

一种基于聚类技术的全文检索与推介系统的构建

张克君¹ 任 鹏² 钱 榕¹ 居荣斌¹ 姜 琛¹ 张国亮¹

(北京电子科技学院计算机科学与技术系 北京 100070)¹

(西安电子科技大学计算机学院 西安 710071)²

摘 要 近年来,搜索引擎的发展可谓突飞猛进,排序算法也日新月异,但相关搜索推介这项功能却进展缓慢,无法为用户提供令人满意的有价值的关键词。本项目是专门为了解决这个问题而进行研究的,采用单词到文档,文档到聚类,聚类再回归单词的语义检索流程,完成了 K-means 聚类算法以及 TFIDF 权重算法的 Java 实现。通过此系统,用户不仅可以找到包含指定关键词的网页,还会收到与该关键词关联最紧密的其他关键词推介,协助用户进一步发掘信息。

关键词 搜索引擎,聚类,关键词,推介

中图法分类号 TP311 文献标识码 A

Construction of One Kind of Full-text Searching & Recommending System Based on Clustering

ZHANG Ke-jun¹ REN Peng² QIAN Rong¹ JU Rong-bin¹ JIANG Chen¹ ZHANG Guo-liang¹

(Department of Computer Science and Technology, Beijing Electronic Science and Technology Institute, Beijing 100070, China)¹

(School of Computer Science and Technology, Xidian University, Xi'an 710071, China)²

Abstract In recent years, the development of the search engine and the sorting algorithm is updating fast. But the recommending system, which cannot provide valuable keywords in the past, is barely evolved. Our project is specialized in order to solve this problem. This project goes from word to document, and from document to cluster, then from cluster to the word which is to be returned. It realized the K-means clustering algorithm and the TFIDF weight algorithm with Java. Users can find not only the web pages including the specific keyword, but also the most valuable keyword recommended which is to help them finding information related.

Keywords Search engine, Cluster, Keyword, Recommending

1 引言

如今,互联网已经成为了世界上最重要的信息发布平台之一,极低的门槛让每个人都能成为信息的发布者。但随之而来的就是信息的大爆炸,使得互联网上的信息大大超过了一个普通人的记忆范畴,再也不可能单凭记忆去寻找信息。这促使了搜索引擎的出现。借助搜索引擎,可以方便快捷地找到几乎任何我们需要的内容。但从另一个角度讲,搜索引擎的出现却加剧了信息的爆炸,虽然信息丰富了,但想得到真正需要的信息却越来越困难了。为了改善这一状况,越来越多的富网络应用(RIA)^[1]出现了我们可以更精确、更有目的地获得信息。无论是从用户体验还是从底层技术上讲,我们似乎离 Web2.0^[2],甚至 Web3.0^[3]越来越接近了。打开 Google、百度这样的大牌搜索引擎,随便搜索什么,都会在页面的下面找到类似“相关搜索”的列表,其中罗列了很多仅仅基于统计学的相关搜索链接。而这就是问题所在,因为任何语言都不可避免地存在二义性,所以这种相关搜索的借鉴价

值值得怀疑,相反地,它甚至还有可能限制住用户的思维,导致用户无法找到理想的信息源。本项目就是针对这个问题而研发的,在搜索过程中,用户可以得到基于文本聚类的关键词推介,拓展搜索思路,更好地获得信息,体会流畅搜索信息的快意。

2 同类产品现状

2.1 基于关键词的搜索引擎

目前,主流的搜索引擎原理是构建倒排索引和排序反馈用户^[4,5],即根据用户提供的关键词,经过分词、过滤等操作,连接到索引库进行一字一句的搜索。目前主流的大型网站中都有基于关键词的搜索引擎的影子,比如百度、Google 以及几乎所有与查询有关的网站。在这些网站中,用户首先需要将自己抽象的思维具象为一个或一组词汇,词汇与词汇之间通常使用空格隔开。通过这样的具象过程最终得到的就是“关键词”。然后用户通过网页表单向网站服务器提交关键词,而网站服务器的任务则是将关键词进行“分词”,即把用户

