情报工程 ISSN 2095-915X
第2卷 第2期 81-87 2016年4月

TECHNOLOGY INTELLIGENCE ENGINEERING
ISSN 2095-915X Vol.2 No.2 81-87 Apr 2016

doi:10.3772/j.issn.2095-915x.2016.02.010

# 基于协同过滤算法的微博平台的信息推荐研究

王迪，王东雨

（河北大学管理学院 保定 071002）

摘要：将改进的协同过滤算法应用于微博平台的信息推荐，拓展微博算法的应用范围，增加微博平台的可用性，提高信息推荐的结果准确性，更好地满足用户的信息需求。首先分析协同过滤技术及其如何应用于微博信息推荐，并将基于微博文本特征的推荐算法与传统的推荐算法相对比，再融入微博用户兴趣度，得出更优的推荐算法。运用改进的协同过滤算法提高微博平台的信息推荐质量，使微博平台信息推荐更加精准、有效。

关键字：协同过滤算法，微博平台，信息推荐

# Research on the Information Recommendation of Micro-blog Platform Based on the Collaborative Filtering Algorithm

WANG Di, WANG DongYu

**( School of Management, Hebei University, Baoding 071002, China )**

**Abstract:** The improved collaborative filtering algorithm is applied to information recommendation of micro-blog platform,which is aimed to expand the range of application of the algorithm and enhance the applicability of the micro-blog platform. The collaborative filtering algorithm also intends to improve the accuracy of informative

---

作者简介：王迪（1993-），女，硕士，研究方向：个性化推荐，Email：1583368426@qq.com，联系电话:18730253937；王东雨（1991-），男，硕士，研究方向：个性化推荐、智慧医疗，Email:389011848@qq.com，联系电话:18730256175。

recommendation and meet the information demands of users. This research analyzed the collaborative filtering techniques and applied it to the micro-blog recommendation. Comparing with the traditional recommendation algorithm, this study added the micro-blog users' interest into the collaborative filtering algorithm, and applied it to improve the quality of information recommendation of the micro-blog platform.

**Key words:** Collaborative filtering algorithm, micro-blog platform, information recommendation

## 引言

随着移动终端和通讯技术的发展，以微博为代表的新兴媒体平台不断涌现，它们具有时效性强、更新速度快和信息分享便捷等特点更好的满足了人们的信息需求，已成为企业和网民互动沟通的新平台[1]。据中国互联网信息中心（CNNIC）于 2015 年 7 月 23 日在京发布的第 36 次《中国互联网络发展状况统计报告》显示，截至 2015 年 6 月，我国网民规模达 6.68 亿，互联网普及率为 48.8%，微博用户 2.04 亿，网民使用率为 30.6%[2]。

微博平台虽然满足了用户的信息需求，但用户准入门槛低、信息更新快、传播速度强等特征致使信息量庞大冗杂、信息需求者难以有效处理和利用信息。传统的微博平台信息推荐方法主要有两种：一是基于关键词精准匹配的信息推荐，即从微博信息中获取关键词并向量化表示，利用传统的向量间相似度的计算方法得出推荐内容；二是基于关键词正规匹配的信息推荐，即通过分析微博内容的文字特征，构造出相应的正规表达式，利用表达式匹配相应的推荐内容[3]。两种方法都是基于微博信息本身的，而微博信息本身字数少、结构弱、稀疏性强等特征使推荐系统很难与微博用户需求的多样性相结合，最终的推荐结果往往很难满足用户个性化的需求。

鉴于微博本身是信息特征稀疏的短文本，推荐过程中需要提取用户的兴趣特征融合到算法中，这对于协同过滤算法的改进和应用提出更新的挑战。为了提高信息推荐的准确性，使微博更好的服务于用户，本文通过分析微博信息中文档特征的表示方法，并研究新的适用于计算微博短文本信息的相似度方法，再融合微博用户兴趣度，在此基础上使基于微博平台的协同过滤算法在微博信息推荐中得以优化。

