

Visualised Indicators of Links in the World Wide Web Biotechnology and Information Science

Marianne Hörlesberger, Edgar Schiebel

*ARC systems research GmbH, Technology Management
A-2444 Seibersdorf*

*Tel: +43 50550 3864, Fax: +43 50550 3888;
marianne.hoerlesberger@arcs.ac.at*

Abstract: The World Wide Web is open to everybody who has access to the Internet. A posted document or link is available to hundreds of millions of people. It is assumed that Cyberspace is a forum of democracy. Everybody's voice can be heard. Albert-Laszlo Barabasi, however, states in "Linked" (2002, Penguin Group) that there is a complete absence of democracy, fairness, and equality in the World Wide Web. A link or a document cannot be found easily. The measure of visibility on the Web is the number of links. To analyse the linkage behaviour is a question of structure. Network Analysis methods have been developed for investigating structures in social co-operation and networks. Such methods are used and adapted for analysing networks in the World Wide Web. In this contribution we focus on the analysis of linkages in the fields of biotechnology and information science in the World Wide Web and on the visualisation of their Web-networks. For visualisation we use the method BibTechMon™. We try to find answers to the questions: Which indicators are adequate for the analysis and can be visualised? Which difference in linkage behaviour is there between the fields of biotechnology and information science in the World Wide Web? How did patterns change from year 2001 to 2002? We are analysing the Web-sites of the universities and research institutes of the European member states of the years 2001 and 2002. The investigated data are generated automatically by a tool developed by Isidro F. Aguillo CINDOC-CSIC Madrid, Spain.

Keywords: web analysis, network, visualisation, hubs, authorities, centrality, density, information science, biotechnology.

Introduction

The discovery of hubs and authorities in networks forces us to give up the idea that nodes (web pages, ULRs) are equivalent in the Internet. Determination of authorities and hubs are important for creating a search engine. The terms "authority" and "hubs" are used to describe the success of finding the relevant documents. Google or Altavista have developed special algorithms so a user of the search engine can find the best documents in the World Wide Web quickly. Their topic is the content of a Web page. In contradiction to that we are interested more in the linkage behaviour. What are the authorities and hubs in biotechnology and information science of universities and research institutes of the European member states? Behind the hubs there is a rather strict mathematical expression, a power law. The computer scientists M. Faloutsos, P. Faloutsos and Ch. Faloutsos (three brothers at three different universities) show in their

paper 1999 that the connectivity distribution of the Internet, a collection of routers linked by various physical lines, follows a power law. Power laws lie behind hubs. Then we have to distinguish between in-degree and out-degree. "Despite the billion documents on the Web, nineteen degrees of separations suggests that the Web is easily navigable. To be sure, if there is a path between two documents, the path is typically short. [...]. The Web is directed." (Albert-Laszlo Barabasi, 2002). In which way can the method BibTechMon™ calculate and visualise the considered indicators? The following contribution represents first steps of applying BibTechMon™ to find indicators like hubs and authorities in Web data and we learn that Bib TechMon™ is an adequate tool for that.

Methodology

We have to distinguish between Internet and World Wide Web. The World Wide Web is a network of Web pages containing information, linked by hyperlinks from one page to another. The Internet is a physical network of computers linked by optical fibre and other data connections (M. E. J. Newman; 2003). Here we consider the World Wide Web, the hyperlinks in the World Wide Web in special fields as said above.

Computer scientists have developed different methods, measurements and indicators to structure, to analyse or to describe the huge amount of information in the World Wide Web and they are trying hard to improve the methods. When we deal with data of the World Wide Web we are faced with very different kinds of data like URLs, html files, pdf files, outgoing links, incoming links, et al. Many books and Web pages have been written to define and describe the objects which can be detected in the World Wide Web. For each object and for the relations of the different objects there are different methods for adequate analysis. In this paper we consider URLs and links of their Web pages to other URLs. Therefore we are interested in "in-degree", "out-degree", "authorities", "hubs", network indicators like density and centrality and the visualisation.

In-degree is the number of links pointing to a page. A Web page with a lot of links into it is an **authority**. A Web page with a lot of links into it is probably a better authority than one without.

