



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Cluster Performance Calculator for High-Performance Distributed Web Crawler

¹Shishir Sarkar, ²Prateeksha Pandey

¹M.Tech. Scholar, ²Asst. Professor
CSE Dept., CSIT, Durg (C.G), India

Abstract: Broad web search engines as well as many more specialized search tools rely on web crawlers to acquire large collections of pages for indexing and analysis. Such a WebCrawler may interact with millions of hosts over a period of weeks or months, and thus issues of robustness, flexibility, and manageability are of major importance. In addition, I/O performance, network resources, and OS limits must be taken into account in order to achieve high performance at a reasonable cost. In this paper, we describe the algorithm through which we can calculate the threshold value of each CPU Node in cluster and then calculate the cumulative threshold value whenever cluster will restart or configuration will change. Using this threshold value we can calculate the performance of cluster.

Keywords: web search engines, crawling, monitor, cumulative threshold.

I. Introduction

Web crawlers, programs that automatically find and download web pages, have become essential to the fabric of modern society. This strong claim is the result of a chain of reasons: the importance of the web for publishing and finding information; the necessity of using search engines like Google to find information on the web; and the reliance of search engines on web crawlers for the majority of their raw data, as shown in Figure 1 (Brin & Page, 1998; search engines is emphasized by Van Couvering (2004), who argues that they alone, and not the rest of the web, form a genuinely new mass media.

Web users do not normally notice crawlers and other programs that automatically download information over the Internet. Yet, in addition to the owners of commercial search engines, they are increasingly used by a widening section of society including casual web users, the creators of email spam lists and others looking for information of commercial value. In addition, many new types of information science research rely upon web crawlers or automatically downloading pages (e.g., Björneborn, 2004; Faba-Perez, Guerrero-Bote, & De Moya-Anegón, 2003;

Mechanism for stopping crawlers from visiting some or all of the pages in their site. Suggestions have also been published governing crawling speed and ethics (e.g., Koster, 1993, 1996), but these have not been formally or widely adopted, with the partial exception of the 1993 suggestions. Nevertheless, since network speeds and computing power have increased exponentially, Koster's 1993 guidelines need reappraisal in the current context. Moreover, one of the biggest relevant changes between the early years of the web and 2005 is in the availability of web crawlers. The first crawlers must have been written and used exclusively by computer scientists who would be aware of network characteristics, and could easily understand crawling impact. Today, in contrast, free crawlers are available online. In fact there are site downloaders or offline browsers that are specifically designed for general users to crawl individual sites, (there were 31 free or shareware downloaders listed in tu cows.com on March 4, 2005, most of which were also crawlers). A key new problem, then, is the lack of network knowledge by crawler owners. This is compounded by the complexity of the Internet, having broken out of its academic roots, and the difficulty to obtain relevant cost information (see below). In this paper, we review new and established moral issues in order to provide a new set of guidelines for web crawler owners. This is preceded by a wider discussion of ethics, including both computer and research ethics, in order to provide theoretical guidance and examples of

II. Introduction to Ethics

The word 'ethical' means, 'relating to, or in accord with, approved moral behaviors' (Chambers, 1991). The word 'approved' places this definition firmly in a social context. Behavior can be said to be ethical relative to a particular social group if that group would approve of it. In practice, although humans tend to operate within their own internal moral code, various types of social sanction can be applied to those employing problematic behavior. Formal ethical procedures can be set up to ensure that particular types of recurrent activity are systematically governed and assessed, for example in research using human subjects. Seeking formal ethical approval may then become a legal or professional requirement. In other situations ethical reflection may take place without a formal process, perhaps because the possible outcomes of the activity might not be directly harmful, although problematic in other ways. In such cases it is common to have an agreed written or unwritten ethical framework, sometimes

called a code of practice or a set of guidelines for professional conduct. When ethical frameworks or formal procedures fail to protect society from a certain type of behavior, it has the option to enshrine them in law and apply sanctions to offenders.

The founding of ethical *philosophy* in Western civilization is normally attributed to ancient Greece and Socrates (Arrington, 1998). Many philosophical theories, such as utilitarianism and situation ethics, are relativistic: what is ethical for one person may be unethical for another (Vardy & Grosch, 1999). Others, such as deontological ethics, are based upon absolute right and wrong. Utilitarianism is a system of making ethical decisions, the essence of which is that “an act is right if and only if it brings about at least as much net happiness as any other action the agent could have performed; otherwise it is wrong.” (Shaw, 1999, p.10). Different ethical systems can reach opposite conclusions about what is acceptable: from a utilitarian point of view car driving may be considered ethical despite the deaths that car crashes cause but from a deontological point of view it could be considered unethical. The study of ethics and ethical issues is a branch of philosophy that provides guidance rather than easy answers.