## 1　基于协同过滤算法的信息推荐技术

协同过滤推荐算法主要包括两种：基于用户（User–based）协同过滤算法和基于项目（Item–based）协同过滤算法[4–5]，都是通过寻找用户间或项目间相关性向用户进行推荐。User–based 协同过滤推荐根据用户对某一内容开展的评价对目标用户进行内容推荐，最终产生的推荐内容是符合用户兴趣且质量较高；而 Item–based 协同过滤推荐是根据用户对其他内容的评价来预测该用户对目标内容的评分，进而得出推荐列表，有助于发现新的信息内容[6]。

以基于用户的协同过滤算法为例，首先构建用户档案，将用户为每个资源项目的评价、浏览等记录转换成用户对各项目的评价矩阵，如下表所示：

表 1 用户—项目评分表

| | 项 目 | | | |
|---|---|---|---|---|
| 用户 | $Item_1$ | $Item_2$ | ··· | $Item_n$ |
| $User_1$ | $r_{11}$ | $r_{12}$ | ··· | $r_{1n}$ |
| $User_2$ | $r_{21}$ | $r_{22}$ | ··· | $r_{2n}$ |
| ··· | ··· | ··· | ··· | ··· |
| $User_m$ | $r_{m1}$ | $r_{m2}$ | ··· | $r_{mn}$ |

其中，$r_{mn}$代表第 m 个用户 U 对项目 N 的评分，分数越高则用户对该项目越认可，所有评分记录构成如上评分矩阵；然后通过计算不同用户的评分之间的相似程度，寻找相似度最高的用户作为最近邻居集，相似性计算可通过向量间的余弦夹角度量或 Pearson 相关系数度量[7]，公式如下：

$$（1）\ Sim(u_1,u_2)=\cos(\vec{n}_1,\ \vec{n}_2)=\frac{\vec{n}_1\bullet\vec{n}_2}{|\vec{n}_1|\bullet|\vec{n}_2|}$$

用户 $u_1 u_2$ 同时对项目 n 进行过评分，用两个 n 维向量 $\vec{n}_1\vec{n}_2$ 表示；

$$（2）\ Sim(i,j)=\frac{\sum_{c\in Iij}\left(R_{c,i}-\bar{R}_i\right)\left(R_{c,j}-\bar{R}_j\right)}{\sqrt{\sum_{c\in Iij}\left(R_{c,i}-\bar{R}_i\right)^2}\sqrt{\sum_{c\in Iij}\left(R_{c,j}-\bar{R}_j\right)^2}}$$

$R_{ci}$表示用户 c 对项目 i 的评分，$R_{cj}$表示用户 c 对项 j 的评分，$\bar{R}_i\ \bar{R}_j$可分别表示对项目 i 和项目 j 的平均评分。最后，根据最近邻居的评价值产生最佳推荐结果。这是协同过滤算法的基本步骤[8]。

## 2 传统的文本建模与基于 TF-IDF 的协同过滤算法

文本特征表示是传统文本信息挖掘的重要内容，也是互联网信息的重要研究项目。传统的文本特征表示方法主要基于布尔模型、向量空间模型和 TF-IDF 模型等[3]。

布尔模型基于二元判定标准，定义了一个二值的变量集合，其中的变量对应文本的特征项。特征项在文中出现时变量赋值为 1，若不出现则为 0。该模型虽然形式简单、语义精确，但限制了文档分级而且信息需求转化为二元结构的操作性不强，致使模型可行性不高。

向量空间模型是把文本映射为特征空间中的由维度和权值组成的特征向量，特征向量 $V(d)=(t_1,w_{1(d)};\cdots;t_n,w_{n(d)})$。其中 $t_i(i=1,2,\ldots,n)$ 为一列互不相同的特征项，$w_i(d)$为$t_i$在 d 中的权值，一般被定义为$t_i$在 d 中出现频率[9]。这样文档集合就可以表示为一个矩阵，矩阵的每一行代表一个文档文本，每一列代表当前文档中的各特征项。通过计算文本特征向量间的距离来度量文本间的相似度。向量空间模型在文本中的执行性、计算性和匹配效率等方面得到了显著提高。但其权重仅以特征向量出现频率来计量会使部分特征项的重要程度出现偏差，而且往往也存在特征向量的维度过高的问题。

