

一种基于内容挖掘的社会网搜索框架

王 文^{1,2}, 王树锋^{1,2}, 庄燕滨^{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: 提出了一种社会网搜索框架, 从用户过去的及其与之连接的朋友新近提交的内容中挖掘用户对不同主题下同一项目的偏好信息, 结合用简单贝叶斯方法确定的、与标记关联的某个项目属于不同主题的概率, 对用户提交的查询进行搜索, 推荐框架将检索得到的项目中排在前面的推荐给用户. 实验表明, 添加用户偏好能很好地改善搜索的性能, 准确率和召回率均有极大的提高. 另外该方法稍做扩展也能成为在社会网中发现领域专家, 从而改进了专家的定位机制.

关键词: 搜索框架; 社会网; 内容挖掘; 专家定位; 用户偏好

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2014)06-0064-04

Search Framework in Social Network Based on Mining Content

WANG Wen^{1,2}, WANG Shu-feng^{1,2}, ZHUANG Yan-bin^{1,2}

(1 Changzhou Institute Technology, Changzhou 213002, China;

2 Changzhou Key Laboratory of Software Technology and Application, Changzhou 213002, China)

Abstract: The framework of social web search mines past activities of the user in his/her social network to identify the user preferences for items of different topics/categories posted by user's contacts. A Naive Bayes approach is used in category determination process and the corresponding probabilities are mined. All items in the list returned by the search process are evaluated and the top items are recommend to the user. Experiments demonstrate that both the precision and recall values of search results are dramatically enhanced compared to only filtering by contacts. One immediate extension to framework can be to discover expertise, and also improve the expertise locator mechanism.

Key words: search framework; social network; mining content; expertise location; user preference

1 引言

近年来, 在社会网站上注册的用户急剧增长, 成千上万的用户参与到社会网站的活动, 面对海量增长的信息, 利用社会网络站点提供的信息检索工具, 用户很难快速找到自己感兴趣的东. 因为这些工具会将符合某个搜索条件的不变的搜索结果提交给所有用户, 不能实现个性化服务. 在社会网络中, 如果能将网络的拓扑结构和用户近邻的信息用于搜索过程, 肯定能够提高搜索的准确度^[1-3].

与找到符合用户要求信息相关的问题是找到掌握这些信息的人, 也就是专家搜索. 设想游客想在旅

游地找一个有地方特色的风味餐馆, 如果游客能够得到跟他自己有直接或间接交往的美食家朋友的推荐, 肯定要比餐馆推荐系统根据匿名用户的评价提供的推荐列表好的多. 搜索专家只需要考虑专家应该具有的经验和水平, 利用社会网络站点中建立的用户之间的链接, 能够方便地查找可信任的专家^[4-5].

本文提出一种社会网搜索框架, 从用户过去的及其与之连接的朋友新近提交的内容中挖掘出用户对不同主题下同一项目的偏好信息, 结合用简单贝叶斯方法确定的、与标记关联的某个项目属于不同主题的概率, 搜索用户提交的查询, 推荐框架将检索

收稿日期: 2013-09-17; **修回日期:** 2013-11-13

基金项目: 常州市应用基础计划研究项目(J20120009); 常州工学院校级科研基金项目(YN1203, YN1316)

得到的项目中排在前面的项目推荐给用户^[6-8].

2 模型结构

2.1 模型框架

该模型框架结构如图1所示,包含偏好学习模块、项目分类模块、查询分析模块和推荐模块^[9-10].

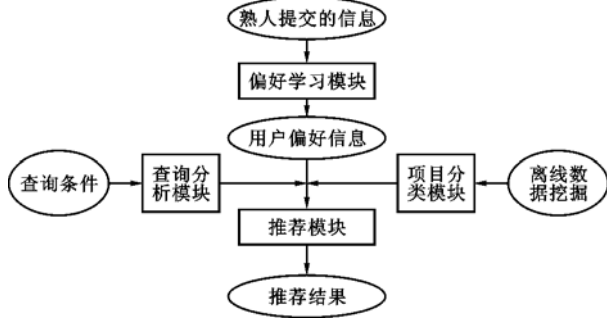


图1 模型框架结构图

偏好学习模块:该模块从描述用户过去活动的数据和在社会网中与用户熟悉的人员所提交给该用户在不同分类下的项目数据中挖掘得到用户的偏好信息,偏好信息描述为由用户 U_b 提交的用户 U_a 也喜欢的项目 i 的概率,用公式 $\Pr(i \in \text{posted}(u_b) | \text{likes}(u_a, i))$ 表示. 这个概率值也用于推荐模块中对搜索结果进行排序.

