基于社会化标签系统的个性化信息推荐探讨

田莹颖
南京大学信息管理系　南京 210093

〔摘要〕针对用户个人特征并向其提供准确恰当信息的个性化信息推荐研究，一直是学术界和产业界所关注的热点。结合后控词表，对用户分散的、个性化的标注进行处理，并将用户兴趣用向量表示，然后借鉴协同过滤算法的思想，寻找出相似用户集及其内部的资源集。在此基础上，采用相对匹配策略，提出一种基于社会化标签系统的个性化推荐方法。

〔关键词〕社会化标签系统　个性化信息推荐　协同过滤
〔分类号〕TP391.4　TP311.13

On Personalized Information Recommendation Based on Social Tagging System

Tian Yingying
Department of Information Management, Nanjing University 210093

〔**Abstract**〕Research on how to provide accurate and appropriate recommendations based on user profile has always been a hot spot. Combining the handling of controlled vocabulary, this paper seeks to deal with users' discrete and personalized tags. Firstly, it expresses the user interest as a user interests vector, then finds similar users and the resources, and draws on the collaborative filtering algorithms. On this basis, by consulting the relative matching method, the paper presents a personalized recommendation method which based on the social tagging system.

〔**Keywords**〕social tagging system　personalized information recommendation　cooperative filtering

社会化标注是Web2.0思想中的一种主要应用，它使用户可以自由地对互联网信息进行标注，从而反映出用户的兴趣和认知偏好。如果将这些联结信息资源和用户的社会化标签运用到个性化推荐系统中去，既简单易行，又能够使信息提供者比较准确地了解用户的需求并有针对性地进行信息推荐，从而大大满足用户的个性化需求。一些论文对此已做了较深入的研究，并提出了许多相关的推荐算法[1-3]。

1 个性化信息推荐

个性化信息服务是随着互联网发展而产生的新型信息服务，是一种根据用户的信息需求、兴趣或行为模式，将用户感兴趣的信息、产品和服务推荐给用户的个性化信息服务模式，是以“用户为中心”的服务原则在网络环境下的具体体现。对于个性化信息服务的概念，目前业界存在多种解释，并没有出现统一的定义。但任何一种解释都体现了个性化信息服务“以用户为中心”的原则：服务时空的个性化和服务方式的个性化、服务内容的个性化。

个性化信息推荐服务中的关键点包括：用户兴趣的获取，用户兴趣与信息类别的匹配。目前，用户个人数据的获取主要分为两种方式：即显式的手工输入用户个性化特征和隐式的通过Web挖掘来跟踪用户的行为，自动获取用户的个性化特征[4]。在获取了用户兴趣之后，我们需要对此建立一个模型，以方便将用户兴趣与资源类别对应起来。用户模型是指对用户的个人兴趣建立的模型，也称为“用户兴趣模型”。相应地，Web文档模型是对文档内容的抽象描述，在Web2.0中不仅有文档内容，还包括音频视频等多种形式，我们暂且称其为资源模型。最后通过将用户模型与资源内容进行匹配，进行个性化信息的推荐。

2 社会化标签

社会化标注是一种以人为本的、灵活的组织和管理在线信息、进行网络信息分类的方式。大众分类更近乎个人的知识体系，它的使用以个人的感性逻辑（个

收稿日期：2009-06-23　修回日期：2009-08-28　本文起止页码：50-53，120　本文责任编辑：杜杏叶

人知识、情感、意志、记忆、素养等等的综合反映)为线索,以个人所需信息的汇集、梳理和查询为目的,以个人的经验为基础。它不同于传统的、针对文件本身的关键字检索,而是一种模糊化、智能化的分类。我们可以为每篇日志、每个帖子或者每张图片、每个视频,甚至我们认为需要或可以添加标签(Tag)的任何网络信息资源都添加一个或多个 Tag。Tag体现了群体的力量,使得内容之间的相关性和用户之间的交互性大大增强[5]。网络用户可以通过添加多个 Tag为网络资源分类,也可以通过搜索某一个或几个 Tag标签发现其它用户具有相同标签的资源。

3 协同过滤算法

协同过滤技术也称为面向用户(user- based)的技术,它的基本原理是利用用户访问行为的相似性来互相推荐用户可能感兴趣的资源,即协同过滤技术通过分析历史数据,生成与当前用户行为兴趣最相近的用户集,将他们最感兴趣的项作为当前用户的推荐结果。基于协同过滤技术的推荐过程可分为 3个阶段:数据表述、发现最近邻居、产生推荐数据集[6]。

