

一种面向主题的用户兴趣挖掘模型研究

郑晓健¹, 庞淑英², 何英³

(1. 昆明理工大学 津桥学院计科系, 云南 昆明 650106; 2. 昆明理工大学 计算中心, 云南 昆明 650093;
3. 昆明学院 计算机与网络技术系, 云南 昆明 650031)

摘要: 用户查询表达式中包含的有效信息对于检索结果影响很大, 利用企业信息系统搜集和挖掘与用户检索兴趣相关的信息, 有助于解决检索信息不足的问题. 为此, 提出一种面向领域主题概念的搜索引擎构架, 据此建立面向主题的复合算子调节用户兴趣趋向的线性规划预测模型, 该模型可预测用户的最大兴趣, 生成用户兴趣查询表达式, 提高检索的查准率和查全率. 另外, 还提出一种用户兴趣演变探测因子重建用户兴趣特征向量的方法.

关键词: 用户兴趣; 兴趣挖掘; 主题检索; 语义网络; 搜索引擎

中图分类号: TP391 **文献标识码:** A **文章编号:** 1674-5639(2010)03-0073-03

Research on a Topic-Oriented Model of User Interest Mining

ZHENG Xiao-jian¹, PANG Shu-ying², HE Ying³

(1. Department of Computer Science and Technology, Oxbridge College, Kunming University of Science and Technology, Yunnan Kunming 650106, China;
2. Center of Computer, Kunming University of Science and Technology, Yunnan Kunming 650093, China;
3. Department of Computer and Network Technology, Kunming University, Yunnan Kunming 650031, China)

Abstract: The efficient information in the query of user affect the results of searching extensively. Collecting user information by the advantages of gathering user interest information from enterprise information systems will help to solve the shortage of information retrieval. So, a model of user interest mining of the multiple-fact, based on the topic-oriented frame of search engines, is established for creating the retrieval query with their greatest interest in search engines based on topic to improve the recall and fallout of retrieval. Besides, a method has been put forward to re-establish the user interest character vector by detecting factor for user interest development.

Key words: user interest; interest mining; topic retrieval; concept semantic network; retrieval engines

0 引言

信息检索系统面临的主要挑战之一是从包含很少有效信息的用户检索要求中去猜测用户查询的意图^[1]. 因此开拓用户兴趣数据源, 从中挖掘用户兴趣来解决检索信息不足问题是目前国内外搜索引擎的研究方向, 由此也产生了所谓个性化搜索引擎. 学者们的研究重点主要放在收集和分析用户个人的检索行为习惯上. 但我们认为用户的检索行为还与其工作环境和内容等因素有密切关系, 特别是企业级搜索引擎, 其服务对象主要是企业用户和相关人员. 我们提出将企业搜索引擎与企业信息系统相结合, 通过它们交换信息来获得额外的信息, 再经过数据挖掘模型的处理就可以产生具有相当可靠度的用户检索兴趣信息, 生成用户兴趣查询表达式, 从而提高

检索的查准率和查全率.

为此本文提出一种以主题概念为导向来挖掘用户兴趣的方法, 即以概念和概念的扩展为基础来产生用户查询要求, 下图 1 为本文提出的面向主题搜索引擎架构. 首先建立领域概念库, 通过主题概念和概念间的语义联系来扩展主题概念, 运用适当的概念扩展策略可以提高查询的性能, 体现出主题检索的智能性. 领域概念库的知识框架按领域主题分类体系构建, 其中每个主题用一组概念来表述, 并形成主题概念特征向量, 用语义网络组织主题概念架构. 然后, 通过分类处理待检索文档, 将它们聚集于相应主题概念下. 各个主题概念间的语义联系使主题内容得到扩展, 例如可以进行同义词扩展、语义蕴涵扩展和语义相关扩展等. 这样还解决了文献[1]提到的领域知识规模的庞大, 使检索效率下降的问题^[1],

收稿日期: 2010-03-17

基金项目: 云南省科技厅专项计划资助项目(2001TJ01)

作者简介: 郑晓健(1963—), 男, 湖北随县人, 讲师, 硕士, 主要从事企业智能化研究; 庞淑英(1958—), 女, 四川资中人, 教授, 硕士生导师, 主要从事地理信息系统研究; 何英(1962—), 女, 江苏江阴人, 副教授, 主要从事算法设计与计算机网络研究.

