

# 基于万有引力的个性化推荐算法

王国霞<sup>✉</sup>, 刘贺平, 李 擎

北京科技大学自动化学院, 北京 100083

✉ 通信作者, E-mail: kdmycevin@sohu.com

**摘 要** 本文把物理学中的万有引力定律引入推荐系统, 提出一种个性化推荐算法, 即基于万有引力的个性化推荐算法。算法把用户使用的标签看作用户喜欢物体的组成颗粒, 标注项目的标签被看作项目物体的组成颗粒, 社会标签的类型就是颗粒的类型, 由此构建了用户喜好物体模型和项目物体模型。喜好物体和项目物体间存在着万有引力, 并且引力大小遵循万有引力定律。计算喜好物体和项目物体间的万有引力, 并把该引力大小作为二者的相似度度量, 引力越大, 二者的相似度就越高, 对应的项目物体就越有可能被用户喜欢。实验结果证明本文提出的算法可以获得好的推荐性能。

**关键词** 推荐算法; 个性化; 万有引力; 社会标签

分类号 TP391

## Gravitation-based personalized recommendation algorithm

WANG Guo-xia<sup>✉</sup>, LIU He-ping, LI Qing

School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China

✉ Corresponding author, E-mail: kdmycevin@sohu.com

**ABSTRACT** A recommendation algorithm is proposed by introducing the universal law of gravitation into a recommendation system. This new algorithm is named as the gravitation-based personalized recommendation (GBPR) algorithm. In the algorithm, social tags used by users are regarded as particles that made up of their preference objects, social tags marking on items are considered as particles that made up of item objects, and the user preference objects and item objects are taken as a user preference object model and an item object model, respectively. Gravitation exists between the user preference objects and item objects, and its strength obeys the universal law of gravitation. The strength of gravitation between the user preference objects and the item objects is computed, and it is regarded as their similarity. The bigger the strength is, the more similar they are, and the corresponding item objects are more probable to be liked by users. Experimental results show that the proposed algorithm can get good performance.

**KEY WORDS** recommendation algorithms; personalization; gravitation; social tags

个性化推荐技术因其在解决信息超载和资源迷向问题上的优越性, 使得它在学术界的研究和商业的应用等方面都取得了良好的发展和应用<sup>[1-2]</sup>, 作为其核心的推荐算法更是学术界的研究热点, 目前大量具有良好推荐性能的推荐算法被提出来。

现有的推荐算法能取得好的推荐性能, 但因算法固有特点而存在一些无法克服的缺点。例如大多算法的推荐依据主要是用户的评分, 而用户的评分仅仅就

是一个数值<sup>[3]</sup>, 分值的高低仅能反映用户对某项目的好恶, 而用户真正喜欢的项目类型及项目自身特征等信息无法从中获得, 所以依据评分信息获得的推荐有失偏颇。另外, 推荐系统存在于被基本物理定律控制的物理世界中, 但现有的推荐算法缺乏物理学方面的深层理解和解释, 这不失为一种严重的遗憾。

Web2.0 技术的发展使得社会标签 (tag) 越来越为用户所熟知。用户在发布或浏览网上资源时, 自由选

收稿日期: 2013-12-23

基金项目: 国家软科学研究计划资助项目 (2013GXSB178)

择词汇对其进行标注,以方便对网络资源进行分类、共享、浏览等<sup>[4]</sup>. 标签作为一种新的网络数据,一方面来源于用户对资源的理解和概括,具有一定的个性化特征;同时标签又可以对资源进行描述和分类,相比较用户评分,标签携带了更多的信息量<sup>[4-6]</sup>. 目前,学术界的很多研究开始利用社会标签进行个性化推荐<sup>[7-9]</sup>、信息检索等,都取得了良好的效果.

