基于用户兴趣子类的协作推荐算法*)

朱征宇　张小林　熊　茜　谢祈鸿
(重庆大学计算机学院　重庆 400044)

摘　要　随着电子商务规模的进一步扩大，用户数目和文档资源急剧增加，导致用户数据的极端稀疏性。传统协作推荐算法都无法很好地解决数据稀疏性问题。本文提出一种基于兴趣子类的协作推荐算法，通过子类处理思想的引入，使得某两个用户即使整体不相似而因为"局部点"的相似产生有用的推荐，"最近邻居"的发现变得更容易更准确。实验结果表明，该算法能有效地解决用户数据的极端稀疏问题，在同等条件下，相对于传统协作推荐算法[9]有更好的推荐质量。

关键词　兴趣子类，兴趣分类树，协作推荐，数据稀疏性，平均绝对误差

An Algorithm of Collaborative Recommendation Based on User's Interest Sub-Class

ZHU Zheng-Yu　ZHANG Xiao-Lin　XIONG Qian　XIE Qi-Hong
(College of Computer, Chongqing University, Chongqing 400044)

Abstract　With the development of E-commerce, the magnitudes of users and Web documents grow rapidly, and result in the extreme data sparseness of users. The traditional algorithms of collaborative recommendation can't solve the problem very well. To address this issue, a novel algorithm of collaborative recommendation based on users' interest sub-classes is proposed. Based on the similarity of the interest sub-classes among the users, the new method makes it more easy and accurate to find the similar neighbors of a user, even if their interests are very different as a whole, and can provide more efficient information recommendation. Our experiment shows that this method can efficiently solve the problem of the extreme data sparseness of users, and can provide better result on information recommendation than the traditional algorithm of collaborative recommendation [9].

Keywords　Interest sub-class, Interest category tree, Collaborative recommendation, Data sparseness, Mean Absolute Error (MAE)

1　相关工作

在当今的个性化推荐系统中，基于最近邻居的协作过滤技术由于具有发现新信息的能力，是应用最为成功的推荐技术之一[3]。它通过分析用户兴趣，在用户群中找到当前用户的相似(兴趣)用户，综合这些相似用户对某一信息的评价，形成系统对该用户对此信息喜好程度的预测。传统的基于最近邻居的协作推荐方法，所使用的用户信息数据集通常可表示为一个 $m\times n$ 的用户-项评估矩阵 $R=(r_{ij})_{m\times n}$，m 是用户数，n 是项数，r_{ij} 是第 i 个用户对第 j 个项的评分。评估值与项的内容有关，如果项是电子商务中的货品，则表示用户订购与否，例如1表示订购，0表示没有订购；如果项是Web文档，则表示浏览与否，用户对它的兴趣度有多高。协作推荐过程可分为3个阶段：数据表述；发现最近邻居；产生推荐数据集。

随着电子商务规模的进一步扩大，用户数目和文档资源急剧增加，导致用户数据的极端稀疏性，最近邻居不易获得。而准确发现目标用户的最近邻居是协作推荐系统成功的关键。针对此问题，许多文献都提出了不同的解决办法。

文[1]提出了基于最近邻用户的协作过滤方法，但在实践过程中其稀疏性和扩展性的缺点也逐渐暴露出来。文[9]提出了基于项目预测评分的协作过滤推荐算法，每次预测都要计算项目的相似性，其计算复杂性，随着用户数目和项目数的增加系统的性能和扩展性仍然较差。

文[8]针对数据稀疏性问题，提出通过奇异值分解(SVD)减少项目空间的维数，使得用户在降维后的项目空间上对每一个项目均有评分，实验结果表明，这种方法可以有效地解决同义词(synonymy)问题，显著地提高推荐系统的伸缩能力。但降维会导致信息损失，降维效果与数据集密切相关，在项目空间维数很高的情况下，降维的效果难以保证。而且，在以内容为基础的网页推荐过程中，忽视了网页内容之间潜在的关系，相似邻居发现并不准确，不能精确地产生推荐集。

