基于复杂网络理论的讨论区发帖分类推荐算法①

李 培

(西安邮电大学 计算机学院, 西安 710061)

摘 要: 在众多的网络应用中, 网络论坛的发展也是非常迅速的, 针对不同的人群的不同的兴趣爱好总会有某些论坛的某些讨论区是我们感兴趣的. 可是论坛复杂的结构对于新的用户来说, 如何找到正确的讨论区发帖成了一件耗时的问题. 本文通过研究复杂网络的相关理论, 针对该问题提出了一种讨论区发帖的分类推荐算法, 并经过验证和测试, 证明该算法是行之有效的, 可以给被该问题困扰的用户提供有用的帮助.

关键词: 复杂网络; 讨论发帖; 分类推荐

Category Recommendation Algorithm for Discussion Forum Posts Based on Complex Network Theory

LI Pei

(School of Computer Science and Technology, Xi'an University of Post and Telecommunications, Xi'an 710061, China)

Abstract: In numerous of network applications, network Forum of development is very quickly. For different of crowd and different of interest hobby always has some forum of some discussion district is we are interested in. But because of complex structuction of forum, new user has a takes of problem for how to find the correct discussion district to post into. Through researching complex network of related theory, for this problem we made a category recommended algorithm for discussion district posts. After validation and test, it is proved that the algorithm is effective, and it could be to provide users with useful help in distress of the problem.

Key words: complex network; discussion forum posts; category recommendation

1 引言

目前网络的普及加上网络技术的发展, 网络所提供的服务越来越丰富, 特别是各类面向广泛群体的信息服务因为具有友好的交互性界面而备受用户欢迎, 但是随之带来的网络信息爆炸式的增长以及网站大型化后各级各类的扩展, 使得用户在访问特定站点时, 不能非常顺利地达到访问的目的地, 或是能够非常顺利的完成相应的操作. 因此, 衍生了各类针对不同应用的推荐算法. 诸如: 电子商务中个性化推荐算法[1]、网络新闻的主题推荐[2]、广告推荐算法[3]等. 但是, 对于网络讨论区的应用问题涉及很少, 比如: 在很多讨论区网站, 不熟悉的用户在浏览时想要查看感兴趣的内容时, 根据第一级的讨论区目录很难决定选定自己需要前往的区域, 可能因为没有选择合适的区域而找不到感兴趣的内容而层层返回, 重新选择; 同样, 用户在发帖进行提问和讨论时, 也可能会找不到合适的讨论区, 也因此达不到进一步交流学习的目的. 这对网站的维护以及更多用户的浏览造成了不便, 也使得网站的推广收到了制约.

这样的问题经过分析主要是由于网站的操作使用过于依赖用户、对用户接触新事物适应性估计过高, 对用户的知识能力水平估计过高等造成的. 而网站的搭建和管理者, 应该基于其对网站结构、分类标准以及已有信息的充分了解给用户提供相应的支持和帮

① 基金项目:国家自然科学基金:(61105064);国家自然科学基金:(61203311);陕西省自然科学基础研究计划:(2011JM8007);西安邮电大学校青年教师科研基金(ZL2013-24)

收稿时间:2013-11-27;收到修改稿时间:2014-01-03

助. 现有的研究中, 多是采用基于词频统计、余弦函数计算相似度的传统分类算法进行推荐[4], 普遍存在推荐不够准确, 且分类过于死板的问题. 而我们所提出并研究的讨论区分类推荐算法, 就是利用网站已有的分类结构以及发帖信息, 在 Web 数据挖掘技术基础上进行文本聚类的研究, 从而为后续发帖用户提供讨论区推荐, 使得原有的静态网页结构有了交互式的帮助信息.

2 Web 数据挖掘技术

Web 数据挖掘技术主要是针对因特网应用相关的大量数据进行处理和分析的技术. Web 数据挖掘的处理对象与因特网相关, 既包括因特网应用自身的数据也包括由于应用相关操作导致的数据, 基于此, Web 数据挖掘的任务可以分为针对网页内容的 Web 内容挖掘、针对网站结构等应用的 Web 结构挖掘、以及针对网络检索、浏览等应用的 Web 使用挖掘.