4 基于社会化标签的信息推荐算法

算法:① 将用户兴趣用向量表示;② 进行不同用户兴趣向量之间的相互比较,找出相似用户集。在这个步骤中会借用协同过滤技术中的部分概念;③ 在这个用户集内部寻找他们最感兴趣的文档或资源,形成一个相似用户集内部的资源集;④将资源集中的每项资源分别用向量表示;⑤将某个用户的兴趣向量与每项资源的向量比较,比较结果按从大到小顺序排列,前 n项即可作为推荐的资源。

算法解释:

第①步:用户兴趣的向量表示。以特征项(包括字、词或短语)作为用户兴趣模型的表示单位,在此处的计算中使用 Tag用词。一个用户兴趣模型可以表示为一个向量: $p(t_1 w_1, t_2 w_2, \cdots, t_n w_n)$,简记为 $p(w_1, w_2, \cdots, w_n)$,向量的维数是特征项的个数 n,每个分量的值 w_i是特征项 t_i在用户所有标注中出现的频率,即某个 Tag在用户所有标注中出现过的次数,用于权衡特征项的重要程度,所以又称为特征项的权重。目前有多种加权方法,本文使用 TF- IUF加权方案。IUF(inverse user frequency)方法来源于著名的 IDF(inverse document frequency)加权方法[7]。IDF定义为 log(Y/y),其中 Y是文档集中的文档数,y是包含某一词的文档数。这样,包含这个词的文档个数越少,idf的值就越大,如果文档集中的每一个文档都包含这个词,那么 idf的值为 0。这表达了一种常识意义上的直觉,如果一个词出现在文档集的每一个文档中,那么这个词在从文档集中区分出该文档时几乎不起任何作用。User- tf是某一词在某一文档中出现的频率,因此,tf是关于文档的统计数据,它因文档的不同而异,其作用是试图度量该词在文档中的重要性。相反,idf是全局统计数据,它度量了该词在整个文档集中分布的广泛性[8]。对应的,TF- IUF(terms frequency- inverse user frequency)方法在对比用户偏好的时候降低了热门 tag的权重,因为这些热门的 tag在比较用户相似程度时的效果不如一般的 tag。

权重的计算运用 TF- IUF公式:

$$u_i(t, p) = tf_i * iuf = tf_i * \log(N / n_i + 0.01) \quad (1)$$

其中 tf_i为标签 t_i在当前用户所有标注中出现的次数,称为项频,N为全部用户集合 U中的用户总数,n_i为 U中使用过标签 t_i的用户数。这个公式的核心思想是一个词被一个用户使用的频率越高,而被其他用户使用的频率越低,则该词对用户兴趣的区分能力越强,故其权重也越大[9]。事实上,如果一个用户经常使用某个词语作为 Tag说明他对这个方面的兴趣是相对更加强烈的,持续时间也更长,同时,这个 Tag越冷门则越能体现这个用户兴趣的特殊性,因此对这种词汇的权重更需要加强。

第②步:基于用户兴趣向量的相似兴趣用户确定。关于用户之间的相似性计算,常用的有 Pearson相关度(correlation)计算方法和目前常用的余弦相似度计算方法。本文采用余弦相似度的计算方法来比较用户兴趣模型,计算用户兴趣的相似度,寻找最近邻居集。两个用户 $user_1$和 $user_2$被看作是向量空间中的两个向量,可以通过计算两个向量的夹角的余弦来衡量相互之间的相似度,夹角越小,相似度越高。

例如将某用户的用户兴趣转化为向量 p,另一用户的用户兴趣转化为向量 r,则通过下面的公式进行计算:

$$Similarity(p, r) = \cos(p, r) = \sum_{k=1}^{n} u_k \sum_{k=1}^{n} g_k \Big/ \sqrt{\sum_{k=1}^{n} u_k^2} \sqrt{\sum_{k=1}^{n} g_k^2} \quad (2)$$

其中 U_k, g_k分别为用户兴趣向量 p和用户兴趣向量 r的第 k个特征项的权重,similarity值越大表明二者内容越相近,设定一个阈值 θ,当 similarity> θ时,p就

被添加进 r的相似用户集[9]。

第③步: 基于相似兴趣用户的资源集确定。确定基于相似兴趣用户的资源, 就是已经找到相似用户集的情况下, 确定这部分用户共同关注的资源。在这步里不考虑 Tag的作用, 直接找出相似用户浏览或发布的资源, 在准备好这些资源之后, 再在算法第 ⑤步里根据 Tag判定每篇资源与用户兴趣的相似性。

