RESEARCH ARTICLE OPEN ACCESS

MANAGING RICH META DATA IN HIGH PERFORMANCE COMPUTING SYSTEM USING A GRAPH MODEL

Mr. S.Sambasivam.,M.C.A.,M.Phil*, Mr. E.P.Pranesh, M C A **

*(Professor, Department of Computer Applications,
Nandha Engineering College (Autonomous),
Erode, Tamil Nadu, India
Email: sammy2173@gmail.com)
** (Final MCA, Department of Computer Applications,
Nandha Engineering College (Autonomous),
Erode, Tamil Nadu, India
Email:praneshep1997@gmail.com)

****

Abstract:

The gold standard internet statistics mining evaluation of net page structure acts as a key element in instructional area which affords the systematic way of novel implementation in the direction of real-time information with exclusive stage of implications. With the rapid improvement and boom in worldwide data on world extensive internet and with expanded and speedy boom in internet customers throughout the globe, an acute need has arisen to enhance and alter or layout search algorithms that enables in successfully and efficaciously searching the specific required facts from the big repository to be had. In current work that use specific web crawlers for obtaining seek consequences efficiently. a few sirs use targeted net crawler that collects distinctive internet pages that commonly fulfill some particular property, by way of successfully prioritizing the crawler frontier and managing the exploration manner for link. A focused net crawler analyzes its move slowly boundary to find the hyperlinks that are in all likelihood to be maximum applicable for the move slowly, and avoids beside the point areas of the web. This ends in good sized savings in hardware and community sources, and helps keep the crawl extra up-to-date. The procedure of proposed I-Spider focused internet malicious web page crawler is to nurture a group set of web documents which can be centered on a few topical subspaces. It identifies the next most important and relevant link to follow by counting on probabilistic models for correctly predicting the relevancy of the file. Researchers across have proposed numerous algorithms for improving performance of focused internet malicious web page crawler. We try to investigate various types of crawlers with their professionals and cons. Principal cognizance vicinity is focused internet malicious web page crawler. destiny instructions for improving performance of centered net crawler had been mentioned. This can offer a base reference for anyone who wishes in getting to know or the use of concept of targeted WebCrawler of their studies work that he/she wishes to perform. The overall performance of a focused WebCrawler depends at the richness of links inside the specific subject matter being searched by using the user, and it usually relies on a well known web search engine for providing beginning factors for looking.

Keywords — Key Element, Internet, Web crawlers, Web search Engine.

****

I. INTRODUCTION

High-performance computing (HPC) systems generate huge amount of metadata about various entities in the system every second. These metadata include both the traditional POSIX metadata which describes the predefined attributes about individual entities such as files and users and the so-called rich metadata which describes detailed runtime information about boarder categories of entities and their complex relationships such as running jobs or A well-known case of rich metadata is data provenance, which describes the relationships among entities such as data files, running jobs, execution context, and their dependencies that contribute to the existence of a data item With rich metadata, we can effectively enable a variety of data management functionality in HPC environment, such as data auditing, result validation, and reproducible For example, the file access history of users can be used to audit users' activities in shared super-computer facilities; the file access history of processes can be used to depict an execution and its configurations; and captured detailed execution history and context can be used to regenerate an environment for reproducing scientific results. While rich metadata can support many attractive data management functionality, the support for rich metadata is still very limited in existing HPC systems. Many of recent studies about HPC data management enhancement, such as spyglass [5] and Magellan [6] are still largely limited to POSIX metadata. The major factor that limits rich metadata management is the lack of necessary facilities to model, store, process, and query the complex rich metadata efficiently in HPC We summarize three key challenges of managing rich metadata in HPC platforms. First, rich metadata are heterogeneous. They can contain predefined attributes and relationships such as POSIX metadata of files, as well as user-defined entities, attributes, and relationships, such as the running execution that creates, reads, and writes the data. This heterogeneity makes modelling rich metadata complicated. Moreover, the rich metadata are typically from different components of HPC systems such as file systems, runtime, and job schedulers. This diversity makes rich metadata management a global service for the entire platform across different system components. These distinct data sources and data formats should be integrated uniformly in order to avoid duplication of functionality across management tools. Second, the volume of rich metadata is large. In addition to the POSIX metadata, rich metadata contain dynamic runtime information about data files, jobs, users, and environmental variables. A leadership super-computer might include millions of processes and computation cores operating billions of files. They all generate dynamic runtime metadata continuously. This scale can lead to a huge volume of rich metadata generated in a short time, and the metadata must be ingested and managed gracefully. Given the already high pressure on parallel file systems for just maintaining the POSIX metadata, storing the rich metadata is easily a big challenge on the storage Third, users need highly efficient methods to query the rich metadata in order to accomplish various data management tasks. These queries lead to access patterns such as locating a single entity and relationship (e.g., load attributes of a file or user); iterating over relationships (e.g., find out all users that have accessed a file); and providing conditional traversal across multiple entities (e.g., locate the initial input data sets for an important result following the read/write relationships between executions and data files). These data accesses need to be efficiently supported, given the large volume of rich metadata and the high concurrency of HPC systems.

