基于语义树与 VSM 的主题爬取策略研究

张 金,倪晓军
(南京邮电大学 计算机学院,江苏 南京 210003)

摘 要:主题爬虫主要用于解决用户的定制化搜索需求,即在日益增长的网络数据中快速、有效、准确地选取用户关注的主题内容进行爬取。提高爬取特定信息的准确性,需要对网页的内容相关度进行主题相关判断,而主题爬虫关注的核心问题就是相关度计算,但现有的改进算法大多采用人工智能和机器学习等技术,不仅引起算法复杂度的提高,而且提升效果有限。为此,提出了一种基于语义树与 VSM 的主题爬取策略,将语义相似度的计算加入到内容相关度计算与链接排序中,并通过对策略中算法细节的改进优化相关度的主题判别。实验结果表明,使用基于语义树与 VSM 爬取策略的主题爬虫可将爬行路线一直保持在相关度较高的网页链接中,对网页链接进行了相关与不相关的有效分类,显著地提高了爬取的准确率。

关键词:主题爬虫;语义树;向量空间模型;内容相关度;链接排序
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2017)11-0066-05
doi:10.3969/j.issn.1673-629X.2017.11.014

Research on Topic Crawling Strategy Based on Semantic Tree and VSM

ZHANG Jin, NI Xiao-jun
(College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: Topic crawler is mainly adopted to solve the customized search needs of users, that is to select the concerning topics of users for crawling quickly, effectively and accurately in the growing network data. In order to improve the accuracy of crawling specific information, the relevance of the content of the page needs to be subject-related judgments while the main concern of the topic crawler is the correlation calculation. But the most of the existing improved algorithms adopt techniques like artificial intelligence and machine learning, which not only improve their complexity, but also own limitations in effect enhancement. Therefore, a topic crawling strategy based on semantic tree and VSM is proposed and the semantic similarity calculation is added to the content relevance calculation and link ranking to optimize the subject discrimination of relevance through the improvement of detail of the algorithm in the strategy. Experimental results show that it can always keep the crawl course in the link of the web page with high relevance, which has effectively classified the web links relevant or not and significantly improved accuracy of crawling.

Key words: topic crawler; semantic tree; VSM; content relevance; link ranking

1 概 述

主题爬取是指遵循一定的规则对相应主题进行爬取操作,有别于传统的爬取策略(爬取所有的页面以供用户后期的检索,信息范围广泛),而主题爬取尽可能多地爬取只与主题相关的网页,避免其他无关页面,信息领域特定,结果专业,提高了爬虫爬取的效率。在互联网的飞速发展下,网络上的信息资源呈指数级增长,爬取的信息量也随之增长,如何在海量数据中为用户提供个性化需求的信息成为当下爬取研究的重点。

主题爬取基于主题爬虫,主题爬虫[1]在 1999 年被 Chakrabarti 等提出,主要用于解决爬取特定主题、个性化需求的网页时查准率不高的问题。主题爬虫在爬取过程中对网页判断主题是否相关,相关则抓取,通过这样的判断减少了无关页面的抓取,从而降低了带宽、时间以及存储空间的需求,既提高了抓取的准确率,又提高了系统的抓取效率。与通用爬虫相比,主题爬虫的网页相关度判断需要解决链接的主题相关性、链接优先级等问题,这就需要实现基于主题搜索场景设计的爬虫。目前传统的主题爬虫对于主题相关性算法主要从两方面进行分析:内容与链接。在内容分析方面,主

收稿日期:2016-11-12 **修回日期**:2017-02-23 **网络出版时间**:2017-07-19
基金项目:教育部专项研究项目(2013116)
作者简介:张 金(1992-),男,硕士研究生,研究方向为大数据、网络爬虫技术;倪晓军,教授,研究方向为嵌入式。

要通过计算主题与页面的相似度来确定抓取队列,主要代表有 Fish-Search[2] 算法和 Shark-Search[3] 算法。而链接分析主要通过链接间的相互引用决定链接的重要性排序,代表算法有 PageRank[4] 与 HITS[5]。这两类算法都得到了研究人员的大量关注,并对此进行了许多改进,但这些算法大都只是对算法的适用度进行增强,因此如何在对用户主题的特定需求下,既准确又高效地进行网页抓取成为研究重点。

