Searching relationships between enterprise websites using graph based web crawling

R. C. F. De Souza, G. M. Caputo & N. F. F. Ebecken
COPPE – Federal University of Rio de Janeiro, Brazil

Abstract

The objective of this paper is to find explicit web relationships using enterprise websites as seeds. We apply a web crawler to find these relationships in a hierarchy starting from the given seed using the external links to construct a Jaccard Score weighted tree. The proposed methodology aims to search related enterprises from the root node based on the link, which are potential partners, suppliers, clients, etc. We crawl the whole site to find external links using the Breadth First Search (BSF) algorithm and build a tree structure containing just the interesting external links. The applied algorithms were programmed with very simple computational components and may produce interesting results to analyze the domain of sites, their structure, and how they link with each other in their acting range.

Keywords: link analysis, BSF algorithm, web crawling.

1 Introduction

Complex behavior can be observed in many systems. In this paper we intend to focus on complex networks.

These networks can present different topologies and follow well-established phenomena, and independent of their nature, can present similarity between some metrics, such as the medium, node degree distribution, clustering coefficient, etc.

Many times, the net can present complex structures that are impossible to take conclusions by simple means (like mere observations on the topology). It is also known that each single network presents a unique structure. For this reason, it can be necessary to use generic models that represent, for instance, not an edge between nodes, but the probability of occurrence of this edge. To these models, it is given the name of random graphs [4].

WIT Transactions on Information and Communication Technologies, Vol 42, © 2009 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA090071

There are a number of behaviors that infer on real life networks. Among them, we can highlight three common phenomena [5]:

- Small world – this phenomenon tells us that the average smallest path of the network is low, taking in account the large number of nodes on real life graphs.
- Friends of a friend – this characteristic deals with the high number of triangles on the graph, measured by the clustering coefficient. It means that it is not rare that redundancy of connections occur.
- Power law – this phenomenon tells us that the node degree distribution of the network follows a power law. It usually means that a large number of nodes in the graph are connected to only a few nodes, but the network can contain a small (but not disposable) number of nodes connected to a lot of other ones.

2 Methodology

The aim of this paper is to present a methodology to search potential partners of companies based on their websites. We used web crawling and BFS technologies, detailed later on.

Various external links also have outlinks not connected to the original website. These ones are the most relevant to the present work. Besides, it is also necessary to find internal links, so we can discover outlinks contained it their HTML source code, which are potential partners.

In a first moment, considering just the initial seed for the crawl, it is expected to find the outlink list that the seed refer. Extending this reasoning for neighborhood external links (pointed by the initial site) it is expected to construct a larger and more complex graph. This second hierarchy has relationships between sites from different companies that have something in common and some triangulation level.

2.1 Web crawling

The implemented web crawler (TBWC – thread-based web crawler) executes the search, initialized by a determined seed or web page. It searches and stores outlinks found in the HTML body. For this reason, it was considered some regular expressions that describe the link between websites in the found HTML source.

The implemented tool was adapted for some behaviors occurred in the website construction, like masks, links written inside JavaScript and password fields. Other characteristics were not implemented because they do not appear in the study case.

2.2 BFS

The Breadth First Search (BSF) Algorithm was chosen because of some interesting characteristics. It`s crawl based on queue means that external sites first detected are crawled first. The search in outlinks crawl one site each time and add the found external links in the end of the queue.

It is very useful for the proposed methodology that the search is executed around the seed. This means that it is desirable to crawl all the pages pointed from the seed, but it shouldn`t start to crawl another websites until the seed is completely crawled. Exploratory techniques are not interesting in this application, so depth-based algorithms were not considered for implementation.

2.3 Jaccard Score

There are many similarity measures based on links for websites. An important and very used one is the Jaccard Score [6] that can be applied to any two pages. The score is given by equation 1.

$$Js(A,B) = \frac{La \bigcap Lb}{La \bigcup Lb} \tag{1}$$

where A and B are two distinct websites and La is the outlink set of A and Lb is the outlink set of B.

The Jaccard Score gives a ratio of similarity of websites based on their outlinks. This means that two websites can be widely different in design and even in content, but the measure only considers the links each one points to give their similarity based on the idea that if both websites point to the same outlinks, they have something in common.

