

在线社会网络中近似网页识别方法研究

张 莉¹, 孙丽娜², 郭 峰¹

(1 郑州成功财经学院 信息工程系, 河南 巩义 451200; 2 河南大学 民生学院, 河南 开封 475001)

摘 要: 为了提高用户的网络体验和页面检索性能, 提出一种基于核心词汇和语义匹配相结合的在线社会网络近似网页识别方法研究. 首先从网页中提取特征核心词汇, 对文档概念进行分类和合并, 并建立相似度特征模型; 基于相似度特征模型, 构建特征向量, 并对特征向量进行检索和存储, 实现社会网络中近似网页的识别. 实验证明提出的方法能够有效地降低网络噪声, 具有良好的识别准确率和召回率.

关键词: 在线社会网络; 近似网页; 识别方法

中图分类号: TP391

文献标识码: A

文章编号: 1000—7180(2017)02—0141—04

Research on Approximate Web Page Identification Method in Online Social Network

ZHANG Li¹, SUN Li-Na², GUO Feng¹

(1 Department of Information Engineering, Zhengzhou Chenggong University of Finance and Economics, Gongyi 451200, China; 2 Minsheng College, Henan University, Kaifeng 475001, China)

Abstract: In order to improve the user's Web experience and page retrieval performance, a new method based on key words and semantic matching for online social network is proposed. The first feature extraction of keywords from web pages, classify and merge of document concept, and establish the similarity model; similarity model based on feature vector is constructed, and the retrieval and storage of feature vector, identify the approximate pages in social networks. The experimental results show that the proposed method can effectively reduce the network noise, and has good recognition accuracy and recall rate.

Key words: online social network; approximate web page; identification method

1 引言

近年来,随着我国通信技术的进步和光纤网络布局的逐步完善^[1-2],互联网得到了飞速的发展,已经深深地融入了社会生活. 在线社会网络^[3-4]是信息加工、制造、处理和发布的重要平台. 但在社会网络中存在大量转载网页,即同样的一篇网页文章可能会以相同或近似的形式出现在其他网页中^[5-6]. 据相关网络管理部门统计得出,目前在线社会网络中,内容相同或近似的网页约占三成以上. 网民利用各种搜索引擎在搜索自

己所需要的网页的过程中,如果搜索相同的主题,就会获取大量内容相同或近似的网页,严重降低搜索引擎的效率,大大占用网络内存,而且浪费大量的网络资源. 另一方面,也给网民的阅读和搜索带来困扰^[7-9]. 这些相同或近似的重复网页一般来自网站间的转载和一些恶意的镜像网站,网页的内容都是一些比较热门的话题和评论. 由此可见,目前用户使用搜索引擎检索网页的最大困扰是重复信息过多,在线社会网络中,对内容近似或相同的网页进行识别和处理具有重要的现实意义^[10-11]. 传统的基于字符串的网页去重方法^[12-13],

收稿日期: 2016—05—07; 修回日期: 2016—06—22

基金项目: 河南省高等学校重点科研项目(教科技[2015]1120 号)“现代化跟踪系统目标成像轨迹模拟研究”(16A520093); 河南省科技厅科技攻关项目“基于 WiFi 的无线存储测试系统设计”(162102210367); 河南省科技攻关项目“基于机器人视觉下的运动目标成像特征提取技术研究”

可实现性强、算法简洁,面对在线社会网络,算法有失效的风险;基于文本结构的方法,尽管能够降低结构布局的复杂度和网络噪声,但计算量大、方法复杂、可通用性低。

本文提出一种基于关键词和语义匹配的在线社会网络近似网页识别方法研究。

2 在线社会网络中近似网页识别方法研究

首先对在线社会网络中的近似网页进行定义,假定某在线社会网络的网页中有两篇文章, X 和 Y 分别表示这两篇文章的段落集合,将 X 和 Y 做交集运算,那么 X 和 Y 的交集 Z 可以表示为

$$Z = X \cap Y \quad (1)$$

如果这两篇文章有近似或重复的段落时, Z 的取值情况有三种可能性,如图 1 所示。

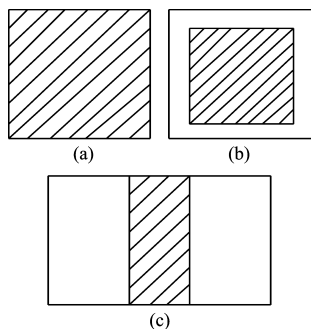


图 1 近似网页运算结果分类

在图 1 中斜线代表的阴影部分,表示 X 和 Y 的交集 Z 可能的取值范围。图 1(a) 表示两篇文章的内容完全一样,这种现象最为常见。如果网页报道或评论的内容较短,而又被大量地转载时,就会出现图 1(a) 所示的情况。

