第 38 卷 第 4 期
Vol.38 No.4

计 算 机 工 程
Computer Engineering

2012 年 2 月
February 2012

• 软件技术与数据库 •　　文章编号：1000—3428(2012)04—0049—03　　文献标识码：A　　中图分类号：TP311

基于效用的个性化推荐方法

吴 兵[1,2]，叶春明[1]

(1. 上海理工大学管理学院，上海 200093；2. 上海电视大学，上海 200433)

摘 要：当前的推荐方法未能从个性化效用角度评价推荐项目，因此用户需按自己的偏好，在推荐结果中进行再次筛选。针对该情况，提出一种基于效用的个性化推荐方法。该方法采用逼近于理想值的排序法(TOPSIS)作为衡量推荐对象效用的基本方法。为克服 TOPSIS 中静态权重设置的不足，采用可变精度粗糙集发现用户对属性的偏好。实验结果表明，该方法能为用户提供更好的个性化效用及准确性的推荐服务。

关键词：个性化推荐；多属性效用；变精度粗糙集；推荐服务

Personalized Recommendation Method Based on Utility

WU Bing[1,2], YE Chun-ming[1]

(1. College of Management, University of Shanghai for Science and Technology, Shanghai 200093, China;
2. Shanghai TV University, Shanghai 200433, China)

【Abstract】Current recommendation methods lack of ability to evaluate utility of the recommendations according to user's preferences. So users have to make a choice among recommendations. On the basis of this, this paper presents a personalized recommendation method based on utility according to user's preferences. This method uses TOPSIS to evaluate utility of recommendations. In order to overcome the shortcoming of static weight in TOPSIS, it adopts variable rough set to discover user's preference for attributes. Experimental results show that the method can provide more appropriate recommendation service with better utility for users.

【Key words】personalized recommendation; multi-attribute utility; variable precision rough set; recommendation service

DOI: 10.3969/j.issn.1000-3428.2012.04.016

1 概述

效用理论认为，某一物品的价值并不以该物品的价格为基础，而是该物品对人的有用程度，或是人对该物品的价值认可程度，即为该物品的效用。目前，推荐技术主要方法包括：基于内容的推荐，关联规则挖掘，协同过滤技术[1]等。虽然不少学者采用各种方法来改进和完善推荐方法，但缺乏从效用角度来评价推荐对象。因此，用户需要花费较多时间和精力在推荐结果进行对比和选择。同时，推荐系统在不能获知用户偏好的情况下，难以为用户提供高质量的推荐服务。本文根据上述问题，提出一种基于效用的个性化推荐方法，并以文档检索推荐服务为背景，探讨基于用户属性偏好的最佳效用推荐。

2 现有技术研究及问题分析

现有的推荐方法为用户过滤了大量的信息，缩小了选择范围。然而，用户对推荐对象的价值衡量不以他人的评价为基础，而是一个依据自己偏好的决策过程。同时，用户面对的大部分决策对象都包含多个属性，属于多属性决策问题。此外，用户的偏好也不是一成不变的。现有推荐服务缺乏基于多属性效用的评价能力，难以从效用角度满足用户的实际需求。因此，需要从如下 3 个方面改善推荐服务：选择适当的效用评价手段；有效地识别用户的偏好；对用户偏好进行动态地学习。

TOPSIS[2]是 Hwang C L 和 Yoon K 于 1981 年首次提出多属性决策研究方法。其基本思想是确立目标的正理想值(Positive Ideal Solution)与负理想值(Negative Ideal Solution)。然后根据有限个评价对象与理想化目标的接近程度对评价目标进行排序、选择的方法。文献[3]对权重策略进行了比较，认为设置合理的权重是提高 TOPSIS 评价准确性的关键。粗糙集(Rough Set)[4]是波兰学者 Pawlak Z 于 1982 年提出的一种数学分析方法，它反映了人们以不完全信息或知识去处理一些不可分辨现象的能力。文献[5]在粗糙集基本原理的基础上提出了可变精度的改进方法。目前，粗糙集在数据挖掘领域有很多成功的应用。

