

一种基于 Folksonomy 的信任预测方法

王树锋^{1,2}, 王 文^{1,2}

(1 常州工学院, 江苏 常州 213002; 2 常州市软件技术与应用重点实验室, 江苏 常州 213002)

摘 要: 社会学认为生活在同一个社会圈子中的用户表现出相似的行为和共同的偏好, 这些有别于信任图拓扑结构的信息也可以用于信任预测中. 由此提出了一种基于 Folksonomy 的信任预测方法 TPMF (Trust Prediction Method based on Folksonomy), 整合社会网中信任信息和 Folksonomy 中的用户资源信息和用户标记信息, 利用优化技术预测社会网络中的信任关系. 在真实数据集上的实验表明, 该方法具有很好的性能.

关键词: 信任预测; 社会网络; 社会标记; 非负矩阵分解方法

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2015)12-0136-05

A Trust Prediction Method Based on Folksonomy

WANG Shu-feng^{1,2}, WANG Wen^{1,2}

(1 Changzhou Institute Technology, Changzhou 213002, China; 2 Changzhou Key Laboratory of Software Technology and Application, Changzhou 213002, China)

Abstract: The people who are in the same social circle often exhibit similar behaviors and tastes in sociology reserach. The ancillary information that difference from trust graph topology can be used for trust prediction. In this article, we propose a novel Trust Prediction Method based on Folksonomy (TPMF). Our new joint method explores the trust information in trust domain and user-sources graph and user-tagged information in folksonomy. By optimize the proposed objective function, the folksonomy information can be utilized to enhance the prediction. The extensive experiments have been conducted on real world data. The empirical results demonstrate the effectiveness of our method.

Key words: trust prediction; social network; folksonomy; nonnegative matrix factorization

1 引言

近年来, 随着社会网络站点的出现并持续增长, 社会网的规模越来越大, 网络中用户之间交互产生了海量的信息. 在这样的环境中, 对可信用户的确定和可靠信息的识别将成为一个挑战性的问题.

许多在线社会网络允许对其他用户及其产生的内容进行简单的评价. 通常将社会网中的这种行为用图模型表示, 将社会网中信任标记转换为信任图 $G = \langle V, E \rangle$, 其中 V 表示用户, E 表示用户之间的信任关系. 事实上, 对一个用户而言, 社会网中只有很少的用户相关联, 能够进行信任标记的就更少了.

然而社会网中对用户产生的内容进行评价和标注要比对用户的可信度评价容易很多. 社会标注系统是使用户能够上传、标注和其他用户共享网络资源的 Web 应用程序, 通过 Folksonomy 这种信息结构, 可以协作创建、标注和分类用户产生的网络内容, 并对网络内容控制分散化, 将 Web 变成一个开放民主的环境. 因此社会标注系统中包含的内容信息能够用于用户之间的信任预测.

Folksonomy 是社会标记系统的底层结构, 定义为四元组: $F = (U, R, T, Y)$, 描述了来源于不同主域的两个异构的图, 具有不同的结构和规模. 利用这些信息, 本文提出了一种信任预测方法, 整合社会网

收稿日期: 2015-04-23; 修回日期: 2015-06-09

基金项目: 常州工学院自然科学基金重点项目(YN1203, YN1316); 常州市应用基础计划研究项目(CJ20120009)

中目标信任主域的信任信息和 Folksonomy 中的用户资源信息与用户标记信息,学习两类信息中用户图结构信息,利用优化技术预测社会网络中的信任关系。

2 文献综述

信任预测是在社会网中用户之间没有交互的情况下预测这些用户的信任程度。根据采用的技术不同,信任预测方法可以分为四类,分别是利用信任扩散属性、利用链接预测、利用协同过滤方法和利用多模型联合进行信任预测。