以及文献[2]降低向量空间维度的问题^[2].

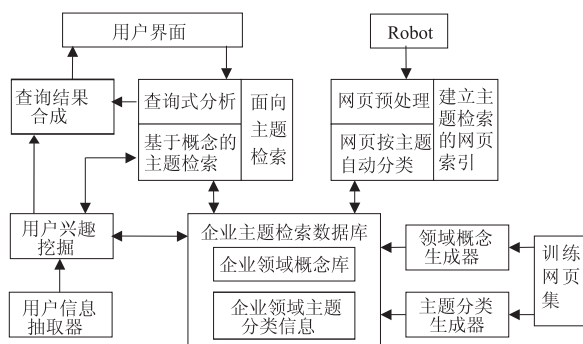


图1 面向主题搜索引擎架构图

企业领域概念库是领域知识的分类体系结构,包含着企业所有业务相关的概念、术语以及它们之间的关系,其功能是将待检索网页中遇到的词汇快速地转换为统一的概念,其中包括同义词的转换和概念的联想等。企业领域主题分类信息库是按照企业的各类业务划分成的各种主题(业务问题与专题)构建的专业领域分类信息结构,它的主要作用是:描述各个主题的逻辑关系;将反映各主题内容的网页与其所属主题联系起来,便于主题检索。企业领域概念库、企业领域主题分类信息库和与检索、建立索引、用户兴趣挖掘等有关的信息共同构成企业主题检索数据库,它是整个系统的核心。Robot从互联网上搜集网页^[3],先经过网页预处理组件进行分词处理,利用企业领域概念库,通过基于概念的标引生成文档特征向量,然后按其描述的特征进行分类,将网页分配到所属主题概念下,并为其建立索引。用户兴趣可以通过用户兴趣挖掘组件得到,与用户查询式综合就得到带有用户个性的查询式,再用它去企业领域主题分类信息库中查找,检索出反映用户兴趣的结果^[4]。

1 用户兴趣模型建模

个性化检索面临着用户兴趣模型建模、相似性计算和模型的更新3个主要问题。其中关键是用用户兴趣模型建模,它直接影响到个性化检索服务的技术性能。

1.1 用户兴趣模型

企业中用户兴趣信息主要来自以下两方面:1)用户本人信息。如:当前的工作内容、近期访问过的网页、个人爱好、所学或从事的专业等;2)用户所在部门信息。如:该部门工作内容和发展目标,最近正进行的工作项目等。由于信息相对丰富,而且查询者在信息源和关注程度上不尽相同,可以引入复合算子加权重调节方式来进行调节,从而预测用户兴趣趋向。我们给出如下定义:

定义1 用户信息类别为三元式 CL :

$$CL = (h, N, V), \quad (1)$$

其中, h 为类别句柄, N 为类别描述, V 为类别特征

项向量 $V = (t_1, t_2, \dots, t_i, \dots, t_m)$, t_i 为类别特征向量的特征项, $t_i \in TG$, $i = 1, 2, \dots, m$, TG 为用户信息集合, t_i 可以用其权重描述,并由多个因子来决定^[5]。

定义2 用户兴趣为一个四元式 I :

$$I = (CL, A, TG, f), \quad (2)$$

其中, CL 为用户信息类别集合 $CL = \{C_1, C_2, \dots, C_n\}$, TG 为用户信息集合, A 为 CL 的权重集合 $A = \{A_{c1}, A_{c2}, \dots, A_{cn}\}$, f 为 CL 到 TG 的映射:

$$f(CL, A) = \begin{cases} \Phi & \text{没有符合 } CL \text{ 的用户兴趣特征} \\ & \text{向量与之对应,} \\ Q & \text{符合 } CL \text{ 的用户兴趣特征向量,} \end{cases} \quad (3)$$

