基于 WWW 缓存的用户长期兴趣发现

Discovering of User Long – term Interest Based on WWW Cache

索红光　杨 涛　（中国石油大学(华东) 计算机与通信工程学院 257061）

摘要：建立用户兴趣模型是实现个性化服务的关键技术之一。利用 Web 挖掘的方法，针对用户的兴趣变化，结合用户浏览 Web 页面的日期和相应 Web 页面特征项的词频，来建立用户长期和短期兴趣，并且通过模拟实验，验证该方法的有效性。

关键词：个性化服务 Web 挖掘 用户兴趣模型

1　引言

随着 Internet 的飞速发展，互联网上的信息资源也在不断地增加。同时，Web 用户数量也在迅猛的增长。因此，Web 已经成为用户获取信息的一个重要途径。面对如此丰富的“信息海洋”，用户如何快速、有效的获取所需信息，就成为一个急需解决的问题。同时由于用户的知识背景、兴趣取向等因素的差异，造成用户获取的信息也是有区别的。但是目前的 Web 系统所提供的信息和服务绝大多数还没有考虑到用户的个性化问题，因此如何针对用户的个性，并向用户提供个性化服务[1]已经成为 Web 技术的研究一个课题。

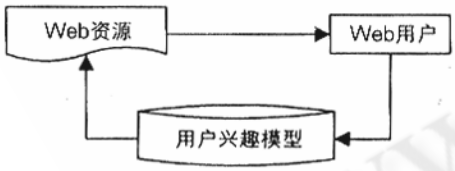

图 1　Web 个性化的过程

从图 1 所示的 Web 个性化的过程中，可以发现建立用户兴趣模型[2][3][4]是实现 Web 个性化的关键，因为只有准确地描述了用户的实际需求，才能根据用户的兴趣来提供个性化服务。在建立用户兴趣模型的过程中，处理和分析用户的各种信息，就要用到 Web 挖掘技术，所以说 Web 挖掘是实现用户建模的基本技术。

2　Web 挖掘技术

由于数据挖掘的绝大部分工作涉及的是结构化数据库，很少有处理 Web 上的异质、非结构化的信息的工作，所以在 Web 上解决这些问题的一个途径就是把传统的数据挖掘技术和 Web 结合起来，进行 Web 挖掘。Web 挖掘是从 Web 文档[4]和 Web 活动中抽取用户感兴趣的、潜在的有用模式和隐藏的信息。

Web 挖掘的一般过程可以分成 3 个阶段：

(1) 预处理，需要对收集的数据进行必要的预处理，例如清洗数据；

(2) 模式发现，应用不同的 Web 挖掘算法发现用户访问模式；

(3) 模式分析，从模式发现的模式集合中选择有意义的模式。

Web 挖掘通常可以分成 3 类，即：Web 结构挖掘、Web 内容挖掘和 Web 使用挖掘。

3　用户的长期兴趣与短期兴趣建模

3.1　数据预处理

利用程序抽取出用户客户端浏览器缓存中的 Web 页面，主要是 html 文件以及对应的用户浏览 Web 页面的日期。对 html 文件进行解析过程中，除去 frame、Script、图片等非文本内容，得到用户浏览的 Web 页面的文本内容。

3.2　Web 文本的特征表示

3.2.1　提取特征词

对 Web 预处理得到的文本进行分词处理，并且根

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn

据停用词表去除停用词，合并数字和人名等词条，得到能反映文本内容的特征词。

然后根据分词处理得到的特征词，计算其在文本中出现的频率即词频。词频分为绝对词频和相对词频，绝对词频使用词在文本中出现的频率表示文本，相对词频为归一化的词频，课题中采用了一种比较常见的 TF－IDF 公式计算：

