中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

情感计算
–情感倾向性分析

中国科学院自动化研究所

刘斌
liubin@nlpr.ia.ac.cn

目录

- 背景及意义
- 文本情感分析词典与数据库
- 文本情感特征
- 文本情感识别
- 舆情分析
- 总结

目录

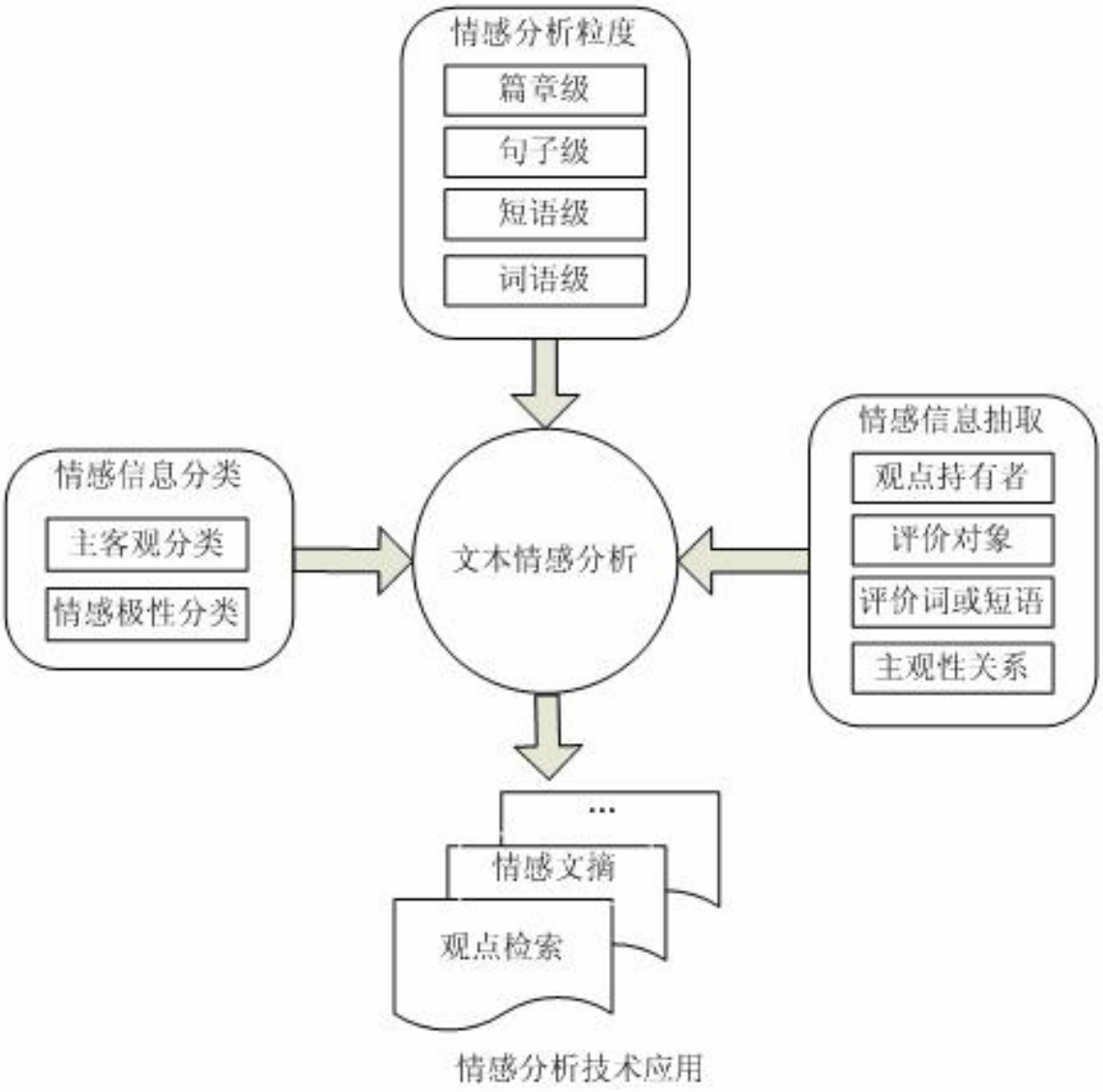
背景及意义

概念

- 情感分析（Sentiment analysis），又称倾向性分析，它是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程；
- 识别用户对事物或人或一句话的看法、态度，即判别用户对评价对象所持有的情感倾向。

背景及意义

- 文本情感分类
  - 情感分析粒度
  - 情感抽取对象
  - 情感信息分类



背景及意义

- 情感分析粒度
  - 按照处理文本粒度的不同，情感分析可分为词语级、句子级、篇章级；
  - 词语级是句子级和篇章级情感分析的基础；
  - 句子级是在词语级上扩展而来，由词语的情感来构成句子的情感；
  - 篇章级情感分析是指定一个整体（如完整的在线评论）的情感极性。

背景及意义

- 情感抽取对象：找到语料中情感的来源或受体，包括：
  - 观点持有者
  - 评价对象
  - 评价词或短语
  - 主观性关系

背景及意义

- 情感信息分类
  - 主客观信息的二元分类
  - 主观信息的情感分类
- 主观信息的情感分类
  - 最常见的褒贬二元分类以及更细致的多元分类
  - 按照极性分类：正向，负向，中性