项目分类模块:对项目、类型数据集进行离线挖掘,确定每个项目所属的类别. 每个项目对应于一个分类标记表,根据标记表中的位置确定每一个标记的权重. 标记表中标记的位置越靠前,项目的分类类别越重要. 设标记 t_j^i 表示列表 i 中第 j 个位置,分类确定算法用一个关于 j 的减函数计算 t_j^i 的权重. 考虑项目 i 中一组标记,根据 $\text{tag}(i)$ 中的标记数量,计算出项目 i 属于任意 $\text{tag}(i)$ 类别的概率.

查询分析模块:该模块将一组标记作为查询条件,在社会网上进行搜索. 该搜索过程一直进行,直到搜到确定数量的项目或达到预先确定的网络深度.

推荐模块:根据用户提交的查询条件,搜索得到一个标记列表,推荐模块将检索得到的项目排序,将排在前面的项目推荐给用户.

2.2 项目分类计算

用简单贝叶斯方法计算每个项目属于不同类别的概率. 设与项目 i 关联的一组标记为 $\text{tag}(i)$,项目 i 属于每一个分类的概率由 $\text{tag}(i)$ 中的标记数量计算得到. $\text{Lookup}(d, t)$ 谓词表示标记 t 是否包含在标记词典 d 中,返回 0 或 1.

C 是分类的集合,项目 i 属于分类 C_x 的计算公

式如下:

$$\Pr(\text{cat}_i = c_x) = \frac{\sum_{k=1}^{l_i} \omega(t_k^i) * \text{lookup}(d_{c_x}, t_k^i)}{\sum_{c_y \in C} \sum_{k=1}^{l_i} \omega(t_k^i) * \text{lookup}(d_{c_y}, t_k^i)} \quad (1)$$

式中, cat_i 指项目 i 的分类, $\text{tag}(i, j)$ 表示标记列表 $\text{tag}(i)$ 中第 j 个位置的标记 t_j^i 的权重 $\omega(t_j^i)$,由分类确定函数计算得到. 该函数是关于 j 的减函数,具体函数由应用领域确定.

2.3 推荐模块的计算过程

根据用户提交的查询条件,搜索得到一组项目,计算出这些项目分属于不同类型的概率. 根据不同类型挖掘出的用户偏好信息,将搜索过程返回的所有项目按对应的概率值进行排序,将排在前面的项目推荐给提交查询条件的用户.

计算 $\Pr(i \in \text{posted}(u_b) | \text{likes}(u_a, i))$ 时,用贝叶斯公理:

$$\frac{\Pr(i \in \text{posted}(u_b) | \text{likes}(u_a, i)) \Pr(\text{likes}(u_a, i))}{\Pr(i \in \text{posted}(u_b))} \quad (2)$$

假设项目的偏好依赖于用户自身的信息和项目的内容. $\Pr(i \in \text{posted}(u_b) | \text{like}(u_a, i))$ 扩展为公式(3):

$$\sum \left[\frac{\Pr(i \in \text{posted}(u_b) | \text{cat}_i = c_x, \text{likes}(u_a, i))}{\Pr(\text{cat}_i = c_x | \text{likes}(u_a, i))} \right] \quad (3)$$

应用条件概率重写式(3),得到式(4):

$$\sum \left[\frac{\Pr(i \in \text{posted}(u_b) | \text{cat}_i = c_x, \text{likes}(u_a, i))}{\frac{\Pr(\text{likes}(u_a, i) | \text{cat}_i = c_x) \Pr(\text{cat}_i = c_x)}{\Pr(\text{likes}(u_a, i))}} \right] \quad (4)$$

用户提交的项目 i 的概率同样依赖于内容. 因为用户提交的习惯总是偏向某些类型. 因此式(2)中的分母 $\Pr(i \in \text{posted}(u_b))$ 可以写为

$$\sum_{c_x \in C} [\Pr(i \in \text{posted}(u_b) | \text{cat}_i = c_x) \Pr(\text{cat}_i = c_x)] \quad (5)$$

将式(3) ~ (5) 代入并化简, $\Pr(i \in u_b | \text{likes}(u_a, i))$ 可以写为

$$\frac{\sum_{c_x \in C} [\Pr(i \in \text{posted}(u_b) | \text{cat}_i = c_x, \text{likes}(u_a, i)) \Pr(\text{likes}(u_a, i) | \text{cat}_i = c_x) \Pr(\text{cat}_i = c_x)]}{\sum_{c_x \in C} \Pr(i \in \text{posted}(u_b) | \text{cat}_i = c_x) \Pr(\text{cat}_i = c_x)} \quad (6)$$

式(6)中未知的概率按下面的方法进行计算:

$\Pr(i \in \text{posted}(u_b) \mid \text{cat}_i = c_x, \text{likes}(u_a, i)) =$ 用户 u_b 提交的属于 c_x 类别的, 并且被 u_a 评论过的项目数 / 所有属于 c_x 类别的, 并且被 u_a 评论过的项目总数.

