A novel Community based Web Crawlers (CWC) for Information Retreival

M.Selvakumar
Research Scholar,
Periyar University,
Salem

D.Prabhu
Assistant Professor,
Mailam Engineering College,
Mailam

A. Vijaya Kathiravan
PhD, Assistant Professor,
Computer Applications, Govt Arts
College, Salem

ABSTRACT

Web communities a recent development of the web 2.0 semantic web will have a huge impact on the web crawlers since web 2.0 bring all web technologies under single roof with the advent of mashups its made possible .use of legacy search techniques only increases time and space cost. Hence an novel approach of applying DSA (deducting search algorithm) is adapted. This narrows the search based on evidences using the Holmes engine which follows the active seeker pattern using the DSA algorithm. Our approach is to extract potential accurate information here in which we diverge from producing relevance information from the web communities .

Keywords

Web Crawler,deducting search algorithm, active seeker pattern

1. INTRODUCTION

Information is a production factor in the present www which provides us with mammoth quantity of valuable information by electronic means accessible through hypertext. This large collection of hypertext is changing dynamically and semantically unstructured, making us finding the related and valuable information difficult. Addresses this issue a web crawler for automatic discovering of valuable information from the web, or web mining is important for us nowadays. In reality this web crawler is a program, which automatically traverses the web downloading documents and following URL from page to page. They are mainly used by search engine to gather data for indexing. Other possible applications include page validation, structural analysis and visualization update notification, mirroring and personal web assistant's agents etc.
There is vital distinctiveness of web that formulates crawling very complicated:

- huge capacity of data still increasing,
- recurrent rate of transform
- Lifeless links.
- newness of information
- Dynamic page production.

The huge capacity implies that the crawler can only download fraction of the web page within a given time, so needs to prioritize downloads. Possible hurdles being faced are recurrent rate of transform of data (information). Still worse is the available of lifeless links (both inter & intralinks).continuous development of gadgets and advancements in all fields produce huge new information which leads to dynamic page production although several web technologies like AJAX can be used for dynamic web content generation all these are possibly for the web . Similarly they are also applicable for web communities to several web communities are present still lot are up in the rise

2. WEB COMMUNITIES

Classification of web communities based on average degree, clustering coefficient, and average path length to analyze a virtual friendship network [4].based on registered users[1]node classification[3].age restriction [5] page views i.e. unique page visits [6] and the list can go on almost all the things named in the world comes under some kind of web community.
Each web community has its own rules and regulations and there is no such thing called as a generalized framework for social networks .hence it give rise to certain bottlenecks in handling information retrieval from these communities. It is an open forum for discussions and has several problems to be addressed.
Web crawling on web communities is done by web scutters A scutter uses the data smushed to create a database where uniquely identifiable objects were linked, even where gathered from different sources.

3. WEB COMMUNITY ANALYSIS

Web community analysis (WCA) is the systematic investigation of social networks. Social network analysis views social relationships in terms of network theory consisting of nodes (representing individual actors within the network) and connections or links (which represent relationships between the individuals, such as closeness, affiliation, organizational position, sexual relationships, etc.)[8][9] These networks are often depicted in a social network diagram, where nodes are the points and ties are the lines.

3.1 METRICS (MEASURES) IN SOCIAL NETWORK ANALYSIS

3.1.1 Bridges

An edge is said to be a bridge if deleting it would cause its endpoints to lie in different components of a graph.

3.1.2 Centrality

Centrality refers to a group of metrics that aim to quantify the "importance" or "influence" (in a variety of senses) of a particular node (or group) within a network.[10][11][12][13] Examples of common methods of measuring "centrality" include between's centrality[14], closeness centrality, eigenvector centrality, alpha centrality and degree centrality.

3.1.3 Clustering coefficient

A measure of the likelihood that two associates of a node are associates themselves. A higher clustering coefficient indicates a greater 'cliquishness'.[15]

3.1.4 Cohesion

The degree to which actors are connected directly to each other by cohesive bonds. Groups are identified as 'cliques' if every individual is directly tied to every other individual, 'social circles' if there is less stringency of direct contact, which is imprecise, or as structurally cohesive blocks if precision is wanted. Structural cohesion refers to the minimum number of members who, if removed from a group, would disconnect the group.

3.1.5 Density

Density measures the proportion of ties in a network relative to the total number possible.

