计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
44(2): 296~301, 2007

适应用户兴趣变化的协同过滤推荐算法

邢春晓[1] 高凤荣[1] 战思南[2] 周立柱[2]
[1](清华大学信息技术研究院 Web 与软件技术研究中心 北京 100084)
[2](清华大学计算机科学与技术系软件研究所 北京 100084)
(xingcx@tsinghua.edu.cn)

A Collaborative Filtering Recommendation Algorithm Incorporated with User Interest Change

Xing Chunxiao[1], Gao Fengrong[1], Zhan Sinan[2], and Zhou Lizhu[2]
[1](*Web and Software Technology Research and Development Center, Research Institute of Information Technology, Tsinghua University, Beijing* 100084)
[2](*Institute of Software, Department of Computer Science and Technology, Tsinghua University, Beijing* 100084)

Abstract Collaborative filtering is one of the most successful technologies for building recommender systems, and is extensively used in many personalized systems. However, existing collaborative filtering algorithms do not consider the change of user interests. For this reason, the systems may recommend unsatisfactory items when user's interest has changed. To solve this problem, two new data weighting methods: time-based data weight and item similarity-based data weight are proposed, to adaptively track the change of user interests. Based on the analysis, the advantages of both weighting methods are combined efficiently and applied to the recommendation generation process. Experimental results show that the proposed algorithm outperforms the traditional item-based collaborative filtering algorithm.

Key words collaborative filtering; personalized recommendation; time-based data weight; item similarity-based data weight

摘 要 协同过滤算法是至今为止最成功的个性化推荐技术之一,被应用到很多领域中.但传统协同过滤算法不能及时反映用户的兴趣变化.针对这个问题,提出两种改进度量:基于时间的数据权重和基于资源相似度的数据权重,在此基础上将它们有机结合,并将这两种权重引入基于资源的协同过滤算法的生成推荐过程中.实验表明,改进后的算法比传统协同过滤算法在推荐准确度上有明显提高.

关键词 协同过滤;个性化推荐;基于时间的数据权重;基于资源相似度的数据权重
中图法分类号 TP391.4; TP311.13

个性化推荐(personalized recommendation)技术通过研究不同用户的兴趣,主动为用户推荐最需要的资源,从而更好地解决互联网信息日益庞大与用户需求之间的矛盾.目前,推荐技术被广泛应用到电子商务[1]、数字图书馆[2]、新闻网站[3]等系统中.

协同过滤(collaborative filtering)是迄今为止应用最成功的个性化推荐技术[4-6].它的基本思想是根据用户兴趣的相似性来推荐资源,把和当前用户相似的其他用户的意见提供给当前用户.其优点是无需考虑资源的表示形式,并能为用户发现新的感兴趣的资源.现有的协同过滤算法存在一个弊端:不能及时反映用户兴趣变化.本文针对这个问题,改进推荐算法,提出了两种新的数据加权度量:基于用户访问时间的数据权重和基于资源相似度的数据

收稿日期:2005-11-10;修回日期:2006-05-26
基金项目:国家自然科学基金项目(60473078)

权重,将这两种权重引入协同过滤算法的生成推荐过程中,更好地反映用户兴趣变化,提高推荐精度.

1 相关工作

1.1 协同过滤推荐算法

在典型协同过滤推荐系统中,输入数据通常可以表述为一个 $m\times n$ 的用户-资源访问矩阵 $\boldsymbol{R}$. 其中 m 是用户数, n 是资源数, R_{ij} 是第 i 个用户对第 j 个资源的访问记录,矩阵值表示用户访问该资源与否,1 表示访问,0 表示未访问. 例如电子商务系统中,可由顾客交易数据产生购买矩阵, $R_{ij}=1$ 表示顾客 i 购买了商品 j.

典型的协同过滤算法是基于用户(user-based)的[1,5],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. 对当前用户 u,系统通过其历史访问记录及特定相似度函数,计算出与其访问行为(购买的产品集合、访问的网页集等)最相近的 k 个用户作为用户 u 的最近邻居集,统计 u 的近邻用户访问过而 u 未访问的资源生成候选推荐集,然后计算候选推荐集中每个资源 i 对用户 u 的推荐度,取其中 N 个排在最前面的资源作为用户 u 的 top-N 推荐集.