其中, $Q = (q_1, q_2, \dots, q_m)$, $q_i \in TG$, $i = 1, 2, \dots, m$ 。

在模型中, A 设置为用户兴趣信息影响度,映射 f 为生成用户兴趣查询式的预测算法,来自企业多个途径的用户信息由算法处理而得到预测的用户兴趣向量 Q 。用户对各类信息重要性的理解可以由 A 来反映。由于用户兴趣不易量化,提出一个精确的算法很难,只能根据实际情况处理,为此引入以下新概念。

定义3 主题兴趣度 SI 指用户对一个给定主题 S 感兴趣的程度,其值域为 $[0, 1]$, 0 表示用户对 S 没有兴趣, 1 表示用户对 S 有兴趣。

定义4 概念兴趣度 CI_k , 指用户对用户信息类别 C_k 中某给定概念 I_k 感兴趣的程度,且

$$CI_k = \frac{tf_{I_k}}{\sum_{k=1}^m tf_{I_k}}, \quad (4)$$

其中, tf_{I_k} 为 C_k 中概念 I_k 出现的频数, $\sum_{k=1}^m tf_{I_k}$ 为 C_k 中所有概念出现的频数总和。 CI_k 的值域为 $[0, 1]$, 0 表示用户对 I_k 没有兴趣, 1 表示用户对 I_k 有兴趣。

定义5 概念影响度 CAF_k , 指用户兴趣特征向量 Q 中概念特征项 t_k 对主题兴趣度 SI 的影响程度,且

$$CAF_k = \frac{tf_k}{\sum_{k=1}^m tf_k}, \quad (5)$$

其中, tf_k 为给定约束条件下,概念特征项 t_k 在用户信息类别集合 $CL = \{C_1, C_2, \dots, C_n\}$ 出现的频数, $\sum_{k=1}^m tf_k$ 为 CL 的概念特征项频数的总和,约定 $0 \leq CAF_k \leq 1$, 且 $\sum_{k=1}^m CAF_k = 1$ 。

定义6 用户信息类别影响度 $CLAF_k$, 指用户信息类别 C_k 对主题兴趣度 SI 的影响程度,且

$$CLAF_k = \frac{tf_{C_k}}{\sum_{k=1}^n tf_{C_k}}, \quad (6)$$

其中, tf_{C_k} 为给定约束条件下, C_k 的所有概念特征项 t_k 在 CL 的总频数, $\sum_{k=1}^n tf_{C_k}$ 为用户信息类别集合 CL 出现的概念频数的总和。约定 $0 \leq CLAF_k \leq 1$, 且 $\sum_{k=1}^n CLAF_k = 1$ 。

用户信息类别影响度 $CLAF_k$ 可以认为是概念兴趣度综合影响的结果。通过以上定义,挖掘用户对某个主题兴趣的问题就是在约束条件下求解最大用

户兴趣度 SI 的问题. 模型的标准形式为:

$$\begin{aligned} \max SI &= CX, \\ \begin{cases} AX = b, \\ X \geq 0, \end{cases} \end{aligned} \quad (7)$$

其中, $C = (C_1, \dots, C_i, \dots, C_n)$,

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix},$$

$b = (b_1, b_2, \dots, b_i, \dots, b_m)^T$, $X = (x_1, x_2, \dots, x_i, \dots, x_m)^T$, $b \geq 0$.

C_i 为概念影响度, 用户兴趣特征向量 Q 中概念特征项 t_k 对主题兴趣度 SI 的影响程度, $X = (x_1, x_2, \dots, x_i, \dots, x_m)^T$ 为用户兴趣向量, $b = (b_1, b_2, \dots, b_i, \dots, b_m)^T$ 为用户信息类别影响度向量, b_i 为用户信息类别影响度, $a_{ij} = b_i CI_{ij}$, CI_{ij} 为概念兴趣度. 模型的求解可以用单纯形法, 而在实际应用中采用市面上现成的线性规划模型库来计算.

另外, 对于用户信息类别特征向量的各概念元素有其它约束, 如要求时间必须在某个期间内等. 在概念语义网络模型中概念顶点的倒排文件中对网页已有约束, 如记录生成和修改时间等, 在检索时按约束查询即可.