背景及意义

- 文本情感分析的应用
  - 商品评论（好评，中评，差评）
  - 电影评论
  - 个性化观点挖掘
  - 用户兴趣挖掘

商品评价

电影评论

目录

情感词典

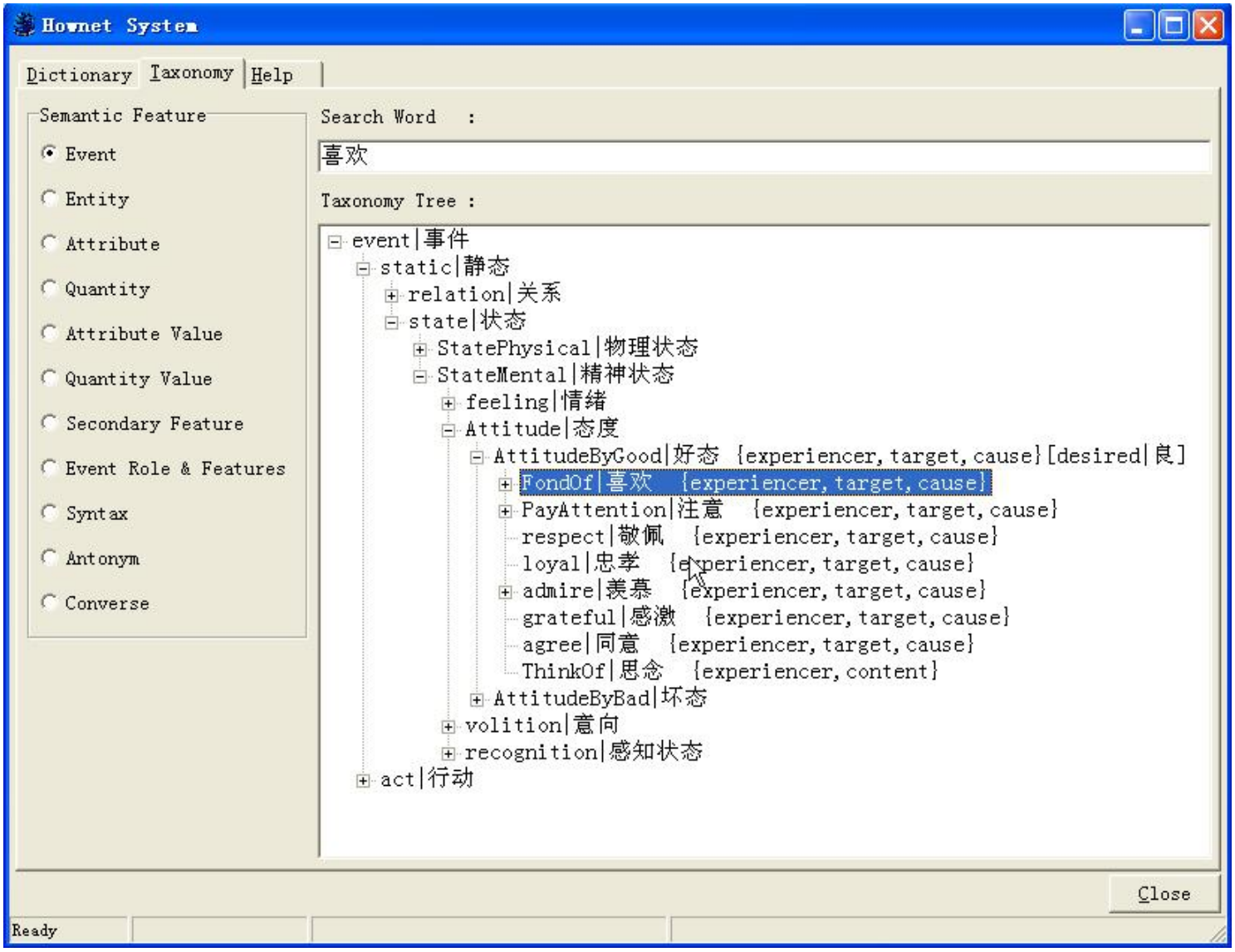
情感词典

- 词典包括情感词典、程度词典、否定词典和连词词典，其中情感词典最为重要，程度词典和否定词典用于识别修饰情感词语的成分，连词词典用于识别句际关系。
- 情感词：高兴、悲伤等
- 程度词：非常、稍微等
- 否定词：没有、木有等
- 连词：然而、不过等
  - 并列连词：前后句子极性一致；
  - 选择连词：前后句子极性一般一致；
  - 递进连词：前后句子极性一般一致，后句稍加强烈；
  - 转折连词：前后句子极性相反，后句更加强烈。

情感词典

- 已有的情感词典主要包括：
  - General Inquirer(GI)词典：该词典1966年开发，在英文文本情感分析中经常被使用；
  - 知网(HowNet)：知网是一个以中、英文词语所代表的概念为描述对象，以概念与概念之间、以及概念所具有的属性之间的关系为基本内容的语言知识库；
  - SentiWordNet：是WordNet英文词典中用于情感分析的词典；
  - 主观词词典：该词典的主观词语来自OpinionFinder 系统；
  - NTU 评价词词典(繁体中文)：该词典由台湾大学收集,含有2,812 个褒义词与8,276 个贬义词。

情感词典


Hownet System
Dictionary
Taxonomy
Help
Semantic Feature
Event
Entity
Attribute
Quantity
Attribute Value
Quantity Value
Secondary Feature
Event Role & Features
Syntax
Antonym
Converse
Search Word :
喜欢
Taxonomy Tree :
event|事件
static|静态
relation|关系
state|状态
StatePhysical|物理状态
StateMental|精神状态
feeling|情绪
Attitude|态度
AttitudeByGood|好态 {experiencer,target,cause}[desired|良]
FondOf|喜欢 {experiencer,target,cause}
PayAttention|注意 {experiencer,target,cause}
respect|敬佩 {experiencer,target,cause}
loyal|忠孝 {experiencer,target,cause}
admire|羡慕 {experiencer,target,cause}
grateful|感激 {experiencer,target,cause}
agree|同意 {experiencer,target,cause}
ThinkOf|思念 {experiencer,content}
AttitudeByBad|坏态
volition|意向
recognition|感知状态
act|行动
Close
Ready

情感词典

- 情感词典的获取
  - 手工方法：Wordnet，Hownet等；
  - 词典的方法：先从种子词典（人工标注的少量情感词典）开始，通过语义相似度计算找种子词典的同义词、反义词；
  - 基于语料库的方法：先从种子词典（人工标注的少量情感词典）开始，通过词语共现度、关系词、Latent Semantic Analysis、double propagation等方法扩展词典。

情感词典

- 英文情感词典
  - General Inquier (http://www.wjh.harvard.edu/~inquirer/)
    - Manually labeled terms (positive, negative)
  - SentiWordnet (http://sentiwordnet.isti.cnr.it/)
    - Extend from WordNet
    - Each synset is automatically labeled as P, N, O
  - OpinionFinder's Subjectivety Lexicon (http://www.cs.pitt.edu/mpqa/)
    - Subjective words provided by OpinionFinder
  - Taboada and Grieve's Turney adjective list
    - Available through Yahoo SentimentAI group. 1700 words
  - IBM Lexicon
    - 1,267 positive words and 1,701 negative words (Melville 2009)

情感词典

- 中文情感词典
  - Hownet (http://www.keenage.com/html/e_index.html)
    - 正面情感、负面情感、正面评价、负面评价、程度级别、主张词语6个子集
  - NTU Sentiment Lexicon (http://nlg18.csie.ntu.edu.tw:8080/opinion/userform.jsp)
    - List the polarities of many Chinese words

文本情感分析数据库

英文情感语料

- MPQA (http://www.cs.pitt.edu/mpqa/databaserelease/)
  - 535 news articles (subjective, objective; P,N,O)
- Movie review data (http://www.cs.cornell.edu/people/pabo/movie-review-data/)
  - IMDB
  - Document level 2000
  - Sentence level 5000
- Custom review data (http://www.cs.uic.edu/~liub/FBS/CustomerReviewData.zip)
  - Product reviews (Product features, P,N)
- Multi product reviews (http://john.blitzer.com/software.html)
  - Book, Electronic, Kitchen, DVD
  - 2000 in each domain
- TREC Blog corpus (http://trec.nist.gov/)
  - Blog data
  - 3,000,000 Webpages
- Multiple-aspect restaurant reviews
  - 4,488 reviews
  - Each review labeled as 1-5 stars

文本情感分析数据库

- 中文情感语料
  - **ChnSentiCorp_htl_all数据集**
    - 7000 多条酒店评论数据，5000 多条正向评论，2000 多条负向评论
  - **waimai_10k数据集**
    - 某外卖平台收集的用户评价，正向4000 条，负向约 8000 条
  - **online_shopping_10_cats数据集**
    - 10 个类别（书籍、平板、手机、水果、洗发水、热水器、蒙牛、衣服、计算机、酒店），共 6 万多条评论数据，正、负向评论各约 3 万条
  - **weibo_senti_100k数据集**
    - 10 万多条，带情感标注 新浪微博，正负向评论约各 5 万条。
  - **simplifyweibo_4_moods数据集**
    - 36 万多条，带情感标注 新浪微博，包含 4 种情感，其中喜悦约 20 万条，愤怒、厌恶、低落各约 5 万条
  - **下载地址：https://github.com/SophonPlus/ChineseNlpCorpus/raw/master/datasets/**

目录

文本情感特征

- 常见的文本表示模型有：
  - 向量空间模型（Vector Space Model）
  - 布尔模型（Boolean Model）
  - 词向量模型（Word Vector Model）

文本情感特征

- 向量空间模型例子

  含40个词，即词表大小为40

1958 2008 奥林匹克 北京 博弈 场地 创 创建
大学 的 第四 第五 东亚 夺冠 高校 计算机 奖
牌 届 锦 标赛 军团 理工 男女 年 排球 设立
是 双双 体育 馆 新高 学子 于 预赛 运动会 在
之一 中 中国 专业 总数 最早

文本情感特征

- 向量空间模型例子
  - 例如 one-hot：

“话筒”表示为：[0，0，0，1，0，0，0，0，0，0，0，…]

“麦克”表示为：[0，0，0，0，0，0，0，0，0，1，0，…]

  - 实现时可以用0, 1, 2, 3等对词语进行计算，这样的“话筒”可以用4表示，麦克可以用10表示实现时

文本情感特征

- 优势
  - 模型结构比较简单，可以进行近似匹配和部分匹配，其匹配结果可以进行排序；
- 不足
  - 维度灾难：维度很大，当词汇较多时，可能会达到百万维；
  - 词汇鸿沟：任意两个词之间都是孤立的，不能体现词与词之间的关系。