图 1(b) 表示 Y 是 X 的一部分,如果 X 的文章属于一个长篇的评论或新闻报道, Y 转载了 X 文章中的某一部分或多个部分。

图 1(c) 表示 X 和 Y 文章的内容有交叉,即文章中有部分内容一致,有部分内容不同,这种情况出现的原因有多种,如转载时网络噪声过大引起的失真;转载时相关人员对文章做了部分修改,但存在相同的部分,修改的不彻底;还有就是对同一事件的不同解读。视不同的情况而言,图 1(c) 所示的情况有可能转化为图 1(a) 或 (b) 所示的情况,有时候这 3 种情况相差不多,较难区分。

2.1 核心词汇的提取与特征相似度模型的构建

在线社会网络中对近似的网页进行识别和检索,首先要提取网页文章中的核心词汇,并对其进行

语义匹配,构建关键词的特征相似度模型。

设在线社会网络中的网页文章的核心词汇为 ω ,分名词类词汇为 n ,动词类词汇为 v ,时间类词汇 t ,再对以上的词汇大类做进一步分类,将名词类词汇进一步分为机构团体类、地名类、人名类和专有名词类,分别用 n_1, n_2, n_3 和 n_4 表示,除以上四种名词外的其他名词词汇使用 n_5 来表示;动词类词汇中将高频使用词汇“有”和“是”分离出来,“有”用 v_1 表示,“是”用 v_2 来表示,其他的动词词汇用 v_3 来表示。那么名词 n 的集合可以表示为

$$\begin{aligned} & \{n | n \in (n_1 \cup n_2 \cup n_3 \cup n_4 \cup \{v' \in n_5\})\} \\ & \{v' \in n \wedge v' \notin (n_1 \cup n_2 \cup n_3 \cup n_4)\} \end{aligned} \quad (2)$$

在线社会网络中的网页文章的核心词汇 ω 的取值范围为

$$\begin{aligned} & \{\omega \in C | t \cup n \cup \{v' \in v_3\}\} \\ & \{v' \in v \wedge v' \neq v_1 \wedge v' \neq v_2\} \end{aligned} \quad (3)$$

式中, C 代表全部词汇, v' 包括 v_1 和 v_2 ,由式(2)、(3)可知,核心词汇所分布的词性类集合分别为 $n_1, n_2, n_3, n_4, n_5, t$ 和 v_3 等 7 类。

如果对全部 7 个元素的集合,综合考虑,会出现两个问题:其中一个词汇量过于庞大,加重了后续评价系统的压力;另外一个在 n_5 和 v_3 的集合中涉及词汇范围广泛,无法突出特色词汇。基于对上述问题的判断,用词频作为参考依据,分别在 7 个集合中选取在原文中词频最高的词汇认定为本集合的核心词汇。

设词频函数为 $f(x), x \in T, T$ 的取值范围为 $\{n_1, n_2, n_3, n_4, n_5, v_3, t\}$,那么核心词汇的范围会进一步缩小为

$$\left\{ \omega \left| \begin{array}{l} \max(t) \cup \max(n_1) \cup \max(n_2) \cup \max(n_3) \\ \max(n_4) \cup \max(n_5) \cup \max(v_3) \end{array} \right. \right\} \quad (4)$$

提取核心词汇后要对核心词汇和文档归类合并,并做出系统评价和识别,近似网页的识别流程图,如图 2 所示。

假设 $T_i = T_i = \{t_i, n_{1i}, n_{2i}, n_{3i}, n_{4i}, n_{5i}, v_{3i}\}$, T_i 为非空集合,另一个非空集合 T_j 与 T_i 同为特定集合。如果存在一个参数 η ,且满足条件 $\eta \in \max(T_i), \max(T_j)$,那么认为集合 T_i 和 T_j 互为参考。

设 m 为集合 T_i 和 T_j 互为参考的个数, $i \neq j, T$ 为非空集合, ζ 为识别判定系数,当以下条件成立时:

$$\frac{m}{k \times \zeta} \geq 1, (0 < \zeta \leq 1) \quad (5)$$

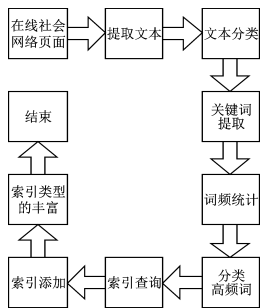


图2 近似网页的识别流程

可以认为 i 和 j 可互为参考. 由此可知, 判定网页内容近似与否, 需要按词性类别分类. 如果有有效的词性类别数与参考性的词性类别数达到一定比例的时候, 可断定在线社会网络的网页文章内容存在近似或重复的现象.