3 TOPSIS 方法

TOPSIS 方法是基于偏好的多属性决策效用评价方法，它的优点包括：可以量化地评价对象的效用；对评价对象有较好的排序能力；算法简单高效等。因此，本文将 TOPSIS 作为评价推荐对象效用的基本方法。通过文献分析，TOPSIS 方法评价效用的准确性在于其权重的设置。对此，本文采用粗糙集方法识别用户对属性的偏好。粗糙集方法优点包括：它是数据驱动的挖掘方法；不需要先验知识；适合在数据库是实现数据挖掘。考虑噪音数据的影响，本文使用可变精度粗糙集模型，引入精度系数提高算法的识别能力。同时，为

基金项目：上海市教育委员会科研创新基金资助项目(11YZ256)；上海市教育委员会重点学科建设基金资助项目(S30504)；上海电视大学基金资助项目(JF1004)；高等学校博士点基金资助项目(20093120110008)

作者简介：吴 兵(1976－)，男，副教授、博士研究生，主研方向：个性化推荐技术，数据挖掘；叶春明，教授、博士生导师

收稿日期：2011-06-20　**E-mail：**wubing@shtvu.edu.cn

提高算法效率，本文采用基于关系演算实现算法。

4 相关定义及算法

4.1 粗糙集属性重要度计算相关定义

定义 1 设一个四元组 $S=\{U,R,V,f\}$ 为一个知识表达系统[4]，其中，U 是对象集合，也称为论域；R 是属性的有限集合，$R=C\cup D$，C 是条件属性集合，D 是决策属性集合；$V=\bigcup V_a, a\in R$，V_a 是属性 a 的值域；$f:U\times R\to V$ 是一个信息函数，它指定 U 中每一个对象 X 的属性值。

定义 2 设 $S=\{U,R,V,f\}$，对于属性 R 的非空子集 B，定义 B 在 U 上的不可分辨关系 $IND(B)$ 为：

$$IND(B)=\{(x,y)\in U\times U: f(x,r)=f(y,r), r\in B\} \tag{1}$$

定义 3 设 $S=\{U,R,V,f\}$，对于属性 R 的非空子集 B，$IND(B)$ 把 U 划分为 k 个不相交的等价类，记为 $U/IND(B)$。

定义 4 可变精度粗糙集模型[5]，设 $S=(U,C\cup D,V,f)$，假定由条件属性集合 C 导出的等价类为 $\tilde{C}=IND(C)=\{\tilde{C}_1,\tilde{C}_2,\cdots,\tilde{C}_k\}$，$\beta(0\leqslant\beta<0.5)$ 是依赖于数据中噪音程度的一个取值，X 为 U 的一个子集，β 正域、β 负域和 β 边界分别为：

β 正域：

$$POS_C(X)=\bigcup_{P(X|\tilde{C}_i)\geqslant 1-\beta}\{\tilde{C}_i\in\tilde{C}\} \tag{2}$$

β 负域：

$$NEG_C(X)=\bigcup_{P(X|\tilde{C}_i)\leqslant\beta}\{\tilde{C}_i\in\tilde{C}\} \tag{3}$$

β 边界：

$$BND_C(X)=\bigcup_{\beta<P(X|\tilde{C}_i)<1-\beta}\{\tilde{C}_i\in\tilde{C}\} \tag{4}$$

其中，$P(X\mid\tilde{C}_i)=\dfrac{P(X\cap\tilde{C}_i)}{P(\tilde{C}_i)}=\dfrac{Card(X\cap\tilde{C}_i)}{Card(\tilde{C}_i)}$，$Card()$ 表示集合的基数。

定义 5 设 $S=\{U,R,V,f\}$，$R=C\cup D$，C 是条件属性集合，D 是决策属性集合，$\beta(0\leqslant\beta<0.5)$，条件属性 C 和决策属性 D 的之间的相关程度定义为：

$$K_\beta(C,D)=Card(POS_C(D)\cup NEG_C(D))/Card(U) \tag{5}$$

其中，$POS_C(D)=\bigcup_i POS_C(D_i), NEG_C(D)=\bigcup_i NEG_C(D_i)$；$K_\beta(C,D)$ 表示决策表中能够确定划分到 β 正域和 β 负域的样本的百分比。