3.1.5.1 Freshness

This is a binary measure that indicates whether the local copy is accurate or not. The freshness of a page *p* in the repository at time *t* is defined as:

$$F_p(t) = \begin{cases} 1 & \text{if p is equal to the local copy at time t} \\ 0 & \text{otherwise} \end{cases}$$

3.1.5.2 Age

This is a measure that indicates how outdated the local copy is. The age of a page *p* in the repository, at time *t* is defined as:

$$A_p(t) = \begin{cases} 0 & \text{if p is not modified at time} \\ t - \text{modification time of p otherwise} \end{cases}$$

Coffman*et al*. wo crawler that is equivalent to freshness, but use a different wording: they propose that a crawler must minimize the fraction of time pages remain outdated. They also noted that the problem of Web crawling can be modeled as a multiple-queue, single-server polling system, on which the Web crawler is the server and the Web sites are the queues. Page modifications are the arrival of the customers, and switch-over times are the interval between page accesses to a single Web site. Under this model, mean waiting time for a customer in the polling system is equivalent to the average age for the Web crawler. The objective of the crawler is to keep the average freshness of pages in its collection as high as possible, or to keep the average age of pages as low as possible. These objectives are not equivalent: in the first case, the crawler is just concerned with how many pages are out-dated, while in the second case, the crawler is concerned with how old the local copies of pages are.

4. TYPES OF CRAWLER

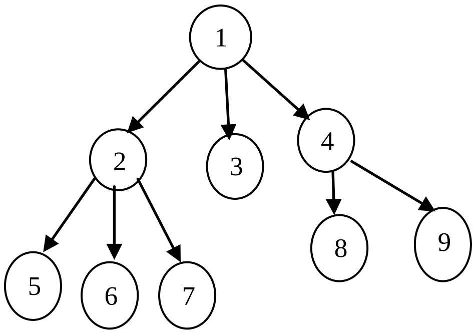
4.1 BREADTH FIRST SEARCH

Breadth-first search is the simplest strategy for crawling. It does not utilize heuristics in deciding which URL to visit next. All URLs in the current level will be visited in the order they are discovered before URLs in the next level are visited. Although breadth-first search does not differentiate Web pages of different quality or different topics, it is well suited to build collections for general search engines. However, recent research showed that breadth-first search could be also used to build domain-specific collections.

The assumption here is that if the starting URLs are relevant to the target domain, it is likely that pages in the next level are also relevant to the target domain. Results from previous studies have shown that simple crawlers that fetch pages in a breadth-first order could generate domain-specific collections with reasonable quality. However, the size of collections built by such simple crawlers cannot be large because after a large number of Web pages are fetched, breadth-first search starts to lose its focus and introduces a lot of noise into the final collection.

Other researchers have tried to use breadth first search and Web analysis algorithms together in focused crawling [5]. In their approach, Web pages are first fetched in a breadth-first order, and then irrelevant pages are filtered from the collection using a Web analysis algorithm. Compared to using breadth-first search alone, this combined method can build much larger domain-specific collections with much less noise.

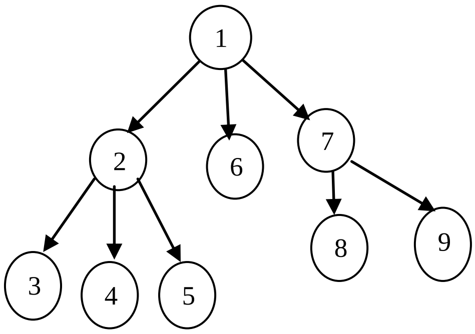
However, since a lot of irrelevant pages are fetched and processed by Web analysis algorithms during the crawling process, this method suffers from low efficiency.



Breadth- First Search

4.2 DEPTH FIRST SEARCH

Depth- First Search is search algorithms that starts with the root node and explores as far as possible along each branch before backtracking Breadth –First Search crawlers need more space to store all visited pages in each node level than DFS crawlers, which only need to store visited pages in one branch of the web graph. DFS crawlers, however, can be trapped by infinite link loop. Depth-first requires memory of only depth times branching-factor (linear in depth) but gets “lost” pursuing a single thread



Depth- First Search

4.3 FOCUSED CRAWLER

A focused crawler or topical crawler is a web crawler that attempts to download only web pages that are relevant to a pre-defined topic or set of topics. Topical crawling generally assumes that only the topic is given, while focused crawling also assumes that some labeled examples of relevant and not relevant pages are available. Topical crawling was first introduced by Menczer. Focused crawlers aim to search and retrieve only the subset of the world-wide web that pertains to a specific topic of relevance. The ideal focused crawler retrieves the maximal set of relevant pages while simultaneously traversing the minimal number of irrelevant documents on the web. Focused crawlers therefore offer a potential solution to the currency problem by allowing for standard exhaustive crawls to be supplemented by focused crawls for categories where content changes quickly. Focused crawlers are also well suited to efficiently generate indices for niche search engines maintained by portals and user groups, where limited bandwidth and storage space are the norm. Finally, due to the limited resources used by a good focused crawler, users are already using personal PC based implementations. Ultimately simple focused crawlers could become the method of choice for users to perform comprehensive searches of web-related materials

4.4 PERIODIC CRAWLER

The periodic crawler can index a new page only after the next crawling cycle starts, but the incremental crawler may immediately index the new page, right after it is found. Given the importance of web search engines (and thus web crawlers), even minor improvement in these areas may enhance the users' experience quite significantly.

