View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Universidade do Minho: RepositoriUM

Geographical Partition for Distributed Web Crawling *

José Exposto
ESTiG - Instituto Politécnico de Bragança
Bragança - Portugal
exp@ipb.pt

Joaquim Macedo
DI - Universidade do Minho
Braga - Portugal
macedo@di.uminho.pt

António Pina
DI - Universidade do Minho
Braga - Portugal
pina@di.uminho.pt

Albano Alves
ESTiG - Instituto Politécnico de Bragança
Bragança - Portugal
albano@ipb.pt

José Rufino
ESTiG - Instituto Politécnico de Bragança
Bragança - Portugal
rufino@ipb.pt

ABSTRACT

This paper evaluates scalable distributed crawling by means of the geographical partition of the Web. The approach is based on the existence of multiple distributed crawlers each one responsible for the pages belonging to one or more previously identified geographical zones. The work considers a distributed crawler where the assignment of pages to visit is based on page content geographical scope. For the initial assignment of a page to a partition we use a simple heuristic that marks a page within the same scope of the hosting web server geographical location. During download, if the analyze of a page contents recommends a different geographical scope, the page is forwarded to the well-located web server.

A sample of the Portuguese Web pages, extracted during the year 2005, was used to evaluate: a) page download communication times and the b) overhead of pages exchange among servers. Evaluation results permit to compare our approach to conventional hash partitioning strategies.

Categories and Subject Descriptors

H.3 [**Information Systems**]: Information Storage and Retrieval

General Terms

Design, Experimentation, Performance

Keywords

Web Mining, Parallel Crawling, Web Partitioning

*Research supported by FCT/MCT, Portugal, contract POSI/CHS/41739/2001, under the name "SIRe - Scalable Information Retrieval Environment".

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
GIR'05, November 4, 2005, Bremen, Germany.
Copyright 2005 ACM 1-59593-165-1/05/0011 ...$5.00.

1. INTRODUCTION

In our days, the enormous number of existent Web pages suggests the use of distributed algorithms and systems for processing the huge volume of related information. In particular, the appearance of distributed crawling system architecture and applications has proved to be well suited to achieve a good coverage and freshness of a target web space.

In this context, an important issue is the partition of the Web space between multiple cooperative crawlers. The most popular partition strategies are the ones based on the definition and application of a specific hash function to the name of a host (or its associated IP address). Other approaches exploit the partitioning of the Web taking into account the network (using for instance Round Trip Times) or the Web link topological proximity (using the number of mutual links). An orthogonal approach to handle the large scale of web is the use of topic focused crawlers.

This investigation is based on a special kind of focused crawler, which centres in a particular geographical scope. It is supported by a distributed architecture made up of multiple collaborative crawlers, each one responsible for a specific geographical scope. As so, during download crawlers collect the information enclosed in its scope, transferring the URL and other page information to more appropriated crawlers, whenever the analysis of the page content determine that the page is related to a different geographical scope. It is assumed the existence of a tool that maps web page contents to corresponding geographical scopes allowing for each page to determine its geographical location and send it to the appropriate crawler.

The assignment of geographical scopes to pages is based on the assumption that p is the probability that the content of a page has the same scope of the corresponding Web server location (S_i). The probability to have a different scope ($S_j \neq S_i$) is assumed to be $\frac{1-p}{N-1}$, where N is the number of total scopes.

The download and exchange time costs are computed by changing the scope probability (p) and the number of geographical partitions. The total space is partitioned according to the geographical proximity of Web servers, using a graph partitioning algorithm.

2. RELATED WORK

Cho and Molina [5] defined a distributed crawl taxonomy and proposed several partitioning strategies, evaluated using a set of defined metrics. The work includes guidelines on the implementation of parallel crawler's architecture, using host name hashing as the base partitioning technique.

Unlike parallelization, focused crawling [4] collects from the Web only the pages relevant for a given set of topics, avoiding unrelated regions of the Web space. Focused crawling reduces network and computing resources and improves the freshness of collected pages.

The work reported in [1] claims a more general crawling strategy than focused crawling that uses the Web link structure to compute the probability of the URL candidate to be assigned to a given crawl. The crawl criteria are given by arbitrary predicates. Web space partitioning based on page classifier tools is another work that is based on collaborative focused crawling [6].

