MINING THE COSTA RICAN WEB

Esteban Meneses
Computing Research Center - Costa Rica Institute of Technology
P.O.Box 159-7050, Cartago, Costa Rica

Keywords: Web mining, local web, web clustering.

Abstract: There is much to say about the structure and composition of a local web. Identification of authorities, topics and web communities can be used to improve search engines, change a portal design or to develop marketing strategies. The Costa Rican web was chosen as a test case for web mining analysis. After the study we obtained several descriptors of the web as well as the answers to typical questions like how many pages on average a site has, which file type is preferred for building a dynamic site, what is the most referenced site, which sites are similar, and many more.

1 INTRODUCTION

The *World Wide Web* is formed by an unusual and heterogeneous set of elements. This property makes the web an important and flexible mechanism for communication and information storage, but makes its analysis hardly easy. Luckily, many recent research efforts have been focused on how to overcome this challenge and provide a set of descriptors from a local web.

This paper presents the results after the analysis of the Costa Rican web. This subset of the global web is formed by every page in a site on the `.cr` domain. Many interesting facts can be discovered from this collection. Things like what is the most referenced site, what is the site with more outgoing links, which sites compose the *core* of the web, how old are web pages, what is the most common file type for dynamic pages, which site is similar to some given web site, and many more, can be revealed by applying several web mining techniques.

When studying a collection of web pages, it is fundamental to analyze it in two dimensions: *web structure* and *web composition*. The former provides the intricate form on which web pages and sites are linked, while the latter offers some description about the topics covered in the web pages or sites. Also, the HTTP protocol generates lots of information with every web transaction. All these statistics are relevant to describe the collection in a physical sense.

The data obtained from the analysis can be applied to a variety of processes: construction of web crawlers, improving web browsing and searching, design of web portals, optimization for web servers, focusing a marketing strategy, among others.

The results presented in this paper were generated using a couple of tools. The Wire set of tools (WIRE) for web mining and analysis and Effimining (PREDISOFT) for web site clustering. Also, it was necessary to develop miscellaneous tools for extracting information from web pages.

However, in the *Klá* research project at the Computing Research Center of the Costa Rica Institute of Technology, we are currently developing several tools for web mining in the `MS .net` platform which are focused on mining a local web. The final objective is to construct a portal for web searching and web collocation analysis.

Section 2 introduces the topic of web mining and many core concepts related with the statistical descriptors of a local web. After that, section 3 presents the results of the analysis made on the Costa Rican web. These statistics can serve as an example of what can be known from a subset of the global web. Section 4 presents some experiments done when clustering some of the Costa Rican web sites. Here, the interest will be focused in the representation used as well as the different strategies for grouping the elements of a collection. Finally, section 5 offers some conclusions and future work.

Meneses E. (2006).
MINING THE COSTA RICAN WEB.
In *Proceedings of WEBIST 2006 - Second International Conference on Web Information Systems and Technologies - Internet Technology / Web Interface and Applications*, pages 414-421
DOI: 10.5220/0001249504140421

2 WEB MINING

This section surveys the basic principles on which web mining is based. These fundamentals come from many disciplines: data structures, information retrieval, formal languages, parsing theory, statistics and many more.

The analysis of a web page collection usually implies at least two dimensions: *structure* and *composition*. The first refers to how pages are linked and that gives a means for authority. The latter is related with the content of a page. These two dimensions are orthogonal and both are usually taken into account in any analysis.

2.1 Structure

The web or a fraction of it can be seen as a big directed graph. If every page is considered as a node and a hyperlink from a url to another is considered as a link, then every collection of web pages is a directed graph, not necessarily connected. However, if sites are considered as nodes, then the same is true.

As a graph, all the theory around it can be applied. Then, shortest paths can be found from a url to another, cliques can be discovered, cycles found, strongly connected components detected, you name it.