本文在充分考虑社会标签信息的基础上,试图在物理学的框架下给出一种新的个性化推荐算法. 该算法把目标用户喜欢项目特征视为物体颗粒,从而虚拟出目标用户喜欢的物体模型,根据项目特征构建项目物体模型. 然后考察用户喜好物体和项目物体间的万有引力的大小,并把它看作是二者相似性的度量,引力越大,二者就越相似,也就是用户喜欢该项目物体的可能性就越大,依此目标用户就可获得推荐.

## 1 相关定义

### 1.1 万有引力定律

万有引力作为自然界四种基本作用力之一,存在于宇宙中任何两个物体之间. 牛顿发现了这一引力并在1687年《自然哲学原理》上发表文章,提出了万有引力定律. 万有引力定律说明万有引力的大小和两物体质量的乘积成正比,与两物体间的距离的平方成反比:

$$F = G \frac{m_1 m_2}{r^2}. \quad (1)$$

式中:  $m_1$  和  $m_2$  分别为物体1和物体2的质量,  $r$  为两物体间的距离,  $G$  为万有引力常量,  $F$  为万有引力.

### 1.2 相关定义

基于万有引力的个性化推荐 (gravitation-based personalized recommendation, GBPR) 算法把物理学中的万有引力定律引入了推荐系统中,就需要把推荐系统中的概念映射到物理学的框架中来.

**定义1(项目物体(item object))** 推荐系统中的项目因其实际存在而被认为是物体,并被命名为项目物体. 项目物体与物理学中的物体一样,具有自身特有的属性以及质量.

**定义2(项目颗粒(item particle))** 项目颗粒是项目物体的组成成分. 项目颗粒有两个重要的要素,分别是质量和类型.

**定义3(项目物体的质量)** 项目物体的质量由组成它的项目颗粒的质量形成的,令  $q_{ig}$  表示项目物体  $i$  中第  $g$  个项目颗粒的质量,并且  $q_{ig} \geq 0$ , 则项目物体  $i$  的质量可用一个质量向量表示  $\mathbf{q}_i = (q_{i1}, q_{i2}, \dots, q_{ig})$ .

**定义4(项目颗粒引力(gravitation between item particle))** 项目颗粒间存在类万有引力的引力,并且其大小遵循万有引力定律. 这里的引力和物理学上的

引力的主要区别是,这一引力仅仅是一种数量上的度量,而没有方向. 同时还有同类型的项目颗粒间存在引力,不同类型的项目颗粒间则不存在引力.

**定义5(项目物体引力(gravitation between item object))** 项目物体间因其组成颗粒间的引力而存在着类万有引力的引力,并且该引力遵循万有引力定律. 同样的,该引力仅仅是数量上的度量而没有方向.

**引理1(叠加原理(superposition principle))** 两项目物体间的引力大小由项目物体的项目颗粒间的引力叠加形成的. 若项目物体  $i$  包含项目颗粒有  $(t_{i1}, t_{i2}, \dots, t_{ig})$ , 项目物体  $j$  包含的项目颗粒有  $(t_{j1}, t_{j2}, \dots, t_{jg})$ , 其中  $t_{ig}$  表示项目物体  $i$  中的第  $g$  个项目颗粒,  $t_{jg}$  表示项目物体  $j$  中的第  $g$  个项目颗粒,同类项目颗粒间的引力分别为  $(f_1, f_2, \dots, f_g)$ , 项目物体  $i$  和  $j$  间的引力为:

$$F_{ij} = \sum_{g=1}^g f_g. \quad (2)$$

式中  $f_g$  为两项目物体中的第  $g$  个项目颗粒之间的引力.

**定义6(引力场(gravitation field))** 推荐系统中的项目物体因引力,形成了一个充满整个数据空间的场,这个场叫做引力场.

**定义7(引力子场(gravitation subfield))** 存在于整个数据空间的引力场是由若干个引力子场组成. 每个项目物体因与其他物体间的引力而在其周围形成了一个小的引力场,该小场叫做引力子场.

对于每个引力子场,GBPR 算法重点考察其三个因素,分别是中心点、作用点和引力强度.

