

doi: 10.3969/j.issn.1007-7162.2014.03.005

大数据环境下竞争情报系统的研究与实现

王 勇, 许钟涛, 王 瑛

(广东工业大学 计算机学院, 广东 广州 510006)

摘要: 为了让企业在大数据的环境下及时获得准确和可靠的竞争情报, 提出了基于大数据的企业竞争情报系统的构建. 通过 B/S(浏览器/服务器) 架构, 运用聚焦爬虫工作原理的网络爬虫进行情报信息的采集, 并用基于 Hadoop 的 KNN 算法进行情报信息的处理和分类, 克服了大数据下 KNN 分类算法时间复杂度高等问题, 同时系统支持用户个性化定制.

关键词: 竞争情报系统; 大数据; Hadoop; K 近邻算法

中图分类号: TP391

文献标志码: A

文章编号: 1007-7162(2014)03-0027-05

Research and Implementation on Competitive Intelligence System Based on Big Data

Wang Yong, Xu Zhong-tao, Wang Ying

(School of Computers, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: To help enterprises to obtain timely, accurate and reliable intelligence information under the environment of big data, it proposed how to build an enterprise competitive intelligence system, based on big data. It used web crawlers of focused crawler works to collect intelligence information, and employed the Hadoop-based KNN algorithm to process and classify the intelligence information via the B/S (Browser/Server) framework, which can solve the problem of high time complexity of the KNN classification algorithm in a big data environment and other issues. The system also allows users some customization, making it more individualized.

Key words: competitive intelligence system; big data; Hadoop; KNN

竞争情报系统被认为是本世纪企业最重要的竞争工具之一,可以帮助企业及时获得准确和可靠的竞争情报.在大数据时代,企业提高利润的支撑点主要是来自数据分析.企业不再局限于对现有数据的分析和监测,而是注重于对企业未来发展趋势的分析和预测,从而察觉行业发展细微的变化,以期增强企业的竞争力.

随着信息技术的飞速发展,网络信息正以指数级数增长,其数量已经超出了人们能够接受的范围,而信息爆炸给竞争情报系统带来了严峻的挑战.当前基于大数据的竞争情报的研究大部分还是处于理论研究阶段,如黄晓斌等^[1]提出基于大数据的企业竞争情报系统模型构建并分析系统模型的结构和作用;顾涛^[2]提出了基于大数据的竞争情报协作分析研究,主要阐明大数据资源协作共享基本原则,从而

提出竞争情报协作分析的模型.本文结合上述的理论研究,实现一个基于大数据的竞争情报系统,主要的内容如下:首先是情报收集问题,信息的结构各不相同并且数量巨大,如果仅靠人工收集或者依靠一般的搜索工具,不但会错过很多有价值的重要信息,而且会浪费时间和资源在无用的信息上,从而导致无法提供有效准确的信息给决策层.竞争情报的信息采集模块就要解决此问题^[3].再者是采集的情报信息的处理和分析,随着信息量急剧增长,对于信息价值的开发利用程度却在下降,因此对于信息的处理和挖掘就显得尤为重要,考虑到本系统用的文本挖掘算法是理论成熟且简单常用的分类算法—KNN 算法,但是由于 KNN 算法带来的时间复杂度较高,对于大数据的分析任务可能无法很好完成.因此为解决传统挖掘算法对海量数据无法处理的问题,本

收稿日期: 2014-04-16

基金项目: 广东省教育部产学研结合项目(2012B040500005);广东省教育部产学研结合项目(2012B091100071)

作者简介: 王勇(1968-),男,副教授,博士,主要研究方向为应急监控和知识管理.

文引入 Hadoop 平台,将 KNN 算法移植到该平台,从而提高处理信息的效率,降低时间复杂度^[4]。

1 系统总体设计

1.1 系统目标

利用 Hadoop 平台对大数据的处理能力,对系统采集的信息用基于 Hadoop 的 KNN 算法进行处理分析,使得处理后的竞争情报信息可以快速、准确地显示在用户操作界面,让用户可以及时了解到自己所关注的信息。管理员通过系统,可以及时进行信息的采集,并对信息进行处理,普通用户通过系统可以个性化的订制竞争对手的信息情报,并进行检索。

