第10章 其他挖掘方法

10.1 文本挖掘技术

10.2 Web挖掘

数据挖掘的研究范围十分广泛，除了前面几章介绍的基本数据挖掘方法外，数据挖掘方法应用到不同的领域形成了与相关领域相结合的各种数据挖掘技术。

本章主要介绍文本挖掘、Web挖掘方法。

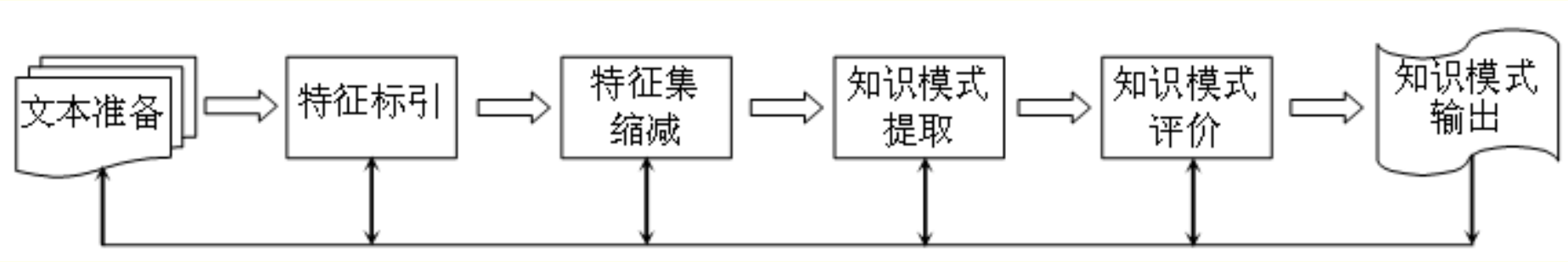
10.1 文本挖掘技术

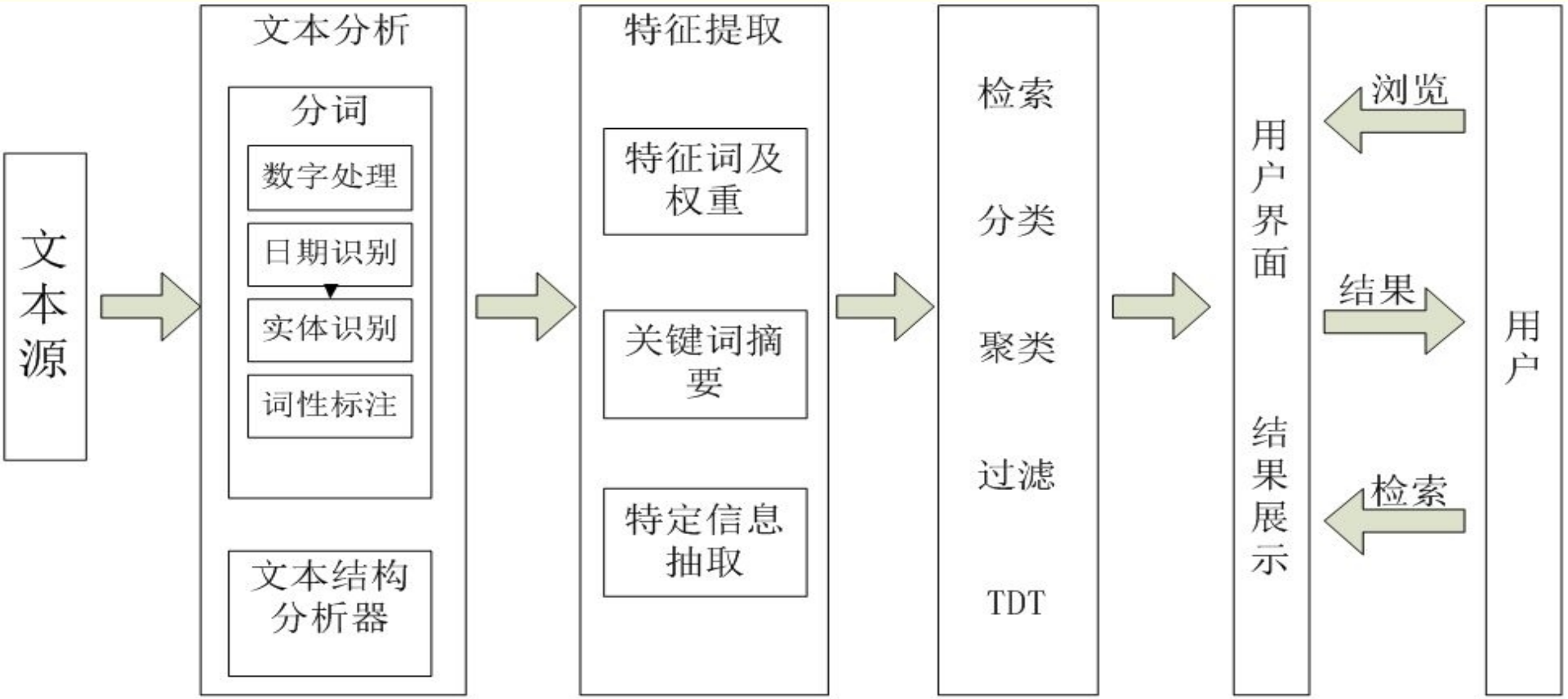
10.1.1 文本挖掘概述

1. 什么是文本挖掘

文本挖掘处理的是非结构化的文本信息，文本挖掘的主要任务是分析文本的内容特征，发现文本中概念、文本之间的相互作用，为用户提供相关知识和信息。

2. 文本挖掘过程


文本准备
特征标引
特征集缩减
知识模式提取
知识模式评价
知识模式输出


文本源
文本分析
分词
数字处理
日期识别
实体识别
词性标注
文本结构分析器
特征提取
特征词及权重
关键词摘要
特定信息抽取
检索
分类
聚类
过滤
TDT
用户界面
结果展示
浏览
结果
检索
用户

2. 文本挖掘过程

（1）输入：结构化和非结构化的数据；

（2）任务1：建立语料库。收集与研究领域相关的文档，包括文本文档、XML文件、电子邮件、网页、录音等。商用的文本挖掘软件能够将它们转换为平面文件；

（3）任务2：创建词项-文档矩阵（TDM）。将语料库转换为TDM过程中，涉及到中文分词、英文词干提取、删除停用词、词频统计、建立面向特定问题的词典、词项的索引表示（例如反文档频率等）、矩阵降维等。

（4）任务3：使用文本挖掘方法提取知识。

（5）输出：用于决策的特定知识。

3. 文本挖掘和数据挖掘的区别

区别项	数据挖掘	文本挖掘
研究对象	用数字表示的、结构化的数据	无结构或者半结构化的文本
对象结构	关系数据库	自由开放的文本
目标	获取知识，预测以后的状态	提取概念和知识
方法	关联分析、*k*-最近邻、决策树、贝叶斯分类、神经网络、支持向量机、粗糙集、聚类算法等	提取短语、形成概念、关联分析、文本分类、文本聚类等

10.1.2 数据预处理技术

1. 分词技术

（1）基于词库的分词方法

