一种基于Context Graph主题爬行算法的改进

高庆芳，蒲宝卿，包 蕾

(陇南师范高等专科学校，甘肃 陇南 742500)

摘要:为了解决传统的基于上下文图形爬行模型爬取效率低、特征词提取不精确的问题,本文采用了在现有的技术支持上,融入机器学习算法的思想,设计并实现了一种改进的基于主题的爬虫算法,结果表明该算法是有效的.

关键词:搜索引擎;Context Graph主题;爬虫;特征权重

中图分类号:TP391.3 **DOI**:10.19789/j.1004-9398.2021.03.003

0 引 言

随着互联网的快速发展和全面普及,通过不同网络搜索引擎,从海量网络信息中获取信息,是互联网应用在学习和生活中的重要表现形式和组成部分,搜索引擎的核心是网络搜索引擎爬行算法.

通过对主题爬行算法的研究[1],传统的算法思想可分为2类:一类是基于Context Graph的主题爬行算法以及基于网络的超链接结构的爬行算法;另一类是早期由学者Diligenti提出的基于Context Graph的主题爬行思想,即利用爬行算法爬取相近主题的页面时,路线都是相似的[2].要实现后者的爬行算法,就要构建Context Graph模型层次结构图,用链接指向表明页面之间的联系,爬行路径让主题爬虫自主爬取.

根据相关文献,机器学习是主题爬虫最突出的特点,即可在爬行时选择其一网页分析算法选择链接对象,对于无关的对象可进行剔除.在这种情况下,在保证爬取效率的同时,也必须确保得到的结果页面具有更好的相关性.在计算机系统中用中文分词系统对中文文本进行词语自动识别,这方面的研究目前已经比较成熟,很多学者都总结提出了相关的算法.根据其特点,可以对现有的中文分词算法分类,主要有基于字符串匹配的、基于统计的和基于语义的3类[1].

传统的主题爬行算法存在不足之处是会降低网络的搜索速度.本文对传统的算法做了一个改进,主要修改了TF-IDF公式,目的是可以提高特征词权重计算的准确性,完成分类器的训练[3].主题爬虫是基于某个主题只爬取与主题属于同一类内容的页面,在建立的主题集上进行特征分类,在诸多量化方法中,下文提出的Context Graph层次模型具有较大优势.

1 Context Graph模型的构建

基于Context Graph的主题爬虫算法可以判断目标页面的路径与新访问的页面一致[3],计算当前页面及页面之间的位置.通过已知位置页面的样本训练,得到一定的链接路径,到达目标页面.

以一个2层的Context Graph模型为例[4],其结构如图1所示.为实现查询而收集的示例页面构成图的中心结点,作为第0层,指向第0层的页面(这些节点也称作父链接或者Backlink)被分到第1层,第1层和第0层的节点之间是直接链接关系,第1层的Backlink节点做为第2层,以此类推,直到构建出所需的模型为止.

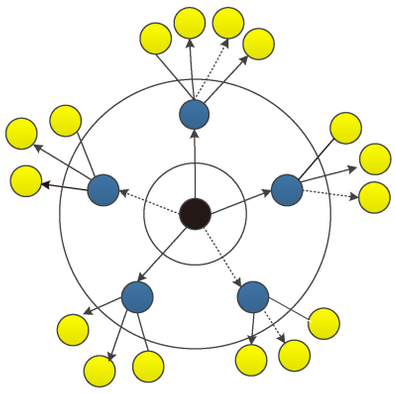
2 对特征权重算法的改进

传统的Context Graph主题爬行算法存在不足,如要基于爬虫爬行得到的已知页面集计算特征词的权值,而计算的值有时并不准确.在此,对原始算法TF-IDF公式做出改进.

2.1 文档特征权重的计算方法

TF-IDF函数用来表示一个特征项是否重要,即计算某个特定向量中一个词项的权重.特征项用词频(term frequency,TF)和逆文本频数(inverse docu-

收稿日期:2020-05-01



注:● 为中心节点第0层;● 代表的节点为第1层;
● 代表的节点为第2层.

图1　2层的Context Graph模型

ment frequency,IDF)2个因素衡量其重要程度[5].

