个性化信息过滤系统中用户兴趣模型建立和更新①

Establishing and Updating of User Profile in Personalized Information Filtering System

费洪晓　戴 弋　穆 珺　黄勤径　肖新华　（中南大学信息科学与工程学院 长沙 410075）

摘要：提出了一种新的用户兴趣模型的建立和更新方法。该方法通过 web 内容挖掘和用户行为分析建立用户模型，而通过移动时间窗口更新用户模型。实验表明，该用户兴趣模型能有效提高检索精度，提高个性化信息服务的效率。

关键词：兴趣模型 web 内容挖掘 行为分析 时间窗口

目前，随着“信息过载”和“信息迷向”问题的加剧，人们对搜索引擎的功能、智能化程度有了更高的要求，希望它们能提供更准确、更精炼和更符合个人需求的检索结果。信息过滤作为对现有信息检索系统的补充，不但设法提供给用户感兴趣的信息，而且利用用户模型（User Profile）[1] 记录每个用户的兴趣，这为信息检索提供个性化、智能化的服务。

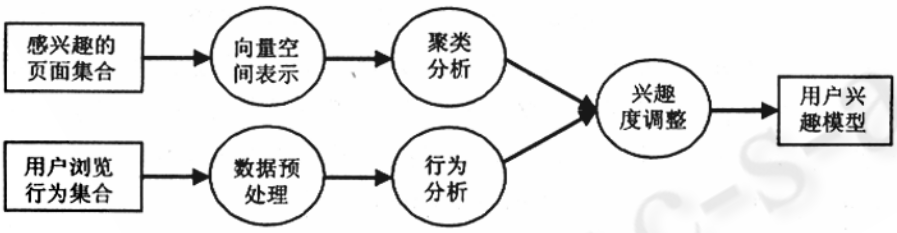

图 1　用户兴趣模型的建立数据流图

本文提出一种新的兴趣模型的创建和更新方法。用 Web 内容挖掘和行为分析创建用户兴趣模型，而用移动时间窗口更新兴趣模型。这样得到的用户兴趣模型更能够适应用户兴趣的变化，提供更加准确的个性化的服务。

1　基于 web 内容挖掘和行为分析的用户兴趣模型的建立

用户浏览 web 主要方法为：打开站点的主页面，点击一系列链接来打开页面。这样，用户为了访问到自己感兴趣的页面（内容页面），都需要经过一系列其它的页面（辅助页面），而辅助页面并不是真正用户感兴趣的页面，所以需要通过 web 预处理和聚类分析得到用户真正感兴趣的页面集合；另一方面，行为分析对用户浏览页面时的行为信息进行分析，得到用户对单一页面的兴趣度，再将两者结合，得到了用户感兴趣的主题类别和每个主题的兴趣度，从而得到用户兴趣模型。如图 1 所示。

1.1　Web 内容挖掘

聚类分析是把不同文本组与相应得文本主题对应的过程。它是一个发现文本集包含内容的方法，适合于基于文本内容的用户兴趣模型的建立。目前，文本聚类方法有很多，包括 k－最近临近聚类法、层次聚类法、平面划分法等。各种聚类算法都有其优缺点，如平面划分法必须先确定聚类参数 K，这违背了无监督分类的意图；层次聚类算法缺点是每个凝聚或分裂过程一旦完成就不能撤销。所以本文用改进的凝聚的层次聚类法与 K－Means 相结合的聚类算法来对 Web 网页进行聚类，得到更为准确的用户兴趣模型。

1.1.1　web 页面向量空间表示

与数据库中结构化数据相比，web 文档的结构有

① 科技项目：湖南省科技计划项目（2006JT1040）

限,或根本没有结构。此外,文档的内容是计算机难以理解的自然语言,需要对文本进行预处理,抽取代表其特征的元数据。特征表示模型有两种,一种是基于语义的方法,它试图完成一定程度的语法和语义分析。另一种是基于统计的方法,它是将每个 Web 页面表示成特定词的集合。常用的文本特征表示方法有布尔逻辑型、向量空间型和概率型。近年来应用效果较好的文本特征表示法是向量空间模型(Vector Space Model,VSM)法,该方法将文档用 n 维空间向量表示,每篇文档都转换为一个 n 维特征向量,本文采用向量空间模型表示法。