TF-IDF 模型是基于向量空间模型中特征项权重以出现频率为计算方式的改进。词频 (TF) 指的是某一个给定的词语在该文件中出现的频率，表示为 $tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}}$，该式分子是某词在文中出现的次数，而分母则是文件中所有字词

表 2　向量空间模型图

| | $t_1$ | $t_2$ | … | $t_m$ |
|---|---|---|---|---|
| $D_1$ | $w_{11}$ | $w_{12}$ | … | $w_{1n}$ |
| $D_2$ | $w_{21}$ | $w_{22}$ | … | $w_{2n}$ |
| … | … | … | … | … |
| $D_n$ | $w_{n1}$ | $w_{n2}$ | … | $w_{nm}$ |

出现次数之和；逆向文件频率 (IDF) 是由总文件数目与包含该词语的文件的数目之比，再取对数得到的，是词语普遍重要性的度量，表示为 $idf_i = \log\frac{|D|}{|\{j: t_i \in d_i\}|}$，如果该词语不在语料库中，就会导致分母为零，所以一般情况下取上式分母部分加 1 作为真正的分母，最后取 TF 与 IDF 之积，从而过滤掉常见的特征项，使特征项权重更合理，进而通过计算文本间的共现词汇来衡量文本之间的相似度，最终得出推荐信息。

相比上述传统的长文本，微博短文本仅含有极少的通用词汇，不同的微博文本之间的共现词汇也是有限的 [2]，这就使得传统的文本建模方法难以适用于微博短文本间的相似性判断，因此微博平台信息推荐应该针对短文本特征，在传统文本自然语言处理方法的基础上，结合互联网信息推荐技术，以提高信息推荐结果的准确性。

## 3　微博短文本特征表示及基于协同过滤算法的微博短文本信息推荐

微博短文本虽语义表达清楚，但很多忽略语法结构，大量出现同近义词，无论是汉语表达还是英语表达，都有类似的现象。而 Word Net 和同义词词林作为英、汉词汇的知识库，清晰呈现了词汇概念之间以及概念属性间的关系，成为处理英汉语中同近义词的理想词库。Nugget 作为 Word Net 和同义词词林的概念集，可作为标准的词库源，更好地适用于微博短文本的表示与关系模型构建。

首先对微博文本的语料库进行预处理，从而得到微博语料库的原始词项集合，然后创建面向微博语料库的词项集合的 Nugget。具体如下：

（1）读取微博语料库词项集合中的词项 $w_i$，并依次遍历 Word Net 中的同义词集，当词项首次出现在 Word Net 的同义词集中时，抽取同义词集并建立对应的 Nugget，即 $N_i$；若 Word Net 的同义词集中没有时，构建对应的 Nugget 即 $N_j$，只包含当前词项 $w_j$；

（2）继续下一个词项 $w_{i+1}$，并遍历当前的 Nugget 集，若包含此项，继续下一项，直到微博语料库的词项空间读取完毕。若不包含此项，则重复第一步。

Nugget 集合创建好后，便是基于 Nugget 完成对微博信息的文本特征建模。

（1）将每条微博文本 i 用向量 $\vec{n}_i$ 表示，向量 $\vec{n}_i$ 的每个元素 $n_{i,j}$ 表示词项 $w_j$ 在当前微博文本中出现的次数；

（2）将元素 $n_{i,j}$ 转换为 $n_{i,jk}$，其中 k 表示词项 $w_j$ 所属的 Nugget 的索引标识；由于每个词项所含信息不尽相同，所以我们需要计算每个 Nugget 中的词项各自的权重值 p（$w_j$，k），公式如下：$p(w_j,k) = \frac{\sum_{i\in R} w_{i,jk}}{\sum_{i\in R, w_j\in N_k} w_{i,jk}}$。其中 R 表示微博文本的

