Self-learning search engines

Schuth, A.

Published in:
Amsterdam Science

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Self-learning search engines

- How does a search engine such as Google know which search results to display? There are many comparing algorithms that generate search results, but what makes Google special? We developed a new probabilistic method for quickly comparing large numbers of search algorithms by examining the results users click on. Our study was presented at SIGIR 2015, the leading international conference on information retrieval, held in Santiago (Chile) last summer.

Interleaving

Developers of web search engines constantly create hundreds of alternative search algorithms, all of which aim to find the best possible match between a user’s information need and web pages. It is vital for both the search engine and the user to know which of these algorithms produces the best results. A new way to compare search algorithms is through interleaving, a search engine technique where the search engine analyses the users’ click behaviour to determine a preference between two alternative algorithms. After the user has typed in a query, the unique results of two search algorithms (blue and red in the Figure) are interleaved alternately (from top to bottom) and displayed to the user as a single list. If the user then clicks on a result found by one search algorithm (red), the algorithm analysis infers that in this particular case the algorithm generating the selected result produces better results than the other one. By scaling up this type of inference to cover millions of users, the search engine automatically learns which algorithms yield the best results.

Multileaving

Multileaving is, however, limited by the fact that only two algorithms can be compared at a time, and a large number of combinations may therefore be required to determine which one of hundreds of existing algorithms really works the most effectively. So-called multileaving methods, which have been developed at the University of Amsterdam, allow multiple algorithms to be compared simultaneously. In earlier work, we did so by combining the results from many lists of results at once (in the example of blue and red lists, imagine also adding orange and green lists, etc.). The multileaved result list that is shown to the user is then a mix of results originating from many search algorithms - a multicoloured list. We keep track of where each of the results came from (their colour), and, as with interleaving, we observe which search algorithm (colour) attracts most clicks from users. Again, the search algorithm that receives most clicks wins. Typical of this method has been established, the search engine will completely switch over to the victorious search algorithm for all its users and queries.

Next step: probabilistic multileaving

Our newest method takes multileaving a step further. While we still combine the results from many search algorithms into a single multileaved result list, we do so probabilistically. Instead of alternatingly picking results from each of the lists, always working from the top-ranked downwards, we now define a (high) probability that the top-ranked result is picked, leaving a non-zero probability that a lower-ranked result is selected instead. By making the multileaved list probabilistic, we ensure that any combination of search algorithms (coloured lists) could have resulted in the multileaved result list that is shown to a user. This has the major advantage that we can retroactively evaluate any search algorithm, using a multileaved result list that has already been shown to a user. In other words, it now becomes possible to reuse old combination of multileaved result lists and users’ clicks to keep evaluating new search algorithms. As can be expected, the search algorithms that originally contributed results to the multileaved result list, or algorithms that are very similar, can be evaluated with higher confidence than very different search algorithms. However, even working at lower confidence levels, it is a major advantage of our probabilistic multileaving method that new search algorithms that were not even invented when the multileaving took place can be evaluated retrospectively. This way, our method can identify the best search algorithms much faster, enabling search engines such as Google to self-improve much more efficiently.

In the early 1900s, it was well known that the same fossils could be found in the rocks of different continents. At first, geologists believed that continents were fixed, but Alfred Wegener, a Prussian meteorologist, proposed that the continents must have shifted. Wegener wrote, “South America must have lain alongside Africa and formed a unified block.” Before their separation, the continents were fused together in what he named the ‘Urkontinent’ – now known as Pangaea. “The parts must have become increasingly separated over a period of millions of years,” Wegener wrote. He suggested that the ‘Urkontinent’ was pulled apart by the centrifugal force from the Earth’s rotation and that the continents drifted apart, with the distance between them increasing over time. Wegener argued that the ‘Urkontinent’ – now known as Pangaea – must have been a single landmass, split apart, leaving a non-zero probability that the top-ranked result is picked, leaving a non-zero probability that a lower-ranked result is selected instead. By making the multileaved list probabilistic, we ensure that any combination of search algorithms (coloured lists) could have resulted in the multileaved result list that is shown to a user. This has the major advantage that we can retroactively evaluate any search algorithm, using a multileaved result list that has already been shown to a user. In other words, it now becomes possible to reuse old combination of multileaved result lists and users’ clicks to keep evaluating new search algorithms. As can be expected, the search algorithms that originally contributed results to the multileaved result list, or algorithms that are very similar, can be evaluated with higher confidence than very different search algorithms. However, even working at lower confidence levels, it is a major advantage of our probabilistic multileaving method that new search algorithms that were not even invented when the multileaving took place can be evaluated retrospectively. This way, our method can identify the best search algorithms much faster, enabling search engines such as Google to self-improve much more efficiently.

In 1915, Wegener published his highly influential paper “The Origin of Continents and Oceans”, which was greeted with great scepticism. The mechanism involving centrifugal forces proved erroneous, and in the end it took until the 1970s for the theory of continental drift (plate tectonics) to be accepted. Mapping the topology of the ocean floor, the geologist Marie Tharp (1920 - 2006) discovered a chain of mountains splitting the large ocean basins in two. Tharp and the geologist Bruce Heezen (1944 - 1977) recognized that the mid-ocean ridges were lines along which the oceanic crust was split apart, creating the continents we walk on today as stone drafts, drifting through the oceans as fast as our nails grow.

The movement of magma beneath the Earth’s crust continuously pushes the continents apart, creating new land bridges. The magma flows up and down cyclically due to radioactive heating in the Earth’s core. Magna flowing outward from the mantle is diverted horizontally when it meets the crust, and this exerts extensional stress, pulling the crust apart in so-called rift zones. The Mid-Atlantic Ridge is a typical rift zone, and in these locations, volcanic eruptions continually add magma to the diverting crust, pushing both sides apart, like slowly peeling apart two pieces of paper. By measuring the electromagnetic signal in the crust parallel to the ridge – imprinted by the Earth’s electromagnetic field in magnetic minerals at the moment the magma cooled down to temperatures below 580°C – we can estimate when each portion of crust was added. The fact that the estimated age of the crust scales with its remoteness from the rift zone confirms the tectonic picture as being correct. Reconstructions suggest that it was around 300 million years ago when Wegener’s ‘Urkontinent’ split.

The second edition of Wegener’s book, in German, is freely available at http://www.gutenberg.org/ebooks/45460

Reference

A. Wegener, Die Origin of Continents and Oceans, translated from German by Brian Brown (New York, pub. 1968).