信息内容安全

王巍

w_wei@hrbeu.edu.cn

泛在网络与信息安全团队

哈爾濱工程大學 HARBIN ENGINEERING UNIVERSITY

第三章 文本内容安全

3.1 文本预处理技术

3.2 文本内容分析

3.3 文本内容安全应用

哈尔滨工程大学 HARBIN ENGINEERING UNIVERSITY

网络信息

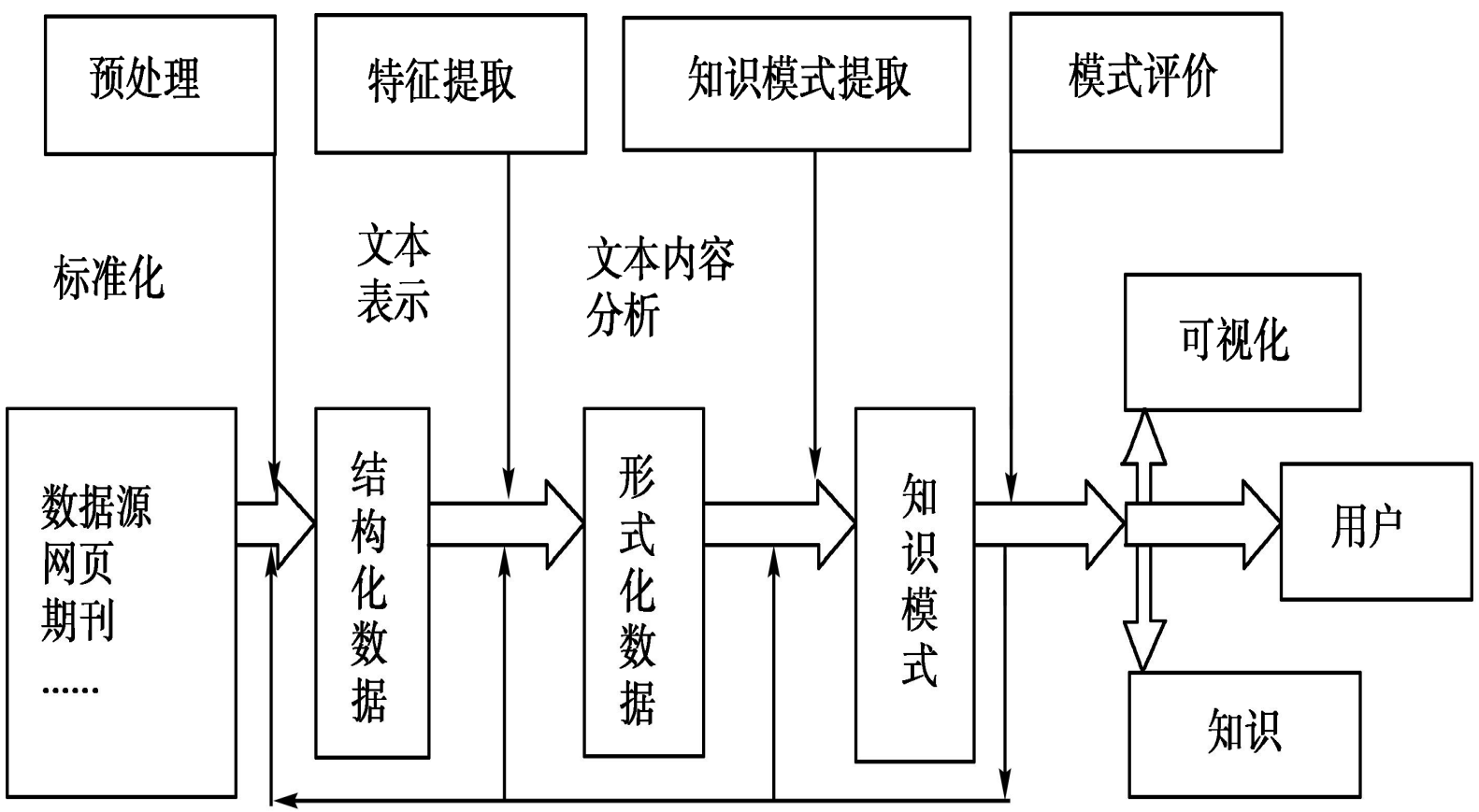
分为文本和多媒体两类。在日常生活中接触到的信息，绝大部分是文本，其呈现方式是印刷品或电子文档。随着互联网的飞速发展，越来越多的文本表现为电子文档形式。

