信息检索技术大作业报告

第二阶段 文档检索

电 61 王铮澄
2016010910

一、作业任务

输入一个或多个关键字，或者一个句子，输出系统检索到的文档集，所用检索模型自选。

二、总体思路

使用概率模型进行相似性评估，并加入单词纠错功能。

具体流程为：

1、 接受输入
2、 对输入进行纠错
3、 构建文档名称列表
4、 加载第一阶段建立的倒排字典
5、 计算初始 P、R 值
6、 第一次计算文档相似度
7、 根据第一次计算出的相似度对相关文档进行排序
8、 根据改进概率模型重新计算 P、R 值
9、 第二次计算文档相似度
10、 返回结果

三、具体实现

定义 Search()类

➢ 拥有属性：

文档总数 N，相似度最高文档数 V，检索目标单词列表，倒排字典，文档名称列表，相似度字典，单词 p 值字典，单词 r 值字典，检索结果列表。

➢ 拥有方法：

获取文档列表，接受输入（其中包含了单词纠错功能），加载倒排字典，计算第一次 P、R 值，计算相似度，计算第二次 P、R 值，结果展示。

定义 SepllCorrect()类

单词纠错部分代码参考了 Peter Norvig 的代码（http://norvig.com/spell-correct.html），为了便于调用，我将其封装成了类，为了加快纠错速度，避免每次重复训练，我对代码进行了部分修改，将训练结果存为字典，每次纠错时直接调用训练结果即可。这样做提高了纠错效率。此外，为了防止检索词中出现未在训练集中出现的单词，我将纠错结果和原单词一并输入目标单词列表，而错误单词对检索系统的结果是不起任何作用的，纠错功能的意义在于将由于错误输入被忽略的单词补回目标单词列表中。

定义 SearchApp()类

将 Search()类中的方法进行整合，便于直接调用。

四、性能测试

➢ 检索时间测试：

测试检索词：human machine computer

次数	耗时/s

1	0.078125
2	0.09375
3	0.109375
4	0.0625
平均	0.0859

由测试结果可见，检索时间很短，完全可以满足检索需要。

➢ 拼写检查测试：

测试输入	输出	用时/s
speling	spelling	0.015625
mistaka	mistake	0.015625
comouter	computer	0.015625
computer	computer	0
happiness	happiness	0

拼写检查的速度满足使用需要，并且当输入单词正确时，不会额外消耗纠错时间。

➢ 检索准确度测试

为了验证检索准确度，每次检索返回准确度前十的文档名称及其准确度。通过以下两种方式进行准确度评估：

1、 人工评判文档内容是否符合检索要求

2、 与使用不同模型的同学对相同检索词进行交叉验证

测试算例（篇幅所限，仅进行列出一组测试）：

检索词：

simplex method

linear algorithm

computer

返回结果：

```
CACM-1905.html Similarity: 11.911909526768943
CACM-2285.html Similarity: 11.911909526768943
CACM-1453.html Similarity: 9.576961003197452
CACM-1066.html Similarity: 9.514254485585994
CACM-1729.html Similarity: 9.47910860667716
CACM-1863.html Similarity: 9.47910860667716
CACM-2223.html Similarity: 9.47910860667716
CACM-3138.html Similarity: 9.47910860667716
CACM-1169.html Similarity: 7.2221080962616995
CACM-1197.html Similarity: 7.2221080962616995
```

选取前三个文档进行测评：

CACM-1905.html：

The Simplex Method of Linear Programming Using LU Decomposition

Standard computer implementations of Dantzig's

simplex method for linear programming are based upon forming the inverse of the basic matrix and updating

the inverse after every step of the method.

These implementations have bad round-off error properties.

This paper gives the theoretical background

for an implementation which is based upon the LU decomposition,

computed with row interchanges, of the

basic matrix. The implementation is slow, but has good

round-off error behavior. The implementation

appears as CACM Algorithm 350.

CACM May, 1969

Bartels, R. H.

Goulub, G. H.

simplex method, linear programming, LU decomposition,

round-off errors, computational stability

5.41

CA690504 JB February 17, 1978 3:49 PM

CACM-2285.html:

Computer Routine for Quadratic and Linear

Programming Problems [H] (Algorithm A431)

A computer program based on Lemke's complementary

pivot algorithm is presented. This can be

used to solve linear and quadratic programming problems.

The program has been extensively tested on

a wide range of problems and the results have been extremely satisfactory.

CACM September, 1972

Ravindran, A.

linear program, quadratic program, complementary

problem, Lemke's algorithm, simplex method

5.41

CA720905 JB January 27, 1978 4:34 PM

CACM-1453.html:

A Nonrecursive Method of Syntax Specification

The use of the Kleene regular expression notation

for describing algebraic language syntax,

in particular of ALGOL, is described in this paper.

A FORTRAN II computer program for carrying out the

elimination algorithm of Gorn,similar to Gaussian elimination

for linear systems of algebraic equations,

is described. This was applied to numerous smaller

languages, including some sublanguage of ALGOL.

A hand calculation result of the application of the algorithm

to all of ALGOL is given, thus expressing

```
the Revised ALGOL 1960 syntax in completely nonrecursive
terms, as far as its context-free portion is
concerned.   This description in many ways is far more
intuitively understood than the previous recursive
description, it is suggested.   The paper also includes
results of the machine program, which does not
include a simplification algorithm.
CACM April, 1966
Carr III, J. W.
Weiland, J.
CA660402 JB March 3, 1978    11:02 AM
```

从以上三个文档中可以看出，前两个文档包含了所有检索词，因此相似度相同，而第三个文档少了一个“simplex”，因此相似度要低于前两个。由此算例可以看出该检索模型的准确度是比较高的。

对输入有误的检索词进行测试：

检索词：

```
simplex methpd
linwar algorithm
computor
```

检索结果：

```
CACM-1905.html Similarity: 11.911909526768941
CACM-2285.html Similarity: 11.911909526768941
CACM-3138.html Similarity: 11.835942432583185
CACM-1453.html Similarity: 9.57696100319745
CACM-2890.html Similarity: 9.536139787920526
CACM-1066.html Similarity: 9.514254485585994
CACM-1973.html Similarity: 9.500993909011694
CACM-2903.html Similarity: 9.500993909011694
CACM-1729.html Similarity: 9.479108606677158
CACM-1863.html Similarity: 9.479108606677158
```

检索得到的结果与正确输入的结果基本一致。

五、文件说明

Search.py：Search 类所在文件
SearchApp.py：SearchApp 类所在文件
SpellCorrect.py：单词纠错类所在文件
big.txt：单词纠错训练语料库
RevIndex.txt：倒排字典
NWORDS.txt：纠错训练结果字典