融合主题与语言模型的个性化标签推荐方法研究

李 慧[1,2] 马小平[1] 胡 云[2,3] 施 珺[2]
(中国矿业大学信息与电气工程学院 徐州 221116)[1] (淮海工学院计算机工程学院 连云港 222005)[2]
(南京大学计算机科学与技术系 南京 210093)[3]

摘 要 随着Web的推广和普及,产生了越来越多的网络数据。广泛应用了标签系统,以便人们使用搜索技术来组织和使用这些信息。这些数据允许用户使用关键字(标签)注释资源,为传统的基于文本的信息检索提供了方案。为了支持用户选择正确的关键字,标签推荐算法应运而生。提出了一种个性化标签推荐方法,该方法综合了用户的资源标签与标签概率模型。该模型利用了简单语言模型和隐含狄利克雷分配模型,并针对现实世界的大型数据集进行了大量实验。实验表明,该个性化方法改进了标签推荐算法,推荐结果优于传统方法。

关键词 标签,推荐,主题,潜在主题模型,个性化
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.8.015

Personalized Tag Recommendation Algorithm Mixing Language Model and Topic

LI Hui[1,2] MA Xiao-ping[1] HU Yun[2,3] SHI Jun[2]
(School of Information & Electrical Engineering, China University of Mining and Technology, Xuzhou 221116, China)[1]
(School of Computer Engineering, Huaihai Institute of Technology, Lianyungang 222005, China)[2]
(Department of Computer Science and Technology, Nanjing University, Nanjing 210093, China)[3]

Abstract More and more content on the Web is generated by users. To organize this information and make it accessible via current search technology, tagging systems have gained tremendous popularity. We introduced an approach to personalized tag recommendation that combines a probabilistic model of tags from the resource with tags from the user. In this models, we investigated simple language models as well as Latent Dirichlet Allocation. Extensive experiments on a real world dataset crawled from a big tagging system show that personalization improves tag recommendation, and our approach significantly outperforms traditional approaches.

Keywords Tag, Recommendation, Topic, Latent topic models, Personality

1 引言

随着Web 2.0技术的发展,越来越多的网络用户能自由选择标签(字词或短语)来标注网络资源,形成了社会化标签。社会化标签允许用户用自造的标签来标注网络资源,对标签的内容、个数和一致性均无限制。用户能够使用任意的词汇也就是标签对自己喜欢的资源进行标注,使用户能够便捷地分享和有效地组织这些标签信息。由此,用户既是信息的接受者,又是信息的发布者。正是由于社会化标签的简单高效性,社会标签网站变得越来越流行,如Flickr、LastFm、YouTube、Delicious、豆瓣网等著名网站都采用了这种方式。但是,被标注过的文档在整个网络空间占据的比例甚小,而网络资源的迅速增长会严重影响利用标签进行网页搜索等应用的服务性能,为此需要更广泛地对网络资源进行标注。但是,由于网络资源非常巨大,人工标注费时费力,并且人工标注中也存在大量错标、标注不一致等问题。因此,利用计算机进行自动、高质量的关键短语抽取和标签推荐成为一个现实的选择。当用户浏览某个产品时,标签推荐系统给出一些相关的标签以便于他能够更好地标注此产品。正是如此,标签推荐系统获得了越来越多的关注。

目前对标签推荐算法的研究主要分为两种,一种是将算法直接进行改进,将其设计成能处理三维关系的算法和模型,如Zhang Zike[1]提出的超图模型等;另一种是将社会标签的三维关系转换为二维关系,直接应用传统推荐系统模型,如基于概率扩展的PLSA方法[2,3]、由Cohn and Hofman[4]提出的基于内容和基于链接相结合的统一框架模型以及当前应用较为广泛的由Peter Mika[5]提出的三部图模型,这些模型的主要目的是将标签系统三维关系转换为二维关系。

本文结合了以用户为中心的推荐算法和以资源为中心的推荐算法来为用户推荐个性化标签。为了实现这个目标,本文使用简单语言模型(LM)和隐含狄利克雷分配模型(LDA)来评估新标签正确的概率。应用LDA的潜在优势会产生一些新的、用户从未使用过的标签,从而增加了标签推荐时可用

