Search engines and bibliographic databases

Today, the internet is the biggest information storage medium in the world. Access to all this information is a great challenge. Search engines and databases are efficient tools for searching and finding this information. In the following we will discuss the basics of search engines and databases, especially in mathematics.

Wolfram Sperber

To look for a needle in a haystack - this metaphor is a synonym for the difficulty and expense to find a certain object. But the internet is much greater challenge: It is huge, and the information is heterogeneous in content, form, structure. Moreover, not always does the user know exactly her/his information needs, sometimes he/she has only vague ideas of what she/he is looking for.

Great challenges need new solutions. The only key technology for better information retrieval is – as in high-tech – mathematics. Methods of structuring and automatically processing information are based on mathematical concepts and methods. In this article, we will discuss some aspects of databases and search engines, especially in the field of mathematics.

Let us start with a short review of information retrieval.

A brief historical review of information retrieval

Information retrieval is an old problem. It has its origin in the development of written languages. The first permanent storage medium for information outside the human brain was stone followed by papyrus. Information management formed the basis of the development and the advance of mankind. More and more information was collected and stored. The papyrus rolls were stored at special institutions, the “libraries”, e.g., the famous library of Alexandria. Egyptians and Romans invented a first method for a systematic analysis and access to the information: tags on the papyrus rolls covering a short summary of the content. This was the pre-form of cataloguing.

Aristotle developed the theory of classification and proposed first classification schemes as an important way to systemize knowledge of a special topic. Also today, classification is a standard technique to learn and structure knowledge.

Librarians took over the part of cataloguing and classification of literature. With the rapid development of industry and sciences the production of scientific documents increased dramatically. Specialized journals were founded, so also the "Annales de mathématiques pures et appliquées", the first scientific journal in mathematics. 1868 the world’s mathematical production covered already 889 publications. The dramatic increase of scientific publications was the reason for developing new information services for a fast and easy orientation about the progress in mathematics, e.g., the "Jahrbuch über die Fortschritte der Mathematik" and later the "Zentralblatt für Mathematik". The idea: A systematic catalogue (completely) covering all publications in mathematics. The catalogue reflects the development of mathematics. The entries in the catalogue cover all bibliographic data of the publications enriched by qualified reviews about the content and possibly an evaluation of the work done by mathematicians. Today, more than 100,000 mathematical journal articles and books are published per year. Today, the reviewing journals – meanwhile in form of an online database – are also an important, high-quality and reliable tool for the mathematical community for searching and finding relevant information for research in mathematics and applied fields. And they are an impressive example of the international collaboration of mathematicians.

The Internet also brought new challenges for knowledge management.

- The Web is the biggest information storage medium.
- New document types have been developed, in mathematics, for example, software, simulations, digital course materials, etc.
- The Internet is decentralized. There is no central control. Every person equipped with a computer having access to the internet can create and publish documents on the Web. Each author decides about the format and the structure of her/his documents.
- There is no quality control.
- Different formats are used. There are no generally accepted standards for structuring and content analysis of documents.
- The information on the Internet is dynamic. Documents can easily be created, removed or transferred to any other location.
- One reason for the triumphant advance of the Internet is hyperlinking between documents as a simple method of associating facts and statements on different Web pages.

The design and architecture of search engines

All search engines have a similar architecture:
1. Collecting documents (Web sites)
2. Indexing
3. Querying module

The crawlers or gatherers are responsible for collecting Web pages. Starting from a seed list of given URLs they analyze the Web sites. The gathered pages are in different formats.
The gatherers extract some information about the Web pages, e.g. the format, the date of creation, and the hyperlinks within a Web page. The destination anchors of the hyperlinks of a Web site are added to the seed list (topic-specific search engines need criteria to decide if a Web page is relevant or not).

Indexing means to evaluate the content of a Web page, e.g., a text is transformed in a list of words. Special methods for the extraction and analysis of the information are important for the quality of the search engines. The extracted information of the Web pages is stored in the index together with a link to the original document.

The querying module is the interface of the search engine for the user. Of course different user scenarios are relevant for searching. One typical scenario: a user is looking for special information but she or he is not a specialist. So, she or he prefers a simple query, e.g. a single word as search term. And she or he expects that the search engine will present a list of relevant results to the query (extracts of the relevant pages). But here the search engine has a problem. A simple search can result in a long list of hits in the index. So we need criteria for the relevance and ranking of the hits.

The classical information retrieval, here we refer to G. Salton, takes as measure the number of relevant terms divided by all terms of a Web site.

Another way was gone by Google. Google started to analyze the link structure to determine a measure for the relevance of a page: The PageRank is a measure of popularity of a Web page. In principle Google's PageRank means the following: The relevance of a page is determined by the number of incoming links and the value (PageRank) of the anchors of the links. Google used the PageRank to sort the hits in a new way. The better ranking was the basis for the phenomenal success of Google.

The roots of Google's PageRank lie in the bibliometric analysis of the scientific literature.

