Nemeslaki, András; Pocsarovszky, Károly

Conference Paper

Web crawler research methodology

22nd European Regional Conference of the International Telecommunications Society (ITS2011), Budapest, 18 - 21 September, 2011: Innovative ICT Applications - Emerging Regulatory, Economic and Policy Issues

Provided in cooperation with:

International Telecommunications Society (ITS)

Suggested citation: Nemeslaki, András; Pocsarovszky, Károly (2011) : Web crawler research methodology, 22nd European Regional Conference of the International Telecommunications Society (ITS2011), Budapest, 18 - 21 September, 2011: Innovative ICT Applications - Emerging Regulatory, Economic and Policy Issues, http://hdl.handle.net/10419/52173

22nd European Regional ITS Conference
Budapest, 18-21 September, 2011

Nemeslaki, András; Pocsarovszky, Károly

Web Crawler Research Methodology

Abstract

In economic and social sciences it is crucial to test theoretical models against reliable and big enough databases. The general research challenge is to build up a well-structured database that suits well to the given research question and that is cost efficient at the same time. In this paper we focus on crawler programs that proved to be an effective tool of data base building in very different problem settings. First we explain how crawler programs work and illustrate a complex research process mapping business relationships using social media information sources. In this case we illustrate how search robots can be used to collect data for mapping complex network relationship to characterize business relationships in a well defined environment. After that extend the case and present a framework of three structurally different research models where crawler programs can be applied successfully: exploration, classification and time series analysis. In the case of exploration we present findings about the Hungarian web agency industry when no previous statistical data was available about their operations. For classification we show how the top visited Hungarian web domains can be divided into predefined categories of e-business models. In the third research we used a crawler to gather the values of concrete pre-defined records containing ticket prices of low cost airlines from one single site. Based on the experiences we highlight some conceptual conclusions and opportunities of crawler based research in e-business.

JEL codes: C8, O3, D83, L86
Keywords: e-business research, web search, web crawler, Hungarian web, social network analyis

Both authors are at Corvinus University of Budapest:
Corresponding author: Nemeslaki, András; andras.nemeslaki@uni-corvinus.hu

Introduction

In economic and social sciences it is crucial to test theoretical models against reliable and big enough databases. The general research challenge is to build up a well-structured database that suits well to the given research question and that is cost efficient at the same time. In this paper we focus on crawler programs that proved to be an effective tool of data base building in very different problem settings.

As the digital universe is expanding one of the major issues on the internet is the staggering amount of data which is available. In 2008 the amount of data on the internet was estimated as 487 exabytes and 5 times that much was expected in 2009 (Ganz, Reinsel, 2009). The drivers of this explosion are coming largely from the vast number of user interactions, the growth of non-traditional devices, mobile internet usage and growing data processing on servers. This includes our so called personal digital shadow in the form of images and files as we produce more and more user generated content, by 2012 this is estimated to be around 1, 741 Exabyte (Ganz, Reinsel, 2009), and as IDC forecasts 98% of data is coming from devices in the form of digital dust most of which we presently do not intend to store and analyse, which desire of course might easily change in the future.

There are several technical challenges regarding the expanding digital universe. For instance the amount of storage space, speed and reliability with which it is up- and downloaded, organizing and searching, responsibility or compliance regarding the integrity of all this information has become ever more important. From a user perspective, apart from all this, it is also essential to create valuable information from the oceans of data. Given the fact that most of it is unstructured, techniques which enable decision makers to search, indentify, structure and analyze them will be in great demand. On top of that, this need is further pressed by the required speed of finding the needle in the haystack, identifying the key patterns in the data set, or the solid foundation for the decision to be made. That is, making sense of bits in preferably in real time is probably the largest challenge from data processing point of view.

Parallel with the data explosion trend the internet also has posed a challenge for the traditional research methods. In social sciences and humanities we see that field research is accompanied, in several cases substituted, with e-mail correspondence, on-line surveys, computer aided interviewing, on-line monitoring and other internet based tools. Virtual methods create a new type of relationship between researchers and informants and they also

enable to create new strategies and indentify new research sites in the cyberspace (Hine, 2005).