本文选取相似兴趣用户发布过的资源作为基于相似兴趣用户的资源集。

第 ④步: 基于相似用户的资源的向量化。这步处理过程与第一步是相似的。先将资源被标注的 Tag经过后控词表处理, 然后按顺序排列。一篇资源可以表示为一个向量: $d(t_1w_1, t_2w_2, \cdots, t_nw_n)$, 简记为 $d(w_1, w_2, \cdots, w_n)$, 向量的维数是特征项的个数 n, 每个分量的值 w_i是特征项 t_i在文资源中出现的频率。权重 w_i的计算运用 TF- IDF公式:

$$w_i(t, d) = tf_i^* idf = tf_i^* log(Y/y_i + 0.01) \quad (3)$$

其中 tf_i为词 t_i在资源 d中出现的次数, idf为反向文档频率, Y为资源集中的资源总数, y_i为资源集中出现特征项 t_i的资源数。

第 ⑤步: 用户兴趣向量与资源向量的比较。考虑到用户的兴趣也是不断变化的, 近期的偏好应该更被重视。因此将用户添加的 Tag按时间顺序分配权重, 对每一项 u_i乘以一个与时间相关的系数, 这个系数是根据时间远近变化的。对于反复出现过的 Tag就根据最近一次标注的时间来确定时间系数。例如将最近标注的 100个 Tag的权重系数设为 a_1, 前 100- 200个 Tag的权重系数设为 a_2, $a_1 > a_2 > a_3 > \cdots a_n$以此类推, 直至最近的 1 000个 Tag。这样就对用户最近标注的 Tag给与更多的重视, 使用户兴趣模型可以随用户兴趣的变化而变化。原先的用户兴趣模型即变为 $P'(u'_1, u'_2, \cdots, u'_n)$。

这样通过计算两个向量的相似度就可以判断文档和用户兴趣的匹配程度。相似度计算多采用余弦公式:

$$Similarity(d, p') = cos(d, p') = \sum_{k=1}^{n} w_k \sum_{k=1}^{n} u'_k \Big/ \sqrt{\sum_{k=1}^{n} w_k^2} \sqrt{\sum_{k=1}^{n} u'^2_k} \quad (4)$$

其中 w_k、u'_k分别为资源向量 d和用户兴趣向量 p'的第 k个特征项的权重, similarity值越大表明两者内容越相近, 设定一个阈值 β, 当 similarity > β时就可以将资源推荐给用户了, 或者直接选取排名最前的 n项资源进行推荐。

对于没有足够线索发现用户喜好的情况 (例如新用户, 或标注的 Tag数量过少的用户), 根据公共兴趣进行推荐。即默认当前用户的兴趣是符合大多数人的看法的, 按照最热门的 Tag进行推荐, 取最热门的若干 Tag当作用户的标注, 依照以上方法进行推荐。

5 实证分析

5.1 实验方法

实验数据集取自豆瓣站点 (www. douban. com), 该站点是一家 Web2 0网站, 用户可以自由发表有关书籍、电影、音乐的评论, 可以搜索别人的推荐。所有的内容、分类、筛选、排序都由用户产生和决定, 甚至在豆瓣主页出现的内容上也取决于用户的选择。截至到 2008年底, 该站点的用户已经超过 300万人。我们通过对于豆瓣网站结构的分析, 得到了豆瓣的用户列表, 利用 htmlparser工具解析出用户的独立 id, 然后通过豆瓣公开的 API, 以用户为参数, 对用户所收集的书籍的标签进行获取, 从该站点得到实验数据集 (包括 64 091个用户和其标签过的图书)。

由于原始数据包含较多的噪音信息, 无法直接使用, 所以做了一些简单的前期处理, 包括: ①剔除 Tag数少于 5个的用户 (信息量太少); ②对 Tag进行排序, 得出 Tag使用的频率, 剔除拥有 80% 热门 Tag的用户 (信息量太大众化, 影响向量空间内分布); ③将用户标注的所有 Tag写入一个临时词表, 再将这个临时词表与后控词表进行比照。后控词表中包含了图书的书名及其相关译名、作者、语言种类、图书分类以及近义词同义词词典。按写入顺序抽取临时词表中的词, 将其与受控词一一比较, 若该词与受控词一致, 则直接使用, 否则将其转化为后控词表中对应的受控词, 如使得 “人工智能”、“人智”、“AI”等标签能统一表示为“人工智能”, 达到降低向量维度的效果。最后, 用经过处理之后的实验数据集进行实验。