2.1.1　TF

TF表示某个文本中某个特征项出现的次数[6],值越大说明该特征项越重要.设特征词为t_i,文档为d_j,TF定义为

$$\mathrm{TF}_{ij}=\mathrm{freq}_{ij}=tf_{i,j}=\frac{n_{i,j}}{\sum_{k=1}^{n}n_{k,j}},\tag{2.1}$$

式中$n_{i,j}$表示在特定文件d_j中某个特征词的出现次数,d_j包含的所有特征词出现的次数之和记做$\sum_{k=1}^{n}n_{k,j}$.特征权重的计算会受到少数高频词的影响,所以要进行归一化处理[7],用$\lg(\mathrm{freq}_{ij})+1$缩小TF范围,防止$tf$值偏向包含更多词项的长文件,则TF最终可表示为

$$\mathrm{TF}_{ij}=\lg(\mathrm{freq}_{ij})+1.\tag{2.2}$$

2.1.2　IDF

IDF反映了特定特征项在不同文档中的区别.假设一个文档集中,某个特征词很少出现,会有这样的情况,即虽然该词频度很小,但实际上该词具有较高的重要性[8].因此,往往包含某个特征词的文档的总数与该特征词的权重成反比,或近似反比[2].IDF表示为

$$\mathrm{IDF}_i=\lg(n/n_i),\tag{2.3}$$

式中n为所有的训练样本数,则特征词t_i出现的训练样本数为n_i.

通过对TF-IDF公式的研究,则有,如果某些特征词在一类文档中出现频率足够高,而在整个文档集合的其他文档中出现频率足够小[5],该词能更好地区别不同文档,因此,用TF和IDF的乘积来标记特征空间坐标系的取值[10].TF-IDF公式为

$$\mathrm{Weight}(T_{ij})=\mathrm{TF}_{ij}\cdot\mathrm{IDF}_i=\mathrm{freq}_{ij}\cdot\lg(n/n_i).\tag{2.4}$$

在文档d_j中,特征词t_i的频度为TF_{ij},n为所有训练样本数,t_i出现过的训练样本数为n_i.

2.2　分类训练及主题分类

目前有较多的基于机器学习分类文本的算法,此处基于贝叶斯定理,用分类准确率较高的朴素贝叶斯(naive bayes classifier)算法[11].假设一个类变量y依赖于特征向量$(x_1,x_2,\cdots,x_n)$,由贝叶斯定理有

$$P(y|x_1,\cdots,x_n)=\frac{P(y)P(x_1,\cdots,x_n|y)}{P(x_1,\cdots,x_n)}.\tag{2.5}$$

假设条件:

每个特征项都是独立的[12],则有

$$P(x_i|y,x_1,\cdots,x_{i-1},x_{i+1},\cdots,x_n)=P(x_i|y),\tag{2.6}$$

基于所有i的关系,可化简为

$$P(y|x_1,\cdots,x_n)=\frac{P(y)\prod_{i=1}^{n}P(x_i|y)}{P(x_1,\cdots,x_n)}.\tag{2.7}$$

算法步骤:

(1)先验概率、条件概率的计算,

$$P(Y=c_k)=\frac{\sum_{i=1}^{N}I(y_i=c_k)}{N},k=1,2,\cdots,K;\tag{2.8}$$

(2)其中$I(\cdot)$是指数函数$I(\mathrm{true})=1$,$i(\mathrm{false})=0$,

$$P(X^{(j)}=a_{jl}|Y=c_k)=\frac{\sum_{i=1}^{N}I(x_i^{(j)}=a_{jl},y_i=c_k)}{\sum_{i=1}^{N}I(y_i=c_k)};\tag{2.9}$$

(3)$j=1,2,\cdots,n$,$l=1,2,\cdots,S_{j,k}=1,2,\cdots,K$,$x_i^{(j)}$表示第$i$个样本的第$j$个特征,$x_i^{(j)}\in\{a_{j1},a_{j2},\cdots,a_{jl},\cdots,a_{jS_j}\}$,$a_{jl}$是第$j$个特征可能取值的第$l$个特征;

(4)对给定的实例数据$x_0=(x_0^{(1)},x_0^{(1)},\cdots,x_0^{(j)},\cdots,x_0^{(n)})^{\mathrm{T}}$计算后验概率为

$$p_k=P(Y=c_k)\prod_{j=1}^{n}P(X^{(j)}=x_0^{(j)}|Y=c_k),k=1,2,\cdots,K;\tag{2.10}$$

