个性化搜索引擎中用户兴趣模型的构建方法①

刘忠宝, 赵文娟

(太原山西大学 商务学院 信息学院, 太原 030031)

摘 要: 在分析个性化搜索引擎的基础上, 提出一种构建用户兴趣模型的方法. 该方法综合考虑用户注册兴趣及浏览行为, 将用户兴趣分为长期兴趣和短期兴趣并通过兴趣树进行存储. 遗忘机制的引入保证模型能够及时准确地反映用户兴趣. 模拟实验表明, 本文提出的用户兴趣模型能够有效地提高检索效率, 使搜索结果更好地满足用户个性化需求.

关键词: 个性化搜索引擎; 用户兴趣模型; 用户兴趣树; 遗忘机制

Construction of User Interest Model in Personalized Search Engine

LIU Zhong-Bao, ZHAO Wen-Juan

(School of Information, Business College of Shanxi University, Taiyuan 030031, China)

Abstract: This paper presents a way to construct user interest model in personalized search engine. The user interest model takes the registrated interest and user browse behavior into consideration. It consists of long-term and short-term which is stored in user interest tree. Forgetting mechanism is provided to timely refresh user interest. Simulated experiments verify the user interest model proposed in this paper is effective and competitive.

Key words: personalized search engine; user interest model; user interest tree; forgetting mechanism

随着互联网的迅猛发展, 各种信息以几何级数的方式增长, 信息量的增大使用户很难找到所需信息[1-3]. 搜索引擎的出现很大程度上解决了这一难题. 搜索引擎以一定的策略在互联网中搜集、发现信息, 对信息进行理解、提取、组织和处理, 并为用户提供检索服务, 从而起到信息导航的作用. 然而目前大多数搜索引擎为用户提供的信息单一, 无法满足用户个性化的要求, 这就需要以用户为中心构建搜索的方法、技术、结果与过程.

实现个性化搜索的关键是建立用户兴趣模型. 只有全面真实地了解用户兴趣, 才能针对不同用户提供个性化服务. 因此, 本文在分析个性化搜索引擎的基础上, 提出一种构建用户兴趣模型的方法. 在该模型的帮助下, 人们能更好地在 Internet 中找到所需信息.

1 个性化搜索引擎理论模型

个性化搜索引擎一般由用户接口、概念提取、查询扩展、检索器、结果排序、网络蜘蛛、索引器及索引数据库、用户兴趣库等部分组成. 个性化搜索引擎理论模型如图 1 所示.

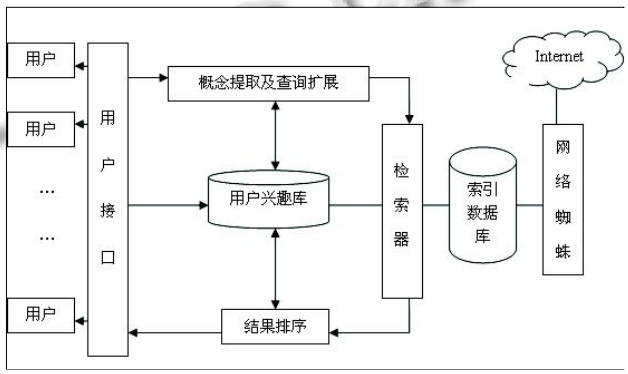

图 1 个性化搜索引擎理论模型

1) 用户接口: 用户提出检索请求并将检索结果返回给用户;

2) 概念提取: 运用中文分词法提取页面信息;

① 基金项目:山西大学商务学院科研基金(XS2011005)

收稿时间:2011-07-11;收到修改稿时间:2011-08-20

3) 查询扩展: 利用已建好的词典库或知识库进行查询词条扩展, 提高搜索的召回率和查准率;

4) 检索器: 从索引数据库中找出与用户查询请求相关的页面;

5) 结果排序: 对满足用户需求的页面排序保证重要页面排名靠前;

6) 网络蜘蛛: 抓取网站页面信息;

7) 索引器: 将页面表示为一种便于检索的方法并存储于索引数据库中;

8) 用户兴趣库: 根据用户兴趣模型, 存放用户兴趣知识.