为解决传统协同过滤算法的可扩展性问题[7],文献[5,8]提出了基于资源(item-based)的协同过滤算法,该算法比较资源与资源之间的相似性,由当前用户已访问的资源集合推荐未访问的资源. 由于资源间的相似性比用户相似性稳定,因此可以离线进行计算存储并定期更新,较好地解决了算法的可扩展性问题.

相比较而言,基于资源的协同过滤算法推荐精度高,实时性好. 对这种推荐算法进行优化更具有现实意义. 算法 1 给出了基于资源的协同过滤算法描述.

算法 **1**. 基于资源的协同过滤推荐算法.

输入:用户 u、与之对应的已访问资源集 I_u、资源近邻模型 M.

输出:用户 u 的 top-N 推荐集.

过程:

Step1. 对每个资源 $i\in I_u$,读取 M 得到它的 k 最近邻居集 $N_i=\{i_1,i_2,\cdots,i_k\}$,合并所有 N_i 得到集合 C;

Step2. 从 C 中删除 I_u 中已经存在的资源,得到候选推荐项集 $Candidate$;

Step3. 对资源 $j\in Candidate$,计算 j 对 u 的推荐度:

$$rec\text{-}w(u,j)=\sum_{i\in I_u} sim(i,j);$$

Step4. 将 $Candidate$ 中的资源按加权推荐度大小排列,其中最前的 N 个资源作为用户 u 的推荐集.

相似度计算是影响推荐算法性能的重要因素. 计算相似度有多种不同的方法,如余弦相似度、*Pearson* 相关系数[7]、条件概率[9]等. 文中我们选用条件概率来计算资源之间的相似性. 对于资源 i 和 j,用 $P(i|j)$ 表示他们被同一用户访问的条件概率,它可以衡量资源间相似性. 计算资源 i 和 j 之间的相似性公式如下:

$$sim(i,j)=\frac{P(i|j)}{Freq(i)^{\alpha}}=\frac{Freq(ij)}{Freq(j)\times Freq(i)^{\alpha}},\tag{1}$$

其中, $Freq(i)$, $Freq(j)$, $Freq(ij)$ 分别表示访问过资源 i, j 的用户数以及同时访问过 i, j 的用户数. α 是一个 0~1 之间的数,称为缩放系数,引入 α 的目的是削弱被访问过很多次的资源在相似度计算中的影响.

1.2 现有算法的不足

现有的协同过滤推荐算法都存在一个问题:只注重用户或资源间的相似性,而忽略了用户兴趣的动态变化. 在现实生活中,用户对资源的需求是随着时间的推移不断改变的,传统的协同过滤算法只利用用户-资源访问矩阵来进行推荐计算,而未考虑用户访问资源的具体时间,因此无法反映出用户的兴趣随时间的变化过程,当用户兴趣发生改变的时候,现有的推荐系统无法及时发现,从而导致系统推荐的资源在很大程度上偏离了用户的需求.

为解决上述问题,我们将两种数据加权策略:基于用户访问时间的数据权重(time-based data weight)和基于资源相似度的数据权重(item similarity-based data weight)引入到基于资源的协同过滤算法的推荐过程中,以解决传统协同过滤算法不能及时反映用户兴趣变化的弊端.

2 改进算法描述

2.1 基于时间的数据权重

现有的协同过滤算法在计算推荐过程中将用户访问过的每个资源同等对待,这显然是不合理的. 一般来说,用户近期访问过的资源对推荐该用户未

来可能感兴趣的资源起比较重要的作用,而早期的访问记录对生成推荐影响相对较小,这是因为用户的兴趣随时间的推移不断变化,而在较短的一段时间内用户的兴趣是相对稳定的,因此一个用户感兴趣的资源最可能和他近期访问过的资源相似. 因此,我们引入基于用户访问时间的数据权重(time-based data weight),以提高最近访问数据在推荐生成过程中的重要性.

