系统设计

手表评论的观点挖掘系统是一套无监督地自动挖掘出用户评论中的用户观点的系统，它能够定时爬取例如电商网站的用户评论，进行分析，生成分析报表，最后以邮件的形式发送给用户。系统整体流程图如下：

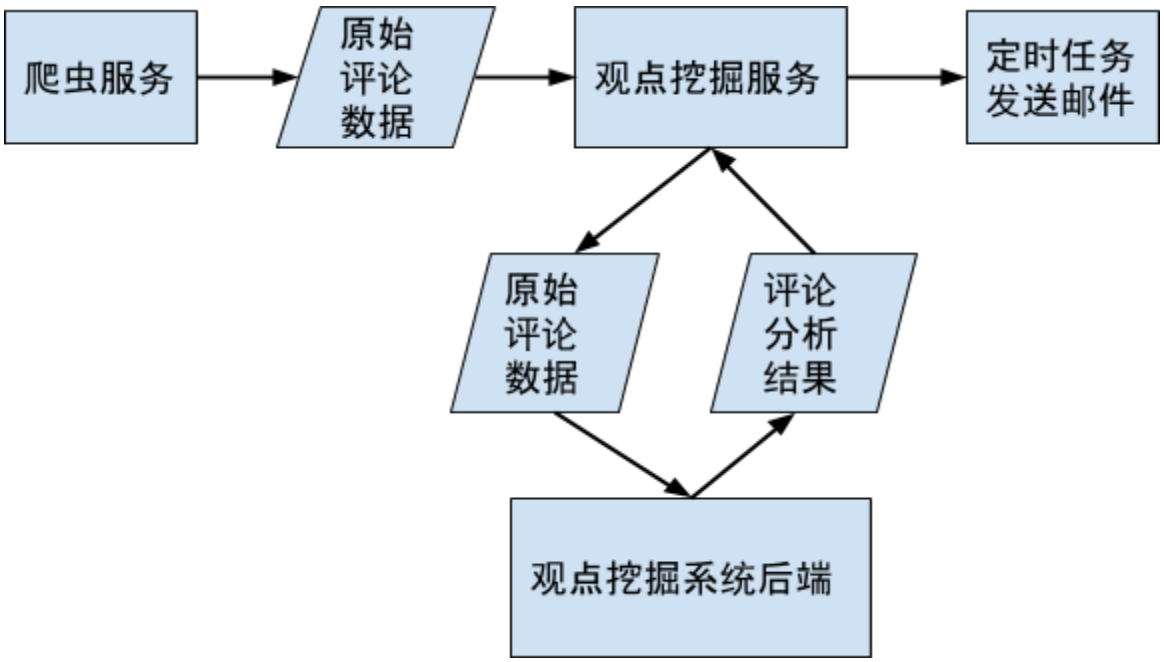

图1 手表评论观点挖掘系统流程图

观点挖掘系统后端的完整流程图如下：

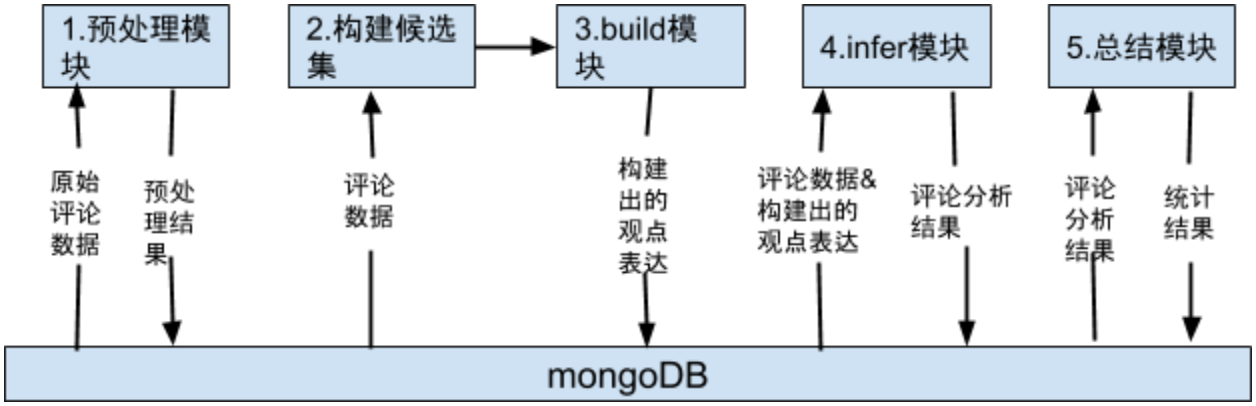

图2 观点挖掘系统后端流程图

如上图所示：

1.预处理：系统首先对评论进行预处理，预处理的内容包括去除噪声字符，对评论做分词处理，以及词性标注处理。

2.构建候选集：在分词以及词性标注的基础上，首先利用通用情感词典中的形容词利用开窗口的方式寻找形容词附近的名词作为候选aspect。然后利用候选的aspect开窗口，寻找候选的（aspect，pattern，opinion）组合。举个例子：用户的原始评论为“客服态度还是非常不错”，那么得到的候选组合为（“客服”（aspect），“还是非常”（pattern），“不错”（opinion））。
3.build：在抽取完所有的候选组合之后利用二部图排序算法进行排序，其中二部图的两边分别为：（aspect，opinion）pair 和 pattern。过滤得到最终的正确的表达集和，也就是build过程。这一排序算法的背后原理可以直观上的解释为：如果大部分人都写出了某种表达，比如“客服态度很好”那么这句表达就是正确的，如果某种表达出现的次数很少，比如“客服态度很便宜”，那么这个表达就会被排序算法排在较后的位置，也就是不正确的表达。
4.infer模块：利用挖掘出的正确观点表达（aspect，opinion）pair从原始评论中找出观点，并根据一张pair的极性表全判断观点的极性，正向或是负向。
5.总结模块：将原始的aspect依据词向量向用户定义的标签做映射，比如“电量”，“电池”，“耗电量”这四个aspect经过计算会被归类到用户定义的标签“续航”上面，然后统计每个标签下面的观点的正向负向占比，输出最终的统计结果。

为何选择二部图排序算法：经过实践，二部图排序算法相比于其他无监督的方法[1][2]在手表领域评论中取得了较好的效果。首先该算法为无监督方法，所以省去了人工标注训练集的繁琐工作，其次，该算法无需复杂的特征，仅需要构建一个候选集即可（流程步骤2）。

[1]Hu, Minqing, and Bing Liu. "Mining opinion features in customer reviews." AAAI. Vol. 4. No. 4. 2004.