3.1 文本预处理技术

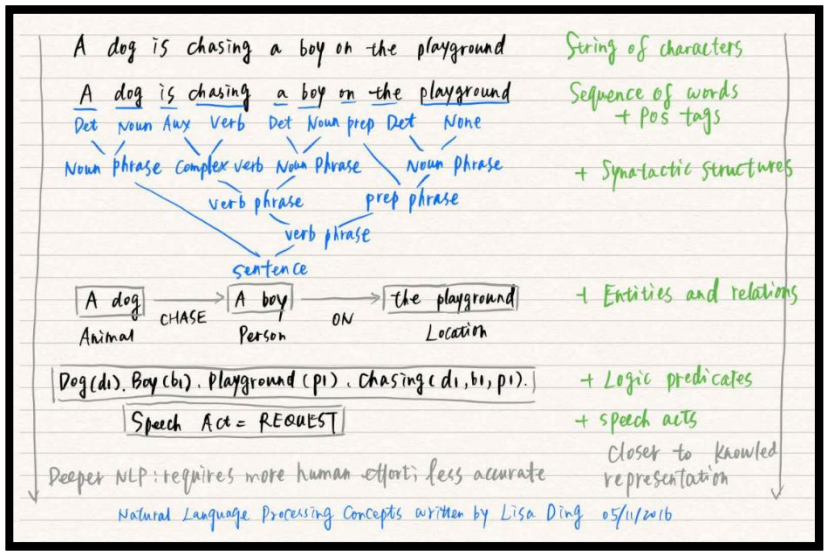
文本概念及处理过程：

文本（text），包括期刊、网页、博客、邮件、短信、微博等外在组织形式的，内涵为纯文字内容的文档对象。文本处理过程一般包括文本预处理、特征提取及缩维、知识模式提取、知识模式评价四个阶段。

