基于 Web 内容和日志挖掘的个性化网页推荐系统

Personalized Web Page Recommendation System Based on Web Content and Log Mining

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： 目前的网页推荐服务大都是基于对查询关键词的匹配来实现的,缺乏自动提取用户兴趣并进行推荐的能力。本文设计和实现了一个基于 Web 内容和日志挖掘的个性化网页推荐系统 Webpage – recommender,该系统采用特征提取技术,能自动识别用户的兴趣,并主动推荐和筛选与用户兴趣主题相关的网页。实验结果证明了该系统的有效性。

关键词： 网页推荐 数据挖掘 文本分类 特征选择 关联规则

随着 Internet 的飞速发展,Web 已经成为当今最大的信息源,在网页浏览中,如何根据用户的浏览内容和路径主动寻找和发掘用户感兴趣的网页,改变用户被动接收信息的单一模式,为用户提供个性化的推荐服务是一个十分有意义的问题[1]**。但是,现有的网页推荐服务大都基于用户搜索的关键词进行匹配,普遍存在以下 3 点不足:(1) 推荐范围受用户指定的关键词的限制。推荐引擎仅仅关注关键词出现率高的网页,这限制了搜索范围,会错过很多用户感兴趣的网页;(2) 关键词的概括需要一定经验的积累,这对普通的网络用户提出了一个比较高的要求;(3) 推荐结果往往忽视了对用户浏览路径的参考,导致推荐的内容与兴趣主题有差距。为解决以上不足之处,笔者设计并实现了一个基于 Web 内容和日志挖掘的个性化网页推荐系统 Webpage – recommender。Webpage – recommender 提供主动式服务,用户无须自己总结感兴趣的关键词,无须通过关键词搜索,系统利用特征提取技术实现对网页内容的自动挖掘,通过关联规则的过滤向用户推荐与其兴趣一致的候选结果。**

1　体系结构

Webpage – recommender 系统对网页的推荐基于以下 2 点假设:(1) 用户的兴趣具有稳定性,网页内容与用户兴趣相关。(2) 具有相同兴趣的网站访问者的浏览方式具有规律相似性,通过分析大量用户的浏览模式,可以发掘其规律。

基于以上假设,可以通过提取用户阅读的网页内容,利用特征提取技术识别出网页的主题,从而挖掘出用户的兴趣,通过关联规则挖掘筛选出用户最有可能感兴趣的网页。Webpage – recommender 系统采用了客户/服务器的体系结构(见图 1)。

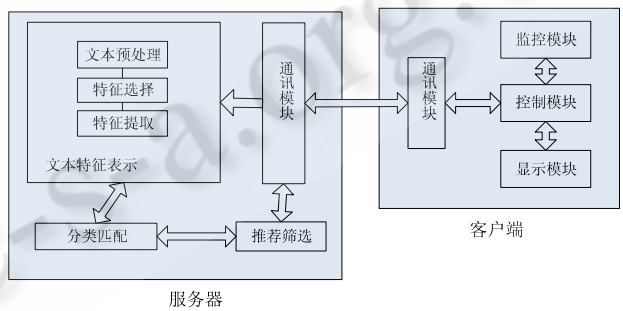

图 1　Webpage – recommender 的体系结构

服务器端由 4 个模块组成:(1) 文本特征表示模块:负责文本预处理、特征选择、特征提取;(2) 分类匹配模块:负责维护分类特征向量和候选网页信息;(3) 推荐筛选模块:对候选网页进行进一步的过滤;(4) 通讯模块:负责与客户端交互,数据的收发。客户端包括 4 个模块:(1) 监视模块作为推荐任务的事件触发者,负责捕捉用户浏览网页的事件,及时发送网页地址给控制模块;(2) 通讯模块负责与服务器进行交互,发送浏览页面的地址,接收推荐结果;(3) 显示模块负责动