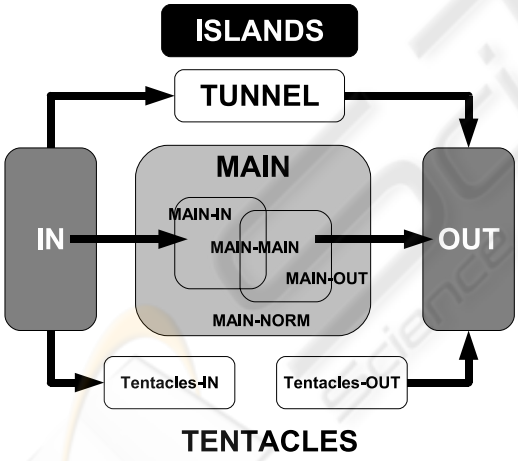

Figure 1: Macrostructure of the Web.

Figure 1 shows the macroscopic structure of the web (Castillo, 2004). It is important to define the following components:

- MAIN: is the set of the biggest strongly connected component in the web graph. It can be divided into MAIN-IN if sites are linked from IN, MAIN-OUT if sites link to OUT. MAIN-MAIN is the intersection of both sets and MAIN-NORM is the rest of set MAIN.
- IN: is the set of all sites not in MAIN that link some site in MAIN.
- OUT: is the set of all sites not in MAIN being linked from MAIN.
- TUNNEL: is the set of all sites linked from IN and that link to OUT.
- TENTACLES: sites linked from IN or that link to OUT, but not both. Tentacles-IN is the set with sites that follow the first condition while Tentacles-Out is the set of sites following the second condition.
- ISLANDS: is the set of sites that neither are linked from IN nor link to OUT.

This characterization of web sites is important since it can give a measure of how important and old a site is. The most important part of a web is in set MAIN, it is the $core$ of the web. Sites in IN are most likely new sites linking to big and important sites. Sites in OUT are typically old sites without maintenance. It is clear that sites in ISLANDS are probably not important because are separated from the rest of the graph.

Continuing with the graph structure, the $degree$ of a node is calculated by the number of links this node has. The *in-degree* is the number of incident links (reach the node) and the *out-degree* is the number of links going out the node. The first value is usually impossible to calculate, because it can not be established how many people link some site. The second value is easy to obtain given a parsing of the HTML code.

The *granularity* of all this analysis can be selected between sites and pages. The structure of links from web pages form a graph where the pages are nodes and there is a link between two pages if there is a hyperlink in the HTML code. Analogously, the structure of web sites can also be a graph if nodes are considered as sites and there is a link between two sites if there is a page in one site linking a page on other site. In this paper both approaches were taken, but it will be clear which one is used in each descriptor.

The number of incoming links and outgoing links in a web collection typically follows a *power law*. In this case the probability that a page has value x for some of the outlined characteristics is directly proportional to:

$$\frac{1}{x^k}$$

Where k is the parameter of the distribution.

2.2 Composition

A web collection can be also analyzed according the content of their documents. This approach follows a typical information retrieval style where documents are represented in a multidimensional space. Thus,

every document is seen as a vector, where entries correspond to terms. The most frequently used model is the TF-IDF.

TF stands for the *term frequency* and it is simply the number of times term t occurs in document d. It can be scaled in a variety of ways to normalize document length and to obtain $TF(d,t)$.

IDF is the *inverse document frequency*. Not all axes are equally important. Coordinates corresponding to function words like a, an and the will be large and noisy irrespective to the content of the document. IDF tries to scale down the coordinates of terms that occur in many documents. There are also many ways to calculate $IDF(t)$ (Chakrabarti, 2003).

Finally, the model joints each of the last computations in some way, a classic formula is:

$$d_t = TF(d,t)IDF(t)$$

Where d_t is the coordinate of document d in dimension t.

Nevertheless, there is a couple of important preprocesses made in the composition analysis. The first one is the elimination of common words that add little to the semantic of the document. They are called *stopwords* and many prepositions fall in this set. After stopword removal a typical action used is *stemming*, which consist in transforming every word to its lexical root or *stem*. For example, all following words {`apply, applying, applies`} will fall in the stem `appl`.

