

挖掘用户标签的增强型社区网页聚类算法

贺秋芳¹, 曾启杰², 蔡延光²

(1 广东轻工职业技术学院 计算机系, 广东 广州 510000; 2 广东工业大学 自动化学院, 广东 广州 510000)

摘 要: 网页的内容信息对于提高聚类质量来说并不完全够用, 针对网络社区网页之间存在的天然链接关系, 本文提出了一种挖掘用户标签的增强型社区网页聚类算法. 本文采用多种距离度量方法, 并挖掘网页链接关系, 然后将网页的内容信息相似度和链接关系结合起来进行聚类. 实验表明, 提出的算法是有效的.

关键词: 社区网页; 链接关系; 网页相似度

中图分类号: TP393

文献标识码: A

文章编号: 1000-7180(2013)02-0074-04

Enhanced Social Web Clustering Algorithm of Mining Information

HE Qiu-fang¹, ZENG Qi-jie², CAI Yan-guang²

(1 Department of Computer, GuangDong Industry Technical College, Guangzhou 510000, China;

2 Faculty of Automation, Guangdong University of Technology, Guangzhou 510000, China)

Abstract: With the development of Internet, textual content is not enough for web clustering sometimes. This paper proposes mining information for social web clustering algorithm. User information of pages is mined, including the link information, tag information. Experimental results show that the proposed social web clustering algorithm using mining information is effective.

Key words: social web; page link; web similarity

1 引言

聚类算法可分为基于层次的聚类算法、基于划分的聚类算法、基于密度的聚类算法、基于网格的聚类算法和基于模型的聚类算法^[1].

文献[2]提出了基于局部标签树匹配的改进网页聚类算法.

文献[3]针对搜索引擎返回的标题和摘要信息, 构造有向图表示, 并在此基础上实现了一种高效的网页聚类原型系统 (efficient web clustering system).

文献[4]提出了根据网页的文档对象模型树结构进行标签路径聚类. 常见的文本聚类方法利用向量空间模型将文档表示成为词语向量, 其中每个词语有一个权值, 然后计算文档向量之间的相似度, 通过向量之间的相似度来把文档聚成类. 虽然可以利

用词语的权值来表示文档向量, 但只是统计了词语的频率而已, 没有刻画词语的语义信息, 加上词语本身一词多义的干扰, 这就造成了有时候聚类的会产生误差. 所以很多时候网页的页面内容信息对于网页聚类来说并不完全够用.

网络社区网页存在着天然的链接关系, 本文提出了一种挖掘信息的增强型社区网页聚类算法. 本文采用多种距离度量方法, 挖掘网页之间的链接关系, 然后将社区网页的内容信息和链接关系结合起来进行聚类.

2 增强型社区网页聚类算法

K-means 聚类算法随机选择 K 个对象作为初始聚类质心, 计算待选数据对象和各个聚类质心的距离, 将该待选数据对象划入离聚类质心最近的那个类里面; 重新计算各个类的聚类中心; 重复计算,

收稿日期: 2012-05-16; 修回日期: 2012-06-28

基金项目: 国家自然科学基金项目(61074147)

直到各个类的聚类中心不再变化,或者满足聚类停止条件. K-means 聚类算法要预先选定集合的初始聚类中心. K-means 聚类算法的详细步骤可以描述为:

①含 n 个元素的集合 x , 假设 I 表示聚类次数, 随机选定 k 个初始聚类中心, 用 $Z_j(I)$ 表示;

②计算数据元素和聚类中心的距离 $D(x_i, Z_j(I)), i = 1, 2, \dots, n; j = 1, 2, \dots, k$; 如果满足判断条件

$$D(x_i, Z_k(I)) = \min\{D(x_i, Z_j(I))\},$$

则将 x_i 划入该类;

③重新计算这 k 个聚类中心:

$$Z_j(I+1) = \frac{1}{n_j} \sum_{i=1}^{n_j} x_i^{(j)};$$

④如果 $Z_j(I+1) \neq Z_j(I)$, 表示聚类还在变化, 则 $I = I+1$, 转到步骤②继续执行; 否则, 结束.

K-means 聚类算法适合数据集规模比较大的应用. 从 K-means 聚类步骤可以看出, 要先选定初始类别数目 k 和初始聚类中心, 有时候 K-means 聚类算法容易受到数据对象中存在的异常点干扰, 抗干扰性不强.