In our paper we present a methodology recommended for researchers in e-business and related fields which enables them to survey the digital universe for relevant data for their analysis and by doing so making it possible to investigate such phenomena which previously were possible only by "small sampling" or approximations. We will also present a research in progress case; a recent study in mapping Hungarian business relationships with social network analysis. Finally, we draw conclusions of how the methodology had worked in these project and show implications and development possibilities.

Webcrawler definition and two basic webcrawler methodologies

Web crawler programs are as old as the world wide web (Risvik, and Michelsen, 2002). They are short software codes sometimes also called as bots, ants or worms written with the objective to download web pages, extract hyperlinks, create a list of its URLs and add them to a local database (Chakrabarti, 2003). Their most widely spread applications are search engines in which form they are familiar to all internet users (Pant et al., 2004). One of the first publications of how effective search crawling and indexing can be on the web was published by Google owners in 1998 as a widely referred academic paper on crawler programming (Brin, Page, 1998).

Building a webcrawler, first of all requires the knowledge of how exactly the users browse a website, and what happens during this process from an information processing point of view. Based on this, the simple idea behind programming a crawler is to (i) imitate the actions of a user visiting a webpage, (ii) extract the information we need, and (iii) repeat these steps. Using this idea we identified three conceptually different type of crawling method each of which is suitable to collect data for different type of analysis and related problem statement.

The first and, we might say, the simplest is when the data we are interested can be found on one site at a well defined place – or technically record – and its value depends on time. For instance this is the logic of storing and presenting market data on some aggregate websites, or showing price information about particular products in a webstore or personal information about employees. In these cases, the data structure is static so the crawler has always visit the same page and same record and periodically download its actual value into a database. In

Figure 1. we show one of our earlier investigation using such a robot which was programmed to collect ticket price data of low-cost airlines in the Budapest-London destination segment.

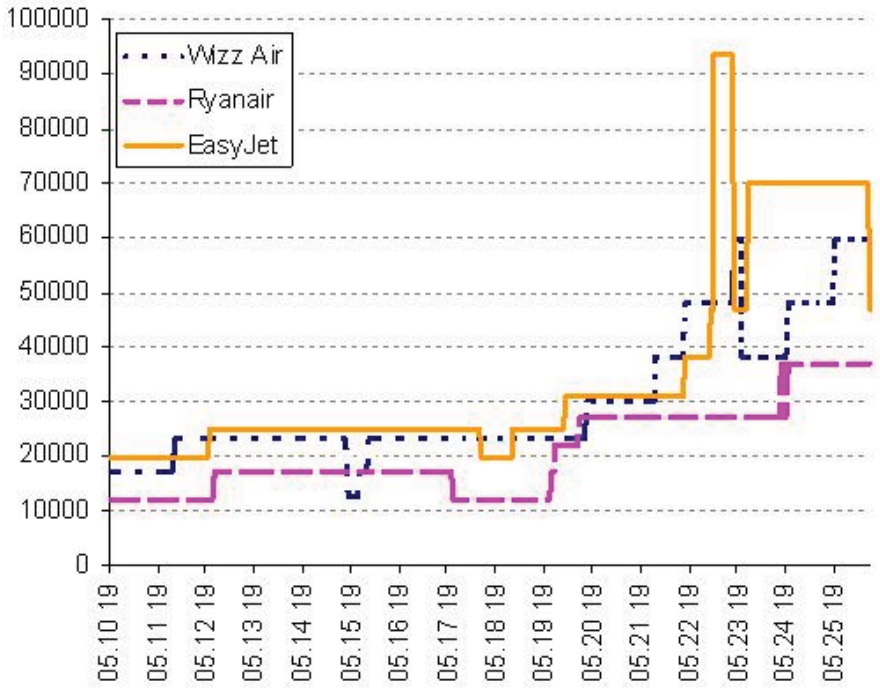


Figure 1. Illustration of a simple webcrawler research: data collection of low cost airline ticket prices

In Figure 1. we can see three low cost airline ticket price trend between May 10 and May 25 2008 collected from a low cost ticket comparison website which was sampled every 15 minutes in the above period.

The second type of crawler is more complex in a sense that it imitates a user who is visiting a list of websites – one after the other – and downloads data, usually different records from each of these sites which after the completion of the visits is grouped, sorted, and analyzed according to the problem or research design.