**定义8(引力子场中心点(centre of gravitation subfield))** 形成引力子场的点即为中心点,所以每个项目物体即为每个引力子场的中心点.

**定义9(引力作用点(action point of gravitation))** 引力场  $f_i$  会对场中项目物体  $j$  施加引力作用,则项目物体  $j$  为引力子场  $f_i$  的其中一个作用点.

**定义10(引力强度(gravitation strength))** 引力强度等于引力子场中心点和引力作用点对应的项目物体间的万有引力大小.

在 GBPR 算法中,引力强度考察的都是针对某个引力子场的某一作用点上的引力强度.

## 2 GBPR 算法

### 2.1 问题描述

当用户选择社会标签对项目进行标注后,形成了如图1所示“用户-项目-标签”网络. 假设系统用户集合为  $U = (u_1, u_2, \dots, u_n)$ ,  $n$  为用户总数,项目集合为  $I = (i_1, i_2, \dots, i_m)$ ,  $m$  为项目总数,标签集合为  $T = (t_1, t_2, \dots, t_g)$ ,  $g$  为标签种类数. 推荐算法就是利用三者之间的关系数据,计算目标用户对未选择项目的喜好程

度,从而把目标用户可能喜欢的项目推荐给他。

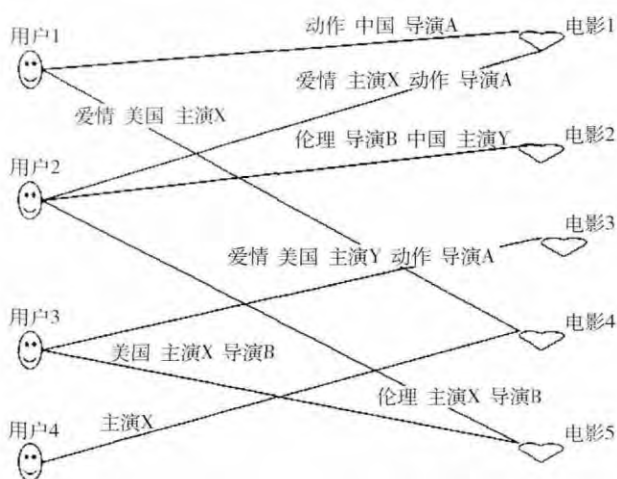


图1 社会标签系统示例

Fig.1 Example of the social tag system

为方便算法以后的计算,根据“用户-项目-标签”之间的关系图,给出几个矩阵的定义如下:

定义 11(用户-项目矩阵  $R$  (user-item matrix))

如果有  $n$  个用户  $U = (u_1, u_2, \dots, u_n)$  和  $m$  个项目  $I = (i_1, i_2, \dots, i_m)$ , 它们之间因选择关系形成一个  $n \times m$  的矩阵, 矩阵的行为用户, 列为项目, 当用户  $u_i$  选择了项目  $j$  并且评分为  $x$  时, 矩阵中的  $r_{ij} = x$ , 否则  $r_{ij} = 0$ .

定义 12(用户-标签频率矩阵  $A$  (user-tags frequency matrix)) 如果有  $n$  个用户  $U = (u_1, u_2, \dots, u_n)$  和  $g$  个标签  $T = (t_1, t_2, \dots, t_g)$ , 它们形成一个  $n \times g$  的矩阵, 矩阵的行为用户, 列为标签, 当用户  $u_i$  使用了  $z$  个标签  $t_g$ , 则  $a_{ig} = z$ , 否则  $a_{ig} = 0$ .

定义 13(标签-项目频率矩阵  $B$  (tags-item frequency matrix)) 如果有  $g$  个标签  $T = (t_1, t_2, \dots, t_g)$  和  $m$  个项目  $I = (i_1, i_2, \dots, i_m)$ , 它们形成一个  $g \times m$  的矩阵, 行为标签, 列为项目, 当有  $y$  个标签  $t_g$  标注了项目  $i$  时  $b_{gi} = y$ , 否则  $b_{gi} = 0$ .