$\Pr(\text{likes}(u_a, i) \mid \text{cat}_i = c_x) =$ 所有属于 c_x 类的被 u_a 用户评论的项目数 / 所有属于 c_x 类的项目总数.

$\Pr(i \in \text{posted}(u_b) \mid \text{cat}_i = c_x) =$ 所有用户 u_b 提交的属于 c_x 类的项目数 / 所有属于 c_x 类的项目总数.

2.4 专家定位

查找专家对有连接关系的每个朋友都进行查询是不现实的. 实际上, 给每一个朋友发送一个查询会引起请求阻塞, 这样会让专家忽略掉最需要应答的请求. 研究表明, 人们十分愿意回答陌生人所提的任何问题. 因此对于任意一个查询, 可以确定一组朋友, 这些朋友提交了与查询匹配的感兴趣的项目. 如果专家定位机制可以在局部区域内确定最合适的朋友应答查询, 系统很快就能检索到相关的信息, 社会网中包含丰富的资源用来实现这个过程.

3 实验和结果评价

3.1 实验数据

flickr.com 网站允许用户上传照片、标记照片、评论照片, 该网站允许用户通过指派他人作为朋友构建社会网. 同时该网站提供了 API, 允许用户编写程序, 从其他帐号中收集信息. 本文利用从 Flickr 系统中收集的数据测试提出的模型.

从 Flickr 的兴趣页的列表上随机选取提交较多评论的 10 个根用户, 对于每一个根用户, 通过调用 API 编程, 访问他的朋友的帐号, 收集包括标记和评论信息的照片. 我们收集了 2011. 6. 1-2011. 9. 30 期间 6 015 个用户上传的信息, 包括 121 953 张照片和 30 040 段评论. 实验用前 3 个月的数据作为训练集, 最后一个月的数据作为测试集. 照片分为 10 类: (Animal, Art, City, Entertainment, Nature, News and Politics, People, Science and Technology, Sports, Travel and Places), 数据集中的一张照片可以属于一种或多种类别. 对于每一种分类, 构建词典包含与该分类相关的一组标记.

为了从分类 C 构建词典, 先根据标记搜索 Flickr, 找到与 C 相关的照片. 将搜索结果中照片的标记, 手工去除不相干项, 然后添加到 C 分类词典

中. 用前面搜索结果中返回照片中共同出现的标记, 输入检索工具进行查询. 重复这个过程, 如果词典 C 达到一定的规模, 利用 Flickr 的 folksonomy 扩充词典并且保持词典的规模不变. 在我们的词典数据库中有 1 213 个具有唯一性的标记. 因为标记聚类不能将标记准确地区分开, 一些标记可能出现在多部词典中, 词典平均由 133 个标记组成. 数据库中大约 80% 的照片有至少 3 个以上的标记.

3.2 性能度量标准

考虑到被推荐的项目是一个集合, 评价采用信息检索文献中评价检索质量的度量标准: 准确率和召回率. 准确率是被用户喜欢的推荐项目的比例, 召回率是用户喜欢的项目被推荐系统推荐的比例.

3.3 推荐项目实验结果

在 5 个不同的标记集上进行了实验, 其中 1、3 和 5 标记从每个标记集中选出来用于查询, 用 5 个不同用户的平均值表示结果. 图 2 和图 3 显示了两种搜索机制下性能的比较. 在这些实验中, 对于任何查询, 返回排序结果表中前 20 项. 第一种框架 S1, 根据用户的熟人过滤结果. 第二种框架 S2, 对用户的熟人按照用户的偏好进一步排序搜索结果. 从图 2 可以看出, S2 框架准确度显著提高, 图 3 是该实验对应的召回率. 本图中 max 表示最大可能召回值, 这里就等于 20. 实验结果表明利用特定用户(用户的熟人)在给定搜索标题下的用户偏好, 能够提高搜索的质量.