通过 Web 数据挖掘技术, 我们可以对网络上的众多应用相关的数据进行分析处理, 从中挖掘出我们所需要的知识, 还可以进一步反过来对各类网络应用的优化和完善提供新的思路和处理办法.

但是 Web 数据挖掘技术不同于一般意义上针对传统数据库的挖掘技术, 它所针对的数据的从形式上来说结构更加复杂, 很多信息属于非结构化信息, 同时很多信息又是动态变化着的, 因此在分析之前必须要进行相应的数据预处理, 从而使得数据信息能够应用于数据挖掘. 同时, 它所针对的数据从数量上也更加巨大, 导致处理数据必须要考虑其时间复杂度, 从而判断其可行性.

针对上述分析, 在 Web 数据挖掘中的一般流程如下:

首先, 获取 Web 数据挖掘的原始数据, 一般有采用网络爬虫、抓包程序或是直接从网络应用的数据库或日志文件中得到.

接下来, 原始数据进行选择和预处理, 主要是针对不同的挖掘目标, 从数据中选择可用和有用数据, 并且将数据中与研究无关的无用信息去除掉, 将数据结构化存储和表示.

再下来, 根据不同的数据挖掘的需要以及处理算法的特点, 将数据整理为标准的特定格式.

最后, 将数据应用相应的数据挖掘算法进行计算, 分析其结果, 获得有效信息, 同时考察算法的可行性.

3 基于复杂网络理论的文本聚类算法

我们所研究的讨论区发帖分类推荐算法, 其对象为网络讨论区的发帖内容, 因此也属于 Web 数据挖掘的研究范围. 按照上一节中所提到的 Web 数据挖掘的一般流程, 针对讨论区发帖分类推荐的关键问题给出解决.

3.1 文本特征选择

网络论坛发帖的内容经过处理后, 最终可以转化为纯文本数据信息. 而在文本信息的表示方面, 大家普遍使用向量空间模型来表征文本特征, 也就是将文本进行分词处理后, 按照分词的结果, 将得到分词作为向量集合中的一个特征项来表征文本. 但是, 众所周知, 一个文本经过分词后得到的分词数量是相当大的, 如果所有的分词都用来作为文本特征, 那整个向量空间模型就存在维度过高, 不利于后期算法处理的问题. 同时, 通过观察分词的结果, 大家会发现很多的分词, 例如: “地、的”等词无意义;有些分词又与文本的主题不是很相关, 因此, 显然没有必要将所有的分词都作为文本特征来表示.

那么, 选择哪些分词既可以很好的表示文本特征, 也可以解决维度过高的问题. 文本的关键词恰恰具有这样的性质, 关键词顾名思义, 是文本主要内容和特征的体现;关键词与整个文本的所有分词相比数量要少很多. 因此, 我们改进传统的文本表示方法, 采用关键字集合来表示文本. 本文的关键字提取采用本课题组提出的基于复杂网络理论[5]中小世界网络模型的关键字提取方法[6].

算法思想如下: 由文本分词构成文档语义连接的文档结构图, 利用出现的小世界现象[7], 选取文档分词集中△Li 和△Ci 值较大的前 n 个词分别组成侯选集合 Lmax 和 Cmax ,然后找出 Lmax 和 Cmax 中在文档中一个句子中为相邻的词, 按位置顺序进行合并形成 Keyword_Set., 最终得到文本的关键词. 其中, △Li 为去掉结点 vi 后图的平均最短路径的变化量, △Ci 为去掉结点 vi 后图的簇系数的变化量. 经验证, 该算法能准确的获取表示文本的主题的关键词.

3.2 文本相似度计算

在采用向量空间模型(VSM)表示文本的前提下, 文本相似性的计算往往是基于向量空间模型中各个向量的相似性计算结果, 也就是基于词语之间的相似性来判断整篇文本的相似性, 这样的设计是符合实际情

况并且也很容易解释的. 那么文本相似度的计算问题主要就集中在如何计算词语之间的相似性.

