DOI:10.3724/SP.J.1238.2013.00683

基于内容过滤推荐的农业信息推荐模型研究

丁德红，方逵*，王娟，朱幸辉

(湖南农业大学信息科学技术学院，湖南 长沙 410128)

摘 要：针对专门的农业知识库，使用基于内容过滤的推荐方法，建立了农民用户兴趣模型和文档特征模型。在用户兴趣模型和文档特征模型中，针对特征项在不同表空间的分布情况，以及 HTML 文档结构对特征项权重的影响，通过改进传统特征项提取算法，提高了推荐模型的精度。结果表明，随着用户数的增加，农业信息推荐模型的查准率和查全率不断加大，说明模型的精确度不断提高。

关 键 词：农业信息推荐模型；内容过滤推荐；特征提取；相似度

中图分类号：TP182 文献标志码：A 文章编号：1007–1032(2013)06–0683–05

Research on agricultural information recommendation models based on content filtering

DING De-hong, FANG Kui *, WANG Juan, ZHU Xing-hui

(College of Information Science and Technology, Hunan Agricultural University, Changsha 410128, China)

Abstract: The adaptive recommendation models in the agriculture information service platform are very important, which provides personalized recommendation information to farmers. Aiming at the special agricultural knowledge bases, by using the recommended methods based on content filtering, we have established the farmer interest model and the document feature model. In the two models, taking account of the influence of distribution of features in the different table space and the effect of HTML document structure on the feature weights, we have improved the accuracy of recommendation model by improving the traditional feature extraction algorithm. The experimental results show that with the increasing number of users of agricultural information recommendation models, the precision and recall rate of them are also increasing, the accuracy of them are also rising.

Key words: agricultural information recommendation model; content filtering recommendation; feature extraction; similarity

1992年，Goldberg等[1]首次提出协同过滤(collaborative filtering，CF)概念。自Rich设计了第1个协同过滤推荐系统之后，相继出现了不少类似的协同过滤推荐系统[2–3]。协同过滤推荐成为目前使用比较广泛的一类推荐技术，它的主要优势是不依赖于项目内容，而通过计算用户行为之间的相似度直接进行推荐，但仍存在稀疏性、冷启动、扩展性、灰羊问题[4–5] 等不足。张峰[6]提出使用BP神经网络来解决稀疏性问题；吴颜[7]提出运用单值分解、聚类的协同过滤算法来解决推荐系统中数据稀疏性的问题；何安[8]提出一种改进的协同过滤方法——基于协同过滤和聚类的组合推荐算法；王娟等[9]曾提出过1个农业信息协同推荐系统的构建方案，对协同过滤推荐系统进行了优化，较好地解决了用户评价矩阵的稀疏性问题。

与协同过滤推荐不同的是，内容过滤推荐是基

收稿日期：2013–07–02

基金项目：国家“十二·五”科技计划项目(2011BAD21B03)；湖南省科技重大专项(2010FJ1006)；湖南省国家农业与农村信息化科技示范省建设项目(2011GA770001)

作者简介：丁德红(1975—)，男，湖南常德人，博士研究生，高级工程师，主要从事农业信息工程研究，dingdehong@qq.com；*通信作者，fk@hunau.net

于信息内容特征和用户兴趣特征的相似性的一种过滤技术[10]，其特点是，为每个用户都建立1个用户兴趣模型(user profile)，对每个项目的内容进行特征提取，并构成特征向量。系统依据目标用户的兴趣模型，通过比较相似度对目标用户推荐信息。基于内容过滤推荐的优点是不考虑用户行为，因而不存在稀疏性、特殊用户等问题，其不足是只能处理文本类信息。

针对以上2种推荐方法的优、缺点，有学者提出了混合推荐技术[11]和基于数据挖掘的推荐技术[12–15]。笔者针对湖南省农村信息化服务平台(简称农信化平台)的农业知识库，基于内容过滤推荐技术，构建了1个新的农业信息推荐模型。