本文受国家自然科学基金项目(61170037),北京电子科技学院科研项目(2014GCYY09)资助。

张克君(1972—),男,博士,副教授,主要研究方向为信息安全、知识发现, E-mail: zkj@besti.edu.cn; 任 鹏(1988—),男,硕士生,主要研究方向为信息安全; 钱 榕(1970—),男,博士,副教授,主要研究方向为信息安全、复杂网络; 居荣斌(1987—),男,硕士生,主要研究方向为信息安全; 姜 琛(1991—),男,硕士生,主要研究方向为信息安全; 张国亮(1990—),男,硕士生,主要研究方向为密文检索。

具象的词组重新分开,得到最细化的词元(Term)。将这些词元输入索引查询器,索引查询器经过倒排索引,向服务器返回搜索结果 ID 等关键信息,再由服务器返回给用户。

2.2 基于统计学的关键词推介系统

因为基于关键词的搜索引擎具有上述局限性,人们在长期的思索后得到了一种完全不同于倒排索引的全新思路,即“基于统计学的搜索引擎”,它是根据长期的数据积累和统计而得到大量有价值的相关信息。基于统计学的搜索引擎在网上购物系统中应用得最为广泛,所有的购物网站中都有它的踪影,即“对该商品感兴趣的用户还对××感兴趣”。相信每一位有过网络购物经验的读者都通过类似这样的链接购买过许多原本没想购买的商品。这是一条新的营销之道,同时也更是一项新的技术,通过它,用户可以很方便地对自己具象出来的关键词进行扩展和完善,以达到最终完全吻合脑中抽象需求的目的。

基于统计学的搜索引擎也有它的优点和缺点。优点是它可以一定程度上优化用户的搜索关键词,进而帮助用户找到真正需求。它的缺点也显而易见,因为基于统计学的搜索引擎往往采用数据库存储的方式,所以搜索速度相比倒排索引要低很多,而且它并不能提供完整的用户需求。它的存在,多数情况都是作为基于关键词的倒排索引搜索引擎的辅助工具。

3 系统设计

3.1 数据流图

将系统模块示意图按照具体任务细化后,得到了如下几个流程模块:用户管理、索引管理、聚类管理、日志管理、词典管理。模块之间的调用关系如图 1 所示。

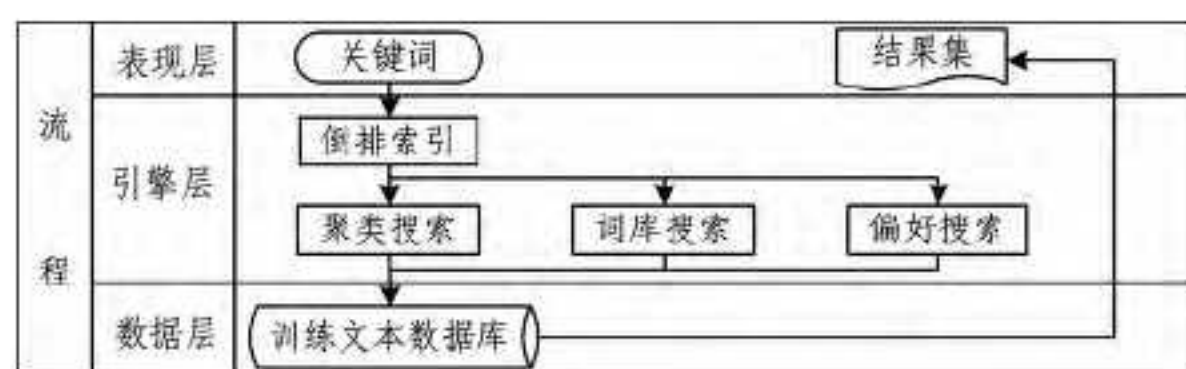


图 1 搜索数据流图

3.2 系统流程

3.2.1 索引搜索

无论是对基于统计学的搜索引擎还是基于语义的搜索引擎,基于关键词的搜索引擎都是一切的基础。本系统也需要构建一个基于关键词的搜索引擎,要求能够做到在用户给系统提供关键词之后得到基于关键词的搜索结果,即不脱离索引的搜索结果^[6]。