LESKOVEC 等利用信任扩散属性研究了在线社会网中正面(友好)和负面(敌意)的关系,提出了一种连接预测模型,能够比较准确地预测社会网络中的连接^[1]。在文献[2]中,作者开发了信任扩散框架,允许根据每个个体信任/不信任的微小变化预测两个个体之间的信任关系并在大规模网络中对框架进行了评价。GETOOR, L. 等认为链接挖掘是在对链接数据构建预测和描述模型时,一种仅考虑链接信息的数据挖掘技术,任务主要包括目标排序、组检测、群智分类、链接预测和子图发现^[3]。文献[4]中把统计关系模型用于链接预测,采用高斯过程对网络中用关系随机变量表示的边进行建模,设计的学习算法能够高效地处理大规模数据。文献[5-6]中通过学习相似用户的偏好模型,利用矩阵因素分解方法找到用户项目矩阵的最小维度的近似矩阵,采用多种协作过滤方法解决信任预测问题。采用多模型联合进行信任预测的方法是将多个相关的模型联合起来,利用这些模型中共同具有的相似性结构,将知识从不同的任务间迁移,采用联合的关系矩阵因素分解法进行信任预测。文献[7]中研究了社会标记系统中多模型搜索方法,对用户社会标记系统中的用户、资源和标记进行检索,从而实现个性化标记推荐。

3 基于 Folksonomy 的信任预测方法

社会网中彼此信任的用户表现出相似的行为和偏好,同时社会网用户因为具有相似的背景和相同的兴趣而成为朋友。因此虽然信任图和用户标记图的表现形式不同,但是两者在结构上具有相似性,这样利用用户标记信息能够提高信任预测的准确度。本文构建的基于 Folksonomy 的信任预测方法

TPMF(Trust Prediction Method based on Folksonomy),根据用户信任图 and Folksonomy 中的用户资源图 and 用户标记图,设计集聚这些信息的目标函数,利用优化技术预测信任关系。

根据 Folksonomy 结构特征,假设用 $G_1 \in R^{n \times m_1}$ 、 $G_2 \in R^{n \times m_2}$ 和 $G_3 \in R^{n \times m_3}$ 分别表示信任图、Folksonomy 中的用户资源图 and 用户标记图。其中 n 是三个主域中相同的用户数, m_1 是接受信任投票的用户数, m_2 是资源数, m_3 是标记数。

3.1 构建目标函数

设计的目标函数是找到三种图中共享的组结构信息,因此将目标函数表示为

$$\min_{U, V_1, V_2, c_1, c_2} \|G_1 - UV_1^T\|_F^2 + \|c_1 G_2 - UV_2^T\|_F^2 + \|c_2 G_3 - UV_3^T\|_F^2 \quad (\text{目标函数 1})$$

式中, $U \in R^{n \times l}$ 、 $V_1 \in R^{m_1 \times l}$ 、 $V_2 \in R^{m_2 \times l}$ 和 $V_3 \in R^{m_3 \times l}$, l 是需要确定的量。 c_1 和 c_2 是调整因子。因为三个图分属于不同的主域,这两个因子用于调整图之间的不一致性。 U 由信任图、用户资源图 and 用户标记图共同决定,因此 U 提供了三种图的共享结构。行表示图中的用户,因此能根据 U 对用户分组,然后利用 V_1 、 V_2 和 V_3 引导信任预测。 U 包含信任图 and Folksonomy 中的知识。

目标函数 1 构成的模型仅考虑了信任图 and Folksonomy 两种度量中用户组的结构,没有考虑社会网络的约束。为了克服这种局限性,引入拉普拉斯规范化因子,将模型改进为

$$\begin{aligned} \min_{U, V_1, V_2, c_1, c_2} & \|G_1 - UV_1^T\|_F^2 + \|c_1 G_2 - UV_2^T\|_F^2 + \\ & \|c_2 G_3 - UV_3^T\|_F^2 + \lambda \text{Tr}(V_1^T L_1 V_1) + \\ & \lambda \text{Tr}(V_2^T L_2 V_2) + \lambda \text{Tr}(V_3^T L_3 V_3) \\ \text{s. t. } & V_1 V_1^T = I, V_2 V_2^T = I, V_3 V_3^T = I \\ & U \geq 0, V_1 \geq 0, V_2 \geq 0, V_3 \geq 0 \end{aligned} \quad (\text{目标函数 2})$$

