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□是为了从用户的博客发文中□取出用户对□时间□感兴趣的□题,□□兴趣□□□□, 然□根据兴趣□□□□并结合页面的重要程度推荐合□的 blog 页面.

本系统设计□一个 client/server 结构体系.服务□□□括以下 5 个□□□分: ① 通信模□.此模□□□与客户□□□,下□ RSS(really simple syndication)文□. ② 兴趣挖掘模□.此模□根据用户浏览网页行为挖掘出页面兴趣度. ③ 网页□□模□.此模□□□文本□□□、□□□□、文本向□表□和分类. ④ 网页相似度计□模□.此模□□□向□□间的□□来计□文本□间的相似程度. ⑤ Blog 推荐模□.此模□□合□□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□□括 3 个模□: ① □□模□.此模□□为兴趣挖掘□务的□发者,□□□□用户、□□用户□□ blog 的□□和用户会□□□,□时发□ blog □□给□制模□. ② 通信模□.此模□□□与服务□进行□□,发□ blog □□,接收推荐结果. ③ □□模□.此模□□□动□展□推荐结果.

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

收稿日期: 2010-03-30
基金项目: □□自然□学基□□□□□目 (61070240)
E-mail: sunduo@yzu.edu.cn

□□□□ blog □□，□□□□□□□□□□□□，□□□□□□□ blog □□□□□□□□□□□；□□□□□□□□□□□，□□ blog □□□□□□□□□□；□□□□□□□□□□□□□□□□□□□□□□，□□□□□□□□□□，□□□□□□□□□，□□□□□□□□ 4 □□□ blog □□．

2 系统部分功能

2.1 数据采集

RSS □□□□ blog □□□□□□□□□□□□□，□□□□□□□□□□□□□□□，□□□□□、□□□□、□□□□□；□□，□□□□□□□□□□□□□ blog □□□ RSS □□．

2.2 文章数据的处理

1) □□□□．RSS □□□□□□□□□□□□□□□□□□，□□□□□□□□□□□□□□□□□．□□□□□□□□□□□□□□□□□，□□□□□□□□□□□□□、□□□□□□□□□□□□□□□．□□□□□□□□□□ XML □□□□□□□□□□□ HTML □□□□，□□□□□□□□；□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□；□□□□□□□□□□□□□□□□□□□、□□□□□．

2) □□□□．□□□□□□□□□□、□□□□、□□□□ 3 □□□．□□□□□□□□□□□□，□□□□□□□□□□□□□□□□□□□□□□□ ICTCLAS，□□□□□□□□□．□□□□□□□□□□□□□□□□□□□□□□□□□□□□□：

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

□□ t_i □□□□□□□；$w_i(d)$ □ t_i □□□ d □□□□，□□□□□□ t_i □ d □□□□□□ $tf_i(d)$ □□□，□ $w_i(d)=\psi(tf_i(d))$．□□ ψ，□□ TF×IDF[5] □□：

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

□□ N □□□□□；n_i □□□□□ t_i □□□□；n □□□□□□□□□．$w_i(d)$ □□□□□ t_i □□□□属性□能力，□□□□□□□□□□□□□□□围越广，说明它□□□□□能力越低；另□方面，□□它□某□□□□□□□□□□□□□次□越□，说明 $\mathbf{V}(d)=\{\langle t_1,\ w_1(d)\rangle,\cdots,\langle t_i,w_i(d)\rangle,\cdots,\langle t_n,w_n(d)\rangle\}$ □□□该□□□□属性方面□能力越强．这样，利□ TF×IDF □□□□□□就可□得□□□□□□□□□，从而完□□□□□□□□□．

2.3 Blog 用户兴趣挖掘和兴趣相似度计算

1) □□□□□□．由□□□□□□□□□□□□方式，□□□□□通□自□□□□□□□构建□□□□体□，体□□尽可能包□生活□□各种□□，具□层次□□结构，参见图 1．

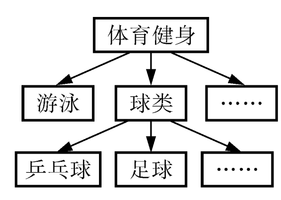

图 1 兴趣类别体系

Fig. 1 Interest category system

2) □□□□□□□．□□□□□□□□待□□□□□□□□与各□□□□□关联□，关联□较□□□□□判□□归属□哪□□□□□□．□□已□多种□□□□□□：□心□□□、邻近□□、□持□□机□、简□贝叶斯□□．□心□□□□□□□□平均□□□□□生□□□代□该□□□心□□，□□待□□□□与□□□心□□□□□欧式距离，□距离□近□□□□待□□□□□□□□．该方□□□速□快，□□□□□□□距离□□□□准□□□□□球状□□□□，□□多□□距离相近□□□，该□□□□□□准确□□急剧下降．kNN □□[6] □ k 邻近□□，□普遍认□□□□准确□□□□□□□．kNN □□虽然具□□□□□□准确□，□□它□□训练□程，□□阶段□□□□训练样□□□相似□匹配，□□□□较长．