特征项 V 用词表示,根据试验结果认为,选取词作为特征项要优于字和词组。其权重计算方法常用的有 TF * IDF 法、信息增益、信息熵、互信息等,由于 TF – TDF 法抑制冗余信息效果比较好,且易于实现,所以这里采用 TF – TDF 法,计算文本 D 中特征词 t 权重的公式如下:

$$w(t,\bar{d}) = \frac{tf(T,\bar{d}) \times \log(N/n_t + 0.01)}{\sqrt{\sum_{t \in d}[tf(T,\bar{d}) \times \log(N/n_t + 0.01)]^2}} \quad (1)$$

其中,TF 表示词条频度因子(Term Frequency),即词条在文本中出现的频率;N 代表文本集中文本总数,n_t 代表出现了 t 的文本数,即词条 t 的文档频度。

1.1.2 Web 页面聚类分析

Web 网页聚类是一种无监督的分类过程。其主要方法有两种,一种是基于概率的方法,其主要以贝叶斯概率理论为基础,用概率分布方式描述聚类结果。另一种是基于距离的方法,用特征向量表示文档,并将文档看作是向量空间中的一个点,通过计算点之间的距离进行聚类。常用的聚类算法有层次聚类法和平面划分法,但是这两种方法都有其缺点,后来提出了凝聚的层次聚类法与 K – Means 相结合的聚类算法,该方法解决了 K – Means 算法要求用户输入参数 k 的问题;但是该算法还存在一些问题,如没有解决孤立点的敏感性问题。

本文采用改进的凝聚的层次聚类法与 K – Means 相结合的聚类算法来对 Web 网页进行聚类,(1)在利用凝聚的层次聚类方法初始化质心向量时,将聚类停止条件设置为:当前具有最大相似度的类之间的相似度与最初合并的类之间的相似度小于某个确定百分比。(2)在 K – Means 算法的每一次进行划分之后,先取得 K 个簇的聚类结果,根据文本与所在簇相似度对聚类结果进行排序,将所有相似度小于特定值的文本视为孤立点从其所在的簇中删除。对于给定的文档集合 $D = \{d_1, \cdots, d_i, \cdots, d_n\}$,聚类过程的算法为:

(1) 将文本集 D 中的每个文本 d_i 视为一个具有单个成员的类 $C_i = \{d_i\}$,构成一个聚类 $C = \{C_1, C_2 \cdots, C_n\}$;
(2) Repeat;
(3) 对于 C 中每对类 (C_i, C_j),计算它们之间的相似度 $sim(C_i, C_j)$,取具有最大相似度的类之间的相似度值 $max = MAX\{sim(C_i, C_j)\}$;
(4) 将 C_i 和 C_j 合并为一个新的类 $C_k = C_i \cup C_j$,从而构成了 D 的一个新的聚类 $C = \{C_1, \ldots, C_i, \ldots C_k\}$;
(5) Until 达到结束条件,当前距离最近文本距离/第一次合并的文本间距离 $\leq \alpha(\alpha = 0.2)$;
(6) 将初始质心向量和 K 值作为 k – means 聚类的种子 $s = \{s_1, \cdots, s_i, \cdots, s_k\}$;
(7) 依次计算 D 中的 每个文档 d_i 与每个种子 S_j 的相似度 $sim(d_i, S_j)$;
(8) 选取具有最大相似度的种子 $MAX\{sim(d_i, S_j)\}$,将 d_i 归入以 S_j 为聚类中心的类 C_j;
(9) 对聚类结果排序,将相似度 $\leq \beta$ 的文本删除;
(10) 重复步骤(6) ~ (9),最后得到较稳定的聚类结果 $C = \{C_1, \cdots, C_k\}$。

1.2 用户浏览行为分析

研究表明,用户的很多动作都能暗示用户的喜好,如:查询、标记书签、拖动滚动条、前进、后退、用户访问时的停留时间、访问次数、保存、编辑、修改等。通过分析发现,主要能揭示用户兴趣的浏览行为两种:浏览时间、保存/加入收藏夹。用多元线性回归方法来描述用户兴趣度与这两种主要行为的关系,其回归方程为:

$$L = a * X + b * Y + c \quad (2)$$

其中 X 表示浏览时间,Y 表示保存/加入收藏夹,L 表示用户对网页的兴趣度,a,b,c 为一组常数(它们的值根据浏览行为对网页兴趣度的影响确定)。

把 web 内容挖掘和行为分析的兴趣度相结合,进而得到网页的兴趣度为:

$$W_i * (d) = W_i(d) + L_i \quad (3)$$

$W_i * (d)$ 是网页的新兴趣度,$W_i(d)$ 是结合前网页的兴趣度,L_i 为用户行为分析的到的该网页的兴趣度。

2 基于时间窗口的用户兴趣模型更新

用户兴趣模型更新(用户兴趣漂移)所处理的是当用户兴趣改变时,如何滤去不感兴趣的信息而加入用户新的感兴趣信息。现在有很多人提出不同方法处

理漂移问题，如：Grabtree[2] 和 Soltysiak[3] 提出时间窗口方法，认为用户只对最近访问的概念感兴趣；Maloof 和 Michalski[4] 采用一种遗忘函数来衰减样本；Koychev[5] 提出的渐进遗忘的方法。