到稿日期:2014-05-25 返修日期:2014-07-29 本文受国家自然科学基金(61403156,61403155),江苏省高校自然科学基金(14KJB520005),江苏省海洋资源开发研究院开放项目(JSIMR201323)资助。

李 慧(1979—),女,硕士,讲师,主要研究领域为数据挖掘,E-mail:shufanzs@126.com;马小平(1961—),男,博士,教授,主要研究领域为智能控制;胡 云(1978—),女,博士,副教授,主要研究领域为社会网络分析;施 珺(1963—),女,硕士,教授,主要研究领域为智能信息化。

的词汇量。结合资源和用户的优势可以根据个人标签喜好过滤一些常见的资源标签。

2 相关工作

2.1 社会化标注系统

近年来，人们对社会化标注系统有着较为深入的研究。Wetzker 等提出一种以用户为中心的标签模型(UCTM)[6]，该算法采用 3 阶张量来模拟用户、标签和资源之间的关系，将个性化标签集映射到相应被标注资源的大众分类方法，UCTM 方法可以应用在基于标签的推荐和标签推荐等方面。Gemmell[7]等提出一种线性加权混合算法来进行基于标签的资源推荐，并在 6 个不同数据集上进行实验分析。文献[8]提出一种综合协同过滤和基于内容的标签的推荐技术，前者是利用用户和社会化标注行为来产生推荐，而后者是利用一些启发式的方式直接从资源的文本内容提取标签。文献[9]提出一种采用图模型进行个性化标签的推荐方法，将用户、标签和资源 3 者之间的关系转换成一个三元无向图，对图中相邻顶点的处理采用一种综合的权重衡量方法，对不相邻顶点之间的关系采用最短路径思想得出。文献[10]以传统的张量分解模型为基础，通过数据缺失值处理，进行局部最优求解获得标签推荐值，以此进行社会标签系统中标签的推荐预测。文献[11]提出用节点的拓扑潜力来表征用户的社交影响力，基于这个指标，可以区分不同用户之间的社会关系，并找出哪些用户对目标用户有真正的影响，这里的标签推荐是基于标注历史和社会化标注网络中影响力最大用户的潜在个性化偏好。

2.2 聚类

标签系统存在一个普遍的问题——稀疏性。为了解决稀疏性问题，大量的以降低稀疏标签空间的维度为目的的聚类算法被陆续提出。

Symeonidis 等人在文献[12]中采用降维实现个性化推荐标签。首先提出将能完整地表示高维数据并且能维持高维空间数据的本征结构信息的 tensor(张量)应用于社会标签系统中，并且利用张量分解方法进行标签预测。将用户-资源-标签(URT)图对应地用 3 个矩阵来表示，对于每个矩阵分别应用奇异值分解，然后将分解后的矩阵再次组合从而构建与原始 URT 图等价的更密集的标签空间。如果标签的权重大于某个预设阈值，则推荐推荐就将该标签推荐给用户。Schmidt-Thieme 在文献[13]中提出了两个更有效的方法，即规范化分解与张量分解。

当资源的内容可用时，标签推荐也可以看成是一个分类问题，只是将其从内容预测转换为标签预测。Song 等人在文献[14]中提出了在这个研究方向上的一个新方法，即在应用一定的降维技术后，对文件-项-标签进行聚类，然后获得了一个对标签进行排序的簇，通过将新资源进行分类，将其归为某一簇后再使用该簇中包含的标签进行推荐。

2.3 LDA 的标签推荐

近几年来，基于隐含主题模型(Latent Topic Models)的方法广泛用于文档内容建模。隐含主题模型中应用最广泛的是 Bid 等人提出的隐含狄利克雷分配模型(Latent Dirichlet Allocation，LDA)及各种在此基础上发展起来的主题模型。LDA 模型是一种无监督的概率生成模型，该模型认为文档中的每个词是由某个隐含主题抽样生成的。通过训练，可以学习到每篇文档的主题分布以及主题在词空间上的分布。