与传统搜索引擎相比, 上述模型的优势主要体现在两方面: 1)用户可以使用灵活多样的描述方式表达信息需求; 2)用户可从多个信息源获取所需信息.

2 用户兴趣模型的构建方法

提高个性化检索服务质量的关键在于全面了解用户兴趣. 本文采用基于文本内容的数据挖掘方法获取用户兴趣, 利用遗忘机制对用户兴趣进行更新.

2.1 用户兴趣模型结构

用户兴趣模型由页面预处理、页面分类、兴趣生成以及兴趣更新四部分组成. 用户兴趣模型总体结构如图 2 所示.

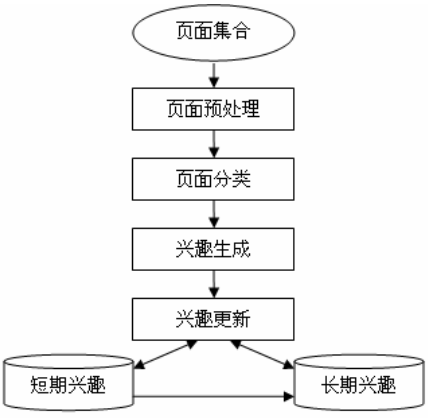

图 2　用户兴趣模型结构图

1) 页面预处理

页面预处理包括: ①页面清洗: 清除掉一些与研究无关的文件, 如图片文件及脚本程序等; ②页面信息提取及向量表示.

2) 页面分类

某页面反映的用户兴趣是偶然的, 但页面归类后反映的兴趣具有很高的确定性. 因此页面分类对于获取用户兴趣至关重要. 本文采用VSM求余弦的方法对页面进行分类.

3) 兴趣生成

兴趣生成包括: ①兴趣表示; ②用户兴趣树建立. 同时引入时间机制, 突出用户兴趣的时效性.

4) 兴趣更新

兴趣更新基于遗忘机制, 周期性地更新用户的短期兴趣和长期兴趣, 保证用户兴趣模型及时准确地反映用户需求.

2.2 用户兴趣挖掘

通过挖掘用户访问的历史记录获取用户兴趣. 面对杂乱无章的页面首先应对其进行特征表示, 然后根据页面特征所反映出来的内容对其进行分类, 最后得到用户兴趣的向量表示.

2.2.1 页面特征表示

目前比较成熟的页面特征表示方法有布尔逻辑模型[4]、向量空间模型(Vector Space Model , VSM)[5]、概率模型[6]等. 本文页面表示方法基于 VSM.

本文对关键词的考察重点在于对其权值的考察, 因此页面特征表示采用加权特征词方法, 该方法包含两步: ①页面特征词提取; ②特征词权值计算.

页面特征词提取方法如下:

① 将页面转化成文本并保留某些重要标记信息;

② 对文本文件切词处理[7];

③ 去掉与页面内容无关的虚词;

④ 去掉低频词, 低频阈值[8]确定方法见表 1.

表 1　低频阈值确定方法

文章长度	低频阈值
(0,200]	2
(200,4000]	3
(4000,10000]	4
(10000,25000]	5
(25000,+∞)	6

⑤ 剩下的词作为特征词, 并保留其出现频率.

页面特征词权值计算方法如下:

① 根据特征词 t_i 在页面中出现的位置和次数 f_i 计算其频率:

$$f_i^{'} = f_i \times s_i \quad (i=1, 2, ..., n) \tag{1}$$

其中 s_i 为特征词 t_i 对应页面标记的权系数 s_i 的取值见

表 2.