基于词库的分词方法是按照一定的策略，将文本中的一部分可能被切成一个词的小段与一个词典（词库）里面的词进行比较，若存在，则划分为一个词。

根据采用的策略不同又分为正向最大匹配和逆向最大匹配等。

例如，一个句子为S=“我们是学生”，长度n=5。

正向最大匹配

S_1=“我们是学”

S_1=“我们是”

S_1=“我们”，找到了

S_2=“是学生”，

S_2=“是学”

S_2=“是”，找到了

S_3=“学生”，找到了

所以S的分词结果是“我们/是/学生”。

例如，一个句子为S=“我们是学生”，长度n=5。

反向最大匹配

S_1=“我们是学生”

S_1=“们是学生”

S_1=“是学生”

S_1=“学生”，找到了

S_2=“我们是”

S_2=“们是”

S_2=“是”，找到了

S_3=“我们”，找到了

所以S的分词结果同样是“我们/是/学生”。

（2）基于无词典的分词方法

这种方法是基于词频的统计，将原文中任意前后紧邻的两个字作为一个词进行出现频率的统计，出现的次数越高，成为一个词的可能性也就越大，在频率超过某个预先设定的阈值时，就将其作为一个词进行索引。

2. 特征表示

文本特征指的是关于文本的元数据，分为描述性特征（如文本的名称、日期、大小、类型等）和语义性特征（如文本的作者、机构、标题、内容等）。

特征表示是指以一定特征项（如词或描述）来代表文档，在文本挖掘时只需对这些特征项进行处理，从而实现对非结构化的文本处理。这是一个非结构化向结构化转换的处理步骤。

特征表示模型中常用的是向量空间模型（Vector Space Model，VSM）。

在向量空间模型中，一个文本集由若干文本组成，每个文本被表示为在一个高维词空间中的一个特征向量：

$$d_i=(t_{i,1}: w_{i,1}, t_{i,2}: w_{i,2}, \ldots, t_{i,m}: w_{i,m})$$

其中d_i为文本，$t_{i,j}$表示第i个文本d_i中的第j个词，$w_{i,j}$表示词$t_{i,j}$在文本d_i中的权重。词的权重一般采用$w_{i,j}=tf\times idf$方法来计算得到。