K-means 聚类模型采用的是词语向量. 文档距离采用文档词语向量之间的相似度来衡量:

$$\text{dist}(d_m, \text{centroid}_k) = \text{dist}(d_m, \text{centroid}_k)^{\text{word}}$$

对于带标签信息的社区网页聚类, 我们可以充分利用标签信息来扩充文档的向量表达.

标签向量模型: 采用文档的标签信息代替词语信息从而建立向量表达式. 文档距离采用文档标签向量之间的相似度来衡量:

$$\text{dist}(d_m, \text{centroid}_k) = \text{dist}(d_m, \text{centroid}_k)^{\text{tag}}$$

(词语+标签)向量模型: 将文档的标签信息分割为词语, 然后加入到该文档的词语表中, 根据新的词语表来建立向量表达式. 文档距离采用文档新向量之间的相似度来衡量:

$$\text{dist}(d_m, \text{centroid}_k) = \text{dist}(d_m, \text{centroid}_k)^{\text{word+tag}}$$

词语向量+标签向量模型: 将文档的标签信息和词语信息分别建立向量表达式. 在聚类的过程中, 分别计算文档标签向量的相似度和文档词语向量的相似度, 根据两者的线性组合值来决定文档的距离.

假设 $D = \{d_i, i = 1, 2, \dots, n\}$ 是网页文档集, 能够用一个无向图 G 来表示. 每篇网页文档 $d \in D$ 代表无向图 G 的一个结点. $c(d)$ 代表网页文档 d 所在的类别, $t(d)$ 代表网页文档 d 包含的词语集合. 假设 $\Phi_{i,d}$ 表示网页文档 d 划到类别 i 的概率. 对于网

页文档 d 的每一个链接文档 $d^* \in L(d)$, 我们只考虑它的类别划分的最大概率 $c_{\max}(d^*)$. 而且, d 和 d^* 之间的链接权重 W 根据它们的链接属性计算, 可以得到:

$$\Phi_{i,d} = \Pr[c(d) = i | t(d)] \prod_{d^* \in L(d)} \Pr[c(d) = i \wedge c(d^*) = c_{\max}(d^*)] * W$$

网页文档之间使用的链接类型有多种^[5]. 比如网页之间的超链接. 本文采用网络社区网页来进行聚类实验, 社区网页存在天然的链接关系, 如图1所示, u 、 r 和 t 分别代表用户、网页和标签. 每次当用户对网页给出标签时, 就建立了一个用户、网页和标签的三元关系.

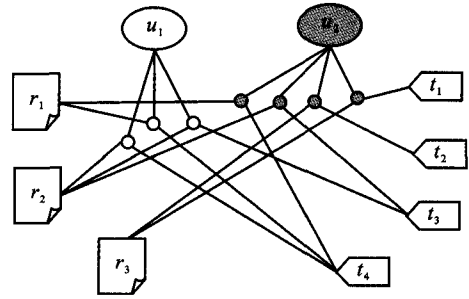


图1 用户、网页和标签的关系图

相同用户链接: 如果两个网页为同一用户, 则在这两个网页之间建立相同用户链接关系.

相同标签链接: 如果两个网页的标签相同, 则在这两个网页之间建立相同标签链接关系.

相同用户相同标签链接: 本文假设相同用户相同标签的网页文档话题更加相关, 划分成一类的可能性更大. 如果两个网页标注为相同用户并且有相同的标签, 则在这两个网页之间建立相同用户相同标签链接关系.

采用传统的向量余弦公式计算网页的内容信息相似度, 对于给定网页文档集 $D = \{d_i, i = 1, 2, \dots, n\}$ 和网页文档包含的词语集 $W = \{w_1, w_2, \dots, w_m\}$, 两文档的内容信息相似度计算公式如下:

$$\text{sim}(d_i, d_j) = \frac{\sum_{m=1}^k p_{i,m} p_{j,m}}{\sqrt{\sum_{m=1}^k p_{i,m}^2} \sqrt{\sum_{m=1}^k p_{j,m}^2}}$$

分别计算网页之间的内容信息相似度 $\text{sim}(d_i, d_j)$ 和网页 d 划入类别 i 的概率 $\Phi_{i,d}$, 根据网页 d 的得分 $\theta * \text{sim}(d_i, d_j) + \omega * \Phi_{i,d}$ 是否大于阈值 λ 来判断网页 d 是否划入类别 i , 重复计算直到每个网页已经划入相应的类别, 或者聚类算法达到最大计算次数.