(5)确定实例数据$x_0=(x_0^{(1)},x_0^{(2)},\cdots,x_0^{(j)},\cdots,x_0^{(n)})^{\mathrm{T}}$进行分类,得出结果$y_0=c_0$,

$$y_0=\arg\max_{c_k}\left(p_k=P(Y=c_k)\prod_{j=1}^{n}P(X^{(j)}=x_0^{(j)}|Y=c_k)\right).\tag{2.11}$$

分类器的构建通过分类训练实现,过程如下:

(1)将待分类项设为$x=\{a_1,a_2,\cdots,a_m\}$,a为特征属性;

(2)存在类别集合 $C=\{y_1, y_2, \cdots, y_n\}$;

(3)计算条件概率 $P(y_1|x), P(y_2|x), \cdots, P(y_n|x)$;

(4)如果 $P(y_k|x)=\max\{P(y_1|x), P(y_2|x), \cdots, P(y_n|x)\}$,则 $x \in y_k$;

(5)计算(3)步骤中的每个条件概率是关键.算法思想是先定好训练样本集,然后对每个类别中的各个特征词的条件概率估值进行统计[13].

$$P(a_1|y_1), P(a_2|y_1)\cdots P(a_m|y_1);\ \cdots P(a_1|y_2),$$
$$P(a_2|y_2)\cdots P(a_m|y_2)\cdots P(a_1|y_n), P(a_2|y_n)\cdots P(a_m|y_n). \tag{2.12}$$

若设定条件独立的各特征属性,按贝叶斯定理有

$$P(y_i|x)=\frac{P(x|y_i)P(y_i)}{P(x)}. \tag{2.13}$$

分母是常值,在此只需最大化分子.基于各条件独立的特征属性得到

$$P(x|y_i)P(y_i)=P(a_1|y_i)P(a_2|y_i)\cdots P(a_m|y_i)P(y_i)=$$
$$P(y_i)\prod_{j=1}^{m}P(a_j|y_i). \tag{2.14}$$

相关的研究显示[14],构建 N 层的上下图文集模型,则参与训练的先验概率就要为 N+2 个类别:$P(c_i)(i=0,1,2,\cdots,N)$、$P(c_{\text{other}})$(若某页面不在 N 层中,则将被归至 c_{other} 类),可估计其先验概率,包含较多文档的层具有较大的先验概率.因此,重要步骤是估算各个特征词在每个层次(类)的先验概率 $P(m_t|c_j)(t=1,2,3,\cdots,|k_j|)$,可用公式为

$$P(m_t|c_j)=\frac{1+\sum_{D_i\in E_j}f(D_i(m_t)P(c_j))}{|V_j|+\sum_{D_i\in E_j}\sum_{s=1}^{|V_j|}f(D_i(m_t)P(c_j))}, \tag{2.15}$$

式中 $f(D_i(m_t)P(c_j))$ 为词 m_t 在文档 D_i 中出现的次数,$|V_j|$ 为词集中词的总数.

N 层 Context Graph 需 N+2 个先验概率类别,则文本分类器要有 N+1 个类别[15].每次爬取新的页面 D_i 时,先对类别 D_i 进行判断,确定是否会被划归至 $P(c_j|D_i)$,根据事先设定的阀值计算最大的类别 C_j,再与阀值比较,若小于阀值,则 C_j 会被归于 $P(c_j|D_i)$ 类.

因此,在一个 N 层上下图文模型中,爬虫得到的页面链接存放需要 N+2 个队列.初始状态下,0~N 队列为空,少量特定的页面会存放在第 N+1 队列中.经判断,对应层次 C_i 的页面会存放在 0~N 队列中,经计算,某个页面分类值小于预定阈值,其将被划归到 c_{other} 类别,就存储在第 N+1 队列中.这个过程是循环的,每一个新的网页被爬取时,会将其从距离模型中心最近的非空队列中取出,经过文本训练后,确定归属的类别,得到概率值,默认需求最近的前驱页面会优先处理.队列内部会根据概率值的大小进行排序存储,确保优先处理的特点[16],这也是主题爬虫算法的重要特征.

