IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Suite 200, Hershey PA 17033, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-gro

ITP4823

# Dynamic Indexing of Information in the Web: the Case of News Sites

Luciano de A. Barbosa[†], Mariano Cravo T. Neto[†], Ana Carolina Salgado[†] and Franklin de S. Ramalho[*]

[†]Centro de Informática – UFPE, Av. Prof. Luiz Freire, s/n
Cidade Universitária, 50740-540
Tel: +55 81 3271.8430, Fax : +55 81 3271.8438
Recife-PE, Brasil
{lab,mctn,acs}@cin.ufpe.br

[*]Departamento de Sistemas e Computação – UFCG, Av. Aprígio Veloso, s/n, 58109-070
Caixa Postal: 10.106
Tel: +55 83 310 1119
Campina Grande-PB, Brasil
franklin@dsc.ufcg.edu.br

## ABSTRACT

*This paper presents a solution to keep available up-to-date information in a search engine whose scope is the content available within news web sites. This solution is based on the use of non-uniform policy to update the documents belonging to this scope. In order to use the non-uniform policy, we identify the most and the least recently updated documents, based on the idea in which it is supposed that the closest documents of the root of a site are the most modified ones. This hypothesis was verified through an experiment within news sites. In order to demonstrate the efficiency of our solution regarding a traditional one, we performed a case study whose results showed that: our solution spent less time to make the new information available, it made fewer requests to the web server, it kept a high freshness of the scope and, finally, it kept the search engine index up-to-date for a much longer time than the traditional solution.*

## 1 INTRODUCTION

The terrorist attacks of September 11 led millions of people to seek recent information in the Web about what had happened. Some people sought it in newspaper sites, others searched this information in search engines. However, the domain general search engines, as Google [8], were not prepared for this kind of demand because they did not have a specialized search in news content [14]. Nowadays, there are already some search engines with specialized search in news content as own Google.

Nowadays, there are already some search engines with specialized search in news content as Google. The main motivation to build a domain specific search engine is to allow the user to query pages that are not likely to be present in more generic search engines and, yet, provide more search capabilities through specialized interfaces. As the indexes of these search engines are smaller than a generic one, the index of a domain specific search engine can be updated more frequently [13]. The update process in generic search engines is a well-known problem due to its size. This feature is considerably interesting in special to news site because of its dynamic nature.

According to [12], a very simple and common way of keeping the index up-to-date is to revisit all the documents indexed by the search engine at the same frequency, regardless of how often they change. This policy is known by uniform policy.

However, if the documents that need to be updated have different modification rates, then the uniform approach is not efficient. Since many of the revisited documents were not yet modified, the uniform approach wastes resources, impairing the quality of the database. As we will present in this paper, it is in that way that the sites news behavior. A more suitable policy to that kind of behavior is the non-uniform policy, in which the elements with faster frequency rate are more frequently updated than slower rate ones [3]. However, to efficiently make use of this approach, we need a method that will be responsible for the identification of the most and the least dynamic documents (sites).

In this work we propose a heuristic for the identification of those documents, that is validated by an experiment. Moreover, it is presented a case study in which it is compared the efficiency of approaches, uniform and non-uniform, using the heuristic we propose to the non-uniform.

## 2 CONTEXT AND MOTIVATION

The data freshness of a search engine has been investigated in the literature but, as far as we know, none of the experiments had news content as the adopted scope. Besides, the commercial search engines do not explain how they make available information to this kind of scope. In order to better present the problem, some concepts are useful:

- Speed of the update process: the faster the process speed is, the smaller will be the time to make the scope up-to-date. There are, however, limits that bind the speed of the process:
- Ethical limit: if the crawlers make constant downloads of pages of the sites by making many HTTP requests per second on the same web server for a long period of time, it will certainly overload it. This practice is not acceptable, once it uses much of the resources that are available [9]. That problem is critical while regarding a scope formed by a reduced number of web servers. Once a great amount of web servers are available, the crawler can alternate HTTP requests among the web servers available, not overloading them.
- Operational limit: when a search engine runs over a limited bandwidth and it has limited hardware resources, it is necessary to avoid overloading it.
- Modification rate of the content of the scope: the larger the modification rate of an element of the scope, more times it will be visited to synchronize the sites information with the search engine database.