3 THE COSTA RICAN WEB

Costa Rica is one of the seven countries that form Central America. It has a population around 4.5 million people and according to UNCTAD it is the second country in Latin America where most people use Internet (Chile is the first). Given that, is sounds interesting to study and characterize the web in Costa Rica. This work was inspired by a study made by our Chilean colleagues (Baeza-Yates, 2003).

For measuring and extracting all the required information about the Costa Rican web, the Wire set of tools (WIRE) was used. WIRE is an open source crawler, repository and analyzer focused on scheduling techniques for crawling (Castillo, 2004). Wire was run on a Pentium 4 machine with 256 Mb RAM and 256 Gb on hard disc. The total crawling took approximately 4 days.

In this section, the results Wire found for the documents and sites will be presented . It will include characteristics as size, age, structure and languages. The search space was delimited to the Costa Rican web, defined by all documents within the `.cr` domain (other definitions of Costa Rican web are also possible). For extracting all site names ending in `.cr` several sources were crawled: Netcraft (NETCRAFT) which is a site with lots of information about web servers around the net, ODP (ODP) the Open Directory Project which lists many *interesting* pages and Google search engine. All this should be done given that the local NIC (CRNIC) was unable to publish the complete site list due to security reasons.

3.1 Documents

The first thing to note is that the total crawled pages were near seven hundred thousand, which is smaller if it is compared with the 3 million pages Chilean web (Castillo, 2004). Table 1 shows a summary of pages in the Costa Rican web. It can be seen that there are few duplicates among pages and that Costa Rican web is mostly static.

Table 1: Summary of Costa Rican web pages.

Total pages	**731,171**	**100**%
Unique	711,504	97.31%
Duplicates	19,667	2.69%
Static	542,247	74.16%
Dynamic	188,924	25.84%

Also, it is a web composed by mostly new documents. Almost one of every five pages has less than 12 months, but this stands for half the pages whose age was known (some web server doesn't return the age of document). Figure 2 shows the distribution of age for pages.

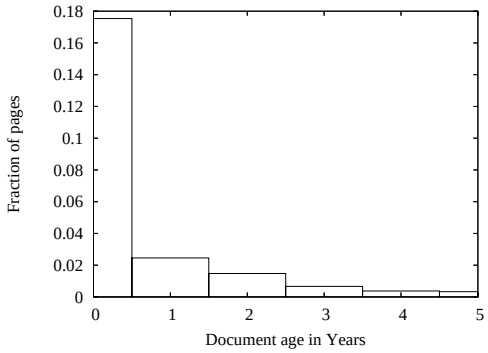

Figure 2: Age of pages.

These results are consistent with the ones found in (Castillo, 2004) from the Chilean web. Another expected result was the number of incoming links, as shown in figure 3.

The number of incoming links is calculated within the collection, counting how many hyperlinks point to

some page. This measure is important as a means for authority. However, as it was mentioned in section 2 it follows a power law, meaning that the probability for a page to have an incoming degree equal to x is proportional to $\frac{1}{x^{1.66}}$. Figure 3 presents this result. The value `1.66` is exactly the same to its analogous value for Chilean web. As this value gets smaller, it is easier to find a page with higher incoming degree. The less this parameter, the more biased the distribution is.

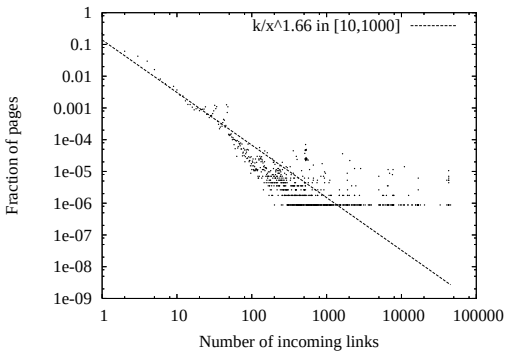

Figure 3: Incoming links.