Our investigation may be viewed as a specialization of the last work, as the geographical scope may be considered as a special topic and it is also supported by a collaborative multiple crawler system. However, because geographical based classification has some singular characteristics, there are still unexploited issues, as is the case of the existent relation between the geographical location of web servers and the geographical scope of the hosted pages.

In a work in progress [8], we are proposing and evaluating a multi-criteria Web partitioning strategy based on IP network latency and Web topology to assist on the task to produce an optimized partition of the Portuguese Web. With the present work we plan to enrich the partition strategy by using geographical information (server location and geographic scope of web pages or servers) as an additional criterion.

3. GEOGRAPHICAL CRAWLING PARTITIONING

Web page geographical scope may be calculated using Web page content parsing [15] or by exploiting the link structure of the Web [7]. Our approach is based on the calculation of the Web site geographical scope on the assumption that enclosed pages have a correlated geographical scope. For determining groups (partitions) of geographically close connected Web sites we use the concept of geodesic distance between Web sites.

In a distributed Web crawler system different sub-spaces of the Web are assigned to each crawler. Each sub-space is a geographical scope containing Web sites within the same scope. This division allows for geographical topic focused crawling and is extended with information about Web link structure. The link structure reveals its importance both minimizing the exchange of information and Web space load balancing among the distributed crawlers.

In normal crawling operation the extracted links can be used to discover the Web topology to create a Web link graph and compute the density of links between pages. To determine the geographical scopes of the Web sites we used a set of heuristics that compute their geographic location. Based on the geodesic distance between two separated locations we build a geodesic distances graph where vertices are host machines and edge weights are the respective distances.

Web pages may be aggregated into coarser entities. Since one single Web server (also designated by IP or server) may accommodate several Web sites, and a Web site may be associated with several URLs, it is possible to unify the distance to a quite large number of URLs and also simplify the Web link graph.

The two graphs may be condensed by just one graph, where the edges have two weights and vertices are single weighted with the number of pages contained in that vertex. Each vertex represents an IP and the edges are weighted with both geodesic distance and Web links between vertices. The partitioning of a multi-weighted edge graph falls into the category of multi-objective partitioning algorithms. The crawling partitioning strategy results in a multi-objective graph partitioning.

3.1 Multi-level partitioning

The graph partition problem corresponds to the need of dividing the vertices of graph into a certain number of parts such that the sum of the weights of the edges is minimized among the different parts. The graph partitioning algorithm used in this work runs at lower level after several stages of coarsening the original graph, by collapsing vertices and edges. Next, the refining phase projects the coarser graphs into the original [10].

We find this type of partitioning algorithms appropriated to the defined problem because it provides for an automatic coarsening mechanism that allows for rapid and good quality partitions. The algorithms also support weighted load balance by assigning weights to the vertices. Edges may also be assigned weights reflecting the affinity between the connected vertices.

In our problem we identify two different partitioning objectives to achieve the overall optimization. In one hand the partitioning of the reduced IP Web link graph which will minimize the inter-partitioning links between IPs, on the other hand, the geodesic distances graph will allow to obtain geographical focused topics by minimizing the proximity between IPs.

3.2 Multi-objective partitioning

To achieve the multi-objective partitioning we followed the procedure suggested in [14] for combining two separate graphs into just one. First, each graph representing each one of the objectives is partitioned separately. Afterwards, a third graph is created by assigning weights to the edges calculated using the sum of the original weights normalized by the partition quality measure of each original graphs called edge-cut. Each normalized weight will be affected by a preferential factor to smooth the differences of magnitude existent between each of the initial graphs. The multi-objective partitioning is the result of the application of the partitioning algorithm to the third graph.

For a Web page link graph a reduced IP graph is created. This reduction works like a coarsening mechanism in which each new IP vertex contains the sum of the pages contained in the Web sites at that IP. The edges are weighted with the sum of the links of the pages belonging to the IPs pointing to pages in other IPs.

In the example of the IP Web graph depicted in Figure 1, each vertex represents an IP with weights associated to the total number of Web pages residing in that IP. The edges weights represent the number of links existent between the connected vertices.