式中, $\lambda > 0$ 是调整因子, L_1 、 L_2 和 L_3 是基于图 G_1 、 G_2 和 G_3 中列的数量的拉普拉斯图, Tr 是求迹操作,得到矩阵主对角线中的元素。

信任图 G_1 描述社会网中用户信任关系, Folksonomy 中的用户资源等级图 G_2 描述了用户对资源偏好的相似性,用户标记等级图 G_3 描述用户对标记偏好的相似性,这些信息用于学习用户之间的隐含关系。因此拉普拉斯图 L_1 、 L_2 和 L_3 表示用户之间显式的和隐含的社会关系。当预测用户间的信任关系时,这些已存在的社会关系得以保留,因此在模

型中添加了三个拉普拉斯规范化因子.

3.2 目标函数优化算法

优化目标函数 2 需要考虑 $U, V_1, V_2, V_3, c_1, c_2$, 为了简化优化过程, 采用分别优化各个参数直到目标函数收敛的算法.

3.2.1 初始化

一般地社会图中包含大量的缺损值, 在进行信任预测时, 首先需要应用相似性计算方法对这些值进行补充. 对应 G_{ij} 中任意的缺损值, 采用 G_{ij} 中行列平均值替代. 对于 Folksonomy 中资源和标记等级关系矩阵, 结合各个用户等级习惯和其他用户对特定资源、标记的等级信息进行初始化. 对于用户信任矩阵, 考虑用户 i 和用户 j 在社会图中的影响度进行初始化.

本文中边权重矩阵 W 定义如下:

$$W_{ij} = \begin{cases} 1; G_{i.} \in N_k(G_{.j}) \text{ or } G_{.j} \in N_k(G_{i.}) \\ 0; \text{otherwise} \end{cases},$$

式中, $N_k(G_{i.})$ 表示 $G_{i.}$ 中 k 个最近邻居. 通过计算图中每个用户之间的欧氏距离, 基于与每个用户最相似的 k 个用户构建对应的 W .

令拉普拉斯 $L = D - W$, 其中 D 是对角矩阵, 其值为 W 中列的和, $D_{ii} = \sum_j W_{ij}$. 对应信任图 G 和 Folksonomy 图 G_2 和 G_3 , 构建 L_1, L_2 和 L_3 .

基于 G_1, G_2 和 G_3 列的 k 均值构建 V_1, V_2 和 V_3 . 对于 V_1 第 i 行, 如果该行属于第 j 个类, $V_1(i, j) = 1$, 第 i 行的其他元素为 0. V_2, V_3 采用同样的方式初始化. 优化目标函数 2 采用固定其余变量优化其中一个变量的方法, 这个过程重复进行, 直到目标函数收敛为止.

3.2.2 优化 U

考虑 U , 将其余变量作为常量, 优化目标函数 2 等价于优化等式:

$$\begin{aligned} J_1 &= \|G_1 - UV_1^T\|_F^2 + \|c_1 G_2 - UV_2^T\|_F^2 + \\ &\|c_2 G_3 - UV_3^T\|_F^2 \\ \text{s. t. } V_i^T V_i &= I, V \geq 0, i = 1, 2, 3 \\ \text{令 } \frac{\partial J_1}{\partial U} &= 0, \text{ 得到} \\ U &= \frac{G_1 V_1 + c_1 G_2 V_2 + c_2 G_3 V_3}{2} \end{aligned} \quad (1)$$