设 D_{ui} 表示用户 u 访问资源 i 的时间与用户 u 最早访问某资源的时间间隔,我们定义基于时间的权重函数 $WT(u, i)$ 表示资源 i 对用户 u 的权重,它是一个和 D_{ui} 相关的函数值. 为了突出用户 u 近期访问过的资源的重要性,权重函数应该设计成关于 D_{ui} 的非递减函数,即对于 $D_{ui} > D_{uj}$ 有 $WT(u, i) \geqslant WT(u, j)$. 本文将基于时间的权重函数作如下定义:

$$WT(u, i) = (1-a) + a\frac{D_{ui}}{L_u}. \qquad (2)$$

式(2)是一个线形函数,其中 L_u 表示用户 u 使用推荐系统的时间跨度,即该用户最早访问某资源的时间与最近访问某资源的时间间隔, $a \in (0, 1)$ 称为权重增长指数,改变 a 的值可以调整权重随访问时间变化的速度. a 越大权重增长速度越快, a 的大小可以影响到算法性能. 根据不同的推荐系统,可以动态调整 a 的值来优化推荐效果.

2.2 基于资源相似度的数据权重

式(2)中, $WT(u, i)$ 的值随着用户 u 访问资源 i 时间间隔 Dui 呈线性变化,用户近期访问数据的权重总是大于早期访问数据的权重,从而突出了近期数据的重要性. 但是在现实中,不同用户兴趣变化速度和规律不同,此外用户的兴趣经常存在反复,所以用户早期访问的资源往往对于生成推荐也很重要,单纯使用基于时间的数据权重,削弱了所有早期资源在推荐计算中的作用,可能对推荐效果产生负面影响. 为此我们引入第 2 种数据加权方法:基于资源相似度的数据权重(item similarity-based data weight)对用户的已访问资源进行加权.

设用户 u 的已访问资源集合为 I_u,通过定义一个时间窗(time window) T,获取用户 u 在最近 T 时段内访问过的资源集合为 I_{uT}, I_{uT} 在一定程度上反映了用户的近期兴趣. 对于资源 $i \in I_u$,无论 u 访问 i 的时间早晚,如果 u 的近期访问资源集 I_{uT} 中很多资源和 i 相似度很高,说明资源 i 和用户的当前兴趣很相关,则在未来一段时间内, u 感兴趣的资源很可能也和资源 i 相似,即资源 i 对生成用户 u 的推荐起比较重要的作用. 因此我们可以定义基于资源相似度的权重函数 $WS(u, i)$ 衡量资源 i 和用户 u 当前兴趣的相关程度,它可以通过 i 和 I_{uT} 的总体相似度 $sim(i, I_{uT})$ 计算,而 i 和 I_{uT} 总体相似度可以通过计算 i 和 I_{uT} 中每个资源 j 的平均相似度来表示:

$$WS(u, i) = \overline{sim(i, I_{uT})} = \frac{\sum_{j \in I_{uT}} sim(i, j)}{size(I_{uT})}, \qquad (3)$$

其中, $size(I_{uT})$ 表示 I_{uT} 中的资源数目. 通过改变时间窗 T 的长短,可以得到不同的近期访问集合 I_{uT},从而影响推荐效果.

从式(3)可以看出,为计算 $WS(u, i)$,需要计算 i 和 I_{uT} 中每个资源的相似度,若集合 I_u 中资源数为 m, I_{uT} 中资源数为 n,则为用户 u 访问过的所有资源进行加权计算的时间复杂度为 $O(mn)$,由此可见,基于资源相似度的数据加权方法计算量比基于时间的数据加权方法要高,但由于一个用户访问过的资源数通常比较小,因此不会对算法的实时性有太大影响.

2.3 两种数据权重的结合

上面介绍了两种数据加权度量,它们各有优点:基于时间的数据权重突出近期数据的重要性,从而能够及时捕捉到用户的当前兴趣,适合处理用户兴趣变化较频繁的情况;而基于资源相似度的数据权重通过计算用户访问过的某资源与该用户当前兴趣的相关度,避免了有价值的早期数据被忽略,适合处理用户兴趣存在反复的情况. 因此考虑将两个权重函数用一定的比例因子结合起来,定义同时基于时间和资源相似度的权重函数:

$$WTS(u, i) = \beta \times WT(u, i) + (1-\beta) \times WS(u, i), \qquad (4)$$

其中比例因子 $\beta \in [0, 1]$, β 和 $(1-\beta)$ 分别代表两种权重值所占的比例. 通过选择合适的 β 值可以将两种加权方法的优点结合起来,从而进一步提高推荐算法的准确率.