[2]Qiu, Guang, et al. "Opinion word expansion and target extraction through double propagation." Computational linguistics 37.1 (2011): 9-27.

通用情感词典的构建：
我们利用NTUSD台湾大学中文情感词典去匹配评论中出现的情感词得到候选情感词，然后再利用形容词词性进行过滤得到通用情感词典。未来工作会利用上面的方式在别的领域不断扩充更新该通用情感词典。

用户定义标签：用户定义标签需要领域专家进行构建。本系统中，手表领域的标签的定义是由我司的手表研发，销售团队根据手表这一产品的特征讨论得出的。

标签映射：标签映射的方法有很多，经过实践，利用词向量计算距离的方法映射的准确率最高。首先计算该待映射的aspect与所有用户定义标签的余弦距离，然后选择距离最近的用户定义标签作为映射对象。

Web 接口

一. 获取产品名称列表

1. 请求地址：http://host:8081/productName

2. 请求方式：GET
3. 请求参数：无
4. 响应示例：

```
{
    "code": 200,
    "data": [
        {
            "id": 1,
            "name": "TicWatch E 时尚"
        },
        {
            "id": 2,
            "name": "TicWatch 2 经典"
        },
        ...
    ],
    "time_cost": 95.065
}
```

二. 获取标签列表

1. 请求地址：http://host:8081/tagList
2. 请求地址：GET
3. 请求参数：domain=手表（string， 目前仅支持“手表”）
4. 响应示例：

```
{
    "code": 200,
    "data": [
        {
            "id": 1,
            "name": "物流"
        },
        {
            "id": 2,
            "name": "续航"
        },
        ...
    ],
    "time_cost": 1.099
}
```