II. LITERATURE REVIEW

C. Gao, L. Wang, C.-Y. Lin et al.,[GAO1] has proposed online boards include a huge amount of valuable individual generated content fabric. on this paintings we cope with the problem of extracting query-solution pairs from forums. Query-answer pairs extracted from forums can be used to assist query Answering services (e.g. Yahoo! Solutions) among different packages. We recommend sequential patterns based classification method to hit upon questions in a dialogue board thread, and a graph based totally propagation approach to detect solutions for questions within the identical We I-

Spider on mining knowledge in the shape of question-answer (QA) pairs from forums. Many forums comprise query-answer knowledge. We investigated 40 boards and observed that ninety % of them contain query-answer know-how. Mining query-solution pairs from forums has the subsequent packages. First, question-answer pairs are essential to many QA offerings, including instant solutions provided by way of sirs, QA are trying to find systems, and community-based totally question Answering (CQA) As an example, CQA services, inclusive of Yahoo! Solutions, Baidu and Naver 1, have those days become very famous. Discussion board has extended records than CQA and consists of lots large purchaser-generated content.

N. Look, M. Hurst, ok. Nigam et al., [GLA 2] has proposed weblogs and message forums offer on-line boards for dialogue that report the voice of the general public. Woven into this mass of debate is a large style of opinion and statement approximately purchaser products. This gives an opportunity for groups to recognize and respond to the client with the aid of using analysing these unsolicited remarks. Given the volume, format and content material of the information, the best method to understand this data is to use big-scale internet and text information mining technology.

This painting describes a stop-to-end commercial tool this is used to support a number of advertising and marketing and advertising and marketing intelligence and business enterprise intelligence applications. In short, we describe a mature device which leverages on-line records to assist make informed and timely decisions with apprehend to manufacturers, merchandise and techniques in the company space. This device strategies online content material for entities interested in tracking the opinion of the internet public (frequently as a proxy for maximum the applications that this data is positioned to variety from:

Early alert - informing subscribers whilst a unprecedented however important, or maybe fatal, scenario takes

Buzz monitoring - following trends in topics of debate and understanding what new topics are forming.

Sentiment mining - extracting aggregate measures of high great vs. horrible options.

Y. Guo, ok. Li, ok. Zhang et al.,[GUO 3] has proposed a logo new approach of Board discussion board Crawling to crawl net dialogue board. This technique exploits the prepared trends of the internet forum websites and simulates human behaviour of travelling net boards.

H.S. Kop pula, ok's. Leela et al., [KOP 5] has proposed of replica files inside the global net adversely influences crawling, indexing and relevance, which is probably the middle constructing blocks of internet seek. on this paintings, we gift a hard and fast of techniques to mine recommendations from URLs and utilize these suggestions for de-duplication the use of simply URL strings without fetching the content material explicitly. Our method is composed of mining the move slowly logs and utilizing clusters of comparable pages to extract transformation policies, which can be used to normalize URLs belonging to each cluster. Keeping each mined rule for de-duplication isn't efficient because of the huge amount of such guidelines. We gift a system studying method to generalize the set of regulations, which reduces the resource footprint to be usable at web-scale.

