《电子商务应用》

实验报告

徐鸣飞、黄梓霖、皮佳宇、陈其阳

2023 年 10 月 27 日

目录

第一章　概述

1.1　关键字广告与竞争性营销

关键词广告 (keyword) 是一种文字链接型网络广告，通过对文字进行超级链接，让感兴趣的网民点击进入公司网站、网页或公司其它相关网页，实现广告目的。链接的关键词既可以是关键词，也可以是语句。

竞争性营销：在营销管理过程中，管理者不仅要考虑顾客的需要，还要考虑企业在本行业中的竞争地位。企业的营销战略和战术必须从自己的竞争实力地位出发，并根据自己同竞争者实力对比的变化，随时加以调整，使之与自己的竞争地位相匹配。由于现代市场营销中竞争的重要性，市场营销不仅包括“产品、价格、促销、渠道”四方面因素，还应让“竞争”成为现代市场营销的第五大因素。竞争意识要在企业的营销决策、营销规划、营销组织中充分体现出来，在营销实践中也要采取有效的策略开展竞争，不断提高企业竞争能力。

1.2　关键字推荐方法

设所有关键词全集为：K，

广告主的种子关键词：$s\,(s \in K)$，

推荐高相关性的关键词：$K_M\,(K_M \in K)$。

- 相关分析法：计算词与词之间的相关程度：

1. 关联相关性：指的是两个关键词之间的相关性，通常是指它们在同一文本中出现的频率或者它们之间的语义关系。例如，在一篇文章中，如果两个关键词经常同时出现，那么它们之间就有很强的关联相关性。
2. 同义词相关性：指的是两个关键词之间的同义词关系。例如，“汽车”和“车辆”就是两个同义词，它们之间有很强的同义词相关性。同义词相关性通常需要通过词汇库或者自然语言处理技术来识别。
3. 竞争相关性：指的是两个关键词之间的竞争关系。例如，在搜索引擎中，如果用户搜索“手机”，那么“苹果”和“三星”就是两个与“手机”相关的竞争关键词。竞争相关性通常需要通过分析用户搜索行为或者竞争对手的关键词来识别。

- 方法：

方法分类	优点	缺点
基于搜索日志	• 方法相对简单 • 能够反映用户的搜索兴趣	• 推荐词与种子关键词高度重复 • 难以处理模棱两可的词
基于网页内容	• 推荐词不与种子关键词高度重复	• 相关性不高 • 推荐质量取决于分析网页内容的范围和规模 • **方法复杂耗时** • 不能反映用户的搜索兴趣
基于语义知识库	• 推荐的关键词语义丰富 • 能推荐出意思相近但不同的词 • 不与种子关键词高度重复	• 推荐质量**高度依赖于语义知识库质量** • 不能反映用户的搜索兴趣 • 更新速度慢 • 方法复杂耗时，需要人工介入

图 1.1: 关键词推荐方法比较

- 通过了解用户搜索意图来获得 K_M

1.3　竞争性关键字与度量方法

构建相应的概率模型来计算关键字之间的竞争度：

$$Comp_a\left(k,s\right)=\frac{|ka|}{(|a|-|sa|)}$$

- s：种子关键字
- k：任意关键词
- a：中介关键词

- 种子关键词 s 与任意关键词 k，存在一个中介关键词 a，其与 s 与 k 都存在联合查询，即 sa 与 ka。
- 度量了在所有查询 ka 的搜索量占 a 搜索量中除去 sa 搜索量的比例，反映了在 a 关键词的维度上，用户不搜索 k 而搜索 s 的概率。

多个中介关键词存在程度的竞争性，a 对于 k 的影响权重：

$$W_a(k)=\frac{|sa|}{|s|}$$

关键词与种子关键词的竞争性程度：

$$Comp(k,s)=\sum_{i=1}^{m}\left\{w_{ai}(k)\times Comp_{ai(k,s)}\right\}$$

1.4　设计与实现的技术线路

1. 操作系统：Windows 10 家庭中文版。
2. 软件环境：JDK 1.8 + word 1.3
3. 设计与实现：
    - **总体流程**：清洗数据，获取所有关键词；compkey 算法开始，输入种子关键词，如：“湖南”；接着从总搜索量中提取出与种子关键词相关的所有搜索信息；对与种子关键词相关的搜索信息进行分词与词频统计；通过对词频统计信息分析确定一定数目的中介关键词；对每个中介关键词查找相应的竞争性关键词，并计算竞争度；最后统计输出，结果保存到文件中。