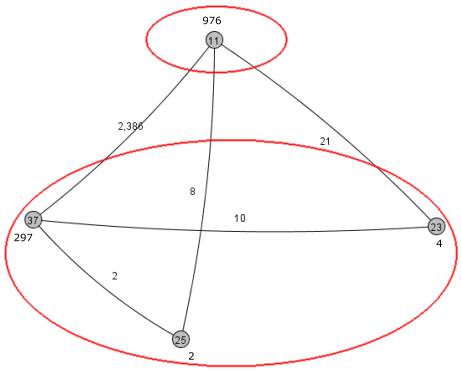

Figure 1: Partitioned IP Web link graph

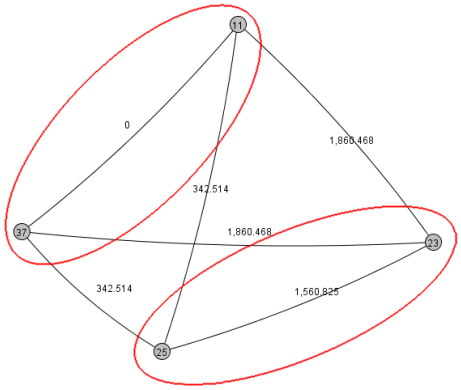

Figure 2: Partitioned distance Graph

In the graph of Figure 1 the ellipses represent the associations obtained by the application of the partitioning algorithm to two partitions. The algorithm automatically produces a balanced Web space. Vertex 11 has a higher number of pages compared to the other vertex, resulting in a sole IP partition.

The geodesic distances between IPs is showed in Figure 2. In this graph, having the same number of IPs of the previous example the geodesic distance weights will be inverted and multiplied by a scaling factor, because the algorithm uses an edge cut minimization scheme. The application of the partitioning algorithm with two partitions is shown in the same Figure.

Finally, Figure 3 depict the weight combination after the normalization of each graph weight using the correspondent edge weight.

4. GEOGRAPHIC AND LINK DATA COLLECTION

To allow the creation of the IP link and IP distance graphs, conventional crawling has to be modified to collect additional data. So, along with the extraction of the URLs belonging to each crawler partition, other kind of information regarding IP physical topology and geographical locations is also collected. Raw link information exposed in the pages is used to build the IP link graph.

IP geodesic distance resultant from IP geographic location is achieved using one of several proposed location heuristics. One of the most accurate sources of IP location is the Location Resource Record (RR LOC) which is available by Domain Name Server (DNS). As we will show, in spite of its popularity it is quite limited.

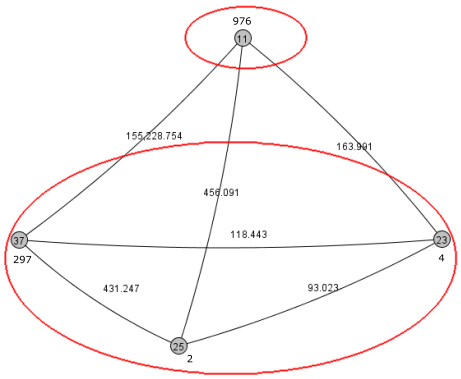

Figure 3: Partitioned Combined Graph

IP hostnames often have semantic context information which may lead to geographic location purpose. Some IPs have airport codes rooted in the hostname. Using a conversion table from airport code to city of 8599 entries, we can easily locate these IPs.

Another simple heuristic that may be used is to take the name of a city also embedded in the name of the host.

Most Web sites belongs to institutions whose name is related with the city name. It is also frequent to find a subdomain related with an institution in the name of the host. We also used a conversion table of 715 institution entries to accomplish this determination.

GTrace [13] is a tool to trace routes including the determination of the location of the traces IPs. In our location determination scheme we also used a collection of 2492 IPs and 9467 hostnames included in GTrace distribution.

Finally, we used NetGeo [3] to determine IP geographic locations. NetGeo parses the address location of Internet resource administrators made available through *whois* service with information of the several Regional Internet Registries. In spite of being very broad the information available in NetGeo is outdated. Nevertheless, since we did not have a similar method to find IP locations we decided to use NetGeo.