定义 6 设 $S=\{U,R,V,f\}$，$R=C\cup D$，C 是条件属性集合和，D 是决策属性集合，$\beta(0\leqslant\beta<0.5)$，$B\subseteq C$，$a\in B$，属性 a 的重要性定义如下：

$$SGF_\beta(a,B,D)=K_\beta(B,D)-K_\beta(B/\{a\},D) \tag{6}$$

4.2 基于可变精度粗糙集的用户偏好识别

利用粗糙集实现用户偏好挖掘需要建立决策信息表。通过数据库中记录的用户访问历史可以建立上述信息表。以文档检索为例，设 $\beta(0\leqslant\beta<0.5)$，用户访问信息表为 S，其中条件属性集合 C 设为{文档类型，文档来源，文档时间，文档引用次数}，决策属性 D 设为{用户行为}。用户行为的值域是：{下载、收藏、浏览}。通过 S 表可以分析用户对文档的使用偏好。

定义 7 基于关系演算实现属性相关度计算。决策信息表 $w\_table=\{a_1,a_2,a_3,d\}$，$C=\{a_1,a_2,a_3\}, D=\{d\}$。给定 $\beta(0\leqslant\beta<0.5)$，利用 SQL 语句实现 $K_\beta(C,D)$ 如下：

select isnull(count(id),0) from w_table t where cast((select isnull(count(id),0) from w_table where d=t.d and a1=t.a1 and a2=t.a2 and a3=t.a3 group by a1,a2,a3,d) as float)/(select count(id) from w_table where a1=t.a1 and a2=t.a2 and a3= t.a3 group by a1,a2,a3)≥1−β or cast((select isnull(count(id),0) from w_table where d=t.d and a1=t.a1 and a2=t.a2 and a3=t.a3 group by a1,a2,a3,d) as float)/(select count(id) from w_table where a1=t.a1 and a2=t.a2 and a3=t.a3 group by a1,a2,a3)≤ β

算法 1 基于可变精度粗糙集识别属性权重

输入 决策表 $S=(U,C\cup D,V,f)$，其中条件属性集合为 $C=\{a_1,a_2,a_3,\cdots,a_n\}$，决策属性集合为 $D=\{d\}$，精度系数 $\beta(0\leqslant\beta<0.5)$

输出 属性重要度权重 w_{a_i}

第 1 步 依据式(5)及定义 7，计算 $K_\beta(C,D)$。

第 2 步 依据式(5)以及定义 7，对于 $\forall a_i\in C$，依次计算 $K_\beta(C/\{a_i\},D)$。

第 3 步 依据式(6)，对于 $\forall a_i\in C$，依次计算属性重要度 $SGF_\beta(a_i,C/\{a_i\},D)$。

第 4 步 计算属性 a_i 的重要度权重，公式如下：

$$w_\beta(a_i)=SGF_\beta(a_i,C,D)\Big/\sum_{i=1}^{n}SGF_\beta(a_i,C,D), i=1,2,\cdots,n \tag{7}$$

第 5 步 输出 $w_\beta(a_i)$，所有属性权重构成一个用户属性偏好向量 $\boldsymbol{W}=(w_{a_1},w_{a_2},\cdots,w_{a_n})$。

4.3 基于用户偏好的效用推荐

在获得用户属性偏好向量后，依据 TOPSIS 方法对推荐候选结果进行效用评价处理，反馈基于个性化效用分析后的推荐结果。以协同过滤 User-based 方法为例，算法设计如下：

算法 2 基于用户偏好的最佳效用推荐

输入 用户可以访问历史数据，用户属性偏好向量 $\boldsymbol{W}=(w_{a_1},w_{a_2},\cdots,w_{a_n})$，最近邻居用户数量 k，推荐项目数量 m