目前, 学术界还没有关于推荐系统效果的标准评价方法, 已有的评价方法多为借鉴信息检索领域或文本分类等其他领域的评价方法。

本文根据实验过程中某测试用户的相似兴趣用户在实验数据集中的发布过的资源为其计算 Top- N推荐资源, 如果 Top- N推荐资源中某个资源 i出现在该测试用户的访问记录里, 则表示生成了一个正确推荐。同时, 我们用信息检索领域中评估系统效果的准确率 (Precision)标准作为我们的算法推荐精度的标准: Precision = Hits/N, 其中, Hits表示算法产生的正确推荐

数, N 表示算法生成的推荐总数。

分别以用户 3310712和用户 3277978为例, 我们首先读入用户 3301712的 Tag文本, 把该用户兴趣用向量表示, 建立数组存储该用户兴趣向量; 然后选取用户群体, 计算该群体内用户兴趣向量之间的相似度, 寻找相应的邻居集; 再通过豆瓣的 API获得相似兴趣用户的资源, 得到该相似用户集的资源集合, 将资源向量化, 进行聚类; 最后通过将用户 3310712的兴趣向量与相似兴趣用户群体资源向量进行比较得出排名。

对用户 3310712的推荐资源取其前 10项得到结果, 如表 1所示:

表 1 用户 **3310712**推荐资源与访问记录

Top- 10推荐资源 (ID)	用户 3310712访问记录
1125731	no
1669898	no
2245914	yes
1463192	no
1766573	no
1054962	yes
1016976	yes
1439100	yes
1396300	no
1041638	yes

则推荐系统对用户 3310712的推荐精度可表示为 Precision = 5/10= 50%。

类似的, 对用户 3277978的推荐资源取其前 15项得到结果如表 2所示:

表 2 用户 **3277978**推荐资源与访问记录

Top- 15推荐资源 (ID)	用户 3277078访问记录
Bettyww	yes
2268878	no
2644930	no
xiachufang	no
sunny_bear	no
ilyly	yes
2982692	yes
2429205	yes
1790723	no
1873321	no
3375107	yes
susanfransisco	no
lamianaiai	yes
2997198	no
3401490	yes

则推荐系统对用户 3277978的推荐精度可表示为: Precision = 7/15= 46 7%。

5.2 结果及分析

根据用户在实验数据集中发布过的资源对测试用户进行推荐后, 统计对 200名用户的推荐精确度, 得到如图 1所示的用户推荐精确度曲线图。

图 1中横轴为用户, 纵轴为推荐算法对该用户的推荐精度。由图可以看出本文所提出的算法对用户个性化信息推荐的精度大约为 50%。

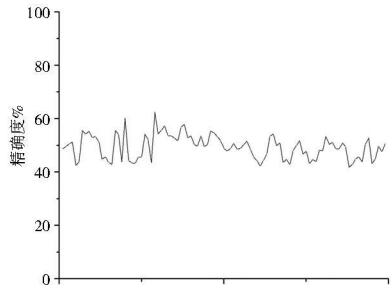

图 1 用户推荐精确度曲线图

由于豆瓣的用户的兴趣较为广泛, 用户对于同一图书的标识各有所爱, 标签的离散性较大, 而且受到受控词表技术支持的限制, 导致用户向量产生了偏差。下一步工作将继续加强完善受控词表, 对于此部分问题进行处理, 化简向量维度, 以达到更好的推荐效果。

6 结 语

本文介绍了一种在社会化标签系统下进行信息推荐的方法。它首先充分利用了 Tag标签能够反映用户偏好的特点, 选取用户自由标注的标签作为个性化推荐的依据; 然后采用后控词表对标签进行预处理, 克服了社会化标签缺乏一致性的缺点; 最后借鉴协同过滤算法的思想, 通过相似用户所标注的资源进行个性化推荐, 并提出在计算用户兴趣向量过程中用赋予不同权重的方法减轻用户兴趣随时间变化造成的影响。

在今后的研究中, 我们一方面要继续对标签预处理的方式进行改进, 将标签与用户浏览、检索行为结合起来, 更深层次探索了解用户心理; 另一方面, 文中采用的协同过滤算法还存在部分不完善之处, 如用户群的划分不准确、冷启动等问题, 我们将更多参考传统推荐算法中的相关做法, 改进用户模型的构建方法。

参考文献:

[1] Linden G, Smith B, York J. Amazon. com recommendations: Item- to- Item collaborative filtering. IEEE Internet Computing, 2003, 7(1): 76- 80.

[2] 李峰, 李军怀, 王瑞林, 等. 基于商品特征的个性化推荐算法. 计算机工程与应用, 2007, 43(17): 94- 197.

[3] 徐国华. 基于 Web挖掘的一种个性化推荐算法. 农业网络信息, 2007(5): 23- 25.

