一个基于用户兴趣的 blog 推荐系统的设计

孙 多
(扬州大学 信息工程学院,江苏 扬州 225009)

摘 要:设计和实现了一个面向 blog 的兴趣挖掘和推荐系统 blog-digger,该系统采用兴趣挖掘技术,主要根据用户在一定时间段对 blog 页面的浏览行为,判断出用户对 blog 网页的感兴趣程度,并采用文本分类技术对用户的兴趣进行挖掘,取得较好的兴趣挖掘结果.另外,结合页面重要度对网页进行排序,以确定最终推荐给用户的 blog.实验表明该系统推荐的 blog 具有较高的主题内容相关.

关键词:blog 推荐;兴趣挖掘;相似度

中图分类号:TP 311.1 **文献标志码**:A **文章编号**:1007-824X(2011)01-0061-04

网络用户通过 blog 共享信息,发表个人观点,这打破了以往只能被动接收信息的单一模式,用户主动寻找和发掘自己感兴趣的 blog,从而促进了 blog 搜索服务的发展.近几年,研究者的兴趣主要从以下 3 个方面展开:内容[1]、结构[2]和使用[3].其中,基于链接结构的分析研究最广泛[4];基于内容的分析是文本分类的一个分支,属于 web 文本挖掘范畴.虽然一些服务商也提供了专门的 blog 搜索功能,但仍存在以下不足:① 搜索范围受到用户指定关键词的限制,搜索引擎仅关注关键词出现率高的网页,这限制了搜索范围,会遗漏很多 blog;② 关键词的概括需要一定的经验积累,这对普通网络用户提出了一个比较高的要求.为此,本文设计了一个基于用户兴趣的 blog 推荐系统 blog-digger,它可以提供主动服务.

1 Blog-digger 系统的体系结构

本系统的设计基于如下思想:首先,用户的兴趣与其发布 blog 内容相关,即系统假设用户在某一时间发表的博客文章的主题为其当时感兴趣的主题;其次,用户对于兴趣的遗忘遵循人类自然遗忘规律,即用户在每篇文章中体现出来的兴趣是随时间衰减的.系统的设计目标是为了从用户的博客发文中抽取出用户对应时间所感兴趣的话题,形成兴趣迁移曲线,然后根据兴趣迁移曲线并结合页面的重要程度推荐合适的 blog 页面.

本系统设计成一个 client/server 结构体系.服务器端包括以下 5 个组成部分:① 通信模块.此模块负责与客户端交互,下载 RSS(really simple syndication)文件.② 兴趣挖掘模块.此模块根据用户浏览网页行为挖掘出页面兴趣度.③ 网页处理模块.此模块负责文本预处理、特征选择、文本向量表示和分类.④ 网页相似度计算模块.此模块借助向量之间的距离来计算文本之间的相似程度.⑤ Blog 推荐模块.此模块综合考虑兴趣度、重要度和相似度,对 blog 网页进行排序,以确定最终向用户推荐的 blog.客户端包括 3 个模块:① 监视模块.此模块作为兴趣挖掘任务的触发者,负责识别用户、捕捉用户阅读 blog 的事件和用户会话识别,及时发送 blog 地址给控制模块.② 通信模块.此模块负责与服务器进行交互,发送 blog 地址,接收推荐结果.③ 显示模块.此模块负责动态展示推荐结果.

Blog-digger 系统实现如下功能:当用户正浏览 blog 网页时,blog-digger 的客户程序会自动捕

收稿日期:2010-03-30
基金项目:国家自然科学基金资助项目(61070240)
E-mail:sunduo@yzu.edu.cn

获此阅读 blog 事件,并进行用户识别和会话识别,将用户浏览过的 blog 网页链接发送到服务器端;服务器进行用户兴趣挖掘,形成 blog 推荐信息并回送客户端;客户端将树形推荐信息浮现在浏览器窗口右侧,根据用户的不同兴趣度,展开对应的兴趣分支,每个兴趣分支给出 4 个推荐 blog 链接.