输出 基于用户偏好的效用推荐结果

第 1 步 依据用户访问历史数据，构建用户-项目评价矩阵。

第 2 步 依据皮尔森距离方法计算相似用户距离，选择最相近的 k 个邻居。

第 3 步 根据这 k 个邻居用户的评价记录，使用预测评分的方法计算当前用户未评分的项目，从中选择评分最高的前 m 项作为推荐候选集合 C。

第 4 步 依据 C 建立如下效用评价矩阵：

$$\boldsymbol{F}=(p_{ij})_{m\times n}=\begin{pmatrix}p_{11}&p_{12}&\cdots&p_{1n}\\p_{21}&p_{22}&\cdots&p_{2n}\\\vdots&\vdots&&\vdots\\p_{m1}&p_{m2}&\cdots&p_{mn}\end{pmatrix} \tag{8}$$

第 5 步 对矩阵进行量纲处理。收益性指标按照式(9)处理，成本性指标按照式(10)处理。

$$v_j=(p[j]-\min([j]))/(\max([j])-\min([j])), 1\leqslant j\leqslant n \tag{9}$$

$$v_j=(\max[j]-p[j])/(\max([j])-\min([j])), 1\leqslant j\leqslant n \tag{10}$$

第 6 步 对矩阵 $\boldsymbol{F}$ 规范化处理，常用方法如下：

$$v'_{ij}=v_{ij}\Big/\sqrt{\sum_{i=1}^{m}v_{ij}^2}, i=1,2,\cdots,m, j=1,2,\cdots,n \tag{11}$$

第 7 步 利用属性权重向量 $\boldsymbol{W}$ 生成加权矩阵 $\boldsymbol{F}'$，方法如下：

$$v'_{ij}=w_j v'_{ij}, i=1,2,\cdots,n, j=1,2,\cdots,m \tag{12}$$

第 8 步 根据权重规格化值 v'_{ij} 来确定正理想解 v_j^* 和负理想解 v_j^-：

$$v_j^* = \begin{cases} \max(v'_{ij}), j \in J^* \\ \min(v'_{ij}), j \in J^- \end{cases} \quad j = 1,2,\cdots,n$$

$$v_j^- = \begin{cases} \min(v'_{ij}), j \in J^* \\ \max(v'_{ij}), j \in J^- \end{cases} \quad j = 1,2,\cdots,n \tag{13}$$

其中，J^* 是收益性指标；J^- 是成本性指标。

第 9 步 依次计算 C 中各元素的“理想解贴近度” C_i^*，公式如下：

$$C_i^* = S_i^- / (S_i^* + S_i^-), i = 1,2,\cdots,m \tag{14}$$

其中，$0 \leqslant C_i^* \leqslant 1$，$C_i^*$ 愈接近 1，表示该评价对象越接近最优水平；反之，C_i^* 愈接近 0，表示该评价对象越接近最劣水平。

第 10 步 根据 C^* 的值排序待推荐对象集合 C。

第 11 步 输出效用推荐结果，算法结束。

5　实验及评价

实验从网上学习系统及相关文献数据库中采集学习资源及用户访问记录。首先，利用 Lucene 对上述资源建立索引。按照 4.2 节方法设置用户访问信息表 S。为便于粗糙集的挖掘工作，对相关属性进行编码处理，例如：“文档来源”按照“四大检索”文献、核心期刊、一般期刊、其他，分别编码为：4~1。对“文档类型”依据期刊、会议出版物、论文、书籍、报纸、其他，分别编码为：6~1。考虑文档的时间效用，对于“文档时间”属性，按照年度进行编码计算，公式如下：

文档时间＝用户检索时间−文档发表时间+1　　(15)

对数值型变量，亦可按照等宽、等频或者设定的方式进行数据离散化处理。这里以设定方式对“引用次数”进行离散化处理，设置 0 次，10 次以内，10 次~50 次，50 次以上，依次对应编码为 1~4。随机选取一个用户 User-A，其历史访问记录如表 1 所示，它是一个包含噪音信息表。设 $\beta = 0.1$，利用算法 1 识别用户属性偏好结果如下：“文档类型”属性的权重为：0.462；“文档来源”属性的权重为：0.115；“文档时间”属性的权重为：0.346；“引用次数”属性的权重为：0.077。