On the other side, the PageRank measure (Page, 1998) discriminates pages according to their authority. It is a measure of how important a page is. Figure 4 shows the power distribution for pages according to their PageRank. In this case the parameter is a standard one, `1.85`. This result is a little bit biased, but represents the reality in the web: few pages are very authoritative. Another measure for authority in the web is the one called HITS (Kleinberg, 1999). This algorithm classify pages into two groups: hubs which references other important pages and authorities which are referenced by many other pages. However, HITS creates a fuzzy classification of pages between these two categories. Thus, every page has its hub score and its authority score. The value for the power distribution of the hub score is `1.87`, while the one for the authority score is `1.85`. These values of parameters were also expected.

Besides all those measures for structure, a simple but interesting counting measure can determine which technology for cgi pages is dominant. Figure 5 presents a chart with all these technologies according to their percent of presence in the Costa Rican web. As it can be seen PHP dominates the construction for dynamic pages, followed by ASP and then for SHTML.

Another statistic regarding the language of the documents is that English is dominant over Spanish in the web pages. In a sample taken by WIRE, `72.58%` of web pages were in English, while only `26.51%` were in Spanish.

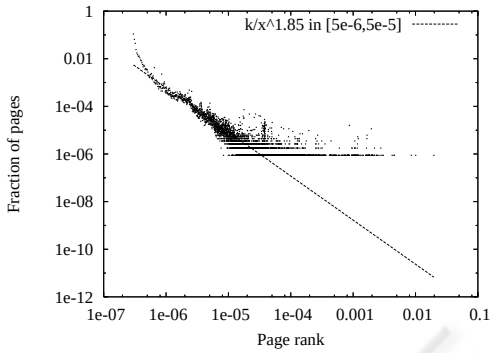

Figure 4: PageRank.

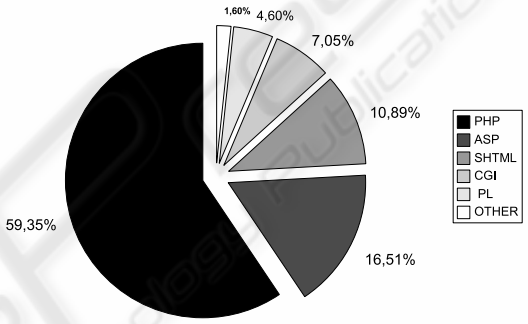

Figure 5: Extensions cgi.

3.2 Sites

In the Costa Rican web every domain must be collocated into one of the following categories: `.co` or commercial, `.or` or organizational, `.fi` of finance, `.ac` or academic, `.go` or governmental, `.ed` or educational and `.sa` or health. The first level domain is restricted, few sites lay into such privileged domain.

According to Costa Rica NIC (CRNIC) there are around 5000 sites in all categories. The initial list for crawling contain less than 3000 sites. However it doesn't seem to be a problem, given that half the sites are working and from the ones working half have more than one page (Castillo, 2004).

As it was reviewed in section 2 the web can be thought as a directed graph. Graphical visualizations are always good to see. Figure 6 shows the graph for all sites present in the Costa Rican web crawled by Wire. Every point stands for a site and the most referenced sites are located in the center of the figure. Watching this representation it is easy to get an idea of how intricated the web structure is.

A more clear representation is found in figure 7

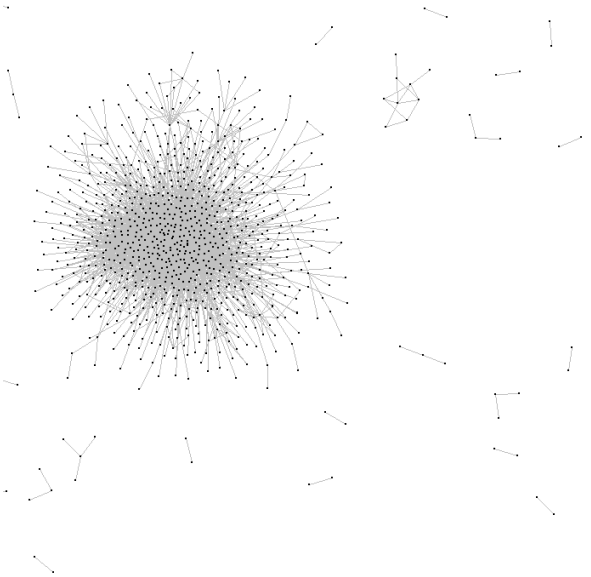

Figure 6: All sites graph of Costa Rican web.