2 系统部分功能

2.1 数据采集

RSS 作为描述 blog 主题和更新信息的最基本途径,其内部蕴涵了大量的博客信息,如发表时间、文章类别、文章内容等;因此,数据采集的目标定位为采集 blog 对应的 RSS 文件.

2.2 文章数据的处理

1) 前期处理. RSS 文件是博客作者最新发布的数篇文章的集合,处理时须将文章逐一抽取形成独立文件.为了使抽取后的文章满足分类器的要求,还须对文本文件进行去标签化、去噪声词以及全角半角转换等操作.去标签化是指将残留的 XML 标签或者文章文本内部的 HTML 标签去除,以免影响分类效果;去噪声词指的是去除那些出现频率很高但是对文本分类却没有太大作用的单词以及中文中的那些虚词;全角半角转换主要针对的是文章内部的数字、标点符号等.

2) 后期处理. 后期处理涉及中文分词、特征选择、特征加权 3 个步骤.对于中文数据集的处理,本文采用中国科学院计算技术研究所的汉语语法分词系统 ICTCLAS,对文本进行分词处理.空间向量模型的基本思想是以一个规范化的特征向量来表示文本:

$$\boldsymbol{V}(d)=\{\langle t_1,\ w_1(d)\rangle,\cdots,\langle t_i,\ w_i(d)\rangle,\cdots,\langle t_n,\ w_n(d)\rangle\},$$

其中 t_i 为特征词条项;$w_i(d)$为 t_i 在文本 d 中的权值,一般被定义为 t_i 在 d 中出现的频率 $tf_i(d)$的函数,即 $w_i(d)=\psi(tf_i(d))$.对于 ψ,采用 TF×IDF[5] 函数:

$$\psi=(\sqrt{tf_i(d)})\mathrm{lb}\ (N/n_i+0.01)/\sqrt{\sum_{i=1}^{n}tf_i(d)^2\mathrm{lb}\ (N/n_i+0.01)},$$

其中 N 为总文档数;n_i 为含有词条 t_i 的文档数;n 为特征向量的维数.$w_i(d)$反映了词条 t_i 区分文档属性的能力,一个词语在文档集中出现的范围越广,说明它区分文档的能力越低;另一方面,如果它在某一个特定的文档中出现的次数越高,说明 $\boldsymbol{V}(d)=\{\langle t_1,\ w_1(d)\rangle,\cdots,\langle t_i,w_i(d)\rangle,\cdots,\langle t_n,w_n(d)\rangle\}$在区分该文档内容属性方面的能力越强.这样,利用 TF×IDF 函数进行计算就可以得到全部特征词的权值,从而完成文档的特征表示.

2.3 Blog 用户兴趣挖掘和兴趣相似度计算

1) 定义兴趣类别. 由于没有统一的兴趣分类方式,所以本系统通过自定义兴趣类别来构建兴趣类别体系,体系中尽可能包含生活中的各种兴趣,具有层次化的结构,参见图 1.

2) 分类博客文章. 用分类算法计算待分类的博客文章与各兴趣类别的关联度,关联度较高的类将被判定其归属于哪一兴趣类别.目前已有多种文本分类算法:中心向量法、邻近算法、支持向量机法、简单贝叶斯法等.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为分类标准将形成类球状类别分布,对于多个类距离相近的文本,该算法的分类准确度将急剧下降.kNN 算法[6] 即 k 邻近算法,被普遍认为是分类准确度很高的算法.kNN 算法虽然具有很高的分类准确度,但是它没有训练过程,分类阶段要对所有训练样本进行相似度匹配,计算时间较长.