1.2 系统总体架构

系统总体架构如图 1 所示。

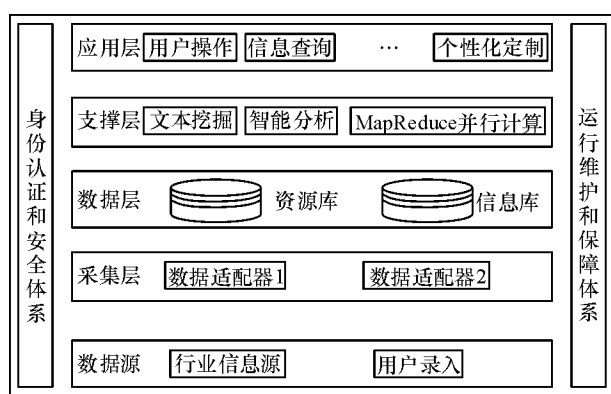


图 1 系统总体架构

Fig.1 System framework

1) 应用层.

在顶层提供统一信息门户,为客户提供各种信息和知识服务的窗口,同时也是平台管理的入口。

2) 支撑层.

提供文本挖掘和智能分析,以及基于 MapReduce 并行计算等功能模块,支持对海量数据的检索、挖掘、分析和检索等功能。

3) 数据层.

数据层汇聚行业资源信息,形成统一规划的数据库,便于应用服务。

4) 采集层.

建立各种数据源相应的数据适配器,实现对各种数据源的数据在线检测和自动采集。

5) 数据源.

数据源是平台建设的重要外部资源,是平台体系架构的一个虚拟层次,包括大量的行业信息源、用户录入企业信息等数据来源。

6) 身份认证和安全体系.

平台安全系统设计除了包含计算机网络、主机、应用系统等进行安全规划之外,系统运行的安全设计也是安全规划的重要内容之一。

7) 运行维护保障体系.

运行维护保障机制包含统一性和标准性、公众服务性、专业性、权威性。因此,运行维护机制要把握 3 个要素:

- (1) 要保持“库”的内容最新的现势程度;
- (2) 要适应社会对“库”多样化的服务要求;
- (3) 要提供简便、规范、畅通的基础数据输入/输出手段。

2 系统功能设计

主要运用文本挖掘技术,将文本挖掘算法在 MapReduce 上实现并行化,提高信息的分析处理效率,以期可以迅速地提供市场行情信息的综合分析结果。系统主要功能模块如图 2 所示。

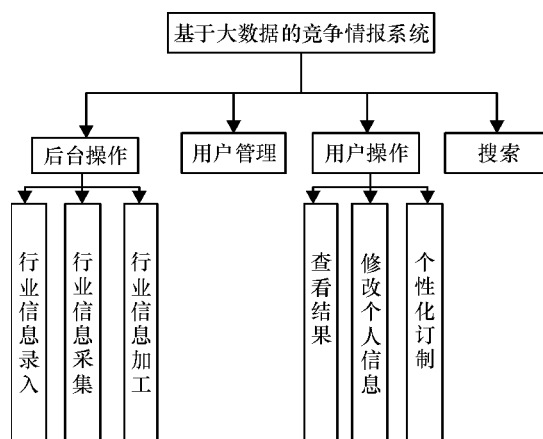


图 2 系统功能设计

Fig.2 Design of system functions

3 系统实现关键技术

3.1 信息采集爬虫技术

本系统采用的是聚焦爬虫,与通用爬虫区别为不追求大范围的覆盖,而是将目标设定为抓取与某一特定主题内容相关的网页,进而为面向主题的用户准备数据来源^[5]。

通用的爬虫是从一个或者多个初始的 URL 开始进行爬取,在获得初始 URL 抓取网页内容的过程中,不断从当前的网页中抽取新的 URL 并放入队列,直到满足系统设定的爬取深度条件后停止。聚焦爬虫的工作相比较为复杂,需要根据一定的网页分析算法过滤去除与主题无关的 URL,存储有用的并将之放入等待队列中,然后根据搜索条件从中选取

下一步要爬取的 URL,重复直到满足停止条件为止^[4]。同时,所有被爬虫爬取过的网页将会被存储起来,然后通过分析、过滤并建立索引,以便后续的查询和检索。这个过程所得到的分析结果还可以对之后的抓取过程给出指导反馈。聚焦爬虫工作流程如图3所示。

