Intelligent Crawler Engine On Cloud Computing Infrastructure

Pratibha Ganapati Gaonkar
M. Tech (Dept. of CS&E)
Bapuji Institute of Engineering and Technology
Davangere, India
prathibha.kwr@gmail.com

Dr. Nirmala C R
Professor and HOD (Dept. of CS&E)
Bapuji Institute of Engineering and Technology
Davangere, India
crn@bietdvg.edu

***Abstract*—This paper is aimed to implement an intelligent crawler engine on cloud computing infrastructure. This approach uses virtual machines on a cloud computing infrastructure to run intelligent crawler engine. The use of Virtual Machine (VM) on this architecture with help for easy setup/installation, maintenance or VM terminating that has been running with some particular crawler engine as needed. With this infrastructure, we have designed an intelligent crawler by making use of Naive Best First algorithm and R-Spam Rank algorithm, which is more efficient compared to the earlier crawlers as per the result and analysis. In order to accomplish this task Amazon public cloud is used with its services, S3 and EC2.**

Keywords—Cloud computing infrastructure; Intelligent Crawler engine;Elastic Compute Cloud(EC2); Simple Storage Service(S3).

I. INTRODUCTION

In the world of Web 2.0, the adage "content is king" remains a prevailing theme. With seemingly endless content available online, the "findability" of content becomes a key factor. Search engines are the primary tools people use to find information on the web. Searches are performed using keywords. When you enter a keyword or phrase, the crawler engine finds matching web pages and show you a search engine results page (SERP) with recommended web pages listed and sorted by relevance. Though it used to be difficult to obtain diverse content, there are now seemingly endless options competing for an audience's attention. As a result, search engines have gained popularity by helping users quickly find and filter the information they want. Google, Yahoo, Bing and Ask have emerged as the most popular search engines in the recent past. Most users have formed searching habits to gain the information they need, as there is no single website that caters to all their needs. Google logs an estimated 2 billion searches per day and an estimated 300 million users use the search facility provided by Google on a daily basis[1].

Web crawlers are the programs or software that uses the graphical structure of the Web to move from page to page[2].Such programs are also called wanderers, robots, spiders, and worms. Web crawlers are designed to retrieve Web pages and add them or their representations to local repository/databases. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages that will help in fast searches. Web search engines work by storing information about many web pages, which they retrieve from the WWW itself. These pages are retrieved by a Web crawler, which is an automated Web browser that follows every link it sees. A crawler for a large search engine has to address two issues. First, it has to have a good crawling strategy, i.e., a strategy for deciding which pages to download next. Second, it needs to have a highly optimized system architecture that can download a large number of pages per second while being robust against crashes, manageable, and considerate of resources and web servers[3][4].

Cloud computing is considered to be a new computing paradigm. However, recently it has gained a lot of attention due to its usefulness in terms of cost, adaptability, variety of services, and computational support to the devices with less computational power [3]. Though some people are worried about the security in the cloud. But in fact, the cloud is more secure than the proprietary infrastructure. As in the cloud computing, we outsource the computation, not the control[4].

Cloud computing can really be a winsome option for an enterprise. Especially for the new enterprises, which want to reduce the upfront cost for their computing infrastructure. Even established organizations can reduce not only the computing infrastructure cost, but also the administrative and operational cost for the infrastructure. Because after purchasing the computing infrastructure, the organization needs human resources, space, energy and many other resources to manage and administer them. Whereas, in the case of opting for cloud computing services, these costs are reduced[5]. Some of the cloud infrastructure/service providers are Amazon[9] , Salesforce , Google App Engine and Microsoft Azure . The major users of these cloud providers are the enterprises.

In the near future, cloud services will be widely used by the enterprises and individuals, using hybrid computing and communication devices. Thus it is required to provide cloud service to the individuals, at a very low cost. It can be possible by creating competition among cloud vendors and reducing infrastructure cost for them. For this purpose, we propose a cloud computing model (i.e. Virtual Cloud) to achieve the low cost objective. Virtual cloud model is mainly aimed to reduce cost for both the cloud user and cloud vendors. Cloud computing claimed that it provides better efficiency in the use of infrastructure [6].