Since the update rate of news site is high, a search engine specialized in news site works with a restricted scope, looking forward to speed up the update process. This work proposes to make available recent information to the end-user, through the identification of the most and the least modified documents within the news sites, respecting the ethical limits mentioned before.

## 3 A HEURISTIC METHOD

As already mentioned, the non-uniform policy is more suitable for our problem than the uniform one. We need then a method that identifies the most and the least dynamic documents. Based on the following works, we have developed a heuristic to identify such dynamic elements:

- In [4], the authors affirm that when a user looks for information in a particular site, the user often starts from the site root page and goes following links. Since the user cannot indefinitely follow these links, if the information is not found within a few links from the root, the user concludes that the information he wants does not exist or has disappeared from that site. Therefore, many users often look at just a small subset of pages from a site, and not the entire site. Knowing that, it is supposed that the editors of a site give preference to update the pages near the root.

Copyright © 2003, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

- In [5], the web connections of two great corporations were monitored in order to evaluate the rate and the nature of the changes of the web resources. One of its conclusions is that pages more frequently accessed are younger than those less accessed. In [7], more specific for the Brazilian Web, they concluded the same thing.

Based on these works, we have the following hypothesis: the probability of a page to be visited is proportional to how closer it is from the root. Also, when a page is frequently visited it has a high probability to be changed. Supported by this hypothesis, and we created a method that identifies the sets of documents most and least modified of a site, based on the following idea: the pages closest to the site root are those most modified.

The basic concept of this method is **anchor**. The anchor of a site is a document that links the most recent documents of the site. In our experiments, we have checked that the anchor was either the initial page of the site or a close link to it. The pages linked by the anchor have depth 1, the ones linked from those have depth 2, and so on.

In order to make this concept clear, we present the following example (Figure 1): we need to find the anchor of a news site to keep it updated in the search engine index. The initial URL of this site is http://www.news.com. Observing this web page, we see that it is just a frame with two references (*hrefs*), one pointing to *menu.html* and the other one to *index2.html*. The initial page, thus, is a static page. Observing the content of these two links we have noticed that the *menu.html* is also static, because its content and the content of its links are not news. Browsing the site, we have also noticed that the index2.html has references to all the content of more recent edition of the site. Hence, this page (http://www.news.com/index2.html) should be chosen to be the anchor of that site. The hyperlink graph of this site is shown in the Figure 1.

The depth of a node (web page) of the hyperlink graph to an anchor will show whether a document is more likely to be modified. To use this heuristic, it must be chosen the depth value *n*. These elements with depth less or equal to *n*, which form the more dynamic group, must be updated more frequently, while the remaining ones, which form the less dynamic group, must be updated with a less periodicity. To update these groups, it must be specified *n*, according to the ethical and operational limits mentioned early, because as larger the value *n*, larger will be the size of the dynamic group, which becomes more difficult the update process.

### Experiment

To check the proposed hypothesis we performed an experiment in the following way:

1. Site selection: it was chosen for specialists of the Radix search engine [11], 61 news sites of the Brazilian Web based on their popularity.
2. Anchors identification: human experts manually chose the anchors of the selected sites. During this process, we have verified that inside these sites, the anchor was either the root of the site or a very close link to it. The results of this process confirm our hypothesis: if the page is near the root site, e.g., with a small depth, it has a great probability to be changed.
3. Monitoring pages[1]: each selected site at step 1 was visited during 7 consecutive days (September, 2002) in a breadth-first search, starting from the anchor until *n*=4.
4. Change identification: to verify whether a page was modified, it was stripped HTML syntax of the page's content, generated its checksum (MD5 algorithm [10]) and it was compared to the page checksum of the previous day with the one of the current day. The MD5 algorithm was chosen due its wide use in digital signature verifications.

After the execution of this experiment, we obtained the percentage values of daily out-of-date pages with depths 1, 2, 3 and 4, starting from the anchor. These values are illustrated in Figure 2.

