基于模糊隶属度的个性化网页推荐系统①

Personal Webpage – Recommender System Based on The Fuzzy Degree of Membership

张培颖　（中国石油大学(华东) 计算机与通信工程学院 山东 东营 257061）

摘　要： 个性化信息服务越来越成为信息检索领域中研究的热点。针对用户模型的构造问题，文章利用用户浏览过的网页历史记录自动进行文本结构分析，获取网页信息的逻辑表示，将段落作为识别用户兴趣的基本要素，利用段落间的聚类分析和对用户兴趣的表达能力，获取最终的用户兴趣特征向量。提出了一种基于主题描述的二级层次用户模型，并给出了用户模型的动态调整算法，构建了一个基于模糊隶属度的个性化网页推荐系统。模拟实验表明，该用户模型和个性化推荐算法能够有效地提高检索结果的准确性，并且具有良好的适应性。

关键词： 模糊隶属度 用户模型 主题特征向量 聚类分析

Web 已成为人们获取信息的一个重要途径，由于 Web 信息的日益增长，人们不得不花费大量的时间去搜索自己想要的信息。搜索引擎是最普遍的辅助人们检索信息的工具，比如 Google、百度等。信息检索技术满足了人们一定程度上的需求，但由于其通用的特性，仍不能满足不同背景、不同目的和不同时期的查询请求。个性化搜索引擎正是为了解决这一问题而提出的，它为不同用户提供不同的服务，以满足不同的需求。个性化搜索引擎通过收集和分析用户信息来学习用户的兴趣和行为，从而实现主动推荐的目的。

个性化搜索引擎的设计和实现需要解决的一个基本问题是：用户知识的搜集、分析、处理和表示。形象地说，就是必须为每个用户提供一个过滤器，根据这个过滤器对检索得来的结果进行过滤，或者对用户的检索需求进行过滤。个性化搜索引擎是指根据用户的兴趣和特点进行检索，返回与用户需求相关的检索结果。能否成功地提供个性化服务的关键，在于个性化搜索引擎中用户模型的建立和更新机制。文章首先介绍了个性化搜索引擎的一般模型，然后提出了基于主题描述的二级层次用户模型，给出了基于模糊隶属度的个性化检索算法，最后通过实验验证了该模型的有效性。

1　个性化搜索引擎的检索模型

传统的搜索引擎系统一般包括 5 个基本部分：搜集器、分析器、索引器、检索器和用户接口。搜集器负责对 Web 进行搜索并下载文档；分析器负责对下载的文档进行分析以用于索引；索引器负责将文档表示为便于检索的方式并存储在索引数据库中；检索器负责从索引中找出与用户查询请求相关的文档；用户接口为用户提供可视化的查询输入和结果输出界面。

个性化搜索引擎一般在传统的搜索引擎系统的基础上增加了用户模型、提问调整、源选择、结果处理等“个性化”处理模块。如图 1 所示：

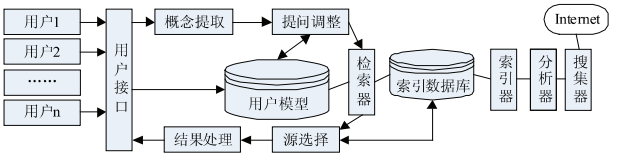

图 1　个性化搜索引擎的检索模型

① 基金项目：中国石油大学(华东)计算机与通信工程学院青年教师创新基金

2　基于主题描述的二级层次用户模型

2.1 用户模型的表示

为了准确地描述用户的兴趣，一般情况下需要为每个用户建立一个用户模型(User Profile)。不同的个性化搜索引擎中用户模型的表示各有其特点，用户模型可以划分为基于兴趣的和基于行为的两种类型。基于兴趣的用户模型可以表示为加权矢量模型、类型层次结构模型、加权语义网模型、书签和目录结构等；基于行为的用户模型可以表示为用户浏览模式或访问模式。在具体实现时可以综合基于兴趣和基于行为这两种表达方式。