In addition, cloud computing technology has been developed so rapidly and could change the implementation or

operation mode in Information Communication Technology. By using cloud computing, the use of information technology infrastructure is claimed to be more efficient and more effective. The infrastructure in this context consists of a number of machines that runs crawler engine to accommodate the high demand and storage servers or disk to store the results of all searches[7].

II. Related Work

In this section we will describe about the state of art of research on crawler engine. Some of specific crawler developed for special purposes. Topical crawlers or focused crawlers were developed to create contextual search engine or more focused result[10][20]. Topical crawlers follow the hyperlinked structure of the Web using the scent of information to direct themselves toward topically relevant pages. For deriving the appropriate scent, they mine the content of pages that are already fetched to prioritize the fetching of unvisited pages. Unlike search engines that use contextual information to complement content-based retrieval, topical crawlers depend primarily on contextual information. This is because topical crawlers need to predict the benefit of downloading unvisited pages based on the information derived from pages that have been downloaded. Topical crawlers have been used in a variety of applications such as competitive intelligence search engines and digital libraries. They allow a higher level application to gather from the Web, a focused collection rich in information about a topic or theme. The Naive Best-First crawler can be used for crawling the web[11].

Spam web pages intend to achieve higher-than-deserved ranking by various techniques. While human experts could easily identify spam web pages, the manual evaluating process of a large number of pages is still time consuming and cost consuming. The R-SpamRank algorithm can be used for detection of spam pages[12].

Cloud computing today has brought a new era in the use of infrastructure more efficient. Currently, cloud computing infrastructure will maximize the use of virtual machine (VM). It means that we can create instance of VM(s) on a single physical computer. The use of VM is expected to increase the efficiency of resource usage because VM can be set, invoked or terminated in accordance with the requirements. Typically, cloud computing architecture consists of the front end and the node as shown in figure 1 [6]. Front end on a specific implementation in Eucalyptus often referred or named with Cloud Controller [13][14]. While node or Node Controller is a physical device that can run single or multiple Virtual Machines (VMs) based on demand. The management of the number allocates for VM(s), and VM capacities perform by the front end. With this capability, then the cloud computing capacity can be managed and allocated efficiently by allocating Virtual Machine (VM) in accordance with necessary needs. For example, at the peak time session, we can determine the VM with a large capacity as the host server or we can also duplicate VM with same specifications.

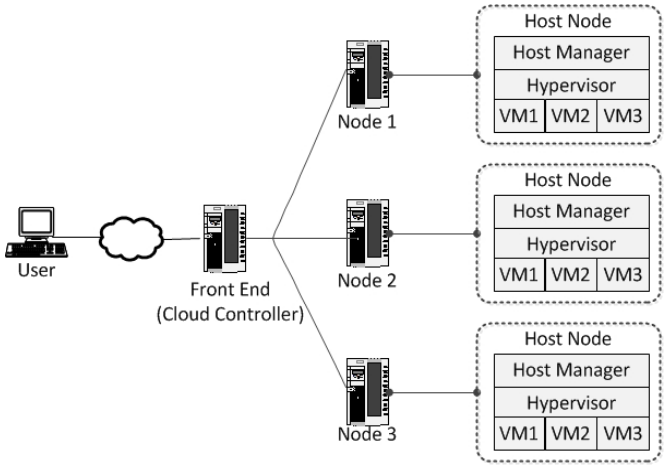

Figure 1 – Typical Cloud Computing Architecture

"Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction*"* [15]. There are three main categories of service models of cloud computing:Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) [16]. Scalability is one of the most prominent characteristics of all three categories [17][18]. The IaaS systems can offer elastic computing resources like Amazon Elastic Compute (EC2) and on demand storage resources like Amazon's Simple Storage Service (S3). Two of the most common deployment models of cloud computing are public cloud infrastructure and private cloud infrastructure. The former are the cloud computing infrastructure provided by 3rd party service provider (such as Google and Amazon) based on pay-as.-you-use model and the latter is the cloud computing infrastructure set up and managed by an organization for its own use.

Key advantages of cloud computing is the use of virtualization so that the users do not need to know where the computation performed by a machine. Also, with the usage of VM(s) will make it easier when running application and operating system installation. By using the VM, we can easily create a master application or service built in an operating system resulting in image. If we need some similar system to run the same program then cloud computing will easily turn on or duplicate the same VM on a particular physical node computer without hardware installation[19].