表 2 HTML 部分标记权重设置表

HTML 标记	权重
<TITLE>	1
<H1> <H2> <H3>	0.8
<B> <STRONG>	0.7
<BODY>	0.5

② 由于页面长短不一，式(2)实现规范化处理.

$$f_i = \frac{f_i'}{\sqrt{\sum_{i=1}^{n} f_i'^2}} \quad (i=1, 2, ..., n) \tag{2}$$

在得到页面特征词及其权值后，可将页面表示为 $p=\{(k_1, tf_1),(k_2, tf_2), \cdots ,(k_n, tf_n)\}$ (i=1, 2, ⋯, n)，其中 k_i 和 tf_i 分别为页面特征词及其权值.

2.2.2 页面分类

本文采用基于VSM求余弦的方法对页面分类. 计算公式如下:

$$sim(p, u_c) = \frac{\sum_{i=1}^{n} p(i) \quad u_c(i)}{\sqrt{\sum_{i=1}^{n} p(i)^2 \quad \sum_{i=1}^{n} u_c(i)^2}} \tag{3}$$

其中, sim(p, u_c)表示页面 p 和用户兴趣类 u_c 之间的相似程度; $p(i)$表示页面中第 i 个特征词的权值; u_c (i)表示用户兴趣类中第 i 个特征词的权值.

2.2.3 用户兴趣类向量表示

用户兴趣类向量表示方法如下:

① 统计用户兴趣模型中所有页面数量 N;

② 求出页面特征词的并集 K={k_1, k_2, ⋯, k_m}作为用户兴趣类向量的特征词集;

③ 统计特征词 k_i 在页面中出现的次数 n_i;

④ 利用式(4)计算各特征词的权值:

$$w_i = TF_i ? IDF_i \quad \sum_{j=1}^{i} tf_{ij} \quad log(\frac{N}{n_i}) \tag{4}$$

在得到页面特征词及其权值后，可得用户兴趣类向量 u_c={(k_1, w_1), (k_2, w_2), ⋯, (k_n, w_n)} (i=1, 2, ⋯, n), 其中 k_i 属于 K, K 为兴趣类特征词集, w_i 为兴趣类特征词的权值.

2.3 用户兴趣存储

借鉴 ODP[9]思想，本文通过建立兴趣树对用户兴趣进行管理. 用户兴趣树主要有两类结点: 用户兴趣结点和特征词结点. 用户兴趣树如图 3 所示.

图 3 中，虚线方框表示的结点是为了表示方便而形成的结点; 粗线方框表示用户结点; 中间两层用于表示用户兴趣类别的结点称为兴趣结点; 最底层的结点表示特征词结点.

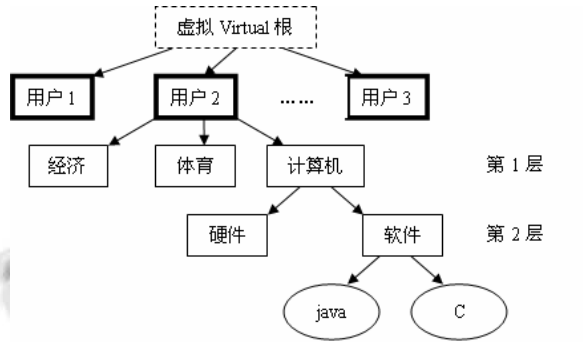

图 3 用户兴趣树示例

2.4 用户兴趣模型建立与更新

2.4.1 用户兴趣模型建立

1) 长期兴趣树的建立

用户可能短期内对某领域感兴趣而忽略了长期感兴趣的领域，这势必会影响用户搜索效果. 为了避免这种情况，将长期兴趣的初始兴趣度设置为 10.

长期兴趣树建立方法如下:

① 将用户帐号作为长期兴趣树的用户结点;

② 根据用户输入的兴趣生成相应兴趣结点，并置用户定制的长期兴趣对应结点 Node(c_i)兴趣度为 10;

③ 借鉴 ODP 分类模型建立长期兴趣树;

④ 利用式(5)计算长期兴趣树各结点的兴趣度:

$$v_j = \sum_{i=1}^{k} v_i \tag{5}$$

其中,v_i($1 \leqq i \leqq k$)表示特征词结点权值或兴趣结点 Node(c_i)兴趣度; k 表示父结点的子类数.

