A cursory search engine : ASBS

2019201411

July 25, 2020

1 Introduction

ASBS is a in-station search engine. It can be used for search function in some small site, like Info school of RUC. And it works as many pouplar search engine, it gets the input information and find the Top 10 websites which are best matches to display. However, it is built by home server. So you can only used it when you download the whole project and built it in proper order. What's more, it's small-scale. So it may be broken down when you use it to deal with a huge station. **All of the content is only for scientific research.**

2 Details

2.1 Crawl

I use the "wget" command in C++ to download the html of the sites. A disadvantage is that I don't use multithread processing so that I set the sleep time as 1s. And it will take 2h to get all pages belonging to http://info.ruc.edu.cn.

2.2 HTML parser and word-seg

i) I view the files downloaded by wget as some "txt", and I extract the content which has label <p> or <title>.

ii) As to word-seg, I use jieba-py to cut the words.

2.3 Scoring

To choose the best mathes, we need to assign a score to each site. We assume that we have N docs(or html of sites).

- First we define $tf_{t,d}$ as the occur times of term t in doc d.
- And then we define df_t as the number of docs which we can find term t in them.It means if t occurs many times in a doc,it will only be counted once.

Define

$$W_{t,d} = (1 + log_{10}tf_{t,d}) * (log_{10}N/df_t)$$

And for a certain query, we calc $W_{t,q}$ as above. To calc the similarity between the query and the docs, we view query and docs as same-order vectors(W_d, W_q), and the dimension is decided by the num of terms in query. We regard the angle between 2 vevtors as the similarity. Closer they are, we suppose they are more similar. So we use the cosine to calc the final Score.

But when we test a few queries, we find that somtimes it doesn't perform well. Considering that the titles really means a lot, so I give the term in the title more weight.

2.4 UI

- Use flask to communicate between server and user.
- Simple css to build the html.

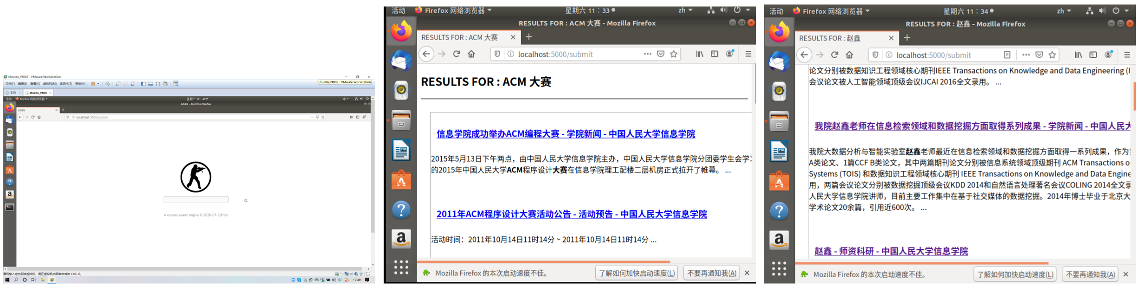
3 Some tricks

Actually they are some problems when coding, I've mentioned them in the presentation. By the way, to strong the key words occur in the abstract in the result-website, I add <strong></strong> around the key word, it's kind of silly but really easy and efficient.

4 Display

Advantages For those news query, ASBS usually get satisfactory results. Because the news report are unique in generally. And if the query is compose of few words, always good answers.

Disadvantages Since I did not set the stop words, so when the input has the similar meaning with the passage but not same or the query has so many key word, then the results may not be that satisfying.



RESULTS FOR : ACM 大赛

信息学院成功举办ACM编程大赛 - 学院新闻 - 中国人民大学信息学院

2015年5月13日下午两点，由中国人民大学信息学院主办，中国人民大学信息学院分团委学生会学
的2015年中国人民大学**ACM**程序设计**大赛**在信息学院理工配楼二层机房正式拉开了帷幕。 ...

2011年ACM程序设计大赛活动公告 - 活动预告 - 中国人民大学信息学院

活动时间：2011年10月14日11时14分 ~ 2011年10月14日11时14分 ...

论文分别被数据知识工程领域核心期刊IEEE Transactions on Knowledge and Data Engineering (I
会议论文被人工智能领域顶级会议IJCAI 2016全文录用。 ...

我院赵鑫老师在信息检索领域和数据挖掘方面取得系列成果 - 学院新闻 - 中国人民大

我院大数据分析与智能实验室**赵鑫**老师最近在信息检索领域和数据挖掘方面取得一系列成果，作为
A类论文、1篇CCF B类论文，其中两篇期刊论文分别被信息系统领域顶级期刊 ACM Transactions o
Systems (TOIS) 和数据知识工程领域核心期刊 IEEE Transactions on Knowledge and Data Engine
用，两篇会议论文分别被数据挖掘顶级会议KDD 2014和自然语言处理著名会议COLING 2014全文录
人民大学信息学院讲师，目前主要工作集中在基于社交媒体的数据挖掘。2014年博士毕业于北京大
学术论文20余篇，引用近600次。 ...

赵鑫 - 师资科研 - 中国人民大学信息学院

(a) homepage (b) result1 (c) result2

Figure 1: pictures about ASBS

In conclusion, I use C++ and python to bulid this task. It's really challenge for me since that I have some problems in C++ web programming and turn to using python(start from very beginning). However, I really learned a lot from it and it's an interesting project that I will continue to solve the problem which I encountered before(the calculate method and use C++ to finish the task). Really a good way to kill time in summer vacation. And thanks for Professor Zhao xin, and TA Menci.

Appendix : How to use ASBS

Requirements:python3.0+, Ubuntu

A pretreat

i) edit crawl.cpp and modify ROOTURL. The default ROOTURL will be http://info.edu.ruc.cn/

ii) edit crawl.cpp and modify the sleep time.

B run

i) crawl.cpp, analyze.cpp, test.cpp, app.py, templates should be in the same catalogue.

ii) compile and run crawl.cpp, analyze.cpp, test.cpp in order. And you will find a number output on the shell, memorize the number as N. Use "python app.py -num N" to run. And the homepage will be "localhost:5000/submit".