定义10.1 词频*tf*（Term Frequency）是指一个词在一个文本中出现的频数，其定义为：

$$tf_{t_{i,j}} = \frac{n_{t_{i,j}}}{N_i}$$

其中，是词$t_{i,j}$在文本d_i中出现的次数，N_i是文本d_i中所有词出现的总数。显然，一个词的*tf*值越大，则对文本的贡献度越大。

定义10.2　逆文本频度*idf*（Inverse Document Frequency）表示一个词在整个文本集中的分布情况，其定义为

$$idf_{t_{i,j}} = \log_2 \frac{N}{m_{t_{i,j}}}$$

其中，*N*是文本集中包含的文本总数，　$m_{t_{i,j}}$ 是包含词$t_{i,j}$的文本个数。

tf×*idf*是一种常用的词权重计算方法，有多种形式。如果一个词或短语在一篇文章中出现的词频*tf*高，并且在其他文章中很少出现，则认为该词或短语具有律好的类别区分能力，适合用来分类。

tf×*idf*结合了两者，从词出现在文本中的频率和在文本集中的分布情况两方面来衡量词的重要性。

3. 特征提取

特征提取算法一般是构造一个评价函数，对每个特征进行评估，然后把特征按分值高低排队，预定数目分数最高的特征被选取。

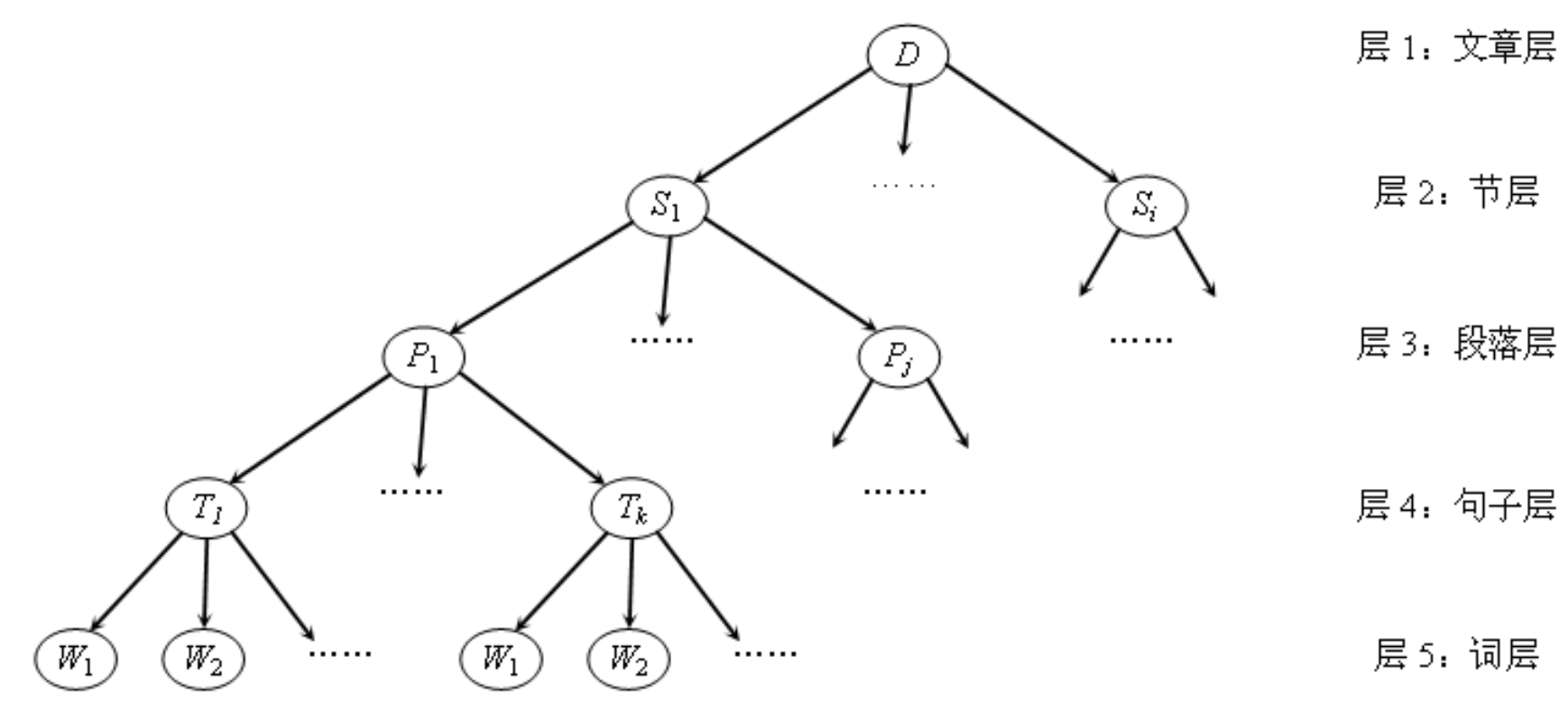
在文本处理中，常用的评估函数有信息增益、期望交叉熵（Expected Cross Entropy）、互信息（Mutual Information）、文本证据权（The Weight of Evidence for Text）和词频等。