## 2.2 BGPR 算法框架

基于项目间引力的考虑, BGPR 算法的框架如下。

(1) 构建目标用户的兴趣偏好模型。对目标用户  $u_i$  根据其使用的标签信息获取其兴趣偏好, 从而构建目标用户  $u_i$  的兴趣偏好模型。该模型也是虚拟的项目物体, 因为它反映了用户的喜好, 所以又被叫做喜好物体。

(2) 构建项目物体模型。根据对项目物体标注的标签信息, 构建出能反映项目自身属性特征的模型, 该模型叫做项目物体模型。

(3) 计算引力强度。在推荐系统中, 以目标用户  $u_i$  的喜好物体为中心点, 以  $u_i$  没有选择的项目物体为作用点, 建立  $u_i$  喜好物体的引力子场  $f_i$ , 计算  $f_i$  对各作

用点的引力强度。

(4) 相似度衡量。根据定义 4 和引理 1 可知  $f_i$  对各作用点的引力强度越大,  $u_i$  的喜好物体与项目物体的相似度越高, 所以引力强度可被视为一种相似度的度量。

(5) 获得推荐。对引力子场  $f_i$  的各作用点的引力强度倒序排列, 引力强度大的前  $N$  个作用点对应的项目物体可能是目标用户  $u_i$  比较喜欢的物体项目, 把它们推荐给目标用户  $u_i$ 。

## 2.3 用户喜好物体和项目物体模型

用户在对项目进行浏览和使用时, 依据个人对项目的理解, 选择简单的词语对自己喜欢的项目进行标注, 这些词语就是社会标签 (social tag)。BGPR 算法主要依据社会标签信息构建用户喜好物体和项目物体模型。

对目标用户而言, 他们使用社会标签对自己喜欢的项目进行分类和描述, 以方便对项目的浏览、组织和分享, 所以标签可以在一定程度上反映用户的喜好。比如用户 Jack 在对电影标注时经常使用“喜剧”、“爱情”、“战争”等标签, Jack 可能喜欢这类电影。如果 Jack 使用的标签中, “战争”类标签使用的次数很多, 那么他会更喜欢战争类的电影。也就是说用户使用的标签类别和频率可以反映用户的喜好特征。

对目标用户  $u_i$ , 用户使用的标签类别可以反映用户喜欢项目类别, 所以该用户使用过的标签被视为组成  $u_i$  喜好物体的项目颗粒。若推荐系统中有  $g$  类标签, 则用户  $u_i$  的喜好物体模型可用一个向量表示:

$$O_{u_i} = (p_1, p_2, \dots, p_g). \quad (3)$$

式中  $p_g$  为用户  $u_i$  使用的第  $g$  类标签的频率。

对项目而言, 社会标签是它们的分类信息, 即标签可以反映出项目的部分属性。比如“张艺谋”、“爱情”标签标注到“红高粱”这部电影上时, 可以知道“红高粱”这部电影为张艺谋执导的爱情类电影。如果某一部电影有很多用户使用了“爱情”这一标签, 那么该电影是爱情类电影的可能性就会更高。总的说来, 对目标标注的标签类别和频率可以从很大程度上反映项目的属性信息。

鉴于上述分析, BGPR 算法选择对项目标注的标签为项目物体的项目颗粒, 若推荐系统中有  $g$  类标签, 则项目物体  $i$  的模型可用一个向量表示:

$$O_i = (s_1, s_2, \dots, s_g). \quad (4)$$

式中  $s_g$  为项目物体  $i$  接收到的第  $g$  类标签的频率。

## 2.4 引力强度的计算

以用户  $u_i$  的喜好物体  $O_{u_i}$  为中心点的引力场为例来说明引力强度的计算。喜好物体  $O_{u_i}$  对作用点  $j$  (即项目物体  $j$ ) 引力强度等于两物体之间的万有引力强度。

根据万有引力的计算方法,如式(1)所示,首先要计算两物体的质量.项目物体和喜好物体都是由项目颗粒组成,它们的质量由组成它的项目颗粒的质量决定的.GBPR算法把社会标签视为项目颗粒,项目颗粒的质量等同于社会标签对其标注的项目物体的重要程度;同理,喜好物体包含的项目颗粒等同于用户使用的标签对自己的重要程度.