这里采用基于主题描述的二级层次用户模型，第一级是用户感兴趣的主题特征向量及其权重，第二级是每个兴趣类型下的特征向量及其权重。每个主题 Ti 下的特征向量采用向量空间模型来表示，关键词最多可包括 5 个，如图 2 所示，为一个用户的兴趣模型。

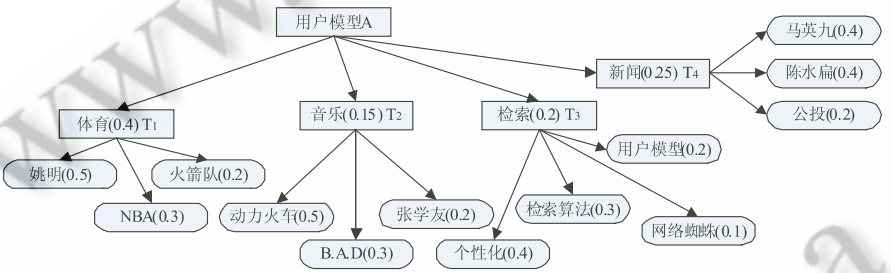

图 2　基于主题描述的二级层次用户模型

基于主题描述的二级层次用户模型能够准确地、完善地描述用户的兴趣所在。根据用户模型的主题词节点，就可以大体知道用户的兴趣类型，及其对每一兴趣类型的兴趣度高低。每个主题词下面的特征词向量节点可以准确地表示用户对该兴趣类型中相关内容的感兴趣程度。第二级节点的特征向量采用向量空间模型表示，这样就把用户兴趣的表示和页面文档的表示统一了起来。

2.2 用户模型的信息来源

目前，获得用户模型的信息有显示收集和隐式收集两种方法。

显示收集：采用显示收集方式时，要求用户指出自己的浏览兴趣，比如提交感兴趣的关键词、主题，喜欢的网页，对网页的反馈信息等。这是一种直接的信息收集方式，直接来自于用户，需要用户的参与。

隐式收集：采用隐式收集方式时，系统通过分析用户的上网数据，如浏览的网页、页面的点击情况、在网页上停留的时间等，获得用户的浏览兴趣。这是一种间接的收集方式，它不是从用户的反馈信息直接获得，而是从用户的上网数据中分析得到的。

这里采用隐式收集的方式，通过对用户最近浏览过的网页信息进行文本结构分析，获取网页信息的逻辑表示，将段落作为识别用户兴趣的基本要素。利用段落间的聚类分析和对于用户兴趣的表达能力，获取最终的用户兴趣特征向量。

2.3 用户模型的更新机制

用户模型的更新主要是用户感兴趣的主题词的更新，以及每个主题词下面的用户兴趣节点的更新。对于 Web 站点的每一个用户，他对该站点的兴趣一般集中于某一个主题或几个主题，这可以由聚类分析得到。因为通过聚类，用户浏览的历史页面自动地分成了若干个聚簇(主题)，每一聚簇的页面体现了用户感兴趣的某一主题。每个主题下用户兴趣节点的更新，可以从用户的浏览行为中体现出来。例如：某一用户的浏览记录为：{信息检索(5 张)，中文分词(3 张)，个性化(8 张)}，那么就可以初步地判断该用户的兴趣度排序为：个性化→信息检索→中文分词；又如：某用户的第一次浏览顺序为：{A→B→C}，那么可以简单地判断其兴趣度排序为：A→B→C，但是隔了一段时间，他的第二次浏览顺序变为{A→C→B}，那么可能他的兴趣度排序发生了变化，C 类的兴趣度比 B 类的要大。

通过对用户浏览过的历史网页按照段落进行聚类分析，找出用户感兴趣的主题词，计算其权重。通过对某一主题词下的页面内容信息和浏览行为信息进行分析，就可以定量地计算出用户在某一类主题词下用户兴趣节点的兴趣度(权重)。有了兴趣度的高低排序，在基于该兴趣模型对用户进行推荐服务时，系统才知道应该先向用户推荐什么，后推荐什么，提高系统的个性化服务效率。