表 1　包含噪音的文档访问信息

ID	文档类型	文档来源	文档时间	引用次数	用户行为	数量
1	6	4	2	3	下载	10
2	6	3	2	3	下载	6
3	5	4	2	3	下载	2
4	4	4	2	3	浏览	1
5	5	3	2	3	浏览	1
6	6	4	4	2	浏览	1
7	6	3	5	3	浏览	1
8	5	2	4	3	收藏	1
9	4	2	5	2	收藏	1
10	5	2	4	1	收藏	1
11	4	2	5	3	下载	1
12	4	1	12	2	浏览	1
13	6	4	2	3	收藏	1

随机选取 10 个用户进行实验。每个用户按照 5 种推荐数量(40~80)分别做实验。对比 User-based 协同过滤方法及本文提出的方法。效用对比方法采用如下公式：

$$k1 = Sum(Value(top40\%)) / Sum(Value(all)) \tag{16}$$

其中，$Value()$是效用计算函数；$k1$ 的作用是对比推荐结果前 40%占总结果的效用比例。

另外，查准率是比较推荐质量的指标之一。利用文献[6]中相似性比较方法来计算查准率。查准率对比采用如下公式：

$$Sim\_\exp(item) = \sum_{i=1}^{n} Sim(item, item\_\exp) \Big/ n \tag{17}$$

$$k2 = Sum(Sim\_\exp(top40\%)) / Sum(Sim\_\exp(all)) \tag{18}$$

其中，$Sim\_\exp(item)$ 是计算推荐文档与用户阅读历史文档的相似性；$k2$ 的作用是对比推荐结果前 40%占总结果的相似度比例。图 1 显示了效用比率及查准率的对比结果。对比结果表明，算法 2 的推荐效用高于协同过滤推荐方法，同时推荐准确率略高于协同过滤方法。

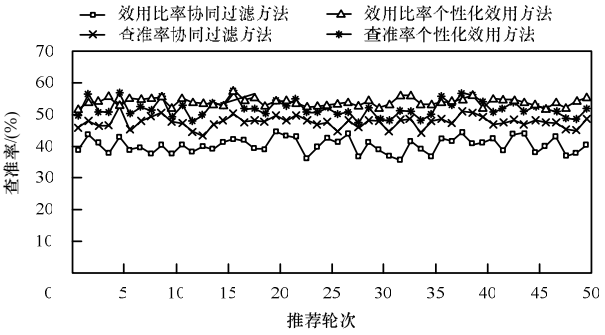

图 1　推荐效用比较

综合实验结果，通过基于用户属性偏好的挖掘，以及对推荐对象的效用评价及排序，本文提出的效用推荐方法在效用及准确率方面都有较好的推荐质量。

6　结束语

传统的推荐方法仅将用户可能喜好的对象呈现给用户，没有从效用角度为用户提供推荐服务。在实际应用中，实现基于用户偏好的效用推荐是十分有价值的。然而，要达到效用推荐的目标需要选择适合的评价方法及识别用户的偏好。对此，本文选取 TOPSIS 方法作为效用评价的基本方法。为提高效用评价的准确性，本文采用粗糙集挖掘用户对属性的偏好。通过相关实验表明本文方法有较好的效用推荐效果。在今后工作中，将对效用评价方法与推荐算法集成问题做进一步研究。

参考文献

[1] 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, 30(7): 1282-1288.

[2] Hwang K Y C L. Multiple Attribute Decision Making Methods and Applications[M]. Berlin, Germany: Springer, 1981.

[3] Olson D L. Comparison of Weights in TOPSIS Models[J]. Mathematical and Computer Modelling, 2004, 40(7): 21-25.

[4] Pawlak Z. Rough Set Theory and Its Applications[J]. Journal of Telecommunications and Information Technology, 2002, 3(2): 7-10.

[5] Ziarko W. Variable Precision Rough Set Model[J]. Journal of Computer and System Sciences, 1993, 46(1): 39-59.

[6] 吴　兵, 叶春明, 陈　信. 基于多代理的个性化推荐学习系统[J]. 计算机工程, 2010, 36(15): 256-258.

编辑　索书志