Out-degree is the number of links coming out of a page. A page with a lot of links out is a **hub**. Then a Web page with a lot of links of hubs is better than a Web page with a lot of links of ordinary pages.

BibTechMon™

BibTechMon™ (bibliometric technology monitoring) is a software tool for structuring, visualising and analysing data. The tool was developed in the Department of Technology Management of ARC systems research GmbH (Kopcsa and Schiebel, 1998) and is still being improved. This method is based on the calculation of co-occurrences of objects. Two objects are connected if they occur together in documents, in patents, or they have a link to the same URL, et al. BibTechMon™ has been applied for analysing a huge amount of documents like articles of journals or patents. Via the key term extraction of BibTechMon™ the phrases of the considered documents are listed. Then other content issues like IPC codes in patents or names of institutes can be of interest for further analysis. The different objects of a dataset can be correlated and visualised via BibTechMon™. Thus we get networks. After the iteration of the network

we have a lot of possibilities for analysing. In this contribution we apply this tool to Web data. BibTechMon™ is very well qualified to calculate networks of link data, to show the connections of URLs to other URLs because of the method of co-occurrence as the following analysis will show.

Actor degree centrality

Actor degree centrality $C_D(n_i)$ is defined as (see Wasserman and Faust 1994, p. 178)

$$C_D(n_i) = \frac{d(n_i)}{g-1}$$

where $d(n_i)$ is the degree of node n_i and g is the total number of nodes n . This index is independent of g and can therefore be compared across networks of different sizes.

The index takes on values from 0, if n_i has no adjacent nodes, to 1, if all remaining nodes in the network are directly adjacent to node n .

The degree of a node is the number of nodes adjacent to it. The centrality shows how often a node is located on geodesics in a network. Nodes with a high centrality play a central role in the network.

Density

The density of a network is defined as the numbers of lines/paths in a network, expressed as a proportion of the maximum possible number of lines/paths. The formula for the density is

$$\frac{l}{n(n-1)/2}$$

where l is the number of edges present and n is the total number of nodes. It can take on values from 0, if nodes are totally unconnected, to 1, if all nodes are connected (Wasserman and Faust 1994, p. 101).

Data

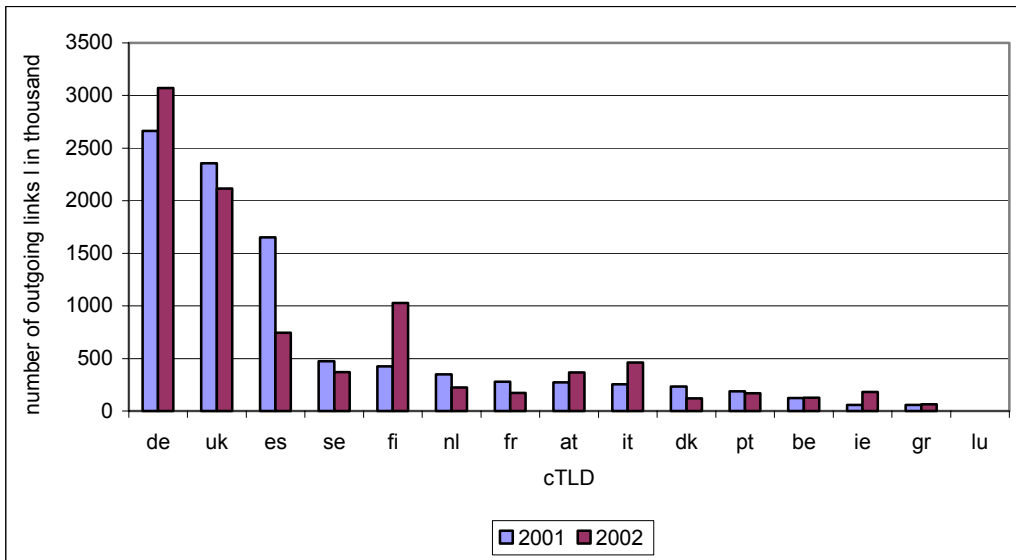
The Web sites of the fifteen European research institutes of the years 2001 and 2002 with their URLs and their outgoing links are the basis of our dataset. Each Web site can be allocated to a code which gives information about the content of the Web site. Isidro F. Aguillo, CINDOC-CSIC, Madrid used the so called UNESCO codes to identify the content of the Web sites. We extracted the URLs allocated to topics of biotechnology (codes of biotechnology: 2302 biochemistry, 2403 biochemistry, 2409 genetics, 2414 microbiology, 2415 molecular biology, 3101 agricultural chemistry, 3202 biochemical technology, 3309 food technology) and information science (codes of information science: 1203 computer science, 1207 operations research, 2203 electronics, 3304 computer technology, 3307 electronic technology).