本文采用自适应的时间窗口法处理兴趣漂移问题，把用户不感兴趣的信息漂移出时间窗，以适应用户兴趣随时间的变化。处理机制如下：如果发现兴趣漂移，则自动调整时间窗口使时间窗内数据预测精度达到最大值，否则增大时间窗口加入新的兴趣项。这里预测精度（Accuracy）作为一个衡量指标，用来感应兴趣发生漂移。预测数据来自时间窗内最近的兴趣（至少 30 个），通过显著性检验方法得到预测精度。当用户兴趣发生变化时，时间窗中包含了许多用户不感兴趣的信息，这样会导致预测精度降低。当检测到预测精度显著降低时，说明发生了兴趣漂移，则通过缩小时间窗把不感兴趣的信息漂移出去；当预测精度提高时，则增大时间窗以加入新的兴趣项，时间窗内的兴趣特征项遵守先进先出的原则。

当发生兴趣漂移时，采用黄金分割法调整时间窗口的大小。假设预测精度在窗口大小[a,b]内是服从正态分布的，也就是说只存在一个窗口大小 L*，使预测精度最大。F(L) 为预测精度，取黄金分割值 T=0.618，在[a,b]上取 l=b−T*(b−a)；r=a+T*(b−a) 两点，当 f(l)>f(r) 时，在[a,r]上取两个点 l_1 和 r_1，且 b 的值用 r 的值代替。否则在[l,b]上取 l_1 和 r_1，a 的值用 l 的值代替，循环取值 T*(b−a) 后，得到的值 L* 即时间窗口的大小。

算法如下：

```
l=a;r=b;
for 1 to n do
   l=b−T*(b−a);
   r=a+T*(b−a);
   If f(l)>f(r) then b=r;
   Else if f(l)<f(r) then a=l;
     Else L*=a;
```

3 实验方法与分析

试验采用的数据是本校数字图书馆 Web 访问日志里面所记录的用户 3 个月的访问数据，包括 3000 张页面，首先，对这些数据进行预处理，除去辅助页面和噪声页面，得到用户感兴趣的页面 2458 张，同时用 Java 编制一组代理模块，能够自动捕获用户浏览行为并记录下行为数据。使用信息检索领域广泛使用的查准率和召回率作为评价推荐效果的度量标准。实验结果得到召回率的平均值为 76.7%，准确率的平均值为 72.1%。考虑到一些无法克服的人为因素，召回率和准确率还是比较理想的。并且，从实验得出，用于计算行为参数的网页数越多，其召回率和准确率就越高。可断定所浏览的网页越多，利用多元线形回归计算的用户行为参数就越准确，恰好符合了统计学原理。实验表明，本文提出的用户兴趣模型能较好的自适应用户兴趣的变化，且最终能较好地满足用户的信息需求。

4 总结和后续工作

本文提出新的用户兴趣模型的建立和更新方法，该方法通过 web 内容挖掘和行为分析建立用户兴趣模型，通过移动时间窗口更新兴趣模型，这样得到的用户兴趣模型通过试验证明更能够适应用户兴趣的变化。对于用户兴趣模型，将来这方面的工作还有很多，如设置用户行为权重回归方程参数的最佳值；建立混合兴趣模型等。

参考文献

1 M J Martin－Bautista. et al. User profiles and fuzzy logic for web retrieval issues. Soft Computing, 6 (2003):365～372.

2 Grabtree I, Soltysiak S. Identifying and Tracking Changing Interests[J]. International Journal of Digital Libraries, 1998,2(1):38～53.

3 Widmer G, Kubat M. Learning in the Presence of Concept Drift and Hidden Contexts [J]. Machine Learning, 1996,23(1):69～101.

4 Maloof M, Michalski S. Learning Evolving Concepts Using a Partial Memory Approach [C]. Working Notes of the AAAI Fall Symposium on Active Learning, Boston, MA, 1995,11: 10～12.

5 Koychev I. Gradual Forgetting for Adaptation to Concept Drift[C]. Proceedings of ECAI 2000 Workshop, 2000:101.