10.1.3 文本结构分析

文本结构分析的目的是为了更好地理解文本的主题思想，了解文本所表达的内容以及采用的方式。

最终结果是建立文本的逻辑结构，即文本结构树。

如图11.2所示是文章的形式结构图，根结点是文章层，依次为节层、段落层、句子层和词层。



10.1.4 文本分类

- 朴素贝叶斯分类算法
- 类中心最近距离分类算法
- *k*-最近邻分类算法
- 决策树分类算法
- 神经网络

分类性能评估

查全率是衡量所有实际属于某个类别的文本被划分到该类别中的比率。查全率越高表明分类器在该类上可能漏掉的分类越少，它体现了分类的完备性“

$$\text{查全率}=\frac{\text{正确分类的样本数}}{\text{应有的样本数}}$$

查准率是衡量所有被划分到该类别的文本中正确文本的比率。查准率越高表明在该类别上出错的概率越小，它体现了分类的准确程度：

$$\text{查准率}=\frac{\text{正确分类的样本数}}{\text{实际分类的样本数}}$$

11.1.5 文本聚类

- 基于划分的方法
- 基于层次的方法
- 基于密度的方法
- 基于网格的方法
- 基于模型的方法

10.1.5 文本自动摘要

文本摘要是指从文档中抽取关键信息，用简洁的形式对文档内容进行解释和概括。这样，用户不需要浏览全文就可以了解文档或文档集合的总体内容。

1. 单文档自动摘要
2. 多文档自动摘要

10.1.6 文本关联分析

采用基于关键字的关联分析是从文本集中收集词或者关键字的集合，将问题转化为事务数据库中事务项的关联挖掘。

其基本过程是，调用关联挖掘算法发现频繁共现的词或关键字，即频繁项集，然后根据频繁项集生成词或关键字的关联规则。

例如，产生这样的关联规则：

{数据挖掘，密度}→{DBSCAN，OPTICS}

（支持度=30%，置信度=50%）

10.1.7 文本挖掘应用

（1）自然语言处理应用：自动问答、机器翻译、语音识别、文本朗读等；

（2）营销应用：挖掘呼叫中心记录和产品评论等识别消费者情感、预测消费者后续购买行为等；

（3）安全应用：舆情监控等；

（4）生物医学应用：医学文献挖掘，发现基因-疾病关系、基因-蛋白质关系等；

（5）学术应用；

（6）其它。例如：专利数据库挖掘，可以获取竞争情报、帮助企业进行收购或者新产品研发的决策、保护企业自身知识产权等。

10.2 情感分析与观点挖掘

10.2.1 情感分析概述

1. 什么是情感分析

对于文本中表达的情感和观点进行分析。

文本信息可以分成事实和观点两类。

普通的文本挖掘和文本检索、情感分析与观点挖掘分别针对这两种信息。

2. 情感分析应用

- 产品比较与推荐
- 个人与机构声誉分析
- 电视节目满意度分析
- 互联网舆情分析
  - 利用文本情感计算技术深入分析人们对社会现实和现象的群体性情绪、观点、思想、心理、意志和要求；

网页 图片 视频 资讯 地图 更多 | MSN | Hotmail

必应 bing Beta

apple ipod

所有结果

购物

SHOPPING

POPULAR FEATURES

全部

Ease Of Use

Screen

Sound Quality

Affordability

Appearance

Battery Life

Size

Video

Speed

RESOURCES

How cashback works

Frequently asked questions

cashback for advertisers

iPod touch 8GB 2nd Generation

from $161 (24 stores) Bing cashback · 2 - 5%

user reviews (378)

expert reviews (4)

Highlights includes groundbreaking technologies such as Multi-Touch, the accelerometer, 3D graphics and access to hundreds of games. Play hours of music. Create a Genius playlist of songs that go great together. Watch a movie.... more...

user reviews | product details | expert reviews | compare prices

user reviews

view: **positive comments** (116) | negative comments (19)

ease of use 86%

Pros: Packed with applications, very handy and easy to use.
Timothij www.ciao.co.uk 8/17/2008 more...