文本情感特征

- 定义
  - 布尔模型：以集合论和布尔代数为基础的进行严格匹配的检索模型；
  - 将向量空间模型中的权重限制为0或1，0表示该特征不存在，1则表示该特征存在。

查询“movie ”
以及 “great”

文档	movie	great	bad	love
D1	1	0	0	1
D2	1	0	1	0
D3	0	1	0	1
D4	1	1	0	0

匹配成功返回D4

文本情感特征

- 优势
  - 结构简单，检索速度较快；
- 不足
  - 不能反映不同特征词语对文档的贡献程度；
  - 匹配较为严格，匹配结果无法排序。

文本情感特征

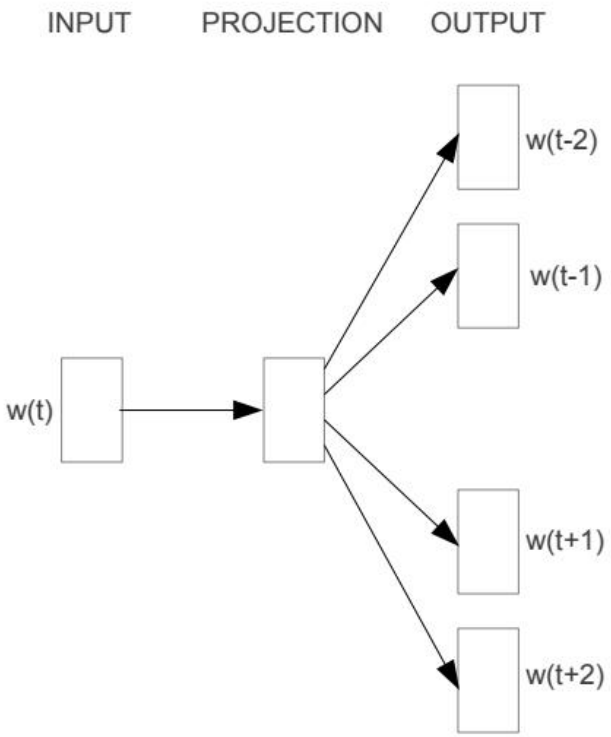
- 词向量定义
  - 词向量（Word Embedding）也叫词嵌入技术或者分布式表示（Distributional Representation）；
  - 词向量是一个将单词转换成向量形式的工具。把对文本内容的处理简化为向量空间中的向量运算，计算出向量空间上的相似度，表示文本语义上的相似度。

文本情感特征

- 词向量表示方法：
  - 将词表示为 [0.793, -0.177, -0.107, 0.109, 0.542, …]的矩阵，通常该类矩阵设置为50维或100维；
  - 通过计算向量之间的距离，来体现词与词之间的相似性，解决词汇鸿沟的问题；
- 词向量训练工具：
  - Word2vec、Gensim、Glove等。

文本情感特征

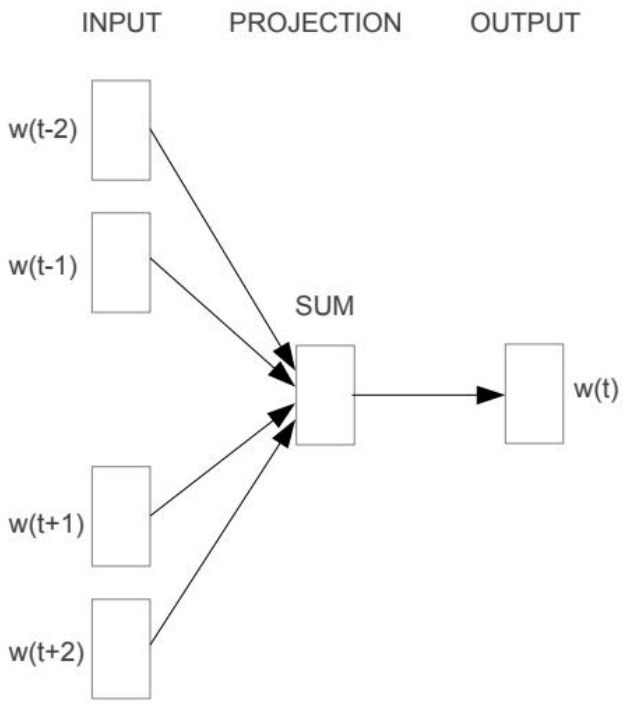
- Word2vec: 常用Skip-gram 和 CBOW 模型计算
  - Skip-gram模型：用一个词语作为输入，来预测它周围的上下文



Skip-gram

文本情感特征

- Word2vec: 常用Skip-gram 和 CBOW 模型计算
  - CBOW模型：拿一个词语的上下文作为输入，来预测这个词语本身



CBOW

文本情感特征

- 特征降维
  - 如果使用文本中所有的字，词语或短语作为特征，那么特征项的个数会达到上万，甚至十几万个，会导致维度灾难；
  - 部分特征对分类结果贡献很小，可以去掉来提高分类器的效率。

文本情感特征

- 文本特征提取的方法
  - 文档频率法
  - 信息增益法
  - 卡方统计法

文本情感特征

- 文档频率法：出现某个特征项的文档频率
  - 当该特征项的DF值小于某个阈值时，该特征项使文档出现的频率太低，没有代表性；
  - 当该特征项的DF值大于另外一个阈值时，该特征项使文档出现的频率太高，没有区分度；
  - 优点：降低向量计算的复杂度，并可能提高分类的准确率，因为按这种选择方法可以去掉一部分噪声特征；
  - 缺点：某些特征虽然出现频率低，但往往包含较多的信息，对于分类的重要性很大；这类特征容易滤除掉。

文本情感特征

- 信息增益法
  - 依据某特征项为整个分类所能提供的信息量多少来衡量该特征项的重要程度，从而决定对该特征项的取舍；
  - 信息增益是不考虑任何特征时文档的熵与考虑该特征后文档的熵的差值；
  - 许多信息增益比较高的特征出现频率往往较低，当使用信息增益选择的特征数目比较少时， 会存在数据稀疏问题。