III. Computer Ethics

The philosophical field of computer ethics deals primarily with professional issues. One important approach in this field is to use social contract theory to argue that the behavior of computer professionals is self-regulated by their representative organizations, which effectively form a contract with society to use this control for the social good (Johnson, 2004), although the actual debate over moral values seems to take place almost exclusively between the professionals themselves (Davis, 1991). A visible manifestation of self-regulation is the production of a code of conduct, such as that of the Association for Computing Machines (ACM, 1992). The difficulty in giving a highly prescriptive guide for ethical computing can be seen in the following very general important advice, “One way to avoid unintentional harm is to carefully consider potential impacts on all those affected by decisions made during design and implementation” (ACM, 1992).

There seems to be broad agreement that computing technology has spawned genuinely new moral problems that lack clear solutions using existing frameworks, and require considerable intellectual effort to unravel (Johnson, 2004). Problematic areas include: content control including libel and pornography (Buell, 2000); copyright (Borrull & Oppenheim, 2004); deep linking (Fausett, 2002); privacy and data protection (Carey, 2004; Reiman, 1995; Schneier, 2004); piracy (Calluzzo & Cante, 2004); new social relationships (Rooksby, 2002); and search engine ranking

Technology is never inherently good or bad; its impact depends upon the uses to which it is put as it is assimilated into society (du Gay, Hall, Janes, Mackay, & Negus, 1997). Some technologies, such as medical innovations, may find themselves surrounded at birth by a developed ethical and/or legal framework. Other technologies, like web crawlers, emerge into an unregulated world in which users feel free to experiment and explore their potential, with ethical and/or legal frameworks later evolving to catch up with persistent socially undesirable uses. Two examples below give developed illustrations of the latter case.

The fax machine, which took off in the eighties as a method for document exchange between businesses (Negroponte, 1995), was later used for mass marketing. This practice cost the recipient paper and ink, and was beyond their control. Advertising faxes are now widely viewed as unethical but their use has probably died down not only because of legislation which restricted its use (HMSO, 1999), but because they are counterproductive; as an unethical practice they give the sender a bad reputation.

Email is also used for sending unwanted advertising, known as spam (Wronkiewicz, 1997). Spam may fill a limited inbox, consume the recipient's time, or be offensive (Casey, 2000). Spam is widely considered unethical but has persisted in the hands of criminals and maverick salespeople. Rogue salespeople do not have a reputation to lose nor a need to build a new one and so their main disincentives would presumably be personal morals, campaign failure or legal action. It is the relative ease and ultra-low cost of bulk emailing that allows spam to persist, in contrast to advertising faxes. The persistence of email spam (Stitt, 2004) has forced the hands of legislators in order to protect email as a viable means of communication (www.spamlaws.com). The details of the first successful criminal prosecution for Internet spam show the potential rewards on offer, with the defendant amassing a 24 million dollar fortune (BBCNews, 4/11/2004). The need to resort to legislation may be seen as a failure of both ethical frameworks and technological solutions, although the lack of national boundaries on the Internet is a problem: actions that do not contravene laws in one country may break those of another.

IV. Research ethics

Research ethics are relevant to a discussion of the use of crawlers, to give ideas about what issues may need to be considered, and how guidelines may be implemented. The main considerations for social science ethics tend to be honesty in reporting results and the privacy and well-being of subjects (e.g., Penslar, 1995). In general, it seems to be agreed that researchers should take responsibility for the social consequences of their actions, including the uses to which their research may be put (Holdsworth, 1995). Other methodological-ethical considerations also arise in the way in which the research should be conducted and interpreted, such as the influence of power relationships (Williamson, & Smyth, 2004; Penslar, 1995, ch. 14).

Although many of the ethical issues relating to information technology are of interest to information scientists, it has been argued that the focus has been predominately on professional codes of practice, the teaching of ethics, and professional dilemmas, as opposed to research ethics (Carlin, 2003). The sociology-inspired emerging field of Internet research (Rall, 2004a) has developed guidelines, however, although they are not all relevant since its research methods are typically qualitative (Rall, 2004b). The fact that there are so many different environments (e.g., web pages, chatrooms, email) and that new ones are constantly emerging means that explicit rules are not possible, instead broad guidelines that help researchers to appreciate the potential problems are a practical alternative. The Association of Internet Researchers has put forward a broad set of questions to help researchers come to conclusions about the most ethical way to carry out Internet research (Ess & Committee, 2002), following an earlier similar report from the American Association for the Advancement of Science (Frankel & Siang, 1999). The content of the former mainly relates to privacy and disclosure issues and is based upon considerations of the specific research project and any ethical or legal restrictions in place that may already cover the research. Neither allude to automatic data collection.

