比较网页结构相似度

Table of Contents

总体介绍

网页网页结构相似度计算通常是网页自动分类的基础，在一般的网页信息提取中，判断网页片断是“噪声”还是“有效信息”通常是个两类分类问题。 简单地，我们可以把一般网页分为三个类，即：

- 目录导航式页面（List\Index Page）
- 详细页面（Detail Page）
- 未知页面（Unknown Page）

由于网页本身就可以抽象成串行的节点或者是DOM树，那么对于串行序列，就可以常用最长公共子序列来衡量相似度

最长公共子序列

最长公共子序列是动态规划的基本问题：

序列a共有m个元素，序列b共有n个元素，如果a[m-1]==b[n-1]，

那么a[:m]和b[:n]的最长公共子序列长度就是a[:m-1]和b[:n-1]的最长公共子序列长度+1；

如果a[m-1]!=b[n-1]，那么a[:m]和b[:n]的最长公共子序列长度就是

MAX（a[:m-1]和b[:n]的最长公共子序列长度，a[:m]和b[:n-1]的最长公共子序列长度）

递归式展示

$$c[i,j]=\begin{cases}0 & \text{if } i=0 \text{ or } j=0,\\ c[i-1,j-1]+1 & \text{if } i,j>0 \text{ and } x_i=y_j,\\ \max(c[i,j-1],c[i-1,j]) & \text{if } i,j>0 \text{ and } x_i\neq y_j.\end{cases}$$

算法实现（python实现）

```python
#params:
# - a : str
# - b : str
#return
# - c : 过程处理矩阵
# - c[x][y] : the lcs-length(最长公共子序列长度)
def lcs(a,b):
    lena=len(a)
    lenb=len(b)
    c=[[0 for i in range(lenb+1)] for j in range(lena+1)]
    for i in range(lena):
        for j in range(lenb):
            if a[i]==b[j]:
                c[i+1][j+1]=c[i][j]+1
            elif c[i+1][j]>c[i][j+1]:
                c[i+1][j+1]=c[i+1][j]
            else:
                c[i+1][j+1]=c[i][j+1]
    return c,c[lena][lenb]
```

网页相似度计算

```python
#-*-coding:utf-8-*-
import lxml.html.soupparser as soupparser
import requests
headers = {
    "User-Agent":"Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like
Gecko) Chrome/55.0.2883.87 Safari/537.36"
}
def get_domtree(html):
    dom = soupparser.fromstring(html)
    for child in dom.iter():
        yield child.tag

def similar_web(a_url,b_url):
    html1 = requests.get(a_url,headers=headers).text
    html2 = requests.get(b_url,headers=headers).text
    dom_tree1 = ">".join(list(filter(lambda e:
isinstance(e,str),list(get_domtree(html1)))))
    dom_tree2 = ">".join(list(filter(lambda e:
isinstance(e,str),list(get_domtree(html2)))))
    c,flag,length = lcs(dom_tree1,dom_tree2)
    return 2.0*length/(len(dom_tree1)+len(dom_tree2))

percent = similar_web(
'http://edmondfrank.github.io/blog/2017/04/05/qian-tan-mongodb/',
'http://edmondfrank.github.io/blog/2017/03/27/emacsshi-yong-zhi-nan/')
print(percent) #相似度（百分比）
```

Posted by EdmondFrank • • python

Tweet

« 浅谈MongoDB

Web正文提取(偏純文本类) »

About Me

Recent Posts

Popular Posts

GitHub: @EdmondFrank

Twitter: @EdmondFrank4

Blog:
https://edmondfrank.github.io

この町、冗談と気まぐれと偶然でてきっているらしい。