态展示推荐结果；(4) 控制模块作为中枢，负责数据流和业务流的调度。

2 关键模块

2.1 文本特征表示

利用特征提取技术来实现对网页主题信息的抽取，由文本预处理、特征选择、特征提取分类模块 3 个部分组成。

图 2 给出了文本预处理的流程。文本预处理将不同的网页格式数据转换为主题分类可以处理的单词序列。本系统的语料包括 HTML 网页和 XML 格式的网页文件。英文单词存在不同时态和复数等变体形式，需要进行单词还原。而中文以字为单位，要进行专门的分词处理。在得到单词集后，还要按照停词表去掉对区分主题无帮助的虚词，如："的"，"the"，"a"等。同时单词可以出现在页面的标题、副标题、正文、超级链接等位置，不同的位置也反映了不一样的重要性，因此在预处理的同时记录单词位置信息。

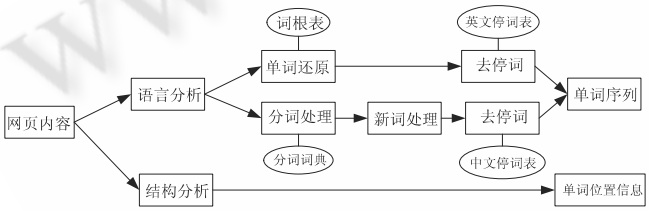

图 2　文本预处理流程

此模块要解决的关键问题是中文分词。常见的分词方法有 3 种：基于字典的字符串匹配法，基于统计的分词，基于句法和自然语言理解的方法[2,3,4,5]。本系统结合了字符串匹配方法和概率统计法进行分词。字符串匹配虽然可以得到大多数词汇，但现在的更新速度远远高于词典的网络语言，"超女"、"快男"、"博客"等网络新词层出不穷。通过统计方法将一些高频出现的未知词汇作为新词加入词表。在汉语的语言习惯中，二字词语占绝大部分，因此，可以将发现二字词作为新词发现的重点。利用下面公式计算任意 2 个汉字 x 和 y 之间的互信息 I(x,y)，将 I 值大于指定阈值的 xy 作为新词添加到词表中。

$$I(x,y)=\lg\frac{p(xy)}{p(x)p(y)}$$

其中，p(x) 和 p(y) 分别是汉字 x 和 y 在汉语中单独出现的概率，而 p(xy) 是 x 和 y 相邻出现的概率。在实际的计算过程中，p(xy)，p(x)，p(y) 是通过语料集的统计数据来估计的。假设以 f(xy)，f(x)，f(y) 分别表示二元组和相应的单字在语料集中出现的频率，而以 N 表示语料集的规模，即语料集中单字的个数，那么 p(x) 和 p(y) 分别为 f(x)/N，f(y)/N，而 p(xy) = f(xy)/N。

特征选择用于从大量已分类的训练文本中抽取出最能代表各自类别特征的一些词汇，组成特征向量。特征选择的优劣直接影响着分类的准确度。选出的特征词的数量，即特征向量维度也决定了分类匹配模块的计算量。本系统采用基于信息收益 (Information Gain，IG) 主成分分析的特征选择方法，目的是用尽量少的词汇尽可能全面体现分类主题的语义。这样可以尽量压缩特征向量的维数，减少匹配计算时间，同时保持较高的匹配精度。首先用信息收益法计算词汇的信息熵。该方法采用信息论中信息相对熵的概念，依据不同词汇对分类系统提供的信息增量的不同进行筛选。如果简单地选择信息收益大于阈值的词汇作为特征，往往忽略了词汇的语义信息，存在同义词，关联词等干扰噪声。在计算词汇信息收益的基础上，再运用主成分分析，选择不相关的的词汇作为特征。

文本表示依据特征选择选出的特征，对文本进行向量化表示。根据向量空间模型，每个文档都可以表示成一个加权向量形式 V(w1，w2，w3，…，wi，…，wn)，n 是特征向量维度，可以采用 TF/IDF 方法赋予每个特征词不同的权重。权值 WTF/IDF 计算方法：TF × IDF，其中 TF(Term Frequency) 表示特征词汇在该文本中出现的次数，IDF(Inverse Document Frequency) 为逆文本频率值，计算方法：IDFi = lg(D/ Di)，D 表示训练文档总数；Di 表示出现第 i 个特征词的文档数。由于网页属于半结构化数据，针对词汇在网页中出现的位置不同，对不同的位置定义相应的权重 Wpos，所以总权重 $Wi=\alpha WTF/IDF+(1-\alpha)Wpos$，$\alpha\in[0,1]$，这样可以综合考虑特征的语义和位置信息。