语料库集合，$w_{i,jk}$表示$w_{jk}$在微博文本 i 中出现的次数；

（3）根据词项的权值，运用公式将微博文本特征向量$\vec{n}_i$转换为基于 Nugget 的新的文本特征向量$\vec{u}_i$，该向量中的元素记为$u_{i,k}$，k 表示 Nugget 集合中 Nugget 的序列标识。运用的公式如下：

$$u_{i,j}=\sum\nolimits_{j\in N_k}T_{i,j,k}$$

$$T_{i,j,k}=\begin{cases}\ln\left(w_{i,j,k}+1\right)\times p\left(w_j,k\right), & w_{i,j,k}\neq 0 \quad ;\\ 0\ , & w_{i,j,k}=0.\end{cases}$$

至此完成了基于 Nugget 集合的向量表达转换，使微博短文本特征进行了有效的向量表示。从原始的微博文本词项空间转换为基于 Nugget 集合的词项空间，可以有效处理微博文本中的同、近义词项，也能达到降维的效果，为短文本信息推荐奠定重要基础。

根据协同过滤算法的基本步骤，接下来计算微博短文本间的相似度。假设两条基于 Nugget 特征空间的微博短文本分别为

$$w=（t_1，t_2，\ldots，t_n）、w'=（t_1'，t_2'，\ldots，t_n'），$$

采用余弦系数法相似度计算方法，公式如下：

$$Sim(w,w')=\frac{\sum_{i=1}^{n}t_i\times t_i'}{\sqrt{\sum_{i=1}^{n}t_i^2\times\sum_{i=1}^{n}t_i'^2}}$$

可以得出当前微博用户的微博文本基于 Nugget 的近邻集合，鉴于微博文本的类型特征和时间特性等因素对信息推荐准确性的影响，在上式基础上融入调节参数 θ，θ 的计算公式如下：

$$\theta=\frac{1}{1+a(date(now)-date(t))}\times flag$$

其中，$date(now)$ 表示当前日期，$date(t)$ 表示用户微博的发表日期，$date(now)-date(t)$ 表示当前用户的微博的发表天数，$\alpha$ 是衰减参数，根据训练集数据可以学习其具体的参数值，$flag$ 表示用户发表微博的类型，如“原创类”、“转发并评论类”和“转发未评论类”，三类的权值由高到低。

最后在两条微博文本进行余弦相似度计算结果的基础上加入调节参数 θ，即余弦系数法相似度计算公式调整为：

$$Sim(w,w')=\frac{\sum_{i=1}^{n}t_i\times t_i'}{\sqrt{\sum_{i=1}^{n}t_i^2\times\sum_{i=1}^{n}t_i'^2}}\times\theta$$

将计算结果由高到低进行排序，系统便可将排列在前的 N 条微博信息作为推荐结果反馈给当前用户，从而完成基于协同过滤算法的微博短文本信息推荐[10]。

## 4　融合用户兴趣度的微博信息推荐

上述微博信息推荐是从微博文本内容的角度出发，而微博用户是微博信息的最终寻求者，他们对信息推荐也是重要的影响因素，通过对用户及用户好友之间关系的深入研究，提出了一种融合用户兴趣度的微博信息推荐方法[11]。

定义用户 U 对微博 W 的兴趣度为$p_{uw}$，那么影响$p_{uw}$大小的因素主要包括微博用户 U 与其关注的好友 V 之间的熟悉程度$f(u,v)$、微博用户 U 与其关注的好友 V 之间的兴趣相似度$Sim(u,v)$以及微博用户 U 所关注的好友 V 对微博 W 的兴趣度$r_{vw}$[12]。