2.4 适应用户兴趣变化的协同过滤推荐算法

我们把上述加权方法引入到传统的协同过滤算法中,提出一种改进的基于资源协同过滤算法. 首先根据前面的式(2)计算并存储每个资源的近邻资源集,设用户的已访问资源集为 I_u,推荐过程中首先读入 I_u 中每个资源 i 的 k 最近邻居集及相应的相似度,生成候选推荐集,根据式(3)或(4)计算 I_u 中每个资源 j 的权重 $W(u, j)$,然后根据计算出的

数据权重计算候选推荐集中每个资源的加权推荐度, 推荐度最大的前 N 个资源作为用户 u 的 top-N 推荐集. 算法的具体描述如下:

算法 **2**. 适应用户兴趣变化的协同过滤推荐算法.

输入: 用户 u、与之对应的已访问资源集 I_u、资源近邻模型 M.

输出: 用户 u 的 top-N 推荐集.

过程:

Step1. 对每个资源 $i \in I_u$, 读取 M 得到它的 k 最近邻居集 $N_i = \{i_1, i_2, \cdots, i_k\}$, 合并所有 N_i 得到集合 C;

Step2. 从 C 中删除 I_u 中已经存在的资源, 得到候选推荐项集 *Candidate*;

Step3. 对每个资源 $i \in I_u$, 根据式(2), (3), 或式(4)计算 $Weight(u, i)$;

Step4. 对资源 $j \in Candidate$, 计算 j 对 u 的加权推荐度:

$$rec\text{-}w(u, j) = \sum_{i \in I_u} Weight(u, i) \times sim(i, j);$$

Step5. 将 *Candidate* 中的资源按加权推荐度大小排列, 其中最前的 N 个资源作为用户 u 的推荐集.

3 实验结果及分析

3.1 测试数据集

我们用 KDD2000 的网上交易数据集[10] 作为测试数据来对本文提出的算法与传统的基于资源的协同过滤算法进行比较. 原始数据文件是某电子商务网站的 Web 日志, 通过对日志文件的预处理, 保留了 2660 个用户对 387 种商品的 90182 个有效访问记录, 时间跨度为 2 个月. 每个访问记录表示为一个三元组〈用户 ID, 商品 ID, 访问时间〉, 由于测试数据分为训练集和测试集, 其中, 把实验数据中每个用户最后 10 天的访问记录隐藏起来作为测试集, 其余的访问记录为训练集. 训练集用于构建用户-资源访问 0/1 矩阵 $\boldsymbol{R}$ 和进行资源相似度计算.

3.2 评价标准

实验过程中根据每个用户在训练集中的访问记录为其计算 Top-N 推荐集, 如果 Top-N 推荐集中某个资源 i 出现在该用户测试集中的访问记录里, 则表示生成了一个正确推荐. 我们用信息检索领域中评估系统效果的准确率(*Precision*)标准作为对比传统算法和我们的算法推荐精度的标准:

$$Precision = \frac{Hits}{N},$$

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.

3.3 实验结果

实验是为了比较本文提出的算法和传统算法之间捕捉用户兴趣变化的能力大小, 最终选择 1000 个使用时间跨度超过 40 天, 访问商品数超过 10 个的用户共计 38125 条记录.

我们设计了 3 组实验, 把它们的推荐效果与传统的基于资源的协同过滤算法(即算法 1, 简称为 Item-based CF)相对比. 在所有实验中, 资源最近邻数 $K=20$, 观察推荐数目 N 从 10 ~ 50 每次增加 10 不同情况下推荐算法的性能比较. 实验中我们为所用用户计算推荐, 以下的实验结果是对所有用户计算结果的平均.

图 1 给出了不同取值的基于时间的数据权重对推荐准确率的影响. 其中权重增长指数 a 分别取 0.3, 0.4, 0.5, 0.6, 0.7.