In order to calculate geodesic distances we extensively use a database of geographic references called GeoNames [16] because some of the presented heuristics use a city mapping and there is a need to use geodesic coordinates. The approach allows the conversion between cities (and, eventually, other geographic resources) and the respective geodesic coordinates.

In short, the final geographic location of an IP is determined by the first hit in the following heuristics:

1. DNS RR LOC for the IP (DNS);
2. Airport Codes in the hostname (Airport);
3. Name of a city in the hostname (City);
4. Domain of an institution in the hostname (Institution);
5. GTrace Database (GTrace);
6. NetGeo service (NetGeo)

5. EVALUATION

Data used to evaluate the proposed approach to geographic graph partitioning results from the fusion of two previous independent Web collections – NetCensus [12] and WPT03 [17] – both retrieved from the Portuguese Web.

The derived collection, comprising 16,859,287 URLs, has been used to support the development and validation of the partitioning algorithms and to produce statistics of the Portuguese Web. From the collection we obtained link and geographic location data, along with information about the related Internet topology entities and the associated geographic entities. As topological entities we identified the Internet address block (Address Block), the address aggregate published by autonomous systems in BGP routing (Address Aggregate) and the autonomous system (AS). The geographical entities identified include the cities and the respective countries, which will be further localized in terms of geodesic coordinates. The amount of entities with known location is as follow: 42,857 Hostnames; 5,468 IPs; 1,645 Address blocks; 583 Address aggregates; 384 ASs; 339 Cities; and 25 Countries.

It is also interesting to point out that 52% of the IPs belonging to the Portuguese Web reside outside of Portuguese territory.

5.1 Sample size and selection

Even considering only the region of the Web delimited by Portuguese Web, URL downloading and geographical localization tasks are very time consuming. The algorithms we are evaluating also take a long time to run. For these reasons, we decided to pick just a small portion of the initial derived collection to use as a collection test data base.

The size of the sample was calculated based on statistical sample size formulas and is based on the Round Trip Times (RTT) of the involved IPs. Assuming a confidence level of 90% for which $Z(\frac{\alpha}{2}) = 1.645$, let $s = 120.76$ be the standard deviation of the RTTs of the population and $e = \overline{X} - \mu = 20$ the acceptable error between the sample mean ($\overline{X}$) and the population mean(μ), the sample size is given by $n = \frac{Z(\frac{\alpha}{2})s}{e}^2 = 99$.

The data sample, obtained from the initial collection, permitted to use the identified topological and geographical entities as agglomeration unities which concentrate several IPs. Choosing a 99 IP sample of elements of one of these entities would result in a more representative sample. Thus, to achieve a good geographical representation we selected a 99 IP sample of the most populated cities.

5.2 Geographic Evaluation

In order to evaluate each one of the heuristics used to determine the IP geographic location, we present in Table 1 the number of IPs discovered by each of the heuristics.

Airport, City and GTrace heuristics did not came out with any IP location discovery. We found out that Airport and City heuristics are more suitable to hostnames of router IPs than server IPs. Unlike the last observation, Institution heuristic showed to be very adequate to these kind of IPs, achieving 1,081 location determinations. Nevertheless, NetGeo achieved a grand slice of the whole location determination, almost 88%.

Table 2 presents the number of intersected and matched IP locations using two different heuristics. The heuristic combinations not listed correspond to zero intersections. From the heuristics pairs 1-4 and 1-6 there is almost nothing to conclude since the number of intersected IPs is very small. However, for the heuristic pair 4-6 there is 96.9% of intersected IPs, but the percentage of successful matches is only 42.6%. We believe that between heuristics 4 and 6, the first is more accurate due to the nature of the determination which is the domain of the institutions included in the hostname. Because heuristic 6 has the largest number of hit locations we may assert that it is accurate at most for 42.6%, assuming heuristic 4 is roughly 100% accurate.

	Heuristic	# IPs	% total IPs (5,468)
1	DNS	6	0.097
2	Airport	0	0.000
3	City	0	0.000
4	Institution	1,081	17.503
5	GTrace	0	0.000
6	NetGeo	5,433	87.968

Table 1: Number of IPs discovered by each heuristic

HA	HB	Intersection		Match Location	
		#	%	#	%
1	4	1	16.667	1	16.667
1	6	5	83.333	0	0.0
4	6	1,047	96.855	460	42.553

Table 2: Intersected and matched IP locations

5.3 Partitioning evaluation

For partitioning evaluation we used the 99 IP sample. We removed 13 IPs whose geographic distance was not possible to determine, remaining a total of 86 IPs.