For the illustration of this crawler, we present Figure 2. which summarizes the most frequently used e-business models in Hungary. For this research we used the list-based robot programming strategy. Starting with the most visited 125 sites according to the Hungarian Web audit, we collected key words which according to business model classifications have been characteristic to unique e-business models (ie. basket, community, subscription, free, download, order placement etc.) The key words for the model classification were identified

based on the e-business model ontology of (Rappa, 2002) and (Osterwalder and Pigneur, 2002) describing the following models. The results of the top visited sites can be seen in Figure 2. 76 pages were classified as advertising (57,41%), 15 as a retail (16,67), 7 as community building (10, 49%), and 5 as marketplace revenue models (3, 09%).

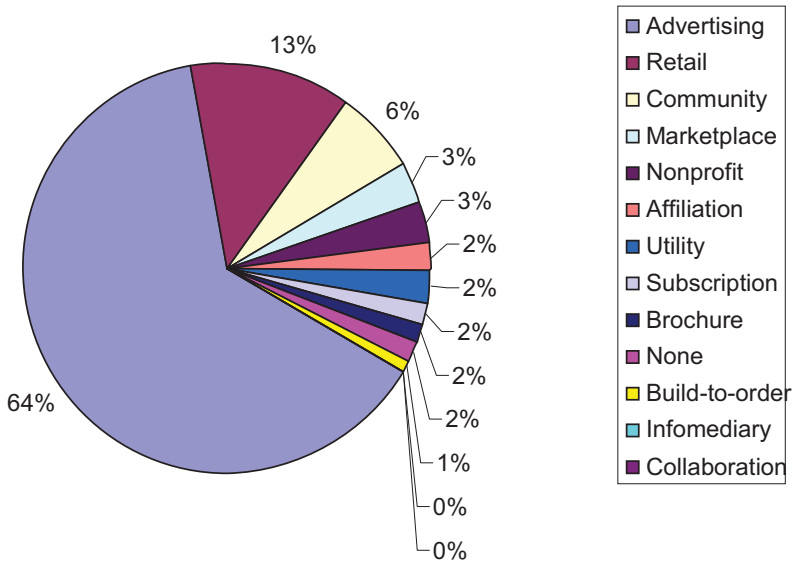


Figure 2. Primary business models in Hungary amongst the Top 125 visited sites in 2008.

The third, and most complex, type of search when the browsing sequence is not pre-determined neither by its sequence of websites nor the actually numbers of them. This happens typically when we explore a domain of a topic, for instance collect information about graduate programs at universities, looking for a particular service, or try to find out the size and connection structure of a community on the net. Since this most complicated case entails both previous once technically and in this present paper we use an example for crawler illustration from this type we describe it in more detail.

Illustration of a complex search process: how the complex crawler works

When a user wants to open e.g. facebook.com the followings takes place on the server and on the client side:

1. Open a web browser

Browser is an interface to communicate with web servers by sending requests and receiving the response packages.

2. Type in http://www.facebook.com

Browser requests the server called facebook.com (69.63.189.11) to generate and forward the source code of the front page. Sending the URL is not the only parameter the client passes to the server. Facebook.com also receives the:

- *Protocol*: what is the "language" of the request?
- *Address*: where to send back the answer?
- *User agent*: who are we? (operating system, browser type, language)
- *Referrer*: what was the previous page we've viewed?
- *Additional parameters*: to control the response (get, post, cookie)

3. Receive the response

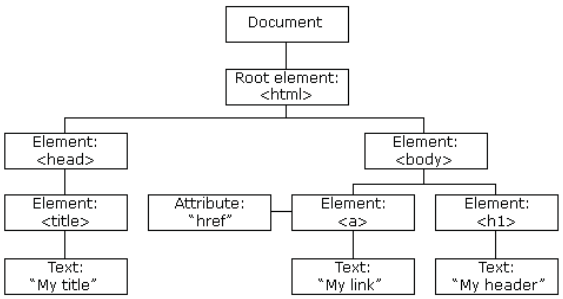
The server according to the request, responses: the *status code* (page not found, access denied, ok…), the *content type* (html, image, video…), the *character encoding* and the *content code* itself. From these information the browser builds and visualizes the webpage.

