

基于主题聚类的 Web 资源个性化推荐研究

张万山, 肖 瑶, 梁俊杰, 余敦辉

(湖北大学 计算机与信息工程学院, 湖北 武汉 430062)

摘 要: 提出了基于主题聚类的 Web 资源个性化推荐算法 PRWRTC(personalized recommendations of Web resource based on Topic clustering), 该算法首先基于 Web 资源的主题提出 Web 资源聚类算法, 从而提高 Web 资源聚类的精确度, 进而提升 Web 资源推荐的准确度; 然后, 基于用户的浏览行为, 提出实时获取用户偏好的算法; 最后, 针对用户偏好的动态演化, 在算法中加入了时效的概念, 实现了对 Web 资源的动态推荐. 并通过实验验证了该算法的有效性.

关键词: 语义网; 主题; 隐式跟踪; 个性化推荐

中图分类号: TP391.1

文献标识码: A

文章编号: 1000-7180(2015)04-0035-05

Personalized Recommendation of Web Resources Based on Topic Clustering

ZHANG Wan-shan, XIAO Yao, LIANG Jun-jie, YU Dun-hui

(School of Computer Science and Information Engineering, Hubei University, Wuhan 430062, China)

Abstract: The PRWRTC(personalized recommendations of Web resource based on Topic clustering) algorithm is proposed in this paper. The Web resources based on the theme of Web resource which can improve the accuracy of Web resources clustering, and can enhance the accuracy of recommended Web resources. On the basis of the user's browsing behavior, the algorithm that gets users' preferences in real-time is given out. According to the dynamic evolution of user preferences, the concept of efficiency is put into the algorithm, and the dynamic recommendation of Web resources realized. And experiments verify the effectiveness of the algorithm.

Key words: Semantic Web; theme; implicit tracking; personalized recommendations

1 引言

随着网络学习的普及, 网络 Web 资源突破传统 Web 资源在人员、地域、时空的多重限制, 而提供大量、全面、开放的资料. 但是, 对于大量的网络 Web 资源, 当前的搜索引擎没有考虑语义信息及用户的个性化特征, 主要是基于关键字匹配来查找并返回大量的超链接信息, 不能准确提供用户真正需要的 Web 资源, 造成了 Web 资源和人力资源严重浪费, 使用户的学习效率低下. 对用户而言, 希望系统能够主动推荐准确的、个性化的 Web 资源, 以提高学习、

工作的绩效.

Web 资源个性化推荐技术因此得到广泛的研究, 传统的基于协同过滤的 Web 资源个性化推荐方法, 是依据用户对资源的评分找到相似用户, 再将相似用户感兴趣的资源推荐给用户. 该方法存在稀疏矩阵和新用户问题. 文献[1-2]中对传统的协同过滤算法做出改进, 在一定程度上解决了传统基于协同过滤推荐算法中的稀疏矩阵问题. 但基于协同过滤的推荐算法没有考虑 Web 资源的语义信息, 导致 Web 资源聚类准确度不高, 进而影响推荐的准确度. 文献[3-5]中提出从语义层次对 Web 资源进行

收稿日期: 2014-06-13; 修回日期: 2014-08-08

基金项目: 国家自然科学基金项目(61272111, 61202031, 61273216, 61202032); 湖北省自然科学基金项目(2013CFB002, 2013CFA115); 武汉市科技攻关技术项目(201210621214, 201210421132)

推荐的方法,大大提高了 Web 资源推荐的准确度,此类算法考虑了 Web 资源的语义信息,但没有深入考虑其主题信息,对包含多主题的 Web 资源的推荐,该类算法的准确度有待提高。

为此本文提出了基于主题聚类的 Web 资源个性化推荐方法。首先,基于主题对 Web 资源进行聚类;然后,利用隐式跟踪的方式发现用户偏好;最后,将所发现的用户偏好与 Web 资源进行匹配,产生个性化 Web 资源推荐集,并根据用户偏好进行变更,实现 Web 资源动态推荐。

2 Web 资源聚类

Web 资源文档聚类首先要对 Web 资源文档进行预处理^[6],即利用已有的分词工具将 Web 资源文档分词,剔除其中的虚词和非名词分词,得到一个仅包含大量名词的集合,称为“主题词集 I ”。