2) 短期兴趣树的建立

用户注册时定制的兴趣使得模型能快捷地建立起长期兴趣树，但这些兴趣只是用户初始兴趣，随着时间推移，用户兴趣还会发生变化，这就需要建立短期兴趣树来及时地反映用户兴趣变化.

短期兴趣树建立方法如下:

① 将用户帐号作为短期兴趣树的用户结点;

② 根据页面分类结果，逐类计算特征词的权值,

同时把各词最早出现的日期作为创建日期;

③ 参考 ODP 分类模型建立短期兴趣树;

④ 利用式(5)计算短期兴趣树中各结点的兴趣度.

2.4.2 用户兴趣模型更新

用户兴趣模型更新即自动地对用户兴趣变化做出判断, 适当地调整用户兴趣权值, 尽量保持原有兴趣的稳定性, 防止用户兴趣反复变化. 本文引入遗忘因子描述用户兴趣逐渐遗忘的过程.

遗忘因子 F(x)[10]定义为:

$$F(x)=e^{-\frac{\log 2}{hl}(cur-est)} \tag{7}$$

其中,*cur* 表示当前日期, *est* 表示兴趣特征词或兴趣类第一次出现的日期, *hl* 表示半衰期, 即经过 *hl* 天后用户的兴趣遗忘一半.

1) 短期兴趣树更新

短期兴趣树的更新包括用户新兴趣的添加以及旧兴趣的遗忘. 短期兴趣更新方法如下:

① 统计页面特征词;

② 根据页面分类结果, 逐类统计特征词, 利用式(1)和式(2)计算特征词权值, 生成最新的兴趣类向量;

③ 对原兴趣树中的特征词进行遗忘, 利用式(4)分别计算各词的遗忘因子, 得到遗忘后各词权值; $w_i^{'}=w_iF(t_i)$

④ 将兴趣类向量中的特征词加到模型中, 若短期兴趣中已存在该词, 则转⑤, 否则转⑥;

⑤ 重新计算该词的权值 $w_i=w_i^{'}+w_i^{''}$, 其中 $w_i^{'}$ 为③计算得到的权值, $w_i^{''}$ 为当前兴趣向量中的权值, 并将该词的 *est* 设定为当前日期;

⑥ 将该词添加到模型, 并将 *est* 设定为当前日期;

⑦ 更新短期兴趣特征词;

⑧ 利用式(5)对短期兴趣树逐层计算各父类结点 Node(c_j)的兴趣度;

⑨ 更新短期兴趣类.

2) 短期兴趣向长期兴趣转化

某段时间内, 若用户经常访问某个类或某个词, 类兴趣度或特征词权值会逐渐增大. 当累积到一定程度, 若类兴趣度大于阈值 th_c 或特征词对类兴趣度的影响程度大于阈值 th_t, 则将其转化为长期兴趣. 短期兴趣向长期兴趣转化方法如下:

① 从短期兴趣树中遍历出兴趣度大于阈值 th_c 的用户兴趣类和权值大于阈值 th_t 的特征词;

② 对①找到的特征词进行遗忘, 利用式(4)分别计算各词的遗忘因子, 得到遗忘后各词权值 $w_i^{'}=w_iF(t_i)$;

③ 将各词添加到长期兴趣树中相应位置, 若该词已存在则转④, 否则转⑤;

④ 重新计算各词权值 $w_i=w_i^{'}+w_i^{''}$, 其中 $w_i^{'}$ 为②得到的权值, $w_i^{''}$ 为该词在原长期兴趣树中的权值, 并将该词的 *est* 设定为当前日期;

⑤ 将该词加到长期兴趣树, 并设 *est* 为当前日期;

⑥ 更新长期兴趣类.