Li .K, Cheng X .Q, Y. Guo, and okay. Zhang, et al [6] [LICHE6] the guideline extraction strategies are sturdy against net-website unique URL conventions. We examine the precision and scalability of our technique with brand new efforts inside the use of URLs for de-duplication. Experimental effects exhibit that our technique achieves 2 instances greater bargain in duplicates with most effective half the tips as compared to the most latest preceding technique.

G.S. Manku, A. Jain et al., [MAN7] has proposed near-replica internet documents are ample. Two such documents vary from every different in a completely small element that displays commercials, for example. Such versions are inappropriate for internet search. So the superb of a web crawler will boom if it may assess whether or not a newly crawled net page is a close to-duplicate of a formerly crawled internet page or now not. Inside the route of developing a near-replica detection system for a multi-billion internet web page

repository, we make research contributions. First, we show off that Charikar's fingerprinting technique is appropriate for this purpose. 2nd, we present an algorithmic technique for figuring out present day f- bit fingerprints that change from a given fingerprint in at most ok bit-positions, for small ok. Documents which may be particular duplicates of each different (due to mirroring and plagiarism) are easy to understand by way of preferred check summing techniques. A more difficult trouble is the identification of close to-replica documents. Two such documents are identical in phrases of content material but vary in a small a part of the record together with advertisements, counters and timestamps.

Elimination of near to-duplicates1 saves network bandwidth, reduces storage costs and improves the exquisite of seek indexes. It also reduces the burden on the faraway host this is serving such net pages. A system for detection of close to-replica pages faces some of stressful situations. First and important is the trouble of scale: engines like Google index billions of net-pages; these quantities to a multi-terabyte database. Second, the crawl engine needs to be able to crawl billions of net-pages in keeping with day. So the choices to mark a newly-crawled web page as a near-duplicate of a modern-day internet page want to be made quickly.

U. Schonfeld and N. Shivakumar et al., [SCH8] has proposed complete coverage of the public net is essential to net engines like Google. Search engines like google and yahoo like google and yahoo use crawlers to retrieve pages and then find out new ones through using extracting the pages' outgoing links. However, the set of pages reachable from the publicly connected web is predicted to be notably smaller than the invisible internet, the set of files that haven't any incoming links and may best be retrieved thru net applications and net forms. The Sitemaps protocol is a fast-growing net protocol supported at the same time via maximum important like Google and yahoo to help content fabric creators and engines like Google unfastened up this hidden information by the use of making it to be had to engines like Google. We feature out an extensive have a look at of how "conventional" discovery crawling compares with Sitemaps, in key measures consisting of coverage and freshness over key consultant websites in addition to over billions of URLs seen at Google. We check that Sitemaps and discovery crawling supplement each different thoroughly, and offer unique tradeoffs.