由于 LDA 对文档建模的能力，近年来出现了越来越多的基于 LDA 的应用，比如主题发现、文档分类和信息检索等方面。LDA 也被用于社会化标签推荐的研究中，Si 和 Sun 在文献[15]中介绍了一种使用 LDA 的集合标签推荐技术。Xiance 等人使用 LDA 从文档(或博文)和相关的标签中抽取出主题。在此基础上，他们对于博客中出现的新文档推荐使用新的标签。krestel 等人利用 LDA 从可用资源的标签中推断主题，然后资源从这些隐主题中推荐额外的标签。本文扩展了那些个人标签信息的个性化标签推荐的推荐方法，实现将简单语言模型与隐含狄利克雷分配模型相合的标签推荐算法，从而使推荐精度获得了显著提高。Bundschus 介绍了一个组合内容、资源标签与用户的 LDA 方法[16]。推荐的基本过程是从资源的标签出发，引出用户特定的主题内容。基于内容的主题和基于标签的主题与用户 ID 一一对应。个性化标签推荐的基础是从资源的内容上识别出用户特定的主题，然后使用对应的基于标签的主题实现标签推荐。

3 混合语言模型与主题模型的个性化标签推荐

本节提出了一种结合用户与资源的标签推荐算法，该方法克服了单独运用用户标签推荐或资源标签推荐算法的缺点。一方面，法综合考虑了用户自身兴趣与对标签的使用偏好；另一方面，算法可以在特定资源下识别最合适的标签。

3.1 目标和方法

标签系统允许用户使用关键字来注释资源。标签推荐旨在帮助用户完成资源的标注。当用户标注一个新资源时，系统会为用户推荐合适的标签来减轻用户构思和录入关键字的负担。由于用户多数会选择系统推荐的标签而不是构思新的标签，因此影响了资源的标签分布。故推荐算法是推荐系统的重要组成部分。下面给出标签推荐任务的正式定义。

给定一组资源 R、标签 T、用户 U，则三元关系 $X\subseteq R\times T\times U$ 代表用户为指定资源分配标签。书签 $b(r,u)$ 表示用户 $u\in U$ 对资源 $r\in R$ 指定的所有标签集合。个性化标签推荐的目的是从书签 $b(r,u)$ 中推荐标签给用户来协助用户标识一个新的资源，从而减少用户在标记资源时的认知负荷量。这可以通过为该资源分配其它标注信息或相似资源标签来实现，甚至可以使用相似用户的标签进行推荐。

总而言之，我们的任务是在给定资源和用户的前提下对可能的标签 T 进行排序。使用基于概率的方法进行标签提名。更确切地说，在给定资源 r 和用户 u 后，为其推荐标签 t 的概率 $P(t|u,r)$ 的计算公式如下：

$$P(t|r,u)=\frac{P(r,u|t)P(t)}{P(r,u)} \tag{1}$$

$$P(t|r,u)\approx\frac{P(r|t)P(u|t)P(t)}{P(r,u)} \tag{2}$$

$$P(t|r,u)=\frac{P(t|r)P(r)}{P(t)}\frac{P(t|u)P(u)}{P(t)}\frac{P(t)}{P(r,u)} \tag{3}$$

$$P(t|r,u)=\frac{P(t|r)P(t|u)}{P(t)}\frac{P(r)P(u)}{P(r,u)} \tag{4}$$

$$P(t|r,u)\propto\frac{P(t|r)P(t|u)}{P(t)} \tag{5}$$

式(1)应用了贝叶斯法则；式(2)假设条件 t 下 r 与 u 相互独立，从而将 $P(r,u|t)$ 进行分解；式(3)对 $P(r|t)$ 与 $P(u|t)$ 再次应用贝叶斯规则；式(4)对式(3)进行约简；式(5)忽略了

对所有标签都相同的因素 $P(r)$，$P(u)$和 $P(r,u)$。

$P(t)$可以对标签在书签集合中出现的频率进行估算。我们调研和结合了两种方法来计算 $P(t|r)$和 $P(t|u)$的值。一种方法是使用简单的语言模型(3.2 节)，另一种是利用潜在狄利克雷分配方法(3.3 节)，该方法能对新的资源和用户以及存在少数有效的标签进行推荐。

