第33卷 第18期 计 算 机 工 程 2007年9月
Vol.33 No.18 Computer Engineering September 2007
·软件技术与数据库· 文章编号：1000—3428(2007)18—0040—02 文献标识码：A 中图分类号：TP311

基于模糊集的主题提取和层次发现算法

周红芳[1,2]，冯博琴[1]
(1. 西安交通大学电子与信息工程学院，西安 710049；2. 西安理工大学计算机学院，西安 710048)

摘 要：从语义相关性角度分析超链归纳主题搜索(HITS)算法，发现其产生主题漂移的原因在于页面被投影到错误的语义基上，提出了一种基于模糊集的主题提取和层次发现算法(FSTH)，通过用户日志扩展查询词，构造符合用户需要的个性化根集和基础集合，达到防止主题漂移的目的。FSTH采用模糊集划分方法，层次地发现与用户查询相关的主题页面集合，利用HITS算法分别计算每个主题页面集合中页面的权威值，返回与查询相关的其他主题权威页面。在14个查询上的实验结果表明，与HITS算法相比，FSTH算法不仅可以减少7%~53%的主题漂移率，而且可以发现与查询相关的多个主题。

关键词：模糊集；超链归纳主题搜索；主题提取；主题漂移；查询扩展

Algorithm for Topic Distillation and Hierarchical Exploration Based on New Fuzzy Set

ZHOU Hong-fang[1,2], FENG Bo-qin[1]
(1. School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049;
2. School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048)

【Abstract】To interpret the procedure of hypertext induced topic search based on a semantic relation model, the reason about the topic drift of HITS is found that Web pages are projected to a wrong latent semantic basis. A new fuzzy set based algorithm for topic distillation and hierarchical exploration (FSTH) is presented to improve the quality of topic distillation. Personalized root set and base set with query expansion is constructed using individual query logs to avoid the topic drift, and applying a hierarchical division algorithm based on fuzzy set to explore relative topics of user query, and then using HITS to evaluate and return authority pages of relative topics to end-users. The experimental results on 10 queries show that FSTH reduces topic drift rate by 7% to 53% compared to that of HITS, and discovers several relative topics to queries that have multiple meanings.

【Key words】fuzzy set; hypertext induced topic search; topic distillation; topic drift; query expansion

在Web信息检索中，很多研究者通过分析页面的链接结构来对返回页面进行排序过滤[1]，其中以HITS(hypertext induced topic search)算法最为著名，可得到协同引用最大的页面团体，但此算法在某些情况下会产生主题漂移。已有很多研究者试图对HITS算法进行改进[2~4]，但他们大多只关注了HITS的计算权威值过程，只能得到最普遍的主题，很难真正符合特定用户的需要。本文提出一种基于模糊集的主题提取和层次发现算法，借助于用户日志，利用迭代算法构造个性化根集和基础集合来改善链接结构分析中的主题漂移问题，并通过模糊集方法发现查询相关主题。

1 用于主题提取和层次发现的查询扩展技术

1.1 HITS算法基础

用户查询由查询词组成，可用向量表示。对于特定的用户查询 Q，HITS 算法首先利用搜索引擎得到包含 Q 的页面集合，构成根集合 R，然后添加所有指向根集合的页面和所有由根集合指向的页面，扩充得到强相关链接的基础集合 T。

假定基础集合 T 对应的图为 $G=(V,E)$，其中 V 为由所有页面组成的顶点集合，数目为 n，E 为所有有向边的集合。HITS 算法根据链接结构发现集合 T 中的重要页面。页面 i 的重要度根据权威值和中心值来确定，权威值基于页面 i 的入度，而中心值基于页面 i 的出度。V 中所有页面的权威值可表示为权威向量 $\boldsymbol{a}=(a_1,a_2,...,a_n)$，中心值可表示成中心向量 $\boldsymbol{h}=(h_1,h_2,...h_n)$，初始时所有的权威值和中心值均设置为 1，指向页面 i 的页面为 j，a_i 的值更新为所有 h_j 之和 $a_i^{(t+1)}=\sum_{j|j\to i}h_i^{(t)}$。若页面 i 指向的页面为 j，则 h_i 的值更新为所有 a_j 之和 $h_i^{(t+1)}=\sum_{j|i\to j}a_j^{(t)}$。经过多次迭代后收敛到不动点 a_i^* 和 h_i^*，则 a_i^* 为页面 i 的权威值，h_i^* 为页面 i 的中心值。若用 $n\times n$ 的邻接矩阵 $\boldsymbol{A}$ 表示图 G，$\boldsymbol{A}^{\mathrm{T}}$ 为矩阵 $\boldsymbol{A}$ 的转置，当 $i\to j$ 时，$A[i,j]=1$，否则 $A[i,j]=0$，那么上面的操作分别对应为：$\boldsymbol{a}=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{h}=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{a}$ 和 $\boldsymbol{h}=\boldsymbol{A}\boldsymbol{a}=\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{h}$。在每一步迭代后进行规范化操作，保持 $|\boldsymbol{a}|=|\boldsymbol{h}|=1$，则最后向量 $\boldsymbol{a}$ 和 $\boldsymbol{h}$ 分别收敛到矩阵 $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ 和 $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ 的主特征向量(即矩阵最大特征值对应的特征向量) $\boldsymbol{a}^*$ 和 $\boldsymbol{h}^*$。本文将 $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ 和 $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ 分别称为权威矩阵和中心矩阵。