Although important aspects of research are discipline-based, often including the expertise to devise ethical frameworks, the ultimate responsibility for ethical research often lies with universities or other employers of researchers. This manifests itself in the form of university ethics committees (e.g., Jankowski & van Selm, 2001),

V. Problem Identification

- i. Mostly web crawler doesn't have any distribute cluster performance system.
- ii. Doesn't have any implemented algorithm.
- iii. Doesn't have any threshold value which help to divert the traffic.
- iv. It's very difficult to implement load balancer.
- v. Doesn't have any dynamic generation technique

VI. Web crawling issues

Having contextualized ethics from general, computing and research perspectives, web crawling can now be discussed. A web crawler is a computer program that is able to download a web page, extract the hyperlinks from that page and add them to its list of URLs to be crawled (Chakrabarti, 2003). This process is recursive, so a web crawler may start with a web site home page URL and then download all of the site's pages by repeatedly fetching pages and following links. Crawling has been put into practice in many different ways and in different forms. For example, commercial search engines run many crawling software processes simultaneously, with a central coordination function to ensure effective web coverage (Chakrabarti, 2003; Brin & Page, 1998). In contrast to the large-scale commercial crawlers, a personal crawler may be a single crawling process or a small number, perhaps tasked to crawl a single web site rather than the 'whole web'. It is not appropriate to discuss the software engineering and architecture of web crawlers here (see Chakrabarti, 2003; Arasu, Cho, Garcia-Molina, Paepcke, & Raghavan, 2001), but some basic points are important. As computer programs, many crawler operations are under the control of programmers. For example, a programmer may decide to insert code to ensure that the number of URLs visited per second does not exceed a given threshold. Other aspects of a crawler are outside of the programmer's control. For example, the crawler will be constrained by network bandwidth, affecting the maximum speed at which pages can be downloaded. Since crawlers are no longer the preserve of computer science researchers but are now used by a wider segment of the population, which affects the kinds of issues that are relevant. Table 1 records some user types and the key issues that particularly apply to them, although all of the issues apply to some extent to all users. Note that social contract theory could be applied to the academic and commercial computing users, but perhaps not to non-computing commercial users and not to individuals. These latter two user types would be therefore more difficult to control through informal means. There are four types of issue that web crawlers may raise for society or individuals: denial of service, cost, privacy and copyright. These are defined and discussed separately below.

VII. Algorithms

1. Start fetching the code details of each CPU when the cluster is restart or configuration is changed.
2. Assign value in each attribute {CPU Clock Speed, Cores per Processor and Operations per Cycle}
3. Putting this value in calculation using below formula

$$\{CPU \text{ Clock Speed} * \text{Cores per Processor} * \text{Operations per Cycle}\}$$
4. Threshold value of CPU Node =
$$\frac{\{CPU \text{ Clock Speed} * \text{Cores per Processor} * \text{Operations per Cycle}\}}{\{CPU \text{ Clock Speed} * \text{Cores per Processor}\} / \text{Cores per Processor}}$$
5. Calculate threshold value of each CPU Node
6. Add all the Threshold value of each node

7. *Cumulative Threshold value(CTV) = {Threshold value of CPU Node 1+Threshold value of CPU Node 2+.....+Threshold value of CPU Node N}.*
8. *Calculate the Response time of each CPU node using below formula*

$$\text{Calculate Start time} - \text{Calculate End Time}$$
9. *Response time =*
$$\frac{\text{Operations per Cycle}}{\text{Calculate Start time} - \text{Calculate End Time}}$$
10. *IF (Response time > Cumulative Threshold value)*
11. *Check the threshold value of another cluster if response time if less than Cumulative Threshold Value divert the traffic to that cluster.*
Else
12. *Continue to same cluster*

IX. Benefits of Proposed Architecture

- i. Able to calculate the *cumulative threshold* value
- ii. Efficiently able to divert the traffic.
- iii. Down time of cluster is very less.
- iv. Efficiently utilize the entire cluster.
- v. Runtime ability to calculate the *cumulative threshold* if server configuration will change or adding new server in cluster
- vi. Reduced the hardware cost.
- vi. Very less required human interference

X. Conclusion

This kind of mechanism should be used in large scale of crawler kind of system, so that we can avoid the system shutdown, fail kind of problem generally we can have a scanner system to scan the complete system but doesn't have a System or mechanism which will take a action according.

References

- [1] Shishir Sarkar and Prateeksha Pandey "Design t he framework for distributed and high performance web crawler"
- [2] S. Chen, B. Mulgrew, and P. M. Grant, "A clustering technique for digital communications channel equalization using radial basis function networks," IEEE Trans. on Neural Networks, vol. 4, pp. 570-578, July 1993.
- [3] J. U. Duncombe, "Infrared navigation—Part I: An assessment of feasibility," IEEE Trans. Electron Devices, vol. ED-11, pp. 34-39, Jan. 1959.
- [4] C. Y. Lin, M. Wu, J. A. Bloom, I. J. Cox, and M. Miller, "Rotation, scale, and translation resilient public watermarking for images," IEEE Trans. Image Process., vol. 10, no. 5, pp. 767-782, May 2001.
- [5] K. Bharat, A. Broder, M. Henzinger, P. Kumar, and S. Venkatasubramanian. The connectivity server: Fast Access to linkage information on the web. In *7th Int.world Wide Web Conference*, May 1998.
- [6] S. Brin and L. Page. The anatomy of a large-scale Hypertextual web search engine. In *Proc. of the Seventh world wide*.
- [7]. M. Burner. Crawling towards eternity: Building an archive of the world wide web. *Webtechniques*, 1997.