Pros: User interface is beautiful and easy to find things.The built in app store store is amazing and very easy to use
Shinrahn reviews.cnet.com 12/30/2008 more...

Pros: Intuitive interface, very easy to use, gorgeous device, very slender profile
Dyonas www.ciao.co.uk 10/13/2007 more...

Pros: Navigation is great, Apps are easy to use and Access, easy to sync Calender and Contacts, volume control buttons, external speaker, and a million other things
anarchy4128 reviews.cnet.com 5/2/2009 more...

It is very quick, simple and easy to use .
Recon3 www.ciao.co.uk 8/30/2008 more...

福克斯

满意度 点击查看详情

统计各大论坛网友的发言，
自动计算获得，仅作参考

油耗	38	满意：402 不满：658
安全性	95	满意：219 不满：10
空间	56	满意：255 不满：199
动力	74	满意：412 不满：147
操控	81	满意：610 不满：142
外观	87	满意：699 不满：106
内饰	33	满意：239 不满：480

概览 | 价格 | 品质 | 外形 | 油耗 | 内饰 | 空间 | 安全 | 配置 | 操控 | 精华

思域 小福 观察之后还是决定小福了

本人大学毕业 家里准备年前买辆车 （因为上班地好远哦） 一直看好小福和思域 但是同事们都说鬼子的车安全性能不行 撞成两半的车怎么能开啊 但是小福的内饰的确比思域难看一点 不过看了两天 觉得其..

汽车之家 发布日:2008-12-30 浏览:91 回复:14 车型pk

一辆马路牙子 一辆水沟里

今天中午出去办事 碰到2丫超我车迟狂按喇叭 我一看是一凯悦 后面还跟一中华 当时70左右的时速 （乡下水泥小路 我不让他们是超不了车的）我马上加速到100 2y也紧跟 ..

汽车之家 发布日:2008-12-30 浏览:98 回复:18 同类话题

豪华尊贵不再是奢望，福克斯引领高性价..

岁末降临，福克斯为了庆祝销量正式突破30w辆特别推出了一款1.8l自动豪华纪念版车型，12月15日，这款车正式上市，颠覆了高配置一定高价格的车市固有模式，进一步肯定了09福克斯在性价比上的突..

新浪汽车论坛 发布日:2008-12-30 浏览:4 回复:1 同类话题

福克斯大灯换市光透镜和远光透镜作业,更新共..

废话不讲，看图雾灯：凯梅瑞近光透镜，4300k 飞利浦，国产安定近灯：市光双光透镜，4300k，飞利浦，松下安定远灯：奔驰红外远光透镜卤素[本帖最后由 小何 于 2008-12-28 2..

东莞车迷网 发布日:2008-12-27 浏览:321 回复:22 同类话题

换胎~~

如题,准备换前面两条胎,各位有也好介绍?普力斯通。价钱??顶下先~~~。不记得了。等于没说。你又话换铃.....换了n年了。..

东莞车迷网 发布日:2008-12-29 浏览:80 回复:6 同类话题

又出问题了~~~我难道是传说中的冤大头?

既昨天晚上拔了钥匙大灯不自动关之后今天早上又有问题,打完球回家拐弯时顿时感觉方向很硬一看熄火了,都没感觉就熄火了,还是正在行驶过程中熄火的,打着了又走了,大约3公里以后我要停车的时候挂的一档..

YAHOO! 中国雅虎 人物

网页 生活服务 黄页 资讯 音乐 人物 更多

雅虎生活服务 | 雅虎首页 | 雅虎大全 | 帮助

周杰伦

搜索

周杰伦 人际网 个人资料 明星大家说 重名信息

人物标签：天王 酷 优秀 天才 害羞

好评： 39%

中评： 58%

差评： 3%

发表评论 228条

好 **周杰伦** 的好评有213条 【查看更多 213条】

评论	可信度	来源
让张伟平对周杰伦刮目相看:"周杰伦的确是个很好的演员	★★★★★	来源出处
她说:「我觉得周杰伦很有才华	★★★★★	来源出处
而王力宏评价周董则是个很优秀的艺人。	★★★★★	来源出处
发现周董非常有想法	★★★★★	来源出处
"之前一直听人说周杰伦很酷	★★★★★	来源出处
杰伦是个聪明的小孩	★★★★★	来源出处
"黄秋生说周杰伦很聪明	★★★★★	来源出处
杰伦看起来是很酷的样子	★★★★★	来源出处
而杰伦是一个情感丰富的人	★★★★★	来源出处
"南季妈妈"四个人异口同声的说周杰伦是个真性情的人	★★★★★	来源出处
评价:周杰伦是一个很勤奋的年轻人	★★★★★	来源出处
我告诉儿子周杰伦是个孝顺的孩子	★★★★★	来源出处