General Statistics

Before we go into detail we describe the huge amount of records with some general statistics. The emphasis is on the links, especially on the outgoing links. Figure 1 and figure 2 represent the number of outgoing links in the two considered fields of the years 2001 and 2002. We see that in almost each country there are at least four times more outgoing links in information science than in biotechnology. The gap between the two

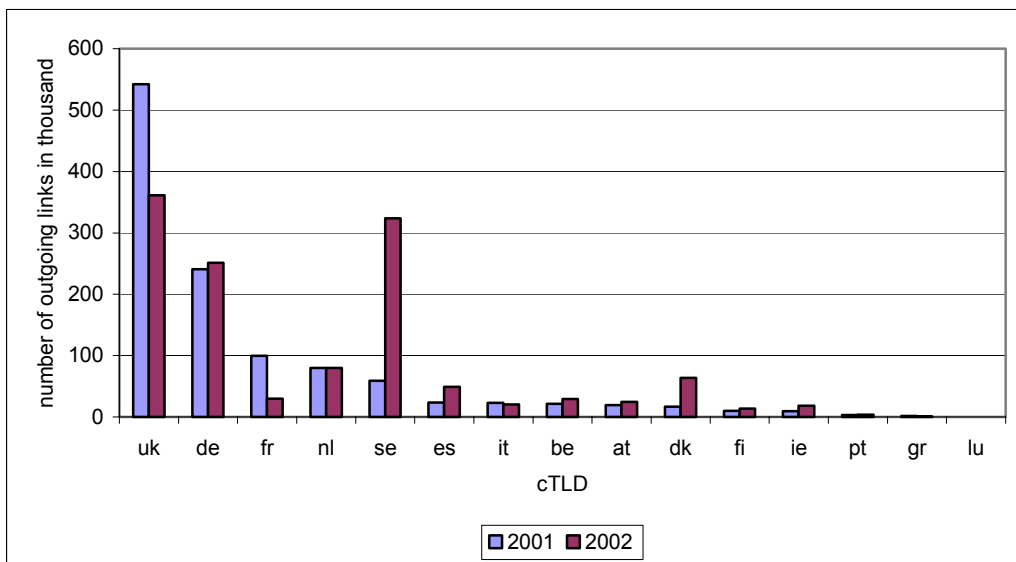
years is not so big (the runaways in Finland in information science, in the United Kingdom and Sweden in biotechnology are not considered in detail).

Figure 1: Number of outgoing links (in thousand) in information science



at: Austria, be: Belgium, de: Germany, dk: Denmark, es: Spain, fi: Finland, fr: France, gr: Greece, ie: Ireland, it: Italy, lu: Luxembourg, nl: Netherlands, pt: Portugal, se: Sweden uk: United Kingdom

Figure 2: Number of outgoing links (in thousand) in biotechnology

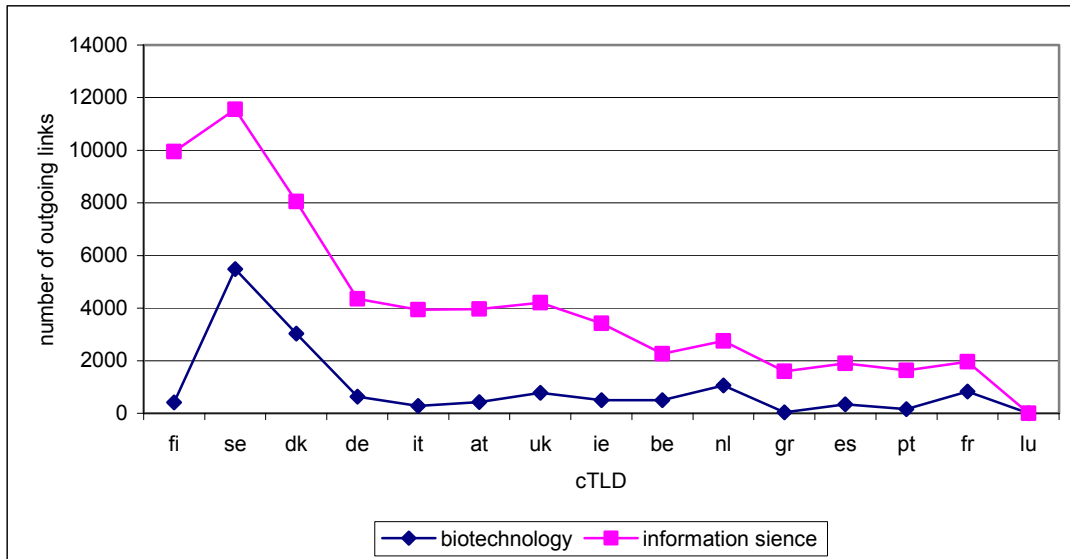


at: Austria, be: Belgium, de: Germany, dk: Denmark, es: Spain, fi: Finland, fr: France, gr: Greece, ie: Ireland, it: Italy, lu: Luxembourg, nl: Netherlands, pt: Portugal, se: Sweden uk: United Kingdom

Figure 3 represents the number of outgoing links per Web site. As we expected it shows that information science is more connected, has a network with a higher density in the World Wide Web than biotechnology,. Although biotechnology is a modern science, was established in the time where Internet “grew up”, biotechnology does not use the World Wide Web in the same intensity as information science does. Are the presence and the density of a science community in the World Wide Web a question of knowledge and capability of handling Internet technologies? Furthermore we learn from figure 3 that the

countries of the northern part of Europe (especially Finland, Sweden, Denmark, Germany) are better represented or connected in the World Wide Web in both fields than the southern part of Europe.

Figure 3: number of outgoing links per Web site



at: Austria, be: Belgium, de: Germany, dk: Denmark, es: Spain, fi: Finland, fr: France, gr: Greece, ie: Ireland, it: Italy, lu: Luxembourg, nl: Netherlands, pt: Portugal, se: Sweden uk: United Kingdom

Networks

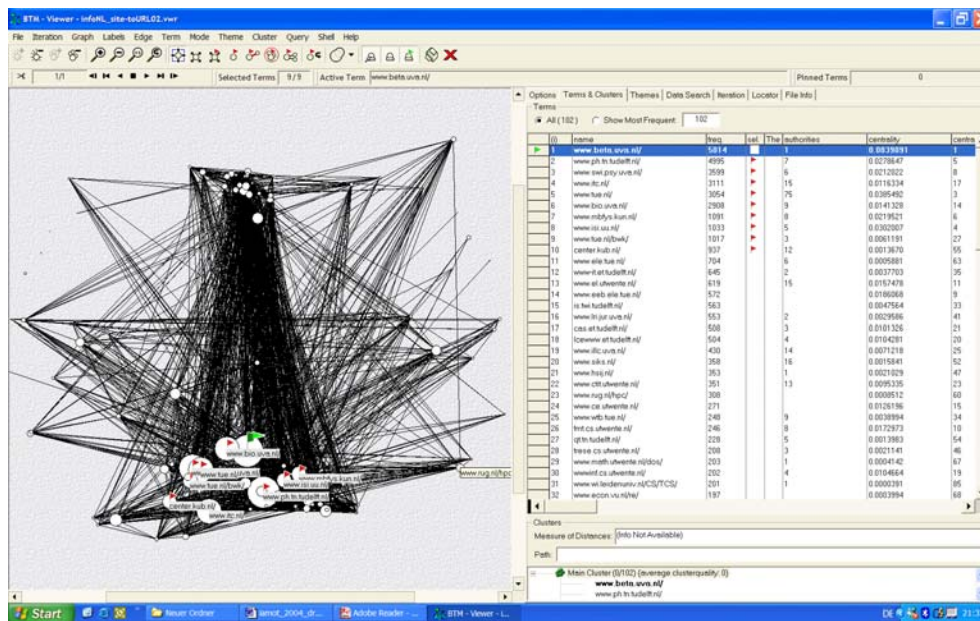
The World Wide Web is cyclic. There is no natural ordering of sites and no constraints that prevents the appearance of closed loops. Our data of the Web come from “crawls” of the network, in which Web pages are only be found if another page points to it. In a crawl that covers only a part of the Web (as all crawls do at present) pages are more likely to be found the more other pages point to them. This suggests for instance that our measurement of fraction of pages with low in-degree might be an underestimated. This behaviour contrasts with that of a citation network. A paper can appear in the citation indices even if it has never been cited.

The World Wide Web delivers a very huge amount of records for analysis, e.g. there were about 1.6 billions outgoing links of Germany in 2002. An excellently working computer could probably calculate a network of all connections of the World Wide Web. But would it be possible for us to recognise any structure in this calculated network on a computer monitor afterwards? We focused on the visualisation of data of different countries in the two named fields. According to an OECD report the Netherlands is most of the active country of the European member states in the field of information technology while Belgium is very busy in biotechnology.

A node in the network is an URL (of a Web site of a university or a research institute) in our considered dataset. Two nodes are connected if they link to the same URL. The position of a node in the network is given by the “forces” to all other nodes. A node with a high frequency and connections to almost each other node is quite in the middle of the network. If there are e.g. two very strong connected groups of nodes in the network, the two groups are pressed to the edge.

The method BibTechMon™ delivers the visualised networks. After calculating the networks there are a lot of different possibilities for representing indicators of the networks available. In order to limit the illustrations we represent our considered indicators in one figure per network.

Figure 4: Network of URLs based on their outgoing links: information science of the Netherlands in the WWW 2002.



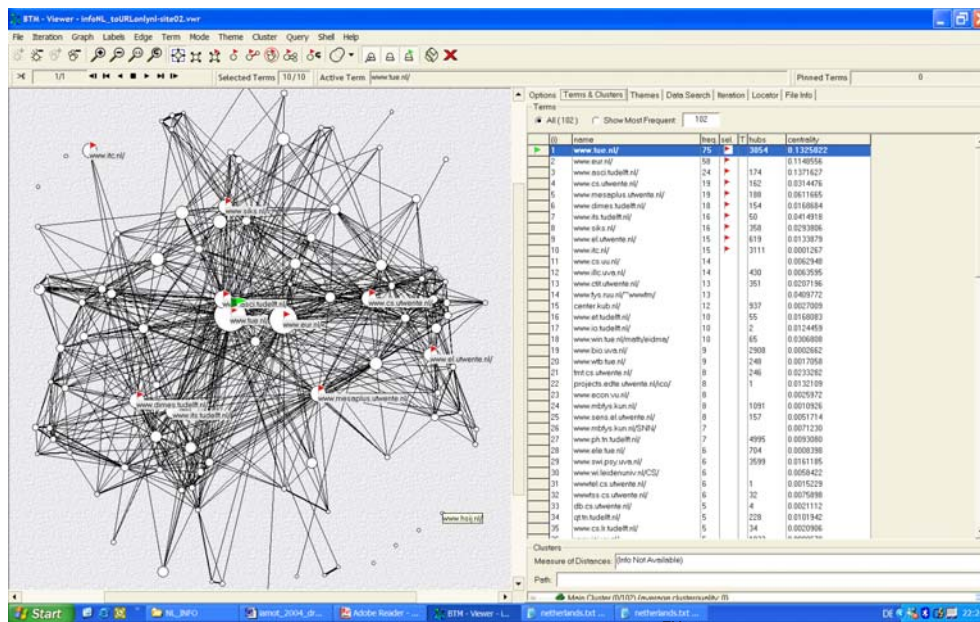
Source: calculated with BibTechMon™

The network in figure 4 has 102 nodes (different URLs) and 1857 edges (under the layer “options” we find the number of connections). The out-degree of each URL is the number of outgoing links. The out-degree is exactly the “frequency” (right in figure 4: “freq”). The size of a node represents the out-degree. The higher the out-degree the bigger is the node. The nodes with the highest out-degree are hubs. They are selected in the network (figure 4: red flags on nodes and in the column). The URLs www.beta.uva.nl/, www.ph.tn.tudelft.nl/, www.swi.psy.uva.nl/, www.itc.nl/, www.tue.nl/, www.bio.uva.nl/, www.mbfys.kun.nl/, www.isi.uu.nl/, www.tue.nl/bwk/, center.kub.nl/ have the highest out-degree as the number in column “freq” shows. Good hubs link to authorities. The entries in column “authorities” (figure 4, right) are the in-degrees of the considered URLs. The highest in-degree of 75 (we considered only the incoming links of URLs from the Netherlands) has www.tue.nl/ followed by www.itc.nl/ with an in-degree of 15 of the selected URLs in the network. The third rank goes to center.kub.nl/ with an in-degree 12. Therefore we conclude that the best hubs of the network in figure 4 are www.tue.nl/, www.itc.nl/, center.kub.nl/.

The values of actor degree centrality are listed in the column “centrality” (figure 4). www.beta.uva.nl/ has the highest actor degree centrality followed by www.sens.el.utwente.nl/ and www.tue.nl/bwk/. This indicator represents another quality of a network. The density of this network is 0.3605125.

The incoming links for “networks of URLs based on their incoming links” are only from the Netherlands. The study of authorities is limited to the incoming links of the Netherlands. We cannot compare the networks in figure 4 and figure 5 directly.

Figure 5: Network of URLs based on their incoming links: information science of the Netherlands in the WWW 2002.



Source: calculated with BibTechMon™

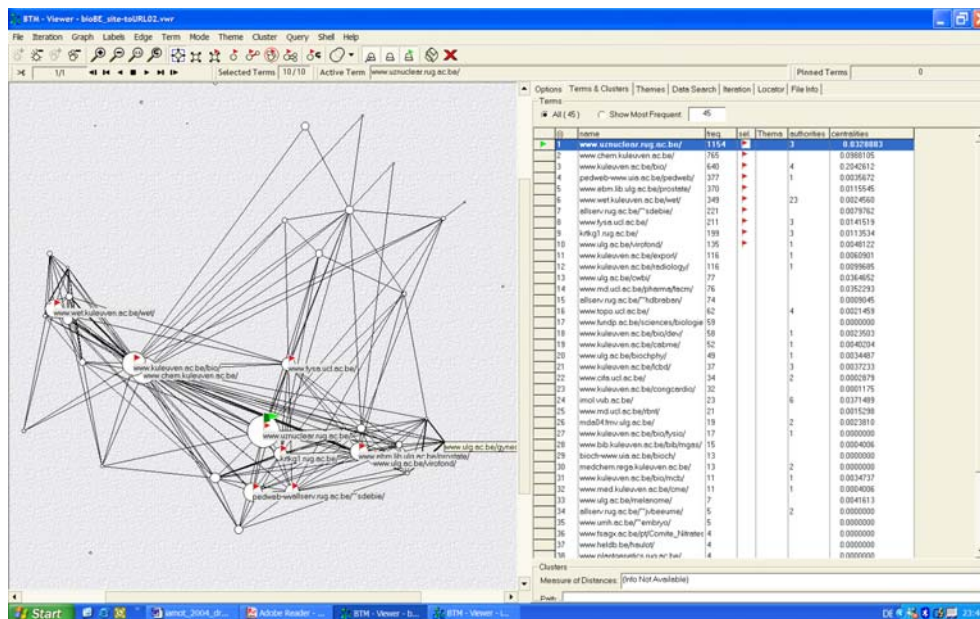
There are 102 nodes (URLs) in this network with 783 edges. The first impression of the network in figure 5 is that the nodes with the highest frequency are quite in the middle. This is because the connections of the nodes with the highest frequency are connected similarly to all other nodes. There is more balance of connection in the network. As described in figure 4 a node is an URL. In this case two nodes are connected if they have they same incoming link. So the in-degree is given by the “frequency” (figure 5: column “freq”). URLs with the best in-degree are authorities. The ten URLs with the best in-degree are www.tue.nl/, www.eur.nl/, www.asci.tudelft.nl/, www.cs.utwente.nl/, www.mesaplustwente.nl/, www.dimes.tudelft.nl/, www.its.tudelft.nl/, www.siks.nl/, www.el.utwente.nl/, www.itc.nl/. Good authorities are “created” by good hubs. The values in column “hubs” in figure 5 represent the quality of the authority (the nodes with red flags in the network). When we consider the selected URLs (nodes with red flags) www.tue.nl/ is the best authority because it is a link of many hubs. On rank two there is www.itc.nl/, then www.el.utwente.nl/ follows.

It seems that www.tue.nl/ was one of the most important Web sites of information science of the Netherlands in 2002.

The actor degree centrality is listed in the column “centrality” (figure 5). The five URLs with the highest values are www.asci.tudelft.nl/, www.tue.nl/, www.eur.nl/, www.mesaplustwente.nl/, www.its.tudelft.nl/. www.tue.nl/ is among the best ones too. The density of this network is 0.1520093.

The field of biotechnology is represented completely differently in the World Wide Web. The results of the general statistics are supported by visualisation. When one considers the network you could get the impression of looseness. We show it on the sample of Belgium.

Figure 6: Network of URLs based on their outgoing links: biotechnology of Belgium in the WWW 2002.



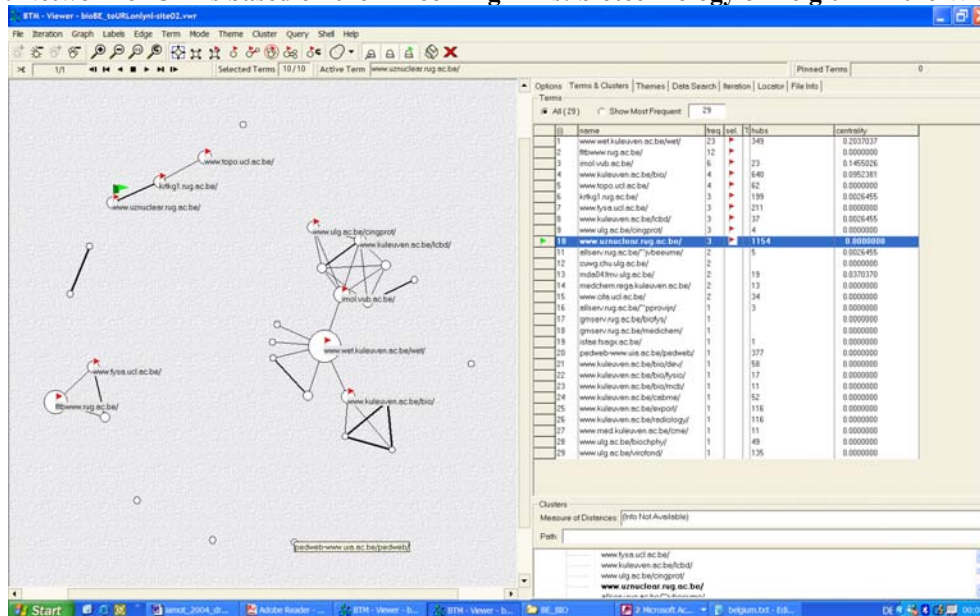
Source: calculated with BibTechMon™

We have 45 different URLs of universities or research institutes as nodes in the network with 201 edges. There are two nodes in the south west of the networks which are strongly connected to most of the other nodes. Two nodes are connected in this network if they link to the same URL. The hubs in this case are www.uznuclear.rug.ac.be/, www.chem.kuleuven.ac.be/, www.kuleuven.ac.be/bio/, pedweb-www.uia.ac.be/pedweb/, www.ebm.lib.ulg.ac.be/prostate/, www.wet.kuleuven.ac.be/wet/, allserv.rug.ac.be/~sdbie/, www.fysa.ucl.ac.be/, krtkg1.rug.ac.be/, www.ulg.ac.be/virofond/. The best hubs are URLs which link to good authorities. The best ones here are www.wet.kuleuven.ac.be/wet/ and www.kuleuven.ac.be/bio/ as column “authorities” in figure 6 shows.

The values with the highest actor degree centrality are allocated to www.kuleuven.ac.be/bio/, www.chem.kuleuven.ac.be/, imol.vub.ac.be/, www.ulg.ac.be/cwbi/, www.md.ucl.ac.be/pharma/facm/. The density of the network is 0.2030303.