传统的词语相似性计算往往基于词根或是只能单纯的比较两个词语是否一样, 这样的方法与中文语言词汇丰富且意义深刻的特点不相符合, 我们都知道在中文文本中很多词汇代表相似的含义, 但是构成完全不同, 而且很多词语具有多种含义. 例如: “计算机”和“电脑”这样的两个词语, 在传统的词与相似性计算中是无法给出两者相似的结论的, 但是实际情况恰恰两者意义是非常相近的, 导致基于传统词语相似性计算的不准确, 从而导致文本分类的最终结果错误.

因此, 如何解决词语相似性判定的问题成了文本相似度计算的关键, 参考文献基于《知网》的词汇语义相似度计算算法[8,9], 利用文献中提出的一个能够计算词语语义的 DLL(动态链接库)供人们使用, 进行词语语义计算, 从而计算文本相似度.

目前关于如何度量两个含有非数值型字段的记录之间的距离的讨论有很多, 并提出了相应的计算方法, 比如欧氏距离以及余弦函数. 这些方法的缺点是没有考虑特征项之间的关系, 认为它们是互相独立的.

本文中用到的具体算法如下:

Step1: 文本 T1 表示为 KeyWord(T1)＝{ w_i | w_i 表示 T1`的第 i 个关键字 };文本 T2 表示为 KeyWord(T2)＝{ $w_i$ | $w_i$ 表示 T2`的第 i 个关键字 };

Step2 查找 Keyword(T1)和 KeyWord(T2)集合中词语相似度大于阈值的词语数量 N.

Step3 计算 N 占关键词语的比率作为文本相似度的度量标准.

这个算法充分考虑词语之间的语义相似性. 也就是说如果一篇文本中大部分的关键字都与另一篇文本相关联, 则这两篇文本就是相似的.

3.3 文本聚类算法

现有的聚类算法种类繁多, 涉及到各个领域, 但是在研究复杂网络理论的过程中, 我们发现, 复杂网络理论中的很多算法与传统的聚类算法相比较, 更适合处理大规模海量的数据, 恰好与当前应用中处理中文文本的需求相符合[5]. 因此我们选取了复杂网络相关理论的奠基人 Newman 所提出的一种基于图的、针对大量网络结点运算、并且具有快速有效特点的聚类算法[10,11]进行改进, 应用于中文文本聚类, 并与实验中验证其准确性、有效性及可行性.

关于 Newman 算法的介绍见参考文献[10], 该算法提出的这种基于图的中文文本聚类算法处理的对象是按照前述处理后得到由关键词语表示的文本, 同时该算法具有参数少和聚类数目可选的特点.

具体算法如下:

Step1: 采用第三节中提出的算法计算文本的相似度.

Step2: 构图. 首先, 每一个文本作为图中的一个结点;接下来, 将文本相似度大于阈值的文本之间进行连线, 构成图的边.

Step3: 聚类. 借鉴 Newman 聚类算法, 按照聚类结果要求类内联系紧密, 而类间联系稀疏的特点, 进行不断地合并.

4 论坛发帖分类推荐算法

基于上一节所提出的基于复杂网络理论和语义相似性的文本聚类算法, 针对网络论坛发帖分类推荐问题, 我们提出一种推荐算法, 通过用户的发帖内容分析, 向用户给出该内容可选择的发帖讨论区作为推荐, 最终由用户选定理想的发帖讨论区.

该算法基于以下几点:

1、大部分发贴内容为纯文本, 且非文本内容, 例如图片、动画、表情等, 对文本内容的表示并非唯一不可取代, 完全可以由其文字性的注释等获得相同的信息, 因此去除发贴内容中的非文本信息, 不影响对发贴内容主题的判断, 也不会对发贴内容的讨论区选择造成影响.

2、绝大部分论坛区的发贴内容是与讨论区内容相关的, 也就是说前人选择的讨论区大部分是正确的.

3、讨论区发帖内容合适的讨论区并不唯一, 也就是说一篇贴在可以发在不同的讨论区, 我们给出推荐排名.