三.获取评论接口

1. 请求地址：http://host:8081/comments

2. 请求方式：POST
3. 请求参数：
    a. 地址：pageNo=1（int），pageSize=10（int）
    b. body：{"spu": "TicWatch E 时尚"（string）, "sku": "朋克黑"（string）, "polarity": 1（int，情感正向是1，情感负向是-1）, "tag": "续航"（string）, "startTime": 1472659200000（毫秒级时间戳）, "endTime": 1535731200000（毫秒级时间戳）}
4. 响应示例：


```
{
  "code": 200,
  "data": {
    "items": [
      {
        "sku": "表壳尺寸:44mm;表带尺寸:适合130-200毫米腕围;表系列:朋克黑",
        "comment_time": "2017-10-21",
        "clean_comment": "还可以，电有点用的快，还有音乐听歌有点复杂，现在都不懂弄，其它的都还可以",
        "pairs": [
          {
            "raw_express": "用的快",
            "opinion": "快",
            "polarity": 1,
            "tag": "续航",
            "aspect": [
              "用的"
            ]
          }
        ],
        "domain": "手表",
        "spu": "TicWatch E 时尚"
      },
      {
        "sku": "表壳尺寸:44mm;表带尺寸:适合130-200毫米腕围;颜色分类:朋克黑",
        "comment_time": "2017-10-09",
        "clean_comment": "电量还可以，带了这么久，两个压表带的环环都变松了，经常脱落，表带会翘出来，而且这个材质，特别容易受伤留下伤痕，材质时间长会变形，表带，有点后悔买，分期没还完，手表戴够了，尴尬",
        "pairs": [
          {
            "raw_express": "时间长",
            "opinion": "长",
            "polarity": 1,
            "tag": "续航",
```

```
                    "aspect": [
                        "时间"
                    ]
                }
            ],
            "domain": "手表",
            "spu": "TicWatch E 时尚"
        },
        ...
    ],
    "page": {
        "totalCount": 26,
        "pageSize": 10,
        "pageNo": 1,
        "totalPageCount": 3
    }
},
"time_cost": 28.333000000000002
}
```


四.获取spu下的sku

1. 请求地址：http://host:8081/productSku
2. 请求方式：GET
3. 请求参数：spu=TicWatch E 时尚（string）
4. 响应示例：


```
{
    "code": 200,
    "data": [
        {
            "id": 1,
            "name": "全部"
        },
        {
            "id": 2,
            "name": "朋克黑"
        },
        {
            "id": 3,
            "name": "摩登白"
        },
        {
            "id": 4,
```

```
            "name": "MLGB定制款"
        }
    ],
    "time_cost": 1.848
}
```


五.获取评论观点情感极性统计（用于单个时间区间画图）

1. 请求地址：http://host:8081/picture
2. 请求方式：POST
3. 请求参数：body：{"spu": "TicWatch E 时尚"（string）, "sku": "朋克黑"（string）, "domain":"手表"（string）, "startTime": 1472659200000（毫秒级时间戳）, "endTime": 1535731200000（时间戳）}
4. 响应实例：


```
{
    "code": 200,
    "data": {
        "外观": {
            "pos": 254,
            "neg": 11
        },
        "软件": {
            "pos": 235,
            "neg": 33
        },
        ...
    },
    "time_cost": 126.42500000000001
}
```


（注：缺失值用空字符代替）

六.获取评论观点情感极性统计（用于多个时间区间比较）

1. 请求地址：http://host:8081/pictureTrends
2. 请求方式：POST
3. 请求参数：body：{"spu": "TicWatch E 时尚"（string）, "sku": "朋克黑"（string）, "domain":"手表"（string）, "startTime": 1493568000000（毫秒级时间戳）, "endTime": 1506787200000（毫秒级时间戳）}
4. 响应实例：


```
{
    "code": 200,
    "data": {
        "外观": {
            "pos": [",6,23,83,90,40],
```

```
            "neg": ["","",5,2,2,2]
        },
        "软件": {
            "pos": ["", "",32,62,86,43],
            "neg": [",1,4,13,10,4]
        },
        ...
    },
    "time_list": [
        "2017-05-01",
        "2017-06-01",
        "2017-07-01",
        "2017-08-01",
        "2017-09-01",
        "2017-10-01"
    ],
    "time_cost": 73.068
}
```