差 **周杰伦** 的差评有17条 【查看更多 17条】

评论	可信度	来源
卓远天成的凯旋•周杰伦是一个在音乐上的绝对自恋的人	★★★★★	来源出处
周杰伦在我的印象中简直就是群魔乱舞的典范。	★★★★★	来源出处
"周董过往给人过度自负的印象	★★★★☆	来源出处
周杰伦一定是个自恋的男生	★★★☆☆	来源出处
周杰伦在我的印象中是比较迟钝的艺人。	★★★☆☆	来源出处
周董很是有些"挂名导演"和好色的嫌疑。	★★★☆☆	来源出处
周董当晚给人最大的印象是木讷	★★★☆☆	来源出处
到奇幻片的过渡总结了一下,发现周董是个极自恋的人!	★★☆☆☆	来源出处

研究框架


领域知识
情感检索
情感摘要
情感问答
应用层
情感要素抽取
情感倾向性分析
主客观分析/观点文本识别
核心层
NLP基本模块
情感资源收集与标注
基础层
产品评论
电影评论
新闻评论
博客
微博

中国知网情感词典

Release of the latest updated version of HowNet

- Today we release "Chinese/English Vocabulary for Sentiment Analysis (VSA)(Beta version)". The VSA includes 12 subsets.

1. "Chinese Vocabulary for Sentiment Analysis" , which contains 6 sub-files:

"Plus Feeling", e.g.爱，赞赏，快乐，感同身受，好奇，喝彩，魂牵梦萦，嘉许...
"Minus Feeling", e.g.哀伤，半信半疑，鄙视，不满意，不是滋味儿，后悔，大失所望...
"Plus Sentiment", e.g.不可或缺，部优，才高八斗，沉鱼落雁，催人奋进，动听，对劲儿...
"Minus Sentiment", e.g.丑，苦，超标，华而不实，荒凉，混浊，畸轻畸重，价高，空洞无物...
"opinion"
"degree"

2. "English Vocabulary for Sentiment Analysis" , which contains 8945 entri

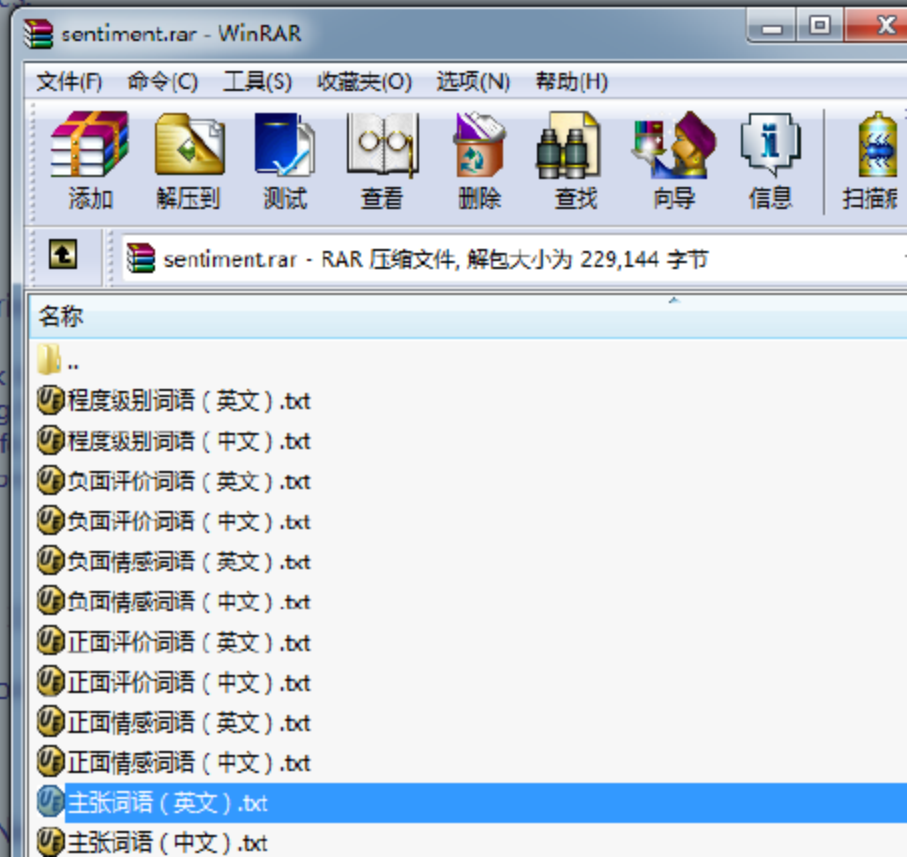
"Plus Feeling", 772 entries, e.g. happy , be jealous , admiration , consent , welcome , look
"Minus Feeling", 1012 entries, e.g. defy , disappointed , fear , criticize , regret , pull a long
"Plus Sentiment", 3596 entries, e.g. good-looking, high-quality , effective , tranquility , saf
"Minus Sentiment", 3562 entries, e.g. grotesqueness , inferior , expensive , expensively , b
"opinion"
"degree"