III. Proposed Architecture

This paper tries to build an intelligent crawler engine service in the cloud computing architecture. The main objectives of this paper are:

- To design an intelligent crawler engine on cloud computing Infrastructure by making use of naïve best first algorithm and R-SpamRank algorithms.
- To design the module to save keyword and for more effective searching.
- To create specific buckets in cloud storage to save, record and indexing the results from crawler engine.

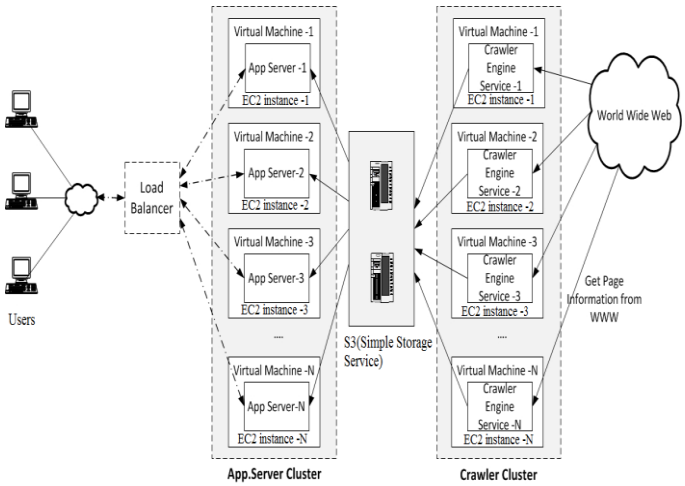

Figure 2: Architecture of Crawler Engine on Cloud Computing Infrastructure.

The proposed architecture of crawler engine in the figure 2 consists of virtual machines running on EC2 instances with crawler engine services and application servers and storing the query based results of crawling in S3.The crawler uses Naive Best-First crawling strategy. The Naive Best-First crawler represents a fetched Web page as a vector of words weighted by occurrence frequency. The crawler then computes the cosine similarity of the page to the query or description provided by the user, and scores the unvisited URLs on the page by this similarity value. The URLs are then added to a frontier that is maintained as a priority queue based on these scores. In the next iteration each crawler thread picks the best URL in the frontier to crawl, and returns with new unvisited URLs that are again inserted in the priority queue after being scored based on the cosine similarity of the parent page. The cosine similarity between the page p and a query q is computed analogous to the equation (1).

$$sim(p,q) = \frac{V_p \cdot V_q}{\|V_p\| \cdot \|V_q\|} \quad (1)$$

Where vq and vp are term frequency (TF) based vector representations of the query and the page respectively.

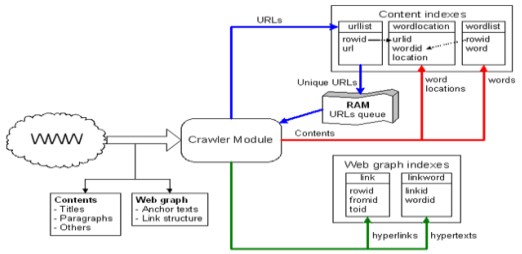

Figure 3: Crawler Structure.

In figure 3 the crawler module crawls the URLs from www and creates the content indexes and web graph indexes. The urllist table stores the crawler web addresses based on the time they were crawled. The rowids become identity of the corresponding web addresses. All the other tables that have to utilize web addresses use the rowids instead of directly using URL names. In wordlist table, rowids become identity numbers for corresponding words. Then word location table utilizes both urllist and wordlist to make list of locations of all words. The link table stores network structure of the crawled web pages. The link word table stores the anchor texts of the corresponding links which are used in anchor text analysis and scoring.

On the heels of the widespread adoption of web services such as social networks and URL shorteners, scams, phishing, and malware have become regular threats. Despite extensive research, spam filtering techniques generally fall short for protecting the web services. To better address this need, we present R-SpamRank, that crawls URLs as they are submitted to web services and determines whether the URLs direct to spam and helps the crawler whether to proceed with crawling or not for that URL.R-SpamRank algorithm aims to detect spam web pages. In this algorithm, the web page gains the spam rank value through forward links, which are the links of reverse direction used in traditional link-based algorithm.Therefore, this algorithm is called as R-SpamRank which means reverse spam rank.