在早先的一些协作过滤系统中，都有一个共同点：将用户当作一个"整体"来寻找其相似用户群，没有考虑到将用户兴趣分类。实际上，每个人的兴趣是很广泛的，所以很难找到与其完全相似的用户，实验也表明，某人对一篇文档感兴趣与该文档所属类别有一种潜在的关系，所以我们应该尽量使文档主题类别与用户兴趣类别相一致。因此，将子类处理思想应用于协作推荐应该是可行的。

原则上讲，将基于子类的处理思想应用于用户群中用户兴趣的相似度度量，应该优于早先基于整体的处理思想。例如，用户甲喜欢[体育][娱乐][文教][IT]等多方面的新闻，用户乙则喜欢[体育][社会][财经]方面的新闻，如果[体育]新闻(或相应特征词的权重在用户的所有特征词中)所占比重又较小，就可能使此两用户相似度较低。但采用基于兴趣子类的方法，只对[体育]子类进行相似度计算，两用户可能因为"局部点"相似而成为在某一子类上的最近邻，这种"按照子类划分的相似用户群、并进行子类信息协作推荐"的处理方法，在多数情形应更加合理。

本文提出一种基于兴趣子类协作过滤的推荐算法。它以网页内容为基础，首先对用户的兴趣分类，然后以兴趣子类为

*)基金项目：重庆大学骨干教师资助计划项目(2003A33)。朱征宇　副教授，研究方向为Web智能检索、电子商务、数据库技术；张小林　硕士研究生，研究方向为电子商务、Web智能检索；熊　茜　硕士研究生，研究方向为Web服务；谢祈鸿　硕士研究生，研究方向为电子商务。

单位来发现某一子类的相似用户群，得到该用户的所有兴趣子类的多组相似用户群，并形成每个子类的“兴趣子类－子类网页项”矩阵。

该方法的优点是：(1)以网页内容为基础来划分用户兴趣，并以兴趣子类为单位来发现某一子类的相似用户群，从而对用户－项矩阵实现降维，能有效地解决传统协作过滤中普通存在的数据稀疏性问题。(2)与传统协作推荐方法相比，基于子类的最近邻更容易获得，也更准确，能产生更为精确的推荐。(3)初步实验表明，本文提出的方法可以有效地解决传统协作过滤方法在用户评分数据极端稀疏情况下存在的不足，显著地提高信息推荐的质量。

2 基于兴趣子类的协作推荐模型

其结构模型如图 1 所示。1)首先通过一组代理捕获用户行为，分析用户对该网页的兴趣度，并生成当前用户的兴趣分类树。2)然后针对用户的每一个兴趣子类寻找其相似邻居，分别得到当前用户的多个兴趣子类的相似用户群，并获得 m（兴趣类别数）个“兴趣子类－子类网页项”矩阵。3)先对当前子类的最近邻居未浏览过网页预测评分，再与当前用户的该子类分类兴趣度相结合，最后求出其 top-N 集。4)进行信息推荐。

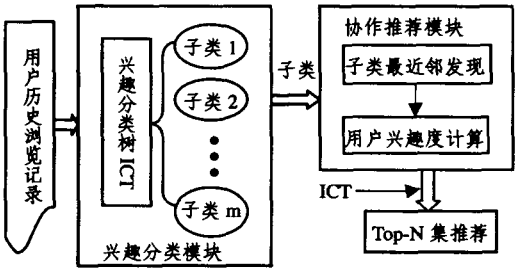


图 1 基于兴趣子类的协作推荐

下面将对模型中的核心模块设计进行说明。

2.1 用户兴趣描述模型(User Profile)

为了获取用户的信息需求并对用户兴趣分类，需要建立用户的兴趣模型。而用户兴趣是通过记录和分析用户对 Web 文档的浏览行为得到的，所以用户兴趣的表示应与 Web 文档的表示相一致，因此采用一个向量来描述用户在某一领域的兴趣。

假设用户至多有 M 个兴趣领域，即兴趣子类（例如本文中将新闻类网页分为军事、体育、政治等十个子类），每个子类用 n 个关键词描述，则应用向量空间模型，可将用户的每一个兴趣子类表示为一个向量 $Q_i=(q_1,q_2,\cdots,q_n)$，而用户兴趣描述文件 U 可表示为一向量矩阵：