对于 $P(t|r)$和 $P(t|u)$的估计，在式(5)中它们被赋予了相同的权重。然而，通常对于特定的用户 u 会比资源 r 存在更多的可用标签。因此应该为 $P(t|u)$赋予比 $P(t|r)$更高的权重。最终，我们利用先验概率 $P(t)$对 $P(t|r)$与 $P(t|u)$进行了平滑：

$$P'(t|r)\propto\log_2(|r|+1)P(t|r)+\log_2(|u|+1)P(t) \quad (6)$$

$$P'(t|u)\propto\log_2(|u|+1)P(t|u)+\log_2(|r|+1)P(t) \quad (7)$$

其中，$|r|$表示资源 r 可用的标签数目，$|u|$表示用户 u 可用的标签数目。当$|r|$小于$|u|$时，$P(t|r)$的值被增大，从而导致 $P(t|r,u)$小于 $P(t|u)$。注意到当 $P(t|r)$的值为 0 时，$P'(t|r)$与 $P(t)$成正比，从而组合概率 $P'(t|r)*P'(t|u)/P(t)$与 $P'(t|u)$也成正比。同样地，当某个资源没有任何标签时，$\log_2(|r|+1)=0$，并且组合概率与 $P'(t|u)$成反比。

3.2 简单语言模型

标签推荐最简单的方法是为每个资源只推荐最常见的标签。但是由于社会化标签的随意性，标签集中会存在一些无效标签：包括多个标签表示同一含义的冗余标签，以及在训练集中出现频率非常少的冷僻标签。这些无效标签的存在严重影响了标签推荐的准确性和高效性，降低了标签推荐的质量。因此，可以基于标签相似度对原始标签集进行筛选和提纯处理，从而去除标签集中冗余和冷僻的标签。

本文使用文本特征加权 TF-IDF 方法计算标签之间的相似度。用 $rel(r,t)$表示待推荐资源 r 与标签 t 的相关系统，用 $n(r,t)$表示标签集合 T 中的一个标签被用于标注资源 r 的次数，R 表示已经被标注的资源集合，$TF(r,t)$表示这个标注次数在所有标注资源的次数中的比例，$IDF(r,t)$表示标签 t 出现的频率。$rel(r,t)$的计算方法如下：

$$TF(r,t)=\frac{n(r,t)}{\sum_{T_i\in T}n(r,T_i)} \quad (8)$$

$$IDF(t)=\frac{\sum_{R_j\in R}\sum_{T_i\in T}n(R_j,T_i)}{\sum_{R_j\in R}n(R_j,t)} \quad (9)$$

$$rel(r,t)=TF(r,t)\times IDF(r) \quad (10)$$

$rel(r,t)$的值越大，说明待推荐资源 r 与标签 t 的关系越紧密。用 $rel(r,T_i)$和 $rel(r,T_j)$表示标签 T_i 与 T_j 标注同一资源 r 的相关系数。若它们的相关系数差值越小，说明它们在标注资源 r 时表现的差异性越小，即两个标签越相似[17]。用 $R(t)$表示被标签 t 标注的资源集合，$n(R)$表示资源集 R 的总数目，$S(T_i,T_j)$表示标签 T_i 与 T_j 之间的相似度，其计算方法如下：

$$S(T_i,T_j)=\frac{\sum_{R_i\in R}(1-|\frac{rel(R_i,T_i)-rel(R_i,T_j)}{rel(R_i,T_i)+rel(R_i,T_j)}|)}{n(R(T_i)\cup R(T_j))-n(R(T_i)\cap R(T_j))} \quad (11)$$

根据标签之间的相似度可以进行冗余标签的筛选。方法是通过预设阈值或选取标签相似度较大的标签进行去除。当两个标签的相似度大于预设阈值时，即可认为它们是相同或意义相近的标签，需要从原始标签集中筛选。根据标签的 IDF 值还可以进行冷僻标签的筛选。当某标签的出现概率小于预设阈值或过小时，即可认为该标签是冷僻标签，也需要从原始标签集中被删除。