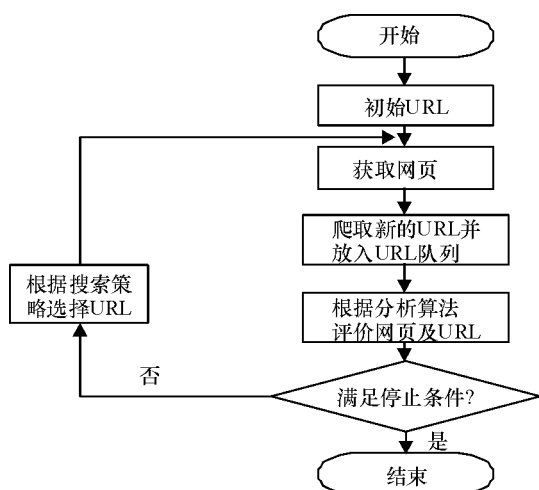


图3 聚焦爬虫工作流程

Fig. 3 Work process of focused crawlers

3.2 KNN 分类算法

最近邻算法^[7-8]是模式识别中广泛使用的分类方法,是模式识别非参数法中最重要的方法之一。K近邻算法是最近邻算法的一个推广,当 $k=1$ 时,就是最近邻算法。NN强调最近点的重要性,而KNN是从总体考虑,是一种更普遍的方法。

KNN的分类思想是给定一个未知文本,在训练数据中找出与其最相似的 K 个训练文本,即是这个未知样本的 K 个近邻。然后根据这 K 个近邻来确定未知样本所属的类别,可以把未知样本分到 K 个近邻最公共的类中,也可以分到 K 个近邻中权重最大的类中。 K 近邻算法的简单示意图如图2所示。

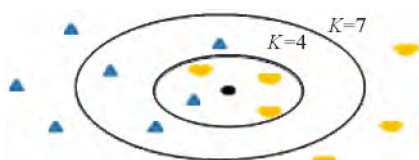


图4 KNN分类示意图

Fig. 4 Sketch map of KNN

3.3 KNN 分类算法的 MapReduce 并行化方法

KNN 算法尽管原理比较简单,但是其计算量很大,对其空间和时间要求都比较高^[9],所以提出将KNN 算法运用到 MapReduce 上进行并行化计算,以提高运行效率。

MapReduce 技术最开始是被 Google 用于大数据并行处理^[10],基本思想是将大数据集分割成无数的小数据集,然后每个数据集分别有集群中的一个 Map 函数执行计算任务,生成中间结果,从而作为 Reduce 函数的输入执行计算任务,得到结果。

MapReduce 的编程模型^[11]如图5所示。

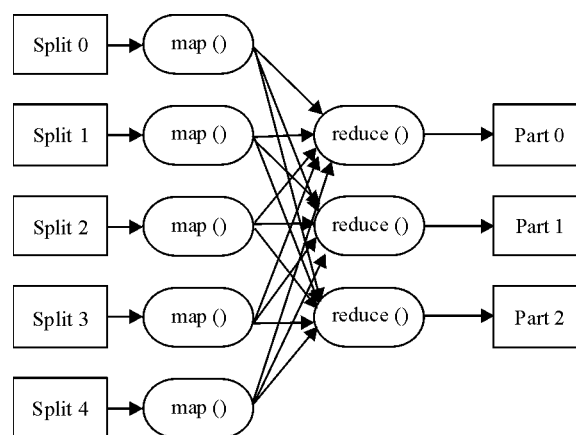


图5 MapReduce 编程模型

Fig. 5 Programming model of MapReduce

3.3.1 文档特征空间和相似度计算

本系统采用向量空间模型^[12]来描述所收集的情报信息文档,每个舆情文档都可以表示成一个特征向量 $D = D((t_1, w_1), (t_2, w_2), \dots, (t_n, w_n))$,其中 T_n 为特征项,即为特征词, w_n 为在 D 中的权值,特征权值计算公式采用 TFIDF 算法^[13-15],如式(1)所示。

$$W_{ik} = TF_{ik} \times IDF_{ik} = \frac{D_k}{A_{ik}} \times \lg\left(\frac{N}{n_k} + 0.1\right). \quad (1)$$

其中 TF 表示一个特征词与某个文档的相关性, IDF 表示一个特征词表示文档主题的权重大小。 D_k 为特征词在文档出现的次数, A_{ik} 为文档所有特征词数。 N 为训练集的文档总数, n_k 为出现特征词的文档数。计算出 TFIDF 后,对于海量数据,其包含的特征词数目比较多,所以必须进行降维,通过对计算出的 TFIDF 值排序然后设定阈值,选取若干个特征词,同时,利用 Hadoop 平台的 MapReduce 进行架构从而实现 TFIDF 的并行计算,进而有效地提高运行效率,但是在 Hadoop 平台中,数据经过 Map 函数处理后默认是按照 key 值升序排序的,因此可以按照 $1/TFIDF$ 值得方法取得升序排列的前若干位数值即 TFIDF 按照降序排序的相应若干较大值^[16]。