若事件 w 发生的概率记做 $p(w)$,则可定义信息论中的信息量公式为

$$I(w)=\lg\left(\frac{1}{p(w)}\right). \tag{2.16}$$

若属于类别 d 的文档的出现概率为 $p(d)$,在可确定的情况下,$p(dm_k)$ 的值差别不大.针对该现象做出一个改进,对应公式为

$$\text{MI}(d,m_k)=\lg\left(\frac{p(dm_k)}{p(d)p(m_k)}\right)=$$
$$\lg p(dm_k)-\lg p(d)-\lg p(m_k), \tag{2.17}$$

可知,若计算得到的 $\text{MI}(d,m_k)$ 值越大,则 $\lg p(m_k)$ 的值就应足够小.

特征选择算法的特点是基于 TF 差异[17].在改进的算法中,将实现对现有的 TF-IDF 公式的改善.

如果用 $M(m_k,d_i)$ 来表示词 m_k 关于类别 d_i 的权重[18],则计算类别权重的公式为

$$M\left(m_k,d_i\right)=\exp\left[\frac{1}{\overline{f_k^d}(N-1)}\sum_{j=1}^{N}\left(f_k^{d_i}-f_k^{d_j}\right)\right]\cdot$$
$$\left[1+\lambda\frac{\sqrt{D(d_i,f_k^{m_j})}}{E(d_i,f_k^{m_j})}\right],\quad j\neq i, \tag{2.18}$$

式中 m_k 为某一文档集中第 k 个特征词,d_i 为类别中的第 i 类,N 为数据集总类别数,$f_k^{d_i}$ 为 m_k 在 d_i 中出现频率,λ 为调整系数,$D\left(d_i,f_k^{m_j}\right)$ 为 $f_k^{d_j}$ 在 d_i 中的方差,$E(d_i,f_k^{m_j})$ 是 $f_k^{d_j}$ 在 d_i 中的均值.

基于以上理论提出改善后的 TF-IDF 公式为

$$v\left(m_k\right)=$$
$$\frac{tf(m_k,d)\cdot\lg(N_i/f(m_k+0.01)\cdot M(m_k,d_i)}{\sqrt{\sum_{m_k\in d}[tf(m_k,d)\cdot\lg(\frac{N_i}{f\left(m_k\right)}+0.01)\cdot M(m_k,d_i)]^2}}, \tag{2.19}$$

与原计算公式的差别是,原公式与特征词的类别权重进行乘运算.因为不同类别的文档,每一个特征词的重要性都应该被区分,也就是存在某个特征词在某个特定类别的文档中更为重要,而该词必

然不会在其他类别中表现得更为重要.

一般使用比较成熟的语料库,优点是样本可直接使用.首先,对样本集所包含的各个分词在每个类别的权值$M(m_k, d_i)$进行计算;其次,计算平均值$\bar{M}(m_k, d_i)$;最后,按照语料库中某个主题对应的类别,用已计算出的类别权值以及改进的TF-IDF公式来进行运算.如果遇到的分词还没计算出类别权值,可以取该词对应类别中所有分词权值的平均值,而后整个集合的特征词集就由计算结果最大的分词构成.

3　结　果

使用搜狗语料库作为数据源来提取文档和特征词,其资源内容主要采用手工整理与分类,目前是很多实验中都在采用的样本.该语料库中有财经类,本文选取"20.txt"这一文档,通过对各类别特征词的权值的计算,再通过后续处理.搜索主题为房地产时,其文档总数为720次,得到的部分结果如表1所示.可知,利息出现过的文档在整个文档集中的比例也不高(24%),如果按照改进前的TF-IDF公式,类似的这些分词不会被重视,对原有思想及公式改进后,其权重值增大,说明在此提出的改进算法是有效的.

表1　提取房地产的部分特征词结果

特征词	出现数/次	百分比/%	权重值	
			改进前	改进后
房价	543	75	1.751	1.814
每平方米	387	54	1.622	1.722
楼盘	226	31	1.376	1.552
开发商	201	28	1.138	1.222
贷款	195	27	1.141	1.227
利息	176	24	1.126	1.214
按揭	298	41	1.152	1.267
炒房	110	15	1.257	1.535

4　结束语

本文提出的改进算法是成功而有效的,可以提高提取特征词的准确性.算法不足之处是特征词相关度的确定还需再提高.由于汉语的结构复杂、语义丰富,怎样理解用户的真正意图并准确定位用户的搜索目标,能否给出其想要的信息,是目前搜索引擎技术亟待进一步解决的问题之一.