在一般主题爬虫的基础上,提出一种改进的语义树[6] 与 VSM[7] 相结合的主题爬取策略,优化主题相关度计算与链接排序,提高抓取的准确率。在向量空间模型计算页面相似度的基础上,发挥语义树在计算内容语义相似度的关键性作用,判断页面相关度,使用多次链接的相似度计算对链接进行候选优先级排序,实现了在个性化场景下爬取的准确与高效。

在信息采集领域,相关度理论的研究一直是焦点,尤其是主题爬虫中的网页相关度算法的研究与改进。传统的相关度算法主要分为基于链接的重要度分析和基于内容的相似度计算分析两大类。基于链接分析主要是通过 PageRank 等算法来建立主题相关度计算模块,PageRank 算法通过链接到页面的链接重要性递归计算来计算页面的等级,链接的页面越多,计算得到的等级也越高。但在 PageRank 算法中,由于出度链接的不确定性、用户点击概率的非均等性等问题,抓取时会出现"主题漂移",导致无效抓取,浪费资源。基于内容分析是依据向量空间模型,首先依据主题特征词生成主题特征向量,再将页面关键词权重用 TF * IDF 表示,然后对页面向量与主题词特征向量进行相似度计算,通过设定相应的阈值对网页进行相关匹配,解决无关页面的资源占用,但是如果页面对关键词进行虚假设置,就会导致相关度计算结果的不准确,从而导致误判现象,同时也忽略了网络结构的作用。文献[8]考虑了关键词出现位置的差异性,根据关键词出现的不同位置赋予不同权重系数,这样能更加精确地描述页面间相关度,抓取更为准确。但是并没有考虑特征词之间的语义差别,而语义上的差别肯定会干扰页面相关度的判断,进而影响相关页面的确定与抓取。

主题爬虫在之前发展的基础上,加入了更多技术手段来优化相关度的计算,比如遗传算法、蚁群算法、神经网络等。文献[9]通过关键词来定义页面信息,使用在线增量学习的方式进行链接爬行,对锚文本、URL 串、父子页面间的关系进行页面综合价值的计算并排序。这类算法使得抓取的精度更高,不容易产生主题漂移现象,但同时加大了算法复杂度,降低了抓取效率。文献[10]通过对 URL 中的关键字出现次数与父页面相关性进行总相关性计算来确定链接的相关得分。在此基础上,文献[11]通过朴素贝叶斯分类算法模型计算链接的相关度,但是无法对关键词不同的页面进行有效处理,不管页面的主题是否相似。文献[12]将语义计算与主题爬虫结合起来,提高了相关度计算的准确度以及抓取的准确性,有效地过滤无关页面,但其时间与空间复杂度相较于其他有了明显提高。文献[13]使用语义来对页面的语义关系进行评估,并对相关度的页面赋予高优先级进行抓取。在语义关系的比较中,最常用的是语义树,文献[14]在对基于语义树的语义计算方法进行大量研究的基础上,提出了计算各个特征词向量之间的相似度来判断词的相似性,但是针对词的相似判断基础是特征词所在的上下文也是相同的,这显然会出现比较明显的错误。

文献[15]通过大量实验研究对比了不同主题算法的效果后发现,无论是基于分类增加型学习的算法,还是基于遗传和神经网络的算法,实际上抓取效果提升平平,并没有对抓取进行实际明显的优化。

为了进一步提升主题抓取的准确率和效率,对语义树在计算相似度方面的算法进行深度的分析与挖掘,将基于语义树的语义判断加入算法中,同时仍采用 VSM 计算页面间基本的相似度,将两者结合以提高相似度计算的准确性。通过对父子 URL 链接的多次相关度计算与语义距离综合计算页面得分并对候选 URL 进行排序,从而达到准确率提高的效果。

2　主题爬取策略

2.1　内容相关度计算

内容相关度计算就是对即将爬取的页面主题相关程度进行计算,对相关度高的页面进行抓取,以尽可能避免发生"主题漂移"现象,主要涉及如何对主题与网页特征的相似比较。首先需要确定的是爬取主题,其次需要将主题信息转化为可计算的模式,通过主题特征词的转化建立主题特征向量:

$$\boldsymbol{T} = \{t_1, t_2, \cdots, t_n\} \tag{1}$$

其中,n 为主题特征值的个数,t_n 为特征值的权重。

主题特征向量可以通过两种方式进行设定,一种是通过人工设定主题的特征值与特征权重来形成向量,这里的特征值指的就是主题特征词;另外一种就是通过对抓取的初始页面进行分析得到主题特征向量。