3.2.3 优化 V_1

考虑 V_1 , 优化目标函数 2 等价于优化等式:

$$J_2 = \|G_1 - UV_1^T\|_F^2 + \lambda \text{Tr}(V_1^T L_1 V_1)$$

$$\text{s. t. } V_1^T V_1 = I, V_1 \geq 0$$

在 s. t. 约束下, 上式不能直接求出 V_1 , 需要采用重复多步升级算法, 引入拉格朗日乘法因子 $\alpha \in R^{l \times l}$, 对应的拉格朗日函数为

$$L(V_1) = \|G_1 - UV_1^T\|_F^2 + \lambda \text{Tr}(V_1^T L_1 V_1) - \text{Tr}(\alpha(V_1^T V_1 - I)).$$

在 s. t. 约束下, 令 $\frac{\partial L(V_1)}{\partial V_1} = 0$, 得到:

$$\alpha = -V_1^T G_1^T U + \lambda V_1^T L_1 V_1.$$

令 $\alpha \cdot V_1 = 0$, 即 $(-V_1^T G_1^T U + \lambda V_1^T L_1 V_1) \cdot V_1 = 0$. 引入 $L_1 = L_1^+ - L_1^-$, $V_1 = V_1^+ - V_1^-$, $U = U^+ - U^-$, 其中,

$$U_{ij}^+ = (|U_{ij}| + U_{ij})/2,$$

$$U_{ij}^- = (|U_{ij}| - U_{ij})/2.$$

用同样的方式定义 L_1 和 V_1 , 得到

$$G_1^T U^- + \lambda L_1^+ V_1 + V_1 \alpha^- - G_1^T U^+ - \lambda L_1^- V_1 - V_1 \alpha^+ \cdot V_1 = 0.$$

由上式可以得到如下的升级公式:

$$(V_1)_{ij} \leftarrow (V_1)_{ij} \sqrt{\frac{[G_1^T U^+ + \lambda L_1^- V_1 + V_1 \alpha^+]_{ij}}{[G_1^T U^- + \lambda L_1^+ V_1 + V_1 \alpha^-]_{ij}}} \quad (2)$$

3.2.4 优化 V_2, V_3

V_2 和 V_3 的计算过程与上述计算过程类似.

$$\beta = -c_1 V_2^T G_2^T U + \lambda V_2^T L_2 V_2$$

$$(V_2)_{ij} \leftarrow (V_2)_{ij} \sqrt{\frac{[c_1 G_2^T U^+ + \lambda L_2^- V_2 + V_2 \beta^+]_{ij}}{[c_1 G_2^T U^- + \lambda L_2^+ V_2 + V_2 \beta^-]_{ij}}} \quad (3)$$

$$\delta = -c_2 V_3^T G_3^T U + \lambda V_3^T L_3 V_3$$

$$(V_3)_{ij} \leftarrow (V_3)_{ij} \sqrt{\frac{[c_2 G_3^T U^+ + \lambda L_3^- V_3 + V_3 \delta^+]_{ij}}{[c_2 G_3^T U^- + \lambda L_3^+ V_3 + V_3 \delta^-]_{ij}}} \quad (4)$$

3.2.5 优化 c_1, c_2

考虑 c_1 , 目标函数 2 等价于

$$J_4 = \|c_1 G_2 - UV_2^T\|_F^2.$$

上式等价于 $\min_{c_1} \text{Tr}(c_1 G_2 - UV_2^T)(c_1 G_2 - UV_2^T)^T$.

可以写为

$$\min_{c_1} AC_1^2 - 2BC_1 + D.$$

式中, $A = \text{Tr}(G_2 G_2^T)$, $B = \text{Tr}(UV_2^T G_2^T)$ 和 $D = \text{Tr}(UV_2^T V_2 U^T)$.