The crawler has to follow the above steps in order to properly receive a site and extract the desired information. There are several scripting environments that are able to fulfill this requirement (PHP, JAVA, .NET…), so without any further specifications:

1. Generally the crawler has to have a connection function to reach a webpage (in PHP, we've used the cURL library's functions)
2. The connection is parameterized with the request data (who we are, what we want, how long the program should wait for the answer, what is the maximum number of redirections…)
3. Has to be able to receive the response, understand the status code, turn the source code to textual information, and with setting the proper character encoding, understand it (store in the memory and decode)

The source code structure is usually defined by the HTML standard, and such as, we are able to understand it by using the document object model (DOM). Every link, form, image…etc is described by this hierarchical system. For example to list all the bold sentences on the site, search for the `<b>*</b>` expression. More complicated tasks (like find the content of a yellow box placed on the right side) are also easy to carry out using the DOM tree. However, this

technique is limited for the expressions that are fall under the HTML standard. Searching for phone numbers, addresses…etc. requires additional methods.




```
<html>
  <head>
    <title>My title</title>
  </head>
  <body>
    <a href="">My link</a>
    <h1>My header</h1>
  </body>
</html>
```

Figure 3 HTML DOM tree

One of the strongest text-mining tools is the Regular Expression. Extracting complex patterns from texts are impossible by using the standard * and ? wildcards. Regular expression extends the limitations of them, and provides a flexible way for matching strings of text. For example the `\b[A-Z0-9._%-]+@[A-Z0-9.-]+\.[A-Z]{2,4}\b finds` all the email addresses.

Using the above techniques, the crawler is ready to download and process a web page. What if the information we're searching for is scattered around thousands of sites? Let us do the same and repeat it. Naturally, the repetition requires some further work, but it's depending on the final purpose as well. In the followings, we present a situation where one domain (facebook.com) is being inspected through thousands of subpages.

1. First of all we need to login to facebook.com with the crawler program as most of the data are available only to authenticated users. When a client signs in, the previous process is repeated except the 2. step where some additional parameters are set. These parameters are derived from the login form (mainly the username and the password) and from cookies, provided by the page. Both of them have to be attached and copied.
2. After successful login, we select a certain person to analyze his profile. From the crawler point of view, it's not as obvious as before, because facebook.com uses AJAX functions. It means the DOM tree cannot be accessed explicitly in the source code, but dynamic JS programs are handle it. However these scripts return HTML, so after all, it is possible to extract the information we need (using connection stream analysis) and store them.

3. Not only one profile, but its connections are also important. At this step, the crawler saves all the friends, belonging to the previous person. The database stores only the user ID-s and the connections (in [parent, child] form), as the profile URLs can be recovered from them.

4. For a given set of connections the crawler uses systematic breadth-first-search procedure to continue the process from the 2. step. It means we analyze the profiles first and not the connections. As long as there are unanalyzed parents, there is no child process happens.

The database structure is shown in Figure 4. as follows:

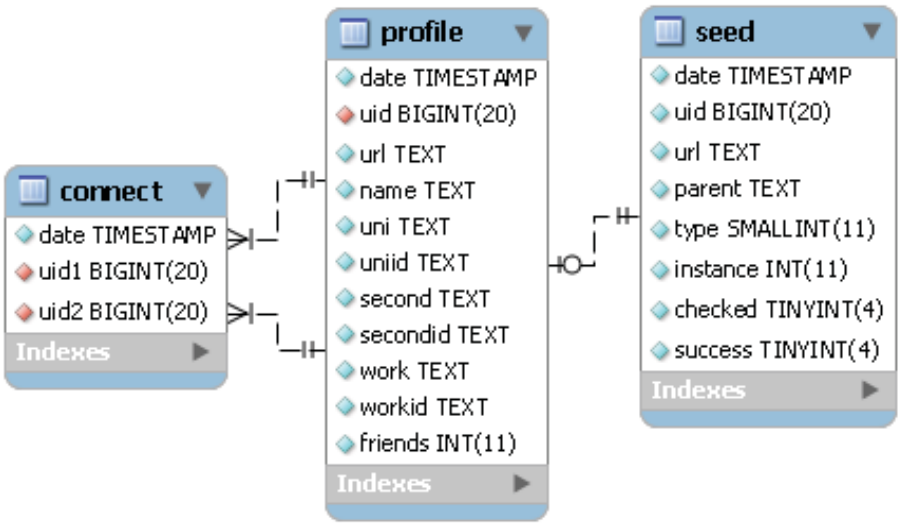


Figure 4. Database structure. Seed table contains the profiles under inspection, profile table is the already downloaded and extracted profiles, while the connect table show the connections between two user

The seed table contains the profiles or web pages to be inspected. Every link, or in our case, every friends is registered there, as the crawler analyze them systematically. Certain flags are set to keep track of the program (what are the downloaded URLs, or was it successful to reach them). The profile table includes all available information we extract. The connection table stores the connection between to link, or two connections.

