Finding web communities for business application

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

Abstract

This work presents a study of web communities, discovering links and associations between business web pages. Those links are drawn using graphs where the nodes are mapped, entities representing business subjects, such as products, services, enterprises name, partnerships, and some characteristics evolving those entities. The edges are the relationships between those mapped entities. In comparing the resulting graphs it could be expected to find differentiations between competitors and their strategies. To achieve this objective, communities' discovery techniques and visual analysis are applied and tested using a set of graph representations and different web crawlers' depths. On further analysis, the obtained bipartite graph aims to discover market trends and help competitive intelligence analysts.
Keywords: link mining, web mining, group detection.

1 Introduction

Competitive intelligence tools have been used to obtain information from the internet. They support the search for valuable information that represents gains and can anticipate the rising demand mapping the competitors' business [1].

All the available information on the internet, access free, are mapped through large data bases and used for a range of ways, like data/text mining, and content, usage and structure mining, as shown in [2].

The company mapped information can be compared with other mapped companies. This process converts sites in nodes and their links to other sites in edges creating a graph. This graph can be reduced using just correlated web sites that represent their subjects and have a correlation between the analyzed companies [3].

WIT Transactions on Information and Communication Technologies, Vol 42, © 2009 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA090031

The objective of this work is to develop a methodology representing these web sites through graphs. For this, data obtaining, information extraction techniques and visualizing ways were used and are significantly important to the analysis [4].

Defining business as a community means to say that the related web sites have subjects, objectives and links in common. A web community was defined by [5] as a collection of web pages such that each member page has more hyperlinks within the community than outside of the community.

Kumar defined in [6] that there are two kinds of communities: explicit communities (found in forums, e-mails groups, social network website, e.g. Facebook, and others) and implicit communities that are more difficult to find on the internet because they do not share a specific set of web pages. They search for similar subjects that are interesting to the same group.

The first step of competitive intelligence is to monitor dynamically the concurrency, and identification about who they are is fundamental for the analysis success. Combining implicit communities and CI techniques can show these differentiations between similar businesses and related ones.

Besides this, it is necessary to identify which information resources would be used, and in this case, the resource is the internet.

This paper is organized as follows: the next part briefly reviews the basic concept of web crawlers, information extraction and data visualization and shows how they were applied. The third part shows a case study where the methodology is applied.

2 Methodology

The suggested methodology consists of crawling the web searching external links (out-links) and storing contents. Then the content is analyzed aiming to reduce the set of phrases, and finding just those ones that represent the business.

2.1 Web crawler

A web crawler tool was used to collect every link related to any other site given a set of seeds. The tool collected information, such as: Title, page content, back-link (the page's link that sends to the link) and URL.

The seeds to start the web crawler are the companies' home pages. Then every found link is crawled using the BSF algorithm (breadth-first search).

The stop criterion was the hierarchy of the web crawler. The smaller spider number of hierarchy was tried. It was noted that:

- The first and the second hierarchies do not have enough links to external pages and do not have enough information to find an implicit community.
- The third hierarchy presents external pages in both pages, but the information is too superficial to find the strategy of the companies.
- The fourth hierarchy contains enough external links and content to define an implicit community,

Based on this information, the fourth hierarchy was chosen showing the relationships between the community content.

2.2 Information extraction

The Information Extraction IE [7] was applied on the content of the web pages obtained in the previous step. Information extraction shows the explicitly content of the web page.

Besides this, for commercial web sites, the content is organized to facilitate finding products and services. It means that the major company strategy has to be explicit somewhere in the page, and it was probably collected in the crawling phase. These organizations are usually separated by product and services in the menus and this characteristic can facilitate the user and consequently, the information extraction tools.

The process of collecting the words and phrases similar to those used in text mining applications. Once the terms are collected, the data are organized and the frequency in terms and relevancy are considered in order to choose whether each word will become part of the process. These relevancies can be done by simple metrics, like term frequency or TFxIDF [8].