1.2 用户兴趣特征向量与网页文档的相似度计算

通过收集各用户兴趣类别的信息, 建立用户信息资料库. 当用户要查询时, 从用户信息资料库中提取用户信息, 通过用户兴趣模型产生用户兴趣特征向量, 将其与用户的查询式进行融合产生经过扩展的用户查询式, 然后再到企业领域主题分类检索信息库中检索符合用户兴趣的网页, 最后排序输出符合要求的网页结果, 见图2.

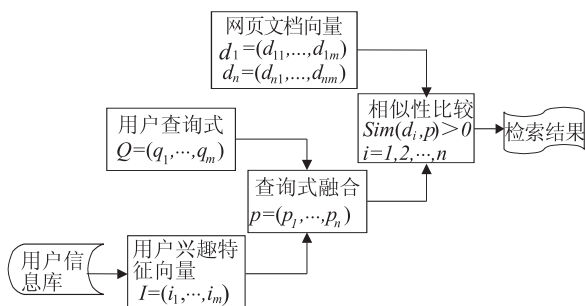


图2 用户兴趣特征向量与网页文档的相似度比较流程图

用生成的综合用户兴趣查询向量到企业领域主题分类检索信息库中检索符合用户兴趣的网页, 就

是用 P 按策略路径去与主题顶点及其网页进行相似性匹配:

$$Sim(d_i, P), \quad (8)$$

其中 d_i 为主题顶点的主题特征向量.

1.3 用户兴趣特征向量的重建

用户兴趣会随时间变化, 必须研究动态调整用户兴趣模型的方法. 例如可以根据每次或者一段时间中用户的检索行为对用户兴趣类别权重进行调整. 依据 *Rocchio* 反馈模型就可以完成动态调整的任务. 还可以设置用户兴趣变化探测因子来实现此功能:

$$\begin{aligned} \text{当 } P_{k+1} - P_k > \tau \text{ 时,} \\ A_{ci} &= R(\tau), \end{aligned} \quad (9)$$

其中, $\tau = (a_1, a_2, \dots, a_n)$, P_{k+1} 为新的用户兴趣特征向量, P_k 为旧的用户兴趣特征向量. 可以根据用户兴趣特征向量的变化量对用户兴趣类别权重因子进行相应的调整. 最简单的办法是计算 P_{k+1} 和 P_k 的相似度 $Sim(P_{k+1}, P_k)$, 如果它超过某个阈值 ε , 就对类别权重进行调整. 通过对阈值 ε 设定不同的值可以调整探测因子的灵敏度.

2 结束语

个性化检索是搜索引擎发展的一个重要方向, 本文提出利用企业信息系统的用户信息来支持用户兴趣挖掘的方法, 在面向主题检索模型基础上建立根据用户兴趣挖掘用户兴趣的技术方法具有一定的参考价值. 由于用户兴趣是多变的, 表达用户兴趣的模型也是变化的, 因此, 提出通用化的用户模型有相当难度. 要解决这个问题可以从研究表达用户兴趣的标准化语言入手, 来实现模型的通用化, 这是未来一个研究方向.

[参考文献]

- [1] 焦玉英. 信息检索进展[M]. 北京: 科学出版社, 2005.
- [2] 邹志文, 柯青. 基于向量空间模型的主动推送系统设计与优化[J]. 现代图书情报技术, 2005(7): 42-45.
- [3] 丁海燕. 利用 Dreamweaver8.0 实现动态网页的数据库访问[J]. 昆明学院学报, 2009, 31(3): 77-79.
- [4] 庞淑英, 付铁威, 胡恒奎, 郑晓健, 等. 挖掘用户兴趣的 Web 智能检索桌面的研究[J]. 成都理工大学学报: 自然科学版, 2003, 30(2): 214-216.
- [5] 陶跃华, 王锡钢, 王云爱. 信息检索向量空间模型中特征提取的研究[J]. 云南师范大学学报, 2000, 20(6): 18-20.