3. "Chinese/English Vocabulary for Sentiment Analysis" which contains

The "Chinese/English Vocabulary for Sentiment Analysis (VSA)(Beta versio

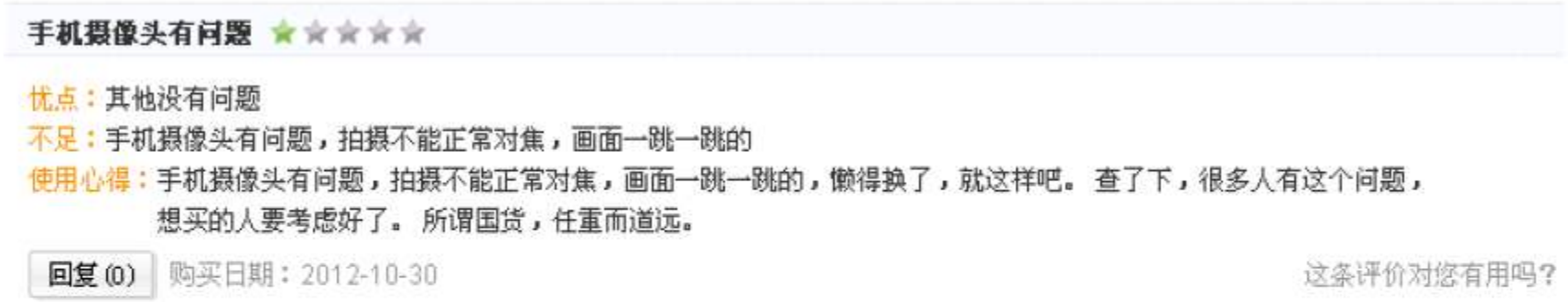
Chinese/English Vocabulary for Sentiment Analysis

- Oct. 08, 2007 Release of the latest updated version of Mini-HowN



利用Web获得评论倾向性

- **利用用户打分获取产品评论倾向**



- **利用表情符号获取文本倾向**

情感分类任务

- **主客观分析/观点文本识别**
  - 客观：反映关于世界的事实信息，“北京是中国的首都”
  - 主观：反映个人情感、信念等，“我爱北京天安门”
- **倾向性分析(可看作主客观分析的细粒度处理)**
  - 对包含观点的文本进行倾向性判断
  - 一般为以下三类
    - 褒义：“外观不错”
    - 贬义：“软件目前不丰富”
    - 中性：“我认为中国需要治理环境”
      - 在一些问题中不考虑中性
- **情绪分析**
  - 用户情绪识别：愤怒、高兴、喜好、悲哀、吃惊，等
- **粒度**
  - 词、句子、文档

情感分类方法

- **基于规则的方法**
  - **利用情感词典、模板**
- **基于机器学习的方法**
  - **利用标注语料训练分类器**

基于情感词计数的情感分类

- **步骤**
  - **正面情感词(如“好”)倾向值为+1**
  - **负面情感词(如“差”)倾向值为-1 (或-2)**
  - **文本的情感倾向值等于词语情感倾向值之和**
  - **如果情感词之前有否定词(如“不”)，那么情感倾向取反**
  - **如果情感词之前有强调词(如“非常”)，那么情感倾向值翻倍**

这个相机拍照效果不好，整体性能非常差。

=> 情感倾向值为-1+2*(-1) = -3

Input: *a review* rev^k *in the kth language. Four lexicons in the kth language: Positive_Dic*^{k}*, Negative_Dic*^{k}*, Negation_Dic*^{k}*, Intensifier_Dic*^{k}*, which are either Chinese or English lexicons;*