文本情感特征

- 卡方统计法：卡方值可以衡量词与类别的相关程度
  - 观察实际值与理论值的偏差来确定理论正确性；

    假设理论值是E，实际值是x, xi表示样本

$$\sum_{\mathrm{i}}^{n} \frac{(x_i - E)^2}{E}$$

  - 如果差值很大，则认为与原假设（独立假设）不符合，认为词与类别很相关；
  - 只考虑了词是否出现，而没有考虑出现了多少次，容易夸大低频词的价值。

目录

- 背景及意义
- 文本情感分析词典与数据库
- 文本情感特征
- 文本情感识别
- 舆情分析
- 总结

文本情感识别

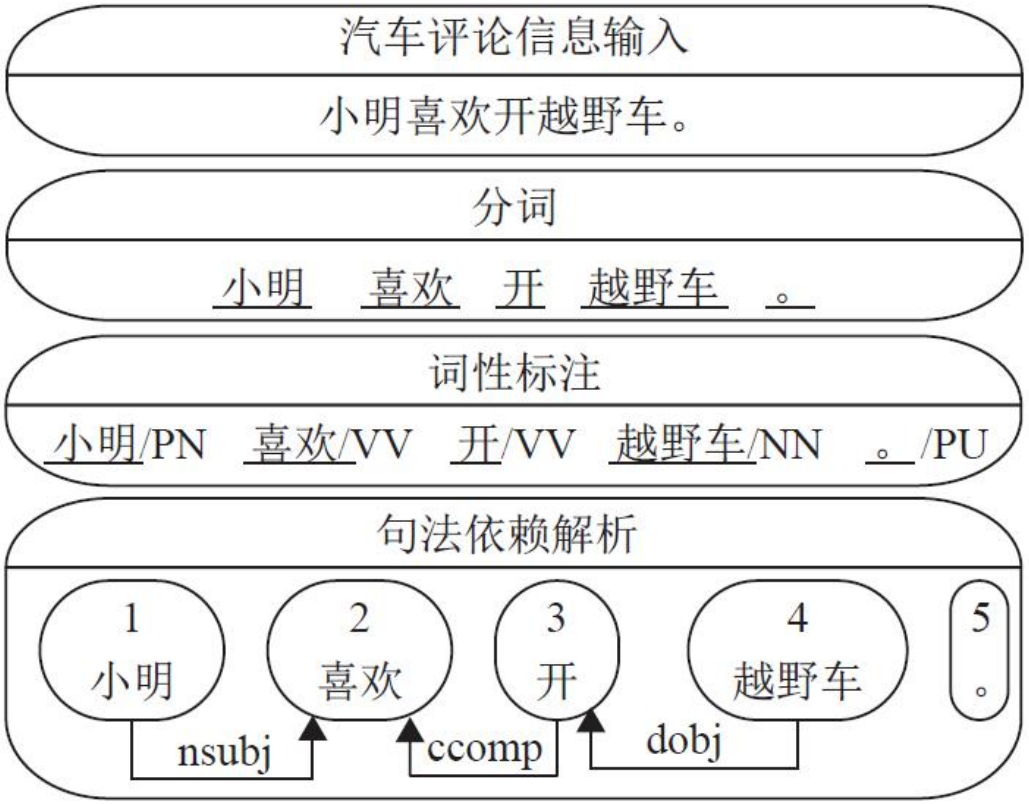
- 文本情感特征分析方法主要包含四种
  - 基于情感词典的文本情感识别方法
  - 基于统计的机器学习文本情感识别方法
  - 基于深度学习的文本情感识别方法
  - 基于预训练模型的文本情感识别方法

文本情感识别

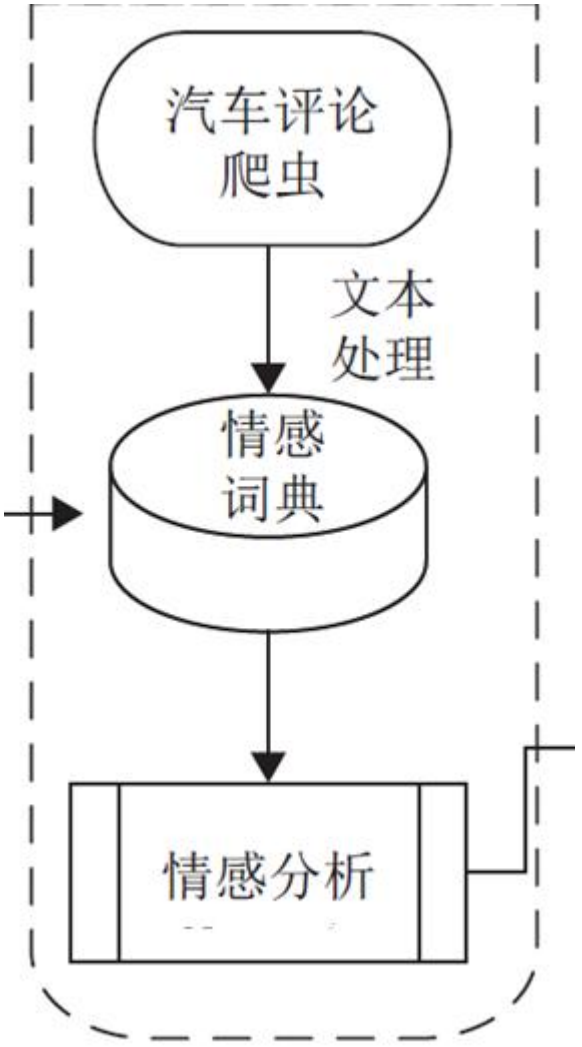
- 基于情感词典的文本情感识别方法
  - 主要以情感词典为基础，通过判断文本中是否出现该情感词来判断文本情感；
  - 需要考虑：不同领域下相同的情感词表达有差异；
  - 不能有效地处理带有否定词的情况；
  - 只能处理句子中带有明显情感色彩的词汇的文本，隐含情感信息的文本效果较差。

文本情感识别

- 基于情感词典的系统框架



预处理



情感识别

文本情感识别

- 基于统计的机器学习文本情感识别方法
  - 采用机器学习算法对已标记情感的语料进行训练，再将训练过的分类器用于对未知文本的情感分类；
- 主要方法
  - 朴素贝叶斯
  - 支持向量机
  - 最大熵模型

文本情感识别

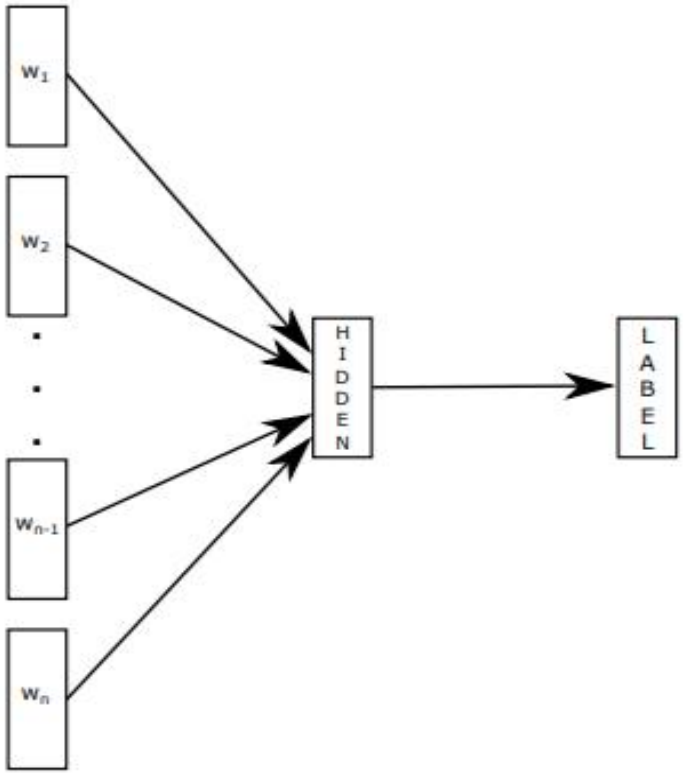
- 优点
  - 不仅考虑情感关键词和其他词汇的倾向性，而且对文本中的标点以及多个词汇同时出现的频率特征进行自动学习；
- 缺点
  - 过于依赖训练语料库的大小，只适用于较长的文档，对句子级别的文本情感分析效果较差。

文本情感识别

- 基于深度学习的文本情感识别方法
  - FastText模型
  - TextCNN模型
  - TextRNN模型
  - TextRNN + Attention模型