采用基于语义相似度与 VSM 的页面特征相似度算法进行相关度计算,首先需要提取页面文本信息,即对网页特征值进行提取并映射成设定的网页特征向量:

$$\boldsymbol{W} = \{w_1, w_2, \cdots, w_n\} \tag{2}$$

考虑到出现在不同位置的关键词,对网页所起的

重要性也不同,对用于特征提取的 TF 计算公式进行改进:

$$w_{\mathrm{tf}} * (1 + \sum_{v_i \in V} \boldsymbol{w}_{v_i} * \boldsymbol{t}_{v_i}) \tag{3}$$

其中,w_{tf} 表示词频;$\boldsymbol{v}_i$ 表示不同位置的特征向量;$\boldsymbol{w}_{v_i}$ 表示对应位置的权重向量;$\boldsymbol{t}_{v_i}$ 表示关键词出现的次数向量。

通过对不同位置的关键词加权,比如锚文本、标题等,从而更加精确地描述了页面的主题,生成合理的特征向量。在得到网页特征向量之后,计算向量余弦距离:

$$\cos(\boldsymbol{T},\boldsymbol{W}) = \frac{\sum_{i=1}^{n} (\boldsymbol{t}_i * \boldsymbol{w}_i)}{\sqrt{(\sum_{i=1}^{n} \boldsymbol{t}_i^2) * (\sum_{i=1}^{n} \boldsymbol{w}_i^2)}} \tag{4}$$

其中,$\boldsymbol{T}$ 表示主题特征向量;$\boldsymbol{W}$ 表示网页的特征向量。

但是单纯地将特征向量的余弦距离作为页面内容相关度得分是不可靠的,因为这里的关键词受到页面噪声的干扰较大,同时也存在人为设置关键词的问题,容易引发误判。因此需要进一步研究语义树在计算相似度方面的作用,考虑语义的影响因素,以此解决 VSM 的局限性,提高内容相关度计算的精确性。将语义树应用到语义相似度计算中,语义树中的每一个节点都是语义相关的,不同的是语义树节点间的父子关系不同,所计算的节点距离不同。这里通过各个特征词在语义树上的节点距离来计算语义相似度,首先计算各个特征词的语义相似度:

$$\mathrm{SemanticSim}(w_i,w_j) = \frac{\alpha}{\alpha + \mathrm{dist}(w_i,w_j)} \tag{5}$$

其中,α 取节点相似度为一半的距离;$\mathrm{dist}(w_i, w_j)$ 计算节点间的距离以对相似度进行校正。首先设定节点 X,Y,P,其中节点 P 为节点 X 与 Y 的距离两个最近的共同祖先节点,有:

$$\mathrm{dist}(w_i,w_j) = \begin{cases} |\mathrm{dep}(w_j) - \mathrm{dep}(w_i)|, w_j \text{ 为 } w_i \text{ 的祖先节点} \\ \mathrm{dep}(w_j,P) + \mathrm{dep}(w_j,P), \text{其他} \end{cases} \tag{6}$$

$$\mathrm{dep}(X,Y) = |\mathrm{dep}(X) - \mathrm{dep}(Y)| \tag{7}$$

在此基础上,对特征向量的 n 个特征值相似度求和,计算网页特征向量的语义相似度:

$$\mathrm{SemanticSim}(W_i,W_j) = \frac{1}{n^2} \sum_{p=1}^{n} \sum_{q=1}^{n} \mathrm{SemanticSim}(w_p,w_q) \tag{8}$$

最终综合特征向量的余弦距离与语义相似度,得出网页特征向量的相似度,作为内容相关度的得分:

$$\mathrm{Sim}(P,W) = \sqrt{\frac{\mathrm{SemanticSim}(\boldsymbol{T},\boldsymbol{W})^2 + \cos(\boldsymbol{T},\boldsymbol{W})^2}{2}} \tag{9}$$