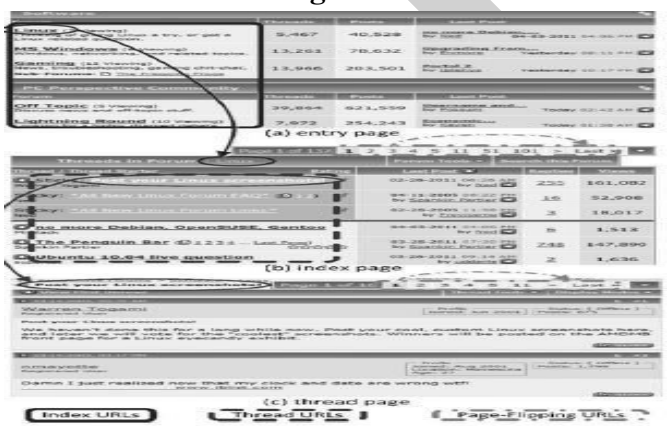
III. EXISTING SYSTEM

For studying normal expression patterns of URLs that lead a crawler from an entry page to goal pages. Target pages had been located thru evaluating DOM trees of pages with a preselected sample goal web page. it's far very effective but it handiest works for the specific site from which the pattern web page is drawn. The identical technique needs to be repeated every time for a new web page. Consequently, it isn't suitable for huge-scale crawling. In evaluation, base line method learns URL styles across more than one web sites and routinely finds a discussion board's access web page given a web page from the discussion board. Experimental outcomes display that base line approach is effective at huge-scale forum crawling through leveraging crawling expertise learned from some annotated discussion board websites. Guo et al. and Li et al. are just like our paintings. However, Guo et did now not point out a way to find out and traverse URLs. Li et al. evolved a few heuristic policies to find out URLs. Instance screenshots of index pages (the pinnacle two pages) and thread pages (the lowest pages) from exceptional boards (the left two pages are from one discussion board and the proper two pages are from. But, their rules are too particular and might only be applied to unique boards powered through the precise software package deal wherein the heuristics had been conceived. Sadly, in keeping with forum Matrix, there may be hundreds of different discussion board software applications used at the net.

Figure:1

Please talk over with for extra information approximately discussion board software program packages. In addition, many forums use their personal customized software. Base line technique pursuits to automatically analyse a forum crawler with minimal human intervention by using sampling pages, clustering them, deciding on informative clusters thru an in formativeness degree, and finding a traversal course by means of a spanning tree set of rules. But, the traversal course selection procedure requires human inspection. Observe up paintings with the aid of Wang et al. proposed an set of rules to address the traversal route selection problem. They added the concept of skeleton link and web page-flipping link. Skeleton hyperlinks are "the maximum crucial hyperlinks supporting the structure of a discussion board web page." importance is determined by means of in formativeness and insurance metrics. Web page-flipping hyperlinks are decided the use of connectivity metric. Via figuring out and most effective consistent with our assessment, its sampling method and in formativeness estimation is not sturdy and its tree-like traversal route does no longer allow a couple of direction from a beginning web page node to a identical finishing page node. for instance, as proven in Fig 1, there are six paths from access to threads. However base line method could best take the primary route (entry ! board ! thread). Base line technique learns URL place records to discover new URLs in crawling, but a URL area would possibly emerge as invalid while the page shape adjustments. Rather than final analysis method, we explicitly outline entry-index-thread paths and leverage web page layouts to perceive index pages and thread pages.

Figure 2



Sitemap is an XML report that lists URLs at the side of extra metadata consisting of replace time, change frequency and many others. Normally speaking, the reason of robots.txt and Sitemap is to enable the website to be crawled intelligently. So they may be useful to discussion board crawling. But, it is difficult to keep such documents for forums as their content continually adjustments. In our experiment in phase, extra than 47 percentage of the pages crawled with the aid of a established crawler that may well apprehend these enterprise standards are uninformative.

IV. PROPOSED SYSTEM

To facilitate presentation within the following sections, we classified discussion board pages into web page kinds.

Access page: The homepage of a forum, which incorporates a list of boards and is likewise the bottom commonplace ancestor of all threads. See Fig. 2 for an example.

Index web page: A web page of a board in a forum, which generally contains a desk-like shape; every row in it includes statistics of a board or a thread. See Figs 2, list-of-board page, list of board and thread page, and board web page are all index pages.

Thread page: A page of a thread in a discussion board that carries a listing of posts with consumer generated content material belonging to the identical dialogue. Fig 2

Other page: A page that is not an access page, index page, or thread page. There are four types of URL.

Index URL: A URL this is on an access web page or index page and factor to an index page. its anchor textual content indicates the identify of its destination board.

Thread URL: A URL this is on an index web page and factor to an thread page. Its anchor text is the identify of its vacation spot thread. Fig 2 displays an instance.

Web page-flipping URL: A URL that leads users to some other page of the same board or the equal thread. Efficiently dealing with web page flipping URLs enables a crawler to download all threads in a large board or all posts in a protracted thread.