在此基础上,本文首先基于 HowNet 抽取 Web 资源的主题集合;然后利用 TF*IDF 算法抽取每个主题中权重较大的主题词,从而降低每个主题的维度,以提高聚类算法的性能;最后基于主题对 Web 资源聚类。

2.1 基于 HowNet 的主题抽取

基于 HowNet 对主题词集 I 聚类,得到 n 个簇,计算各个簇的权重,选取权重最大的 k 个簇 $T_1, T_2 \dots T_k$ 作为 Web 资源文档 D 的主题集合 T ,具体实现流程见图 1。

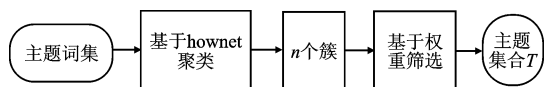


图 1 基于 HowNet 抽取主题 T

具体算法描述如下:

输入:主题词集 $I = \{(\omega_1, f_1), (\omega_2, f_2) \dots (\omega_n, f_n)\}$, 其中, ω_i 是主题词, f_i 是主题词 ω_i 的词频。

输出:Web 资源文档 D 的主题 $T = \{T_1, T_2 \dots T_k\}$ 。

算法:

(1) 基于 HowNet, 计算主题词集 I 中 $\omega_1, \omega_2 \dots \omega_n$ 的两两相似度:

$$s_{ij} = \text{sim}(\omega_i, \omega_j) \quad (1)$$

式中, $\text{sim}(\omega_i, \omega_j)$ 是主题词 ω_i, ω_j 基于 HowNet 的相似度, $1 \leq i \leq n, 1 \leq j \leq n, i \neq j$ 。取 $s_{ij} (s_{ij} \geq \alpha, \alpha$ 相似度阈值), 组成集合 S

$$S = \bigcup s_{ij} \quad (2)$$

(2) 基于相似度,借助 kruskal 算法思想对 $\omega_1,$

$\omega_2 \dots \omega_n$ 聚类。具体算法如下:

① 从主题词集 I 中提取主题词 ω_i , 得新的集合 $I' = \{\omega_1, \omega_2 \dots \omega_n\}$ 。

② 从 S 中选择最大的 s_{ij} , 记为 $\max s_{ij}$ 。从 I' 中选择 ω_i, ω_j (其中, $\max s_{ij} = \text{sim}(\omega_i, \omega_j)$) 并将 ω_i, ω_j 聚类。

③ 从 S 中删除 $\max s_{ij}$, 从 I' 中删除 ω_i, ω_j 。

④ 从 I' 中选择 ω_k , 如果 $s_{ik} \in S$ 或 $s_{jk} \in S$, 则将 ω_k 与 ω_i, ω_j 聚类。其中 $s_{ik} = \text{sim}(\omega_i, \omega_k), s_{jk} = \text{sim}(\omega_j, \omega_k)$ 。

⑤ 从 I' 中删除 ω_k , 从 S 中删除 s_{ik}, s_{jk} 。

(3) 重复步骤(2), 直到 I' 为空。最终得到 n 个簇, 称为 n 个“主题”, 记为 $T_1, T_2 \dots T_n$ 。其中, $T_r \subset I' (1 \leq r \leq n)$, 对任意 $\omega_i, \omega_j \in T_r, \text{sim}(\omega_i, \omega_j) \geq \alpha$ 。

(4) 计算主题 T_r 的权重, 记为

$$W(T_r) = \frac{\sum_{T_r} f_i}{\sum_i f_i} \quad (3)$$

式中, $\sum_{T_r} f_i$ 是主题 T_r 中所有词频之和, $\sum_i f_i$ 是 Web 资源文档 D 的主题词集 I 中所有词频之和。

(5) 取所有 $W(T_r) \geq \beta$ (β 为权重的阈值) 的主题 T_r , 得到 Web 资源文档 D 的主题集合:

$$T = \{T_1, T_2 \dots T_k\} \quad (k \leq n) \quad (4)$$

2.2 基于 TF*IDF 的特征词集提取

基于 TF*IDF 计算主题集合 $T = \{T_1, T_2 \dots T_k\}$ 中主题词的权重, 分别抽取主题 $T_1, T_2 \dots T_k$ 中权重最大的若干个主题词组成 Web 资源文档 D 的特征词集 M , 实现流程如图 2 所示。