3 Implemented tool

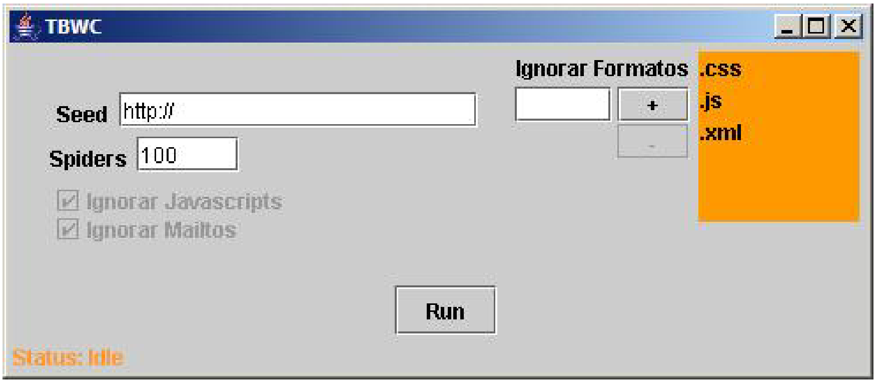
The implemented tool was designed to be easy to use. Actually, with the large range of free software available nowadays, it is also quite simple to program a tool like the one used on this paper.

The tool is based on multithreading technology. That means various instances (spiders) can crawl different websites at the same time. To do so, we used the Java technology, which, the language itself and some of the best IDEs available, are free. In Java it is very simple to design thread-safe applications and avoid traps, since to do a single thread class all you have to do is extend the correct superclass and the communication between the instances are quite simple.

The spider implements a BFS-like algorithm, what means it maintains a public and static row of sites to be crawled. Using this row, the methods programmed to connect the next websites are called. To establish the connection is also easy with the URL and URLConnection classes, well detailed on the Java 2 EE 1.4.2 documentation. Using a simple input stream, the website HTML is retrieved, parsed and its links are used to feed the BFS row and to calculate the Jaccard Score.

Figure 1 shows the main interface of the developed tool.

The implemented tool has also some options to avoid undesirable outlinks, since the process of discover a new link is based on HTML tags, which can



TBWC

Seed http://

Spiders 100

☑ Ignorar Javascripts

☑ Ignorar Mailtos

Ignorar Formatos

.css
.js
.xml

Figure 1: Main interface.

contain not only links to pages, but also *mailtos*, JavaScript, etc. We also found it useful to do a simple way to skip determinate file formats such as images, CSS files, etc as shown in figure 1.

3.1 Visualization

The developed web crawler exports its results in a formatted file to be read under Graphviz [3]. To do so, we print a text file indicating links and outlinks.

Graphviz is a free software capable to draw graphs in a variety of manners and algorithms, exporting the results to a common JPEG image file. The software is easy to use and its documentation and related examples makes it simple to understand how to format the input files correctly.

4 Case study

The case study aims to present the methodology given an URL seed. To do so, we adopted the Scientific Foundation Site (http://www.coppetec.coppe.ufrj.br) of Federal University of Rio de Janeiro. This foundation manages the development and project resources. This website was chosen because its facility to read its HTML, i.e., the linked sites are available and organized in the code and are easy to recover.

From this site, we execute the crawling to find adjacent websites based on the links exposed on their HTML. The stopping criterion for the crawling algorithm was not to overpass the second hierarchy from the seed. It was obtained a subset of the connected component created by outlinks from http://www.coppetec.coppe.ufrj.br

At the end of the crawl and after a brief selection and data cleaning of the data, it was obtained 546 records, in which 81 were from the first hierarchy, i.e., contain inlinks directly from the seed. Figures 2 and 3 show the links of the first and second hierarchies, respectively.

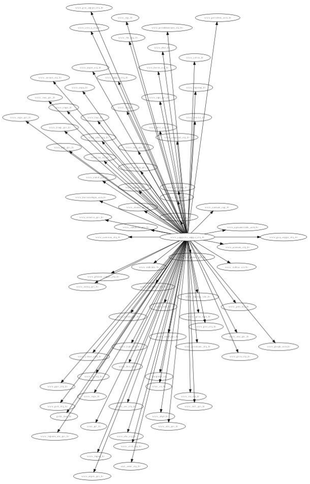


Figure 2: Outlinks from COPPETEC.