索引搜索的机制是先建立倒排索引再进行搜索,搜索过程与源文档完全分离,所以在用户进行搜索之前需要在后台完成索引的创建、添加、更新和优化等过程,在这些之后方可允许用户进行索引搜索。所以用户进行搜索时实际是在搜索倒排索引,这个模块完成的就是对倒排索引进行搜索的操作,并将搜索结果返回给前台,最终显示在浏览器中。

3.2.2 聚类搜索

该部分是本语义搜索引擎系统的重点内容,它可以通过聚类实现相关关键词的提示。如图 2 所示,首先建立训练文本库,并对文档中的每个分词计算权重,进而对整个训练文本库进行聚类,相关的权重和聚类信息均保存在数据库中。在完成索引搜索之后,系统会自动在数据库中进行聚类搜索,即

在索引搜索得到的结果集中找到与当前关键词关联最大的文档,并找到这个最大关联度文档所在的聚类,最后将该聚类的特征词提取并反馈。

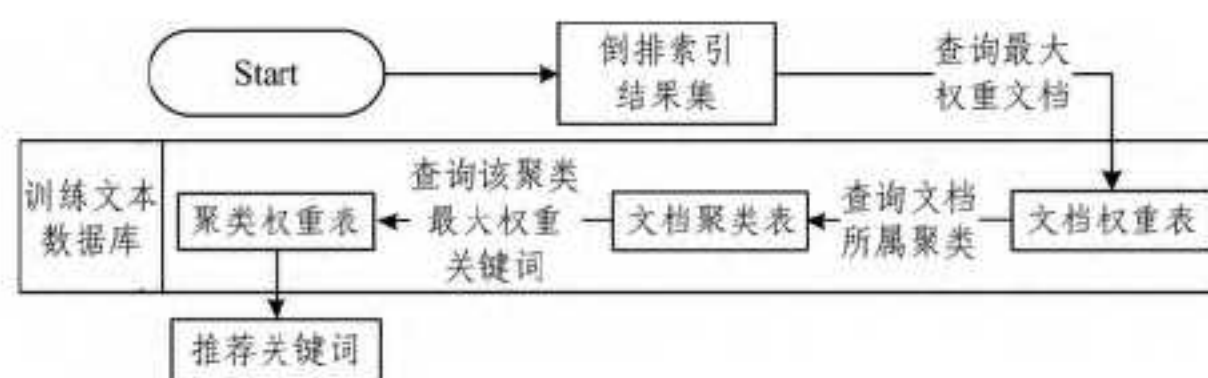


图 2 聚类搜索流程

3.2.2.1 TFIDF 权重算法实现

本系统中的词汇_文档和词汇_聚类的特征值权重采用 TFIDF 算法^[7]计算。TF-IDF(Term Frequency-Inverse Document Frequency)是一种用于资讯检索与资讯探勘的常用加权技术。TF-IDF 是一种统计方法,用以评估一个字词对于一个文件集或一个语料库中的一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加,但同时会随着它在语料库中出现的频率成反比下降。总之,我们可以通过 TFIDF 算法得到某个词对于某一段文字的权重值,作为其在这段文字中重要程度的衡量标准。其计算公式如下:

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (1)$$

$$IDF_i = \log \frac{|D|}{|\{j: t_i \in d_j\}|} \quad (2)$$

$$TFIDF_{i,j} = TF_{i,j} \times IDF_i \quad (3)$$

由这 3 个公式可以得到任意一个分词对于任意一篇文档的权重值,然后借助此权重值进入文本相似度计算的流程。

3.2.2.2 文本相似度的计算

文本相似度是用来衡量两个文本对象之间的相似度,文本之间的相似度越高,就越容易被划分到一个聚类中。可以说,度量文本相似度可以将所有的文本看作一个二维坐标系内独立的点,相似度越高,两点之间距离越近,反之距离越远,计算公式如式(4)。其中 w 代表使用 TFIDF 算法计算出来的词汇_文档特征值权重。

$$sim(d_i, d_j) = \frac{\sum_{k=1}^n (w_{ik} \cdot w_{jk})}{\sqrt{\sum_{k=1}^n w_{ik}^2 \cdot \sum_{k=1}^n w_{jk}^2}} \quad (4)$$