1 农业信息推荐模型

构建农业信息推荐模型，目的是利用农信化平台有倾向性地向农民推荐有用的知识。推荐模型整体架构如图1所示。

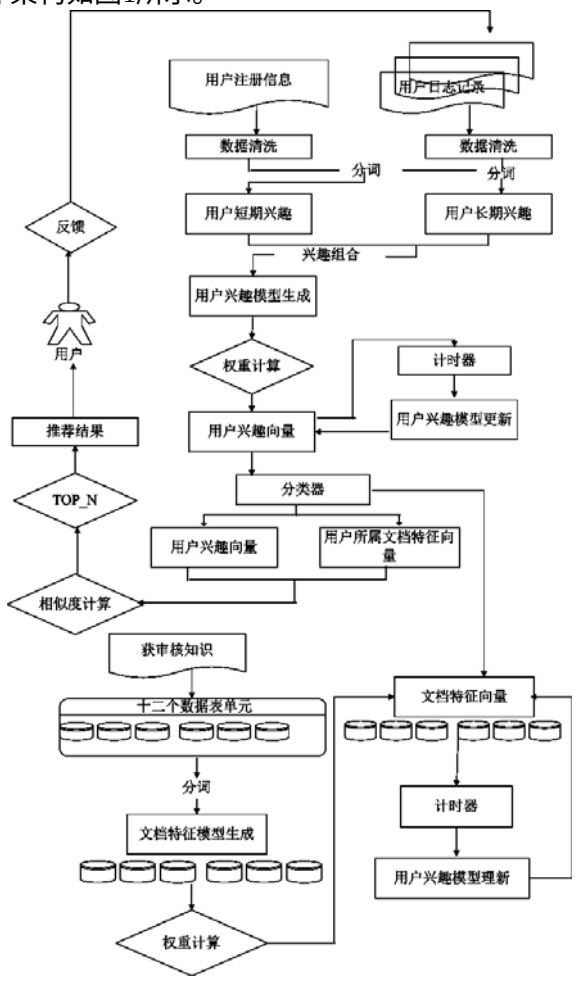

图1 模型的架构

Fig.1 Framework of the model

农业信息推荐模型主要内容有：①收集信息和信息预处理。用户兴趣向量模型，通过搜集农民用户个人信息和日志信息来建立。文档特征向量模型，依据农业知识库中10大优势产业划分的10大表空间建立。②专业分词。分词器提取关键词的准确性对提高推荐的质量以及对检索系统和查重系统十分重要。用一般的词典无法准确地切分农业专业词汇和农业方言词汇，且分词准确率比较低，因此使用专门设计的农业智能分词器[16]。③用户兴趣。在模型中，用户兴趣不仅仅是用户当前的兴趣偏好，还应包括用户的长期兴趣偏好，因此将用户兴趣分为短期兴趣偏好和长期兴趣偏好。短期兴趣偏好包括搜索到的相关字符串和浏览过的网页；长期兴趣偏好包括个人注册信息和浏览过的网页。④特征提取。鉴于目前使用最广的特征提取方法TF–IDF和 TF–IDF–IG都存在很多不足，特提出改进的特征提取算法。⑤用户兴趣分类。通过对用户兴趣分类处理，用户兴趣可以映射到10大表空间中的一类或几类，避免了用户兴趣分类的重复计算，减少了相似度的计算量。⑥相似度计算和信息推荐。通过比较用户兴趣向量与文档特征向量的相似度，向用户推荐其感兴趣的农业信息内容。模型使用余弦相似度公式[18]计算相似度，将相似度从大到小排序，把前面N条信息推荐给用户。⑦反馈修正。系统可将用户的每一个操作结果自动反馈给模型，模型可以跟踪用户的短期兴趣和长期兴趣。在模型中还设计了时间窗口，用来对用户长期兴趣模型进行修正。

2 特征提取算法

从3个方面对TF–IDF–IG算法进行改进和优化。

2.1 基于产业类型对算法进行改进

针对农信化平台中的10大产业类型表空间，将文本所属的不同产业类别引入到信息增益中，即考虑了特征词在不同产业类型的分布情况对权重的影响，并且在TF–IDF–IG公式中引入产业类型表空间。不带有分类信息的特征词被赋予较小的权重；对于分布不均匀的特征词频繁出现，被赋予较大的