3 实验结果与分析

3.1 实验结果评测方法

对聚类效果进行评测并不是一件容易的事情,准确率、召回率、F-Measure 值、Purity 和 NMI 这五个指标可以用来评测聚类效果。

假设聚类 i 和一个分类 j , N 为在文档集 i 中属于类别 j 的个数, M 为聚类 i 中所有文档的个数, K 为分类 j 中所有文档的数目, 则召回率 R 可以定义为:

$$R = N/K$$

准确率 P 可以定义为:

$$P = N/M$$

本文将准确率和召回率综合在一起考虑. F-Measure 值的定义^[6]. 分类 j 的 F-Measure 定义为:

$$F-Measure = \frac{2 \times P \times R}{P + R}$$

本文还采用了是 purity^[7] 和 normalized mutual information (NMI)^[8] 这两个评测指标. NMI 是根据即将要归类的变量和已经分类的数据之间的互信息值来计算的. NMI 的公式是:

$$NMI(X, Y) = \frac{I(X; Y)}{(\log k + \log c)/2}$$

其中, 变量 X 是即将要归类的随机变量, 变量 Y 是已经分类的随机变量, 变量 k 是聚类的数目.

3.2 实验过程与结果

本文采用从 <http://www.delicious.com/> 下载的真实社区网页作为语料进行实验. 本文下载了从 2010 年 1 月到 2010 年 12 月的 257,142 个网页. 聚类算法迭代次数设定为最大 50 次.

K-means 算法中的一个关键参数是预先选定的类别数目 k , 我们利用随机选择的 20000 个网页对类别数目 k 的取值进行实验, 如表 1 所示.

表 1 类别数目 k 的取值实验

k 值	P (%)	R (%)	F-Measure (%)	Purity (%)	NMI (%)
8	57.17	55.34	56.08	52.14	51.86
9	59.12	61.09	60.36	53.27	55.92
10	62.38	69.74	65.73	62.07	63.49
11	71.03	75.84	73.46	72.19	75.39
12	82.69	85.14	83.94	81.97	82.78
13	81.42	83.31	82.57	80.48	80.34
14	79.12	80.25	79.64	78.41	76.58

从表 1 可以看出, 最初选定 k 的取值对聚类效果有影响, 当 k 过小时, 不相关的网页聚在一起; 当 k

过大时, 相关的网页被迫分开, 影响聚类质量. 当 $k = 12$ 时, 聚类效果较好. 当然, 如果网页语料不同, 数目 k 的取值需要调整.

本文对 K-means 聚类的向量模型进行了三种改进, 并且分别采用 tf 和 tf · idf 权重公式进行对比实验, 结果如表 2 所示.

表 2 不同向量模型的聚类评测结果

向量模型	F-Measure (%)		Purity (%)		NMI (%)	
	tf	tf · idf	tf	tf · idf	tf	tf · idf
词语向量	58.12	63.76	60.12	62.71	59.36	58.29
标签向量	64.57	68.49	62.54	71.38	65.38	67.91
(词语+标签)向量	68.59	71.45	72.83	75.17	67.98	68.24
词语向量+标签向量	78.25	81.37	78.15	82.34	72.59	73.86

从表 2 可以看出, 词语向量+标签向量模型取得的评测结果较好. 阈值 λ 对聚类结果的影响如表 3 所示.

表 3 阈值 λ 对聚类结果的影响

λ	P (%)	R (%)	F-Measure (%)	Purity (%)	NMI (%)
0.2	52.25	57.23	55.41	52.16	53.18
0.3	64.24	67.15	65.62	67.84	67.55
0.4	74.26	77.85	76.03	77.24	78.69
0.5	83.16	81.37	82.59	80.01	79.28
0.6	78.36	79.45	78.97	78.96	75.62
0.7	62.37	64.85	63.95	61.79	64.73
0.8	55.73	58.79	56.84	56.52	52.77

从表 3 可以看出, 阈值 λ 对聚类效果有影响, 阈值大小影响到文档是否被排除在类之外. 当 $\lambda = 0.5$ 时, 聚类效果较好.

为了使实验结果具有可比性, 本文将传统 K-means 聚类算法作为基准和本文提出的增强型网页聚类算法进行对比实验. 实验中的参数设定采用上面实验的结果, 即 $k = 12, \lambda = 0.5$, 实验结果如表 4 所示.