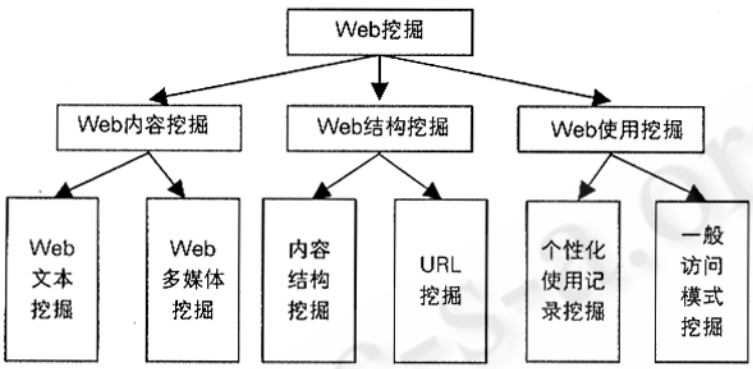

图 2　Web 挖掘的分类

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

其中，W(t,d) 为词 t 在网页 d 中的词频，而 tf(t,d) 为词 t 在网页 d 中的绝对词频，N 为训练网页总数，nt 为训练集中出现 t 的网页数，分母为归一化因子。

另外，在计算特征词词频的时候，考虑到不同文档的关键词数是不一样的，所以我们还需要考虑文档的关键词数这个因素，对于得到的 W(t,d) 再作一步处理。

$$W'(t,d)=W(t,d)/M \quad (2)$$

其中 M 表示文档 d 中关键词的总个数。

这样，Web 页面的文档可以用特征词向量（<T1,W1'>,<T2,W2'>,…,<Tn,Wn'>）来表示了。

3.2.2　通过概念映射获得特征概念向量

具体做法是：查询概念词典 HowNet 可以获得特征词对应的概念，完成概念映射。由于文档中一些特征词存在多重语义，可能对应多个概念；此外还存在概念词典中未标注的新词，也具有较强的提示作用，这两种情况需要特殊处理。

（1）对于一些未在概念词典中标注的新词，通常的做法是直接保留其作为特征概念，加入概念向量。

（2）对于多重语义的特征词可以计算它们之间的共现频率。根据事先设置好的两个概念之间的关系（主要有上下位，部分和整体等）的词频来计算两者的共现频率，选择隶属度大的那个特征词对应的概念作为最后的概念。

3.3　对于文档进行聚类分析

Web 文档聚类是一种无监督的文档分类，它的目标是将文档集分成若干类，要求同一类内文档内容的相似度尽可能大，而不同类之间的相似度尽可能地小。目前文档聚类的方法很多，包括层次聚类法、平面划分法、简单贝叶斯聚类算法等。但是各种算法都有它的优缺点，平面划分法虽然有较理想聚类结果，但是它必须先确定聚类参数 K，而且聚类结果要受到聚类中心初始值的影响。

本文将层次聚类法和平面划分法（K－means 算法）相结合，首先利用层次凝聚法进行初始聚类确定初始聚类中心和 k 值，然后用 K－means 算法进行聚类分析。

对于给定的文档集合 D＝{d1,…,di,…,dn}，整个文档聚类过程的算法为：

（1）用层次凝聚法进行初始聚类；

（2）将层次凝聚法得到的初始聚类中心作为 K－means 聚类的种子；

（3）依次计算 D 中的每个文档与每个种子的相似度；

（4）选取具有最大相似度的种子 MAX{sim(di,Sj)}，将 di 归入以 Sj 为聚类中心的类 Cj；

（5）重复步骤（2）－（4），最后得到较稳定的聚类结果 C＝{C1,…,Ck}。

其中计算特征向量之间的相似性采用夹角余弦相似度函数。

根据聚类的结果，采用基于概念的向量模型来表示 Web 文本，这样能降低 VSM 的维数，便于以后发现用户的兴趣，并且把一些同类的特征词都用一个概念来表示，消除了同义词和近义词对建模的影响。

3.4　用户的长期兴趣和短期兴趣

随着时间和环境的改变，Web 用户的兴趣也会有所变化，用户往往会因为一个短期的动机而去捕获相关信息，我们可以称之为用户的短期兴趣，例如用户要

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn

购买一个 DV，他可能在购买之前的一段短期的时间段内，会浏览感兴趣的 DV 促销或者采购信息，而在购买完之后，可能关注这方面的信息就少了；相对应的，就是称为用户的长期兴趣，比如用户对影视感兴趣，那么他会在网上长期关注这方面的信息，不会因为时间的改变而有很大的变化。所以，为了能更好的为用户提供个性化的服务，我们应该捕获用户的短期兴趣的变化，还有区分长期和短期的兴趣。