$U=(Q_1,Q_2,\cdots,Q_m)$

2.2 兴趣分类树 ICT(Interest Cate-gory Tree)

为建立用户的兴趣分类树，先给出网页兴趣度的概念及其运算。

网页兴趣度是指用户对一个网页内容的感兴趣程度，采用 0-1 间的实数表示，0 和 1 分别表示无兴趣和最大兴趣。大量研究表明用户对网页的兴趣度与他在该网页上的浏览行为密切相关。综合文[2～5]的研究，认为用户对某网页的兴趣度 $P_{interst}$ 起关键作用的两种行为是：在网页 P 上的浏览时间 $t(P)$（简称 T 行为）和翻页/拉动滚动条次数 $v(P)$（简称 V 行为）。我们用多元回归方法[10]来表示 T、V 行为与网页兴趣度 $P_{interst}$ 之间的定量关系。即用户对某网页的兴趣度 $P_{interst}$ 可表示为：

$$P_{interst}=a*t(P)+b*v(P)+c \quad (1)$$

其中 $P_{interst}$ 是与自变量 $t(P)$、$v(P)$ 有关的随机变量，a 和 b 称为回归系数（本文称 a,b,c 为行为影响因子）。

用户随意浏览一组网页（一般只需 30～40 个），系统自动捕获相应的一组 $t(P)$、$v(P)$ 值，并通过调用 Matlab 中的多元线性回归分析函数[10]，可以自动计算出每个不同用户的行为影响因子。从而对以后用户浏览的任何网页，根据式(1)，系统即能自动计算出用户对该网页的兴趣度。

兴趣分类树 ICT 就是将一个人的兴趣划分为几个子类，并描述用户对哪些子类感兴趣，以及对子类兴趣程度的一种量化。其生成过程主要分为两步：一是采用分类算法对用户浏览的网页按照主题类别分类，获得用户的兴趣分布；二是计算用户对各网页的兴趣度，从而计算出各主题类别的兴趣度，进而生成 ICT。

特别指出，ICT 的确定与应用领域有关。由于本文主要针对新闻文本领域，参照图书分类法，将新闻分为军事、体育、财经等十类。对标准分类树 SCT(Standard Category Tree)，其分类粒度需根据应用要求做一定调整。因为，如果分得过细，会使基于子类向量的相似用户群太少，从而不能为用户发现新信息。如果分得过粗，则子类思想的优势得不到体现，退化为基于用户整体的协作过滤。此外，为能在网页自动分类时采用序列算法[6]（该算法省略掉训练过程易于实现，并借助语义分析分类精度较高），要求为 SCT 中各子类都确定一个本地字典[8]（一组代表本分类的关键词），可根据领域知识或专业词典得到。

本文对用户所浏览过的网页采用序列算法[6]与新闻文本的各个子类本地字典进行匹配，直至最终得到用户的兴趣分类树 ICT。我们用 Interest (C_i) 来描述用户对子类 C_i 的兴趣大小，可表示为：

$$\text{Interest}(C_i)=\sum_j(P_{ij})/\sum_i\sum_j(P_{ij}) \quad (2)$$

P_{ij} 表示用户对子类 C_i 中某网页 j 的兴趣度，$\sum_j(P_{ij})$ 表示 C_i 类中网页的兴趣度之和，$\sum_i\sum_j(P_{ij})$ 表示所有类中网页的兴趣度总和。本文基于新闻文档的 ICT 模型如图 2 所示。

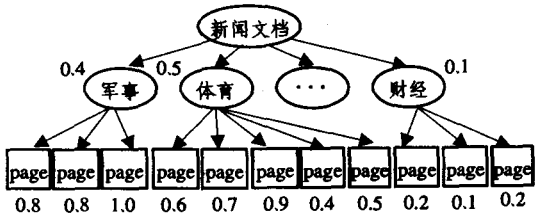


图 2 基于新闻文档的 ICT 示意图

2.3 基于兴趣子类的用户兴趣分类算法

首先，用代理捕获用户一段时间内（如浏览 40 个网页）的浏览行为，采用 Matlab 中的多元线性回归分析函数，通过式(1)，计算出该用户的行为影响因子。

算法 1：基于兴趣子类的用户兴趣分类算法。

输入：用户的历史浏览网页记录

输出：用户的兴趣描述文件 U、网页兴趣度 P、子类兴趣度 Interest(C_i)。

步骤：

step1：设置用户兴趣描述文件向量集 U，初始为空。

Step2：获得兴趣分类树：

(1)对每个历史浏览网页文档 d_i 做预处理。解析出该文档的词语,然后进行相关的处理,如根据停止词表(stop words list)删除停止词处理等,再计算文档的 $tf*idf$[10]表示为向量 d_i。

