The mathematics of ranking

People, objects, and all kind of items are constantly evaluated and ranked in all kind of ways. Books are evaluated on Amazon, papers are evaluated by citations, authors by a g- or h-index, journals by impact factors, web pages by Google's PageRank etc.

The one-day symposium that was jointly organized by the *Belgian Mathematical Society* and the *National Committee for Mathematics* in the prestigious *Royal Academy* in Brussels, on October 15, 2008 wanted to bring together several lecturers who make a scientific study of how ranking can be obtained in a truthful and acceptable way. Analysis of large networks, data mining, collecting the data, and ranking criteria studied in *scientometrics*, or even more general *infometrics*, are hot topics nowadays in which mathematical methods play an increasingly important role. Four of the five lecturers at this symposium were prepared to write up a paper of their lecture which is included in this issue.

The success of Google was to some extent due to PageRank that was able to sort the list of websites according to their importance. Finding this ranking requires the solution of an eigenvalue problem for a matrix of size that runs into the billions.

But the Internet and PageRank is a relatively simple ranking problem. There is just a huge set of sites and a huge set of links between them and all these sites rank each other. It becomes much more complicated when we want to rank the interconnected networks of citations by papers written by authors and published in journals where each of these have some impact on the importance of a particular paper, a journal or an author. The contribution by *Gianna Del Corso* and *F. Romani* illustrates how weighting of the different components can influence the ranking.

However in practice, simpler indices like the h- or g-index are still the most widely used indicators for measuring the quality of research. However to compute a proper index value based on citations, one has to know the dynamics of the citation process. How long does it take for a paper to be widely cited and how long will it last that people will continue citing that paper. The distribution of first citations of a paper, or more generally the dynamics of an information producing process (IPP) is governed by certain exponential laws called Lotkaian laws. Their properties and applications are studied in the paper by *Leo Egghe*.

Networks are not restricted to the ranking of research by citations. We all know e-commerce companies where customers can rate other customers or products. So these earn some reputation thanks to votes of other players in the system, just like in citation analysis. But how can one detect "dishonest" players in an auction system like eBay, or spammers in reviewing databases like MovieLens or Amazon, or trust rates in peer-to-peer systems? So one has to decide not only on the reputation of the objects being rated but also on the credibility of the raters. How should we interpret a vote of country X given to country Y in a European song contest? Answering such question requires to solve a large optimalization problem for a graph. Such methods were considered in the talk by *Paul Van Dooren*. In his paper together with *Catherine Fraikin*,

he discusses another optimalization problem related to detecting similarities in graphs. For example nodes of the graph could be words in a dictionary and the edge (u, v) corresponds to word v being used in the definition of u. The similarity would then rank automatically synonyms.

Whether it is an advantage for the company who wants to sell products or for the consumer to buy things can be a topic of discussion, but everybody who has been surfing the Internet recognizes phrases from recommender systems like "Customers who bought x were also interested in $y_1, y_2, \ldots, y_n$". This may or may not be something you appreciate, but when searching web pages, you may be most happy with a page you "stumbled upon". It may have the answer to the question that you didn't know you were looking for. That page may not be there by accident! The underlying question is: when is a ranking the right one? Thus how should we evaluate the quality of the ranking given by a search engine or a recommender system? The challenges and possible solutions when trying to give an answer to these questions is discussed in the paper by *Bettina Berendt*.

The text of *Hugo Zaragoza*'s lecture is not included here[1]. He dealt also with some of the questions addressed in the other lectures. How do we measure the performance of an information retrieval system? Performance measures are proposed based on probabilities about what the user is really expecting. Optimization of the ranking over several features, PageRank just being one of them, but also the probability of spam, relation with similar queries, etc may all be taken into account. How do we analyze link graphs and detect similarities? But also the analysis of semantics of natural language is an enormous challenge. If the query is for example "Pablo Picasso", what pages is one going to get and how are they related to each other and to the query? Can it be that a page not containing the string "Pablo Picasso" can be more relevant than another page that does?

It is obvious that optimization, statistics, graph theory, and many other mathematical tools are needed, and new algorithms should be designed to deal with the overwhelming amount of information. It will need a lot of engineering and creativity before some of the answers to these questions can be turned into our daily practice.

There were about 60 participants at the symposium mathematicians as well as engineers as computer scientists. This was one of the goals of the symposium, to bring people from different disciplines together to learn from each other. We thank the Royal Flemish Academy of Belgium for Science and the Arts for its hospitality. We thank the Belgian Mathematical Society, and the Scientific Research Network FWO Fundamental Methods and Techniques in Mathematics for the financial support, enabling this organisation.

Adhemar Bultheel and Freddy Dumortier

[1]His lecture, as well as the lectures of the other participants can be found at `www.cs.kuleuven.be/conference/ranking`.