如果同一词性类别中有一个元素相同时, 那么就判定这一词性类别具有参考性. 提取关键词来组建完成互为参考的特征相似度模型后, 继续构建特征向量, 并在在线社会网络中, 对特征向量进行检索和存储, 实现社会网络中近似网页的识别.

2.2 在线社会网络中近似网页识别的实现

提取在线社会网络中近似网页的关键词汇后, 引入聚度和节点平均路径长度的概念, 以此来构建特征向量, 提高在线社会网络中文档相似度的检测水平, 并能够降低网络噪声.

聚度能够反映局部网络结构的特征, 为网络中两两相连的比例. 对于网络节点 P 而言, 假定存在 p_i 个邻近节点, 那么该节点与其邻近节点最多可以组成 $\frac{p_i \times (p_i - 1)}{2}$ 条边线, λ 为节点间实际存在的边数. 节点 P 的聚度 ξ 可以表示为

$$\xi_i = \frac{\lambda_i}{k_i \times (k_i - 1) / 2} \quad (6)$$

整个在线社会网络中的聚度 ξ' , 可以描述为

$$\xi' = \frac{1}{k_i} \sum_{i=1}^n \xi_i \quad (7)$$

平均路径长度为在线社会网络中, 点与点之间的最短距离. 设两点间的平均路径为 d_i , 那么两点间的平均路径 d_i 及整个在线社会网络的平均路径 d' 分别为

$$d_i = \frac{1}{k_i - 1} \left(\sum_{i,j=1}^n d_{\min}(i, j) \right) \quad (8)$$

基于概念语义匹配的两个基本特征, 聚度和路径, 对关键网络节点进行识别, 并构建特征向量. 可以利用特征向量表, 用来记录关键文档的位置. 使用

向量表 Q_i 来记录特征向量的位置, 其中 $q_{i,j}$ 为第 i 个关键词在第 j 次出现的位置:

$$Q_k = \begin{bmatrix} Q_1 \\ Q_2 \\ \dots \\ Q_k \end{bmatrix} = \begin{bmatrix} q_{11} & q_{12} & \dots & q_{1j} \\ q_{21} & q_{22} & \dots & q_{2j} \\ \dots & \dots & \dots & \dots \\ q_{i1} & q_{i2} & \dots & q_{ij} \end{bmatrix} \quad (9)$$

特征向量代表一组网页概念及所处的位置特征. 如果搜索到前面多维特征向量, 计算其相似度, 就能够识别出文档是否近似, 相似度 ζ 计算过程为

$$\begin{aligned} \zeta &= \frac{d_1 \times d_2}{\|d_1 \times d_2\|} \\ &= \frac{\sum_{i=1}^m d_1(i) \times d_2(i)}{\sqrt{\sum_{i=1}^m d_1(i)^2} \times \sqrt{\sum_{i=1}^m d_2(i)^2}} \end{aligned} \quad (10)$$

在线社会网络中, 两个网页特征项距离矩阵表示:

$$Q_k - Q_k' = \begin{bmatrix} q_{11} - q'_{11} & q_{12} - q'_{12} & \dots & q_{1j} - q'_{1j} \\ q_{21} - q'_{21} & q_{22} - q'_{22} & \dots & q_{2j} - q'_{2j} \\ \dots & \dots & \dots & \dots \\ q_{i1} - q'_{i1} & q_{i2} - q'_{i2} & \dots & q_{ij} - q'_{ij} \end{bmatrix} \quad (11)$$

式中, $q_{i,j}$ 为首篇文章, 即第 i 关键词在第 j 次出现的位置; $q'_{i,j}$ 为第 i 关键词在第 j 次出现的位置, 两者之差用 τ 表示, $\tau = q_{ij} - q'_{ij}$. 第 i 个概念的平均距离为

$$\text{avg}_i = \frac{\sum_{i=1}^k \tau q_{ij}}{k} \quad (12)$$

使用方差 σ_i 求解关键词的分布, 求解过程为

$$\sigma_i = \frac{\sqrt{\sum_{i=1}^k \tau q_{ij} - \text{avg}_i}}{k} \quad (13)$$

在线网络中完整网页的整篇文章关键核心词汇的距离分布可以表示为

$$\sigma = \frac{\sum_{i=1}^n \sigma_i}{n} \quad (14)$$

通过对特征向量进行检索和存储, 确定文章核心词汇距离的分布, 完成在线社会网络中近似网页的识别.

3 实验结果与分析

为了验证本文提出的基于核心词汇和语义匹配相结合的在线社会网络近似网页识别方法的正确性和召回率, 本文设计了一系列仿真实验, 分别采用基

于字符串的去重方法和本文方法,进行在线社会网络近似网页的识别效果对比.首先,在两个学科领域分别采用传统方法和本文方法进行近似网页的识别准确率效果对比,如图 3 和图 4 所示.