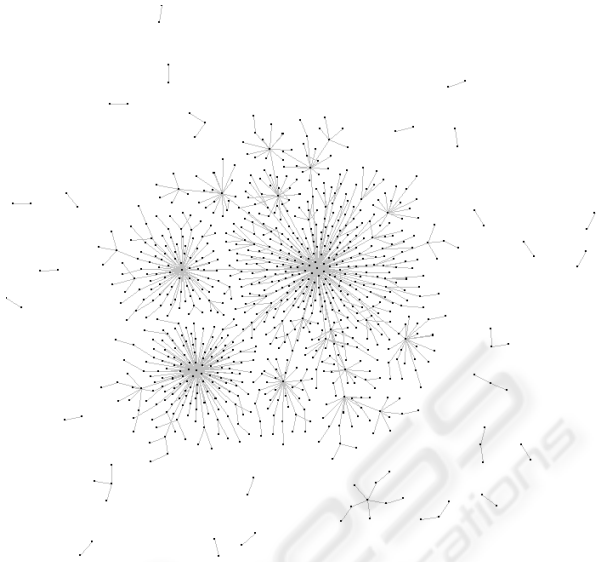

Figure 7: All sites minimum spanning tree of Costa Rican web.

where the minimum spanning tree has been drawn. The center for any *star* correspond to a very important site which is linked to many others. This can be because its authority or hub condition. For example, the center of the big *star* correspond to the University of Costa Rica (which will be mentioned below). These two figures were generated using LGL software (LGL).

Besides the structure of web, there are more tacit things to observe. A few basic statistics are shown in table 2. The total number of sites "alive" is `2,152`. The average number of pages per site is more than five hundred, which means sites are big, but not so complex given that the depth average is less than four levels. Also, average site is not big in size, given that less than 5MB are enough to contain it all.

Table 2: Site summary.

Number of sites ok	**2,152**
Average internal links	3,957.65
Number of sites with valid page age	1,682
Average pages per site	514.96
Average static pages per site	325.85
Average dynamic pages per site	189.11
Average of age of oldest page in months	18.91
Average of age of average page in months	12.83
Average of age of newest page in months	10.12
Average site size in MB	5.53
Average site max depth	3.12
Average in-degree	2.02
Average out-degree	2.02

Table 3: Top 5 sites according to number of documents.

Top sites by count_doc	**count_doc**
www.fcj.or.cr	100,000
www.ecyt.ac.cr	100,000
ftp.ucr.ac.cr	100,000
www.emate.ucr.ac.cr	88,313
www.ulasalle.ac.cr	76,031

Following those statistics some questions arise almost immediately, like *what is the ranking of sites?*. It depends on what it is considered. Tables 3, 4 and 5 presents three rankings according to number of documents, incoming degree and outgoing degree, respectively.

The first three places of table 3 have 100,000 documents because this was the maximum documents crawled by site. The first two places in this table are sites with an online calendar technology for showing events for every day of year. This software creates an html for every day and then the number of web pages counted are huge, reaching easily the limit of web pages crawled for a site. The third site is an immense site with thousands of documents, located into the University of Costa Rica. The fourth place is also located on the University of Costa Rica and contains big online manuals for doing math typesetting.

In table 4 the first three places are national universities of Costa Rica, which tend to be authoritative enough according to (Castillo, 2004). The very first place correspond to the University of Costa Rica,

Table 4: Top 5 sites according to in degree.

Top sites by in_degree	in_degree
www.ucr.ac.cr	138
www.una.ac.cr	84
www.itcr.ac.cr	62
cariari.ucr.ac.cr	55
www.inbio.ac.cr	53

Table 5: Top 5 sites according to out degree.