(2)通过式(1)计算用户对该文档 d_i 的网页兴趣度。

(3)采用序列算法[6]对该文档自动分类,形成用户的兴趣分类树 ICT。

(4)根据计算出的所有网页兴趣度运用式(2)得到子类兴趣度 Interest(C_i),并标注兴趣分类树(如图 2)。

Step3:建立用户兴趣模型:

(1)分别合并用户 ICT 中同一个子类下的所有网页向量,得到当前用户对应的该子类向量 Q_j,并写入 U,$Q_j=d_1+d_2+\cdots+d_k$(k 表示该子类中的网页数)。这里的"+"运算是:如果某一个词在两个以上的网页中,那么把该词的权值相加所得的和作为它在 Q_j 中的权值;如果某一个词只在一个网页中,那么把该词和其所对应的权值加入到向量 Q_j 中。

(2)将新得到的向量 Q_j 中的词按其权重大小从大到小进行排序,取前 n 个元素作为向量 Q_j 中的元素。

(算法结束)

3 基于兴趣子类的协作推荐

下面介绍基于兴趣子类的协同过滤推荐算法。算法主要分为两步:寻找最近邻居和产生推荐。由于协同过滤技术的用户-项矩阵的数据表述方法所带来的稀疏性严重制约了推荐效果,又忽视了网页内容之间潜在的关系,导致相似兴趣的用户难以建立联系,相似用户群不够准确。所以,本文在前面将兴趣分类后,以兴趣子类为基础,建立 m(兴趣子类数)个"兴趣子类-子类网页项"矩阵 R。可达到对传统的用户-项矩阵维数简化的目的,而不会导致信息损失,如图 3 所示。

$User_i \supseteq Q_1$ 表示包含兴趣子类向量 Q_1 的用户,$Web_j \in Q_1$ 表示属于子类 Q_1 的网页,$R_{i,j}$ 表示用户 i 对网页 j 的兴趣度。相对于传统的 $m\times n$ 用户-项评估矩阵,本文矩阵中 $L\leqslant m$,$K\leqslant n$。

$Web_j \in Q_1$

$User_i \supseteq Q_1$	Web_1	…	Web_j	…	Web_K
$User_1$	$R_{1,1}$	…	$R_{1,j}$	…	$R_{1,K}$
…	…	…	…	…	…
$User_i$	$R_{i,1}$	…	$R_{i,j}$	…	$R_{i,K}$
…	…	…	…	…	…
$User_L$	$R_{m,1}$	…	$R_{L,j}$	…	$R_{L,K}$

图 3 兴趣子类-子类网页项矩阵

3.1 最近邻居发现

采用向量空间方法计算用户间在某一子类上的相似性,来寻找在每个子类的最近邻。这里分析的对象是经过兴趣分类算法分解后的"兴趣子类－子类网页项矩阵",我们计算的是不同用户 Profile 在同一分类上的子类向量的相似度,而不是用户兴趣描述文件(User Profile)整体上的相似度。例如,设 Q_{si} 为用户 i 的体育子类向量,Q_{sj} 为用户 j 的体育子类向量,这样,二者的体育兴趣子类的相似度 $S_{i,j}$ 表示为:

$$S_{i,j}=\cos(Q_{si},Q_{sj})=\frac{Q_{si}\cdot Q_{sj}}{|Q_{si}\|Q_{sj}|} \tag{3}$$

注意:因为我们已经建立了用户兴趣模型,所以,这里采取的是基于内容的两个子类向量的比较。这样,就能够更准确更多地找到基于子类的最近邻,使得本文提出的相似邻居发现算法优于早先的基于用户评分向量的余弦相似性算法,第 4 节进行了实验验证。