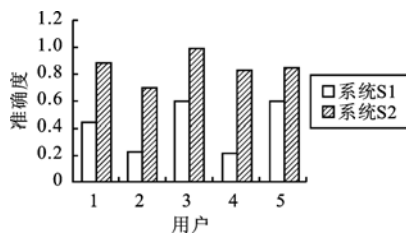


图 2 系统 S1 和 S2 准确度比较图

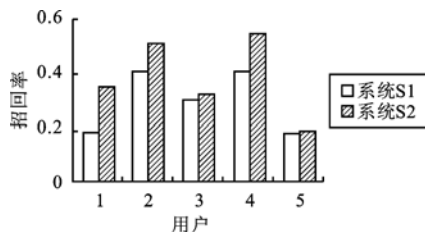


图 3 系统 S1 和 S2 召回率比较图

3.4 专家定位实验及结果分析

试验验证给定一组标准, 系统是否能够确定正确的专家. 将根据熟人过滤照片, 然后进行排序的方案替换为: 先限制熟人的数量为常数 k , 从这 k 个用

户中选取匹配的照片,然后进行排序.试验中取 $k=10$,在两种系统上度量准确度. E1 中朋友随机选择, E2 中朋友根据特定标题下用户偏好选择. 在专家定位系统中,确定一个具有特定知识的用户要比找到具有一点知识的人重要的多. 利用具有 5 个标记的 5 个不同的查询开始实验,计算 5 个不同的用户准确度的平均值. 图 4 表明偏好选择比随机选择准确度有显著的提高. 对某些用户,平均准确度达到 80%.

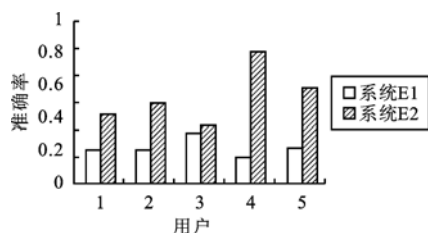


图4 两种不同的专家搜索方法比较图

4 结束语

本文提出了一种社会网搜索框架^[11-13],根据用户过去的及其与之连接的朋友新近提交的内容,挖掘出用户对不同主题下同一项目的偏好信息,结合用简单贝叶斯方法确定的、与标记关联的某个项目属于不同主题的概率,对用户提交的查询进行搜索. 推荐框架将检索得到的项目中排在前面的项目推荐给用户. 实验表明,添加用户偏好,能很好地改善搜索的性能,准确率和召回率有极大的提高. 另外该方法稍做扩展也能用于在社会网中发现领域专家,改进了专家的定位机制.

参考文献:

- [1] Schafer J B, Konstan J, Riedi J. Recommender systems in ecommerce[C]// Proceedings of the 1st ACM conference on Electronic commerce. New York, ACM, 1999: 158-166.
- [2] Jamali M, Abolhassani H. Different aspects of social network analysis[C]// Web Intelligence. IEEE/WIC/ACM International Conference on. Washington: IEEE, 2006: 66-72.
- [3] Brin S, Page L. The anatomy of a large-scale hypertextual web search engine[J]. Computer Networks, 1998,30(1): 107-117.
- [4] Sinha R R, Swearingen K. Comparing recommendations made by online systems and friends[C]// DELOS workshop: personalisation and recommender systems in digital libraries. Ireland: Dublin City University, 2001:106.
- [5] Lerman K. Social networks and social information filtering on digg[J/OL]. (2006-11-07). [2013-04-10]. arXiv preprint cs/0612046.
- [6] Adamic L, Adar E. How to search a social network[J]. Social Networks, 2005, 27(3): 187-203.
- [7] Lerman K, Plangprasopchok A, Wong C. Personalizing image search results on flickr[J]. Intelligent Information Personalization, 2007(4):20-30.
- [8] Golbeck J, Hendler J. Filmtrust: movie recommendations using trust in web-based social networks[C]// Proceedings of the IEEE Consumer communications and networking conference. Washington: University of Maryland, 2006(1):282-286.
- [9] Yu B, Singh M P. Searching social networks[C]// Proceedings of the second international joint conference on Autonomous agents and multiagent systems. New York, ACM, 2003: 65-72.
- [10] Kautz H, Selman B, Shah M. Referral web: combining social networks and collaborative filtering[J]. Communications of the ACM, 1997, 40(3): 63-65.
- [11] Guandong Xu, Yanchun Zhang, Lin Li. Web mining and social networking techniques and applications[M]. New York, Springer, 2011.
- [12] Jannach D, Zanker M, Felfernig A, et al. Recommender systems: an introduction[M]. Cambridge University Press, 2010.
- [13] Shapira B. Recommender systems handbook[M]. New York, Springer, 2011.

作者简介:

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

王树锋男, (1968-), 博士, 副教授. 研究方向为智能信息处理.

庄燕滨男, (1964-), 硕士, 教授. 研究方向为嵌入式系统、基于领域的信息化应用.