经过从原始标签集中去除了相似标签和冷僻标签等无效标签后，再将每个资源中最常见的标签推荐给用户。即给定资源 r，为其推荐标签 t 的概率为：

$$P_{bm}(t|r)=\frac{c(t,r)}{\sum_{t_i\in r}c(t_i,r)} \quad (12)$$

其中，$c(t,r)$表示资源 r 中标签 t 的数量。用户 u 使用标签 t 的概率 $P_{bm}(t|u)$与上式定义相似。

3.3 隐含狄利克雷分配模型

对于那些只有少量标签分配的新的资源和用户，简单语言模型无法实现有效的标签推荐，因为用户想用来标注的标签可能并不存在于资源标签词汇中。为了解决这个问题，可以使用隐含狄利克雷分配模型(LDA)进行基于主题的标签推荐。LDA 主要是基于衍生模型而提出的，即当用户在寻找资源时，他首先选择一个与资源相关的一个主题，然后从该主题中选择一个标签。

LDA 方法的处理过程可以描述为：对每一个资源 r 寻找一个组合主题 z，如 $P(z|r)$，用标签 t 描述的满足另一个概率分布的主题，如 $P(t|z)$。可用如下公式表示：

$$P_{ida}(t|r)=\sum_{z=1}^{Z}P(t|z)P(z|r) \quad (13)$$

其中，$P_{ida}(t|r)$表示在给定资源 r 和资源涉及的潜在主题 z 的前提下，推荐标签 t 的概率，$P(t|z)$表示从主题 z 中推荐标签 t 的概率(见式(15))，$P(z|r)$表示从主题 z 中选取标签 r 的概率(见式(16))。潜在主题集合 z 中的主题数量必须预先定义且允许调整。

LDA 方法从一个使用狄含利克雷先验分布的无标签文档集计算主题-标签分布 $P(t|z)$与资源-标签分布 $P(z|r)$。吉布斯采样[18]是一种可行的方法：它将资源 r 的每一个标签 t_i 多次迭代，使用式(14)中基于概率 P 为标签选择一个新的主题 z，直到 LDA 模型中的参数达到收敛。

$$P(z|t_i,r,z_{-i})\propto(C_{rz}^{RZ}+\alpha)\frac{C_{t_iz}^{TZ}+\beta}{\sum_{T}C_{tz}^{TZ}+T\beta} \quad (14)$$

其中，C^{TZ}表示分配的所有主题-标签的数量，C^{RZ}表示分配的资源-主题的数量，z_{-i}表示除了当前为标签 t_i 分配的主题 z 以外的所有主题-标签和资源-标签。α 和 β 是狄利克雷先验参数，作为计数的平滑参数。

式(13)中基于计数的后验概率可用式(15)与式(16)得到，即：

$$P(t|z)=\frac{C_{tz}^{TZ}+\beta}{\sum_{t_i}C_{t_iz}^{TZ}+T\beta} \quad (15)$$

$$P(z|r)=\frac{C_{rz}^{RZ}+\alpha}{\sum_{z_i}C_{rz_i}^{RZ}+Z\alpha} \quad (16)$$

用同样的方法可以求得 $P_{ida}(t|u)$，即将操作对象由原来的资源标签集合改为用户标签集合即可。

3.4 结合 LDA 和 LM

由于 $P_{bm}(t|r)$和 $P_{ida}(t|r)$都属于标准化的概率分布，可以直接应用线性插值法将这两个概率组合起来($P(t|u)$的计算也是如此)：

$$P(t|r)=\lambda \cdot P_{lm}(t|r)+(1-\lambda) \cdot P_{lda}(t|r) \quad (17)$$

本文通过实验验证了 λ 的取值，结果表明当 λ 在范围[0.2,0.8]之间进行取值时系统会取得较好结果。

4 实验评估

为了验证本文提出的基于主题的个性化标签推算算法的有效性，将其(记为“LM+LDA 方法”)与标签推荐领域中的主流的 FolkRank 算法[19]和 UCTM 算法[20]进行对比。FolkRank(FR)算法是一种随机游走推荐算法。该推荐算法可以获得很高的推荐精准度，但是计算代价太高。与本文方法相比，FolkRank 算法没有利用潜在主题信息，它是基于图论理论的。实现 LM+LDA 方法的过程中大量采用了 Phan 等[11]提供的使用 Java 语言编写的利用吉布斯抽样进行 LDA 参数估计的工具 JGibbLDA 的源代码。实验硬件平台为3.17GHz Core 2 Duo 处理器。