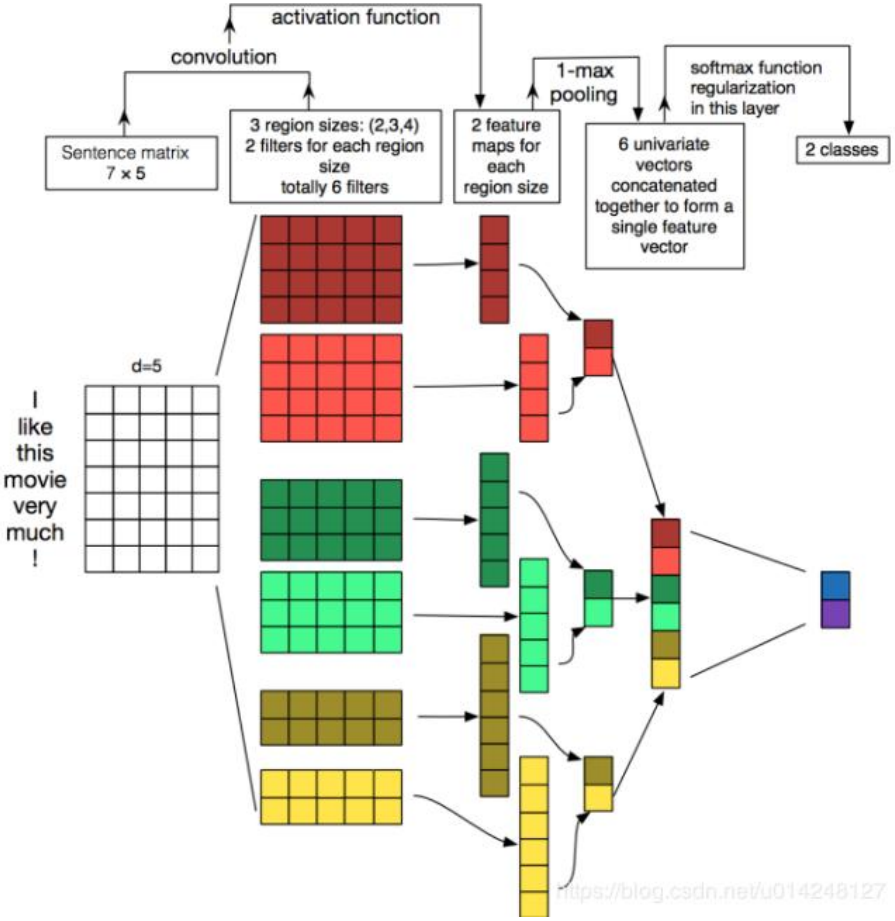
文本情感识别

- FastText模型
  - 句子中所有的词向量进行平均（某种意义上可以理解为只有一个avg pooling特殊CNN），然后直接连接一个softmax 层进行分类
  - 完全没有考虑词序信息



文本情感识别

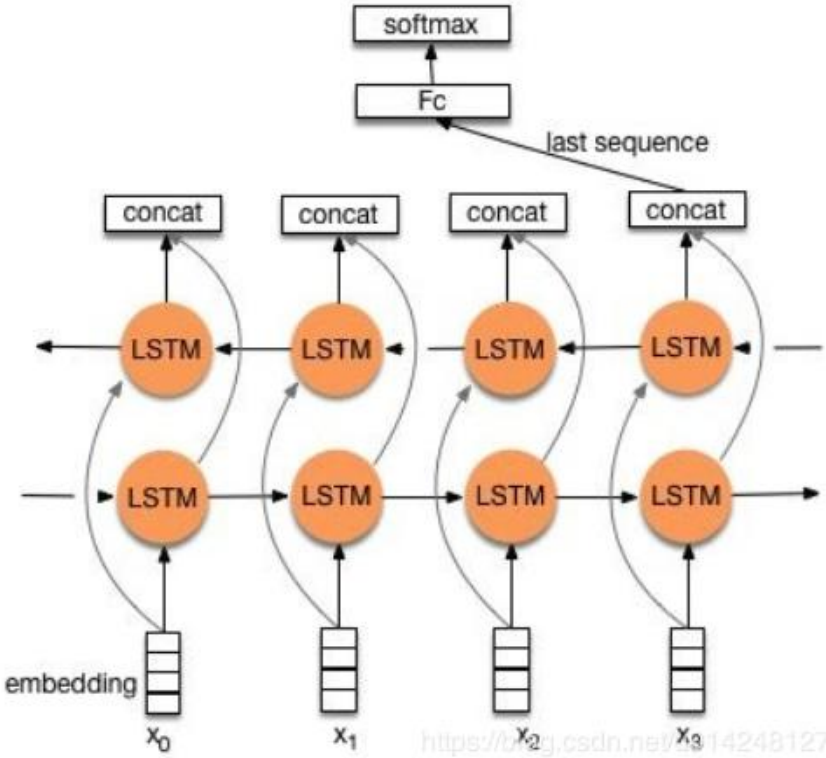
- TextCNN模型
  - 利用CNN来提取句子中类似 n-gram 的关键信息



文本情感识别

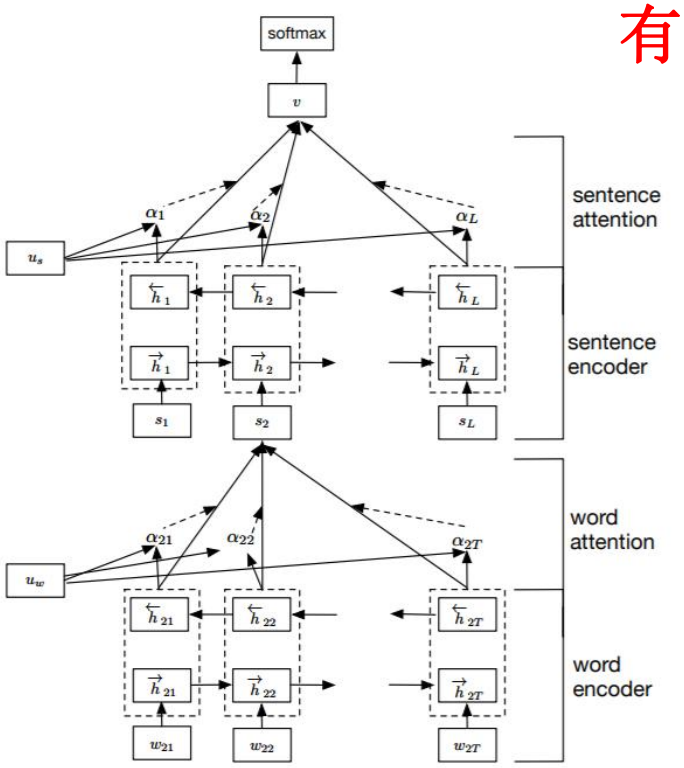
- TextRNN模型
  - 双向LSTM从某种意义上可以理解为可以捕获变长且双向的的 “n-gram” 信息

有效利用了序列中的上下文信息



文本情感识别

- TextRNN + Attention模型
  - 注意力（Attention）机制是自然语言处理领域一个常用的建模长时记忆机制，能够直观的给出每个词对结果的贡献