论坛发贴分类推荐算法的具体实现步骤如下:

Step 1: 将论坛中原有的文章进行处理, 去除其中非文本的内容, 存储为纯文本格式并进行编号命名. 并建立和存储该文本与所在讨论区编号的对应信息. 例如:

Tji 代表讨论区 j 中的第 i 个文本.

{ j 为论坛中讨论区的编号, i 为原始文本在该讨论区的编号 }

Step 2: 按照前一节所提到基于小世界网络模型

的关键词提取算法对原论坛中所有纯文本进行关键词提取，并使用关键词对文本进行表示，即

Keyword(Tji)={wi|wi 为文本 Tji 的第 i 个关键词}

Step3: 对用户的发贴内容进行处理，去除其中非文本的内容，存储为纯文本格式. 同样采用基于小世界模型的关键词提取算法进行关键词提取，并使用关键词表示该文本. 即

Keyword(P)={ vi|vi 为文本 P 的第 i 个关键词}

P 为发贴内容的纯文本.

Step 4: 对论坛原始文本集合和发贴内容文本按照前面提到的基于复杂网络理论的图聚类算法，进行文本聚类.

Step 4.1: 计算所有文本之间的文本相似度

Step 4.2: 文本相似度大于文本相似度阈值的文本对应的结点之间加边. 文本相似度阈值的选取原则是保证图的密度是 d_{thred} . 图的密度计算公式是图的边数除以 C_n^2 .

Step4.3: 对构成的文本关系网络图进行聚类，初始时候，每个结点被当作一个簇. 然后重复合并其中的两个簇，合并的前提条件是仅当合并会使得 Q 的增量是最大的. 直到 Q 不再增大时候终止.

Q 被称为模块度，是 GN 聚类算法中给出了一个衡量网络划分质量的标准:

$$Q=\sum_i (e_{ii}-(\sum_j e_{ij})^2)$$

其中 e_{ij} 代表连接类 i 中的结点和类 j 中的结点的边数占总边数的百分比，e_{ii} 代表类 i 中的边数占总边数的百分比.

Step 5: 观察上一步的聚类结果，找出与文本 P 分为一类的文本 Tji，并查询并记录 Tji 所归属的讨论区. 鉴于有可能存在与文本 P 归为一类的文本不会都属于同一个讨论区，需要计算文本 P 所归属讨论区 j 的概率. 即

$$p_j=\frac{\text{与文本P归为一类的文本中属于讨论区 j的文本总数}}{\text{与文本P归为一类的文本总和}}$$

{j 为论坛中讨论区的编号}

Step 6: 将上一步计算所得概率排名前三的讨论区作为推荐发贴讨论区推荐给用户，并将概率值作为其推荐权值.

按照上述算法，通过在“西安妈妈网”论坛中提取的若干讨论发贴内容文本作为实验数据进行验证如下:

分别从讨坛的三个典型讨论区: “环游世界”、“美味厨房”、“早教幼教”中提取各 30 篇共 90 篇的文章，并对三个讨论区按照顺序编号为 1、2、3，进行算法的验证测试. 以撰写的一篇与教育有关的文章 p 为例，进行论坛发贴分类推荐模拟. 按照前文中的六步，最终 p 与五篇属于编号为 3 开头即“早教幼教”讨论区的文本分为一类. 显然，该文章会推荐它发表在“早教幼教”讨论区，且推荐概率为 1.

下图是为了能够体现该算法的执行过程，将处理文本集合数量大幅减少，分别用三个讨论区中各两个文本组成集合与待推荐文本 P 采用上述算法进行处理图示如下. 文本编号的第一位代表其所属讨论区编号，后两位为讨论区内顺序号，即 303.txt 为属于 3 号讨论区“早教幼教”的第三个文本.