3) 长期兴趣树更新

长期兴趣相对稳定, 但随着时间推移, 用户对长期兴趣亦会逐渐遗忘. 长期兴趣树更新方法如下:

① 对长期兴趣中的所有词进行遗忘, 分别计算各词的遗忘因子, 同时调整各词权值 $w_i^{'}=w_iF(t_i)$, 并将各词的 *est* 设为当前日期, 更新长期兴趣特征词;

② 利用式(5)逐层计算各结点 Node(c_j)兴趣度;

③ 淘汰兴趣度小于阈值 th_c 的兴趣类;

④ 更新长期兴趣类.

3 实验与分析

实验目的是考察用户兴趣模型能否正确理解用户需求并有效提供个性化检索服务.

3.1 个性化模型建立

实验数据来源于某用户 2011 年 3 月 1 日至 2011 年 3 月 15 日访问的 192 个页面(见表 3). 页面可归为六类: 网球、编程、数码产品、操作系统、心理健康、礼品, 类型 ID 分别为 100、101、102、103、104、105. 用户定制的兴趣是: 网球、编程、数码产品.

表 3　测试页面分布表

ID / 批次	100	101	102	103	104	105
1	14	0	0	20	0	16
2	4	24	30	6	0	0
3	0	0	20	4	20	0
4	0	16	18	0	0	0

经多次实验可得如下参数经验值: 短期遗忘因子

hl_s=2, 长期遗忘因子 hl_l=7, 兴趣度阈值 th_c=10, 特征词权值 th_t=0.01.

用户兴趣模型分 4 批学习, 实验过程如下:

1) 处理第一批数据(2011 年 3 月 1 日)得到的短期兴趣见表 4.

表 4 第一批数据处理后的结果(短期兴趣)

ID	类名	兴趣度
100	网球	53.4
103	操作系统	70.7
105	礼品	54.0

2) 处理第二批数据(2011 年 3 月 5 日)得到的短期兴趣和长期兴趣分别见表 5 和表 6.

表 5 第二批数据处理后的结果(短期兴趣)

ID	类名	兴趣度
100	网球	30.9
101	编程	82
102	数码产品	69.3
103	操作系统	34.2
105	礼品	9.6

由表 5 可知: 新增兴趣有编程、数码产品. 礼品的兴趣度下降, 遗忘速度较快; 而用户对编程、数码产品较为感兴趣.

将满足条件的短期兴趣转化为长期兴趣. 由表 5 可知只有礼品的兴趣度 9.6<th_c. 除礼品外, 将表 5 中其他类型的短期兴趣转化为长期兴趣.

表 6 第二批数据处理后的结果(长期兴趣)

ID	类名	兴趣度
100	网球	39.5
101	编程	69.9
102	数码产品	69.1
103	操作系统	30.3

表 6 中, 网球长期兴趣比短期兴趣大的主要原因是用户定制了网球兴趣, 保证该兴趣不会过早地从长期兴趣中淘汰.

3) 处理第三批数据(2010 年 3 月 10 日)得到的短期兴趣见表 7.

表 7 第三批数据处理后的结果(短期兴趣)

ID	类名	兴趣度
100	网球	5.5
101	编程	14.5
102	数码产品	59
103	操作系统	19.5
104	心理健康	45.8
105	礼品	1.7

由表 7 可知: 增加的兴趣是心理健康. 网球、编程、数码产品、操作系统、礼品等兴趣均有不同程度的下降.

4) 处理第四批数据(2010 年 3 月 15 日)得到的短期兴趣和长期兴趣分别见 8 和表 9.

表 8 第四批数据处理后的结果(短期兴趣)

ID	类名	兴趣度
100	网球	1
101	编程	54.8
102	数码产品	65.2
103	操作系统	3.4
104	心理健康	8.1
105	礼品	0.3

由表 8 可知: 用户短期兴趣发生较大变化: 编程和数码产品的兴趣度上升, 网球、操作系统、心理健康、礼品的兴趣度下降. 将符合条件的短期兴趣转化为长期兴趣, 并及时更新长期兴趣得到表 9.