解方程:

$$c_1 = \frac{Tr(UV_2^T G_2^T)}{Tr(G_2 G_2^T)} \quad (5)$$

采用同样的方法过程得到:

$$c_2 = \frac{Tr(UV_3^T G_3^T)}{Tr(G_3 G_3^T)} \quad (6)$$

对目标函数 2 采用重复多步升级算法,如图 1 所示。目标函数不是凸函数,因此不能保证算法全局收敛,然而目标函数的正交约束保证算法能够得到唯一解。

```

算法 TPMF
输入:  $G_1$ 、 $G_2$ 和 $G_3$ , 最大重复次数;
输出: 收敛的 $U$ 、 $V_1$ 、 $V_2$ 和 $V_3$ 
利用行列平均值初始化 $G_1$ 、 $G_2$ 和 $G_3$ 中的缺损值;
利用 $k$ 平均聚类初始化 $V_1$ 、 $V_2$ 和 $V_3$ 。
根据图之间可区分度初始化 $c_1$ 和 $c_2$ 。
While (未收敛且重复次数小于 $T$ ) do
{
按照式(1)计算 $U$ ;
按照式(2)计算 $V_1$ ;
按照式(3)计算 $V_2$ ;
按照式(4)计算 $V_3$ ;
按照式(5)计算 $c_1$ ;
按照式(6)计算 $c_2$ ;
}
Enddo

```

图 1 基于 Folksonomy 的信任预测 TPMF 算法

4 实验与结果评价

Epinions^[8]数据集包括等级和信任两部分内容。数据集中包含 49 290 个用户,139 738 个项目,664 824 段用户对项目的评论,187 181 条用户之间信任评价语句。利用这些信息构建了信任图、用户资源图。通过统计用户项目的评价和用户之间的信任评价语句数据中出现单词的词频,确定一个阈值,将出现频率高于阈值的词汇作为标记用词,计算这些标记用词在每个用户评论中出现的次数作为用户标记等级数据,得到一个用户标记等级数据作为用户标记图。

本文将基于 Folksonoly 的方法和 k 最近邻居方法(KNN)、SimRank 和 SVD 近似三种方法在上述步骤构造的数据集上进行信任预测的性能比较。

在实验过程中,选出发出和接收投票数最多的前 3 000 名用户,选择与这些用户相关的超过 68 个等级的项目,得到 $3\,000 \times 3\,000$ 的信任图 G_1 ,包含 249 256 信任等级,用 1 表示,大约占有所有信任等级的 5.25%,所有不信任 and 不知道的等级用 0 表示。用户项目等级图 G_2 是 $3\,000 \times 136$ 的矩阵,包含 14 235 个从 1 到 5 的等级,大约占有所有可用等级的

6.23%,所有的缺损值表示为 0。在这些等级中,等级 1、2、3 的数量基本相同,等级 4 的数量是前者的 2 倍,等级 5 的数量是前者的 4 倍。这是因为用户不愿意对不满意的资源进行评价造成的。按照前面的方法,构建用户标记信息等级。得到用户标记图 G_3 是一个 $3\,000 \times 86$ 的矩阵,包含 7 426 个从 1 到 4 的等级。

由于二进制的信任投票表现为偏态分布,因此利用准确率和召回率之比 F1 的值以及 AUC 的值描述性能。为了更好地评价信任预测的结果,我们分别计算了信任和不信任情况下的 AUC 值和 F1 的值。

由于 G_3 图是构造得到的,我们先不考虑模型中的用户标记图,仅考虑用户资源图,用上述几种方法进行处理,实验结果如表 1 所示。

然后考虑用户资源图 and 用户标记图,用同样的过程进行处理,对比构建得到的用户标记图对信任预测的影响,实验结果如表 2 所示。

表 1 不包含用户标记信任时信任预测性能表

信任类型	方法	AUC	F1
信任	KNN	0.184	0.209
	SimRank	0.186	0.209
	SVD	0.125	0.167
	TPMF	0.212	0.221
不信任	KNN	0.916	0.977
	SimRank	0.916	0.976
	SVD	0.583	0.971
	TPMF	0.992	0.991

表 2 包含用户标记信任时信任预测性能表

信任类型	方法	AUC	F1
信任	KNN	0.189	0.213
	SimRank	0.191	0.214
	SVD	0.131	0.171
	TPMF	0.217	0.231
不信任	KNN	0.921	0.982
	SimRank	0.920	0.983
	SVD	0.587	0.976
	TPMF	0.993	0.994

从表 1 可以看出,TPMF 跟其他方法比具有更好的性能。表 2 中列出了根据 AUC 和 F1 标准得到的性能值,从表中可以看出 TPMF 的性能最好。由

于信任图和 Folksonomy 数据中彼此共享有价值的信息, 这些信息有助于降低社会网数据集中数据稀疏问题, 提高预测的准确度.