Three key considerations when planning a crawler research

Beside the technical challenges of automated data retrieval there are other serios dilemmas even ethical problems with crawler based research. There are three types of issue that web crawlers may raise for society or individuals: denial of service and related cost, privacy and copyright (Thelwall and Stuart 2006).

Denial of service

Poorly written or managed web crawlers may slow down web servers by too frequently requesting pages, or may use up limited network bandwidth. This results in denial of service, by analogy to the denial of service attacks by hackers and viruses. Sometime they can also crash routers, disrupt network performances and create server overload. Needless to say, some these technical problems have serious cost consequences for the site owners.

Privacy

Although privacy issues seem to be self explanatory due to the fact that everything on the web is in the public domain, still crawlers may invade privacy basically when information is aggregated on a large scale over many web pages. A typical example for this is the generation of spam lists from email addresses or directories.

Copyright

Copyright is perhaps the most important legal issue for search engines, since they make permanent copies of copyright material (web pages) without the owner's permission (Polansky, 2006). There several practically accepted (but legally not regulated) ways to treat copyright ownership for instance using opt-out or opt-in policies with webpages. Non owners can state the case to remove offending pages and owners can apply the robot.txt protocol. This widely followed standard was developed to address the two most critical issue of webcrawling: the service denial and stop robots to index undesired content areas (Koster, 1993).

Illustrative case of crawler research: mapping business relationships using social media platforms

For the illustration of a complex crawler based search methodology in this section we present a research-in-progress case. Our objective is to focus on demonstrating how the crawler works and we plan to communicate the theoretical and practical details of our enquiry in a different paper discussing enablers and impediments of business ecosystems in Hungary. In order to understand, however, the main context we present the broader motivation of our research.

Recent research in R&D economics has argued that collaborative capability of enterprises play a key role in effectiveness of innovation. This resulted in the conceptualization of cluster creation, establishment of living laboratories and also with the investigation of business ecosystems. All of these structures are rooted in intertwined connection and co-operation of smaller and bigger companies, research organizations, educational institutions in regional and broader environments. The purpose of this collaboration is creating, innovating and delivering such products and services which are competitive and sustainable for the partners and produces substantial enough profit for their survival. Sharing of knowledge, best practices, and working in common technology platforms are key characteristics of such ecosystems.

Coming from this theoretical background there is a strong argument that major impediments for innovative capabilities originate from the lack of willingness to create such ecosystems in the Hungarian economy. Since collaboration is not only a matter of economic necessity but also deeply rooted in socio-cultural traditions as well. Several research in the area of building trust, in the role of alumni networks, and also in exploring individual ties among business relations have argued for the importance of such social network relations. Our research motivation originates from these ideas and intends to contribute the mapping of such relations in the Hungarian economy and by doing so, contributing to the improvement of innovation capabilities.

Given the high rate of internet penetration and the popularity of social media, specifically Facebook, we assumed that connections identified here are good enough indicators of basic collaboration in business. We assumed that if two individuals declare relationship between each other then their employers are potentially ready for collaboration. The more employees know each other between two firms the stronger the tie and greater the potential for rich collaboration.

The crawler which we developed visited 15000 users of Facebook from which we used 5000 who publicly displayed their employers. To illustrate the complexity of the ties we show the density of user connections in Figure 2.