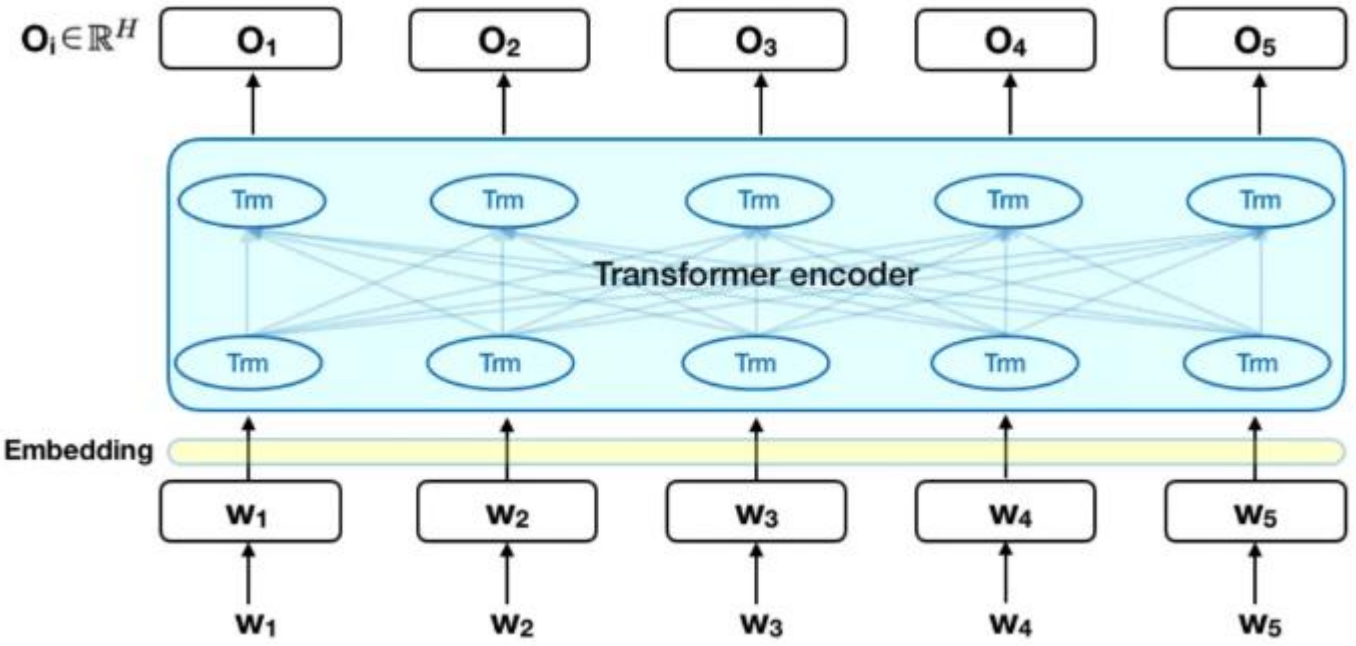
有效融合不同词在序列中的贡献度

文本情感识别

- 基于预训练的文本情感识别方法
  - 基于Bert模型的文本情感识别
  - 基于XLNet模型的文本情感识别

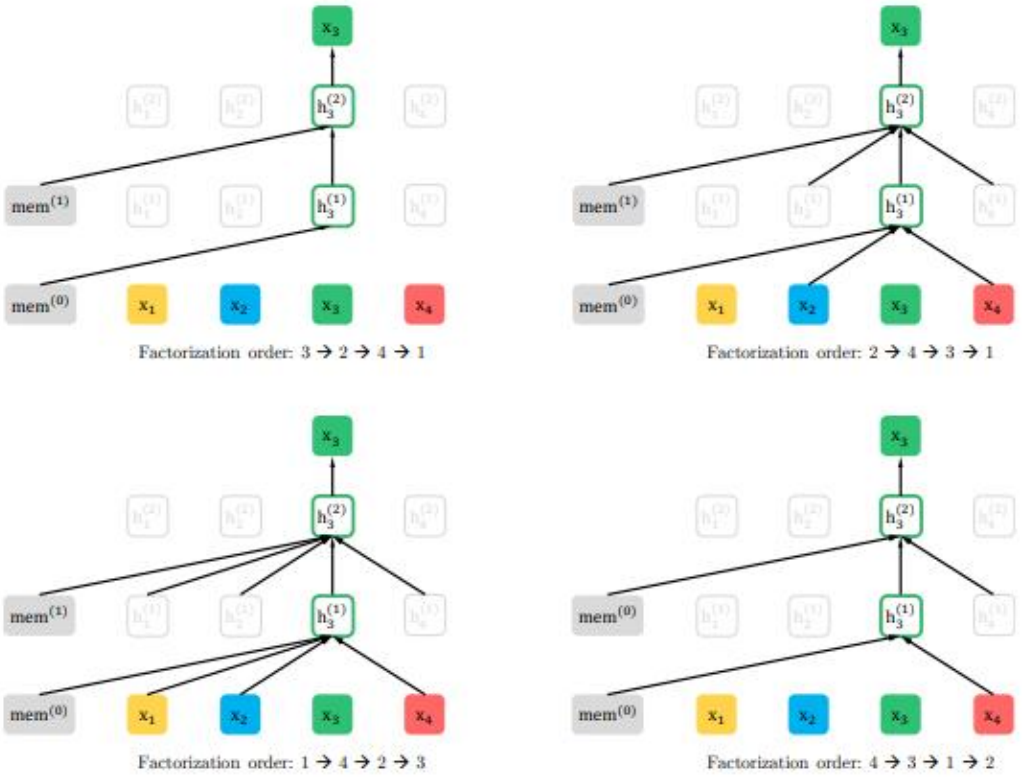
文本情感识别

- 基于Bert模型的文本情感识别
  - BERT是一种预训练语言表示的方法，在大型文本语料库（例如Wikipedia）上训练通用的“语言理解”模型；然后将该模型用于情感识别任务



文本情感识别

- 基于XLNet模型的文本情感识别
  - XLNet模型是在BERT模型基础上 改进，提出的一种泛化自回归预训练方法



目录

- 背景及意义
- 文本情感分析词典与数据库
- 文本情感特征
- 文本情感识别
- 舆情分析
- 总结

舆情分析

- 舆情分析的概念
  - 舆情分析，又称为社交媒体情感分析，基于新闻媒体的情感分析多用于舆论分析，服务于政府部门

舆情分析

- 舆情分析与情感分析的区别
  - 用途不同：基于产品评论的情感分析多用于商业，舆情分析多用于政府部门；
  - 复杂性不同：舆情分析是个比较复杂的系统，涉及更多的技术；
  - 信息来源更广泛：例如，新闻评论、BBS、聊天室、博客、RSS等。

舆情分析

- 舆情分析具有突发性、直接性和偏差性的特点
  - 直接性：通过BBS，新闻点评和博客网站，网民可以立即发表意见，下情直接上达，民意表达更加畅通；
  - 突发性：网络舆论的形成往往非常迅速，一个热点事件的存在加上一种情绪化的意见，就可以成为点燃一片舆论的导火索；
  - 偏差性：由于发言者身份隐蔽，并且缺少规则限制和有效监督，网络自然成为一些网民泄愤情绪的空间。

舆情分析

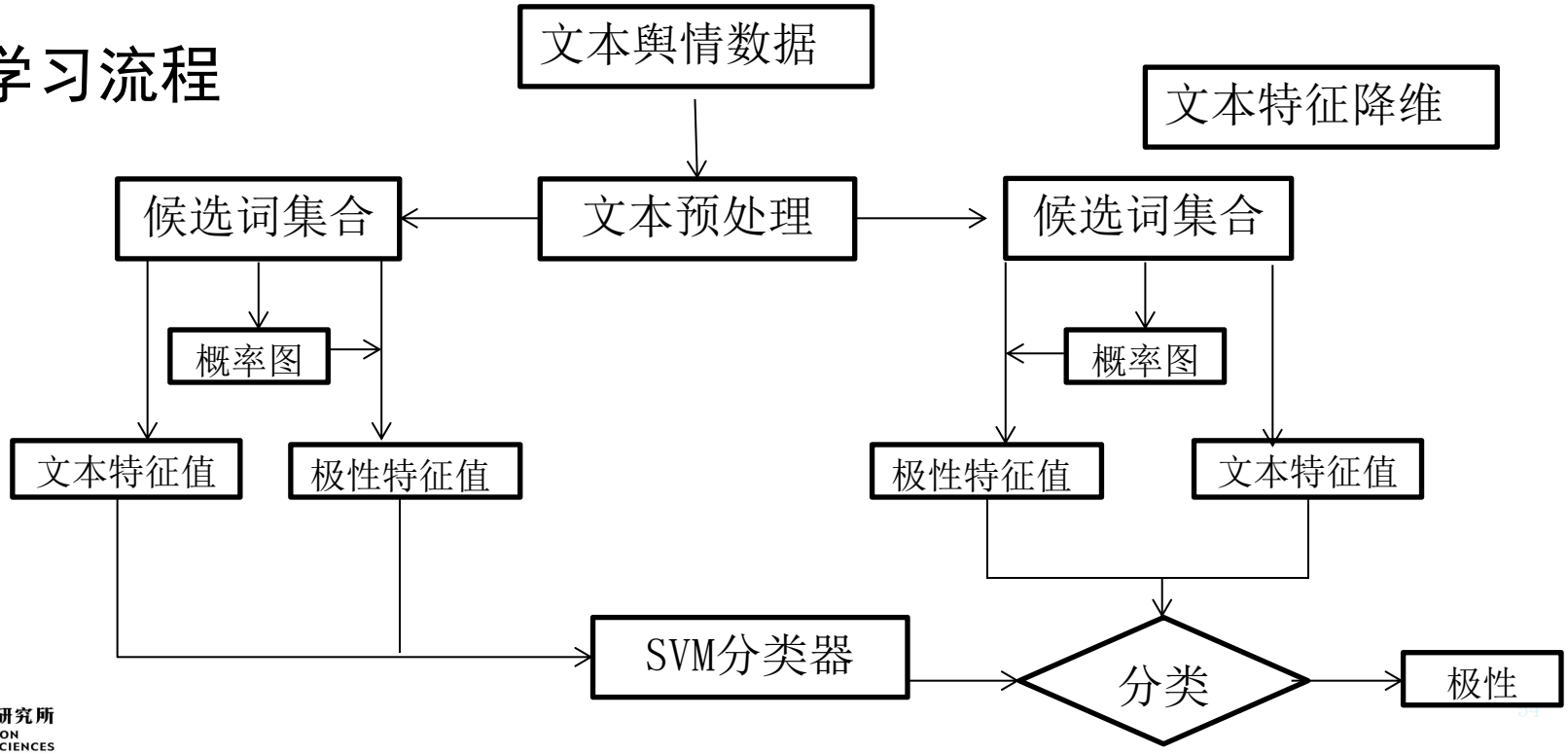
- 舆情分析系统框架
  - 数据采集层：负责从社交媒体中采集资源；
  - 数据处理层：负责对采集的原始数据进行预处理；
  - 报告展示层：舆情分析的结果最终以报告、统计图表等形式展示给用户，为用户下一步决策提供指导依据。