3.2 用户兴趣度的计算

通过对子类向量运用余弦相似性度量方法得到目标用户的几组最近邻居群,设定选取的最近邻个数 K 值,并删除"兴趣子类-子类网页项"矩阵中的非最近邻居。为了简化算法的复杂性,本文将用户对未评分项目的评分设为用户的平均评分。由于在本文提出的矩阵中,都是属于同一子类的网页,因此用户都有可能感兴趣,能为用户发现新信息。实验也表明,这种改进方法可以有效地提高协同过滤推荐系统的推荐精度[1]。

下一步需要计算用户对矩阵中任意项的兴趣度。设用户 x 的最近邻居集合用 NBS_x 表示,则用户 x 对网页 i 的预测评分 $P_{x,i}$ 可以通过用户 x 对最近邻居集合 NBS_x 中网页的评分得到,计算方法如下[1]:

$$p_{x,i}=\bar{R}_x+\frac{\sum_{y\in NBS_x} sim(x,y)\times(R_{y,i}-\bar{R}_y)}{\sum_{y\in NBS_x}(|sim(x,y)|)} \tag{4}$$

$sim(x,y)$ 表示用户 x 与用户 y 之间的相似性,$R_{y,i}$ 表示用户 y 对网页 i 的评分,$\bar{R}_x$、$\bar{R}_y$ 分别表示用户 x 和用户 y 对项目的平均评分。

3.3 top-N 推荐集的产生

因为用户对各子类的分类兴趣度不一样,所以在给用户推荐网页时应结合用户的兴趣分类树,以推荐不同子类的网页数目多少来表征用户的不同兴趣分布。如确定 top-N 集 $N=10$,则每个子类推荐的网页数目 C 值为:

$$C=interest(C_i)\times N \tag{5}$$

Interest(C_i)表示用户对第 i 个子类的子类兴趣度。

3.4 基于兴趣子类的协作推荐算法

算法 2:基于兴趣子类的协作推荐算法。

输入:用户兴趣模型 U 和兴趣分类树。

输出:推荐给用户的 top-N 集网页。

步骤:

step1:发现兴趣子类的最近邻居

for 1 to m

// m 表示当前用户的兴趣子类数,$m\leqslant M$(用户兴趣的最多子类数,本文为 $M=10$)。

{

(1)取出当前用户的兴趣描述文件 User Profile; //由算法 1 获取;

(2)生成 m 个"兴趣子类-子类网页项"矩阵 R,并把当前用户对该网页的兴趣度 P 值写入矩阵 R;

(3)对每一个兴趣子类向量 Q_i 运用式(3),求出 Q_i 的最近邻并输出其相似度 S,设定选取的最近邻个数 K 值;

//寻找子类的最近邻

(4)删除"兴趣子类-子类网页项"矩阵 R 中非最近邻居的行,并将用户对未评分项目的评分设为用户的平均评分。

(5)预测用户对矩阵中任意子类网页项的兴趣度,利用式(4)计算获得;

}

step2: 产生 top-N 推荐集

(1)对每个矩阵中用户未浏览过网页的兴趣度按照从大到小排序。

(2)利用式(5)计算出每一个兴趣子类的推荐网页数目

C。

(3)按照每个子类的 C 值输出前 10 项(可由系统设定)作为 top-N 集推荐给用户。

(算法结束)

4 实验结果及其分析

4.1 数据集

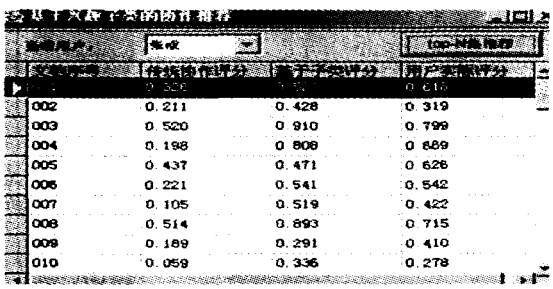
本文实验语料是选自中国新闻网(http://www.chinanews.com.cn)和中搜新闻中心(http://news.zhongsou.com)搜索出的部分新闻集合,它共包含 495 篇标注好类别的文档。由于部分类别样本数量太少而不具有测试的可靠性,因此只选取军事、体育、财经、科教、娱乐 5 个主题类别鲜明的类别进行测试。实验语料中属于这 5 个类别的文本数分别为 108、130、69、68、120 篇,每个类别的本地字典(一组代表该子类的关键词)是从相应的专业字典中获得,其中含有的元素个数分别为 136、159、112、107、148。我们采用序列算法[6]对用户兴趣分类(因该算法的 BEP 值可达 83%)。