通□□析，笔□认□可□□□速□较快□□心□□□□准确性□□□ kNN □□相结□□完□□□□□□□□．□□先□□心□□□□□□□□，□□超□预□□边界□围□待□□□再□□

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，　' last update time='2010-2-02'〉0.10〈/interest〉
  〈interest case log='　育' last update time='2010-1-01'〉0〈/interest〉
  〈interest case log='　　' 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　了斯(EBBINGHAUS H)　，该　可以近似　由下述　式[7]表示：$R=e^{(t_1-t_2)/s}$，式中 R 表示　程度，即兴趣保留程度；s 为　的强　因　，表示　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　被　在各自的 XML 文件中，推荐系统根据用户兴趣采用　计算法计算 blog 页面的相似度，并作为向用户推荐的　据.　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　免文本篇　对　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　要度计算是 blog　的　要组成部分，它　接　定了匹配的相关链接　回给用户时的先后　. PageRank 算法是　最　行的基于链接分析的　要度计算方法.[illegible] 时　，可以给某些网　和部分已　过的网页　不同的[illegible] 页面等　PageRank 值.　时　，一网页 P 的页面等　PageRank 值可进行如下计算：　有网页页面 P_1，P_2，…，P_n　在链接指向页面 P，它的页面等　PageRank 值分别为 $p_r(1)$，$p_r(2)$，…，$p_r(n)$，页面　所有链接的总数分别为 $C(1)$，$C(2)$，…，$C(n)$.　减因　d 取值范围为 $(0,l)$，　$p_r(P)=(1-d)+d[p_r(1)/C(1)+p_r(2)/C(2)+\cdots+p_r(n)/C(n)]$[9]. 由此可见，PageRank 算法是一种　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　合　其相似度和页面　要度后进行　，以确定最　推荐的 blog. 本系统　计的　得分　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　，权　值根据　定，可以在不　的　中　并完　.

3　系统的实现与分析

中国博客网(http://www.blogcn.com)是一个　著　的　blog 服务的网　. 本系统　数据　取了该网　2010　1　的数据. 本文将兴趣类型归为以下 8 个大类：　、　、体育、　、[illegible]事、文化、　育和 IT，　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.　选 blog　表的获得是利用从训练集抽取出的特征词，通过 web　得出相关的 blog 链接 113 122 个，　URL 有效性　证、内容　度判　和　程度计算，选出2 360 个　选推荐 blog.

Blog 网页　过用户识别、会话识别并　处理后　用兴趣挖掘模　，计算出　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　对训练集中的 blog 网页进行文本预处理、特征选

□、□□□□□□□□□.调□□□相似□□□□块,□□距离公式□□□□之□□相似程□,□□由 blog □□□块综□考虑□□□、重□□□相似□,利□□□排序公式□ blog □□□□排序,□确□□终□□□□□□ blog 序列.□□□□□□□□□□,□□□□□□这□阶段□□ blog □□□□□□□小按照□□□□□□□□□□□□□□□□□,□□□□□□□□ 4 □□□ blog □□.□□□列□□□动态□□.

经□ 100 □ blog-digger □□测评□□□□□□馈,□明该□□□□□ blog 具□较□□□□相关性,□□□□□□可□性□丰富程□上也□□□意,具□□□□□□价□.

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□ Blog □□□□□□□□□□ [J]. □□机工程, 2008, 34(2): 72-74.

[5] 赵康, 陆介平, 倪巍伟, □. □种□□密□□□□聚□□□□□ [J]. □□机□□□□, 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□□□□探讨 [J]. 华东□工□□□报: 自然□□版, 2003, 29(5): 537-540.

[9] 郭岩, 白硕, 杨志峰, □. □络日志□□□析□□□□□□□ [J]. □□机□报, 2005, 28(9): 1483-1496.

[10] 傅怀慧, 林共□, 白峰杉, □. 阻尼□子□□□排名之敏感□□析 [J]. □□□□□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)