Different URL: A URL that isn't an index URL thread, or web page-flipping URL.

EIT course: An entry-index-thread course is a navigation course from an entry page via a sequence of index pages (thru index URLs and index web page-flipping URLs) to string pages (through thread URLs and thread page-flipping URLs).

ITF normal expression: An index-thread-page-flipping regular expression is a regular expression that can be used to apprehend index, thread, or page-flipping URLs. ITF regular expression is what I-Spider objectives to analyze and applies without delay in online crawling. The found out ITF ordinary expression are site particular, and there are four ITF regular expression in a domain: one for recognizing index URLs, one for thread URLs, one for index web page-flipping URLs, and one for thread web page-flipping URLs.

Advantages

- Couldn't become aware of the horrific URL within the website.
- Does now not identify type of protocol used for any net page.
- Retrieve the net pages, we observe pattern popularity over text and pattern symbolizes test textual content only.
- Take a look at how much textual content is available on web page.

V. CONCLUSION

Web shape Mining is a powerful technique used to extract the records from beyond conduct of net structure Mining to on this work we applied I-Spider Crawling that focusing at the category of internet structure mining for our pattern test we diagnosed the university internet portal is extra emphasized on educational hyperlinks instead of with the individual college links.

Considering this is a large area, and there a whole lot of work to do, we are hoping Our proposed method make it as an smooth system via the unconventional view of periodic net facts stage garage and retrieval combos, further focusing in their mutual proportion together with variant outcomes we done an data analysis method. In close to destiny these studies will extend its variety in the direction of web usage evaluation.

REFERENCES

[1] C. GAO, L. WANG, C.-Y. LIN ET AL , AND T. WANG-CHIEW, "WHY AND WHERE: A CHARACTERIZATION OF DATA PROVENANCE," IN DATABASE THEORY ICDT 2001. SPRINGER, 2001, PP. 316– 330.

[2] K.-K. Muniswamy-Reddy, D. A. Holland, U. Braun, and M. I. Seltzer, "Provenance-Aware Storage Systems," in USENIX Annual Technical Conference, General Track, 2006, pp. 43–56.

[3] Y. L. Simmhan, B. Plale, and D. Gannon, "A Survey of Data Provenance in e-Science," ACM Sigmod Record, vol. 34, no. 3, pp. 31–36, 2005.

[4] C. T. Silva, J. Freire, and S. P. Callahan, "Provenance for Visualizations: Reproducibility and Beyond," Computing in Science & Engineering, vol. 9, no. 5, pp. 82–89, 2007.

[5] A. W. Leung, M. Shao, T. Bisson, S. Pasupathy, and E. L. Miller, "Spyglass: Fast, Scalable Metadata Search for Large-Scale Storage Systems," in FAST, vol. 9, 2009, pp. 153–166.

[6] A. Leung, I. Adams, and E. L. Miller, "Magellan: A Searchable Metadata Architecture for Large-Scale File Systems," University of California, Santa Cruz, Tech. Rep. UCSC-SSRC-09-07, 2009.

[7] D. Dai, P. Carns, B. R. Ross, J. Jenkins, K. Blauer, and Y. Chen, "GraphTrek: Asynchronous Graph Traversal for Property Graph Based Metadata Management," in IEEE International Conference on Cluster Computing, IEEE CLUSTER. IEEE, 2015.

[8] D. Dai, R. B. Ross, P. Carns, D. Kimpe, and Y. Chen, "Using Property Graphs for Rich Metadata Management in HPC Systems," in Parallel Data Storage Workshop (PDSW), 2014 9th. IEEE, 2014, pp. 7–12.

[9] A. S. Tanenbaum and A. Tannenbaum, Modern Operating Systems. Prentice Hall, Englewood Cliffs, 1992,

[10] D. Dai, P. Carns, R. B. Ross, J. Jenkins, N. Muirhead, and Y. Chen, "An asynchronous traversal engine for graphbased rich metadata management," Parallel Computing, vol. 58, pp. 140–156, 2016.