The obtained data were nouns and phrases representing the company products, services, offers, partners, and some characteristics of all of them [9].

Mainly the content field and title field of each site were used.

2.3 Data visualization

By the obtained information we have two kinds of data and two kinds of relationship to represent: the sites and the relationship between them (showing the hierarchy) and the seeds and the characteristics between them. In this paper, two visualization data were created. This simple structure can be represented as several distinct graphs.

The first visualization shows the hierarchy between web pages. The nodes are sites, and the edges are oriented when a site refers other site.

The second visualization shows the relationship between the main pages and a range of things that they want to sell or offer to their clients. It is represented by a bipartite graph, with a set of actor nodes (sites), a set of dependent nodes (products, and others), and edges that represent an actor's dependency.

Using bipartite graphs means that one site never refers to the concurrent web site.

The graphs were created using the GrafViz, a simple and Open source graph (network) visualization project from AT&T Research. It can generate a range of graphs according to the chosen perspective [10].

3 Case study

The tested web sites are about competitors. Both sites represent a company where their main business is about mobile products and services. They share a common interest and dispute the same market.

To win the concurrency, they adopted different strategies and invested in innovation to their main business. To protect both companies we will call them Company1 and Company2.

Figure 1 shows a bipartite graph created by the intersection of both companies and their links up to the fourth hierarchy. The graph is called bipartite because both sites refer to important other sites in common but they never refer to each other, as a strategic concurrency.

The most cited sites were ANATEL, the "National Agency of Telecommunications". Both sites refer ANATEL that refers every other Mobile company in Brazil.

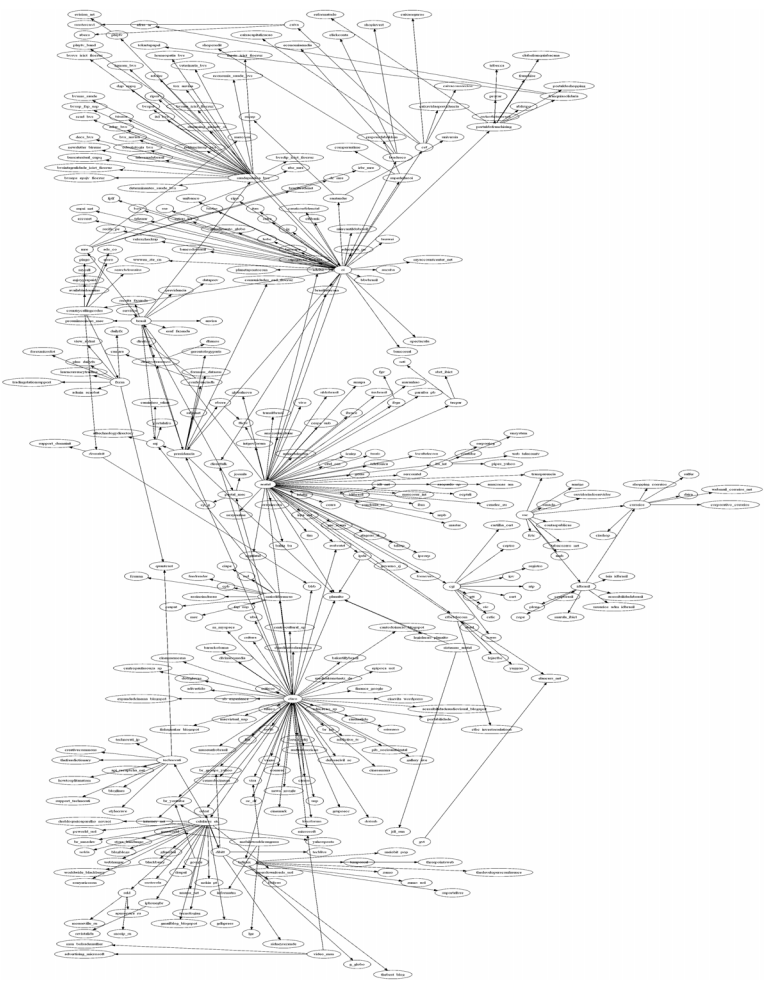

Figure 1: The concurrent web sites and the hierarchy using the BSF algorithm.