Top sites by out_degree	out_degree
www.infoweb.co.cr	165
www.directorio.co.cr	156
sibdi.bldt.ucr.ac.cr	126
www.asamblea.go.cr	117
www.racsa.co.cr	102

which is the oldest and biggest university in Costa Rica, formed by many faculties and research centers. The second and third places are also public universities, which group several departments. The fifth place is the National Biodiversity Institute, which is very important for Costa Rica, a country in the top 20 countries with more biodiversity in the world.

In the other hand, web directories, like the first two places in table 5 act as hubs in the web. The third place in this table represents an instance of the government of Costa Rica that link many other departments. The fifth place correspond to RACSA, which centrally administers the internet service in Costa Rica.

The same as PageRank is calculated for documents, SiteRank can be calculated for sites. Figure 8 shows the distribution for sites according to their SiteRank. The parameter for the power law is `1.81` which again represents an expected parameter.

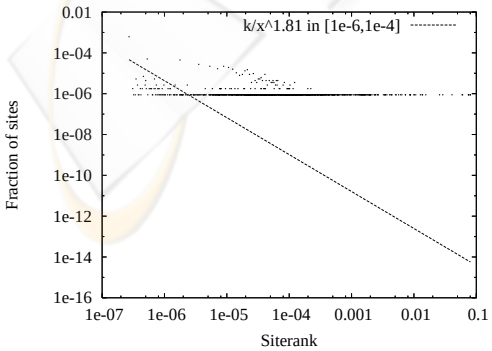

Figure 8: SiteRank.

Recalling what was discussed in section 2 regarding location of sites in the graph, table 6 presents some measures for each component. It can be seen that the principal component has `336 sites`, which represents almost the 15% of total sites alive. Also, it is interesting to note that the IN component has few sites, but the OUT component has many sites for a suspicious big old web section. Also, the $ISLAND$ size is huge, meaning that a big portion of the web is disconnected from the *core*.

Table 6: Graph component sizes.

Component name	Number of sites	Percent
MAIN_NORM	107	4.97%
MAIN_MAIN	84	3.90%
MAIN_IN	36	1.67%
MAIN_OUT	109	5.07%
IN	100	4.65%
OUT	388	18.03%
TIN	22	1.02%
TOUT	28	1.30%
ISLAND	1,278	59.39%

Last but not least, it can be analyzed which are the domains that are more commonly referenced by sites in the Costa Rican web. Figure 9 shows the top domains referenced from sites in Costa Rica. The `.cr` domain is the most referenced because is the same domain. Domain `.com` is the second one and Argentina (`.ar`) domain is the third one, followed by Brazil (`.br`).

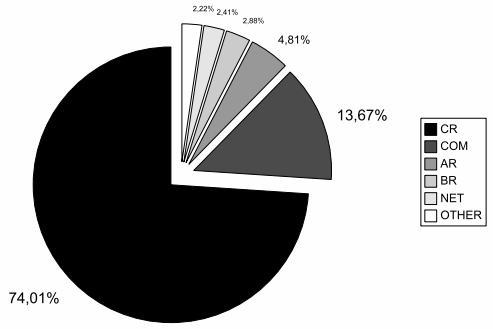

Figure 9: Top level domains.

4 WEB CLUSTERING

This section discusses the alternatives for clustering web objects and presents a method for clustering sites

using the classic vectorial representation. Some visualizations of results are also showed.

4.1 Approaches

The utility of clustering for text and hypertext information retrieval relies on the *cluster hypothesis*, which states that *given a suitable clustering of a collection, if the user is interested in document d, he is likely to be interested in other members of the cluster to which d belongs to* (Chakrabarti, 2003). There are many alternatives for clustering web objects: pages or sites (Chakrabarti, 2003; Jain, 1999).