由于很多项目颗粒可能存在于不同的项目物体中,它们在不同的项目物体中时,其重要性可能是不相同的,有必要定义它们独立于具体项目物体的质量,即平均质量.GBPR算法中,项目逆频率被定义为颗粒的平均质量,即

$$\overline{q(p_g)} = \ln\left(\frac{m}{\ln_g}\right). \quad (5)$$

式中: $\overline{q(p_g)}$ 是第 $g$ 个项目颗粒的平均质量; $m$ 推荐系统中的项目总数; $\ln_g$ 是项目颗粒 $g$ 的总数,即社会标签的总数.

项目物体 $j$ 中的第 $g$ 个项目颗粒的质量,即第 $g$ 个项目颗粒对项目物体 $j$ 的重要程度,计算方法如下:

$$q_{jg} = w(p_g, i_j) \times \overline{q(p_g)}. \quad (6)$$

其中 $q_{jg}$ 是项目物体 $j$ 中的第 $g$ 个项目颗粒的质量, $w(p_g, i_j)$ 是颗粒 $g$ 对项目物体 $j$ 的重要性参数, $\overline{q(p_g)}$ 是第 $g$ 个项目颗粒的平均质量.

式(6)中的重要性参数采用(TF × IDF (term frequency-inverse document frequency))方法来计算<sup>[10-11]</sup>.该方法通常用来评价一词汇对一个文件集中某分文件的重要程度,其中TF (term frequency),即词频,在本文中用来表示项目颗粒的在某项目物体中出现频率,如果 $\ln_{i_j}^j$ 是项目物体 $j$ 中项目颗粒 $g$ 的数目, $\ln_r^j$ 是项目物体 $j$ 中所有项目颗粒的总数,那么TF的计算式为

$$\text{TF} = \frac{\ln_{i_j}^j}{\ln_r^j}. \quad (7)$$

IDF (inverse document frequency),即逆向文件频率,在本文用来表示项目颗粒的类别区分能力.如果 $m$ 是项目物体总数, $\ln_{i_g}$ 是包含第 $g$ 个项目颗粒的项目物体总数,则IDF的计算式为

$$\text{IDF} = \log\left(\frac{m}{\ln_{i_g}}\right). \quad (8)$$

根据定义3可得项目物体 $j$ 的质量向量 $q_j$ ,同理可得用户 $u_i$ 喜好物体的质量向量 $q_{u_i}$ .

然后要计算用户喜好物体和项目物体之间的距离.欧几里得距离又叫欧式距离,通常用来计算两向量之间的距离,也可以认为是二者之间的差距.GBPR算法计算用户 $u_i$ 使用的标签频率向量和标注项目 $j$ 的标签频率向量间的欧式距离,并把该距离作为用户 $u_i$ 的喜好物体和项目物体 $j$ 之间差距:

$$d_{ju} = \sqrt{\sum_{i=1}^g (b'_{jg} - a_{jg})^2}. \quad (9)$$

式中 $b'_{jg}$ 为矩阵 $B$ 第 $j$ 行的转置, $a_{jg}$ 为矩阵 $A$ 中用户 $u_i$ 对应的第 $j$ 行, $d_{ju}$ 为喜好物体和项目物体 $j$ 间的距离.

那么,项目物体 $j$ 和用户 $u_i$ 的喜好物体间的引力强度的计算如式如下:

$$F_{ji} = \frac{q_j \cdot q_{u_i}}{(d_{ju})^2}. \quad (10)$$

其中 $F_{ji}$ 为项目物体 $j$ 和用户 $u_i$ 喜好物体间的引力强度.