Other important highlights are mobile factories and banks that support online payment. Close partners were cited in the text, so it was highlighted in the other generated graph.

The graph shows the common interest between both companies and where they have similar and different strategies related by a range of subjects. As an example, both sites sell mobile cell phones and have a range of plans to offer to their clients. Available download of ring tones, wall paper, and a range of other services, including special offers.

Figure 2 shows the Vendor1 and Vendor2 graphs joined in similar subjects. The data was highlighted to show which subjects are common to both and which subjects they differ in. The dark nodes are the same subjects found in both, and the bright nodes are the subjects that are not of interest to both. Remembering that these subject are obtained by the most important phrases and words acquired by the information extraction.

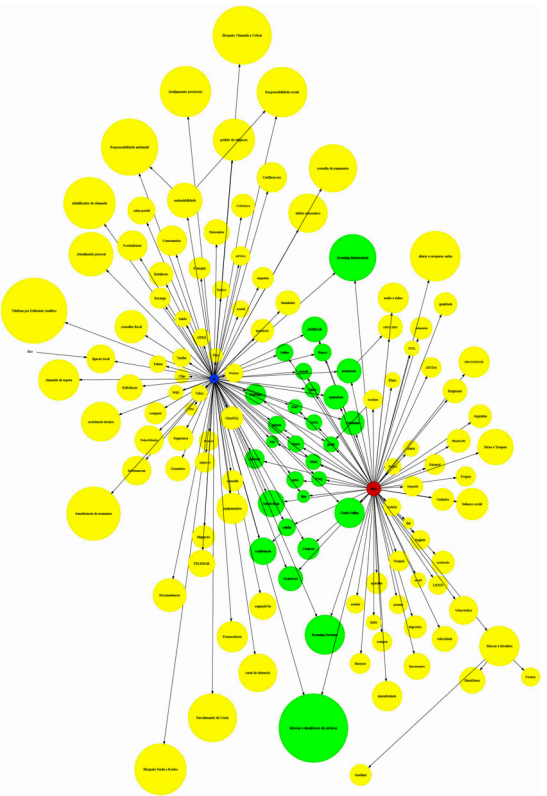

Figure 2: Content graph – information differentiations between the analyzed web pages.

The graph shows that, although they dispute the same market, they have a lot of different strategies, and a close analysis can show the difference between products, services and partners.

Table 1 shows the 29 terms that are common to both vendors. It is very important to keep in mind that some word seams do not have any sense alone, but, analyzed by a specialist and combined with other terms, they then reveal great knowledge.

Table 1: Terms in common for both companies.

operadora	Controle	PrePago
banda larga	Download	Promocao
Roaming Internacional	GGG	Site
Roaming Nacional	gratis	Conta Online
Assinatura	Internet	conferencia
ativacao e desativacao de servicos	Linha	FAQ
Banco	loja	GSM
cartao	movel	Jogos
celular	Online	SAC
Cobertura	planos	

For the first company (Vendor 1), 73 terms were found, where 44 do not appear in the second Vendor list. Table 2 lists all these terms, including partners, characteristics, products and services, that analyzed by a specialist, the combination.