About citation analysis and impact factor

The origin of bibliometry is the evaluation of journals by statistical means. For this purpose the citations of publications of a journal are analyzed.
The measure for the relevance of a journal is the impact factor. The impact factor of a journal in a year is defined as the quotient of the number of all citations of the articles in the two preceding years and the number of all articles in the journal that year. A journal, where the citation of each article is 1 in average, has also the impact factor 1. The ranking of a journal is increasing with increasing impact factor.
The impact factor analyzes the directed graph which is given by the articles of a journal as nodes and the citations as edges.

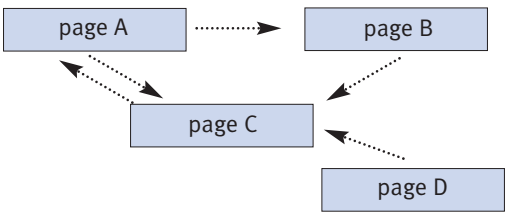
The "Institute of Scientific Information" (ISI) started to calculate the impact factors of more than 6,000 scientific journals in the sixties. The results are published in the "Journals Citation Reports" (http://scientific.thompson.com/products/jcr).
The use and importance of the impact factor and the relevance of bibliometric (statistical) methods is discussed controversially in the scientific community. The advantage: It is easy to calculate the impact factors (and other bibliometric measures). Also non-specialists can interpret the results and get an impression of the relevance of a journal (or publication). On the other hand bibliometric methods only analyze some formal data of publications, especially the citations. They cannot evaluate the content of a publication. Bibliographic methods are a measure for the popularity and the influence of a journal (publication).
Gabriel Pinski and Francis Narin have defined a recursive impact factor. The recursive impact factor assigns citations in journals with a high impact factor a higher weight than citations in journals with low impact factor, for more information see the Wikipedia page on the impact factor http://en.wikipedia.org/wikipedia/impact_factor.
The recursive impact factor is the immediate origin of Google's PageRank.

Google's PageRank – Tell me who speaks about you, I'll tell you how important you are

Google's PageRank evaluates - in analogy to the citation analysis of scientific journals - the directed graph which is defined by the hyperlinks of the Web.

The weighted links to a Web site are a measure for the popularity (= relevance) of a Web page. The weights are given by the relevance of the anchor Web pages.

We demonstrate Google's model for a simple example. Let us look at the graph consisting of only four given Web sites with the following link structure (the arrows mark hyperlinks from the Web page i to the Web page j)



The links are given in the following table

	A	B	C	D	(source anchor of the link Links)
A	-	-	x	-	
B	x	-	-	-	
C	x	x	-	x	
D	-	-	-	-	

(destination anchor of the link)

Based on this table we construct a matrix by the following method:
o is the value of a[i,j] if there is no link from the Web page i to the Web page j, 1/|.| is the value of the element of the matrix from if there is a link from the Web page i to the Web page j with |.| the number of all links going out from the page i (So we have |B| = |C| = |D| =1 and |A| = 2). In other words: The weight of a page is uniformly distributed to all outgoing links of a Web page.

So we come to the following matrix representation of a linear system of equations

$$\begin{bmatrix} p(A) \\ p(B) \\ p(C) \\ p(D) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1/2 & 0 & 0 & 0 \\ 1/2 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} p(A) \\ p(B) \\ p(C) \\ p(D) \end{bmatrix}$$

where p(.) denotes the PageRank of a Web page.

As said above the system of equations for the whole Web is huge. Some people believe that Google's model of the Web is the largest system of equations solved until now.

Page and Brin, the inventors of the PageRank, experimented with the iteration algorithm to solve this system of equations.

Some remarks:

1. **First simulations and tests**

First numerical simulations were encouraging. But the simulations had led also to some problems and open questions with the iteration process:

- Convergence of the iteration process
  So, the solution can lead to cycles and the iteration procedure doesn't converge (e.g., look at a graph of two Web pages where each page is hyperlinked with the other page).

- Dangling nodes:
  What is the PageRank of dangling nodes (Web pages with no outgoing links)?
- Rank sinks:
  The whole PageRank is assigned to some Web pages, other Web pages get the PageRank 0.
- Sensitivity:
  What about the dependency of the iterations from the starting point?
  What about convergence speed?
  Numerical problems and simulation are enough reasons to reflect Google's approach and model.

2. Modification of the mathematical model.
Page and Brin have discussed the behaviour of a surfer in the Web in more detail. A surfer is starting his walk through the Web on any Web page. She/he is clicking randomly on a link and thus switches to another Web page. And the same process continues. But what happens if a chosen Web page doesn't contain any link? Then we can go back to a former Web page, or the surfer starts a new walk with an URL from the bookmark list.
This typical behaviour of Web surfing should be inserted in the model. This is possible by a modification (extension) of the coefficient matrix of the model. Instead of the matrix above, Page and Brin used

$$G = \alpha \times S + (1-\alpha)\ 1/n\ \mathbf{e}\ \mathbf{e}^T$$

where S is the matrix of hyperlinks introduced above. G is called Google matrix in the following.
For $\alpha = 1$ we get the system of equations (1).
For $\alpha = 0$ we have a stationary process. The original PageRank of a Web page is not changed by iteration.
In principle α can be an arbitrary value in the interval (0,1), but numerical simulations have shown that $\alpha = 0.85$ is a good choice.