如果微博用户 U 和微博用户 V 很熟悉，那么一般情况下他们应该会有较多的共同好友，基于这一事实，假设用户 U 的好友集合用$friend(u)$表示，那么计算公式如下：

$$f(u,v)=\frac{|friend(u)\cap friend(v)|}{|friend(u)\cup friend(v)|}$$

如果微博用户 U 和微博用户 V 的兴趣相似度高，那么一般情况下他们共同喜欢的微博数量较多，基于这一个事实，我们假设用户 U 喜欢的微博集合用$C(u)$表示，那么计算公式如下：

$$Sim(u,v)=\frac{|C(u)\cap C(v)|}{|C(u)\cup C(v)|}$$

如果微博用户 V 是微博 w 的创建者，此时令 $r_{vw}$ =1；当用户 V 对微博 w 是“评论并转发”的，此时令 $r_{vw}$ =0.9；当用户 V 对微博 w 只是“转发未评论”时，此时令 $r_{vw}$ =0.7；当用户 V 对微博 w 是“评论未转发”时，此时令 $r_{vw}$ =0.5；而如果用户没有对微博 w 有任何行为，则令 $r_{vw}$ =0。得到以上度量数据后，便可得用户 U 对微博 W 的兴趣度 $p_{uw}$，公式如下：

$$p_{uw}=\sum_{v\in friend(u)} r_{vw}\times f(u,v)\times Sim(u,v)$$

计算结果进行排序便可以给微博用户推荐其信息墙上其他好友微博信息中感兴趣的前几条微博，这也是个性化推荐应用的重要表现[13]。

## 5 改进的协同过滤算法在微博信息推荐中的效果检验

微博在提供推荐服务时，一般是给用户推送个性化列表，这种推荐叫做 TopN 推荐。TopN 推荐的预测准确率一般通过准确率（precision）/召回率（recall）度量[12], $F_1$是由两率构成的综合性评价指标。

$$\Pr ecision=\frac{|R(u)\cap T(u)|}{|R(u)|},$$

$$\operatorname{Re}call=\frac{|R(u)\cap T(u)|}{|T(u)|},$$

$$F_1=\frac{2\times(\Pr ecision\times\operatorname{Re}call)}{\Pr ecision+\operatorname{Re}call}$$

其中，$R(u)$是推荐算法根据训练集上的数据在测试集上给用户 U 的推荐结果列表，$T(u)$是用户 U 在测试集上的实际行为结果。

本实验的训练数据集是收集到的整个语料库共计 10000 条微博数据记录，整理后包含的数据项有用户粉丝列表、关注列表、微博正文、评论者列表、发布时间及是否转发。测试集是应用推荐算法的微博用户数据集及语料库中统计出发布微博数量最高的 50 个微博用户的相关数据。为了全面评测 *TopN* 推荐的准确率和召回率，一般会选取不同的推荐列表长度 N（*N=10*，*20*，*30*）。

首先对传统的基于 TF–IDF 的协同过滤算法进行试验，对测试集数据构建基于 TF–IDF 的向量空间，然后与微博用户的微博文本向量进行相似度计算。对最终推荐结果的评价如下表所示：

**表 1 基于 TF-IDF 的协同过江算法实验结果**

| 评价指标 | precision | recall | $F_1$ |
|---|---|---|---|
| N=10 | 10% | 3.3% | 4.8% |
| N=20 | 10% | 6.8% | 8.0% |
| N=30 | 6.7% | 6.7% | 6.7% |

基于 Nugget 近邻的微博短文本推荐算法，首先根据微博语料库和同义词林建立 Nugget 集合，依据 Nugget 集对测试集的微博数据构建向量空间，然后将微博用户的微博数据与测试集的其他微博数据进行余弦相似度计算，加入时间衰减和类型参数，计算得到推荐结果。各评价指标如下表所示：

**表 2 基于 Nugget 近邻的微博短文本推荐算法实验结果**

| 评价指标 | precision | recall | $F_1$ |
|---|---|---|---|
| N=10 | 20% | 6.7% | 9.9% |
| N=20 | 15% | 10% | 12% |
| N=30 | 13.3% | 13.3% | 13.3% |