## 2.5 获得推荐

依据定义4和引理1,项目物体 $j$ 和目标用户 $u_i$ 喜好物体间的引力强度可以被认为是二者的相似度,它们之间的引力强度越大,目标用户的喜好物体中包含的项目颗粒和相应项目物体中相同类型的项目颗粒含量相同的就越多,那么该项目物体就越有可能被用户喜欢.

对目标用户 $u_i$ ,倒序排列其引力子场中各作用点的引力强度,前 $N$ 个作用点对应的项目物体被认为是用户可能喜欢的,从而把它们推荐给目标用户 $u_i$ .

## 3 实验结果

### 3.1 数据集和评价矩阵

本文采用的数据集为MovieLens.该数据集为GroupLens小组整理所得,从网站www.groupLens.org下载.由于该数据集被大多数推荐系统测试使用,所以用它来测试的算法性能比较具有说服力.当然,本文采用MovieLens数据集中带有tag的数据集.该数据集中用户数为2113,电影数为10197,标签数为13222.随即选择数据集中的20%作为测试集,80%作为训练集来进行测试.

由于该算法预测结果的依据是用户的喜好物体和项目物体间的引力强度,也即推荐结果没有项目的评分,故测试的性能选择为准确率(Precision)和排名准确率(Hit\_rank).

准确度为用户喜欢的项目被预测正确的比例,计算方法如下<sup>[4,12-13]</sup>:

$$\text{Precision} = \frac{\text{Hits}}{N}. \quad (11)$$

式中,Hits为预测的命中个数, $N$ 为推荐个数.

排名准确率为推荐正确的项目在推荐列表中排名情况的度量,排名越靠前说明推荐效果越好,计算方法如下<sup>[4]</sup>:

$$\text{Hit\_rank} = \frac{1}{n} \sum_{i=1}^h \frac{1}{c_i}. \quad (12)$$

式中 $n$ 为用户数, $h$ 推荐列表中命中的个数, $c_i$ 为排名位置.

### 3.2 实验结果和结论

协同过滤推荐是推荐算法中最获得认可的算法之一,所以本文把 BGRP 算法和协同过滤推荐中的基于用户的推荐和基于项目的推荐性能进行对比。同时,最近提出了很多基于标签的协同过滤推荐,从中选择一个较好算法的性能也拿来作性能的对比。在实验时,随即抽取 10 个目标用户,Top- $N$  中  $N$  取 30 个,在不同的运算规模下计算其推荐性能,最后取每一运算规模下的平均值来获得最后结果。图 2 是精确度的对比,图 3 是排名准确度的对比。

