NDCG(Normalized Discounted Cumulative Gain)对算法进行评价. 对一个查询 Q,如果第 i 个结果网页相关则 $r_i=1$,否则 $r_i=0$. MAP 为所有查询的平均准确率(AP)的均值,

$$AP_Q=\frac{1}{N_R}\sum_{i=1}^{N_L}\frac{r_i}{i}\sum_{j=1}^{i}r_j,$$

式中 N_R 为相关网页数,N_L 为检索结果网页数,实验中取 $N_L=100$. $NDCG$ 为排名靠前的检索结果赋予较大的权重,因此适合作为 Web 检索结果的评价

① http://dmoz.org/

指标，

$$NDCG_Q=\frac{1}{I_Q}\sum_{i=1}^{N}\frac{2^{r_i}-1}{\log(i+1)},$$

式中 I_Q 为查询相关的归一化常数，以使理想排序结果的 $NDCG$ 值为 1，N 为待评价的检索结果数.

4.3 查询模型比较

在计算查询词的相关性 $P(t|d)$ 时，可以采用基于 MLE 的 Jelinek-Mercer 平滑方法(MLE，式(6)，取 $\lambda_{JM}=0.7$)，或者单独采用 LDA 模型(LDA，式(7))，或者同时结合主题模型和 MLE 的平滑方法(LDA+MLE，式(8)，取 $\lambda_{JM}=0.7$，$\lambda_{TM}=0.5$). 为了验证本文提出的主题模型及其平滑方法的有效性，基于相同的互增强模型(见第 3.2 节，在式(14)中取 $\lambda^{UD}=\lambda^{DU}=0.8$)，表 1 和图 5 对不同查询模型的性能进行了比较.

从表 1 和图 5 中可以看出，本文采用的同时结合主题模型和 MLE 的平滑方法获得了最好的性能. 对比 MLE 和 LDA，由于后者通过对隐含主题的学习，提取可以被人理解的、相对稳定的隐含语义结构，从而可以有效地检索到内容相关但并不直接包含查询词的文档，提高了检索性能. 但是，由于在主题模型中，每个词语表示为其在不同主题上的分布. 这种表示在有限的主题数目显得相对粗糙. 因此，对比 LDA+MLE 和上述两种查询模型的性能，可以看出，通过结合 LDA 和 MLE，有机融合查询词在不同隐含主题、不同文档和整个文档集中的分布，有效地提高了检索性能.

表 1 不同查询模型的 MAP 结果比较

查询模型	MAP	比较 MLE/%
MLE	.1009	—
LDA	.1194	+18.3
LDA+MLE	**.1314**	**+30.2**

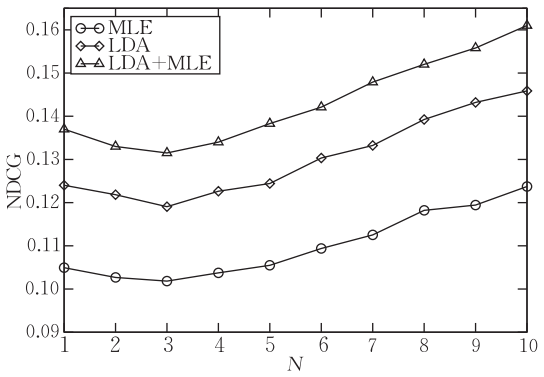


图 5 不同查询模型的 NDCG 结果比较

4.4 互增强模型比较

如何合理表示网页和用户之间的互增强关系也是本文研究的重点. 和以往的研究中对互增强关系权重采用的：(1)平均赋值[7,8](Uniform)，即为同一用户对不同网页和不同用户对同一网页的标注赋予相同的权重；(2)按标注时间赋值[8](Time，见第 2 节)，即按照用户标注网页的时间先后赋予不同的权重等方法不同，本文结合网页和用户的主题模型，采用了标注滑动系数来量化特定标注与用户兴趣和网页内容的匹配度(Sliding Ratio). 为了验证本文采用的方法的有效性，基于相同的查询模型(LDA+MLE)和参数设置(在式(14)中取 $\lambda^{UD}=\lambda^{DU}=0.8$)，表 2 和图 6 对不同互增强模型的性能进行了比较.