表 9 第四批数据处理后的结果(长期兴趣)

ID	类名	兴趣度
100	网球	21
101	编程	58.9
102	数码产品	70
103	操作系统	11.3

经过 15 天对用户访问页面的跟踪, 获得用户的长期兴趣和短期兴趣如图 4 和图 5 所示.

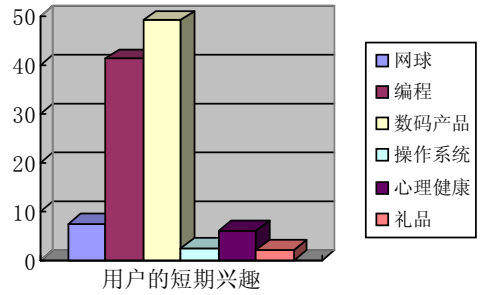

图 4 短期兴趣比重示意图

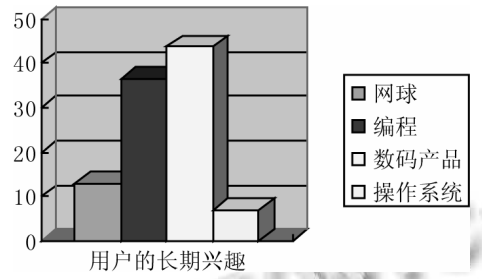

图 5 长期兴趣比重示意图

4.2 个性化检索

先后输入检索词: “索尼爱立信”、“苹果”、“windows、“抑郁症”, 实验结果见表 10.

表 10 个性化检索结果

检索词	检索序号	检索内容	类别
索爱	1-8	索爱数码产品	数码产品
	9-10	索爱网球公开赛	网球
苹果	1-5	苹果牌数码产品	数码产品
	6-7	水晶苹果礼品	礼品
windows	1-6	windows 编程相关内容	编程
	7-18	windows 操作系统	操作系统
抑郁症	1-3	抑郁症相关内容	心理健康

由表 10 可以看出: 检索结果与用户兴趣一致. 由此可见: 用户兴趣模型能正确理解用户需求并有效提供个性化检索服务.

4 结论

在分析个性化搜索引擎的基础上, 提出一种用户兴趣模型的构建方法. 该方法综合考虑用户注册兴趣及浏览行为, 巧妙地将用户兴趣划分为长期兴趣和短期兴趣并通过兴趣树存储用户兴趣. 此外, 随着时间推移, 遗忘机制的引入保证模型能够及时准确地反映用户兴趣. 模拟实验表明, 本文提出的用户兴趣模型能够有效地提高检索的查准率, 使搜索结果更好地满足用户个性化需求.

参考文献

1 Ying XM. The research on user modelting for interest personallized services. National University of Defense Technology, 2003.

2 Xue GR, Lin CX. Scalable collaborative filtering using cluster-based. Hong Kong University of Science and Technology, ACM, 2005.

3 曾春,刑春晓,周立柱.基于内容过滤的个性化搜索算法.软件学报,2003,14(5):999–1004.

4 刘红泉,张亮峰.布尔逻辑检索模型的分析探讨.现代情报,2004,9:4－6.

5 唐明伟,卞艺杰,陶飞飞.基于语义向量空间模型的文档检索系统研究.情报杂志,2010,5:167–170.

6 刘华.文本分类相似度模型和概率模型的实现与比较.现代图书情报技术,2006,4:53–55.

7 许林杰.中文文本分词研究.山东师范大学,2003.

8 顾立帆,王永成.联想树分析法及其在无词库中文自动标引中的应用.情报学报,1992,11(5):354–360.

9 陈一峰,赵恒凯,余小清,万旺根.基于本体的用户兴趣模型构建研究.计算机工程,2010,36(21):46–51.

10 蒋萍,崔志明.智能搜索引擎中用户模型分析与研究.微电子学与计算机,2004,21(11):24–26.