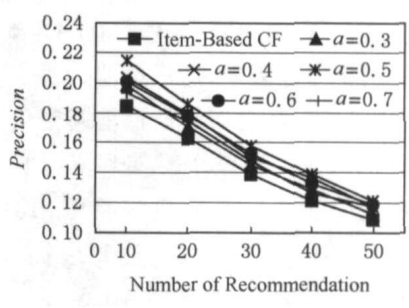

Fig. 1 Comparison of recommendation quality of traditional item-based CF algorithm and our improved algorithm using time-based data weight

图 1 基于时间的数据权重算法和传统协同过滤算法推荐效果对比图

从图 1 可以看出, 将基于时间的数据权重引入传统的协同过滤推荐算法中, 推荐精度会有较大的提高, 尤其当推荐数目较少时, 准确率提高尤为明显(例如当 a 取 0.5, 推荐数目为 10 时, 准确率比传统算法提高约 16%). 由此可见, 引入基于时间的数据权重有效地突出了用户近期访问数据对生成推荐的重要性, 从而使得算法生成的推荐更好地满足用户的当前兴趣. 同时我们可以看到, 权重增长指数 a 的设置对算法准确率有比较大的影响, 对于不同的用户和不同类型的资源, 用户兴趣的变化速度和变化规律也不同, 权重增长过快或过慢都会对推推荐效果产生负面影响.

图 2 给出了基于资源相似度的数据权重算法中不同的时间窗 T 取值对推荐结果的影响. 其中 T 分别取 5, 10, 15 天.

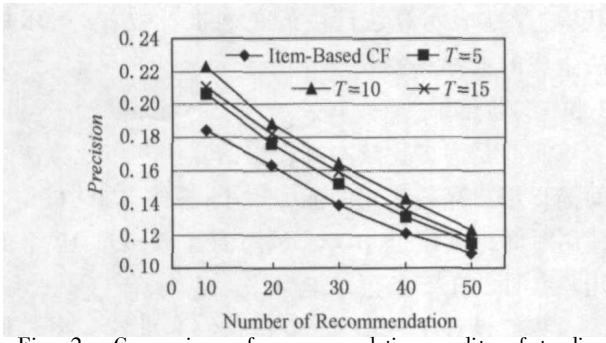

Fig. 2 Comparison of recommendation quality of traditional item-based CF algorithm and our improved algorithm using item similarity-based data weight.

图 2 基于资源相似度的数据权重推荐算法和传统基于资源协同过滤算法的推荐准确率对比

实验结果表明,基于资源相似度的数据权重整体上比基于时间的数据权重在推荐效果上更好一些.这是由于基于资源相似度的数据权重避免了早期重要数据被忽略,可以有效地处理用户兴趣反复的情况.此外,我们得到 $T=10$ 时算法性能最好,这说明时间窗的长短对推荐准确率有一定影响,过长则无法反映用户当前兴趣,过短会使计算出的数据权重有很大的随机性,都会对算法性能起到负面作用.

最后我们测试同时基于时间和资源相似度的综合权重算法的性能.在前两组实验中,我们分别得到 $a=0.5$, $T=10$ 时算法性能达到最优,因此在计算综合权重过程中我们就设 $a=0.5$, $T=10$.为了获得式(4)中最合适的 β 值,我们进行了一组改变 β 的实验,让 β 从 0.2 变动到 0.8,每次增加 0.1,结果如图 3 所示:

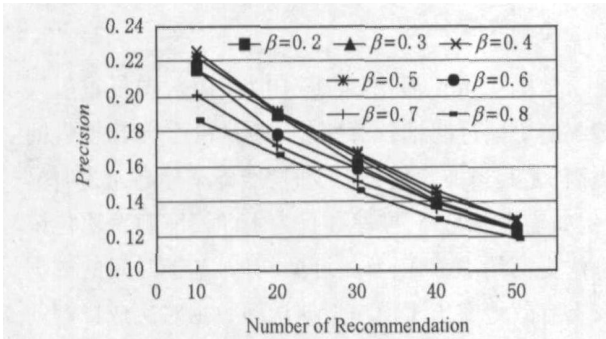

Fig. 3 Recommendation precision using different β on combined data weight.

图 3 不同比例的综合加权对算法准确率的影响

由图 3 可以看出,如果 β 值选取适当,综合数据权重比两种数据权重单独使用效果又有进一步提升,这是由于它结合了两种加权方法优点,不仅能突出近期数据的重要性,又避免了早期数据被忽略,从而更准确地反映了用户的兴趣变化趋势,使生成的推荐有更好的准确度.

4 结论和未来工作

本文针对现有协同过滤算法不能快速发现用户兴趣变化的问题,提出两种改进方法:基于时间的数据权重和基于资源相似度的数据权重,并将两种权重有机结合.在此基础上,将本文提出的数据权重方法引入基于资源的协同过滤推荐算法中,以反映用户兴趣的动态变化,克服了传统算法的弊端.