- **中介关键词的选取**：对于一个确定的种子关键词，对清洗后的数据进行提取，得到与种子关键词有关的所有的搜索量保存到文件中，使用 word 分词插件对该文件进行分词与词频统计，出现的次数越多说明相关性越大，权重越大，则可以选取出现频率高的词作为中介关键词。
- **竞争关键词的选取**：每一个中介关键词都会对应一个竞争性关键词。对于每一个确定的中介关键词，从清洗后的数据中进行提取出于中介关键词有关的所有的搜索量保存到文件中，使用 word 分词插件对该文件进行分词与词频统计，出现的次数越多说明其与中介关键词相关性越大，则可以选取出现频率最高的词（该词不是种子关键词和中介关键词）作为竞争关键词。

第二章　关键词竞争算法数据预处理

2.1　数据来源与数据特征

1. 数据来源：C:\Users\24964\Desktop\mavenproject1 \data\搜狗比赛数据 \user_tag_query.10W.TRAIN

2. 数据特征（字段说明）：数据特征（字段说明）

   - **ID**：加密后的 ID
   - **age**：0：未知年龄; 1：0-18 岁; 2：19-23 岁; 3：24-30 岁; 4：31-40 岁; 5：41-50 岁; 6：51-999 岁
   - **Gender**：0：未知 1：男性 2：女性
   - **Education**：0：未知学历; 1：博士; 2：硕士; 3：大学生; 4：高中; 5：初中; 6：小学

3. 数据示例：00627779E16E7C09B975B2CE13C088CB 4 2 0 钢琴曲欣赏 100 首 一个月的宝宝眼睫毛那么是黄色 宝宝右眼有眼屎 小儿抽搐怎么办 剖腹产后刀口上有线头 属羊和属鸡的配吗

2.2　种子关键词的选取

”作文”, ”小说”, ”大学”, ”软件”, ”工资”, ”诗句”, ”电视剧”, ”广场舞”, ”手机”, ”壁纸”

2.3　数据获取与关键代码实现

```
1  public class MainDataClass {
2      public static void data(String wordKey) throws
           FileNotFoundException,
           UnsupportedEncodingException, IOException,
           Exception {
3          PathClass pa = new PathClass();
4          InputStreamReader inStream = new
               InputStreamReader(new FileInputStream(
               new File(pa.wordOut)), "utf-8");//读取总
               搜索量文件
5          OutputStreamWriter outStream = new
               OutputStreamWriter(new FileOutputStream(
               new File(pa.wordRelated)), "utf-8");
6          BufferedReader bf = new BufferedReader(
               inStream);
7          BufferedWriter bw = new BufferedWriter(
               outStream);
8          String valueString = null;
9          while ((valueString=bf.readLine())!=null){
               //与种子关键字相关的搜索信息
```

```
10                  if(valueString.contains(wordKey)) {
11                      bw.append(valueString);
12                      bw.newLine();
13                  }
14              }
15              bw.close();
16              File f = new File(pa.wordRelated);
17              if(f.length() == 0) {
18                  System.out.println("搜索日志中不含关键
                        词“" + wordKey +"”，请重新设置种子
                        关键词！！！");
19                  System.exit(0);
20              }
21              System.out.println("加载Word分词器...");
22              System.out.println("开始对与种子关键字相关
                    的搜索信息进行分词与词频统计...");
23              new StatisticsDataClass().statistic(pa.
                    wordRelated, pa.wordApart, pa.
                    wordStatistics);//对相关信息进行分词和词
                    频统计
24          }
25      }
26      public class StatisticsDataClass {
27          public static void statistic(String wordOut,
                String wordApart, String wordStatistics)
                throws Exception {
28              //词频统计设置
```

```
29            PathClass pa = new PathClass();
30            PrintStream ps = new PrintStream(pa.log);/*
                  过滤屏幕信息*/
31            PrintStream out = System.out;
32            System.setOut(ps);
33            WordFrequencyStatistics
                  wordFrequencyStatistics = new
                  WordFrequencyStatistics();
34            wordFrequencyStatistics.setRemoveStopWord(
                  true);//去掉虚词和一般的连词
35            wordFrequencyStatistics.setResultPath(
                  wordStatistics);
36            wordFrequencyStatistics.
                  setSegmentationAlgorithm(
                  SegmentationAlgorithm.MaxNgramScore);
37            wordFrequencyStatistics.seg(new File(
                  wordOut), new File(wordApart));
38            wordFrequencyStatistics.dump();//输出词频统
                  计结果
39            System.setOut(out);
40            wordFrequencyStatistics.dump();
41        }
42    }
```