2.2 分类匹配

分类匹配的任务是将待分析的网页经过分类算法的计算后，判定其归属于哪一主题。在系统中，用分类算法计算待分类文本与各主题的相似度，相似度较高

的类将被认定为分类结果,进一步得到相似度高的候选网页。目前已有多种文本分类算法:中心向量法、k邻近算法、支持向量机、简单贝叶斯等。中心向量法是根据算术平均为每类文本生成一个代表该类的中心向量,计算待分类文本与每类中心向量间的欧式距离,以距离最近的类作为待分类文本的类别。该方法分类速度快,但是,以向量空间距离作为分类标准将形成类球状类别分布,对于与多个类距离相近的文本,该算法的分类准确度将急剧下降。由于系统对分类速度要求较高,为了不影响用户的使用体验,采用了速度较快的中心向量法,同时为了弥补准确度下降的缺陷,在推荐筛选模块中利用关联规则对候选结果进行过滤。

2.3 推荐筛选

基于具有相同兴趣的网站访问者的浏览方式具有规律相似性的假设,在推荐筛选模块,对网站的日志文件进行挖掘,发现关联规则,利用这些规则和用户当前的浏览路径对候选结果进行筛选,进一步提高推荐的准确性。筛选流程如图 3 所示:

图 3 筛选流程

在 Web 日志中按时间顺序记录了不同用户对站点的访问数据,通过数据预处理可以将日志数据整理成用户会话的形式,即在规定的超时时间限制内,用户对站点连续、完整访问的序列形式。预处理过程一般包括 4 个步骤:数据清理、用户识别、会话识别、路径补偿。考虑到用户在页面停留时间长短同样反映了用户的兴趣,在本系统中还应根据各页面访问时间,计算出页面的停留时间。因此最后用户会话整理成为以访问页面和停留时间组成的二元组为元素的序列。通过改进的 Apriori 算法[6,7,8],挖掘具有一定可信度和支持度的正逆关联规则,以此对照用户当前的浏览路径对候选列表中符合逆关联规则的结果过滤掉,对符合正关联规则的结果提高排序。

3 Webpage – recommender 系统运行

测试运行中将 Webpage – recommender 系统部署于某网站的新闻频道,网页训练集分别来自 BBC(英文)和新浪网(中文),共采用不同主题的英文文档 4312 篇,中文文档 6410 篇。中文文档通过分词处理得到不同词条 16713 条。采用信息增益算法取得英文特征词 4505 个,中文特征词 6789 个。通过对最近 34M 的 Web 日志进行清理获得用户会话 415132 条。经关联规则挖掘,在可信度为 0.8 支持度为 0.7 发掘出 4731 条正逆关联规则。此关联规则集每月动态更新一次。当用户当前窗口为网站网页时,Webpage – recommender 的客户程序会自动捕获此阅读网页事件,并将此时的网页链接发送到服务器端;服务器进行用户兴趣挖掘,形成网页推荐信息回送客户端,客户端将树型推荐信息以半透明方式浮现在浏览器窗口右侧:以当前网页为中心,展开多个兴趣分支,每个兴趣分支延伸出多个推荐网页链接。

4 结束语

本文介绍了基于 Web 日志和内容挖掘的 Webpage – recommender 系统的设计和实现,重点分析了关键组件——特征表示组件和推荐过滤组件的结构及其关键技术。本文采用特征提取技术对用户感兴趣的网页进行了推荐,较关键词匹配方式,达到了模糊识别主题的效果;利用 Web 日志挖掘获得的正逆关联规则对候选推荐列表进行筛选,提高了符合正向关联规则结果的排名,进一步提高了推荐的精度。Webpage – recommender 采用多主题推荐方式,在内容的可读性和丰富程度上考虑了个性化,具有一定的实用值和研究价值。

参考文献

1 Kazienko P, Kiewra M. Personalized Recommendation of Web Pages. In: Nguyen T, ed. Intelligent Technologies for Inconsistent Knowledge Processing. Advanced Knowledge International, Adelaide, South Australia, 2004, 163 – 183.