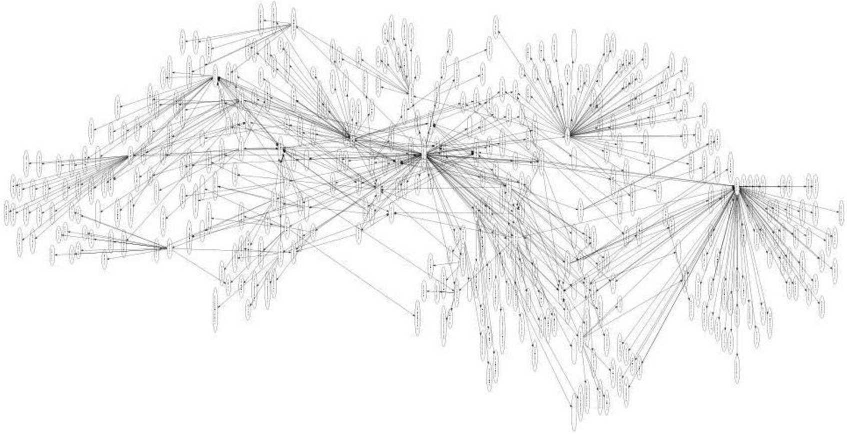


Figure 3: The second hierarchy from the seed.

The Jaccard Score was applied in both sites of COPPETEC (first hierarchy) and its outlinks (second hierarchy). Table 1 shows the top 10 sites that have the highest Jaccard Score. It can be seen that the most similar sites are from the second hierarchy and COPPETEC site just appears in the 10th position. The first one, with 25% of links similarities are the CAPES and MEC sites. These sites are respectively the Graduate Programs Founding and Evaluation Agency and Educational Ministry. The first one subsidizes the second in post-graduation politics formulation. This high relationship shows the similarity between both sites.

Observing the data shown in Table 1, one thing that is clearly noticed is that the MCT (Brazilian Ministry of Science and Technology) website appears in almost all the rows. By analyzing the inlinks it is possible to see how its prestige degree [1] is much higher than the others, in other words, we can conclude that the MCT website acts as an authority [2] in this piece of data, which means a high level of confidence in who is pointing to MCT and the website content. MCT is referred by INPA (Amazonas Research National Institute), EMBRAPA (Brazilian Ministry of Agriculture, Livestock and Supply), ON (national observatory), FINEP (Study and Projects Funding), IBICT (Brazilian Institute of Science and Technology Information), CETEM (Mineral Technology Center).

Table 1: Overall top 10 highest Jaccard Scores between the visited sites.

Site	Outlink	Jaccard
www_capes_gov_br	www_mec_gov_br	0.25
www_inpa_gov_br	www_mct_gov_br	0.125
www_embrapa_br	www_mct_gov_br	0.111111
www_on_br	www_mct_gov_br	0.105263
www_finep_gov_br	www_mct_gov_br	0.095238
www_mct_gov_br	www_finep_gov_br	0.095238
www_ibict_br	www_mct_gov_br	0.093023
www_ibict_br	www_finep_gov_br	0.085714
www_cetem_gov_br	www_mct_gov_br	0.08
www_coppetec_coppe_ufrj_br	www_metalmat_ufrj_br	0.068966

Table 2: Top 10 highest scores from the seed.

	Seed	Outlink	Jaccard
1	www_coppetec_coppe_ufrj_br	www_metalmat_ufrj_br	0.068966
2	www_coppetec_coppe_ufrj_br	www_peq_coppe_ufrj_br	0.055556
3	www_coppetec_coppe_ufrj_br	www_cos_ufrj_br	0.047619
4	www_coppetec_coppe_ufrj_br	www_biorio_org_br	0.044444
5	www_coppetec_coppe_ufrj_br	www_ibict_br	0.036697
6	www_coppetec_coppe_ufrj_br	www_planeta_coppe_ufrj_br	0.035294
7	www_coppetec_coppe_ufrj_br	www_producao_ufrj_br	0.034483
8	www_coppetec_coppe_ufrj_br	www_oceanica_ufrj_br	0.034091
9	www_coppetec_coppe_ufrj_br	www_inpe_br	0.030303
10	www_coppetec_coppe_ufrj_br	www_pee_ufrj_br	0.02439

Table 2 contains the highest similarities of links between the given seed and its outlinks. We observed that there are no expressive similarities in these sites based on their link distance, since the highest Jaccard Score obtained is not greater than 7%.

However, analyzing the institutional nature of the websites showed in table 2, it is possible to conclude that from the top ten most similar websites from the seed, seven correspond to institutions (labs and programs) directly maintained and/or funded by COPPETEC Foundation. These sites include several departments from Federal University of Rio de Janeiro: 1 – METALMAT (Metallurgical Engineering and Materials), 2 – PEQ (Chemical Engineering), 3 – COS (Engineering and Computer Systems), 6 – BIORIO (Biotechnology Pole in Rio de Janeiro), 7 – PRODUCAO UFRJ (Industrial Engineering), 8 – OCEANICA (Ocean and Naval Engineering) e 10 – PEE (Electrical Engineering).