本文构建了测试集来测试算法的性能。测试集的构建方法是将一定数量的标签数大于 8 的用户的标签删去，用本文提出的方法来给这些删除标签的用户打上标签，最后将算法得到的标签与用户自己标记的原始标签进行比较。下面分别介绍实验所使用的数据集与评价指标。

4.1 数据集

本文选用 Delicious 作为实验数据集进行测试。Delicious 是一个帮助用户共享他们喜欢网站链接的流行网站。该数据集由 Delicious 用户的多种多样的 url 标签组成。对于 Delicious 网站的数据集，采用 Wetzker 等人在文献[20]中提供的数据集，该数据集由 2007 年 12 月至 2008 年 4 月的将近 100 万个用户组成，检索处理表明在 2003 年 11 月至 2007 年 12 月分配了大约 132000000 条书签或者是 420000000 个标签。大约 700 万条不同的标签都包含在这个数据集内，该数据集里还包括大约 55000000 条 url。

出于内存和时间方面的考虑，本文仅使用了整个数据集中的部分数据。经过观察发现，数据集中的标签、资源、用户等数据存在少量的重叠。为了获得稠密子集，本文计算了 Batagelj 和 Zaversnik 在文献[21]中提出的在不同级别上分别计算的 p-cores 值。

对于 $p=20$ 的情形，本文从每一个资源中抽取了定量的书签将它们拆分为训练集和测试数据集(比例为 9∶1)。20-core 保证每个标签、每个资源、每个用户都在分配过程中至少出现 20 次，对于测试集中 10%的数据，只将前 n 个用户提交的资源放到训练集中(其中 $n\in 1,3,5,7,10,20$)，这与现实情况相符。当一个资源被多个用户标注时，该资源就具有一个稳定的标签分布，从而降低了推荐系统的推荐难度。

4.2 评估指标

MRR(Mean reciprocal rank)——平均倒数排名，是一个国际上通用的对搜索算法进行评价的机制，即第一个结果匹配分数为 1，第二个匹配分数为 0.5，第 n 个匹配分数为 $1/n$，没有匹配的句子分数为 0。最终的分数为所有得分之和。

F-Measure 又称为 F-Score，是 IR(信息检索)领域常用的一个评价标准。

召回率(recall)表示推荐列表中相对于用户测试集中实际跟踪对象的命中数与用户测试集中实际跟踪的对象数之比，该值越高，代表系统性能越好。对于测试集中每一个用户资源对(u,i)，这里都会推荐 N 个标签给用户作参考。令 $T(u,i)$为给用户 u 推荐的应该在物品 i 上标记的标签集合，里面包含用户可能会对资源标记的标签集合。$R(u,i)$是测试集中用户 u 实际给物品 i 标记的标签集合。召回率的定义如下：

$$recall=\frac{\sum_{(u,i)\in Test}|R(u,i)\cap T(u,i)|}{\sum_{(u,i)\in Test}|T(u,i)|}$$

5 实验结果

图 1 显示当书签数量在 1 到 20 之间比较各种方法的平均倒数的排名(MRR)结果。图中“FR”代表 FolkRank 算法，“LM”代表简单语言模型，“LDA+LM”代表本文提出的混合简单语言模型与隐含狄利克雷分配模型的标签推荐算法。结果表明本文提出的组合算法明显优于其他两种方法，但值得注意的是当书签数超过 7 时，LM 算法的性能优于 FolkRank；并且当书签数超过 20 时所有方法的 MRR 有所降低的原因是因为实验的配置问题。当书签数目达到 20 时，导致可用测试数据变少，从而使 MRR 值下降。

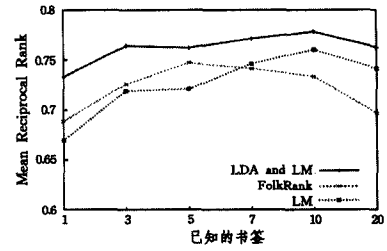

图 1 不同数量书签的平均倒数排名结果

图 2 显示了随着推荐标签个数的增加，F-measure 的变化趋势。由图可知，当推荐标签数量为 3 时，所有的推荐算法在查全率和查准率之间达到了最佳平衡。这也表明用户对某一资源平均给出 4.3 个标签。由图 2 可知，本文提出的方法在 F-measure 作为评估指标时也明显优于 FolkRank 和平滑的简单语言模型。