2 代六玲, 黄河燕, 陈肇雄. 中文文本分类中特征抽取方法的比较研究. 中文信息学报, 2004, 18(1): 26 – 32.

3 单松巍, 冯是聪, 李晓明. 几种典型特征选取方法在中文网页分类上的效果. 计算机工程与应用, 2003, 39(22): 146 – 148. (下转第 112 页)

```
    r. Bottom := trunc( w * Height/Width) ; //高度按比例调整
    end;
    //设定描绘区域大小
    img. Picture. Bitmap. Width := r. Right;
    img. Picture. Bitmap. Height := r. Bottom ;
    //可伸展方式描绘,事实上是按比例缩小或放大
    img. Canvas. StretchDraw( r,TGraphic( sbmp) ) ;
end;
```

图 4 所示是标本采集前的检验医嘱处理程序主界面,其右侧所列的是与执行检验申请有关的标本采集信息提示,包括标本容器的实物图像和有关说明文字。

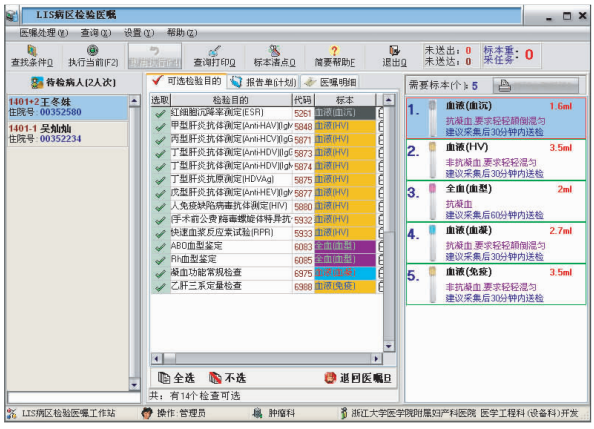

图 5　用户界面

4　结束语

可视化的标本采集提示,以醒目的彩色图文界面突出了临床检验标本采集环节的处理重点,增强了应用程序的亲和力,从而提高了软件的易用性,也方便了标本核对工作,有助于减少差错发生。显示图像均由素材图像动态处理生成,而不需要在系统中维护数量很多的不同容器图像,既可以减少存储空间和传输信息的量,也方便了相关维护工作。系统所需的素材图像是经过专业图像处理软件裁剪、背景和颜色调整的图像文件,虽然素材准备工作不够方便,但是,系统一旦启用后增加图像的情况并不经常需要。

参考文献

1　朱美芹. 分析前质量控制的影响因素与流程再造. 中国误诊学杂志,2007,7(18): 4271 – 4272.

2　田玉敏,梁若莹. 计算机彩色输入输出设备常用颜色空间及其转换. 计算机工程,2002, 28(9):192 – 200,274.

3　刘骏. Delphi 数字图像处理及高级应用. 科学出版社,2003.

(上接第 11 页)

4　Salton G. Automatic Text Processing. The Transformation, Analysis, and Retrieval of Information by Computer. Addison – Wesley, Reading, MA ,1989.

5　Mooney R J, Roy L. Content – based book recommending using learning for text categorization. In: 5th ACM conference on Digital Libraries, pp. ACM Press, New York,2000:195 – 204.

6　Chen Z, Fu A W – C, Tong F C – H. Optimal Algorithms for Finding User Access Sessionsfrom Very Large Web Logs. World Wide Web: Internet and Web Information Systems 6,2003. 259 – 279.

7　Alata B, Akin E. An efficient genetic algorithm for automatedmining of both positive and negative quantitative association rules. SoftComputing – A Fusion of Foundations, Methodologies and Applications 10(3),2005. 230 – 237.

8　Antonie M – L, Zaïane O R. Mining Positive and Negative Association Rules: An Approach for Confined Rules. In: Boulicaut J – F, Esposito F, Giannotti F, Pedreschi D,eds. PKDD 2004. LNCS (LNAI)3202, pp. Springer, Heidelberg ,2004. 27 – 38.