Output: *Polarity Value* $f^k_{SO}(rev^k)$*;*

Algorithm Compute_SO:

1. *Tokenize review* rev_k *into sentence set S and each sentence* $s \in S$ *is tokenized into word set* W_s*;*
2. *For any word w in a sentence* $s \in S$*, compute its SO value SO(w) as follows:*
   1) *if* $w \in$ *Positive_Dic*^{k}*, SO(w)=PosValue;*
   2) *If* $w \in$ *Negative_Dic*^{k}*, SO(w)=NegValue;*
   3) *Otherwise, SO(w)=0;*
   4) *Within the window of q words previous to w, if there is a term* $w' \in$ *Negation_Dic*^{k}*, SO(w)= –SO(w);*
   5) *Within the window of q words previous to w, if there is a term* $w' \in$ *Intensifier_Dic*^{k}*, SO(w) =*$\rho \times SO(w)$*;*
3. $f^k_{so}(rev^k) = \sum_{s \in S} \sum_{w \in W_s} SO(w)$;

Figure 2. The algorithm for semantic orientation value computation

Input: *a review* rev^k *in the kth language. Four lexicons in the kth language:* $Positive\_Dic^k$, $Negative\_Dic^k$, $Negation\_Dic^k$, $Intensifier\_Dic^k$, *which are either Chinese or English lexicons;*

Output: *Polarity Value* $f^k_{SO}(rev^k)$;

Algorithm Compute_SO:

1. *Tokenize review* rev_k *into sentence set S and each sentence* $s \in S$ *is tokenized into word set* W_s;
2. *For any word w in a sentence* $s \in S$, *compute its SO value SO(w) as follows:*
   1) *if* $w \in Positive\_Dic^k$, $SO(w)=PosValue$;
   2) *If* $w \in Negative\_Dic^k$, $SO(w)=NegValue$;
   3) *Otherwise,* $SO(w)=0$;
   4) *Within the window of q words previous to w, if there is a term* $w' \in Negation\_Dic^k$, $SO(w)= -SO(w)$;
   5) *Within the window of q words previous to w, if there is a term* $w' \in Intensifier\_Dic^k$, $SO(w) = \rho \times SO(w)$;
3. $f^k_{SO}(rev^k) = \sum_{s \in S} \sum_{w \in W_s} SO(w)$;

Figure 2. The algorithm for semantic orientation value computation

基于机器学习的情感分类

- 看作是特殊的文本分类任务
- 文档采用标准的特征向量表示
  - 特征包括：unigram, bigram, POS, sentiment lexicon, etc.
- 实验
  - Data : movie reviews (Internet Movie Database), rating -> negative, neutral, positive
  - Naïve Bayes, Maximum Entropy, Support Vector Machine

Features	# of features	Frequency or presence?	NB	ME	SVM
unigrams	16165	freq.	78.7	N/A	72.8
unigrams	16165	pres.	81.0	80.4	82.9
unigrams+bigrams	32330	pres.	80.6	80.8	82.7
bigrams	16165	pres.	77.3	77.4	77.1
unigrams+POS	16695	pres.	81.5	80.4	81.9
adjectives	2633	pres.	77.0	77.7	75.1
top 2633 unigrams	2633	pres.	80.3	81.0	81.4
unigrams+position	22430	pres.	81.0	80.1	81.6

观点提取

"I bought an iPhone a few days ago. It was such a nice phone. The touch screen was really cool. The voice quality was clear too. Although the battery life was not long, that is ok for me. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, and wanted me to return it to the shop. …"

Feature Based Summary:

Feature1: **Touch screen**

Positive: 212

- *The touch screen was really cool.*
- *The touch screen was so easy to use and can do amazing things.*

…

Negative: 6

- The screen is easily scratched.
- I have a lot of difficulty in removing finger marks from the touch screen.

…

Feature2: **battery life**

…

Note: We omit opinion holders

10.3 Web挖掘

10.3.1 Web挖掘概述

1. 什么是Web挖掘

Web挖掘是指从大量的Web文档集合中发现蕴涵的、未知的、有潜在应用价值的、非平凡的模式。

它所处理的对象包括静态网页、Web数据库、Web结构、用户使用记录等信息。

2. Web挖掘与数据挖掘的区别

Web挖掘和数据挖掘有着不同的含义。

Web挖掘的研究对象是以半结构化和无结构文档为中心的Web网页，这些数据没有统一的模式，数据的内容和表示互相交织，数据内容基本上没有语义信息进行描述，仅仅依靠HTML语法对数据进行结构上的描述，可以说Web网页的复杂性远比任何传统的文本文档大。