2.2 链接队列排序

确定链接队列顺序,对于优先爬取相关度高的页面来说尤为重要,同时也是主题爬虫研究的关键问题之一。优先级队列的确定可以保证主题抓取始终保持在高相关度的页面中。传统的链接重要分析使用在 PageRank 算法基础上进行改进的策略,PageRank 算法建立在用户的点击操作不仅是随机的,而且对每一个链接来说点击的概率是均等的基础上,这种情况在实际当中并不普遍存在,而且依此分析出来的页面也并非都是主题相关的。因此,提出依据页面子链接的分析为基础的链接排序算法。根据大量的研究表明,子链接通常与页面内容也有一定关系,因此可以考虑子链接的相关度对当前的链接相关度进行加权,但是页面中的子链接并不都是有用的,有些页面含有大量的广告链接、导航链接等,这些无效的链接需要剔除掉,以减少对当前链接的影响,因此最终选择的链接的优先级得分计算公式为:

$$\mathrm{Score}_{\mathrm{url}} = \begin{cases} \delta * \mathrm{Sim}(\boldsymbol{T},\boldsymbol{V}), \text{大部分子链接相关度低} \\ \delta * \mathrm{Sim}(\boldsymbol{T},\boldsymbol{V}) + (1-\delta)(\dfrac{\sum_{i=1}^{m} \mathrm{Sim}(\boldsymbol{T},v_i)}{m}), \text{大部分高} \\ \delta * \mathrm{Sim}(\boldsymbol{T},\boldsymbol{V}) + (1-\delta)(\dfrac{\sum_{i=1}^{n} \mathrm{Sim}(\boldsymbol{T},v_i)}{n}), \text{其他} \end{cases} \tag{10}$$

其中,$\boldsymbol{T}$ 为主题特征向量;$\boldsymbol{V}$ 为当前链接特征向量;v_i 为子链接;m 为有效链接数;n 为总链接数;δ 为权重因子;使用式(9)计算相似度。

对于上文提出的问题,分三种情况进行计算。当子链接大部分相关度高时,抛弃相关度低的链接后再计算链接的得分;当大部分子链接都不相关时,忽略子链接对当前链接的加权;其他情况则正常计算链接的得分值。

通过加权计算所得的得分进行链接队列的排序与更新,保证了抓取时的爬行路线可以保持在较高相关度的链接中,确保抓取的都是主题相关高的页面,从而提高了爬取的准确性,避免了无效抓取。同时由于加权使链接延伸,有利于爬虫进行隧道穿越。

2.3 算法流程实现

主题爬虫的特点就是使得爬虫永远在主题相关的页面上爬行,抛弃不相关的页面。基于语义树与 VSM 的内容相关度计算成了主题爬虫的重点,采用当前链接与子链接的相关度计算对链接进行排序,从而优化

爬行路线。提出的爬取策略主要在两个部分进行改进,一个是内容相关度计算,一个是链接队列排序,依据该算法的爬取结构如图 1 所示。

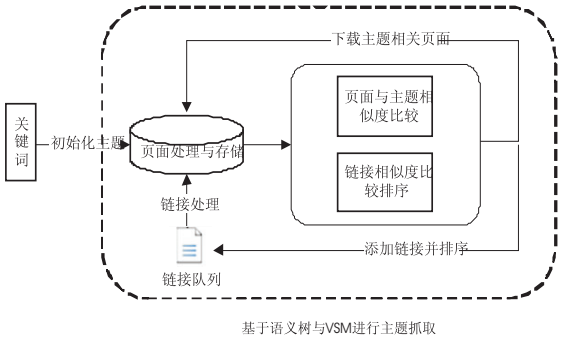

图 1　爬取结构图

首先确定搜索主题,生成可用于相关计算的主题特征向量,根据链接队列爬取页面并抽取特征词生成页面的特征向量,同时加入语义判断计算判断页面是否相关,并将主题相关的页面进行存储,同时提取页面中的子链接,根据链接的相关度得分进行优先级排列并更新链接队列,然后重复此抓取过程直到达到设定的停止条件。

算法流程如下:

(1)初始化主题特征向量,人工指定或使用训练集训练;

(2)从链接队列按序取得网页链接抓取页面;

(3)对各个页面进行特征向量提取,通过式(9)计算相似度并进行比较,取得相关性高的页面并抓取存储,更新页面库;

(4)对页面中的种子链接通过式(10)遍历计算与主题的相似度得分,并更新排序链接队列;