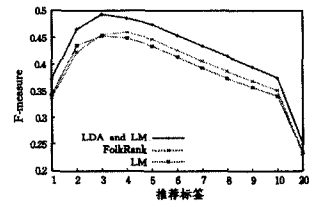

图 2 不同推荐标签数量下的 F-measure

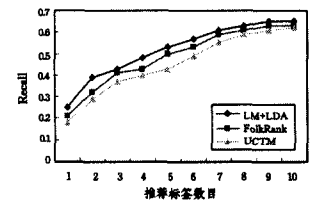

图 3 3 种算法召回率的对比结果

图 3 示出本文提出的 LM+LDA 方法和 FolkRank 算法与 UCTM 算法在算法召回率性能上的对比结果。实验过程是将数据集按 9∶1 的比例分为训练集和测试集，通过训练集学习用户的标注模型，对于测试集中的每个用户资源对(u,r)，都会推荐 N 个标签给用户 u 作参考。将上面的实验进行 10 次，对比每次实验中 3 种方法的召回率结果。由图 3 可知，在不同的推荐标签数目下，本文提出的 LM+LDA 方法相对于其他两种方法的召回率有明显提高。

为了直观地比较不同标签推荐方法与用户实际使用的标签的差异，从数据集中随机挑选择一个用户，将用户实际使用的标签与推荐系统推荐的标签进行对比，表 1 给出了一个运

用 FolkRank 和本文介绍的方法实现标签推荐的结果。表中用粗体表示正确预测的标签。结果表明，本文提出的方法在前 6 个标签中正确地预测出 4 个标签，并在前 20 个中正确预测了 6 个，而 FolkRank 在前 20 个中只预测了 4 个。

表 1　推荐系统与用户实际标签的对比结果

用户最初的标签	FolkRank		LDA&LM	
	标签	分数	标签	分数
Webdev 的编程参考 Web2.0 网页 XHTML 格式的教程	**Microformats**	0.0138	**Microformats**	50.6
	Howto	0.0078	Howto	15.1
	Standards	0.0070	**Tutorial**	12.8
	Tutorial	0.0068	Standards	11.5
	Collection	0.0066	**Programming**	9.6
	Information	0.0064	**Reference**	8.4
	Resources	0.0061	Semantic	7.2
	Tutorials	0.0060	Development	5.3
	Webdev	0.0024	Software	4.0
	Tool	0.0021	Web	3.4
	Development	0.0020	xml	3.2
	Programming	0.0018	**Webdesign**	3.0
	Web	0.0013	Code	2.7
	html	0.0011	Tool	2.4
	Code	0.0009	**Webdev**	2.3
	Software	0.0007	Information	2.3
	Javascript	0.0006	css	2.0
	Python	0.0005	Design	2.0
	Snippets	0.0004	Tips	2.0
	Optimization	0.0004	Tutorials	1.6

结束语　在当今以用户贡献内容为核心的社交网络中，标签成为用户对资源进行标记和分类的重要依据。在社会网络中，用户可以自由地给自己打上标签以表明自己的兴趣和特征等，用户标签在舆情分析与监测、广告推送和网络营销等应用中起到了非常重要的作用。本文探讨了以用户为中心、以资源为中心的个性化标签推荐方法，比较并且使用了一个语言模型方法和一个基于狄利克雷分配的方法。研究表明，组合简单语言模型与隐含狄利克雷分配方法(LDA 和 LM)在各种性能指标上优于随机游走标签推荐算法。下一步的研究工作可以将该方法应用于基于上下文的标签推荐，如可以利用时间、位置等上下文进行更加个性化的标签推荐。

参 考 文 献

[1] Zhang Z, Liu C. A Hypergraph model of social tagging networks [J]. Journal of Statistical Mechanics: Theory and Experiment, 2010(10): 1-14

[2] Aleksandra K M, Alexandros N, Mirjana I. Social tagging in recommender systems: a survey of the state-of-the-art and possible extensions[J]. Artificial Intelligence Review, 2010, 33(1): 187-209