舆情分析

- 舆情具有突发性，通常会用到网络爬虫技术，在社交媒体网站上爬取开源数据
  - 首先从社交媒体网页中抓取用户的链接地址并存放如网页链接地址队列中；
  - 从网页链接地址队列中依次读取待抓取链接地址，访问并下载该页面；
  - 通过解析下载页面，把需要的文本数据以及对应图片保存，同时检测是否有其他用户链接地址；
  - 跳转步骤（2），直到网页链接地址队列为空。

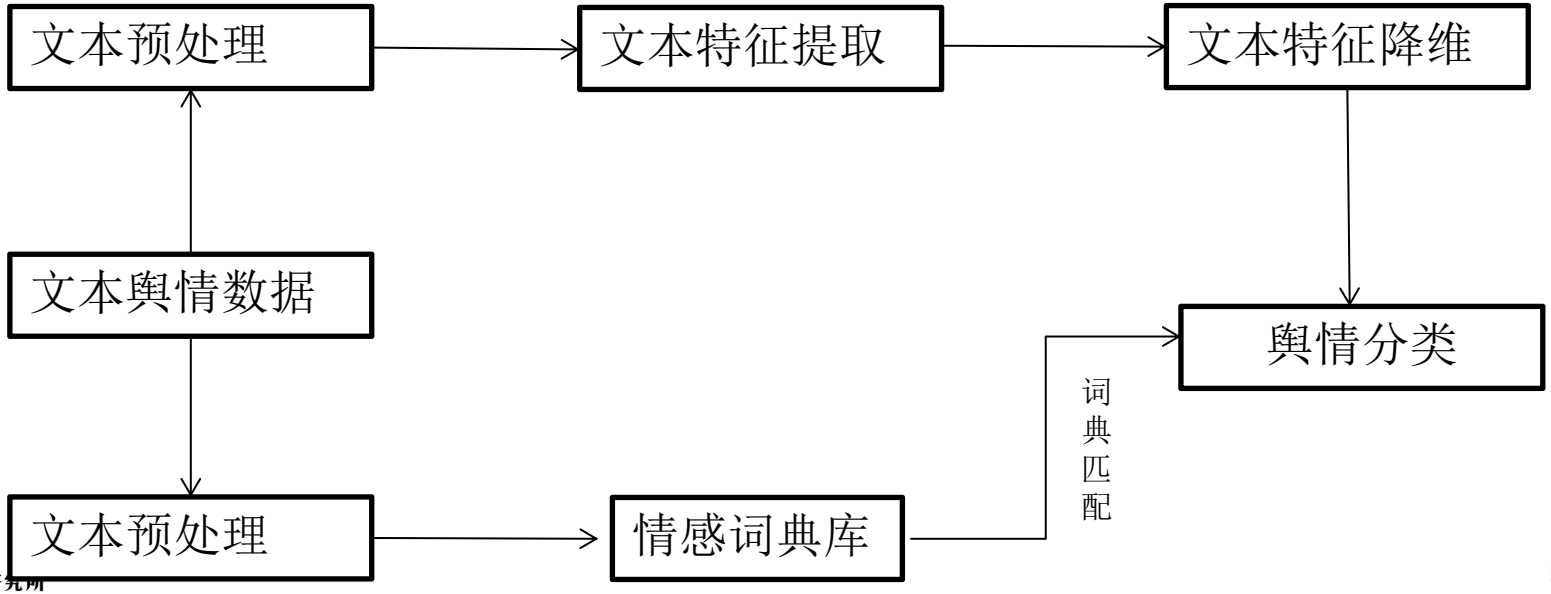
舆情分析

- 基于概率图模型的舆情分析
  - 通过分析训练语料建立一种具有先验概率的图模型，来计算语料中词语的情感概率值，再利用信息熵将概率值归一化为情感特征值，最后用分类器来分类
- 图学习流程



舆情分析

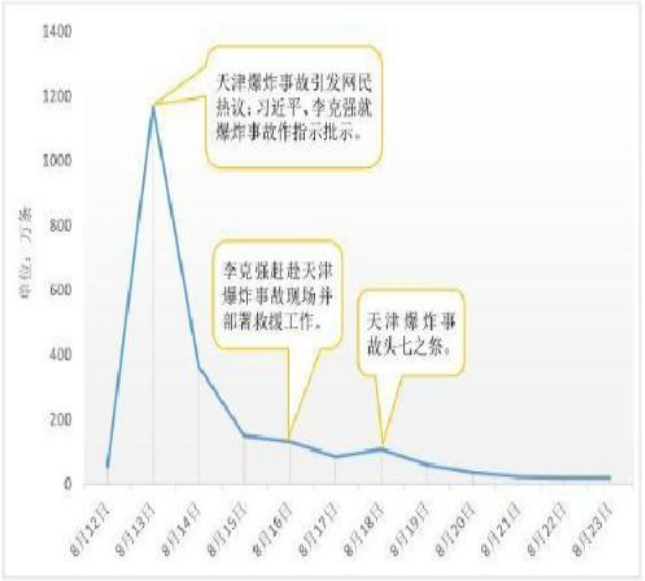
- 定义
  - 混合学习指不再单独采用一种方法进行分析，融合多种方法期望获得更好效果，如利用前文介绍的情感词典和机器学习结合的方法
- 词典—机器学习混合学习流程



舆情分析

舆情分析的应用

热点敏感话题识别

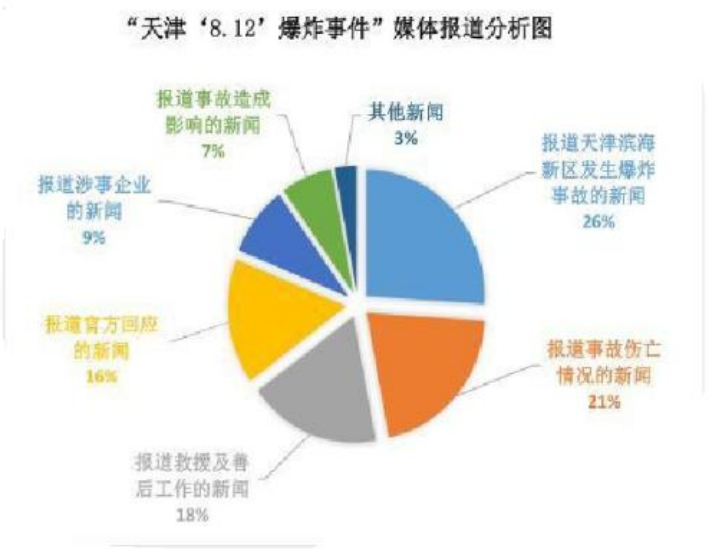


主题跟踪

舆情分析

舆情分析的应用

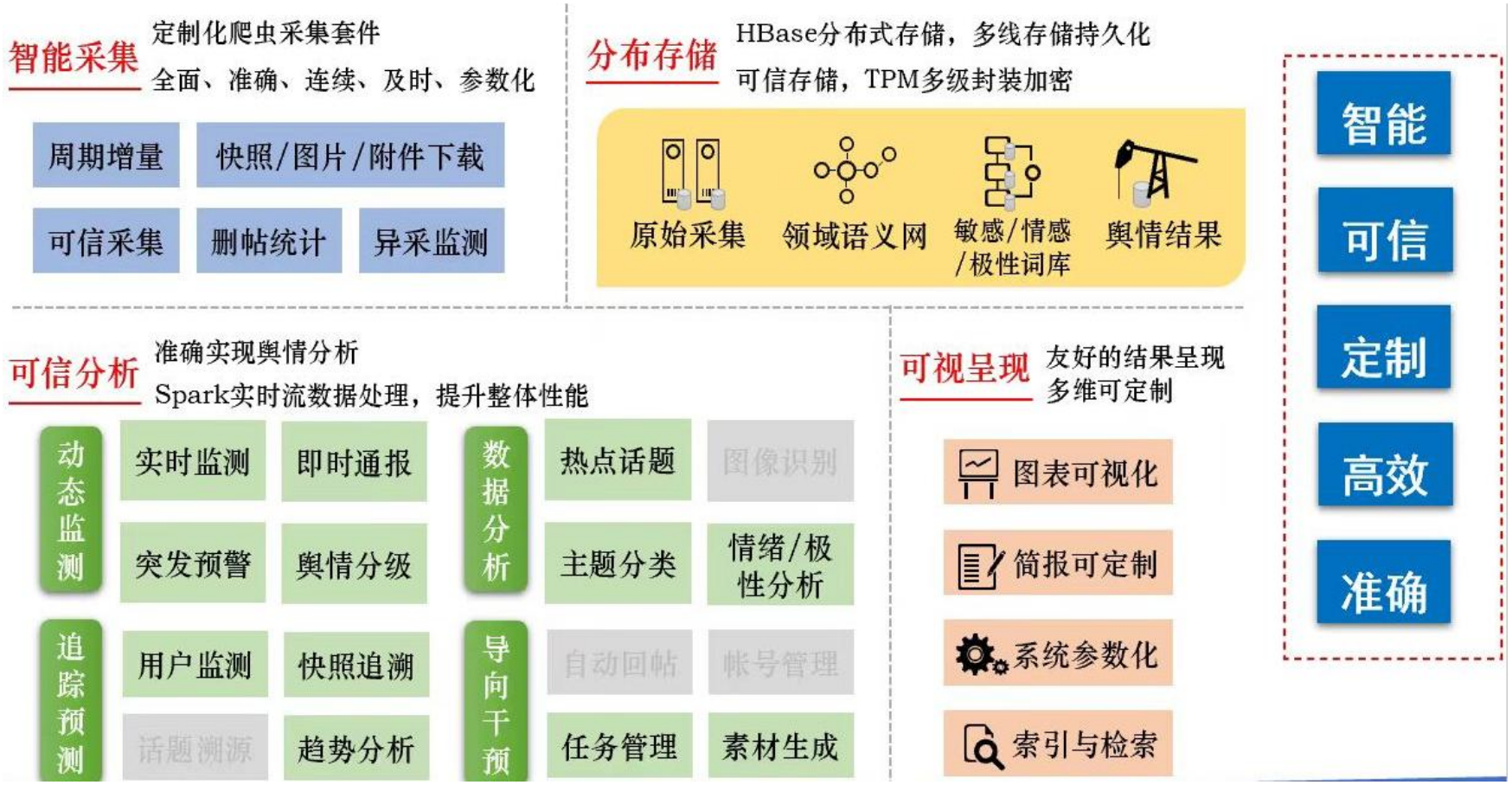
突发事件分析



可视化统计分析

舆情分析

舆情分析的应用



舆情分析

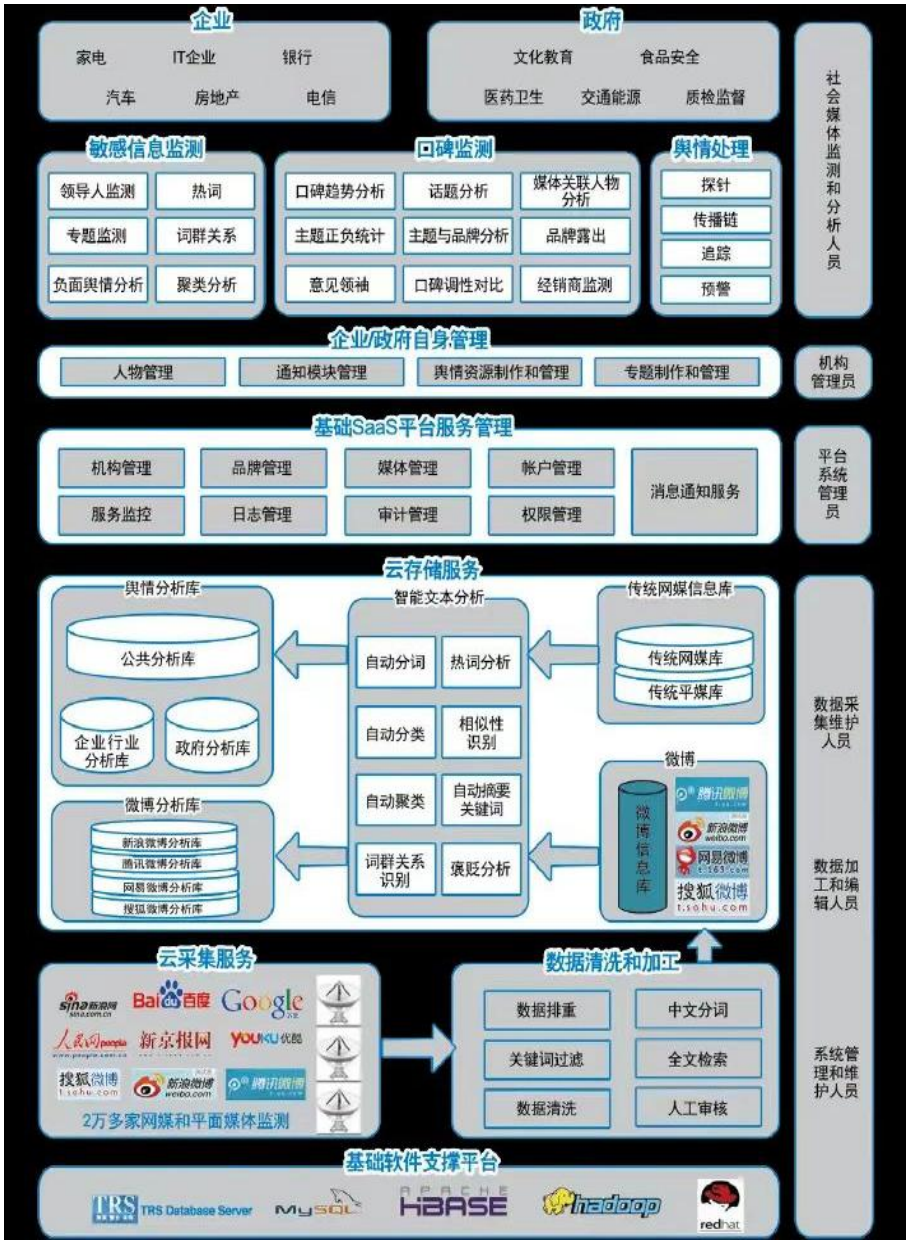
- 舆情分析的应用

舆情分析

- 舆情分析的应用

舆情分析

- 舆情分析的应用



目录

- 背景及意义
- 文本情感分析词典与数据库
- 文本情感特征
- 文本情感识别
- 舆情分析
- 总结

总结

- 文本倾向性分析在商业和政府舆情上都有着很好的应用前景；
- 情感信息的抽取需要充分考虑语境信息；
- 进一步探索融合语义信息的情感分析；
- 面向开源碎片化文本的情感倾向性分析仍然存在着挑战。

Thanks