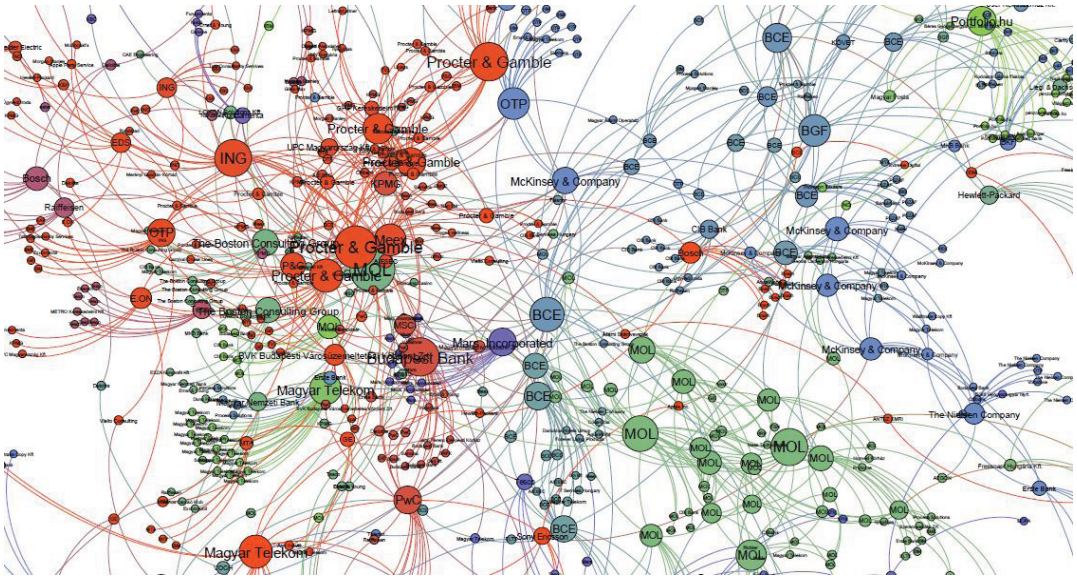


Figure 5 Network created based on crawling through 5000 user pages in Facebook

In Figure 5. we depicted the users in the edges and their relationship in the vertices. The bigger the node the more connections the particular user has, and different shading represents the different type of employer.

We then transformed this network into a relationship graph of employers ie. companies. The more individuals knew each other between firms the stronger the relationship became and this is described in Figure 6.

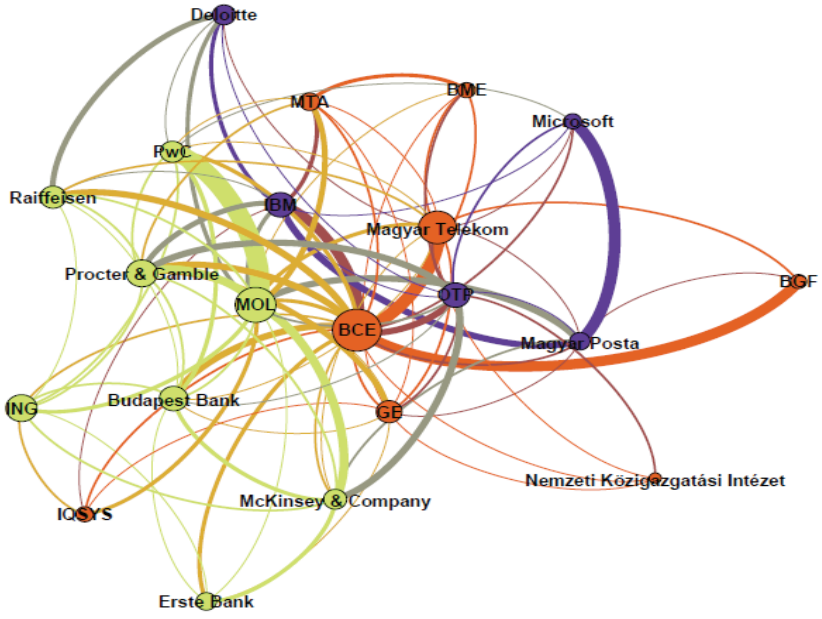


Figure 6 Summarized and filtered relationship of Hungarian employers based on crawling 5000 Facebook users

The results in Figure 6. can be interpreted with many ways but naturally with caution considering the limitation of the dataset and the crawler methodology. Since the source of the search started from one of the authors of this paper in there is a natural strength of centering the university employers in the center, but also indicating the manyfold relationship with several industries: banks, consulting firms, post, telecommunication and the software industry. We can say in general that the graph in Figure 6. represents a dense small world network with an intertwined core. We were are also able to visualize that connected to the main firms in Hungary (OTP, MOL, GE, IBM, Magyar Telecom) we see consulting firms who seem to have specialized connection with these main ones (IQSYS, Deliotte, McKinsey, PWC).