1.2 用户查询扩展

相关反馈是信息检索领域中的一个有效的学习方法，最常用的就是 Rocchio 公式。

$$q_{k+1}=\hat{q}_k+\frac{\alpha}{|R|}\sum_{d_i\in R}\hat{d}_i-\frac{\beta}{|N|}\sum_{d_j\in N}\hat{d}_j \quad (1)$$

基金项目：国家“863”计划基金资助项目(2001AA113182)；陕西省教育厅2006年专项科学研究计划基金资助项目(06JK229)
作者简介：周红芳(1976－)，女，博士研究生，主研方向：数据仓库与数据挖掘，知识发现，粗糙集等；冯博琴，教授、博士生导师
收稿日期：2006-10-09 **E-mail：**zhouhf@xaut.edu.cn

其中，文档集合 R 和 N 分别表示搜索结果中的相关文档集合和不相关文档集合；在每次迭代后，q_k、d_i 和 d_j 都要进行规范化操作，并且 $\hat{q}_k$ 表示上一次检索时使用的查询，d_i 和 d_j 分别表示文档集合 R 和 N 中的文档；参数 α 和 β 分别表示正反馈参数和负反馈参数。在此，通过相关性判断可以将式(1)的迭代公式修改为

$$q_{k+1}=\hat{q}_k+\frac{\alpha}{\left|\cup_{G_n\in NC}G_n\right|}\sum_{G_n\in NC}\sum_{d_j G_n}d_j \tag{2}$$

其中，RC 和 NC 分别表示相关簇 G_r 和不相关簇 G_n 的集合；参数 α 和 β 由式(3)和式(4)给出：

$$\alpha=\begin{cases}1/\left(c_1^{\alpha}+c_2^{\alpha}\cdot p_{k,r}^{c}\right) & (p_{k,r}^{c}\le\alpha)\\ 2 & (p_{k,r}^{c}>\alpha)\end{cases} \tag{3}$$

$$\beta=\begin{cases}0.5 & (p_{k,n}^{c}\le b)\\ c_1^{\beta}+c_2^{\beta}\cdot p_{k,n}^{c} & (p_{k,n}^{c}>b)\end{cases} \tag{4}$$

在式(3)和式(4)中，$p_{k,r}^{c}=<q_k,s(\cup_{G_r\in RC})>$，$p_{k,r}^{c}=<q_k,s(\cup_{G_n\in NC})>$。$\cup_{G_r\in RC}$ 和 $\cup_{G_n\in NC}$ 分别表示 RC 中的文档集合 G_r 和 NC 中的文档集合 G_n 的全集。

1.3 基于模糊集发现相关主题

实际查询中，用户可能无法对自己感兴趣的内容进行准确描述，所以需要用模糊的思想来解决问题。

首先删除出现频率很高但很少与文档的内容有关的虚词，同时合并具有相同词根且含义相近的词；然后利用 FTIDF 方法计算词在每个文档 d_i 中的重要程度；根据阈值 T 选择一组词 W_i，满足 $\forall w\in W_i$，$tfidf(w,d_i)\ge T$；合并抽取的词汇 $W=\sum_{i=1}^{n}W_i$ 作为代表词汇。

而一个词对文档的重要性不仅取决于它本身的出现频率，还取决于它的子概念的出现频率以及子概念的隶属度。给定文档 d 和词 w，w 在 d 中出现的次数称为绝对频率，记作 $val(w,d)$。w 对祖先 v 的相对频率等于 $\beta*val(d,w)*belong(w,v)$，其中 $\beta\in[0,1]$ 是衰减因子。词的频率等于绝对频率和相对频率之和。矢量转换以 W 及其祖先为属性集合，属性在文档中出现的频率为属性值，经过标准化后映射为一个矢量，从而把文档数据库 D 转换为矢量数据 库 D'。