当计算两个网页的相似度时,需要计算两个文档间的相似度。本文采用的是余弦相似度^[17],如式(2)所示:

$$\text{Sim}(d_i, d_j) = \frac{\sum_{k=1}^M W_{ik} \times W_{jk}}{\sqrt{(\sum_{k=1}^M W_{ik}^2)(\sum_{k=1}^M W_{jk}^2)}} \quad (2)$$

通过对 d_i 和 d_j 两篇文档的特征向量进行计算,可以得到文档间的相似度. 相似度越大,说明两篇文本相关程度越高,反之,相关程度越低. 相似度在 $[0,1]$ 之间取值,当两篇文本无关时,相似度为 0,相似性越高,则相似度趋向于 1.

3.3.2 KNN 分布式计算

KNN 算法根据不同文档之间计算的余弦相似度进而通过选取适当的 K 值,可以求出最近邻的 K 个文档. 由于任意两篇文档间的相似度计算和排序都互不相关,所以可以并行进行.

求任一类中的文本在不是自己本身类别中的 n 个最近邻的文档.

Map 阶段输入: (file, filename₁ = Sim₁; filename₂ = Sim₂; ...; filename_i = 0; ...; filename_n = Sim_n)

格式转换后,输出为 ((filename_i, Sim_{ij}), filename_j; type_j). 经过排序,Reduce 得到经过排序的上一阶段的结果,然后选择前 k 个样本,输出为 (filename, type).

所有的 Map 节点读取余弦相似度结果,并根据每个文本的类别,选出与自身类别不同的文本相似度并进行格式转换,结果输出为 ((filename_i, Sim_{ij}), filename_j; type_j). 其中 Sim_{ij} 表示文本间的相似度, type_j 表示测试文本分到的类别.

Map 输出结果中的 key 值是一个自定义的数组形式 Key(String, float), MapReduce 框架中 Map 过程和 Reduce 过程之间根据 Key 值进行排序,即按照 filename 排序,但因为需将同一个 filename 相关的 Sim 分配到同一个 Reduce 节点,所以要将相同 filename 的 Sim 值进行降序排列. 经过两次排序后,Map 的输出作为 Reduce 的输入,即 ((filename_i, Sim_{ij}), filename_j; type_j), 根据排序的结果,选择每个文档的前 K 个最近邻文档,从而将测试文档归入到 type 类中^[18].

4 运行结果

本系统没有针对某一特定行业,实验主要通过网上爬虫采集网页内容来进行,如计算机类、体育类等. 对于某一具体行业只要将相关的分类信息做修改配置即可.

4.1 用户界面

用户界面分为管理员界面和企业用户界面,管

理员界面比普通企业用户界面多一个后台管理模块. 企业用户主要的操作功能是查询结果、对竞争情报的信息录入以及用户信息的相关操作等,而管理员则除了具备企业用户的功能外,还可以对信息处理,如定期爬取用户订制的情报内容,对爬去内容进行预处理并做分类处理,同时将结果保存起来以备用户查询.

4.2 查看初步结果

通过用户操作模块可以查看用户所关注的竞争对手的情况,也可以查看系统自整理的信息. 当用户选择自己关注的选项时,可以查看整理归纳后的信息,了解自定义的信息. 当选择查看系统自整理信息时,企业用户可以看到除了自定义的领域外,系统其他用户所共同关注的信息,了解多点信息,拓宽自己的领域.

4.3 查看详细的内容

企业用户点击经过分类后的信息,可以查看信息的具体内容,内容显示在弹出框中. 如需查看信息的网页的初始内容,点击信息的 URL 可以跳转到信息的原页面. 同时也可以看到该信息被归纳后的类别.

4.4 分析可视化界面

通过对情报信息的一系列操作后,根据结果生成关于某一具体类别的技术网络图,通过技术网络图可以帮助企业了解信息间的关联,从而更好地作出决策,网络图如图 6 所示.



图6 技术网络图

Fig. 6 Diagram of technological networks

5 结束语

对大数据的处理不仅仅对企业竞争决策起到重

要作用,也可以提高企业的竞争力.而竞争情报系统更能为企业在大数据环境下提供有效的决策和有价值的信息.本文提出将基于 Hadoop 的 KNN 算法用于企业竞争情报系统中,从而可以缓解对于竞争情报的传统处理方式带来的弊端,降低系统运行的时间和空间复杂度,进而使本系统更加有效地适应当今大数据环境.