5 结束语

本文开发了一种基于 Folksonomy 信息作为辅助信息进行信任预测的方法, 将用户信任图与 Folksonomy 中用户资源图 and 用户标记图中共同的组结构信息知识利用起来, 作为辅助信息用于信任预测, 提高了信任预测的能力. 在真实数据集上的测试表明, 我们提出的 TPMF 方法在信任预测方面性能要好于经典的信任预测方法. 改进该方法, 利用用户信任关系信息进行 Folksonomy 中用户标记推荐是接下来需要做的工作.

参考文献:

- [1] Leskovec J, Huttenlocher, D, Kleinberg, J. Predicting positive and negative links in online social networks[C] // Proceedings of the 19th International Conference on World Wide Web. New York: ACM, 2010: 641-650.
- [2] Guha R, Kumar R, Raghavan P, et al. Propagation of trust and distrust[C] // Proceedings of 11th International Conference on World Wide Web. New York: ACM, 2004: 403-412.
- [3] Getoor L, Diehl C. Link mining: A survey[J]. ACM SIGKDD Explorations Newsletter, 2005(7): 2, 3-12.
- [4] Yu K, Chu W. Gaussian process models for link analysis and transfer learning[C] // Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2007: 1657-1664.
- [5] Srebro N, Jaakkola T. Weighted low-rank approximations[C] // Proceedings of the 20th Annual International Conference on Machine Learning. Palo Alto, CA: AAAI Press, 2003: 720-727.
- [6] Salakhutdinov R, Andmih A. Probabilistic matrix factorization[C] // Advances in Neural Information Processing Systems. Cambridge, MA, MIT Press, 2007: 1257-1264.
- [7] Andreas Hotho, Robert Jäschke, Christoph Schmitz, et al. Information retrieval in folksonomies: Search and ranking[C] // Proceedings of the 3rd European Conference on the Semantic Web: Research and applications. Berlin: Springer, 2006: 411-426.
- [8] Massa P, Avesani P. Trust metrics in recommender systems[M]. London: springer, 2009: 259-285.
- [7] Wieselthier J E, Nguyen G D, Ephremides A. Energy-limited wireless networking with directional antennas: the case of session-based multicasting[C] // Proceedings of IEEE INFOCOM. [s. l.]: IEEE, 2002: 190-199.
- [8] Rappaport T S. Wireless communications: principles and practice [M]. New Jersey: prentice hall PTR, 1996.
- [9] Sohrabi K, Gao J, Ailawadhi V, et al. Protocols for self-organization of a wireless sensor network [J]. IEEE Personal Communications, 2000, 7(5): 16-27.
- [10] HOU Y T, Shi Y. Variable bit rate flow routing in wireless sensor networks[J]. IEEE Transactions on Wireless Communications, 2007, 6(6): 2140-2148.

作者简介:

王树锋 男, (1968-), 博士, 副教授. 研究方向为智能计算与语义网技术. E-mail: 15961254516@163.com.

王文 男, (1974-), 硕士, 讲师. 研究方向为信息检索与智能信息处理.

作者简介:

薛亮 男, (1982-), 博士, 副教授. 研究方向为无线传感器网络、无线认知技术.

王燕龙(通信作者) 男, (1991-), 硕士研究生. 研究方向为无线传感器网络、网络优化. Email: 764437873@qq.com.

李志华 女, (1978-), 博士, 副教授. 研究方向为物联网、无线传感器网络.

(上接第 135 页)