Classification scheme for crawler research

Although social scientists have recognized the value of crawler supported research, it is surprising to realize that it has been relatively few documented applications for research in e-business. For instance for the keywords, "web search" there are 7 hits in the Bled e-Conference Proceedings database, and 1 hit for the keyword "crawler" (Polansky, 2006)

Reviewing the relevant topics from these we find for instance that Riemer and Brüggemann presents the use of search tools to support different kind of personalization methods in the web (Riemer, Brüggeman, 2006). Advertising and the connection of sponsored search is also an important topic in e-commerce, especially since the wide spread of Adwords (Stepanchuk, 2008). We also found illustration for search issues in corporate intranet effectiveness using the technology acceptance model (TAM) (O´Boyle at al, 2009), and a case based discussion of on-line knowledge intermediaries as new business models building on searching methodology (Bolisani, Di Biagi, Scarso, 2003).

In the e-business domain we propose three, characteristically different type of crawler based research method according to the nature of the study and the data which is needed for the analysis.

Exploration

The closest to the commercial search problem is when we need to explore the web to find and identify objects for further analysis, basically to collect descriptive data about the informant. In this case we have to browse through many, usually unknown, sites in a given web domain,

download URL addresses, links and put some records into a database. The population size is quite often unknown and the first main objective is to find out the key characteristics of an unexplored field.

Classification

Quite often we have classification schemes into which we need to categorize observations in order to find out the size and relevance of each the particular categories. For instance we need to find out the geographic distribution of visitors of an e-business service, or we have to group the existing web shops of a region by size, product range or payment method. Usually in the web business context the population is limited by the network access and boundaries can be created by setting the IP address domains, but aggregators or on-line lists also can serve as the population limit. In these instances the task is to crawl through the lists, check the matching of category keywords and download them into a database.

Data panel

The most specifically determined web search situation is when the research domain is one single site (usually an aggregator) with a well known and stabile data structure from which we need to fetch the value of a concrete record. Usually this happens when we analyse trends (time series), but in general this approach is also applicable when we want to follow the changes of an object (a person, an organization, a team etc). For instance we might be interested what the most recent CEO´s name is, or occupation of a person or profit of a company at the end of the income statement. Data panels can also be created with this kind of crawler technology, when we collect data for both time series and cross section analysis (correlation of two or more variables).

Table 1 Summary of three different web crawler-based research cases

Type of research	Topic of investigation	Sites crawled	Unit of analysis	Verification	Result
Exploration	Web agency industry	10.900 sites	948 companies	189 filled out surveys 12 interviews	Descriptive data about web agencies, Key applied technologies
Classification	Business Models	Top 125 Hungarian sites 6800 IP address	12 identified business models 1409 sites have relevant business models	30 sites tested manually	Identification of primary and secondary business models in Hungary

Type of research	Topic of investigation	Sites crawled	Unit of analysis	Verification	Result
Data panel	Low-cost airline pricing	1 site 15 airlines	66 observations 6 Gbyte data	Data cleaning, deleting damaged records	Price correlation between airlines (dictating, following)

Issues of the methodology and implications

The major learning we gained during our projects regarding crawler development was how important it is to find the good and appropriate crawling strategy. In all the cases we eventually went back to some key data aggregators to build the crawler and start the search. Multithread seeds and starting from different servers might not only be complicated but the sites downloaded at the end might also be off target.

As Table 1 shows the results and the processed data in all the three domain was staggering, and clients of all three projects were happy that they are able to look at really relevant numbers and information regarding the web agencies, business models, and pricing strategies on the internet.

From the development point of view, the research team also felt that the automated data gathering relieved their attention and focus to the more added value analyisis and interpretation of information. We have to emphasize however that the manual checks and traditional surveys still have a pivotal role in verifying the input data.

On the negative side denial of service was the most painful experience we gained with skyskanner.com where the hosts considered our fifteen minute interval requests as an attack against the site. To prevent such instances it is advisable to contact and negiotiate the research concept with the targeted sites and get their approval for launching the crawler.

Web crawling involves a number of stakeholders whose interests and problems have to be taken into consideration. There are the web site owners, the crawler developers´organization and the clients of the gathered data. A responsible researcher using crawlers should assess the costs and benefits of all of these stakeholders and raise these issues and discuss with them in order to minimize the risks of unethical and illegal research conduct.