The first we can conclude from these numbers is that the news pages have not all the same update frequency, as noticed by [4] in the whole Web study. Therefore, the non-uniform policy is much more suitable for being the update policy of the news site specialized search engine. And second, the result of the experiment confirms our hypothesis, and the heuristic proposed shows a way to identify the most and least dynamic documents (pages) of the news site.

## 4 CASE STUDY

As we introduced in last section, the proposed heuristic really identifies the most and least dynamic elements of news sites. In this case study, we will show how its use, together with a non-uniform policy, is most efficient than traditional update policy. To compare the efficiency between these approaches, we will use the following measures:

- Success rate: it is the percentage of visited pages that were modified. This value indicates how much the computational resources are wasted.
- Time to make information available: the interval of time to update the pages and make them available to the users.

In this study, the scope was the 61 news sites that were used in experiment presented in last section. In order to update and obtain the data, we used the Radix search engine. In the database used to this study, there was 1,475,000 documents belong the 61 news sites.

### Uniform Policy

In the uniform policy all database elements are visited at the same rate, so it is necessary to visit at least once a day, in order to update the information. Using this approach, we executed the update process of the scope at the rate of 141 visited pages per minute, speed in which the entire site would be revisited in 5 days and 16,7 hours. However, due to the ethical limits mentioned early, we just allowed the process to be executed for 6 hours. The percentage of success was 22%. These num-

*Figure 1 - Example of hyperlink graph of a web site and its anchor.*

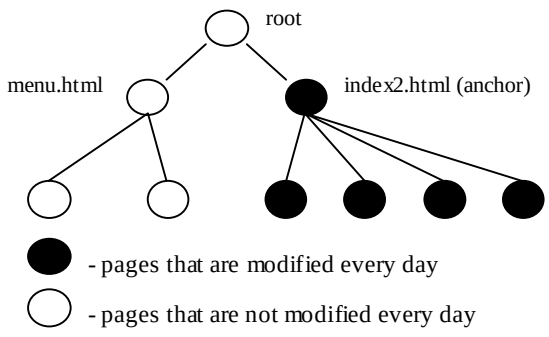
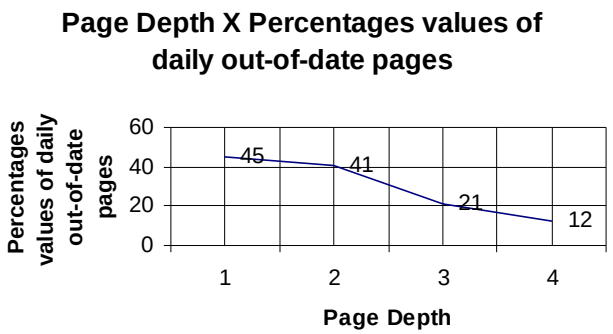



*Figure 2 - Percentages of modified documents to different levels of depth in relation to anchor.*




Copyright © 2003, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

bers show that the traditional update policy was not appropriate to this kind of scope, because the information will not be available daily and due to the waste of resources.

**Non-Uniform Policy**

In order to adopt the non-uniform policy, we have to identify the group of the more and the less recently modified pages. For that identification, we used we used the presented heuristic in the following way:

1. Anchor choice: the anchors used in this case study were the same used in the experiment in previous section.
2. Depth definition in relation to the anchors[2]: starting of the anchors, it was visited pages until levels $n<=1$, $n<=2$ e $n<=3$, obtaining the values presented in Table 1. With $n<=2$, there is a lower waste of resources from Web servers than $n<=3$, and a bigger number of modified pages, 6,244 pages to $n<=2$ and only 1,580 to $n<=1$. Therefore, for this case study, the chosen depth was $n<=2$.

Each 6 hours the pages belong to $n<=2$ were visited at rate of 141 pages per minute. The expended time to execute it was 1.7 hour. The success rate was 42%. This process can be accomplished in a lower speed, to decrease the number of requisitions to Web server. There is, however, a trade off between time to make available the recent information and the number of requisitions. The rest of the database (less dynamic pages) can be updated in a lower rate, to not overload he Web servers. Thus, the main goal, to make available the most recently modified documents to the end user through the search engine index, was successfully reached without overloading the Web servers.