This algorithm uses a blacklist containing spam web pages as seeds. The blacklist is manually collected in the experimental system. We assigned an initial R-SpamRank value for each page in the blacklist, and these values would expand in the iterative computation to the web pages linking to them. The formula of the algorithm is shown in equation (2) below.

$$RSR(A) = (1-\lambda)I(A) + \lambda \sum_{i=1}^{n} \frac{RSR(T_i)}{C(T_i)} \quad (2)$$

$$I(A) = \begin{cases} 1 & if\ A\ in\ blacklist \\ 0 & otherwise \end{cases} \quad (3)$$

where *RSR(A)* is the R-SpamRank value of page A; λ is a damping factor, which is usually set to 0.85;I(A) is the initial value for page A, it is set to 1 if page A in the original blacklist, otherwise 0 as shown in (3); n is the number of forward links of page A, and *Ti* is the ith forward link page of page A;C(*Ti*)is the number of in links of Page *Ti*; *RSR(Ti)* is the R-SpamRank value of page *Ti*.

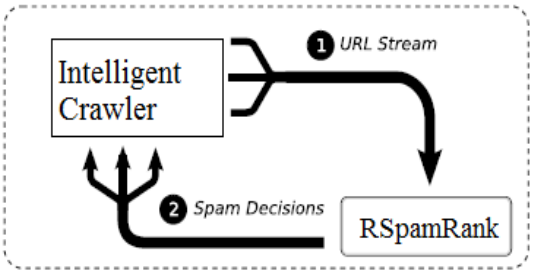

Figure 4: RSpamRank with Intelligent crawler.

IV. Experimental Setup

In order to run service engine crawlers on cloud infrastructure the instances of amazon EC2 are launched, to use the virtual machines to run the crawler engines.Amazon Elastic Compute Cloud (Amazon EC2) provides resizable computing capacity in the AmazonWeb Services (AWS) cloud. Using Amazon EC2 eliminates your need to invest in hardware up front, so we can develop and deploy applications faster. We can use Amazon EC2 to launch as many or as few

virtual servers as you need, configure security and networking, and manage storage. Amazon EC2 enables us to scale up or down to handle changes in requirements or spikes in popularity, reducing our need to forecast traffic.

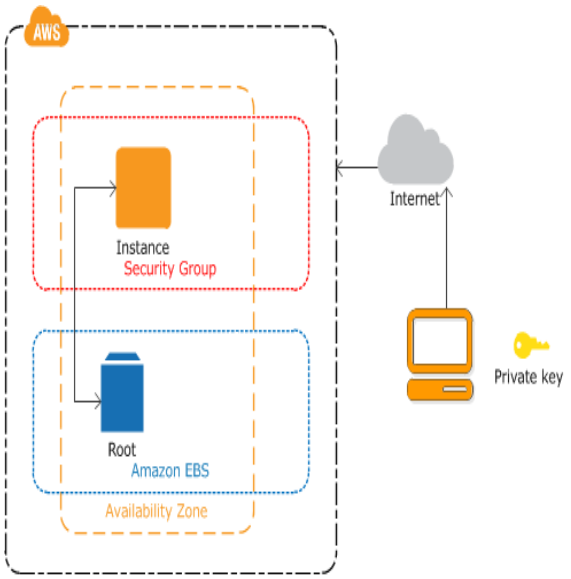

Figure 5: Amazon EC2

The instance is an Amazon EBS-backed instance (meaning that the root volume is an Amazon EBS volume) running Windows Server. We can either specify the Availability Zone in which our instance runs, or let us select an Availability Zone for us. When we launch our instance, we secure it by specifying a key pair and a security group. When we connect to our instance, we must specify the private key of the key pair that we specified when launching our instance. Our instance looks like a traditional host, and we can interact with it as we would with any computer running Windows Server.

After the instance is been launched and connected by using the key pair, the crawler engine services are run on the virtual machines.Then the indexed score values of crawling are stored in the Amazon Simple Storage Service(Amazon S3). We can use Amazon S3 to store and retrieve any amount of data at any time, from anywhere on the web. We accomplish these tasks using the simple and intuitive web interface of the AWS Management Console.