The presented data on table 3 shows the other links from the seed counting from the tenth position, sorted decreasingly by each Jaccard Score**.**

From table 3 we can highlight some important implication:

- Among the 16 illustrated items in the table, five are funding organizations, which is the acting area of COPPETEC Foundation: 16 – CNPQ (), 19 – CAPES (National Counsel of Technological and Scientific Development), 20 – FAPEAM (The Research Foundation for the State of Amazonas Support), 22 – FAPERGS (Foundation for the Research Support of the Rio Grande do Sul State) e 24 – FINEP (Studies and Projects Funding).
- We can also notice, by the data shown on tables 1 and 3, the strong relationship between COPPETEC Foundation and MCT (Brazilian Ministry of Science and Technology). Besides, there is only one direct link between these two websites, there are a lot of paths between both of them by organizations directly linked to MTC: 11 – ON , 12 – EMBRAPA, 13 – MCT, 17 – INPA, 24 – Science Institute, 26 - CETEM).

Table 3: The rest of the non-zero relationships in the first hierarchy.

	Seed	**Outlink**	**Jaccard**
11	www_coppetec_coppe_ufrj_br	www_on_br	0.023529
12	www_coppetec_coppe_ufrj_br	www_embrapa_br	0.021277
13	www_coppetec_coppe_ufrj_br	www_mct_gov_br	0.020833
14	www_coppetec_coppe_ufrj_br	www_ufrj_br	0.02027
15	www_coppetec_coppe_ufrj_br	wikipedia_org	0.014815
16	www_coppetec_coppe_ufrj_br	www_cnpq_br	0.012048
17	www_coppetec_coppe_ufrj_br	www_inpa_gov_br	0.012048
18	www_coppetec_coppe_ufrj_br	www_abc_org_br	0.011905
19	www_coppetec_coppe_ufrj_br	www_capes_gov_br	0.011905
20	www_coppetec_coppe_ufrj_br	www_fapeam_am_gov_br	0.011905
21	www_coppetec_coppe_ufrj_br	www_impa_br	0.011905
22	www_coppetec_coppe_ufrj_br	www_fapergs_tche_br	0.011765
23	www_coppetec_coppe_ufrj_br	www_ciencia_org_br	0.011628
24	www_coppetec_coppe_ufrj_br	www_finep_gov_br	0.011364
25	www_coppetec_coppe_ufrj_br	www_agenciaestado_com_br	0.011111
26	www_coppetec_coppe_ufrj_br	www_cetem_gov_br	0.01087

Analyzing the data generated by the applied methods it can be concluded that it is possible to understand relationships from the link structure of websites. It is also possible to establish which are the institutions directly and indirectly linked to the analyzed sites, where the similarities between sites can be highlighted based on their outlinks.

5 Conclusions

The present paper deals with a site analysis methodology based on their outlinks. Through its similarity it is possible to establish which are the sites that have more common interests weighted by a well defined metric.

It is also shown how simple it is nowadays, using powerful, scalable and free software such as Java, to create a web crawling tool. We also remember that all the software used in this paper is not computationally heavy. For instance, the implemented tool and Graphviz run on a 1.2GHz Pentium 3 with a 512MB RAM.

Actually such implementations (as the web crawler) are commonly known as light applications, but usually network-bound because of lots of connections being established and page information being retrieved.

The methodology was applied in a preliminary study case obtaining satisfactory results. However, it can be applied to other cases of general interest. That means, in sites where it is aimed to find relationships based on the similarities of their links to automatically understand the business behind a website, find potential partners and suppliers, etc.

For further work, it is aimed to add text mining techniques to obtain more accurate similarities, combining subjects amongst a range of pages and the used methodology. It is also intended to use the presented methodology to compare independent website structures based on their outlinks.

Acknowledgements

We are grateful to the Brazilian Research Agencies CNPq, FAPERJ and ANP for their financial support.

References

[1] S. Wasserman and K. Raust. Social Network Analysis. Cambridge University Press, 1994.

[2] J. Kleinberg. Authoritative Sources in a Hyperlinked Environment. In *Proc. of the 9th ACM SIAM Symposium on Discrete Algorithms* (*SODA'98*), pp. 668–677, 1998. S.

[3] Graphviz website <http:// www.graphviz.org/>

[4] R. Albert, A. Baraba. Statistical mechanics of complex networks. *Reviews of Modern Physics*, volume 74, January 2002.

[5] B. Liu. *Web Data Mining Exploring Hyperlinks, Contents, and Usage Data*. Springer 2007.

[6] L. Kauffman, and P. J. Rousseeuw. *Finding Groups in Data: an introduction to cluster analysis*. John Willey & Sons, 1990.