2.4　预处理后的数据格式

换行后结果：见图2.1 分词后结果：见图2.2 统计词频率结果：见图2.3

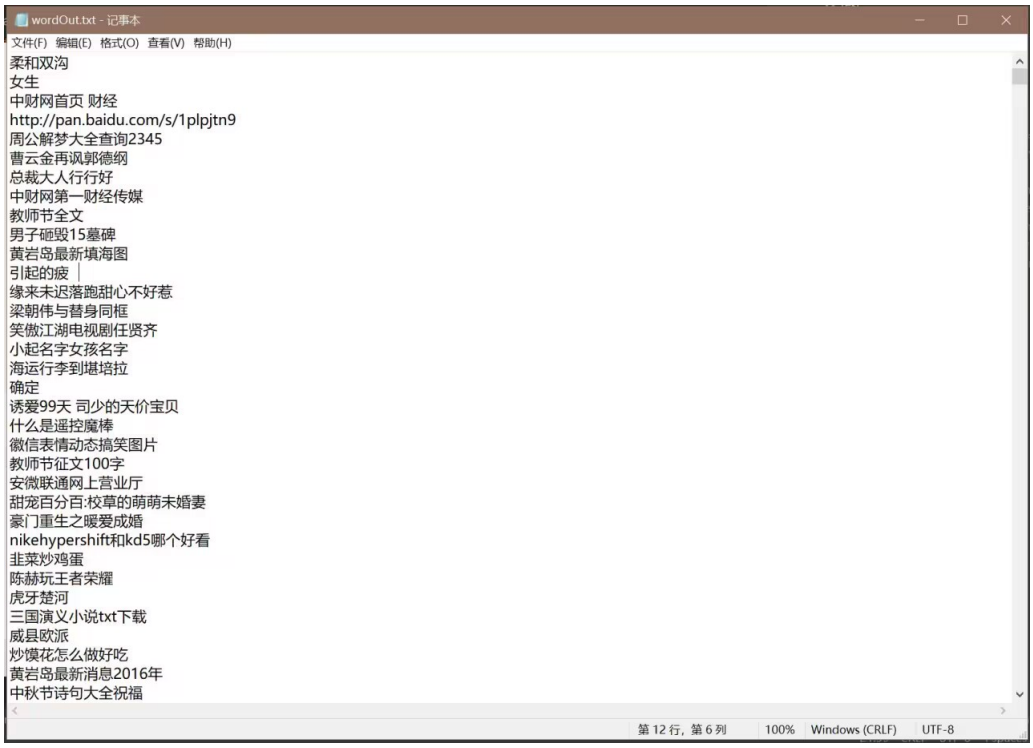

图 2.1: 换行结果

图 2.2: 分词结果


wordStatistics.txt - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
壁纸 261
手机 91
高清 35
图片 30
大全 18
动漫 17
电脑 15
动态 15
屏 12
桌面壁纸 11
锁 9
图片壁纸 8
超 8
美女壁纸 7
清 7
换 6
权志龙 6
亚 6
好看 5
长草 4
美女 4
艾米莉 4
情侣 4
设置 4
猫 4
金立 4
颜 4
林允儿 4
唯美 4
头像 4
萌 4

图 2.3: 词频统计结果