表 4 不同聚类方法的聚类结果对比

聚类方法	P (%)	R (%)	F-Measure (%)	Purity (%)	NMI (%)
基于词语向量的 K-means 聚类	75.48	78.17	77.72	78.02	78.39
基于词语向量+标签向量的 K-means 聚类	76.65	80.45	78.92	79.54	80.29
本文算法	77.82	81.63	79.85	81.63	83.74

从表4可以看出,本文提出的算法得到的聚类结果优于传统K-means算法。

4 结束语

本文提出了一种挖掘信息的增强型社区网页聚类算法。采用多种距离度量方法,挖掘网页之间的链接关系,将社区网页的内容信息相似度和链接关系结合起来进行聚类,可以提高社区网页聚类的效果。针对参数 k 和阈值 λ 对聚类效果的影响,本文专门进行了实验和分析。实验结果证明了本文提出的算法是有效的。

下一步继续研究该聚类算法的聚类速度,还将研究各参数因子对聚类算法的影响程度,寻找最优的参数组合,进一步提高聚类的效果。

参考文献:

- [1] 刘凯鹏,方滨兴. 基于社会性标注的本地学习方法[J]. 计算机学报,2010,36(10):34-40.
- [2] 李睿,曾俊瑞,周四望. 基于局部标签树匹配的改进网页聚类算法[J]. 计算机应用,2010,30(3):818-820.
- [3] 谢艳玲,何丕廉,于鹏,等. 一种高效的网页聚类方法[J]. 计算机工程与设计,2007,28(17):78-81.
- [4] 刘云峰. 基于标签路径聚类的文本信息抽取算法[J]. 计算机工程,2010,36(12):92-96.
- [5] Zhang X, Hu X, Zhou X. A comparative evaluation of different link types on enhancing document clustering[C]//The 31st Annual International ACM SIGIR Conference on Research & Development on Information Retrieval. Singapore: ACM, 2008:135-139.
- [6] Dou Z, Song R, Wen J R. A large-scale evaluation and analysis of personalized search strategies[C]//The 16th international conference on World Wide Web. Banff, Alberta, Canada, 2007: 112-114.
- [7] Qiu F, Cho J. Automatic identification of user interest for personalized search[C]//The 15th international conference on World Wide Web. Edinburgh, Scotland: ACM, 2006: 256-259.
- [8] Chirita P A, Firan C S, Nejdl W. Personalized query expansion for the web[C]//The 30th annual international ACM SIGIR conference on Research and development in information retrieval. Amsterdam, The Netherlands, ACM, 2007:326-329.

作者简介:

贺秋芳 女,(1977-),硕士,讲师.研究方向为计算机应用、人工智能、嵌入式系统及物联网。

曾启杰 男,(1977-),博士研究生,讲师.研究方向为网络控制技术、智能系统、网络信息自动化。

蔡延光 男,(1964-),教授,博士生导师.研究方向为组合优化、人工智能、决策支持系统、智能交通。

(上接第73页)

- [5] Garbayo J M A, Valderas M G, Garcia M P, et al. A recovery mechanism for SET protection using standard-cells[C]//Radiation and Its Effects on Components and Systems (RADECS), Spain: Sevilla, 2011: 128-131.
- [6] Kim E P, Shanbhag N R. Soft N-modular redundancy[J]. IEEE Transactions Computers, 2012, 61(3): 323-336.
- [7] Zhou Q, Mohanram K. Cost-effective radiation hardening technique for combinational logic[C]//Proc. Int'l Conf. on Computer-Aided Design. USA: San Jose, 2004:100-106.
- [8] Kumar J, Tahoori M B. A low power soft error suppression technique for dynamic logic[C]//Defect and Fault Tolerance in VLSI Systems. USA: Monterey, 2005:454-462.
- [9] 刘必慰,陈书明,梁斌. 一种新型的低功耗SEU加固存储单元[J]. 半导体学报,2007,28(5):755-758.
- [10] Garg R h, Jayakumar N, Khatri S P, et al. A design approach for radiation-hard digital electronics[C]//Design Automation Conference. USA: San Francisco, 2006:773-778.

作者简介:

张金艺 男,(1965-),教授,博士生导师.研究方向 SOC DFT与故障容错设计、无线通信类ASIC设计。

唐夏 男,(1982-),硕士研究生.研究方向纳米级工艺下故障容错设计。