Frankly, we realized that these are such new questions in the academic research methodology toolkit, that we proposed in the Ph.D. programs of Corvinus University to amend the well known quantitative and qualitative research methodology curriculum with this third area of automated internet research. We are convienced that this is a very promising are for the academic community and deserves a lot of attention in the future, as not only the number of data show it in the internet but also the increasing relevance and penetration of the network economy.

Conclusions

In e-business and related fields it is always important to support analytical arguments and conclusions with up-to-date and relevant data. If we seriously consider the richness of the digital universe, and assume based on the number of bits available on the internet that relevant and up-to-date data is available on-line, then using virtual research methods we can produce results built on almost "real time" information.

With a complex illustrative example and a reference to three other cases we argued in our paper that with the help of crawler based search programs we can support social and economical studies of internet based business models.

We proposed different types of crawler search an exploratory one for indentifying populations in a given web environment; another for classification purposes to investigate and verify the practical relevance of conceptual models; and a third for assembling concrete datapanels for detailed time series and correlation analysis.

Crawler based data collection in our opinion will be a very important tool for future academic research in the internet for economics and social sciences as well. Our experience has shown that beside the effectiveness and efficiency of automated search manual verification and the combination of "traditional" survey and interview methods are necessary to verify the findings. Also there are some new ethical and legal dilemmas of crawler development like privacy, copyright, bandwidth occupation and cost issues for the environment.

With paying careful attention to these details, however, we are convinced that crawler based methodologies might be the next frontier of e-business research tools deserving a lot of attention by the information system research community.

Acknowledgement

The research is part of the TÁMOP 4.2.1/B09/1/KMR 2010 0005 project, which is a research and innovation program of the Corvinus University of Budapest aiming ICT based analyses of knowledge transfer, sharing and knowledge codification fields.

References

Bolisani E., Biagi Di M., Scarso E. (2003): Knowledge Intermediation: New Business Models in the Digital Economy, 16th. International Bled eConference, Bled, Slovenia, June 9–11.

Brin, S., Page L. (1998): The anatomy of a large-scale hypertextual Web search engine, Computer networks and ISDN systems, 30 (1-7). 107–117.

Chakrabarti, S. (2003). Mining the web: Analysis of hypertext and semi structured data. New York: Morgan Kaufmann.

Ganz J., Reinsel D. (2009): As the Economy Cont racts, the Digital Universe Expands, IDC-Multimedia White Paper.

Hine (2005): Virtual Methods: Issues in Social Research on the Internet, Berg, London.

Koster, M. (1993). Guidelines for robot writers. Retrieved February 23, 2005 from http://www.robotstxt.org/wc/guidelines.html

O´Boyle, P., Acton T., Campion, M., Conboy K., Scott, M. (2009): Towards a Toolset for Intranet Evaluation, International Bled eConference, Bled, Slovenia, June 14 - 17.

Osterwalder, Alexander and Yves Pigneur (2002), An e-Business Model Ontology for modelling e-Business, 15th Bled Electronic Commerce Conference. E-Reality: Constructing the e-Economy (june).:1-11.

Pant G., Srinivasan P., Menczer F. (2004), "Crawling the Web", in Levene M., Poulovassilis A. Web Dynamics: Adapting to Change in Content, Size, Topology and Use, Springer, pp. 153–178

Polansky P. (2006): Intellectual Property Law versus Customs and Values of the Internet Community, International Bled eConference, Bled, Slovenia, June, 5-7.

Rappa, .M (2002) Business Models on the Web. http://digitalenterprise.org/ models/models.html, 2002

Reimer K., Brüggemann F. (2006): Personalisation of eSearch Services – Concepts, Techniques, and Market Overview, International Bled eConference, Bled, Slovenia, June, 5-7.

Risvik, K. M. and Michelsen, R. (2002). Search Engines and Web Dynamics. Computer Networks, Vol. 39, pp. 289–302, June 2002.

Stepanchuk T. (2008): An Empirical Examination of the Relation between Bids and Positions of Ads in Sponsored Search, 21st International Bled eConference, Bled, Slovenia, June 15 – 18.

Thelwall M., Stuart D. (2006) "Web crawling ethics revisited: Cost, privacy and denial of service", Journal of the American Society for Information Science and Technology, Volume 57 , Issue 13 Pages: 1771 – 1779.