3　基于模糊隶属度的个性化推荐算法

本文提出的基于主题描述的二级层次用户模型，旨在提高搜索引擎的个性化服务质量。其算法的流程描述如下：

（1）对用户浏览过的网页历史记录进行结构化分析，将段落作为识别用户兴趣的基本要素。把用户浏览过的网页表示为一个段落的集合 Pset；

（2）根据段落间的聚类分析，形成表达用户兴趣的若干个主题段落集合 Tset；

（3）分别对每一个主题段落集合进行特征向量的提取，计算出每个段落集合的特征向量表示；

（4）在每个段落集合中选择前 k 个权重较大的特征向量作为用户的主题特征向量，权重的计算按照 TF. IDF 公式重新计算；

（5）在每个段落集合中选择前 m 个权重较大的特征向量作为用户的兴趣节点向量，权重的计算按照 TF. IDF 公式在每个段落集合内部重新计算；

（6）这样就可以构建出用户的兴趣模型，在进行个性化推荐的时候，首先，计算网页和主题特征向量的模糊隶属度 FUi，然后再计算该主题特征向量下的兴趣节点和该网页之间的距离 DISTi，最后的权重计算公式如下：

$$Weight(webpage)=\sum_{i=1}^{n} FU_i \times DIST_i$$

其中 n 为主题特征向量的个数；FU_i 表示网页和第 i 个主题特征向量的模糊隶属度；$DIST_i$ 表示网页和第 i 个主题特征下的兴趣节点之间的距离。

（7）根据上面的公式计算各个网页的权重，按照权重的大小顺序推荐给用户。

4　实验结果及分析

根据上面的个性化推荐算法构造了一个基于客户端的 webpage – recommender 系统。在其中采用了用户内容浏览和用户行为浏览相结合的方法来获得用户的兴趣度，对用户浏览过的网页历史记录自动进行文本结构分析，构建基于主题描述的二级层次用户模型。在进行个性化检索的时候，采用基于模糊隶属度的个性化推荐算法。

我们从新浪网下载了体育类、新闻类、娱乐类、学术类网页各 100 篇作为用户搜索的源数据库。参加实验人员为计算机专业一年级两个班学生共 73 人。实验要求在 2 个月内，按照自己的兴趣浏览网站内容，浏览时间共计不得少于 60 小时。经过两个月时间的实验，经统计共有 60 人的浏览时间超过 60 小时。然后我们重新下载上面的四类网页各 100 篇，对这 60 个人进行实验测试。测试分 2 步进行。第 1 步让参加测试的 60 位同学按自己的意图进行 15 分钟的浏览，然后对系统推荐出的前 5 篇文档按符合本人意图程度以 100 分制进行打分；第 2 步是经过 30 分钟的浏览后再进行同样的测试。最后统计所得分值。第 1 步平均得分 73. 8 分，第 2 步平均得分 86. 7 分。可见系统实际测试结果还是比较满意的，说明我们的用户模型能够比较准确地描述用户的兴趣，基于模糊隶属度的个性化推荐算法切实可行。

5　结语

本文构建的个性化网页推荐系统 webpage – recommender，在对用户浏览过的网页历史记录进行分析的时候，将段落作为识别用户兴趣的基本要素，通过聚类分析，计算出用户感兴趣的若干个主题段落集合。根据 TF. IDF 公式构建基于主题描述的二级层次用户模型，在个性化检索阶段使用基于模糊隶属度的个性化推荐算法进行网页推荐。实验表明，该用户兴趣模型和个性化推荐算法能够有效地提高检索结果的准确性，并且具有良好的适应性。

参考文献

1 秦春秀，赵捧未. 基于用户兴趣的个性化检索. 情报学报，2005，24(4)：449 – 452.

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 宋擒豹，沈钧毅. 基于关联规则的 Web 文档聚类算法. 软件学报，2002，13(3)：417 – 422.

4 陈汉深，李卫忠. 基于 C/S 的新一代智能化、个性化搜索引擎. 情报学报，2006，25(1)：70 – 73.

5 杨武剑，王泽兵，冯雁，武新玲. 网站个性化服务的研究. 浙江大学学报，2003，37(3)：278 – 282.

6 胡健，陆一鸣，马范援. 基于 HTML 文档结构的向量空间模型的改进. 情报学报，2005，24(4)：433 – 437.