The Hierarchical Agglomerative Clustering (HAC) starts with all the documents and successively combine them into groups within which interdocument similarity is high, collapsing down as many groups as desired. This kind of clustering is called *bottom-up* or *agglomerative*.

On the other side, a very well known *top-down* method is *k*-Means, where the user preset a number k of desired clusters. The initial configuration is arbitrary, consisting of a grouping of the documents into k groups and k centroids computed accordingly. The basic step is moving to the nearest centroid for every document. Then centroids are again calculated and the algorithm follows until no change is made on the assignment of document to clusters.

Although HAC and *k*-Means are very simple and fast, many other algorithms have been presented: self organized maps, multidimensional scaling, latent semantic indexing, probabilistic approaches, through summarization (Shen, 2004), hyperlink-based (He, 2001), partitioning-based (Boley, 1999) and many more.

However, all these techniques were developed under the scalar vector space model. Here, documents are represented as vector in multidimensional Euclidean space. Each axis in this space corresponds to a term (token). Such representation makes possible the TF-IDF model.

Apart from these techniques based on composition, there are other techniques for web clustering based on structure, which applies graph theory for building the clusters (He, 2001).

4.2 Web Site Clustering

This section provides the results for clustering web sites using a standard vectorial representation, as the one proposed in section 2 for TF-IDF model. We selected 16 sites from the top 20 sites according to their SiteRank. This set contains scientific sites, business sites, universities sites and one sport site. For each site, 10 random web pages were crawled and from these 10 pages a vector is obtained for representing each site.

Using the 160 crawled pages, the 200 most common tokens were identified, after stopword elimination and stemming application. Then, a normalized 200 long vector is associated for each site. This normalization doesn't take into account the *IDF* part of the *TF-IDF* model. With this representation we applied the HAC procedure, discussed in the previous section.

Nevertheless, for visualization purposes, a PCA (Principal Component Analysis, a dimension reduction technique) was applied to our web site database. Figure 10 shows the result after a discrete PCA (Buntine, 2004) is applied. This 2D graphic shows how web sites are distributed and it can be seen that on the upper right corner are located the universities and scientific sites, while the business sies are on the bottom right corner.

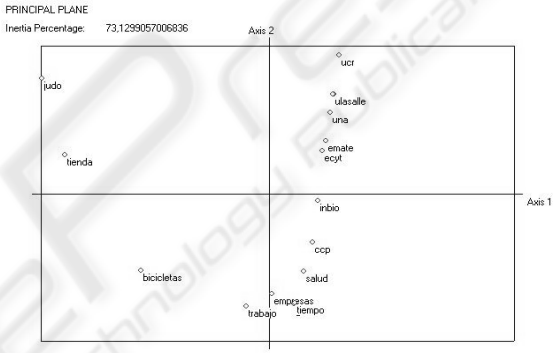

Figure 10: Principal component analysis for web sites.

Figure 11 presents the *dendogram* resultant from the HAC procedure. Again, the scientific sites are located in the bottom branch on the tree, while the business sites are located in the middle branch.

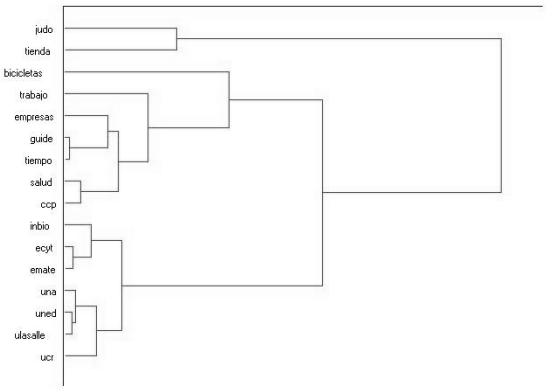

Figure 11: Hierarchical agglomerative web site clustering.

Both figures were generated using Effimining product of Predisoft (PREDISOFT).

5 CONCLUSION AND FUTURE WORK

This work is part of the effort done in the *Klá* research project, which is under development in the Computing Research Center at the Costa Rica Institute of Technology. Its website is *www.kla.ic-itcr.ac.cr*.