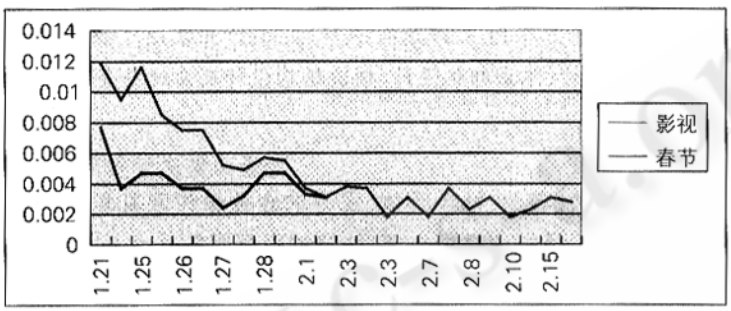

图 3 短期兴趣与长期兴趣对比图

通过分析可以发现，在区分用户兴趣变化时，可以加入用户浏览 Web 页面的时间段这个因素。初步考虑是统计某个词在一周的时间段内出现的频率。利用下面的公式来表示。

$$L = a * X_1 + b * X_2 + c \quad (3)$$

其中 L 表示特征词的最后权值，X_1 表示某个词在一周的时间段内出现的频率，X2 表示在特征词在文档中出现的频率，a、b、c 是一组常数。

4 实验验证

我们抽取出用户客户端浏览器缓存中的 html 网页数据，一共采集到 35 天的数据。对于这些数据，用上述方法进行处理，得到 35 天内用户每天浏览的网页对应的文本信息，并且建立 VSM 向量模型，表示各个文本的特征，删除词频低于某个域值（事先设定好）的关键词，对于剩下的关键词进行分析，用公式（3）计算出每个主题关键词对应的最终权重。并且根据用户浏览该网页日期和该网页对应文本特征项的词频划出图形，形式如图 3 所示。

例如“影视”和“春节”两个主题，它们都是从用户浏览的 Web 页面中抽取出来的，并且能代表浏览 Web 页面的特征内容。在 1.21 到 2.15 的时间段之内，它们的词频变化是不一样的。因为用户在 1.21 到 2.1 之间浏览了关于“春节”的信息，所以它的词频在这个时间段内是比较稳定的，不在这个时间段内，用户就没有浏览“春节”的相关信息了，所以只能说明用户在这段时间内对该主题比较感兴趣，可以称之为短期兴趣；而“影视”则在整个实验时间段内都出现过，并且词频变化不是很大，所以可以确定是用户的长期兴趣。

5 结束语

建立用户兴趣模型是实现个性化服务的关键部分。由于 Web 用户的兴趣经常发生变化，因此基于用户的兴趣进行个性化推荐就应该能捕获到这种变化并相应地为用户推荐出感兴趣的信息，同时也能把短期兴趣与长期兴趣区别开来，这样向用户推荐的内容就有所区别，并且是更有针对性的，对于用户短期感兴趣的内容，只在短期的时间段内向用户推荐就可以了。根据用户浏览 Web 页面的时间和表示页面特征项的词频等信息，来区分用户的长期兴趣和短期兴趣，这样能较好地反映了用户的兴趣变化，为用户推荐的信息也更加准确。

参考文献

1 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.
2 林鸿飞、杨元生，用户兴趣模型的表示和更新机制，计算机研究与发展，2002，39 (7)：843～847.
3 M J Martin－Bautista, et al. User profiles and fuzzy logic for web retrieval issues. Soft Computing 6 (2003):365－372.
4 Kim, H. R., Chan, P. K. (2003): Learning Implicit User Interest Hierarchy for Context in Personalization.
5 Sugiyama K, Hatano K, YoshikawaM. (2004): Adaptive Web Search Based on User Profile Construction without Any Effort from Users.

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn