6e Séminaire SciScI - OST
Vendredi 12 octobre 2012
Isidro F. Aguillo (CSIC, Cybermetrics Lab), Espagne

"Information metrics (iMetrics): New Developments"

RESUME

Bibliometrics still play a very important role in information metrics, though the situation has been changing in the recent years: we now have relevant and free citation databases and a new generation of indicators. There is maturity in the area, but problems remain. Perhaps one of the most important is the freshness of data. Now everybody wants to know the current situation, but the data we are collecting is generally two or three years old. We need a citation analytic window that is less than one year old. The other thing is bibliometric data is regarding only a close environment formed by the academics, university professors, scholars and scientists. We need to know the impact on the rest of society. Non-academic sectors are very important as additional stakeholders. If we intend to better explain the motivations behind impact and visibility of scientific activities, we need to build a better theory that explains the resource for traditional scientific citations and non-scientific linking in the environment of Academia and research centers.

I will focus this presentation on the emerging iMetrics sub-disciplines, which are only two or three years old. Their main purpose is to explore larger and more diverse audiences, even in the scientific field. It is for example crucial to take into consideration not only your peers in Western countries but also the new actors in Asia, in Latin America, in Africa. Bibliometrics currently focus on formal communication (papers, publications in the international or high-impact journals), using two main indicators, the number of papers and citations. We are now recognizing the role and impact of the informal communication and its new processes. We should combine all of them and build a new portfolio of indicators. Here is the key of my presentation: our current view of the iMetrics of the new century: traditional bibliometrics and link analysis, and the emerging webometrics, altmetrics (the mention-analysis based manifesto is only two or three years old), and usagemetrics (analyzing the consumption of scientific information available on the web), that is so new there are very few contributions in this most promising area. We are talking about possible hundreds of variables involved, especially if we have access to the academic websites.

I'm going to focus on profiling, a new development of bibliometrics, and talk about two new free to use citation databases: Google Scholar and Microsoft Academic Search.

WoS and Scopus have approximately 14 million records, whereas Scholar has over one hundred million. Academic Search data may be more reliable, as it is closer to the source than what was accepted for Wos and Scopus. Google Scholar is collecting data from far different sources and adding information every day, but there is a serious problem of quality control. For me, the most interesting is the two profiling systems those two giants developed. Google Scholar Citations is made on a voluntary basis: you need to register to create a profile that is automatically populated. MS Academic Search generates profiles you can tune up if you register. The total number of profiles in GSC is about 100 000 researchers, and the total corpus of Academic Search is 20 million (authors and institutions). The MS Academic Search alternative has a lot of noise, but the presentation is richer: you have graphs, information about coauthors, the topics, the journals he was published in, as well as the relationship between the researcher and other scientists. The topics are provided automatically.

The web is already the main scholarly communication tool. We are considering huge audiences, the formal and informal information is rich and diverse (blogs, social networks, even Wikipedia). It takes into account not only the close environment of scientists, but also people benefiting, using, or exploiting the results of research (economical and industrial sectors for instance), but also the implications for policy; not only the technological transfer, but the knowledge transfer. On the web, you have access to the people on the other side of this process. We are going to try and use webometrics as a complementary tool to bibliometrics. Our most important advantage is we can consider commercial search engines as a source. The main tool for those huge public databases (about 800 billion webpages and billions of links only in Europe) is link analysis (richer, updated, more diverse, but also more difficult to interpret). From the point of view of authors and content producers it is a universal tool that is very easy to edit and cheap to access. There are also major disadvantages. We need a theory related to the network organization of the academic world and new epistemological developments. We also have problems regarding the lack of articulation of the network theory in STS. On the practical side, public research engines aren't easy to crosscheck since they are based on secret algorithms. The geographical coverage is wildly uneven, geographical and linguistic biases are strong, and open data sometimes provide inconsistent or irregular results since there is no quality control regarding the contents. We develop strategies to try and solve this.

We now have new indicators of activity — webometrics use webpages, rich PDF files, media files, profiles, all kinds of documents. For building networks, you can acknowledge and use co-authorship, co-citations, bibliographic coupling and co-words as well as co-links and co-outlinks, mentions, coined interlinks etc. Usage measurement is only starting to develop. You can develop composite indicators. You know the most important rankings of universities are based on bibliometric plus survey data. We are also using the Web to develop rankings.

Regarding link analysis, we can evaluate visibility by counting the links via backlists, inlinks and outlinks, explore link motivations through anchor analysis and link rot, build networks, or develop your own tools and use your own sources. There are three main providers of link data. The American system and British Majestic SEO both provide easy and free access to basic data. Ahrefs comes from Ukraine (a very interesting new actor in this area), and shows the distribution of links according to the top-level referring TLDs.

Altmetrics were introduced by Jason Priem, a very young researcher at the University of North Carolina. It has been a great success because everybody knows that today informal communication on the Web is increasingly important. Altmetrics are especially taking into account mentions in the Web 2.0; the most relevant for academic purposes is social bookmarking, Social Peer-review and Social sharing. The tool for making altmetrics is mention analysis. We can use traditional bibliographic units, but there are also new units to use, for example codes. Today, most of the editors are provided individual DOI for each paper they publish, so it is becoming more and more popular. You can also use many other units. “Hot topics” have clearly identified names or key-sentences that reduce the noise of the results. They offer full collections or individual items, mention of events, even email addresses and postal domains, or coded and standardized objects (scientific names, chemistry formulas etc.). The biggest problem and limitation is that Google doesn’t really support Boolean operators. Jason Priem developed the ImpactStory free database. You provide your list of publications and they give indicators about impact on scholars and the general public. A more established proposal is PLoSone, already the largest journal in the world (six thousand papers published in 2011). One of its big contributions is to automatically provide article-level metrics. It is an API that you can customize, if you are an editor, to obtain the same level of reach in statistics environment.

I will conclude with usagemetrics. Unfortunately, this is the newest development and it is hard to obtain data to analyze, but I think it is very promising. You can update data directly from the server (which can be a problem because of privacy issues) and have access to huge log-files that require specialized analytical software. You can use external tools, the most important being Google Analytics. It is the *de facto* standard, because most people are using it. Another way of getting information is being a registered user of specially developed in-house or commercial software, that are mostly private. The American MESUR project is using data from several universities, contracting these electronic libraries to several providers and collecting a lot of information. The tool of that discipline is usage analysis (traffic, global presence of your institution, etc.). There are several options but only one source is really valid and secure, although not totally reliable; it is called *Alexa* and it is strongly biased from a geographical point of view, but it is still possible to make an analysis. You can also count data directly from your profile (number of visitors, number of visits, origin of the visitors, world and national rank. You can see the terms and codes the visitors used to

find you, which is a very rich environment, from a semantic perspective, but also from a practical point of view.

We are going to change the motto of research from “Publish or perish” to “Web-publish or perish”. If we do that, the quality of the tools I introduced here are going to better reflect the true situation. I am proposing a combination of link, mention and usage analysis, used in conjunction with citation analysis to extend (large scale populations), accelerate (immediacy) and deepen (increased diversity of actors and actions) the description of scholarly communication processes. We are working on the reliability of the data, of the data sources. More research is required about the data sources, specially regarding the quality of the information and the way it is obtained, indexed and provided. There are enormous opportunities for working in this area, because we are offering a set of tools that are answering traditional problems that weren’t that easy to address using bibliometrical tools.