从表 2 和图 6 中可以看出，本文采用的基于标注滑动系数方法获得了最好的性能. 对比 Uniform 和 Time 可以看出，相比于不加区分的平均分配，通过区分网页的发现者(discoverer)和跟随者(follower)，可以合理量化网页和用户之间的互增强关系，从而提高检索性能. 然而，用户标注网页的时间先后可能取决于浏览时间、标注习惯，甚至地理位置(不同时区)等因素，并不一定能够直接反应用户对网页的质量评价的权威程度. 另一方面，如在第 3.2 节中所述，标注与用户兴趣和网页内容的匹配度是衡量标注权威性的有效标准. 对比 Sliding Ratio 和上述两种互增强模型的性能，可以看出，利用表征标注与用户兴趣和网页内容的匹配度的标注滑动系数来反映该标注的权威程度，能够有效的提高检索性能.

表 2 不同互增强模型的 MAP 结果比较

互增强模型	MAP	比较 Uniform/%
Uniform	.1181	—
Time	.1294	+9.6
Sliding Ratio	**.1314**	**+11.3**

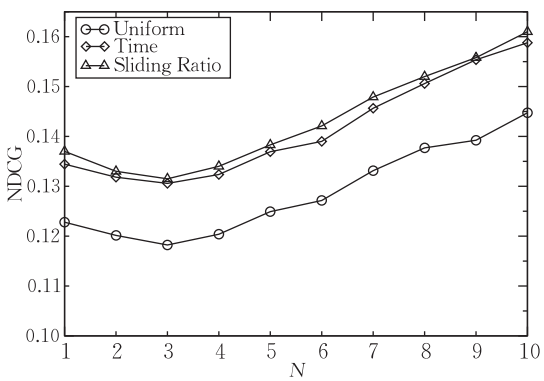


图 6 不同互增强模型的 NDCG 结果比较

4.5 不同算法比较

下面通过比较本文提出的算法(以下称 CoRank)和另外两个代表性算法——SocialPageRank 算法[8]

和 SPEAR 算法[9]，对算法性能进行评价. 对本文提出的算法，在计算网页和用户的查询似然时，在式(8)中取 $\lambda_{JM}=0.7$，$\lambda_{TM}=0.5$；在迭代计算时，在式(14)中取 $\lambda^{UD}=\lambda^{DU}=0.8$. 对 SocialPageRank 和 SPEAR，因其均为查询无关的排序算法，本文通过利用查询词构造相关标注集合，然后进行排序的方法得到查询结果. 另外，设置 Baseline 算法，将网页按照其被查询词标注的次数排序.

表 3 和图 7 分别为 MAP 和 NDCG 的实验结果比较. 从中可以看出，所有的算法的性能均比 Baseline 有明显提高. 这说明在社会性标注系统中，简单的使用标注频率进行排序是不能获得令人满意的效果的.

对比 SocialPageRank 和 SPEAR，后者比前者有明显优势. 这两者的主要区别是，前者基于三部图模型，而后者则基于二部图模型. 由于 SocialPageRank 将标签评分也包括在迭代计算过程中，使得被重要标签标注的低质量网页获得提升. 这种情况产生的影响在垃圾标注日益增多的社会性标注系统中非常明显. SPEAR 则没有在迭代计算过程中涉及标签的重要性，而只利用了网页和用户间的互增强关系，并根据用户发现网页的先后顺序来为网页和用户间的互增强关系赋予权重，从而获得了较好的排序结果.

对比同样基于二部图模型的 SPEAR 和 CoRank，后者获得的优势主要在于更好地结合了查询相关性和互增强关系. 由于 SPEAR 仅仅考虑不同用户标注网页的时间顺序，而忽略了用户标注网页时使用的标签信息，使得它在计算排序时无法有效地融合查询相关性，进而影响了排序结果的质量. 而本文提出的 CoRank 算法则利用语言模型将标签之间的隐含的语义关系引入到查询相关性中，并将其作为排序初始值有机地融合到排序算法中. 同时，CoRank 还利用标注时使用的标签与用户兴趣和网页内容的匹配度来为用户与网页之间的互增强关系赋予权重，使标签中蕴含的语义信息得到了合理利用，从而产生较好的排序结果.