我们人工模拟了 30 个用户,并且每个用户至少浏览过 30 篇文档。这里,通过编制的代理模块,能够自动捕获用户浏览行为并记录下行为数据并计算得到了所有用户的行为影响因子,从而能够计算各网页的兴趣度值。用户评分数据共有 1000 条,整个数据集的稀疏性定义为用户-项矩阵中未评分文档所占的百分比,为:

$1-(1000/30\times495)=0.9327$

当前用户张成(为 30 个用户之一)浏览过 70 篇网页(具有实际评分),我们只把其中 30 篇看作他的历史浏览记录,把其中 40 篇作为测试数据集(testing set)。

图 4 为本文实验系统快照。用户数目为 30,评分数据 1000 条时,采用传统面向用户的协同过滤技术和本文提出的基于兴趣子类的协作推荐算法,当前用户张成对 001 至 010 号新闻文档的评分对照结果。



002	0.211	0.428	0.319
003	0.520	0.910	0.799
004	0.198	0.808	0.689
005	0.437	0.471	0.626
006	0.221	0.541	0.542
007	0.105	0.519	0.422
008	0.514	0.893	0.715
009	0.189	0.291	0.410
010	0.059	0.336	0.278

图 4 实验系统快照

评价推荐系统推荐质量的试题标准主要包括统计精度度量方法和决策支持精度度量方法两类[7]。参照文[9]本文采用平均绝对偏差 MAE 作为度量标准,平均绝对偏差 MAE 通过计算预测的用户评分与实际的用户评分之间的偏差度量预测的准确性,MAE 越小,推荐质量越高。设预测的用户评分集合表示为 $\{p_1,p_2,\cdots,p_N\}$,对应的实际用户评分集合为 $\{q_1,q_2,\cdots,q_N\}$,则平均绝对误差 MAE 定义为[7]:

$$MAE=\frac{\sum_{i=1}^{N}|p_i-q_i|}{N}$$

4.2 数据稀疏性对算法的影响

为了检验本文提出的算法的有效性,我们选择文[9]中面向用户的协同过滤技术(下简称"Tadition CF"),作为对照。同时,对本文提出的基于用户兴趣子类的协作推荐算法分别采取基于序列算法和人工区分两种方式来进行兴趣分类(分别简称"序列算法分类"和"人工兴趣分类")。两种算法都以余弦相似性作为相似性度量标准,计算其 MAE。图 5 是在用户数为 30,最近邻 K 值为 16 时,用户评分数据从 1000 增加到 3500 条,间隔数为 500,三种方法 MAE 值的变化情况曲线图。

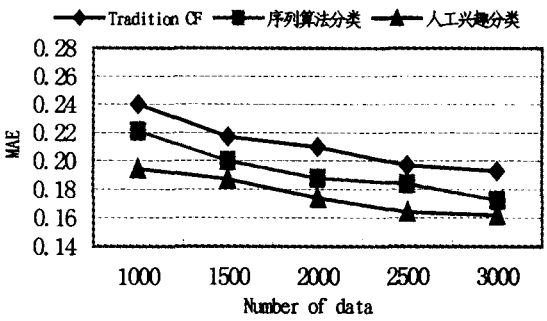


图 5 算法的推荐精度比较

从图中可知,随着评分数据增多,本文提出的基于兴趣子类的协作推荐算法具有较小的 MAE 值;尤其是在数据极端稀疏的情况时,并且用户兴趣分类较准确情况下(人工对兴趣分类,其准确率可达 98%),本文提出的算法有明显的优势。相对传统方法 Tadition CF,本文提出的方法可以更有效地提高系统的推荐质量。