By the time of the writing of this paper, information about link structure was still unavailable. To overcome this deficiency, we generated a random graph based in evolving scale-free networks [2] using the Java Universal Network/Graph Framework (JUNG) [9]. The generated graph was defined to have an average out-degree equal to 10 [5] taking into account the number of pages of the initial collection. Further we used the partitioning algorithm available in the *Metis* software package [11].

The resulted partitions were evaluated to measure: i) download time; ii) exchange time; and iii) relocation time, using the RTT between Web sites and crawlers [8].

The download time estimates for a set of partitions is the maximum time taken to download the pages assigned to the partitions. The download time for a server i by the crawler j may be approximated by the formula:

$$dts_i = \frac{M_i}{L_j}(2RTT_{ij} + \frac{L_j \cdot ps_i}{BW_{ij}} + PT_i) \qquad (1)$$

where L_j is the number of pages downloaded in pipeline by `http` persistent connections between crawler j and server i; M_i is the number of pages of the server i; RTT_{ij} and BW_{ij} is the RTT and available bandwidth between the crawler j and the server i, respectively; ps_i is the average page size of the server i. PT_i is a politeness wait interval between consecutive connections to the same server. It is assumed that, at each moment, each crawler may have only one connection to the same server.

Considering a total load of S_j servers for crawler j and N_j `http` simultaneous connections, the total download time for the crawler j is given by equation:

$$dt_j = \frac{1}{N_j} \sum_{l=1}^{S_j} dts_l \quad (2)$$

The maximum total download time is defined by the time taken by the slowest crawler of p partitions, or by the expression: $\max(dt_1, \ldots, dt_p)$.

Because each partition j has several links to other Web sites assigned to different partitions, the associated URLs are forwarded to the crawlers responsible for the corresponding partitions. The estimated time necessary to forward the foreign URLs is given by the equation:

$$et_j = \frac{1}{N_j} \sum_{l=1}^{P} 2RTT_{jl} + \frac{su \cdot nl_{jl}}{BW_{jl}}, l \neq j \quad (3)$$

where RTT_{jl} is the RTT between crawlers l and j, nl_{jl} the total number of links from the partition j to the partition i, su the average size of a URL, BW_{jl} the bandwidth between crawlers and N_j the number of simultaneous links sent.

The total time for all partitions corresponds to the maximum value of all partitions.

The relocation time estimates the amount of time needed to move URLs assigned to a partition within a specific geographic scope to the partition that corresponds to the real geographic scope of those URLs. This metric is similar to Equation 3, except for nl_{jl} which is the number of pages of the partition assigned to crawler l moved to the partition assigned to crawler j having the real geographic scope of the URLs.

Because, at the moment of writing the paper, we were not able to calculate the real geographic scope of a page, we used a variation of the probability (p) of a page geographic scope as an approximately value to the geographic scope assigned to the IP hosting the page. Once more, a maximum value for all partitions is calculated.

It was mentioned already that the position of the crawlers is known in advance, but for the calculations above crawlers are assigned to the geographic centre of partitions. Metrics were evaluated by varying the number of partitions, however, due to the nature of the graphs themselves and the partitioning algorithm, some experiments could not reach the desired number of partitions. Figure 4 shows the variation of the desired partitions compared to the number of partitions accomplished for SIRe.

Results will be used to compare geographic based partitioning and site-hash based partitioning schemes. To achieve a fair comparison the evaluation metrics will use the number of partitions effectively achieved. For the following experiments we used $L_j = 10$, $BW_{ij} = 16Kbps$, $ps_i = 10KB$, $PT_i = 15$, $N_j = 10$, $su = 40$ bytes. For nl_{ji} we used 10% of the 10 average links per page [5].

Figure 5 presents a comparison for download time estimation using Geographic, site-hash based partitioning schemes and a centralized approach.