权重，从而体现出特征词权重在10大产业类型的分布情况。

2.2　从 HTML 结构上对算法进行改进

基于产业类型对算法的改进未体现特征词在页面中的位置关系，在算法中加入了HTML结构加权因子，从而进一步优化TF–IDF–IG算法。

2.3　数量级差别导致的数据不平衡问题

当不同产业类型表空间中的文档数量存在数量级差异时，TF–IDF–IG算法中的IDF可能严重影响改进算法的计算准度，引入修正系数Wc加强IDF的抑制作用。

2.4　算法描述

2.4.1　相关定义

定义1　设d_i为文档，则总文档集记为

$D=\{d_1,d_2,d_3,\cdots,d_n\}$，$n$为总文档集$D$中的文档数，记$|D|=n$。

定义2　设d_i^t为包含了特征词t的1个文档，包含了特征词t的所有文档集合记为

$D_t=\{d_1^t,d_2^t,\cdots,d_N^t\}$，$N$为$D_t$中的文档数，记$|D_t|=N$。

定义3　设d_{ik}为产业类别i中的文档k，k=1，2，…，M，则总文档集D分为m个产业类别(m个表空间)，记表空间为

$C_i=\{d_{i1},d_{i2},\cdots,d_{iM}\}$，$i$=1,2,…,$M$。$M$为$C_i$类别中的文档数目，记$|C_i|=M$。

定义4　设d_{ik}^t为C_i中包含了特征项t的文档，则C_i^t为C_i中所有d_{ik}^t的集合，记为

$\left|C_i^t\right|=sd_{ik}^t,k=1,2,\cdots,s$，$s$为$C_i^t$中的文档数目，$\left|C_i^t\right|=s$。

定义5　TF–IDF计算公式为：

$$W(t,d)=TF\cdot IDF=f(t,d)\cdot\log(\frac{|D|}{|D_t|}+0.01)。$$

式中：$|D|=n$，$|D_t|=N$，0.01是修正系数，$f(t,d)=\frac{a}{A}$(a为特征词t在文档d中出现的次数，A为文档d中特征词的总数。

定义6　TF–IDF–IG计算公式为：

$$W(t,d)=\frac{TF\cdot IDF\cdot IG_t}{\sqrt{\sum_{t=1}(TF\cdot IDF\cdot IG_t)^2}}。$$

式中：D为文档集，$|d|$表示文档d中词语集合的个数。

$$IG_t=-\sum_{t\in D}P(d)\log P(d)+P(t)\sum_{t\in D}P(d|t)\log P(d|t)，$$

$$P(d)=\frac{|d|}{\sum_i|d_i|}。$$

2.4.2　改进的 TF–IDF–IG 算法

在传统TF–IDF方法中，当不同类别中的文档数量出现数量级的差异时，IDF会失效，从而严重影响改进的TF–IDF公式的准确度。需加入修正系数W_c加强IDF的抑制作用，修正由于数量级差异引起的数据失衡问题，其改进后的TF–IDF算法为：

$W(t，d)=TF`\cdot IDF\cdot W_c$，其中$W_c=\log\frac{|D|}{|C_i|}$。

信息增益体现特征词在分类中的重要程度，引入表空间后，其每一个表空间记为C_i的信息增益公式[17]为：$IG(C_i，t)=H(C_i)-H(C_i|t)=-\sum_{i=1}^{M}P(C_i)\log P(C_i)+P(t)\sum_{i=1}^{M}P(C_i|t)\log P(C_i|t)$。

式中：M为文档表空间数；$P(C_i)$为C_i类文档在总文档集D中出现的概率；$P(t)$为特征项t在总文档集中出现的概率；$P(C_i|t)$为文档包含特征项t时属于C_i的概率。

引入产业类型表空间后，改进的TF–IDF算法如下：

$$\bar{W}(t,d)=\frac{TF\cdot IDF\cdot IG(C_i,t)W_c}{\sqrt{\sum_{t=1}(TF\cdot IDF\cdot IG(C_i,t)W_c)^2}}。$$

但是特征词在文本结构中的位置关系没有得到体现，例如页面标题出现的特征词以及页面中加粗的特征词并没有被赋予更大的权重，因此有必要对不同的位置关系的特征词语赋予不同的权重(表1)。