3. Properties of the modified model
The modified model has some nice properties. Such a kind of matrices are well-known from the theory of Markov chains. They are used to model special stochastic processes, namely the transition between states, the situation as given for surfing on the Web. Such models characterize a process without memory. It is interesting that Page and Brin did not use the theory of Markov chains in their model.

4. Mathematics behind Google
The Google matrix (for $\alpha \neq 1$) is a matrix with nonnegative coefficients. Such mathematical objects were investigated in the Perron-Frobenius theory. It can be shown that the Google matrix is irreducible and primitive. These properties guarantee the fast convergence of the iteration process.
For details, see the book of Langville and Meyer, 2006. For further publications on the topic PageRank check our database ZMATH (www.zentralblatt-math.org/zmath/en) and look for "PageRank" or "Google".

5. Computation of PageRank
Google does indeed use the iteration process to calculate the PageRank. The huge dimension of this linear system of equations is a big challenge for computing.

6. What measures the PageRank?
The measure for the evaluation of a Web page is – analogously to the impact factor in the citation analysis – the popularity of a Web page. More links to a Web page increase the PageRank.

7. Google's success
Google's PageRank has dramatically changed the search on the Internet. Google, a typo of Googol, is today a synonym for the Web search. Google has paved the way for analyzing huge sets of information by automatic (mathematical) methods and means. The new quality of ranking made Google the most popular search engine within just a few months. As we said before, Google is based on mathematics. The success of Google is also a success for mathematics.

Google and ZMATH

What are the perspectives of searching the Web? Does Google make bibliographic databases superfluous? Of course, there is no definite answer to this question but there are some reasons against such a scenario. Google analyzes neither the content nor the quality of the information. Google's aim is to be the best universal search engine, a search engine which can be used intuitively and which provides the best information.

- Google is a commercial company and has to make profit. Small markets like mathematical information are not very interesting from a commercial point of view.

- Mathematics has developed an own language. During the centuries mathematics has developed an own vocabulary and is formula-based. The notation of mathematics in the Web requires special formats, e.g., TeX and MathML.
- Google's retrieval is text-based. Up to now Google does not develop special retrieval methods for single scientific disciplines.
- Google's PageRank is only one aspect of content analysis. Further aspects must be integrated into the sophisticated search engines of the future.
- The driving force behind mathematical information and communication is the mathematical community. Up to now mathematical information and communication is strongly involved and developed by people and institutions; examples are the development of TeX and MathML, the development and updating of mathematical classification schemes, the development and the operating of specialized services, the development and discussion about requirements to mathematical information and communication in the future, the concept of a Digital Mathematical Library, etc.
- Google is restricted to information which is available on the Web. This is not the case for all mathematical knowledge published up to now (but of course the percentage of mathematical publications which are also available over the Web is increasing).

Publications from mathematics and applications are the core of the database ZMATH. The database ZMATH almost exhaustively lists the mathematically relevant publications. Completeness is the first criterion for ZMATH.
The mathematical publications are articles from mathematical journals, lectures of conferences or books. The content of ZMATH is quality-proof information.
The entries within the databases ZMATH starts with a comprehensive content analysis. Processing of mathematical publications in ZMATH means

- to analyze the bibliographic data of a publication (authors including an author's identification, title (original and possibly translated titles), source, publication year, language of the publication
- to describe the mathematical content, e.g. a classification of the mathematical subjects according to the Mathematical Subject Classification (MSC), an overview of the content and the results of the publication given by a review of an independent expert or the abstract of the publication
- linking to full texts or document delivery services.

The retrieval functionalities within ZMATH are a further important aspect of ZMATH. The quality of ZMATH consists of the quality of its entries plus the quality of the content analysis plus the quality of retrieval functionalities.
Also the methods of creating and updating the database ZMATH are under permanent change. More and more automatic procedures will be used, e.g., author identification. New automatic methods and procedures are discussed and checked, e.g., automatic classification. As we said above, automatic processing of information means the use of mathematical methods. So, mathematics and especially ZMATH have the potential to develop new and better methods for information retrieval in mathematics and sciences.

In my opinion the question is not: "Google versus ZMATH". I think that the future will be "Google and ZMATH" (where ZMATH stands for specialized services). ZMATH is a special filter for mathematical information which refines and complements the information retrieval of universal search engines. It could also be in the interest of Google to cooperate with topic-specific information services. A cooperation between universal and topic-specific search engines/databases would generally improve information retrieval on the Web.

Both search engines and such databases are part of the dynamic development of intelligent, powerful and user-friendly information services.

References

Brin, S., & Page, L. (1998). The anatomy of a large hypertextual web engine. Computer Networks and ISDN Systems , S. 107 - 117.

Brin, S., Page, L., Motwami, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Stanford: Stanford University, Computer Science Department.

Langville, A. N., & Meyer, C. D. (2006). Google's PageRank and Beyond TheScience of Search Engine Rankings. Princeton and Oxford: Princeton University Press.

Dr. Wolfram Sperber
FIZ Karlsruhe, Mathematics and Computer Science
Zentralblatt MATH