As expected due to the parallelization processes the increase in number of partitions allows the reduction of the download time for both partitioning methods thus lowering the values obtained by the centralized approach. Site-hash based curve is notoriously higher than the geographic based curve, suggesting that geographic approach is a good alternative to optimize page download times.

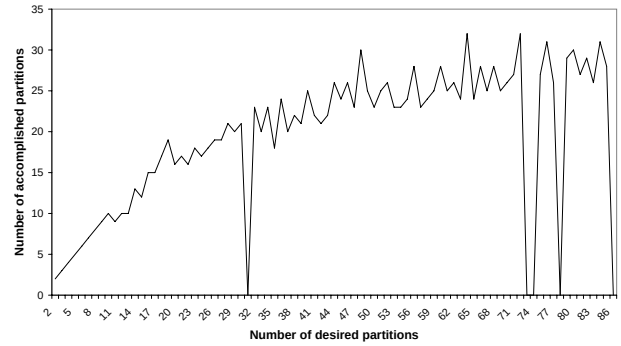

Figure 4: Number of desired vs. accomplished partitions

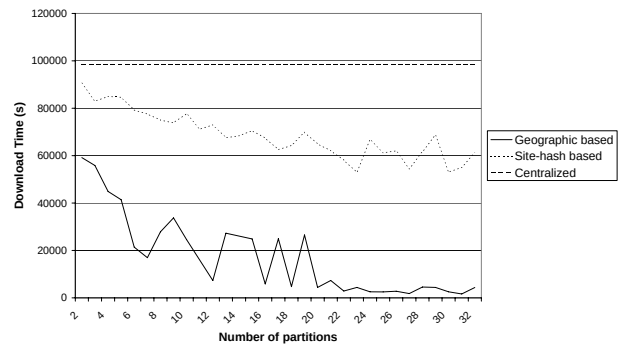

Figure 5: Download Time for Geographic and site-hash based partitioning

Figure 6 presents the results obtained for the estimation of the exchange time. The results are also promising for the geographic based curve.

The two previous charts show a stability tendency for the number of partitions above 20 for the Geographic based curve. For this configuration the observation suggests that is pointless to increase the number of partitions.

Figure 7 presents the results for the estimated relocation time of the URLs by fixing the number of partitions at 20 and varying the probability (p) of the geographic scope of a page be the same of the geographic scope of the partition.

This chart shows an evidence already predictable, the re-

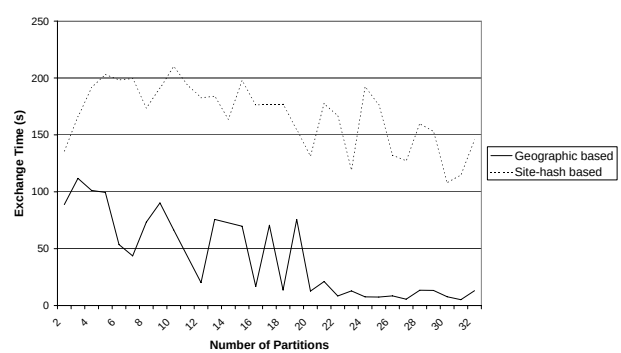

Figure 6: Exchange Time for Geographic and site-hash based partitioning

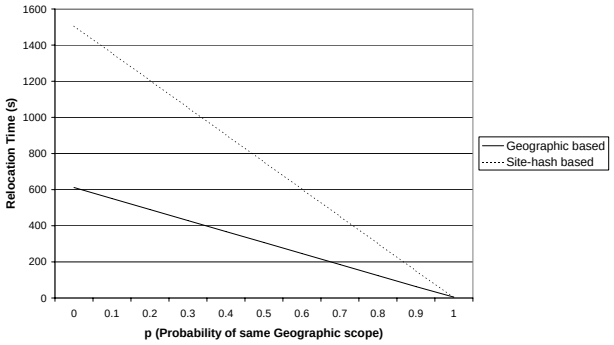

Figure 7: Relocation Time for Geographic and site-hash based partitioning

location time decreases as p increases. But it is also evident the lower times the Geographic partitioning scheme achieves to relocate pages, resulting in a better overall partitioning strategy.