图 2 基于 TF*IDF 的特征词集提取

具体算法描述如下:

(1) 对各个主题 T_r 权重 $W(T_r)$ 归一化, 记为

$$W_n(T_r) = \frac{W(T_r)}{\sqrt{\sum_{T_r \in T} W^2(T_r)}} \quad (5)$$

(2) 计算主题 T_r 中应该抽取的特征词个数 m_r

$$m_r = |M| * W_n(T_r) \quad (6)$$

式中, $|M|$ 是特征词集 M 的个数, $W_n(T_r)$ 是主题 T_r 的归一化权重。

(3) 计算主题 T_r 中主题词 ω_i 的权重 $f(\omega_i)$

$$f(w_i) = V(w_i) \times \lg\left(\frac{N}{N_k} + \lambda\right) \quad (7)$$

式中, w_i 表示主题 T_r 中的主题词, $V(w_i)$ 表示主题词 w_i 的词频, N 表示 Web 资源文档训练集的个数, N_k 表示包含主题词 w_i 的 Web 资源文档的个数, λ 是权重的调节因子, 本文中其值为 0.01.

(4) 分别将各主题 T_r 中的主题词按权重 $f(w_i)$ 降序排列, 抽取前 m_r 个主题词作为特征词.

(5) 将抽取的特征词取并集得到 Web 资源文档 D 的特征词集, 记为

$$M = \bigcup_{i=1}^{m_1} w_i \bigcup_{j=1}^{m_2} w_j \cdots \bigcup_{k=1}^{m_k} w_k \quad (8)$$

式中, $\bigcup_{i=1}^{m_1} w_i$ 表示从主题 T_1 中抽取的 m_1 个主题词, $\bigcup_{j=1}^{m_2} w_j \cdots \bigcup_{k=1}^{m_k} w_k$ 分别表示从主题 $T_2 \cdots T_k$ 中抽取的 $m_2 \cdots m_k$ 个主题词.

2.3 基于 HowNet 的 Web 资源聚类

(1) Web 资源表示模型. 相较于单个特征词 w_i 的权重, 主题 T_r 的权重对 Web 资源文档的聚类和推荐产生的影响更大, 故本文在表示 Web 资源文档 D 时采用了主题对应的权重. Web 资源文档 D 表示如下:

$$D = \bigcup_{w_i \in M} (w_i, W_n(T_n)) \quad (9)$$

式中, w_i 表示 Web 资源文档 D 的特征词, $W_n(T_n)$ 表示特征词 w_i 所属的主题 T_n 的归一化权重.

(2) 计算两 Web 资源文档 D_1, D_2 的相似度:

$$\text{sim}(D_1, D_2) = \frac{\sum \text{sim}(w_i, w_j) \times W_{1n}(T_i) \times W_{2n}(T_j)}{\sum \text{sim}(w_i, w_j)} \quad (10)$$

式中, $\text{sim}(w_i, w_j)$ 是 Web 资源文档 D_1, D_2 中任意两个特征词 w_i, w_j 基于 HowNet 的相似度, $W_{1n}(T_i)$ 是 Web 资源文档 D_1 的特征词 w_i 所对应主题 T_i 的归一化权重, $W_{2n}(T_j)$ 是 Web 资源文档 D_2 的特征词 w_j 所对应的主题 T_j 所对应的归一化权重.

(3) 对文档进行聚类. 如果 $\text{sim}(D_1, D_2) \geq \gamma$ (γ 是相似度阈值), 则将 Web 资源文档 D_1, D_2 聚类, 记为

$$C = \bigcup_{w_i \in D_1 \cup D_2} (w_i, W_n(T_i)) \quad (11)$$

从式(11)可知, 随着多个 Web 资源聚类, 聚类 C 中的特征词个数会增多, 聚类特征向量维数会增大. 采用上述特征词提取算法提取聚类的特征词, 从而降低了聚类的特征向量维数.