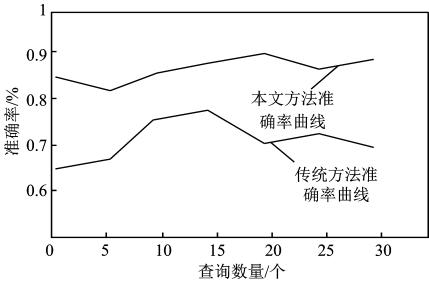


图 3 经济学学科两种方法下的识别准确率曲线对比

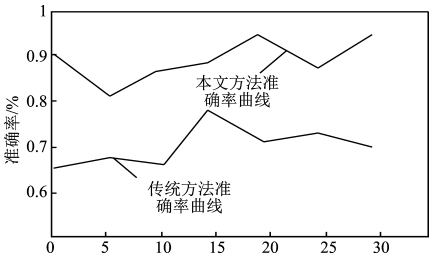


图 4 法学学科两种方法下的识别准确率曲线对比

对经济学学科和法学学科的近似网页进行识别,从两种方法的准确率曲线来看,本文提出的方法的准确率均可以达到 80% 以上,曲线的波动情况相对于传统方法也比较平缓.

利用这两种方法,对经济学学科和法学学科的近似网页召回率曲线进行比较,如图 5 和图 6 所示.

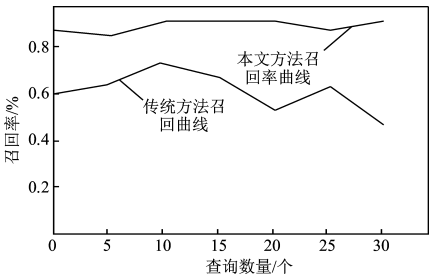


图 5 经济学学科两种方法下的召回率曲线对比

从图 5 和图 6 也可以看出,本文方法无论是在经济学学科还是法学学科,其召回率方法都具有较大优势.

4 结束语

本文提出一种基于核心词汇和语义匹配相结合的在线社会网络近似网页识别方法研究,能够提高近似网页识别的准确率和召回率,实验证明了本文

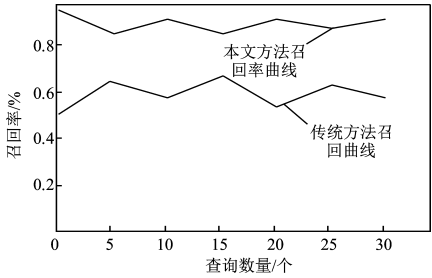


图 6 法学学科两种方法下的召回率曲线对比方法的有效性.

参考文献:

[1] 李栋,徐志明,李生,等. 在线社会网络中信息扩散[J]. 计算机学报, 2014, 37(1):189-206.

[2] 黄蓝会. 在线社会网络的数据采集方法的研究[J]. 信息技术, 2014(2):53-55.

[3] 蔡红云,马晓雪. 在线社会网络中基于关系强度的访问控制机制[J]. 山东大学学报(理学版), 2016, 51(7):34-55.

[4] 王栋,李振宇,谢高岗. 在线社会网络无偏采样技术[J]. 计算机研究与发展, 2016, 53(5):16-20.

[5] 李献礼. 在线社会网络多维度链式信任计算算法[J]. 西南大学学报(自然科学版), 2016(8):26-30.

[6] 袁圆. 基于内嵌字幕的近似视频网页去重研究与实现[D]. 北京:北京工业大学, 2015.

[7] 沙泓州,刘庆云,柳厅文,等. 恶意网页识别研究综述[J]. 计算机学报, 2016, 39(3):529-542.

[8] 孙汇中. 网页木马检测关键技术研究[D]. 北京:北方工业大学, 2016.

[9] 杨洪娇. 基于机器学习的校园网恶意网页检测方法[J]. 信息与电脑, 2016, 28(11):46-50.

[10] 卢晓勇,陈木生. 基于随机森林和欠采样集成的垃圾网页检测[J]. 计算机应用, 2016, 36(3):731-734.

[11] 问延安,赵婷婷. 基于 CSSCI 的国内社会网络研究可视化分析[J]. 潍坊工程职业学院学报, 2016, 29(1): 67-73.

[12] 孙立晟,何东之. 改进无标度网络模型研究[J]. 电子设计工程, 2016, 24(6):115-117.

[13] 曹海,孙婧,史喜斌. 基于特征迭代的短文本去重算法[J]. 计算机工程, 2015, 41(12):54-57.

作者简介:

张莉女, (1981-), 硕士, 讲师. 研究方向为计算机网络、数据库技术. E-mail: amberzhangli@126.com.

孙丽娜女, (1982-), 硕士, 讲师. 研究方向为数据挖掘.

郭峰男, (1983-), 讲师. 研究方向为 Web 应用技术.