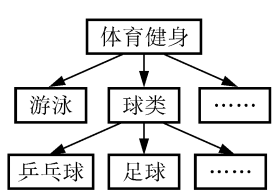

图 1 兴趣类别体系

Fig. 1 Interest category system

通过分析,笔者认为可以采用速度较快的中心向量法和准确性很高的 kNN 算法相结合来完成博客文章的分类.系统先采用中心向量法进行分类,对于超过预定义边界范围的待分类向量再采用

kNN 算法进行补充分类，以保证其分类准确性. 因为在大多数情况下中心向量法即可完成分类，所以该方式显著减少了分类算法的平均计算时间.

通过文本分类技术从用户发表的博客文章中抽取出用户的兴趣取向就构成了用户的兴趣集合，每个集合都由 3 个部分组成，即扫描时间、兴趣类型、兴趣值. 用户兴趣集合的 XML 文档表示如下：

```
〈user username='＊＊＊'〉
  〈scan time value='2010-2-30'〉
  〈interest case log='计算机' last update time='2010-2-12'〉0.52〈/interest〉
  〈interest case log='体育，射击' last update time='2010-2-02'〉0.10〈/interest〉
  〈interest case log='教育' last update time='2010-1-01'〉0〈/interest〉
  〈interest case log='旅游' last update time='2010-1-22'〉0.67〈/interest〉
〈/scan time〉
〈/user〉
```

其中扫描时间 scan time 为用户兴趣生成器定期计算用户兴趣的时间. 在计算兴趣值时引入了艾宾浩斯(EBBINGHAUS H)遗忘曲线，该曲线可以近似地由下述公式[7]表示：$R=e^{(t_1-t_2)/s}$，式中 R 表示记忆程度，即兴趣保留程度；s 为记忆的强健因子，表示遗忘速率的高低；t_1 表示扫描器扫描的时间，t_2 表示文章发布时间.

每个博客用户的兴趣记录被记录在各自的 XML 文件中，推荐系统根据用户兴趣采用余弦计算法计算 blog 页面的相似度，并作为向用户推荐的依据. 余弦相似度[8]定义为 $\mathrm{sim}(\boldsymbol{V}_1,\boldsymbol{V}_2)=\boldsymbol{V}_1\times\boldsymbol{V}_2/(|\boldsymbol{V}_1|\times|\boldsymbol{V}_2|)$，其中 $\boldsymbol{V}_1$，$\boldsymbol{V}_2$ 为 2 个文档的向量. 为了避免文本篇幅差异对聚类结果的影响，本文对所有文本向量都作归一化处理.

2.4　页面重要度的计算

Blog 页面重要度计算是 blog 检索的重要组成部分，它直接决定了匹配的相关链接返回给用户时的先后顺序. PageRank 算法是当今最流行的基于链接分析的重要度计算方法. 初始时刻，可以给某些网站和部分已经考察过的网页赋予不同的初始页面等级 PageRank 值. 任意时刻，任一网页 P 的页面等级 PageRank 值可进行如下计算：假设有网页页面 P_1，P_2，…，P_n 存在链接指向页面 P，它们的页面等级 PageRank 值分别为 $p_r(1)$，$p_r(2)$，…，$p_r(n)$，页面上所有链接的总数分别为 $C(1)$，$C(2)$，…，$C(n)$. 设衰减因子 d 取值范围为 $(0,l)$，则 $p_r(P)=(1-d)+d[p_r(1)/C(1)+p_r(2)/C(2)+\cdots+p_r(n)/C(n)]$[9]. 由此可见，PageRank 算法是一种迭代算法.

2.5　Blog 的排序

根据用户的兴趣，将训练集中的 blog 页面综合考虑其相似度和页面重要度后进行排序，以确定最终推荐的 blog. 本系统设计的排序得分公式[10]为页面总得分＝相似度得分×0.6＋PageRank 得分×0.4，其中 0.6 和 0.4 表示项目的权重，权重值根据经验设定，可以在不断的试验中修改并完善.