3.2.2.3 K-means 聚类算法实现

K-means 算法^[8,9]是一种经典的基于划分方法的聚类算法。该算法的主要过程是对若干个对象进行划分,最终形成 K 个簇,每一个簇就代表一个聚类。第一步,选择任意 K 个对象作为这几个簇的质心, K 就代表了最终希望得到的聚类数,然后将每一个点都指派到离它本身最近的质心。第二步,根据指派到同一个簇的所有点的坐标,更新这个簇的质心。然后重复第一步,直到所有点所属的簇都不再变化为止。到这时,被指派到同一个聚类下的点就属于同一个聚类了,这些点构成了 K-means 算法的最终成果。下面给出 K-means 算法的伪代码实现。

选择 K 个点作为初始质心

repeat

将每个点指派到最近的质心,形成 K 个簇

重新计算每个簇的质心

until 质心不再发生变化

(下转第 512 页)

快速算法的网状聚类改进算法。该算法由于在每个合并点都计算了模块性,这就保证了聚类结果的正确性。但是该算法的时间复杂度较高是一个致命的弱点,今后的研究工作将主要围绕这一点展开。

参考文献

- [1] 孙吉贵,刘杰,赵连宇. 聚类算法研究[J]. 软件学报, 2008, 19(1): 48-61

(上接第 490 页)

图 4 为图 3 所示算法流程的图形方式。其中 A、B、C、D、E 5 个点表示 5 个已经被 TFIDF 算法向量化的文档,两个实心灰点表示两个聚类的质心。由于 TFIDF 算法和相似度计算公式的存在,因此可以计算出任意两个点之间的距离,K-means 算法就是根据这个距离计算每个文档距离聚类质心的距离,并将每个文档分配给离它最近的聚类(第二幅图),分配之后将每个聚类的质心按它所拥有的文档坐标更新(第三幅图),并重复第一步(第四幅图),直到所有文档所属的聚类都不再变化为止(第五幅图)。

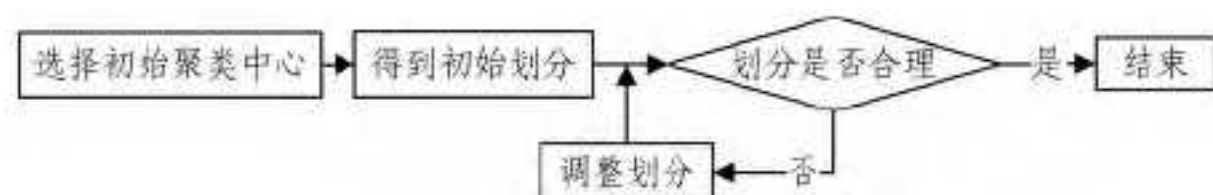


图 3 K-means 算法流程

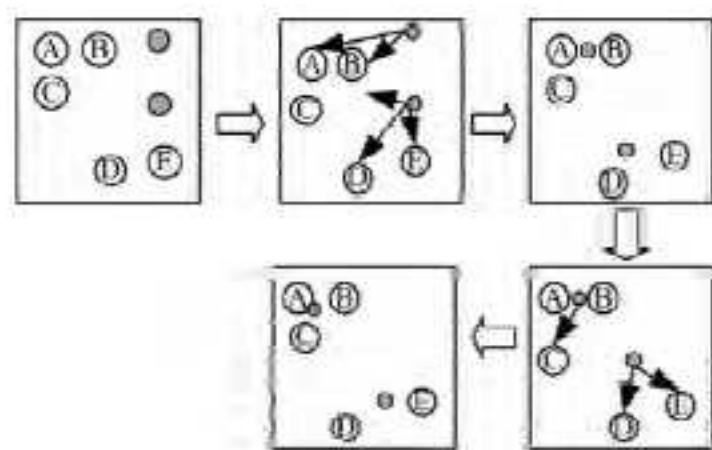


图 4 K-means 算法的模型演示

3.2.3 词库搜索

词库搜索是语义搜索引擎的扩展功能之一。它是根据用户提交给系统的关键词,连接 JWordNet^[10-12] 开源接口,并向 WordNet^[13] 词库发起请求,WordNet 会根据关键词和词性检索符合语义的相关词,该结果即为基于词典搜索引擎的搜索结果。