[3] Cai Y Z, Zhang M, Chris H D. Closed form solution of similarity algorithms[C]// Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2010: 709-710

[4] Cohn D, Hofmann T. The missing link: a probabilistic model of document content and hypertext connectivity[M]// Leen T K, Dietterich T G, Tresp V, eds. NIPS, MIT Press, 2000: 430-436

[5] Peter M. Ontologies are us: a unified model of social networks and semantics[J]. Web Semantics: Science, Services and Agents on World Wide Web Archive, 2007, 5(1): 5-15

[6] Wetzker Z. I tag, you tag: Translating tags for advanced user models[C]// Proceedings of 3rd ACM International Conference Web Search and Data Mining. New York: ACM, 2010: 71-80

[7] Gemmell, Schimoler. Tag based resource recommendation in social annotation applications[M]// User Modeling, Adaption and Personalization: Proceedings of the 6th European semantic web conference on the semantic Web: research and applications. Berlin: Springer-Verlag, 2011: 195-206

[8] Lops P, Gemmis M D, Semeraro G. Content-based and collaborative techniques for tag recommendation: an empirical evaluation [J]. Journal of Intelligent Information Systems, 2013, 40(1): 41-61

[9] Wu X L, Tu F H. Personalized tag recommendation method using graph-model[J]. Computer Engineering and Applications, 2013, 11(1): 1-6

[10] Liao Z F, Wang C Q, Li X Q. Tag recommendation and new user tag recommendation algorithms based on tensor decomposition [J]. Journal of Chinese Computer Systems, 2013, 34(11): 2473-2476

[11] Hu J, Wang B, Liu Y. Personalized Tag Recommendation Using Social influence [J]. Journal of Computer Science and Technology, 2012, 27(3): 527-540

[12] Symeonidis P, Nanopoulos A, Manolopoulos Y. Tag recommendations based on tensor dimensionality reduction[C]// Proceedings of the 2008 ACM Conference on Recommender Systems (RecSys'08). ACM, New York, NY, USA, 2008: 43-50

[13] Rendle S, Schmidt-Thieme L. Pairwise interaction tensor factorization for personalized tag recommendation[C]// Proceedings of the Third ACM International Conference on Web Search and Data Mining(WSDM'10). ACM, New York, 2010: 81-90

[14] Song Y, Zhuang Z, Li H, et al. Real-time automatic tag recommendation[C]// Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR'08). ACM, New York, 2008: 515-522

[15] Si X, Sun M. Tag-LDA for scalable real-time tag recommendation[J]. Journal of Computational Information Systems, 2009, 6(1): 23-31

[16] Bundschus M, Yu S, Tresp V, et al. Kriegel, Hierarchical Bayesian models for collaborative tagging systems[C]// Proceedings of the 2009 Ninth IEEE International Conference on Data Mining(ICDM'09). IEEE Computer Society, Washington DC, USA, 2009: 728-733

[17] 赵亚楠，董晶，董佳梁. 基于社会化标的博客标签推荐方法[J]. 计算机工程与设计，2012，33(12)：4609-4613
Zhao Ya-nan, Dong Jing, Dong Jia-liang. Tag recommendation for blogs based on social tagging[J]. Computer Engineering and Design, 2012, 33(12): 4609-4613

[18] H Heng-zhen, Y Jin-yu. Functionally induced priors for componentwise Gibbs sampler in the analysis of supersaturated designs [J]. Computational Statistics and Data Analysis, 2014, 72(1): 1-12

[19] Hotho, Jäschke, Schmitz. Information retrieval in folksonomies: Search and ranking[M]// The Semantic Web: Research and Applications: Proceedings of the 3rd European semantic web conference. Berlin: Springer-Verlag, 2006: 411-426

[20] Wetzker R, Zimmermann C, Bauckhage C. Analyzing social bookmarking systems: A del. icio. us cookbook[C]// Proceedings of ECAI 2008 Mining Social Data Workshop. 2008: 26-30

[21] Carpineto C, Romano G. Semantic search log k-anonymization with generalied k-cores of query concept graph[C]// 35th European Conference on IR Research(ECIR 2013). 2013: 110-121