(5)重复步骤(2)~(4),直到达到系统指定的结束条件,抓取的总页面数或者抓取深度。

3　实验与分析

通过查全率与查准率对主题爬取策略进行衡量。查全率是指网络中所有相关网页中被主题爬取的网页所占的比例,要计算查全率,就需要知道整个网络中的网页资源,而这个在实际实验中基本上是不能的,虽然可以通过公式模拟计算查全率,但意义不大。查准率是指在所有已爬取的相关页面中真正主题相关的页面所占的比例。计算公式为:

$$S_c = \frac{p^*}{p} \tag{11}$$

其中,p 为已爬取的所有页面数;p^* 为其中主题相关的页面数。

除查准率之外,算法提升对时间效率上的影响也是评价一个主题爬取策略优越性的重要指标。

综上所诉,在比较不同策略查准率的同时,也考虑对同一时间内不同策略抓取的相关页面的数量进行比较,综合比较策略的准确性与效率。在确定好评价标准的基础上,将基于传统 VSM 策略的爬虫与提出的策略进行比较,以验证该策略的有效性。

实验以教授、专家、学者、报告、讲座、汇报为关键词集,通过搜索引擎获取部分链接与人工设定链接作为初始链接,同时设定爬取深度为 3,设定权重因子 δ 为 0.8。保证在两种策略的主题、初始链接、停止条件相同的情况下进行比较实验,结果如图 2 和图 3 所示。

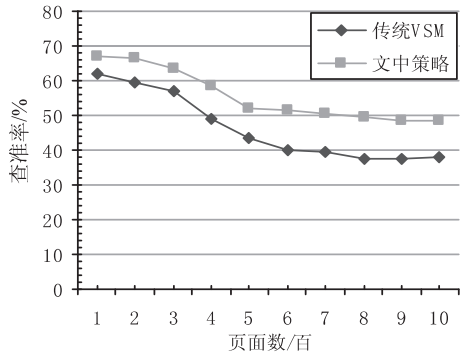

图 2　查准率对比

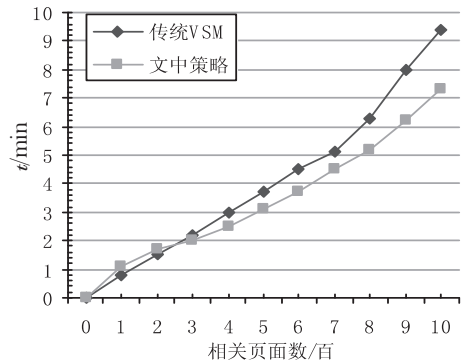

图 3　时间对比

从图 2 中可以看出,改进策略在爬取相同页面数时,爬取准确率普遍提升。传统策略平均查准率为 46.37%,改进策略平均查准率为 55.55%,提高了 9.18 个百分点。同时还可以发现,随着页面数的增加,查准率随之下降,而传统策略下降得更快,这是因为提出策略加入了语义相似度的计算,保证了爬虫可以在相关度更高的页面中爬行。

从图 3 中可以看出,改进策略在获取相同相关页面数的情况下,一开始所耗时间高于传统策略,而后逐渐缩小差距,这是因为该策略改进了相关度算法、链接排序算法,提高了算法复杂度,继而增加了开销,使得在查准率高的情况下,总数会相对减少,但是随着页面数的增加,优势也逐渐凸显。

4　结束语

为了进一步提高主题爬虫的抓取准确率,提出了

一种基于语义树与 VSM 的主题爬取策略,以优化爬虫的爬行路线,尽可能多地避开无关页面。通过将语义树应用于内容相关度计算,解决了使用传统向量余弦距离计算页面相似度没有考虑语义的问题。另一方面,分析子链接的相关度对当前链接相关度得分的影响,通过对链接进一步的分析,使得链接排序更加合理,有利于爬虫穿越隧道。实验结果及其分析均表明,该策略进一步提高了抓取的准确性,减少了无关的爬取存储操作。但语义相似度的计算需要依赖于语义树的构建,且该策略本身也没有涉及对爬取效率的提升。由于爬取效率与准确率一样,对主题爬取至关重要,因此提升主题爬取准确率将成为下一步工作中的研究重点。

参考文献:

[1] Chakrabarti S, van den Berg M, Dom B. Focused crawling: a new approach to topic-specific web resource discovery[J]. Computer Networks, 1999, 31(11-16): 1623-1640.

[2] de Bra P M E, Post R D J. Information retrieval in the world-wide web: making client-based searching feasible[J]. Computer Networks and ISDN Systems, 1994, 27(2): 183-192.

[3] Hersovici M, Jacovi M, Maarek Y S, et al. The shark-search algorithm. An application: tailored Web site mapping[J]. Computer Networks and ISDN Systems, 1998, 30(1-7): 317-326.