TABLE 1: Instance Details

Parameter	Value
Instance ID	i-8e405987
Instance state	running
Instance type	t1.micro
Private DNS	ip-172-31-30-159.us-west-2.compute.internal
Private IPs	172.31.30.159
VPC ID	vpc-a05647c2
Subnet ID	subnet-1be3d56f
Source/dest.check	true
AMI ID	Loading ami-bc92f08c...
Root Device Type	ebs
Public DNS	ec2-54-186-92-55.us-west-2.compute.amazonaws.com
Public IP	54.186.92.55
Availability zone	us-west-2b
Owner	756238232463
Virtualization	hvm
Reservation	r-2676672f
Security groups	launch-wizard-1
Key pair name	Crawler.pem
Network interfaces	eth0

V. Experimental Results

In figure 6 we have plotted graph for crawled pages for News domain by taking domain versus processed pages for two minutes. In the graph for the domain of news (Rediff ,The Hindu, Times of India(TOI), Deccan Herald) the processed pages are 78,58,107 and 219 respectively.

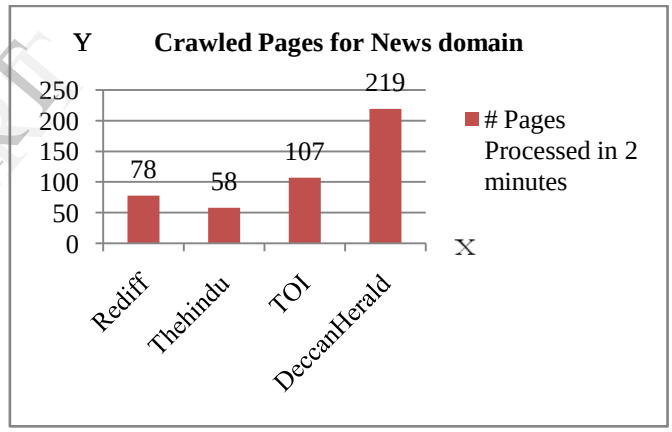

Figure 6 : Crawled Pages for News domain.

X:domain. Y:pages processed.

In figure 7 we have plotted graph for crawled pages for keywords in domain by taking domain keywords versus processed pages for two minutes. For the keyword India the news domain(Rediff, The Hindu, Times of India(TOI), Deccan Herald) the processed pages are (75,52,88,211).For the keyword Sports the news domain(Rediff , The Hindu, Times of India(TOI), Deccan Herald) the processed pages are (53,56,65,217).For the keyword Elections the news domain(Rediff, The Hindu, Times of India(TOI), Deccan Herald) the processed pages are (16,10,60,150).For the keyword 2014 the news domain(Rediff , The Hindu, Times of India(TOI), Deccan Herald) the processed pages are (70,46,106,50).

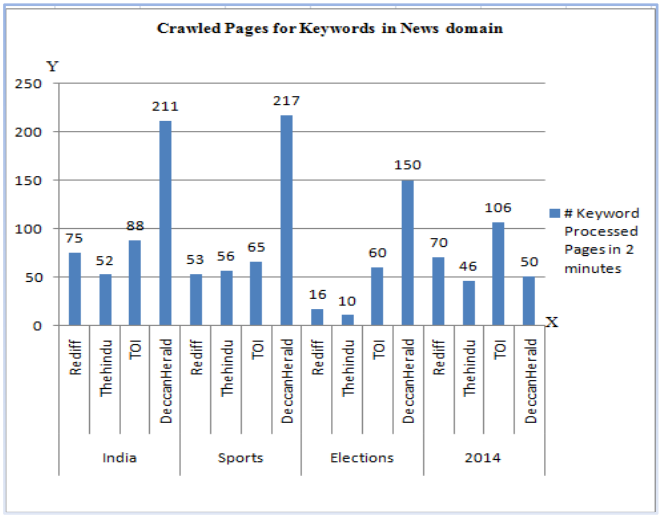

Figure 7: Crawled Pages for Keywords in News domain.

X:domain keywords. Y:pages processed.

By analyzing the graphs in figure 5 and figure 6 we conclude that the news domain Deccan Herald is more popular than the other three news domain Rediff, The Hindu and Times of India(TOI).