**Discussion**

In Section 3 we showed that the uniform policy was not adequated to this kind of scope, due to the fact that pages are modified at different rates. This case study confirms that fact through the numbers presented. For instance, the time to make the information available is about 7 days, what would be a long term to search service in news sites, since the users are interested in the most recent information. Otherwise, using our approach, several updates can be done at same day, making available the information to users in 1.7 hour.

Finally, the analysis of both measures show that, with our approach, even visiting the pages in a short interval, we obtained a success rate bigger than the traditional approach, that, visited the pages in a longer interval (7 days).

Therefore, this study case demonstrated the viability of our approach to update the news content, because it allowed to make available the recently released information several times a day, contrary to the uniform policy.

## 5 RELATED WORKS

There are several works that focus on the area of freshness in search engine databases [1, 2, 3, 4, 6].

The approach [1] treats the problem using Web server information to update the database. It proposes that the Web server keeps a file with an URL list and their respective modification dates. Thus, initially, the crawler downloads the list file, identifies the URLs that were modified since its last visit, and requests only the modified pages. That approach avoids the waste of Web server resources but, on the other hand, it demands modifications to the Web server implementation. However, Web server implementations do not follow any standard, which makes difficult its use. Our approach, on the contrary, is easily used in practice, but it does not know exactly which pages were modified and which were not in the Web Server.

*Table 1 - Values of success rate and number of pages to the depth 1,2 and 3.*

| value of *n* | success rate | number of pages |
|---|---|---|
| $n<=1$ | 45% | 3,510 |
| $n<=2$ | 41% | 15,230 |
| $n<=3$ | 21% | 42,460 |

The work [3] proposes an optimal frequency of revisitation to the crawler that is closer to the uniform policy than to the non-uniform policy. The authors [2] formulate a model to schedule the update process using uniform policy. However, the uniform policy is not appropriated to this kind of scope.

The others approaches [4, 6] use the non-uniform policy as solution to update the search engine databases. They try to identify the pages' behavior based in their historical information of changes. However, they do not make any *a priori* assumptions about the distribution of page change rates.

In [4], to learn how a document behaves according to its modification rate along the time, it was proposed the use of statistical estimators based on historical information of changes. In [6], the authors used a mathematical model to learn the page's behavior.

The main problem of those approaches is the need of waiting a certain period of time to learn in which class a document is, which is not interesting to the scope news sites, because, for instance, much of the new content of a edition may be found in pages that did not exist before. This problem does not occur in our approach, because it can find the recent pages and update this information in the database without any historical information of changes.

## 6 CONCLUSIONS AND FUTURE WORK

It was presented in this work a proposal for the problem of maintaining as up-to-date as possible the information of news sites of the Brazilian Web in the search engine database.

The proposal is easy to be implemented. It is based on building a domain specific search engine that uses the non-uniform policy and on the identification of the most and the least recently modified documents according to the idea that the closest documents of a site root are the most modified ones. It was shown a better and efficient solution compared to the traditional way of update search engine databases, as stated by the numbers presented in the study case.

Comparing to the others update policies presented, we can conclude that our approach is more efficient than them to update this kind of scope.

One of the limitations for the use of that heuristic is when the scope contains many news sites, because even with a small depth, there will be a very big number of elements to update. An alternative to treat this problem is to split that scope in smaller scopes.

As future work, it is proposed to implement an automatic method for the selection of the anchor, and also to verify if the heuristic to identify the most dynamic documents is valid to the whole Web.

## ENDNOTES

1 The one-week period of the experiment and the *n* value equals to 3 was limited due to the ethical limits mentioned previously because in some sites it was crawled about 5,000 links daily.

2 As we wish to update the pages every 6 hours, we limited the depth to $n<=3$, because a bigger depth would generate a lot of pages to update, what it would not be possible to update completely due to the operational limits showed in Section 3.