[4] Page L. The PageRank citation ranking: bringing order to the web[D]. California: Stanford University, 1998.

[5] Kleinberg J M. Authoritative sources in a hyperlinked environment[J]. Journal of the ACM, 1999, 46(5): 604-632.

[6] 张 亮,尹存燕,陈家骏. 基于语义树的中文词语相似度计算与分析[J]. 中文信息学报,2010,24(6):23-30.

[7] 刘冬明,杨尔弘. 话题内相关文本的内容计算[J]. 中文信息学报,2015,29(5):98-103.

[8] Pal A, Tomar D S, Shrivastava S C. Effective focused crawling based on content and link structure analysis[J]. International Journal of Computer Science and Information Security, 2009, 2(1): 1-5.

[9] Aggarwal C C, Al · Garawi F, Yu P S. On the design of a learning crawler for topical resource discovery[J]. ACM Transactions on Information Systems, 2001, 19(3): 286-309.

[10] Hati D, Kumar A. An approach for identifying URLs based on division score and link score in focused crawler[J]. International Journal of Computer Applications, 2010, 2(3): 48-53.

[11] Hati D, Kumar A, Mishra L. Unvisited URL relevancy calculation in focused crawling based on Native Bayesian classification[J]. International Journal of Computer Applications, 2010, 3(9): 23-30.

[12] Ehrig M, Maedche A. Ontology-focused crawling of web documents[C]//Proceedings of the 2003 ACM symposium on applied computing. [s. l.]: ACM, 2003: 1174-1178.

[13] Ganesh S, Jayaraj M, Kalyan V, et al. Ontology-based web crawler[C]//International conference on information technology: coding and computing. [s. l.]: IEEE, 2004: 337-341.

[14] 于甜甜. 基于语义树的语句相似度和相关度在问答系统中的研究[D]. 济南:山东财经大学,2014.

[15] Mencaer F, Pant G, Srinivasan P. Topical web crawlers: evaluating adaptive algorithms[J]. ACM Transactions on Internet Technology, 2004, 4(4): 378-419.

(上接第 65 页)

[2] 凌 云,周华锋. 面向异构集群系统的动态负载均衡技术研究[J]. 计算机工程与设计,2008,29(12):3068-3070.

[3] Rai I, Alanyali M. Uniform weighted round robin scheduling algorithms for input queued switches[C]//IEEE international conference on communications. Helsinki, Finland: IEEE, 2001: 2028-2032.

[4] Kim J S, Lee D C. Weighted round robin packet scheduler using relative service share[C]//IEEE military communications conference. [s. l.]: IEEE, 2001: 988-992.

[5] 陈 伟,张玉芳,熊忠阳. 动态反馈的异构集群负载均衡算法的实现[J]. 重庆大学学报:自然科学版,2010,33(2):73-78.

[6] 朱虹宇,李 挺,闫健恩,等. 基于动态负载均衡的分布式任务调度算法研究[J]. 高技术通讯,2014,24(12):1261-1269.

[7] 尹向东,杨 杰,屈长青. 云计算环境下分布式文件系统的负载平衡研究[J]. 计算机科学,2014,41(3):141-144.

[8] 邓志飞,应良佳,王军威. 基于 IODA 算法 MongoDB 负载均衡的改进[J]. 现代电信科技,2013(7):9-13.

[9] Karger D R, Lehman E, Leighton F T, et al. Consistent hashing and random trees: distributed caching protocols for relieving hot spots on the world wide web[C]//ACM symposium on theory of computing. [s. l.]: ACM, 1997: 654-663.

[10] Sivasubramanian S. Amazon dynamoDB: a seamlessly scalable non-relational database service[C]//ACM SIGMOD international conference on management of data. [s. l.]: [s. n.], 2012: 729-730.

[11] 张聪萍,尹建伟. 分布式文件系统的动态负载均衡算法[J]. 小型微型计算机系统,2011,32(7):1424-1426.

[12] 赵 见. 高性能高可用键值存储系统的设计与实现[D]. 成都:电子科技大学,2010.

[13] Schintke F, Reinefeld A, Haridi S, et al. Enhanced paxos commit for transactions on DHTs[C]//IEEE/ACM international conference on cluster, cloud and grid computing. [s. l.]: IEEE, 2010: 448-454.

[14] 田浪军,陈卫卫,陈卫东,等. 云存储系统中动态负载均衡算法研究[J]. 计算机工程,2013,39(10):19-23.