3 用户个性化偏好捕获

捕获用户的偏好可以采用显式跟踪和隐式跟踪的方法^[7-8]. 显式跟踪方法就是通过用户的评价信息直接获取用户的偏好, 但是获取用户评价信息, 对用户的浏览行为会产生影响; 隐式跟踪方法是在服务站点观察用户的浏览行为, 通过这些行为推测出用户的偏好, 隐式跟踪方法不会影响用户的浏览行为, 获取的用户偏好更客观.

隐式跟踪方法获取用户偏好, 包括两个过程.

(1) 通过用户的浏览行为获取用户感兴趣的 Web 资源页面. 用户的浏览行为有很多, 本文只考虑常用的、对用户兴趣有明确指示作用的行为. 这些行为包括下载、收藏、加入书签、复制、浏览时间、拖动滚动条次数、点击次数等^[9-10].

从 Web 资源页面的特征考虑, 下载、收藏、加入书签、复制等行为可以独立反映用户对 Web 资源页面的偏好程度, 其偏好程度用权值 C_v 表示; 浏览时间、拖动滚动条、点击次数等行为必须综合考虑才能反映用户对 Web 资源页面的偏好程度, 其偏好程度用 F_1 表示:

$$F_1 = \frac{N_c}{L_{pi}} \times e^{-|T_c - \text{MAX}|} \quad (12)$$

式中: N_c 是拖动滚动条次数与点击次数之和, L_{pi} 表示 Web 资源页面的长度, T_c 是 Web 资源页面浏览时间, MAX 是 Web 资源页面的合理浏览时间, 如果浏览时间 T_c 大于这个合理浏览时间 MAX , 用户很可能没有正常浏览页面(或已离开), 从而浏览时间 T_c 对用户偏好的指示作用降低.

用户对 Web 资源页面 p_i 的偏好程度 I_{pi} , 记为

$$I_{pi} = \lg F \quad (13)$$

式中, $F = \min\{10, F_1 + F_2\}$, F_1 见式(12), $F_2 = \sum_{v \in V} C_v \times f_v(p_i)$, 其中, $V = \{\text{下载, 收藏, 加入书签, 复制}\}$, $f_v(p_i)$ 是一个二值函数, 如果用户对 Web 资源页面 p_i 有 v 行为, $f_v(p_i)$ 的值取 1, 否则取 0.

若 $I_{pi} \geq \theta$ (θ 为阈值), 则将页面 p_i 加入页面集 P .

(2) 通过用户感兴趣的 Web 资源页面获取用户的偏好. 首先, 对页面集 P 中的 Web 资源页面进行预处理, 提取 Web 资源页面的文本信息; 然后, 利用本文上述特征词提取和 Web 资源聚类的方法, 提取文本信息的特征词集并聚类, 获取用户兴趣偏好:

$$U = \bigcup (w_j, W_n(T_j)) \quad (14)$$

式中, w_j 表示用户兴趣特征词, $W_n(T_j)$ 表示用户兴趣特征词 w_j 所对应主题的归一化权重.

4 基于资源与偏好匹配度的 Web 资源推荐

首先,基于 HowNet 计算 Web 资源文档聚类 C 与用户偏好的 U 相似度:

$$\text{sim}(C, U) = \frac{\sum \text{sim}(\omega_i, \omega_j) \times W_n(T_i) \times W_n(T_j)}{\sum \text{sim}(\omega_i, \omega_j)} \quad (15)$$

式中, $\text{sim}(\omega_i, \omega_j)$ Web 资源文档聚类 C 和用户偏好 U 中任意两个特征词 ω_i, ω_j 基于 HowNet 的相似度, $W_n(T_i)$ 是特征词 ω_i 所对应主题 T_i 的归一化权重, $W_n(T_j)$ 是特征词 ω_j 所对应的主题 T_j 所对应的归一权重。

然后,取相似度最大的 Web 资源聚,将该聚类中与用户偏好最匹配的前 k 个 Web 资源文档推荐给用户。

5 基于用户偏好变更的动态推荐

用户的兴趣往往会随时间的迁移发生变化^[11-12],最近访问的 Web 资源更能指示用户的偏好,而很早以前所访问的 Web 资源对用户兴趣的表征度相对更低,为此,借助于 LRU(最久未访问资源被淘汰)算法,根据用户访问 Web 资源的先后时间动态捕获用户的偏好兴趣,每隔一个时间周期 τ ,就更新现有的用户偏好信息 U 。具体实现方法如下:

(1) 对旧用户偏好信息 U_o 的权值进行衰减,即:

$$U_o = \bigcup (\omega_j, W_n(T_j) \times \eta) \quad (16)$$

式中, $0 \leq \eta \leq 1$ 。

(2) 利用上述获取用户偏好的算法获取最近一个周期 τ 内用户的偏好信息:

$$U_r = \bigcup (\omega_k, W_n(T_k)) \quad (17)$$

(3) 将 U_r, U_o 合并。对任意 $\omega_j \in U_o$, 如果 U_r 中存在 $\omega_k = \omega_j$, 则 ω_k 的权值为 $W_n(T_k) = W_n(T_k) + W_n(T_j) \times \eta$, 否则, $U_r = U_r \cup (\omega_j, W_n(T_j))$, 直到合并完成。

(4) 将合并后的 U_r 权重归一化,取权重最大的前若干个特征词—权重对,作为新的当前用户偏好信息 U 。根据新的用户偏好信息 U ,再次利用本文上述 Web 资源推荐算法对用户进行 Web 资源推荐,从而实现 Web 资源的动态推荐。

6 实验结果与分析

为了验证 PRWRTC 算法的有效性,设计了实

验,并选择基于 k-means 的推荐算法和基于语义词典的推荐算法(HowNet)与之性能进行比较。

在本文的实验中,随机从 Google 网站选择计算机领域 7 类共 10 000 篇文档作为测试语料,分别进行五组实验,实验中语料情况如表 1 所示。

表 1 测试语料概况

	T1	T2	T3	T4	T5
类别数	7	7	7	7	7
文档数量	1 000	3 000	5 000	7 000	10 000
最少文档数	100	300	600	800	1 200
最多文档数	358	534	861	1 119	1 495

试验中文本的特征向量维度为 50。使用准确率和召回率对 PRWRTC 算法进行评价,准确率和召回率的定义如下^[13]:

召回率: $\text{Recall}(i, r) = n(i, r) / n_i$,

准确率: $\text{Precision}(i, r) = n(i, r) / n_r$ 。

式中, $n(i, r)$ 是聚类 r 中包含的类别 i 中的文档个数, n_r 是聚类 r 中文档的个数, n_i 是预定义类别 i 的文档的个数。

依据实验可得如图 3、图 4。

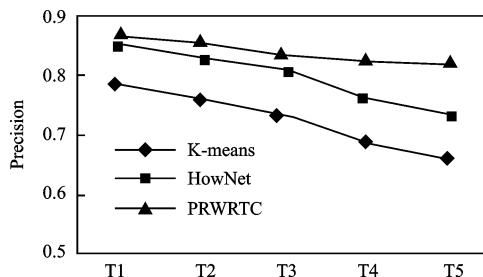


图 3 三类算法准确率比较

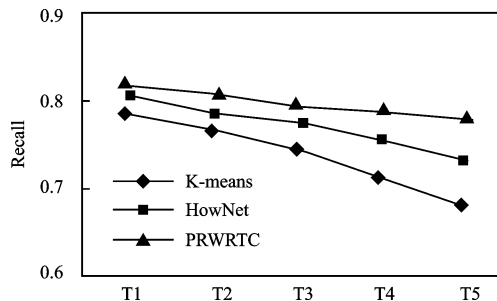


图 4 三类算法召回率比较

实验结果表明,基于主题聚类的 Web 资源推荐算法(PRWRTC)相比其他两类算法,准确率和召回率都有所提高;随着资源数量的增加,三类算法的准确率和召回率都呈下降趋势,PRWRTC 算法的下降速度比其他两类算法缓慢,即该算法的随着资源数量的增加,PRWRTC 算法的稳定性更高。

7 结束语

本文提出了基于主题聚类的Web资源个性化推荐方法.该方法在推荐过程中考虑了Web资源的主题信息,并基于主题信息对Web资源聚类;然后,基于用户的浏览行为,隐式获取用户的兴趣,考虑到用户兴趣自身特点,在算法中加入时效的概念,真正动态获取用户偏好,在此基础上,实现Web资源的动态推荐;最后通过实验验证了该算法的有效性和稳定性.本文提出的方法局限于文本型Web资源的推荐,对音频、视频等Web资源的推荐还需要进一步研究.