5. COMMUNITY BASED CRAWLERS (CWC)

In the community based crawlers we are going to first categories the community .once categorized we form pattern through the help of data mining techniques.

By applying this pattern we can structure the web communities thereby classify them in cluster based on the gender, age, area, web usage and lots more.

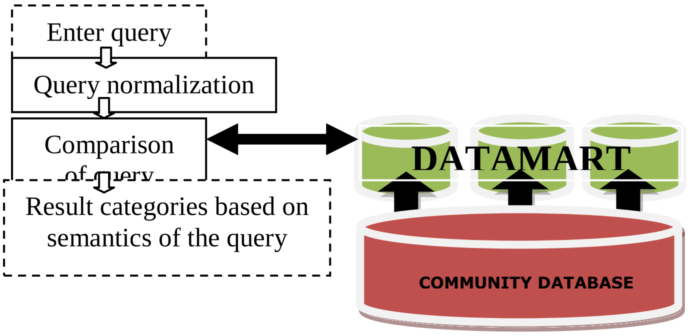


Figure 1 PROPOSED CWC ARCHITECTURE

we form data marts from the existing community database. which reduces the time taken for searching the whole database .in the data mart the information are categorized based on the semantics hence the crawling of data in these is much faster when compared to that of the existing crawlers.

6. CONCLUSION AND FUTURE WORK

Here we have come across the various types of crawlers their comparison with the existing methodologies taking into account the metrics of the crawlers. a framework is been proposed for the development of community based crawlers with much importance for retrieval time . After the query normalization here we take the semantics of the query thereby making it suitable for effective data retrieval from the web communities. Further this can be developed for a multilingual search query input and optimized for better semantic search data results.

7. REFERENCES

[1] "List of Social Networking Websites." *Wikipedia*. Wikimedia Foundation, Inc., 2011. Web. 28 April 2012.

[2] Moffatt, K.. "Top 15 most popular social networking sites."*The ebusiness knowledgebase* . eBizMBA Inc, 2012. Web. 29 April 2012.

[3]Node Classification in Social Networks Bhagat, Smriti; Cormode, Graham; Muthukrishnan, S.Social Network Data Analytics, by Aggarwal, Charu C., ISBN 978-1-4419-8461-6. Springer Science+Business Media, LLC, 2011, p. 115

[4]CLASSIFICATION OF SOCIAL NETWORKS Ibrahim Sorkhoh(1), Maytham Safar(2), and Khaled MahdiISBN (Book): 978-972-8924-68-3. Proceedings of the IADIS International Conference on e-Commerce Amsterdam, The Netherlands 25-27 July 2008

[5] "Myspace." *Wikipedia*. Wikimedia Foundation Inc., 2012. Web. 2 May 2012.

[6]Techtree News Staff (2008-08-13). "Facebook: Largest, Fastest Growing Social Network". Techtree.com. ITNation. Retrieved 2008-08-14

[7] Giuseppe Di Battista, Peter Eades, Roberto Tamassia, and Ioannis G. Tollis" Graph Drawing: Algorithms for the Visualization of Graphs"Prentice Hall (1998) **ISBN-10: 0133016153**

[8] Pinheiro, Carlos A.R. (2011). Social Network Analysis in Telecommunications. John Wiley & Sons. p. 4. ISBN 978-1-118-01094-5.

[9]D'Andrea, Alessia et al. (2009). "An Overview of Methods for Virtual Social Network Analysis". In Abraham, Ajith et al.. Computational Social Network Analysis: Trends, Tools and Research Advances. Springer. p. 8. ISBN 978-1-84882-228-3.

[10] Hansen, Derek et al. (2010). Analyzing Social Media Networks with NodeXL. Morgan Kaufmann. p. 32. ISBN 978-0-12-382229-1.

[11] Liu, Bing (2011).Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data.Springer. p. 271 ISBN 978-3-642-19459-7.

[12]Hanneman, Robert A. & Riddle, Mark (2011). "Concepts and Measures for Basic Network Analysis". The Sage Handbook of Social Network Analysis. SAGE. pp. 364-367. ISBN 978-1-84787-395-8.

[13] Tsvetovat, Maksim & Kouznetsov, Alexander (2011). Social Network Analysis for Startups: Finding Connections on the Social Web. O'Reilly. p. 45. ISBN 978-1-4493-1762-1.

[14]Wasserman, Stanley, & Faust, Katherine. (1994). Social Networks Analysis: Methods and Applications. Cambridge: Cambridge University Press. A short, clear basic summary is in Krebs, Valdis. (2000). "The Social Life of Routers." Internet Protocol Journal, 3 (December): 14–25

[15]Hanneman, Robert A. & Riddle, Mark (2011). "Concepts and Measures for Basic Network Analysis". The Sage Handbook of Social Network Analysis. SAGE. pp. 346-347. ISBN 978