[4] 余强, 张海盛. 个性化 Web信息服务技术研究. 计算机应用研究, 2006, (2): 198- 200.

[5] 踏莎而行. 小 Tag有大智慧. 电子商务世界, 2006(5): 84- 85.

(下转第 120页)

借阅模块中, Spring在 web. xml中配置一个应用上下文文件和一个 Servlet。在应用上下文文件中,配置持久层的定义。持久层的定义包括 Readers. hbm. xml和 Books. hbm. xml,通过这两个 xml文件, Readers. java, Books. java类分别与数据库中的 Readers, Books关系表对应。

4 OSBC整合框架优势分析

第一,三个框架均为开源框架,有丰富的文档和开发背景,非开源框架无法比拟;且基于这种架构的 J2EE应用是基于模块的,利于系统业务的重用和改动。

第二, Struts框架和 Hibernate框架都实现了很明确的层概念,能达到层次之间低耦合,便于实现系统的大规模开发、管理和维护。原有编码过程中,业务逻辑层常被忽略,业务处理的代码经常出现在表示层或持久层中,这将导致程序代码的紧密耦合,难以维护。OSBC引入 Spring开源框架来专门管理业务逻辑层,利用 AOP思想,集中处理业务逻辑,减少重复代码。

第三,设计思想清晰。Struts可以很好地把业务逻辑和表示层分离; Spring为 Hibernate提供良好的事务支持,通过封装 Hibernate Session,事务管理,利用 AOP的 method interceptor或者 Java代码层面显式的 template包装类,透明地创建和绑定 Session到当前的线程; Hibernate通过对象关系映射,把面向对象的设计与关系数据库联系起来。

第四,采用延时注入思想组装代码。利用 Spring的 IoC代替以往开发中使用的工厂模式(Factory),将类之间的依赖关系转移到外部的配置文件中,进行延时注入,组装代码,提高系统的扩展性、灵活性,实现插件式编程。

5 结 语

本文通过整合轻量级的开源框架 Struts, Spring和 Hibernate,分析各个环节的关键技术和实现问题,对应用程序分层解耦,构建高质量 J2EE应用架构,不仅最大化使用了资源,也使得图书借阅系统开发简洁、结构清晰,具有良好的稳定性和重复性。

OSBC的图书借阅和审批子系统已经程序实现,在学校图书馆使用且运行稳定,运行界面如图 2所示:

图 2 图书借阅系统运行界面

项目组根据开源框架特点,挑选具有代表性的页面,给出相应接口说明,让馆员和学生在参考模块源代码的基础上,自行设计页面,替换原有页面,充分展现系统可扩展性和可维护性。在今后的工作中,项目组将继续完善系统其他功能,在学校图书馆中推广使用。

参考文献:

[1] 焦玉英,袁静. 基于用户个性化需求的数字图书馆集成服务研究. 图书情报工作, 2009, 53(3): 54-56

[2] 阮莉萍. 图书馆自动化系统开源软件的比较研究. 图书馆论坛, 2009, 29(1): 78-80

[3] 奉国和. 开源软件与图书馆知识管理探讨. 图书馆论坛, 2008, 28(2): 58-61

[4] 周相兵,杨小平,杨兴江,等. 基于 Web分层结构的通用框架实现及应用. 计算机工程与应用, 2008, 29(7): 59-62

[5] Colin J N. LIMBS: Open source, open standards, and open content to foster learning resource exchanges. Advanced Learning Technologies, 2006(7): 682-686

[6] Antkiewicz M. Automatic extraction of framework-specific models from framework-based application code. Proceedings of the twenty-second IEEE/ACM international//Blanksher C L. Automated Software Engineering Issues. Atlant Bartlomei, 2007: 189-192

〔作者简介〕杨传明,男, 1979年生,讲师,硕士研究生,发表论文 13篇。

(上接第 53页)

[6] 赵亮,胡乃静. 张守志. 个性化推荐算法设计. 计算机研究与发展, 2002, 39(8): 986-991.

[7] Weng L T, Xu Y, Li Y F, et al. An improvement to collaborative filtering for recommender systems//Proceedings of international conference on computational intelligence for modelling control and automation, 2005. New York: IEEE, 2006

[8] 孙铁利,杨凤芹. 根据用户隐式反馈建立和更新用户兴趣模型. 东北师大学报自然科学版, 2003, 35(3): 99-104.

[9] 汪勤,安贺意,秦颖. 网络信息过滤和个性化服务,情报科学, 2007, 25(6): 858-863

〔作者简介〕田莹颖,女, 1985年生,硕士研究生。