参考文献:

- [1] Mojtaba Salehi. Application of implicit and explicit attribute based collaborative filtering and BIDE for Web resource recommendation[J]. Data & Knowledge Engineering, 2013, 87(9): 130-145.
- [2] Manouselis N, Vuorikari R, Van Assche F. Collaborative recommendation of e-Web resources: an experimental investigation[J]. Journal of Computer Assisted Web, 2010, 5(3/4): 347-370.
- [3] Li Dun, Ma Yongtao, Guo Jian li. Words semantic orientation classification based on HowNet[J]. The Journal of China University of Post and Telecommunications, 2009, 16(1): 106-110.
- [4] Holanda Olavo, Ferreira Rafael, Costa Evandro. Educational resources recommendation system based on agents and semantic web for helping students in a virtual Web environment[J]. International Journal of Web Based Communities, 2012, 8(1): 333-353.
- [5] 刘志勇, 刘磊, 刘萍萍,等. 一种基于语义网的个性化Web资源推荐算法[J]. 吉林大学学报:工学版, 2009, 39(2): 391-395.
- [6] Perugini Saverio, Marcos Andre Goncalves, Edward A Fox. Recommender systems research: a connection-centric survey[J]. Journal of Intelligent Information Systems, 2004, 23(2): 107-143.
- [7] Matias Nicoletti, Silvia Schiaffino, Daniela Godoy. Mining interests for user profiling in electronic conversations[J]. Expert Systems with Applications, 2013, 40(2): 638-645.
- [8] Li Jun, Zhang Peng. Mining explainable user interests from scalable user behavior data[J]. Procedia Computer Science, 2013(17): 789-796.
- [9] Liu Xiaozhong, Turtle Howard. Real-time user interest modeling for real-time ranking[J]. Journal of the American Society for Information Science and Technology, 2013, 64(8): 1557-1576.
- [10] Li Zhengwei, Xia Shixiong, Niu Qiang. Research on the uUser interest modeling of personalized search engine[J]. 武汉大学学报:自然科学版,英文版, 2007, 53(5): 893-896.
- [11] 尹春晖, 邓伟. 基于用户浏览行为分析的用户兴趣获取[J]. 计算机技术与发展, 2008, 18(5): 37-39.
- [12] Cheng Chen, Gui Pingliao, Xiao Huishi, et al. The model of user interest update and user classification in personal information push service[J]. Procedia Environmental Sciences, 2010, 2(10): 262-268.
- [13] Pluis Martinez, Luis G Perez, Manuel Barranco. A multigranular linguistic content-based recommendation model: research articles[J]. International Journal of Intelligent Systems, 2007, 22(5): 419-434.

作者简介:

张万山 男, (1973-), 硕士, 讲师. 研究方向为 web 信息挖掘.

肖 瑶(通讯作者) 女, (1987-), 硕士研究生. 研究方向为 web 信息挖掘. E-mail: hubuxiaoyao@163.com.

梁俊杰 女, (1974-), 博士, 副教授. 研究方向为数据分析与云计算.

余敦辉 男, (1974-), 博士后, 副教授. 研究方向为服务计算与大数据.

(上接第34页)

- [9] Kislyakov M, Mosin S. A processor development in-programmable logic basis [C]// The Experience of Designing and Application of CAD Systems in Microelectronics-CADSM 2007. Lviv-Polyana, Ukraine, 2007: 182-185.
- [10] 冯海涛, 王永纲, 石江涛, 等. 基于FPGA的32位整数微处理器的设计与实现[J]. 小型微型计算机系统, 2005, 26(6): 1113-1117.

作者简介:

孙巧稚 女, (1988-), 硕士研究生. 研究方向为计算机系统结构. E-mail: 404914613@qq.com.

施慧彬 男, (1966-), 博士, 副教授. 研究方向为计算机体系结构、多核微处理器设计、可重构计算、多核技术、嵌入式系统、编译技术、软件工程.