VI. CONCLUSION

In this paper we have implemented intelligent crawler engine on cloud computing infrastructure. This approach uses virtual machines on EC2 to run crawler engine services and stores the crawled indexed results on S3.We have conducted experiments for News domain like Rediff, The Hindu, Times of India (TOI) and Deccan Herald with keywords India, Sports, Elections and 2014. Based on experiments conducted we conclude that the news domain Deccan Herald is more popular than the other three news domain Rediff, The Hindu and Times of India (TOI). With the huge data processed on the cloud, the Intelligent crawler engine designed, efficiently crawls, processes and stores the results on S3.

REFERENCES

[1] Akassh A Mishra, ChinmayKamat, "Migration of Search Engine Process into the Cloud", *International Journal of Computer Applications*, Volume 19– No.1, April 2011.

[2] GautamPant,Padmini Srinivasan, and FilippoMenczer: "Crawling the Web".

[3] J. Cho, H. Garcia-Molina, and L. Page. Efficient crawling through url ordering. In *7th Int.World Wide Web Conference*,May 1998.

[4] M. Najork and J.Wiener. Breadth-first search crawling yields high-quality pages.In*10th Int. World Wide Web Conference*,2001.

[5] J. Hurwitz, R. Bloor, M. Kaufman, F. Halper, Cloud Computing for Dummies, Wiley Publishing, Inc. 2009.

[6] R. Chow, P. Golle, M. Jakobsson, E. Shi, J. Staddon, R. Masuoka, J. Molina, "Controlling Data in the Cloud: Outsourcing Computation Without Outsourcing Control", Proceedings of the 2009 ACM Workshop on Cloud Computing Security, Chicago, USA, November 2009.

[7] S. Zhang, S. Zhang, X Chen, X Huo, "Cloud Computing Research and Development Trend", Proceedings of the Second International Conference on Future Networks 2010, China, January 2010.

[8] Sheheryar Malik, FabriceHuet, "Virtual Cloud: Rent Out the Rented Resources", 6th IEEE International Conference for Internet Technology and Secured Transactions 2011.

[9] Amazon Elastic Compute Cloud, https://aws.amazon.com/ec2/.

[10] Gautam Pant and Padmini Srinivasan, "Link Contexts in Classifier-Guided Topical Crawlers", *IEEE Transactions On Knowledge and DataEngineering, Vol.18, No.1, January 2006*.

[11] C. Olston and E.H. Chi, "ScentTrails: Integrating Browsing and Searching on the Web", ACM Trans. Computer-Human Interaction,vol. 10, no. 3, pp. 177-197, Sept. 2003.

[12] Chenmin Liang1, Liyun Ru2 and Xiaoyan Zhu1,"R-SpamRank: A Spam Detection Algorithm Based on Link Analysis".

[13] Daniel Nurmi et al, "Eucalyptus : A Technical Report on an Elastic Utility Computing Architecture Linking Your Programs to Useful Systems".

[14] White Paper , "Intel® Cloud Builder Guide to Cloud Design and Deployment on Intel® Platforms".

[15] P. Mell and T. Grance, "The NIST Definition of Cloud Computing," National Institute of Standards and Technology, USA2009.

[16] A. Lenk, *et al.*, "What is Inside the Cloud? An Architectural Map of the Cloud Landscape," presented at the Workshop on Software Engineering Challenges of Cloud Computing, Collocated with ICSE 2009 Vancouver, Canada, 2009.

[17] R. Grossman, "The Case for Cloud Computing," *IEEE Computer*, vol. 11, pp. 23-27, 2009.

[18] Q. Zhang, *et al.*, "Cloud computing: state-of-the-art and research challenges," *Journal of Internet Services and Applications*, vol. 1, pp. 7-18, 2010.

[19] Sinung Suakanto, Suhono H. Supangkat, Suhardi, Roberd Saragih, " Building Crawler Engine On Cloud Computing Infrastructure ".

[20] Soumen Chakrabarti, Martin van den Berg, Byron Domc. "*Focused crawling: a new approach to topic-specific Web resource discovery*",1999.

[21] Amazon Simple Storage Service, https://aws.amazon.com/s3/.