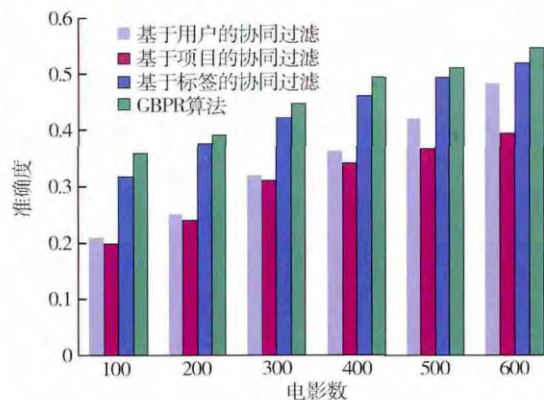


图2 精确度对比

Fig.2 Comparison of precision

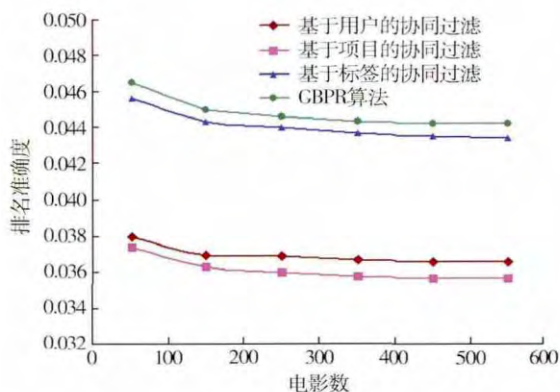


图3 排名准确度对比

Fig.3 Comparison of hit\_rank

GBPR 算法与其他的推荐算法相比复杂性也降低了很多,无论是时间复杂度还是空间复杂度都较低。一旦确定目标用户,需要存储一个长度为  $t$  (社会标签个数) 的一维向量,存储空间较其他算法大大减少。并且需要的存储空间仅仅随着社会标签个数的增加而增加,与系统中用户个数和项目个数无关。这一点对算法复杂度的影响很关键,因为在推荐系统实际应用时,用户的数量以及用户感兴趣的项目数量都会迅速增长,其增长速度会远远超过其他各项的增长速度。

在进行推荐计算的过程中,喜好物体和项目物体间万有引力计算也仅仅与标签数量相关,所得的项目间质量和引力矩阵在实际应用中为稀疏性矩阵,大部

分元素为 0,所以其计算的时间复杂度和标签数量的平方成正比。在对目标用户推荐时,所需就是用户的喜好向量,并且该向量的元素大多数为 0,所以推荐时所需的计算复杂度和其他的算法比也较低。

### 参 考 文 献

- [1] Xu H L, Wu X, Li X D, et al. Comparison study of internet recommendation system. *J Software*, 2009, 20(2): 350 (许海玲, 吴潇, 李晓东, 等. 互联网推荐系统比较研究. 软件学报, 2009, 20(2): 350)
- [2] Wang G X, Liu H P. Survey of personalized recommendation system. *Comput Eng Appl*, 2012, 48(7): 66 (王国霞, 刘贺平. 个性化推荐系统综述. 计算机工程与应用, 2012, 48(7): 66)
- [3] Jin Y A. *Research on Technologies and Methods of Social Tag Recommendation* [Dissertation]. Wuhan: Huazhong University of Science and Technology, 2011 (靳延安. 社会标签推荐技术与方法研究[学位论文]. 武汉: 华中科技大学, 2011)
- [4] Zheng N, Li Q D. A recommender system based on tag and time information for social tagging systems. *Expert Syst Appl*, 2011, 38: 4575
- [5] Kim H N, Ji A T, Ha I, et al. Collaborative filtering based on collaborative tagging for enhancing the quality of recommendation. *Electron Commer Res Appl*, 2010, 9(1): 73
- [6] Duraõ F, Dolog P. A personalized tag-based recommendation in social web systems // *Workshop on Adaptation and Personalization for Web 2.0, UMAP09*. Trento, 2009: 22
- [7] Wang W P, Zhang L J. Collaborative filtering based on similarity fusion of tag and rating under the background of SNS. *Comput Syst Appl*, 2011, 20(10): 78 (王卫平, 张丽君. SNS 背景下基于 Tag 和 Rating 相似度融合的协同过滤. 计算机系统应用, 2011, 20(10): 78)
- [8] Feng W, Wang J Y. Incorporating heterogeneous information for personalized tag recommendation in social tagging systems // *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Beijing, 2012: 12
- [9] Adomavicius G, Sankaranarayanan R, Sen S, et al. Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Trans Inf Syst*, 2005, 23(1): 103
- [10] Yan X L, Liu Y Q, Ma S P, et al. Study on website keyword extraction for browsing recommendation. *CAAI Trans Intell Syst*, 2013, 7(5): 398 (闫兴龙, 刘亦群, 马少平, 等. 面向浏览推荐的网页关键词提取. 智能系统学报, 2013, 7(5): 398)
- [11] Song Y, Liu Z L, Li L J. Research on social tag recommendation techniques based on content. *J Harbin Inst Technol New Ser*, 2013, 20(2): 74
- [12] David M, Pennock E H, Lee C. Social choice theory and recommender systems: analysis of the axiomatic foundations of collaborative filtering // *Proceedings of the Seventeenth National Conference on Artificial Intelligence*. Austin, 2000
- [13] Kristina Lerman. Social networks and social information filtering on digg // *Proceedings of the Int'l Conference on Weblogs and Social Media*. Washington D. C., 2006