6. DISCUSSION AND FUTURE WORK

The presented work evaluates scalable distributed crawling by means of the geographical partition of the target Web Space. The approach is based on the existence of multiple distributed crawlers each one being responsible for the pages belonging to one or more previously identified geographical zones.

Two types of geographic information are used to achieve the desired load balancing: i) the geographical scope of web pages contents and ii) the geographic location of the hosting web servers. It is assumed that the geographical scope of pages is highly correlated with the geographical position of the servers.

The heuristics used to calculate a Web server's geographic location determine each IP server location without taking into account the topological and geographical relations existent within the located nodes. We hope to improve web server geographic location quality using the routers location and network topology.

NetGeo heuristic although extensive in the number of geographic locations determined has an uncertain accuracy mainly due to the lack of periodically maintenance. It would be desirable to implement a parsing mechanism to the address locations in the registers of the Regional Internet Registries.

The evaluation showed that the estimated results for download, exchange and relocation times have a clear efficiency improvement which is introduced by the use the geographical based crawling whenever a set of crawlers want to collect pages within a pre-established geographical scope over the site-hash based scheme.

These preliminary results must be expanded with the use of real Web page content and link information, instead of a generated random graph. A geographical scope classifier should be used in a more realistic scale. The mentioned estimated results need also to be compared with real crawling results.

In the future we plan to use geographical partitioning strategy as an additional criterion to the existent multi-criteria Web partitioning approach [8].

7. REFERENCES

[1] C. C. Aggarwal, F. Al-Garawi, and P. S. Yu. Intelligent crawling on the world wide web with arbitrary predicates. In *World Wide Web*, pages 96–105, 2001.

[2] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *SIAM Journal on Scientific Computing*, 286(5439):509 – 512, 1999.

[3] CAIDA. NetGeo - The Internet Geographic Database. http://www.caida.org/tools/utilities/netgeo/, 2002.

[4] S. Chakrabarti, M. van den Berg, and B. Dom. Focused crawling: A new approach to topic-specific web resource discovery. In *Proceedings of the 8th International WWW Conference*, 1999.

[5] J. Cho and H. Garcia-Molina. Parallel crawlers. In *Proc. of the 11th International World–Wide Web Conference*, 2002.

[6] C. Chung and C. Clarke. Topic-oriented collaborative crawling. In *11th International Conference on Information and Knowledge Management (CIKM'02)*, pages 34–42, 2002.

[7] J. Ding, L. Gravano, and N. Shivakumar. Computing geographical scopes of web resources. In *26th International Conference on Very Large Databases, VLDB 2000*, Cairo, Egypt, September 10–14 2000.

[8] J. Exposto, J. Macedo, A. Pina, A. Alves, and J. Rufino". Multi-objective web graph partitioning for efficient distributed web crawling. Technical report, University of Minho, Department of Computer Science (Work in progress), 2005.

[9] Jung the java universal network/graph framework. http://jung.sourceforge.net/, 2005.

[10] G. Karypis and V. Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, 20(1):359 – 392, 1998.

[11] G. Karypis and V. Kumar. A software package for partitioning unstructured graphs, partitioning meshes, and computing fill-reducing orderings of sparse matrices – version 4.0, 1998.

[12] J. Macedo, A. Pina, P. Azevedo, O. Belo, M. Santos, J. J. Almeida, and L. Silva. NetCensus Project. http://marco.uminho.pt/m̃acedo/netcensus/, 2001.

[13] R. Periakaruppan and E. Nemeth. GTrace: A graphical traceroute tool. In *13th Conference on Systems Administration (LISA-99)*, pages 69–78, 1999.

[14] K. Schloegel, G. Karypis, and V. Kumar. A new algorithm for multi-objective graph partitioning. Technical report, University of Minnesota, Department of Computer Science, 1999.

[15] M. J. Silva, B. Martins, M. Chaves, N. Cardoso, and A. P. Afonso. Adding geographic scopes to web resources. In *ACM SIGIR 2004 Workshop on Geographic Information Retrieval*, 2004.

[16] US National Geospatial Intelligence Agency. Geographic Names Database. http://earth-info.nima.mil/gns/html/index.html.

[17] XLDB Group. WPT03. Linguateca, http://www.linguateca.pt, 2003.