So you want to build a search engine?

Benji Altman

May 5, 2018

Contents

Abstract

There are many different types of search engines, the most complicated of which are where you do not initially have knowledge of what is being searched and that information may change. In order to tackle this we need some way of discovering possible results, matching these results to a query and ranking possible results.

1 Introduction

So you want to build a search engine? Now there are a couple different types of search engines and not all search engines need work the same, however there are two steps that are absolutely necessary for any searching to make sense. First you are going to need to have some set A to search through and second you will need a search function $s : Q \to \mathcal{P}(A)$[1] where Q is the set of all possible search queries. If we step away from mathematical formalities this makes sense as we simply are finding a way to match a search query to set of possibly results.

Now what one might notice is that none of this talks about ranking the results, and this is obviously an important part of any modern search engine. The problem of ordering may, in fact, be the most interesting part of a search engine. This paper however will spend very little time talking about ordering a search result as Page Rank is already well understood and after that most of the complexity really has more to do with understanding the query which we will cover when talking about our search function.

Now this paper will mostly focus on the discovery of set A and assumes and Internet like structure to A. That is to say that A is a large set with a directed graph structure to it. Additionally A is not static, edges in A's graph structure may change, members of A may be created or destroyed, as may the content of those members. Finally A is not initially known to us.

Now while I try to keep this paper in the abstract and I talk about A as if it's a set and does not have to be the Internet, I think it's important to talk about some of the tools that we have built to help us work with constrains the Internet put on us. This is because the Internet is so well known and what most people have experience searching. Leaving out practical information, while it may take away from the abstractness, it is the authors opinion that it is to0 important to leave out. In this vein we will be focusing on Google's architecture as it is so well known and information is readily available on it.[2]

2

The discovery of A is done through a process called web crawling. We will

[1] $\mathcal{P}(A)$ is the power set of A.

[2] This is not fully accurate, there is some very good information on Google's inner workings that comes from before 2000, however much of of modern Google's inner workings are kept as trade secrets or simply not released.