Table 2: Vendor1 exclusive terms.

noticia	modem	Aderente
lancamento	LIGUE	Ajuda
VideoMaker	Kbps	aparelho
Uruguai	interatividade	Argentina
velocidade	impostos	Bluetooth
UOL	Fotos	Cadastro
torpedo	fiat	compra
Toques	carro zero	conteudo
sorteio	premio	email
SMS	conferencia	BlackBerry
Sinopse	alterar e recuperar senha	Foston
Ringtones	audio e vídeo	Gradient
qualidade	balanco social	English
Provedor	Dicas e Truques	Espanol
oferta	Marcas e Modelos	

For the second company (Vendor 2), 90 terms were found where 61 do not appear in the first Vendor list. Table 3 lists all these terms, including partners,

characteristics, products and services, that analyzed by a specialist, the combination.

Although these companies have a lot of implicit similar interest, it's not so easy to find all information in the site and organize it using these tools, or they are not available in the site. Some other information was considered irrelevant on a superficial analysis.

A close analysis shows that while one company offers more online services; the other one is developing another kind of technology in the mobiles. The first company offers internet, online shopping and residential phone. The second one mentions more partners and offers free-car.

If the graph is more opened, we can see the difference between fees and help to decide which company better fits the customer's needs.

Table 3: Vendor 2 exclusive terms.

Saldo	aquisicao	Postos
PulsoMinuto	Certificacoes	Recarga
Conveniados	CheckUp	segundaVia
Dados	Chip	Seguranca
Pedido de religacao	Cobranca	servico
chamada de espera	compare	Simulador
ligacao local	Consertos	Solicitacao
Responsabilidade ambiental	consulta	sugestao
Responsabilidade social	Descontos	sustentabilidade
conferencia	Documentacao	Tarifas
Atendimento pessoal	equipamentos	TELEMAR
caixa postal	Fatura	Velox
canal de denuncia	Fixo	assistencia tecnica
conselho fiscal	FM	bloqueio perda e roubo
consulta de pagamento	Fornecedores	Wi-Fi
Desligamento provisório	Franquia	debito automatico
identificador de chamada	GPRS	transferência assinatura
Parcelamento de Conta	Informacoes	negocio
Telefone pra Deficiente Auditivo	Instalacao	bloqueio chamada cobrar
Acesso	Migracao	
Anatel	Portabilidade	

4 Conclusions

This paper shows the viability of the use of web information resource to find information about competitors, comparing markets and helping the competitive intelligence.

The use of a community technique helped to find the exact intersection where both companies share the same strategy and where they have completely different market plans.

Further work should use text mining to compare the content of each web page and detail their main objectives. Through this methodology the companies are

able to compare their market plans with many concurrencies with low time consuming and specialist support.

Acknowledgement

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

References

[1] SCIP – Society of Competitive Intelligence Professionals - www.scip.org.
[2] LIU, Bing. Web Data Mining - Exploring Hyperlinks, Contents and Usage Data, Springer, December, 2006.
[3] CHAU, M., Shiu, B., Chan, I. and Chen, H. Automated identification of Web Communities for business intelligence analysis. Proceedings of the Fourth Workshop on E-Business (WEB), 2005.
[4] GETOOR, Lise; Diehl, P. Christopher. Link Mining: A Survey. SIGKDD Explorations, Vol 7, Issue 2, Pg: 3–12, 2005.
[5] FLAKE W.G., S. Lawrence, C.L. Lee, and F.M. Coetzee, "Self-Organization and Identification of Web Links", IEEE Computer Journals, 2002.
[6] KUMAR, R., Raghavan, P., Rajagopalan, S., Tomkins, A. Trawling the Web for emerging cyber-communities. Proc. Of the 8th International World Wide Web Conference. 1999.
[7] MOENS Marie-Francine, Information Extraction: Algorithms and Prospects in a Retrieval Context, Springer 2006.
[8] ZANASI, A., 2005, Text Mining and its Applications to Intelligence, CRM and Knowledge Management. 1 Ed. Great Britain, WIT Press.
[9] PAZIENZA, Maria Teresa Information Extraction in the Web Era: Natural Language Communication for Knowledge Acquisition and Intelligent Information Agents, Springer, 2003.
[10] Graphviz – http:// www.graphviz.org/.