融合用户兴趣度的微博信息推荐算法是基于微博用户的好友列表，提取其好友的相关信息，然后根据兴趣度计算公式得到与微博用户兴趣最相近的几个好友，将这些好友的新发微博推荐给当前微博用户。对算法给出的好友兴趣相似度结果的评价结果如下表所示（以 10 个好友为准）：

表 3 融合用户兴趣度的微博信息推荐算法实验结果

| 评价指标 | precision | recall | $F_1$ |
|---|---|---|---|
| N=10 | 25% | 25% | 25% |

实验对传统的基于 TF-IDF 的协同过滤算法和基于 Nugget 近邻的微博短文本推荐算法再融合用户兴趣度进行比较，可以看到基于 TF-IDF 的协同过滤算法的评价结果偏低，利用 Nugget 集构建微博的文本向量后，计算结果的精度较传统的方法有所提高，而且融合用户兴趣度的微博信息推荐算法可以得到相对满意的推荐结果。

## 6 总结

随着互联网技术的发展，微博应用范围不断扩大，如何更好地为微博用户推荐有用的微博信息以提高微博服务质量是本文研究的出发点和落脚点。本文首先提出协同过滤技术的应用方法以及对微博信息推荐的适用性；然后将传统文本表示方法及推荐算法同微博短文本表示方法及推荐算法加以对比，突出文本构建方式以及协同过滤算法在微博应用中的改进和提高；最后融入微博用户的兴趣度，是对以微博信息文本内容的相似度作为推荐依据的有效补充，提高了协同过滤算法在微博信息推荐中的有效性和信息推荐的准确性。

当然，本文局限于协同过滤算法中的一种算法在微博信息推荐中的应用，对于多种算法混合运用还有待进一步探讨；而且文中参照标准的选择具有主观性，数据支撑力度不够，结论的实验检验还有待规范，这是本文的主要不足之处。但改进的协同过滤算法应用于微博信息推荐，有助于提高推荐结果的准确性，改善微博服务质量，使微博用户更充分享受技术成果带来的利益。

### 参考文献

[1] 闫幸，常亚平．微博研究综述 [J]. 情报杂志，011(9):61-62.

[2] 第 36 次中国互联网络发展状况统计报告 [EB/OL]. http://news.xinhuanet.com/politics/2015-07/23/c_128051995.html, 2015-07-23.

[3] 朱亚涛．基于微博平台的信息推荐技术研究 [D]. 首都师范大学，2013.

[4] Herlocker J L, Konstan J A, Terveen L G, et al. Evaluating collaborative filtering recommender systems[J]. ACM Transactions on Information Systems, 2004, 22(1):5-53.

[5] 罗文．协同过滤推荐算法综述 [J]. 科技传播，2015(4):115-196.

[6] 郭艳红．推荐系统的协同过滤算法与应用研究 [D]. 大连理工大学，2008.

[7] 马宏伟，张光卫，李鹏．协同过滤推荐算法综述 [J]. 小型微型计算机系统，2009(7):1283-1284.

[8] Herlocker, Jonathan L, Konstan, Joseph A, Riedl, John. Explaining collaborative filtering recommendations[C]// ACM Conference on Computer Supported Cooperative Work. ACM, 2001:5-53.

[9] 黄正．协同过滤推荐算法综述 [J]. 价值工程，2012(7):226-227.

[10] Goldberg D, Nichols D, Oki B M, et al. Using collaborative filtering to weave an information tapestry[C]// Communications of the ACM, 1992:61-70.

[11] 孙多．基于兴趣度的聚类协同过滤推荐系统的设计 [J]. 安徽大学学报（自然科学版），2007, 31(5):19-22.

[12] 项亮．推荐系统实践 [M]. 人民邮电出版社，2012.

[13] 王国霞，刘贺平．个性化推荐系统 [J]. 计算机工程与应用，2012, 2012(7):66-68.