At last we represent the network URLs based on their incoming links (the incoming links are only with a country top level domain of Belgium).

Figure 7: Network of URLs based on their incoming links: biotechnology of Belgium in the WWW 2002.



Source: calculated with BibTechMon™

The network in figure 7 seems very loose. But we should take into account that for networks based on incoming links we can only use the incoming links of Belgium in this case. My computer could not manage to extract the incoming links of the whole WWW. The network in figure 7 has 29 nodes with 33 connections. The ten URLs with the highest in-degree are www.wet.kuleuven.ac.be/wet/, fltbwww.rug.ac.be/, imol.vub.ac.be/, www.kuleuven.ac.be/bio/, www.topo.ucl.ac.be/, krtkg1.rug.ac.be/, www.fysa.ucl.ac.be/, www.kuleuven.ac.be/lcbd/, www.uznuclear.rug.ac.be/. Which of these authorities are the best ones? The column “hubs” represents this: www.uznuclear.rug.ac.be/, www.kuleuven.ac.be/bio/ and www.wet.kuleuven.ac.be/wet/. The actor degree centrality is listed in the column “centrality”. The five URLs with the highest values are www.wet.kuleuven.ac.be/wet/, imol.vub.ac.be/, www.kuleuven.ac.be/bio/, mda04.fmv.ulg.ac.be/, www.kuleuven.ac.be/lcbd/. The density of the network is given with 0.0812808.

When we consider both networks of Belgium we see that www.kuleuven.ac.be/bio/ and www.wet.kuleuven.ac.be/wet/ were the most important URLs of biotechnology in Belgium 2002.

Conclusions

After handling, analysing, structuring and calculating an incredible huge amount of records in many different databases and finding adequate indicators it is hard to imagine that Internet started with only one Web site, the Web site of Tim Berners-Lee. The random model of Erdős and Rényi is based on two assumptions. First, one starts with an inventory of nodes. Having all the nodes available from the beginning, one assumes that the number of nodes is fixed and remains unchanged. All nodes are equivalent. Unable to distinguish between the nodes, they are linked to each other. But the discovery of hubs and the power laws that describe them shows that the nodes are not equivalent. “Rich get richer – law” is in force. Therefore scientists are developing new methods and tools to detect the best links, URLs. For which kind of analysis of

Web data would BibTechMon™ be an adequate method, was one of our questions. We found that BibTechMon™ is a powerful tool to visualise out-degree and in-degree and therefore hubs and authorities. The next steps in improving the tool will lie in integrating algorithms for computing weights algorithm for hubs and authorities.

This contribution shows that there are differences in the representation of different research fields and there is a geographical difference. Those fields which have had a “natural” access to the technology Internet like mathematics, physics, electronic technology et. al. are represented very well in the World Wide Web. The universities of the northern part of European are represented better in the WWW than the southern part. We assume that it is a question of access to technology.

Further analysis of the available dataset of the URLs of the European universities and research institutes and their linkage behaviour is represented in the EU project EICSTES.

References

Aguillo, I. F. (2002). Cybermetrics, Definitions and Methods for an Emerging Discipline. Séminaires de l'ADEST, Paris, 14 February 2002. www.upmf-grenoble.fr/adept/seminaires/ISIDRO/Cybermetrics.ppt.

Barabási, A. L. (2002). Linked. How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life. *A Plume Book*.

Faloutsos, M, Faloutsos, P and Faloutsos, Ch. (1999). On Power-Law Relationships of the Internet Topology *Computer Communication Review*, **29**, 251

Faust, Katherine - Wassermann, Stanley (1994): Social Network Analysis: Methods and Applications. *Cambridge University Press*.

Kleinberg J.,(1998). Authoritative Sources in a Hyperlinked Environment, *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*.

Kopcsa, A., Schiebel, E. (1998). Science and Technology Mapping. A New Iteration Model for Representing Multidimensional Relationships. *Journal of the American Society for Information Science (JASIS)*, **49**, 1, 7-17.

Newman, M. E. J. (2003). The Structure and Function of Complex Networks, *lanl.arXiv.org e-Print archive mirror*, March 2003