Topic: Searching engine

Student : Shu Yang

BU id: 03134397

Email: conlany@bu.edu

**Description:**

The 21 century is a century of the internet, with billions of people uploading, editing, referring to data, a huge deluge of data shall submerge the content you want to get without a useful searching tool. There are mainly three parts in the whole process of document searching, they are inquiry service, page searching and pre-processing, afterwards we can show the documents to the customer in queue.

There main tasks of the three phases are as follows: Firstly, we need to correct the question given by the customer, not just because they can make spelling mistakes, but also they often describes the question unsharply. For instance, they may type in “Boston is the province of which country?” However, Boston is not a province but a city, so we should replace the key word province by city. Secondly, after we got precise key words, we need to use a fast and stable tool to help searching the database provided by advanced companies or the one we created by ourselves. And thirdly, when having the access to the storage, we need to filter the documents, and make a good ranking towards the content we considered proper then present them to the customer.

**Technology route:**

In this part, we will study the mysterious procedures behind the simple and easy type-in then search operation, digging out the exact things happened then we will analyze the possibility of realize all the functions and chose the proper method and algorithm to achieve our goal.

When it comes to the first phase we mentioned above, we use the technology called inquiry editing. We use editing distance to solve the misspelling problem. When a misspelling occurs, we cannot find the word in our dictionary. Then we shall compare all the letters in this word and try to find the word which in our dictionary while having most letters alike that of the customer typed. It can be harder to solve the problem if customer makes the mistake like the city or province occasion we mentioned above, normally it can be solved using the machine learning skills. LSTM and GAN algorithms are advised, LSTM have the memory and forget module, which is similar to the language function area of our brain, by train the weight of different kinds of individual words to be remembered or forgotten, the machine can have a better understanding of the entire sentence or even article. And GAN is proved to have the best performance in most cases of creating, which means we can do the translation work, like translate English to another language, to translate the question we received to something easier for the engine to understand and search.

A good searching tool can also get the key words in your sentence. key word exist because the frequency of the appearing of the data that have some relations to the individual word in your sentence is different, so some words may have stronger links directs to more documents. The nouns may be more important than prepositions in linking to the document. On the other hand, key words do not necessarily be the ones appears a lot in our database. For instance, you type in “how can I know the important functions of quantum mechanics when I just started learning it” There are some words, like “ I, know, just started learning” have a lot of relevant topics, but will be little relevance with quantum mechanics, then these results can be useless.

The second thing is that how to get access to the database with minimum cost of time. The first thing is to build a database, which should have some basic models to protect itself from being attacked by hackers. IT should also provide a accessible port to allow the users to upload some materials to the database,which will definitely enrich the content we here have in the database. In order to protect our database from being polluted by some uncomfortable content, we may include a machine learning part to help distinguish the bad from the good. We also need a security part to make sure the we will not be attached by crawlers. But due to the limitation of time and fund, we shall also use the database provided by the advanced searching machine companies to ensure we will not be disturbed by the limitation of our amount of documents we possess.

If we want to be authorized to used the database online, we have to use the crawler. Like the human clients do, they search and browsers the entire net with minimum cost of time restlessly. Advanced companies like Google will also use crawlers, but the number of theirs shall be trillions, which help them to upload, download and supervise the entire net. Fast and low cost as it is, burdens like zombie crawlers or evil crawlers may also slow down the service, harm the security or even cause a collapse. That why most websites uses the anti-crawler. We choose to use Selenium library to have access. Selenium provided functions to operate the driver of the browser, which let crawler have the talent of acting alike human to fool the anti-crawler mechanism.

Fast as the crawler is, it still takes a long time for a single crawler to download all the documents. So a multiple thread strategy is applied to eliminate that time cost. If we divide the continuous time in to many slices and allocate all of them into many tasks, then we can create a illusion that we can do multiple tasks simultaneously. In doing so, we can start many crawlers and arrange them to download different documents.

When it comes to the preprocessing of the document, there are four steps to achieve the major requirements, they are repetition reduction, data saving key word ranking and indexing.

In order to enlarge the scale of the database, we can’t have complete supervising toward each document, which makes a lot of repeating resources. During the uploading period of each document, we design and assign a fingerprint of each, which includes the basic and major information of it. What we do is to check the relevant documents, and see if their fingerprint are closer enough or are the same, if yes, they will be filtered.

Data saving is the step that may make a difference to the performance of an engine. Although it seems simple to save data and retrieve them, when the amount arouse, thing are different. Key word ranking is the step to tell the key word of the document we found. The algorithms we choose are to be decided.

And indexing is the step to build a strong link between the documents and their key words. We can train the engine by learning from the click operation of the customer. For instance, if a customer searched for something, and he clicked, not the first link we provided, but the one on behind. Then next time, we may consider to put such answer ahead of the queue.

**Method to test the capability of retrieving:**

The paper, An Evaluation Of Retrieval Effectiveness For A Full-Text Document-Retrieval System (DAVID C. BLAIR and M. E. MARON, 01 Mar 1985) had provided a very simple but effective method to help test the performance of a searching algorithm.

In the research, David et al.(1985) chose the IBM STAIRS, or Storage and Information Retrieval System as their platform to carry out their research. The STAIRS was a program providing storage and online free-text search of text data.

The characteristic of the STAIRS are as follows: STAIRS uses full-text retrieve technology. Compared to other retrieve technology, full-text can test the content of a very document not only by check the existence of key words on titles, abstracts, selected sections, or bibliographical references, or in conclusion, metadata or on parts of the original texts represented in databases, the key words hidden in the article can also be checked. STAIRS also provided interface function allowing the researchers not only can easily store the data, but also using the packaging functions to do the searching and ranking of the target document. Although we may not necessarily use the full-text technology as our algorithm, it is still a very classic method used by lot of web pages searching today.

David et al.(1985) chose two factors to evaluate the performance of a searching routine, they are Recall and Precision. The two factors are taken into consideration since Recall represents the percentage of documents we retrieved to all that relevant and Precision stands for the ratio of document that is really useful to all that we retrieved. When taking practical issues in to consideration, we have few skill to help accelerate the process.

When doing the calculating, we need to either find by ourselves the relevant document or to label all the documents their key word precisely ahead of time, the same is to judge the level of relevance of the document searching engine provided. So we need to decrease the database to some extent. A sub database will be constructed.

And when it comes to evaluating, we need some professional volunteers, in each turn of searching , they will decide weather something interests them appears, if not, the searching will start again while the number of spam documents will adds up to decrease the precision.

**Papers to refer:**

1. Dawei Yin†, Yuening Hu†, Jiliang Tang† . Ranking Relevance in Yahoo Search，KDD 2016 . Relevance Science, Yahoo! Inc.
2. <https://github.com/Justin3go/xiu-search>
3. Using Migrating Agents in Designing Web Search Engines and Property Analysis of Available Platforms .Niraj Singhal1 , Ashutosh Dixit2, R. P. Agarwal3 , A. K. Sharma4 1,3Faculty of Electronics, Informatics and Computer Engineering Shobhit University, Meerut, India 2,4Deptartment of Computer Engineering YMCA University of Science and Technology, Faridabad, India

4)An Evaluation Of Retrieval Effectiveness For A Full-Text Document-Retrieval System . DAVID C. BLAIR and M. E. MARON, 01 Mar 1985