参 考 文 献

[1] 邓尧.社交网络中用户生成内容隐含位置的推测方法研究[D].西安:西安科技大学,2019.

[2] 方志坚,张瑞林,童小素.搜索引擎综合分析[J].计算机工程与设计,2007,28(16):4038-4041.

[3] 史昊天.网络搜索引擎搜索策略及算法研究[D].天津:天津工业大学,2018.

[4] 高庆芳.主题搜索引擎搜索策略的研究及算法设计[D].兰州:兰州大学,2017.

[5] 陈丹雯.基于文本驱动的虚拟新闻生成技术研究[D].长沙:国防科学技术大学,2006.

[6] HEYDON A,NAJORK M. Mercator:a scalable,extensible web crawler[J]. World Wide Web,1992,2(4):219-229.

[7] 张环.垂直搜索引擎中主题网络爬虫算法研究[D].济南:山东师范大学,2015.

[8] BURNER M. Crawling towards eternity:building an archive of the world wide web[J]. Web Techniques Magazine,1997,2(5):198-202.

[9] 叶育鑫,欧阳丹彤.基于语义的主题爬行策略[J].软件学报,2011,22(9):2075-2088.

[10] 姜华.基于Lucene面向主题搜索引擎的研究与设计[D].上海:华东师范大学,2007.

[11] 梁冰.基于CSI的室内定位与跟踪技术研究[D].哈尔滨:哈尔滨工业大学,2019.

[12] 彭涛,孟宇,左万利,等.主题爬行中的隧道穿越技术[J].计算机研究与发展,2010,47(4):628-637.

[13] 罗丽姗.垂直搜索引擎发展概述[J].图书馆学研究,2006(12):68-70+55.

[14] 罗欣,夏德麟,晏蒲柳.基于词频差异的特征选取及改进的TF-IDF公式[J].计算机应用,2009,25(9):2031-2033.

[15] 黄莉,王成良,杨铮.面向主题网络爬行的智能隧道穿越算法研究[J].计算机应用研究,2009,26(8):2931-2933.

[16] SALTON G,WONG A,YANG C S. A vector space model for automatic indexing[J]. Communications of the ACM,1975,18(11):613-620.

[17] 陈星.基于Context Graphs的主题爬虫的研究与实现[J].计算机工程与设计,2011,32(3):914-917.

[18] CHAU M,CHEN H. Personalized and focused web spiders[J]. Web Intelligence,2003(2):197-217.

An improvement of topic crawling algorithm based on Context Graph

GAO Qingfang, PU Baoqing, BAO Lei

(Longnan Teachers College, Longnan Gansu 742500)

Abstract: To solve the problems of low crawling efficiency and imprecise extraction of feature words in the traditional crawling model based on context, this paper adopts the idea of integrating machine learning algorithm with existing technical support, designs and implements an improved crawler algorithm based on topic, and the results show that the algorithm is effective.

Keywords: search engine; Context Graph topic; the crawler; feature weighting

(责任编辑:马田田)

(上接第6页)

[13] QI L Q. Symmetric nonnegative tensors and copositive tensors [J]. Linear Algebra and its Applications, 2013, 439(1): 228-238.

[14] WANG Y, HUANG Z H, BAI X L. Exceptionally regular tensors and tensor complementarity problems [J]. Optimization Methods & software, 2016, 31(4): 815-828.

[15] 韩继业,修乃华,戚厚铎. 非线性互补理论与算法[M]. 上海:科学技术出版社,2006.

Study on properties of solutions of the tensor complementarity problem defined on the cone K

ZHENG Hongru, GU Weizhe

(School of Mathematical Science, Tianjin University, Tianjin 300350)

Abstract: This paper extend the tensor complementarity problem from the nonnegative cone $\mathbf{R}_+^n$ to a more general pointed closed convex cone $\boldsymbol{K}$. This paper also introduces some new structured tensors, also proves the uniqueness, boundedness and compactness of solutions of the corresponding tensor complementarity problem.

Keywords: tensor complementarity problem; pointed closed convex cone; structured tensors; uniqueness; boundedness; compactness

(责任编辑:马田田)