3.2.4 偏好搜索

该模块同目前的基于统计学的搜索引擎原理相似,它的实现建立在这样一个事实基础之上:同一个用户在相近一段时间内的搜索关键词同该用户的需求是呈正相关的。其实现原理是将用户提交的关键词进行分词,然后到以前的日志记录中进行比对,如果发现有符合要求的结果,则寻找那次搜索中用户在短时间内还搜索过的其它关键词,然后把这些关键词返回给用户。

结束语 本文使用了 Lucene 搜索引擎作为索引搜索的核心类库,使用 K-means 算法和 TFIDF 算法作为聚类搜索引擎的核心算法,并且引入了偏好搜索和基于词典的语义搜索,可以完成高价值的关键词推介,并具备了一定程度上的“语义搜索引擎”^[14] 特征。但它的使用体验依然很大程度上依赖于

- [2] Newman M E J. Fast algorithm for detecting community structure in networks[J]. Phys Rev E, 2009, 69(6): 66-69
- [3] Acqueen J. Some methods for classification and analysis of multivariate observations [C] // Proceedings of the fifth Berkeley Symposium of Math and Statist. [S. l.]: s[.n.], 2012: 281-297
- [4] 李东琦. 聚类算法的研究[D]. 成都: 西南交通大学, 2010
- [5] 李明华, 刘全, 刘忠, 等. 数据挖掘中聚类算法的新发展[J]. 计算机应用研究, 2010, 25(1): 60-62

训练文本库的选取。在本系统中,使用了训练文本库与索引文本库相分离的做法,用以提高聚类效率,但这都必须建立在训练文本库足够“精选”的基础上,即聚类搜索结果完全取决于训练文本库是否足够精炼,这其中的差距很大。另外,更新聚类需要耗费大量的计算机资源和时间,所以训练文本库并不能频繁刷新,这也就导致了如果在一段时间内出现某个相当风靡的词汇,聚类搜索引擎是无法将这些词及时列入聚类数据库的,时效性较差。

总之,聚类搜索是一个值得研究和期待的技术,但同大多数技术一样,它也有自己的缺陷和不足,需要不断地去弥补和完善才能满足用户的需求。

参考文献

- [1] 谷照升. RIA 技术解析[J]. 长春工程学院学报: 自然科学版, 2010, 11(1): 85-88
- [2] 侯丽. Web 2.0 的特性及对信息服务的创新性思考[J]. 图书馆建设, 2008(1): 66-69
- [3] 熊回香, 陈姗, 许颖颖. 基于 Web 3.0 的个性化信息聚合技术研究[J]. 情报理论与实践, 2011, 34(8): 95-99
- [4] 刘兴宇. 基于倒排索引的全文检索技术研究[D]. 武汉: 华中科技大学, 2004
- [5] 吴洁明, 冀单单, 韩云辉. 基于 Web 的 DCI 垂直搜索引擎的研究与设计[J]. 计算机工程与设计, 2013, 34(4): 1481-1487
- [6] Tan Pang-ning, Steinbach M, Kumar V. 数据挖掘导论[M]. 北京: 人民邮电出版社, 2011
- [7] 施聪莺, 徐朝军, 杨晓江. TFIDF 算法研究综述[J]. 计算机应用, 2009, 29(1): 167-170
- [8] 吴凤慧, 成颖, 郑彦宁, 潘云涛. K-means 算法研究综述[J]. 现代图书情报技术, 2011, 205(5): 28-35
- [9] 张睿. 基于 k-means 的中文文本聚类算法的研究与实现[D]. 西安: 西北大学, 2009
- [10] 郑廷, 郑诚. 基于 Lucene 的语义检索系统[J]. 计算机工程, 2008, 34(16): 92-94
- [11] 王学松. Lucene+nutch 搜索引擎开发[M]. 北京: 人民邮电出版社, 2008
- [12] 徐会生, 康爱媛, 何启伟. 深入浅出 Ext JS[M]. 北京: 人民邮电出版社, 2009
- [13] 翟延冬. 基于 WordNet 的短文本语义网挖掘算法研究[D]. 长春: 吉林大学, 2012
- [14] 张体首, 蔡明. 语义搜索引擎概念模型[J]. 微电子学与计算机, 2007, 42(3): 171-174