表 3　不同算法的 MAP 结果比较

算法	MAP	比较 Baseline/%
Baseline	.0983	—
SocialPageRank	.1104	+12.3
SPEAR	.1233	+25.4
CoRank	**.1314**	**+33.7**

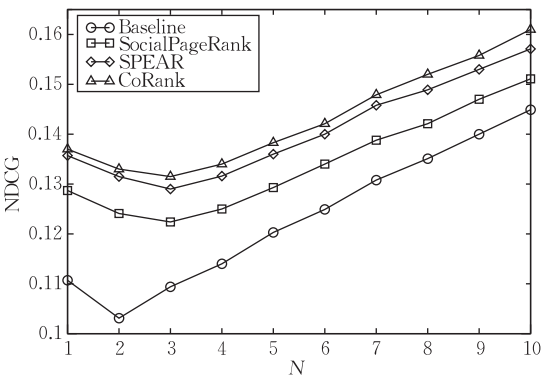


图 7　不同算法的 NDCG 结果比较

为了更加深入地观察不同算法产生的排序结果，表 4 给出了对查询“programming＋javascript＋ajax”，不同算法给出的排名前 10 位的结果，其中相关结果用粗体表示. 从表中可以看出，基于标注次数的 Baseline 方法，将较多的不相关网页排在前面，效果较差. 在 SocialPageRank、SPEAR 和 CoRank 3 种算法中，SocialPageRank 效果较差，将 3 个不相关结果排到了前 10 位. 而 SPEAR 和 CoRank 的排序效果较好，都没有或极少将不相关结果排到前面. 但对比 SPEAR 和 CoRank，在具体的排序质量上，CoRank 有明显优势，因其产生的排名靠前的网页结果均为著名的 Javascript 和 Ajax 项目或技术网站.

总的来说，本文提出的算法在不同的评价指标上均比其他算法有较大提高. 这表明，通过有机结合刻画用户兴趣和网页内容的主题模型和刻画网页和用户间互增强关系的互增强模型，能够有效地提高网页排序结果的质量.

表 4　查询“programming＋javascript＋ajax”的排序结果

Baseline	SocialPageRank	SPEAR	CoRank
wordpress.org	**script.aculo.us**	**script.aculo.us**	**script.aculo.us**
www.pandora.com	**www.go2web20.net**	**www.go2web20.net**	**www.go2web20.net**
www.last.fm	**jquery.com**	**jquery.com**	**jquery.com**
www.w3schools.com	www.alvit.de/handbook	**code.google.com**	**prototype.conio.net**
www.cssbeauty.com	**code.google.com**	**dojotoolkit.org**	**www.djangoproject.com**
digg.com	www.oswd.org	**prototype.conio.net**	**extjs.com**
www.technorati.com	kuler.adobe.com	**www.djangoproject.com**	**dojotoolkit.org**
code.google.com	www.colourlovers.com	www.w3schools.com	**developer.yahoo.com/yui**
www.oswd.org	www.instructables.com	**www.ajaxload.info**	**code.google.com**
www.csszengarden.com	**extjs.com**	processing.org	**www.ajaxload.info**

5 结 论

本文研究了基于社会性标注数据的网页排序算法,以改进网页检索性能.利用相关标签,基于主题模型定义了网页和用户的语言模型,并对查询模型进行平滑,计算查询相关性.使用二部图模型刻画网页和用户之间的互增强关系并对其进行了量化.最后结合查询模型和互增强模型计算网页迭代计算网页评分.实验结果表明,该算法比目前的代表性算法在性能上有较大提高.作为一种算法框架,抽象出语言模型、查询相关性的计算、互增强关系量化等细节,本文提出的算法可以很容易地应用到其他类型的共享资源检索中,具有良好的可拓展性.

参 考 文 献

[1] Page L et al. The pagerank citation ranking: Bringing order to the web. Stanford University, Stanford, CA, USA: Technical Report 1999-66, 1999

[2] Kleinberg J M. Authoritative sources in a hyperlinked environment. Journal of the ACM, 1999, 46(5): 604-632

[3] Koutrika G et al. Combating spam in tagging systems//Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb'07). Banff, Canada, 2007: 57-64

[4] Koutrika G et al. Combating spam in tagging systems: An evaluation. ACM Transactions on the Web, 2008, 2(4): 1-34

[5] Heymann P, Koutrika G, Garcia-Molina H. Fighting spam on social web sites: A survey of approaches and future challenges. IEEE Internet Computing, 2007, 11(6): 36-45

[6] Krause B et al. The anti-social tagger: Detecting spam in social bookmarking systems//Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web(AIRWeb'08). Beijing, China, 2008: 61-68

[7] Hotho A et al. Information retrieval in folksonomies: Search and ranking. The Semantic Web: Research and Applications, 2006, 4011: 411-426