## REFERENCES

[1] Brandman, O.; Cho, J.; Garcia-Molina, H e Shivakumar,N. "Crawler-Friendly Web Servers". In Proceedings of the Workshop on Performance and Architecture of Web Servers (PAWS) , Santa Clara, 2000.

[2] Brewington, B. E. e Cybenko, G. "How Dynamic is the Web?". In Proceedings of the 9th Worldwide Web Conference (WWW9), Amsterdam, 2000.

[3] Cho, J. e Garcia-Molina, H. "Synchronizing a Database to Improve Freshness". In Proceedings of 2000 ACM International Conference on Management of Data (SIGMOD), Dallas, 2000.

[4] Cho, J. e Garcia-Molina, H. "The Evolution of the Web and

Copyright © 2003, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

Implications for an Incremental Crawler". In Proceedings of 26th International Conference on Very Large Databases (VLDB), *Cairo, 2000.*

[5] Douglis, F.; Feldman, A e Krishnamurthy, B. "Rate of Change and other Metrics: a Live Study of the World Wide Web". In Proceedings of the USENIX Symposium on Internetworking Technologies and Systems, Boulder, 1999.

[6] Edwards, J.; McCurley, K. e Tomlin, J. "An Adaptive Model for Optimizing Performance of an Incremental Web Crawler". In Proceedings of the 10th World-Wide Web Conference (WWW10) , Hong Kong, 2001.

[7] Fonseca, N. M.; Resende, R. S. e Pádua, C. I. P. S. "Aspectos Dinâmicos da Web Brasileira". XXVII Seminário Integrado de Software e Hardware (Semish), Curitiba, 2000.

[8] Google Search Engine. http://www.google.com.

[9] Koster, M. "Guidelines for Robot Writers". http://info.webcrawler.com/mak/projects/robots/guidelines.html.

[10] Krause, M. e Tipton, H. F. "*Information Security Management Handbook*", Quarta edição, Volume 1, Editora Auerbach.

[11] Radix Search Engine. http://www.radix.com.br.

[12] Silva, A. S.; Veloso, E. A.; Golgher, P. B.; Ribeiro-Neto, B.; Laender, A. H. F. e Ziviani, N. "CoBWeb – A Crawler for the Brazilian Web". In Proceedings of String Processing and Information Retrieval Symposium & International Workshop on Groupware, 1999.

[13] Steele, R. "Techniques for Specialized Search Engines". In Proceedings of Internet Computing '01, Las Vegas, 2001.

[14] Wiggins, R. W. "The Effects of September 11 on the Leading Search Engine". First Monday Peer-Reviewed Journal on the Internet, 2001.

Copyright © 2003, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

0 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/proceeding-paper/dynamic-indexing-information-web/32004

## Related Content

E-Business Supply Chains Drivers, Metrics, and ERP Integration
Jean C. Essila (2018). *Encyclopedia of Information Science and Technology, Fourth Edition (pp. 5345-5356).*
www.irma-international.org/chapter/e-business-supply-chains-drivers-metrics-and-erp-integration/184238

Parallel and Distributed Pattern Mining
Ishak H.A Meddahand Nour El Houda REMIL (2019). *International Journal of Rough Sets and Data Analysis (pp. 1-17).*
www.irma-international.org/article/parallel-and-distributed-pattern-mining/251898

Understanding the Context of Large-Scale IT Project Failures
Eliot Richand Mark R. Nelson (2012). *International Journal of Information Technologies and Systems Approach (pp. 1-24).*
www.irma-international.org/article/understanding-context-large-scale-project/69778

Competitive Intelligence from Social Media, Web 2.0, and the Internet
Sérgio Maravilhas (2015). *Encyclopedia of Information Science and Technology, Third Edition (pp. 558-566).*
www.irma-international.org/chapter/competitive-intelligence-from-social-media-web-20-and-the-internet/112369

Semantic Measures
Yoan Chabotand Christophe Nicolle (2015). *Encyclopedia of Information Science and Technology, Third Edition (pp. 4690-4698).*
www.irma-international.org/chapter/semantic-measures/112911