参考文献:

- [1] 黄晓斌,钟辉新. 基于大数据的企业竞争情报系统模型构建 [J]. 情报杂志,2013,32(3):37-43.
Huang X B, Zhong H X. Build enterprise competitive intelligence system model based on big data [J]. Journal of Intelligence, 2013, 32(3): 37-43.
- [2] 顾涛. 基于大数据的竞争情报协作分析研究 [J]. 情报科学,2013,31(12):114-118,135.
Gu T. Research on collaboration analysis of competitive intelligence based on big data [J]. Information Science, 2013, 31(12): 114-118, 135.
- [3] 王卫平,郭长旺. 基于文本挖掘的企业竞争情报系统 [J]. 现代情报,2004,9(9):188-222.
Wang W P, Guo C W. Corporation competitive intelligence system based on text mining technology [J]. Journal of Modern Information, 2004, 9(9): 188-222.
- [4] 文思. 基于 Hadoop 的 K 近邻分类算法研究 [D]. 广州: 华南理工大学计算机科学与工程学院,2011.
- [5] Pinkerton B. Finding what people want: Experiences with the web crawler [C] // Proceedings of the Second World-Wide Web Conference. Chicago Illinois [s. n.], 1994.
- [6] 周立柱,林玲. 聚焦爬虫技术研究综述 [J]. 计算机应用,2005,25(9):1965-1969.
Zhou L Z, Lin L. Survey on the research of focused crawling technique [J]. Computer Applications, 2005, 25(9): 1965-1969.
- [7] 程显毅,朱倩. 文本挖掘原理 [M]. 北京: 科学出版社,2010:21-22.
- [8] 滕少华,樊继慧,陈潇,等. 基于 KNN 的多组合器协同挖掘局域气象数据 [J]. 广东工业大学学报,2014,31(1):25-31.
Teng S H, Fan J H, Chen X, et al. The application of KNN-based multi-classifiers in mining local area meteorological data [J]. Journal of Guangdong University of Technology, 2014, 31(1): 25-31.
- [9] 刘慧. 基于 KNN 的中文文本分类算法研究 [D]. 成都: 西南交通大学信息科学与技术学院,2010.
- [10] Lammel R. Google's mapReduce programming model-revisited [J]. Science of Computer Programming, 2008, 70(1):1-30.
- [11] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters [J]. Communications of the ACM, 2008, 51(1):107-113.
- [12] 王之鹏. Web 文本分类系统中中文预处理技术的研究与实现 [D]. 南京: 南京理工大学计算机科学与技术学院,2009.
- [13] Chang Y H. Automatically constructing an effective domain ontology for document classification [J]. Computer Technology and Application, 2011, 2(3):182-189.
- [14] Kim C, Shim K. Automatic template extraction from heterogeneous web pages [J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(4):612-626.
- [15] 邹丽娜,凌捷. 一种基于特征提取的二级文本分类方法 [J]. 广东工业大学学报,2012,29(4):65-68.
Zou L N, Ling J. A two-level text classification based on feature extraction [J]. Journal of Guangdong University of Technology, 2012, 29(4): 65-68.
- [16] 徐雅斌,李卓,陈俊伊. 基于改进 K 最近邻分类算法的不良网页并行识别 [J]. 计算机应用,2013,33(12):3368-3371,3379.
Xu Y B, Li Z, Chen J Y. Parallel recognition of illegal web pages based on improved KNN classification algorithm [J]. Computer Applications, 2013, 33(12): 3368-3371, 3379.
- [17] 张宁,贾自艳,史忠植. 使用 KNN 算法的文本分类 [J]. 计算机工程,2005,31(8):171-172.
Zhang N, Jia Z Y, Shi Z Z. The text classification used KNN algorithm [J]. Computer Engineering. 2005, 31(8): 171-172.
- [18] 闫永刚,马廷淮,王建. KNN 分类算法的 MapReduce 并行化实现 [J]. 南京航空航天大学学报,2013,45(4):550-555.
Yan Y G, Ma T H, Wang J. Parallel implementing KNN classification algorithm using MapReduce programming mode [J]. Journal of Nanjing University of Aeronautics & Astronautics, 2013, 45(4): 550-555.