4.3 最近邻居 K 值对算法的影响

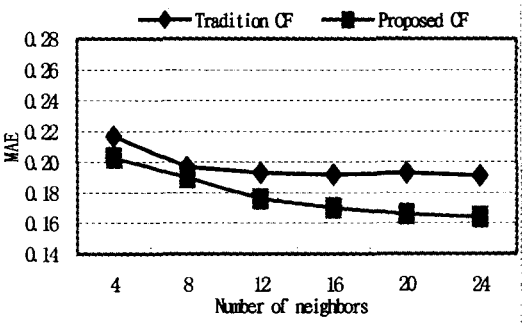


图 6 最近邻居 K 值分析

最近邻居集 K 值大小,关系到系统的推荐的效果,应适当大些,但又不能过分影响计算效率,而本文算法的优势正是在于较容易获得最近邻。所以,为了进一步验证本算法的有效性,我们分析在数据较为稀疏情况下,最近邻 K 值大小对两种算法执行效果的影响。用户评分数据为 3000 条,最近邻个数从 4 增加到 20,间隔数为 4,计算其 MAE。如图 6 所示,当 K 值较小时,两种算法的 MAE 相差不大;但当 K 值较大时,其 MAE 相差较大,其原因在于:面向用户的 Taditon CF 方法的最近邻居不易获得,用户间相似度较低,而本文提出的协作推荐算法(Proposed CF)的子类最近邻居更容易获得,子类间相似度也较高。因此,本文提出的算法能更准确地预测用户兴趣度。

4.4 实验分析

(1)系统用户较少时,本文所提出的方法相对传统协作推荐效果应更明显。因为系统用户少,基于整体思想的相邻用户更少且不易获得,而基于子类思想是要发现具有部分相同

兴趣领域的用户，所以更容易获得，MAE 值更小。

(2)基于兴趣子类的协作推荐在解决矩阵稀疏性问题上比传统协作推荐方法更为有效。这是因为基于整体处理思想在共同评分数据较少时，不易成为最近邻居。而基于兴趣子类就有可能因为某局部点相似而成为最近邻产生推荐。

(3)兴趣分类的准确与否对推荐效果有一定的影响。在同等条件下，本文提出的推荐方法也优于传统的基于"整体"协作过滤的效果，但兴趣分类越准确，其子类的最近邻居也就越准确，MAE 值也更小，越能体现本文提出的子类思想。

结论 由于传统协作过滤推荐算法都无法完全解决数据稀疏性问题，而这个问题的关键是数据极端稀疏情况下，其最近邻居不易获得。针对此问题，本文提出将子类处理思想应用于新闻文档的协作推荐领域，使得某两个用户即使整体不相似，而因为"局部点"的相似而产生有用的推荐，使得最近邻居的发现变得更容易更准确。实验结果表明，基于兴趣子类的协作过滤推荐算法能够有效地解决数据极端稀疏性问题。在同等条件下，相对于传统面向用户的协同过滤推荐算法[9]有更好的推荐质量。

参 考 文 献

1 Breese J, Hecherman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering. In: Proc. of the 14th Conf. on Uncertainty in Artificial Intelligence(UAI 98), 1998. 43～52
2 3Claypool M, Le P, Waseda M, et al. Implicit interest indicators. In: Campbell M., ed. Proc. of the ACM Intelligent User Interfaces Conf. (IUI), New York: ACM Press, 2001. 14～17
3 曾春，邢春晓，周立柱. 个性化服务技术综述. 软件学报，2002，13(10)：1952～1960
4 Lieberman H. Letizia: an agent that assists web browsing. In: Burke, R., ed. Proc. of the Intl. Joint Conf. on Artificial Intelligence. Menlo Park, CA: AAAI Press, 1996. 54～61
5 谭琼，李晓黎，史忠植. 一种实现搜索引擎个性化服务的方法. 计算机科学，2002，29(1)：23～25
6 解冲锋，李星. 基于序列的文本自动分类算法. 软件学报，2002，13(4)：783～789
7 Sarwar B, Karypis G, Konstan J, Riedl J. Item-Based collaborative filtering recommendation algorithms. In: Proc. of the 10th Intl. World Wide Web Conf. 2001. 285～295
8 赵亮，胡乃静，张守志. 个性化推荐算法设计. 计算机研究与发展，2002，39 (8)：986～991
9 邓爱林，朱扬勇，施伯乐. 基于项目评分预测的协同过滤推荐算法. 软件学报，2003，14 (09)：1621～1628
10 宋斌，方小璐. 基于网页特征的 TFIDF 改进算法[J]. 微计算机应用，2002，23(1)：18～20