（注：缺失值用空字符代替）

七.数据预处理接口

1. 请求地址：http://host:8081/preprocess
2. 请求方式：GET
3. 请求参数：domain=手表（string）,startTime=1493568000000（毫秒级时间戳），endTime=1506787200000（毫秒级时间戳）
4. 响应实例：mongoDB的log

八.pair build接口

1. 请求地址：http://host:8081/build
2. 请求方式：GET
3. 请求参数：domain=手表（string）
4. 响应实例：mongoDB的log

九.批量infer接口

1. 请求地址：http://host:8081/inference
2. 请求方式：GET
3. 请求参数：domain=手表（string）spu=string(特殊为’none’)
4. 响应实例：mongoDB的log

十.单条infer接口

1. 请求地址：http://host:8081/infer
2. 请求方式：GET
3. 请求参数：text=首先发货真的快 第二天就收到 用了两天了 出了电池续航不好之外 手表真的很好用 暂时没有发现有什么问题
4. 响应实例：


```
{
   "code": 200,
   "data": [
      {
         "aspect": [
            "发货"
         ],
         "raw_express": "发货真的快",
         "polarity": 1,
         "opinion": "快",
         "tag": "物流"
      },
      {
         "aspect": [
            "续航"
         ],
         "raw_express": "续航不好",
         "polarity": -1,
         "opinion": "不好",
         "tag": "续航"
      },
      {
         "aspect": [
            "手表"
         ],
         "raw_express": "手表真的很好用",
         "polarity": 1,
         "opinion": "好用",
         "tag": "外观"
      }
   ],
   "raw_comment": "首先发货真的快 第二天就收到 用了两天了 出了电池续航不好之外 手表真的很好用 暂时没有发现有什么问题",
   "time_cost": 20.475
}
```

数据库 MongoDB

一.collection：opinion_resources现在包含以下5种doc

1.通用情感词典，document字段：{“doc_type”：”general_opinion”，“lexicon”：{包含极性的通用情感词典 }}
2.停用词表，document字段：{“doc_type”：“stopwords”，“title”：“人物关系”（停用词类别），“lexicon”：[停用词]}
3.标签词表，document字段：{“doc_type”：“product_tag”,“domain“：”手表“，”lexicon“：[]}
4.spu词表，document字段：{“doc_type”：“product_list”，”lexicon”：[]}
5.sku词表，document字段：{“doc_type”：“sku_list”，“lexicon”：{每个spu对应的sku}}

二.collection：opinion_build_pairs包含1种doc

1.系统构建出的包含情感的观点pair：{“domain”：“手表”，“pair_polarity”：{包含极性的观点pair}}

三.collection：comments包含1种doc

1.一个包含全部字段的完整实例：
{
"_id" : ObjectId("5b8641ab143cee198efa5ca9"),
"comment_time"（评论时间，只精确到天）: ISODate("2017-10-19T00:00:00Z"),
"domain" : "手表",
"sku" : "表带尺寸:适合130-200毫米腕围;表系列:纯黑（蓝宝石屏3G版）精钢表头配 黑色尖尾真皮表带",
"spu" : "TicWatch 2 蓝宝石",
"raw_comment"（原始评论）: "tic2好手表漂亮，比我以前摩托罗拉360手表好多了，物流好快第二天就收到了。总之可以推荐需要买的网友们！",
"seg"（分词结果）: "好手 表 漂亮， 比 我 以前 摩托罗拉 手表 好多 了， 物流 好快 第二 天 就 收到 了 。 总之 可以 推荐 需要 买 的 网友们 ！",
"pos" （词性标注）: "NN NN VA PU P PN NT NR NN CD SP PU NN VA OD M AD VV AS PU AD VV VV VV VV DEC NN PU",
"clean_comment" （清洗过后的评论）: "好手表漂亮，比我以前摩托罗拉手表好多了，物流好快第二天就收到了。总之可以推荐需要买的网友们！"
}