Dr. Makiko Mase
Tokyo Metropolitan University
.
Part of August+September 2018 guest at the
School of Business Informatics and Mathematics
University Mannheim

Block Seminar in September 2018

Linear Algebra in Search Engines

The seminar will be held in english. It will follow the book

[LM06] A.N. Langville, C.D. Meyer: Google's PageRank and beyond. The science of search engine rankings. Princeton University Press, 2006.

The seminar talks will report on the central chapters 4-11 in this book, with mathematical insertions from the collection of mathematics in chapter 15. The main tools are linear algebra and (some modest part of the theory of) Markov chains.

The seminar will take place on

Friday, 07.09.2018, all day, and Friday, 14.09.2018, all day.

Here *all day* means four talks of 90 minutes, probably 8:30-10:00, 10:15-11:45, 12:45-14:15, 14:30-16:00.

The talks can be held alone or shared by up to two people. Therefore the number of participating students should be between 8 and 16.

The seminar aims at students in the bachelor *Mathematics in Business and Economics* and in the bachelor *Education Mathematics*, but it is also open for master students in these areas.

My home is the Tokyo Metropolitan University. I will visit the WIM faculty part of August+September 2018. My host is Prof. Dr. C. Hertling. He will hold a

Preparation and information session with distribution of talks

Tuesday, 15.05.2018, 15:30-16:30, A5, C015.

If you miss this meeting and are still interested in a talk, please contact Prof. Hertling, hertling@math.uni-mannheim.de .

In the early 90ies, the World Wide Web consisted already of a huge amount of web pages and contained a lot of informations of all kinds. Though it was not so easy to search for useful sources for specific questions. One had to know or guess his way to suitable web pages. One would need advice of colleagues or friends, or one had to scroll through thousands of not well ranked web pages which early search engines would return.

This changed with the Google search engine, which was created in 1995 by Sergey Brin and Larry Page. As doctoral students at Stanford University, they started their business at their dormitory rooms. They had clever business ideas which in connection with the search engine made them rich and Google huge. But the heart of their search engine was an idea how to use the hyperlink structure of the WWW in order to give each web page a *score*. It arises as a combination of a *content score* and a *popularity score*. The content score depends on the user's question. The popularity score is computed (probably?) once a month for billions of web pages together and is then stored and available for each user's question. This score is also called *PageRank*. The book [LM06] and this seminar are mainly abot the PageRank.

The PageRank of a web page comes from the page ranks of the pages which link to the given page. The basic formula is (4.1.1) in [LM06]. Though modifications are necessary. Chapter 4 in [LM06] gives the basic ideas. The chapters 5 to 10 in [LM06] refine these ideas in various ways.
Chapter 11 presents an alternative method, invented by Jon Kleinberg (later professor at Cornell University) in 1998. He also used the hyperlink structure of the WWW, but he defined two scores for each web page, an *authority score* and a *hub score*. The authority score of a page comes from the hub scores of all pages linking to the given page, and the hub score comes from the authority scores of all pages to which a given page links (so it is easy to get a high hub score), see the equations (11.1.1) in [LM06]. The chapters 12, 13 and 14 in [LM06] give further interesting information, but not of a mathematical nature, namely on other link methods for ranking web pages, on the future of search engines, and some useful web pages about web information retrieval. The chapters 1, 2 and 3 put the PageRank method of Google into the context of the whole search engine. They talk about the history of search engines, the way how to become aware of the billions of web pages (crawling) and the basic idea of ranking web pages by popularity.

The long (48 pages) chapter 15 is an excellent source for the mathematics used in the book. It is mainly about the relevant linear algebra and the relevant part of the theory of Markov chains. Though one can read the chapter 1 to 14 essentially without having digested chapter 15. If a notion is put in italic, this says that chapter 15 contains information on it. This allows the chapters to be read by interested laymen (ignoring the notions in italic) and by mathematicians (looking up notions in italic if they want).

For the seminar, this means that the majority of the talks can essentially follow the presentation in the chapters 4 to 11, with sometimes using material from chapter 15. Only two talks, those about the chapters 5 and 7 will have to present material from chapter 15 in a systematic way, as the chapters 5 and 7 are very short.

The seminar has several aims.
(1) One is that the participants give a good talk and during preparation learn, how to achieve this. This means that one has to digest the material well (which requires to be able to read it and understand it), to choose well what to tell in detail and what not, and how to tell it. The talks shall take 90 minutes. Longer is forbidden absolutely, but much shorter is also bad. There is definitely for each talk enough material to fill 90 minutes.
(2) All participants shall learn from all talks (not only their own one). It is good to prepare also for the other talks, by reading the relevant chapter. Doing that one could note some good questions which one can then pose during the talk if they are not answered anyway in the talk.
The second aim requires presence at all talks.

The following talks are proposed, the first four on Friday 07.09.2018, the second four on Friday 14.09.2018, at the times 8:30-10:00, 10:15-11:45, 12:45-14:15, 14:30-16:00.

Talk 1:
The mathematics of Google's PageRank.
Chapter 4. (And please glance at chapter 15.)

Talk 2:
Parameters in the PageRank model.
Chapter 5 and Theorem 6.1.1 with proof, and for the proof material in chapter 15 on the Perron-Frobenius theorem, that means all or a part of the material in section 15.2.

Talk 3:
The sensitivity of PageRank.
Chapter 6 without Theorem 6.1.1 and its proof.

Talk 4:
The PageRank problem as a linear system.
Chapter 7 and the material on M-matrices on the pages 166 and 167 at the end of section 15.1. As this is probably not enough material for 90 minutes, a second part of the talk could report on (parts of) the material in section 15.3 on Markov chains. This would not be closely related to the first part, but it would be useful for the whole seminar.

Talk 5:
Issues in large-scale implementation of PageRank.
Chapter 8. (And please glance at chapter 15.)

Talk 6:
Accelerating the computation of PageRank.
Chapter 9. (And please glance at chapter 15.)

Talk 7:
Updating the PageRank vector.
Chapter 10. (And please glance at chapter 15.)

Talk 8:
The HITS method for ranking webpages.
Chapter 11. (And please glance at chapter 15.)