3　系统的实现与分析

中国博客网(http://www.blogcn.com)是一个非常著名的提供 blog 服务的网站. 本系统实验数据抓取了该网站上 2010 年 1 月份的数据. 本文将兴趣类型归为以下 8 个大类：财经、旅游、体育、健康、军事、文化、教育和 IT，共采用中文文档 4 110 篇. 中文文档通过分词处理得到不同词条 18 210 条，中文特征词 2 474 个. 候选 blog 列表的获得是利用从训练集抽取出的特征词，通过 web 搜索引擎得出相关的 blog 链接 113 122 个，经 URL 有效性验证、内容丰富度判断和热门程度计算，筛选出2 360 个候选推荐 blog.

Blog 网页经过用户识别、会话识别并消重处理后调用兴趣挖掘模块，计算出当前用户各个会话时段浏览的 blog 网页的兴趣度. 由网页处理模块对训练集中的 blog 网页进行文本预处理、特征选

择、文本向量表示和分类.调用网页相似度计算模块,根据距离公式计算文本之间的相似程度,最后由 blog 推荐模块综合考虑兴趣度、重要度和相似度,利用前述排序公式对 blog 网页进行排序,以确定最终向用户推荐的 blog 序列.将推荐信息回送客户端,客户端将用户这一阶段浏览 blog 网页兴趣度的大小按照用户会话个数展开对应的兴趣分支,每个兴趣分支给出 4 个推荐 blog 链接.此推荐列表定期动态更新.

经过 100 位 blog-digger 系统测评用户的信息反馈,表明该系统推荐的 blog 具有较高的兴趣相关性,同时在内容的可读性和丰富程度上也表示满意,具有使用和研究价值.

参考文献:

[1] CHAU M, XU J. Mining communities and their relationships in blogs: a study of online groups [J]. Int of Human-Computer Studies, 2007, 65(1): 57-70.
[2] BRIN S, PAGE L. The anatomy of a large-scale hypertextual web search engine [J]. Comput Networks & ISDN Syst, 1998, 30(1/7): 107-117.
[3] CHIRITA P A, OLMEDILLA D, NEJDL W. Finding related pages using the link structure of the WWW [C]// Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence. Washington, DC: IEEE Computer Society, 2004: 632-635.
[4] 康楠, 金蓓弘, 李京. 面向 Blog 的兴趣挖掘和推荐系统 [J]. 计算机工程, 2008, 34(2): 72-74.
[5] 赵康, 陆介平, 倪巍伟, 等. 一种基于密度的文本聚类挖掘算法 [J]. 计算机应用研究, 2009, 26(1): 124-126.
[6] WOOD L. Programming the web: the W3C DOM specification [J]. Int Comput, 1999, 3(1): 48-54.
[7] CHEN Yun, TSAI F S, CHAN K L. Machine learning techniques for business blog search and mining [J]. Expert Syst Appl, 2008, 35(3): 581-590.
[8] 宋建康, 张礼平. Web 结构挖掘算法探讨 [J]. 华东理工大学学报: 自然科学版, 2003, 29(5): 537-540.
[9] 郭岩, 白硕, 杨志峰, 等. 网络日志规模分析和用户兴趣挖掘 [J]. 计算机学报, 2005, 28(9): 1483-1496.
[10] 傅怀慧, 林共进, 白峰杉, 等. 阻尼因子对网页排名之敏感度分析 [J]. 中国统计学报, 2005, 43(2): 145-164.

A recommendation system for blog

SUN Duo
(Sch of Inf Engin, Yangzhou Univ, Yangzhou 225009, China)

Abstract: According to the scanning behavior of the users to judge their interesting degree, the author describes the design and implementation of an interest mining, digs the users' interest and gets the better result by the document classification technology. Moreover, combining page importance for web sorting, the system determines the final recommended blog to the user. According this result, the recommended blog has high content related topics.

Keywords: blog recommendation; interest mining; similarity

(责任编辑 史 实)