Klá is focused on characterize the Costa Rican web, based on its structure, composition and clustering. All this knowledge can be applied to develop several tools for improving the web experience. A new search engine has been developed under the `MS .net` technologies. It will be available soon at *www.kla.ac.cr*.

In this paper it has been showed that Costa Rican web has around seven hundred thousand documents. Most of them are static and relatively new. Almost 82% of documents are html text. Distributions of incoming links, outgoing links, pagerank, hubrank and authorityrank follow expected power laws.

It was also presented a clustering of 16 sites of Costa Rican web according to their popularity. This technique builds a classic vector for each site and then applies a HAC or *k*-means for clustering. Currently, we are working on a symbolic representation for web sites, using histograms (Rodríguez-Rojas, 2000) for measuring the frequency of terms in different axis of a web document: text, title, links and bold. Symbolic representations of web objects are not new, Schenker et al have been applied a graph representation for clustering web documents (Bunke, 2003).

ACKNOWLEDGEMENTS

The author would like to thank several people that help in the construction of this paper: Dr. Oldemar Rodríguez-Rojas for his guidance in the very first experiments on symbolic web clustering, Dr. Carlos Castillo for his help using WIRE system and Dr. Francisco J. Torres-Rojas for his support in the establishment of the *Klá* research project.

REFERENCES

Baeza-Yates R., Poblete B. and Saint-Jean, F. (2003). Evolution of the Chilean Web: 2001-2002 (In Spanish). In *Proceedings of the Jornadas Chilenas de Computación*. Chillán, Chile, November 2003.

Boley D., Gini M., Gross R., Han S., Hastings K., Karypis G., Kumar V., Moore J. and Mobasher B. (1999). Partitioning-Based Clustering for Web Document Categorization. In *Decision Support Systems*. 1999.

Bunke H., Last M., Schenker A. and Kandel A.(2003). A comparison of two novel algorithms for clustering web documents. In *Proceedings of the 2nd International Workshop on Web Document Analysis (WDA)*. 2003.

Buntine W., Perttu, S. and Tuulos V. (2004). Using Discrete PCA on Web Pages. In *ECML/PKDD 2004, Proceedings of the Workshop on Statistical Approaches for Web Mining*. Pisa, Italy, September, 2004.

Castillo C. (2004). Effective Web Crawling. PhD Thesis, University of Chile, 2004.

Chakrabarti, S. (2003). Mining the Web. Morgan Kaufmann Publishers, 2003.

COSTA RICA NIC: Domain name services for Costa Rica. http://www.nic.cr.

He X., Zha H., Ding C. and Simon H. (2001). Web Document Clustering Using Hyperlink Structures. Technical Report CSE-01-006, Dept. of Computer Science and Engineering, Pennsylvania State University, 2001.

Jain A.K., Murty M.N. and Flynn P.J. (1999). Data Clustering: A Review In *ACM Computing Surveys*. September, 1999.

Kleinberg J. (1999). Authoritative sources in a hyperlinked environment. In *Journal of the ACM*. 46(5):604632, 1999.

LGL: Large Graph Layout http://apropos.icmb.utexas.edu/lgl/

NETCRAFT. http://news.netcraft.com/.

ODP: Open Directory Project. http://dmoz.org.

Page L., Brin S., Motwani R. and Winograd T. (1998). The PageRank citation ranking: bringing order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

Predisoft: Scientific Prediction. http://www.predisoft.com.

Rodríguez-Rojas O. (2000). Classification and Linear Models in Symbolic Data Analysis. PhD Thesis, University of Paris IX-Dauphine, 2000.

Shen D., Chen Z., Yang Q., Zeng H., Zhang B., Lu Y. and Ma W. (2004). Web-page Classification through Summarization. In *SIGIR 2004, Proceedings of the ACM Conference on Research & Development on Information Retrieval*. South Yorkshire, United Kingdom, July, 2004.

WIRE: Web Information Retrieval Environment. http://www.cwr.cl/projects/WIRE/.