(上接第 102 页)

Client 客户端对应 WebService 服务接口。客户端的显示设计采用 SVG 进行显示，并采用 XSLT 来进行 GML 和 SVG 之间的格式转换；

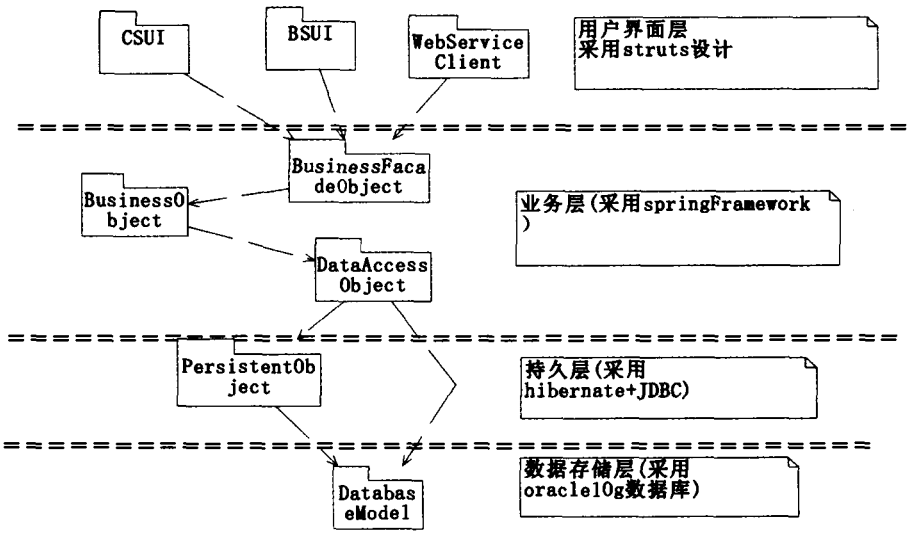


图 5 GIS Server 服务器的部署图

第二层是业务层，主要负责 GIS Server 的业务处理，采用 SpringFrame 轻型框架进行设计，分为 3 个包，BusinessFacadeObject 包主要包括接受用户请求的类，负责用户负载均衡的类，调用业务类的接口等类；BusinessObject 包是 GIS Server 设计中的核心包，包括 GIS Server 中所有业务类，包括数据管理类，数据绘制类，GML 表示类，空间分析类，矢量数据类，栅格数据类，地址类，空间坐标系类，服务管理类等；DataAccessObject 包是负责连接数据层或持久层之间的接口类；

第三层是持久层，有一个 PersistentObject 包，存放各种持久对象；采用 Hibernate 技术进行设计；

第四层是数据层，有一个 DatabaseModel 包，存放地址地理编码数据模型，矢量数据模型，影像数据模型等类和表对象；采用 oracle 10g 进行存储。

结语 GIS Server 服务器是基于 SOA 框架进行设计的中间件平台系统，提供符合业届标准的地理服务和接口，具有良好的封装性、松散耦合性、分布性、标准性等特点；由于其体系架构的优越性使得 GIS Server 服务器必将成为新一代网络 GIS 发展的方向，为数字城市的建设添砖加瓦。

参 考 文 献

1 刘纯波. 数字城市空间信息资源管理与集成调度技术研究：[学位论文]. 北京大学，2003
2 邬伦，唐大仕，刘瑜. 基于 Web Service 分布式互操作的 GIS. 地理信息科学，2003，19(4)：29
3 OpenGIS Consortium, Inc . OpenGIS Web Map Server Cookbook2003. 2003-8
4 OpenGIS Consortium, Inc . OpenGIS Web Feature Server Specification. 2001－11
5 OpenGIS Consortium, Inc . OpenGIS Web Coverage Serice . 2002-4