[8] Bao S et al. Optimizing web search using social annotations//Proceedings of the 16th International Conference on World Wide Web(WWW'07). Banff, Canada, 2007: 501-510

[9] Noll M G et al. Telling experts from spammers: Expertise ranking in folksonomies//Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR'09). Boston, MA, USA, 2009: 612-619

[10] Hofmann T. Probabilistic latent semantic indexing//Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR'99). Berkeley, CA, USA, 1999: 50-57

[11] Griffiths T L, Steyvers M. Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(Supplement 1): 5228-5235

[12] Griffiths T, Steyvers M. Prediction and semantic association//Becker S et al. eds. Advances in Neural Information Processing Systems 15. Boston, MA, USA: MIT Press, 2003: 11-18

[13] Blei D M et al. Latent dirichlet allocation. Journal of Machine Learning Research, 2003, 3: 993-1022

[14] Steyvers M, Griffiths T. Probabilistic topic models//Landauer T et al eds. Handbook of Latent Semantic Analysis. Mahwah, NJ, USA: Lawrence Erlbaum Associates, 2007: 424-440

[15] Jelinek F, Mercer R L. Interpolated estimation of Markov source parameters from sparse data//Gelsema E S, Kanal L N eds. Pattern Recognition in Practice. Amsterdam, the Netherlands: Elsevier B V. 1980: 381-402

[16] Wei X, Croft W B. Lda-based document models for ad-hoc retrieval//Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'06). Seattle, WA, USA, 2006: 178-185

[16] Pollack S M. Measures for the comparison of information retrieval systems. American Documentation, 1968, 19(4): 387-397

[17] Rosen-Zvi M et al. The author-topic model for authors and documents//Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI'04). Banff, Canada, 2004: 487-494

[18] Wu X, Zhang L, Yu Y. Exploring social annotations for the semantic web//Proceedings of the 15th International Conference on World Wide Web (WWW'06). Edinburgh, Scotland, 2006: 417-426

[19] Zhou D et al. Exploring social annotations for information retrieval//Proceedings of the 17th International Conference on World Wide Web(WWW'08). Beijing, China, 2008: 715-724

LIU Kai-Peng, born in 1981, Ph. D. candidate. His current research interests include information retrieval, data mining and machine learning.

FANG Bin-Xing, born in 1960, Ph. D., professor, member of Chinese Academy of Engineering. His current research interests include information security, information retrieval and distributed systems.

Background

With the rapid development of Web 2.0, social tagging systems became highly popular in recent years. These systems allow collaborative users to submit shared resources and to annotate them with descriptive tags, forming the so-called folksonomies. The rapidly increasing popularity of social tagging systems and growing amount of users and resources make it a difficult task to find expert users and relevant resources in folksonomies. In this paper, the authors focus on improving search performance in folksonomies by developing a dynamic ranking algorithm.

Previous studies on ranking in folksonomies usually adopted the tripartite graph model of social annotations. Such methods consider tags as an equal part with users and resources, and compute scores for all of them in a symmetric way. Though they are easy to understand and effective to compute, ranking algorithms based on tripartite graph model may suffer from some problems. Tags are no more than a piece of descriptive metadata; they themselves are not directly related to the quality of resources: the use of important tags does not mean the related users and resources are also important. In addition, malicious users often post spam annotations to the system with popular tags to attract more attention from other users. Therefore, algorithms taking tag importance into account are more susceptible to spam annotations. In fact, as they serve as the semantic entities connecting users and resources, tags are the best source to understand user interest and resource content. Thus, the authors adopted a bipartite graph model of social annotations by removing tags from the tripartite graph model, moreover leveraged the tag information to model user interest and resource content. To this end, the authors adopted the methodology of statistical language modeling and developed a probabilistic generative model to demonstrate the tagging scheme of users and resources. Based on the query likelihood derived from this model, the authors built their algorithm upon the mutual reinforcement between users and resources, and also noticed that the mutual reinforcing relations are not equally important and assign each one of them with a weight according to the coherence between the annotating tags and the corresponding user and resource model.

This work was partially supported by the National Natural Science Foundation of China under grant No. 60703014 and No. 60933005, the National Basic Research Program of China (973 Program) of China under grant No. G2007CB311100 and the National High Technology Research and Development Program (863 Program) of China under grant No. 2006AA010105-02, No. 2007AA01Z416, No. 2007AA01Z442 and No. 2009AA01Z437. These projects aim to study the mechanisms within a virtual computing environment and build a Web information retrieval and data mining system with an emphasis on network monitoring.