表 1 HTML 标签权重

Table 1 Weight table of HTML label

标签	权重
<Title> </Title>	5
<H1> </H1>	4
<H2> </H2>	3
<B> </B>	2
<EM> </EM>	2
<IMG> </IMG>	2
<A> </A>	2
其余	1

记$W_t^{'}$为特征词t在其所在页面中的位置权重，引入文本结构后，改进后的TF–IDF–IG计算公式为：

$$\overline{W}(t,d)=\frac{TF\cdot IDF\cdot IG(C_i,t)W_cW_t^{'}}{\sqrt{\sum_{t=1}(TF\cdot IDF\cdot IG(C_i,t)W_cW_t^{'})^2}}。$$

式中：$IG(C_i, t)$为特征词t在类别C_i上的信息增益。

3 试验设计

验证农业信息推荐模型以及特征提取改进算法的试验流程如图2所示。

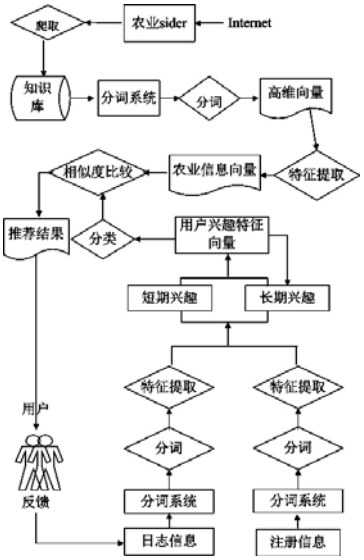

图 2 试验流程

Fig.2 Flow chart of experiment

数据来自农信化平台10大优势产业知识库，试验数据分为训练集文档和测试集文档(训练集文档2 903篇，测试文档2 591篇)，如表2所示。

表 2 试验数据来源分布

Table 2 Distribution of test data source

类别	文档数量/篇									
	粮食	畜禽	茶叶	油料	水产	蔬菜	水果	棉麻	竹木	中药材
训练集	420	384	128	147	310	405	573	196	166	174
测试集	396	350	120	127	280	370	450	168	145	185

采用农信化平台提供的农业专业分词器验证改进的算法。以对用户信息进行分类为例，设计流程如图3所示。

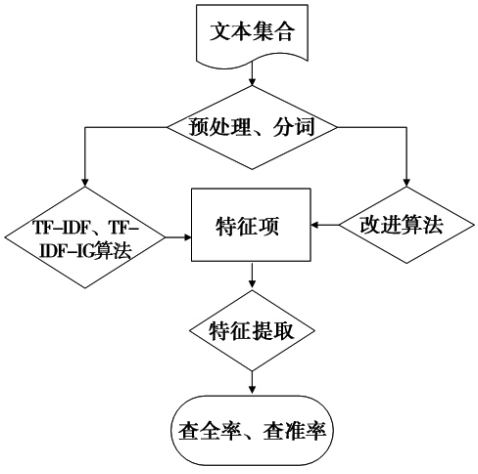

图 3 特征提取流程

Fig.3 Flow chart of feature extraction

4 试验结果

4.1 特征提取算法的评估

使用查全率和查准率[18]来衡量特征词权重算法的性能[19–21]。对改进算法的查准率和查全率进行评估(表3)，结果表明，改进算法的查准率和查全率较之TF–IDF算法分别提高了13.7%和11.2%，比TF–IDF–IG算法分别提高了5.7%和3.3%。

4.2 推荐结果

分别取N=25、50、100、200时(N表示用户数)，通过测试，得到农业信息推荐模型的查全率、查准率如表4所示。结果表明，随着用户数成倍增加，农业信息推荐模型的查全率和查准率也在不断提高，说明模型的精确度也在不断提高，具有很好的适应性。

表 3　算法的查全率和查准率

Table 3　Comparison of precision and recall　%

表空间	查准率			查全率		
	TF–IDF–IG 算法	TF–IDF 算法	改进算法	TF–IDF–IG 算法	TF–IDF 算法	改进算法
粮食	82.53	74.63	88.06	80.34	77.99	90.06
畜禽	83.66	79.34	86.10	82.95	80.48	90.12
茶叶	86.55	80.21	90.52	88.40	77.81	86.39
水果	77.97	79.69	83.91	84.31	80.92	84.24
油料	82.58	72.38	88.12	89.68	79.35	88.30
蔬菜	81.41	76.52	88.30	86.47	80.43	87.24
水产	78.19	80.54	85.36	85.71	81.1	92.68
棉麻	84.56	78.56	88.67	88.67	78.42	90.65
竹木	82.02	77.49	87.98	87.29	81.90	88.06
中药	81.24	63.69	80.62	80.03	74.51	84.29
平均	82.07	76.31	86.76	85.39	79.35	88.26