加权算法简单易行、效率高、实时性好.对比实验表明,引入两种数据权重后,如果参数设置得当,算法性能会有较大提高.从实验结果还可以看出,将两种加权方法加以综合能够更有效地捕捉用户兴趣,因此推荐精度更高.

我们未来的工作包括以下两方面:一方面,不同的用户兴趣的变化规律也不同,针对不同用户选取不同的方案并设置不同的参数比设置固定参数有可能得到更好的结果,因此我们将对权重函数中参数的自动确定方法做进一步研究.另一方面,不仅用户兴趣随时间变化,特定类型的资源受欢迎程度也可能对时间比较敏感(比如电器类产品).因此下一步我们考虑将用户兴趣度按时间分段预测,比如以季度为单位,适合对时间较敏感的产品,从而形成一个分段连续的预测模型.

参 考 文 献

[1] J Schafer, J Konstan, J Riedl. Recommender systems in e-commerce[C]. In: Proc of ACM E-Commerce. New York: ACM Press, 1999. 158-166

[2] Champa Jayawardana, K Priyantha Hewagamage Masashito Hirakawa. A personalize information environment for digital libraries[J]. Information Technology and Libraries, 2001, 20(4): 185-196

[3] J Konstan, B Miller, D Maltz, *et al*. GroupLens: Applying collaborative filtering to Usenet news[J]. Communications of the ACM, 1997, 40(3): 77-87

[4] Gao Fengrong. Research on the key techniques of personalized recommender systems: [Ph D dissertation][D]. Beijing: Renmin University of China, 2003 (in Chinese)
(高凤荣.个性化推荐系统关键技术研究:[博士论文][D].北京:中国人民大学, 2003)

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

[6] G Adomavicius, A Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions[J]. IEEE Trans on Knowledge and Data Engineering, 2005, 17(6): 734—749

[7] Chun Zeng, Chun-Xiao Xing, Li-Zhu Zhou, *et al*. Similarity measure and instance selection for collaborative filtering international[J]. Journal of Electronic Commerce, 2004, 4(8): 115—129

[8] G Karypis. Evaluation of item-based top-*N* recommendation algorithms[C]. In: Proc of CIKM 2001. New York: ACM Press, 2001. 247—254

[9] Brendan Kitts, David Freed, Martin Vrieze. Cross-sell: A fast promotion-tunable customer-item recommendation method based on conditional independent probabilities[C]. In: Proc of ACM SIGKDD Int'l Conf. New York: ACM Press, 2000. 437—446

[10] KDD2000 Dataset[OL]. http://www.ecn.purdue.edu/KDDCUP/data/, 2005

Xing Chunxiao, born in 1967. Professor and director of Web and Software Technology Research and Development Center of Tsinghua University. His main research interests include database technology, software engineering, personalized servise, and digital library, digital government.

邢春晓, 1967 年生, 研究员, 清华大学 Web 与软件研究中心主任. 主要研究方向为数据库技术、软件工程、个性化服务、数字图书馆、电子政务.

Gao Fengrong, born in 1975. Postdoctor in the Department of Computer Science and Technology of Tsinghua University. Her current research interests include personalized service, digital library and e-government.

高凤荣, 1975 年生, 博士后, 主要研究方向为个性化服务、数字图书馆、电子政务.

Zhan Sinan, born in 1980. Master. His current research interests include personalized service, data mining, and digital library.

战思南, 1980 年生, 硕士, 主要研究方向为个性化服务、数据挖掘、数字图书馆.

Zhou Lizhu, born in 1947. Professor and Ph. D. supervisor. His main research interests include database technology, digital library, and etc.

周立柱, 1947 年生, 教授, 博士生导师, 主要研究方向为数据库技术、数字图书馆等.

Research background

Our work is supported by the National Natural Science Foundation of China No. 60473078, named "Research on Theory and Technology of Personalized Service Based on Information Filtering". In the construction of large application systems, such as digital library, E-government, and E-commerce, how to provide users with efficient personalized service has became an important and challenging work. This project analyzes the key technologies and the related works of personalized service. We will conduct research on the theory and methods of personalized service based on information filtering, including content-based filtering and collaborative filtering. Based on the above research, we will design and implement a recommender system for personalized search and automatic recommendation.