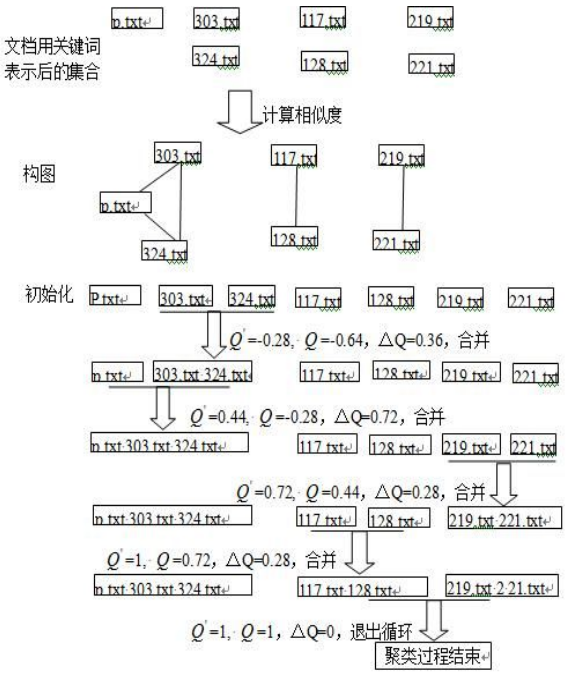

图1 论坛发贴分类推荐算法图示

显然，如图所示，文本 P 与属于 3 号讨论区“早教幼教”归为一类，其推荐讨论区则为“早教幼教”，概率为 1.

5 分析与总结

经过我们的测试，验证了这种基于复杂网络理论的讨论区发帖分类推荐算法的正确性，同时，由于我

们采用的算法都认真的考虑其算法的时间和空间复杂度，采用降低维度等方式增加其可行性，加之，论坛发贴内容大部分文本内容短小精炼，所以处理起来虽然文本数量偏多，但是效率还是比较高的，具有其可行性.

鉴于数据源有限，我们的算法没有在更大的数据集合上或是真实的论坛上进行测试，不能不说是个遗憾，不过，如果数据集合过大或是论坛文章众多的话，我们在算法第一步中处理的文本集合可以不用将所有的文章都作为对象，而可以动态的按照特定的标准选取其中的部分文章，比如: 回帖数量多的、充分反映讨论区特点的，最新发表的等等. 这样同样使得我们的算法适应其应用.

该算法还有很多不成熟的需要进行深入研究的地方，但是由于该研究课题的提出是有着其真实需求的，也是来自我的自身应用中遇到的问题，相信是有着其研究的必要性，而且会给网络论坛用户的发帖应用带来更多的便利，因为目前对于该问题的研究相对较少，而且能够既有理论又有实际测试的研究更少，希望我们的研究能够将更多 web 数据挖掘中的优秀聚类算法引入到该研究领域，取得更好的结果.

同时，该算法除了可以应用在讨论区网站的用户发贴推荐问题中，也可以用于信息发布网站、分类搜索网站的类似问题中，总之，作为用户个性化推荐服务的一个重要方面，其具有很明确的应用背景和应用需求.

参考文献

1 刘利民,刘晓莉.混合智能算法在电子商务个性化推荐中的研究.内蒙古工业大学学报(自然科学版),2011,(3).

2 陈宏,陈伟.基于多主题追踪的网络新闻推荐.计算机应用,2011,(9).

3 涂丹丹,舒承椿,余海燕.基于联合概率矩阵分解的上下文广告推荐算法.软件学报,2013,24(3).

4 丁智斌,杜念.基于 Web 内容挖掘的论坛发贴分类推荐技术.华北科技学院学报,2011,(1).

5 汪小帆,李翔,等.复杂网络理论及其应用.北京:清华大学出版社,2006.

6 周雅夫,马力,董洛兵.基于 SMW 理论提取复合关键字系统的设计与实现.西安邮电学院学报,2007,(5):82–85.

7 Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature, 1998, 393(4): 440–442.

8 刘群,李素建.基于《知网》的词汇语义相似度计算.第三届汉语词汇语义学研讨会.台北.2002.

9 唐歆瑜,乐文忠,李志成,等.基于《知网》语义相似度计算的特征降维方法研究.科学技术与工程,2006,6(21):3442–3446.

10 Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Phys. Rev. E, 2004, 69: 026113.

11 Matsuo Y, Sakaki T. Graph-based Word clustering using a Web Search Engine. 2006.