表 4　模型的查准率和查全率

Table 4　Precision and recall in model　%

用户数/户	查准率	查全率
25	75.3	79.1
50	79.8	80.4
100	83.2	86.4
200	90.6	93.8

参考文献：

[1] Lieberman H，Dyke N V，Vivaequa A．Let’browse：A collaborative web browsing agent[C]//Maybur M，Szekely P，Thomas C G．Proceedings of International Conference on the intelligent User Interfaces．Los Angeles：ACM Press，1999：65–68．

[2] Konstan J，Group Lens．Applying collaborative filtering to use net news[J]．Communications of the ACM，1997，40(3)：77–87．

[3] Mladenie D．Machine Learning for better web browsing [C]//Rogers W，Iba W．AAAI Spring Symposium on Adaptive User Interfaces．California：AAAI Press，2000：82–84．

[4] Ansari A，Essegaier S，Kohli R．Internet recommendation ystems[J]．Journal of Marketing Research，2000，37(3)：363–375．

[5] Schafer J B，Konstan J A，Riedl J．E–Commerce recommendation applications[J]．Data Mining and Knowledge Discovery，2000，5(1)：115–152．

[6] 张锋，常会友．使用 BP 神经网络缓解协同过滤推荐算法的稀疏性问题[J]．计算机研究与发展，2006，43(4)：667–672．

[7] 吴颜，沈洁，顾兰竺，等．协同过滤推荐系统中数据稀疏问题的解决[J]．计算机应用研究，2007，24(6)：94–97．

[8] 何安．协同过滤技术在电子商务推荐系统中的应用研究[D]．杭州：浙江大学，2007．

[9] 王娟，方逵．一种优化的基于协同过滤的农业信息推荐系统研究[J]．农机化研究，2011，33(7)：194–197．

[10] Schwab L．Proceedings of the International Conference on Intelligent User Interfaces[M]．New York：ACM Press，2000．

[11] Balabanovic M．Collaborative recommendation[J]. Commun ACM，1997，40(3)：66–72．

[12] Lin W．Efficient adaptive-support association rule mining for recommender system[J]．Data Mining and Knowledge Discovery，2006，6(1)：83–105．

[13] Mobasher B．Efective personalization based on association rule discovery from web usage data[C]// Mobasher B．3rd Int Workshop on Web Information and Data Management．New York：ACM Press，2001：9–15．

[14] Fu X．Mining navigation history for recommendation[C]// Riecken D．Proceedings of the International Conference on Intelligent User Interfaces．New York：ACM Press，2000：106–112．

[15] Mobasher B．Proceedings of the First International Conference on Electronic Commerce and Web Technologies[M]．Berlin：Springer–Verlag，2000．

[16] 方逵，罗武，朱幸辉．农业知识库系统设计与实现[J]．农机化研究，2013，35(5)：8–11．

[17] Mladenic D，Crobelnik M．Feature Selection for Unbalanced Class Distribution and Naive Bayes[C]// Ivan Bratko，Saso Dzeroski．Proceedings of the Sixteenth International Conference on Machine Learning．Bled：Morgan Kaufmann，1999：258–267．

[18] Ricardo，Baeza-Yates，Berthier，et al．Modern Information Retrieval[M]．London：Pearson Education，2003．

[19] 白若鸦，董渊，张素琴，等．研究中文文本分类技术的辅助平台[J]．清华大学学报：自然科学版，2008，48(7)：110–113．

[20] 王宇．基于 TFIDF 的文本分类算法研究[D]．郑州：郑州大学，2006：20–22．

[21] 奉国和．自动文本分类技术研究[J]．情报杂志，2007，26(12)：108–111．

责任编辑：罗慧敏

英文编辑：张　健