令 $D'=\{X_1,X_2,...,X_n\}$，$A_1,A_2,...,A_m$ 为 D' 的一组属性。从直观上看，两个文档包含的公共词越多，出现的频率越高，它们的相似程度就越高，距离也就越小，因此两个文档之间的距离用矢量的 Jaccard 系数

$$X,Y\in D',\quad d(X,Y)=1-\sum_{i=1}^{m}x_iy_i\Big/\left(\sum_{i=1}^{m}x_i^2+\sum_{i=1}^{m}y_i^2-\sum_{i=1}^{m}x_iy_i\right)$$

来衡量。采用模糊 K-Means 算法，对文档进行聚类。

2 算法描述

主题提取与层次发现算法(FSTH)从 2 个方面对 HITS 算法进行了改进：基于用户日志构造符合用户需要的个性化基础集合；采用层次划分方法发现与用户查询相关的主题。设选择的查询扩展词的个数为 m，根集合的页面数目为 n，选择的文本搜索引擎为 E，则构造个性化基础集合的算法如下。

算法 1 构造个性化基础集合(CPBS)

(1)基于个人查询历史，根据式(1)计算查询 $\hat{q}_k$。

(2)选择相邻两次查询的差小于预先制定的阈值 τ 时的查询作为扩展查询 Q'。

(3)利用扩展查询 Q' 在搜索引擎 E 中检索，选择最前面的 n 个页面构成个性化根集合 R'。

(4)对个性化根集进行链接扩展，得到基础集合 T'。

预先设定图划分算法的阈值为 S，则层次地发现相关主题的算法如下。

算法 2 相关主题层次发现算法(RTHD)

(1)随机地选取 k 个初始聚类中心，计算相应的隶属矩阵；

(2)修改当前的聚类中心，并重新计算隶属矩阵；

(3)如果新的聚类的目标函数值小于原先的聚类，则用新聚类代替原聚类；

(4)重复执行这一过程直至目标函数值不再减小为止。

迭代过程结束时得到的中心矩阵和隶属矩阵分别记作 $\boldsymbol{Z}^*$，$\boldsymbol{W}^*$。如果对象 X_i 与 Z_l^* 的距离不小于 Z_l^* 与其他聚类中心的最小距离或者 X_i 对 Z_l^* 的隶属度不大于最小阈值 λ，则从 Z_l^* 对应的聚类中删除 X_i，并重新计算聚类中心。

(5)每个聚类对应一个主题，用聚类的中心矢量表示。对每个主题运行 HITS 算法计算页面权威值，返回权威值最大的 k 个页面。

3 实验及分析

3.1 实验设计

分别采用 FSTH 算法和 HITS 算法进行实验。实验中，采用人工方法获得扩展查询，然后得到个性化的根集和基础集合。部分参数选择如下：根集页面的数目为前 300 个，扩展查询词选择前 15 个，最后返回的权威页面为前 50 个。一共设计了 10 个查询，如表 1 所示。

表 1 实验中使用的查询及其基础集合的大小

查询编号	查询词	基础集合页面数/个	
		HITS	PTHHE
1	Intelligence	3 265	2 653
2	Data mining	4 562	3 210
3	Knowledge discovery	6 253	4 986
4	Virus	5 103	4 350
5	Software	2 130	1 982
6	Hardware	2 210	1 230
7	Java	6 508	5 610
8	Programming	3 462	2 103
9	Computer	2 891	2 018
10	IC	2 237	1 205

3.2 实验结果及分析

将实验中的 FSTH 算法主特征向量对应权威页面集合与 HITS 算法权威页面集合进行比较，检测 2 种方法的主题漂移程度，实验结果如图 1 所示。

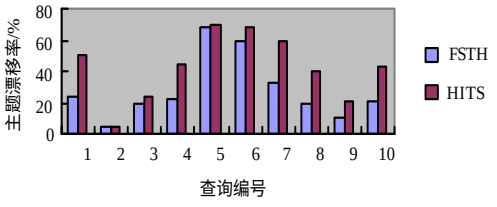

图 1 FSTH 算法与 HITS 算法的查询主题漂移率

从实验结果可以看出，对于语义比较确定的查询，如编号为 2、3、5、6 的查询不太容易出现主题漂移问题，而对于具有多种含义的查询，如编号为 1、7、8、10 的查询，利用用户的个性化信息可以过滤掉一些无关的页面，避免非常普遍但却与查询无关的主题添加到基础集合中。

在实验中通过为 FSTH 算法设置不同的阈值，可以得到不同层次的独立主题，以避免主题遗失。下面以查询 Virus 的实验结果为例进行说明，Virus 包含了多个不同主题，假定

(下转第 44 页)