文本处理过程



文本预处理技术



- 分词，去停用词
- 特征子集选择与特征重构
- 语义特征提取
- 向量生成和文本内容分析

如何去除和减弱文本信息噪声和变形的影响是文本信息处理软件所遇到的一个重要的问题

停用词

对一篇文本进行中文分词后，其结果是由一组独立的词组成，其间有一些常用的、高频的、对文章的内容判别不起作用的词称为停用词。

a	啊
about	阿
above	哎
across	哎呀
after	哎哟
afterwards	唉
again	俺
	俺们

去停用词

输入：一份文档；输出：去除停用词的文档。

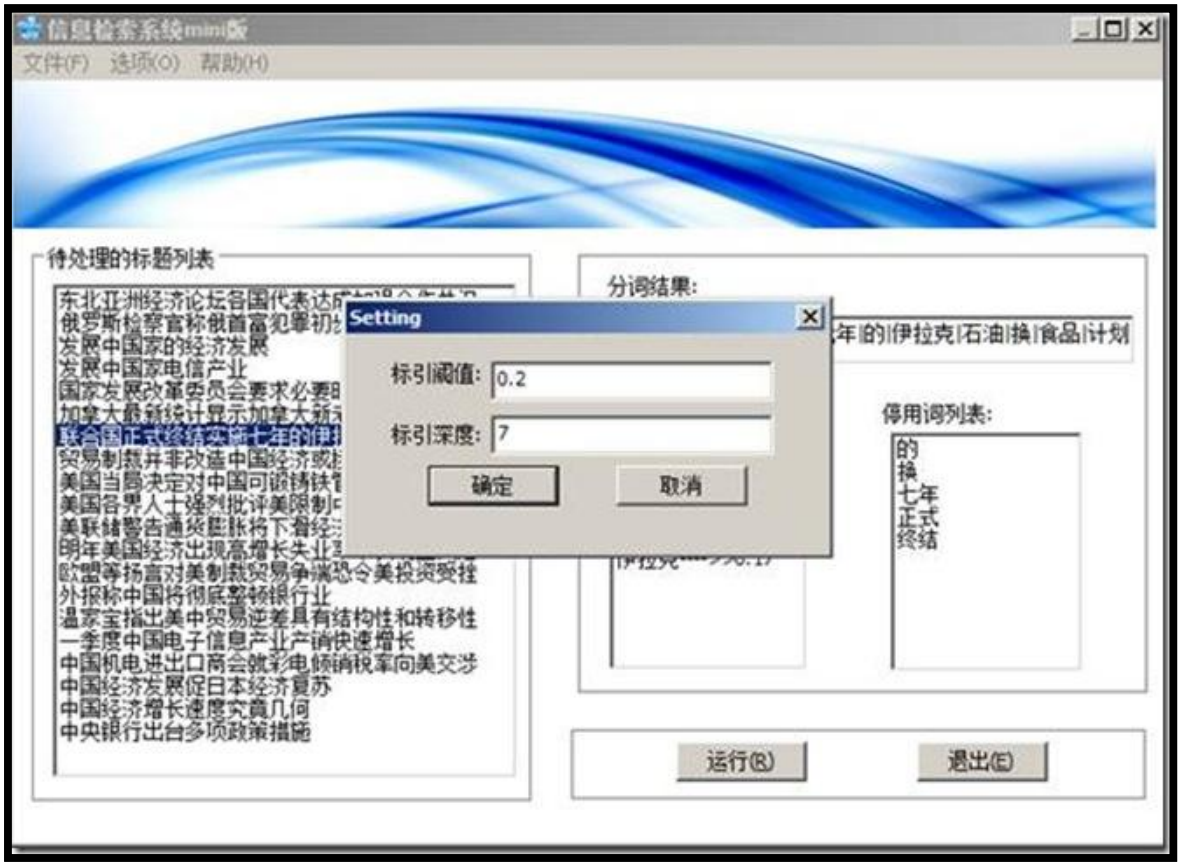
Step 1：中文分词，保留所有的名词、动词、形容词以及副词；

Step 2：保留出现在文档标题或者主题中的数词，过滤其他部位的数词；

Step 3：用停用词表过滤剩余结果；

Step 4：输出结果文档。

去停用词



- 计算机对其处理不但是没有价值的工作，还会增加运算复杂度，通常文本的停用词处理中可采用基于词频的方法将其除去

3.1.1 分词技术

自动分词

中文是以字为基本书写单位，单个字往往不足以表达一个意思，通常认为词是表达语义的最小元素。因此须对中文字符串进行合理的切分。

词是最小的能够独立活动的有意义的语言成分，汉语是以字为基本的书写单位，词语之间没有明显的区分标记，因此，中文词语分析是中文信息处理的基础与关键。

随着搜索引擎技术的广泛应用，全文检索技术和中文分词技术也逐步受到广泛的研究和应用。

自动分词

面临的问题1：分词规范问题

一方面是单字词与词素之间的划界，另一方面是词与短语（词组）的划界。此外，对于汉语“词”的认识，普通说话人的语感与语言学家的标准也有较大的差异。有关专家的调查表明，在母语为汉语的被试者之间，对汉语文本中出现的词语的认同率只有大约70%，从计算的严格意义上说，自动分词是一个没有明确定义的问题。

自动分词

面临的问题2：歧义切分问题

交集型切分歧义：汉字串AJB如果满足AJ、JB同时为词（A、J、B分别为汉字串），则称作交集型切分歧义。此时汉字串J称作交集串。

如“结合成”、“大学生”、“师大校园生活”、“部分居民生活水平”等等。

组合型切分歧义：汉字串AB如果满足A、B、AB同时为词，则称作多义组合型切分歧义。

如“起身”：（a）他站|起|身|来。（b）他明天|起身|去北京。

自动分词

面临的问题3：未登录词问题

未登录词又称为生词（unknown word），可以有两种解释：一是指已有的词表中没有收录的词；二是指已有的训练语料中未曾出现过的词。在第二种含义下，未登录词又称为集外词（out of vocabulary,OOV），即训练集以外的词。通常情况下将OOV与未登录词看作一回事。

如博客、超女、恶搞、房奴、给力、奥特等，尤其在网络用语中这种词汇层出不穷。

特定领域的专业名词和新出现的研究领域名称。如数字中国、区块链等。

自动分词

中文是以字为基本书写单位，单个字往往不足以表达一个意思，通常认为词是表达语义的最小元素。因此须对中文字符串进行合理的切分。

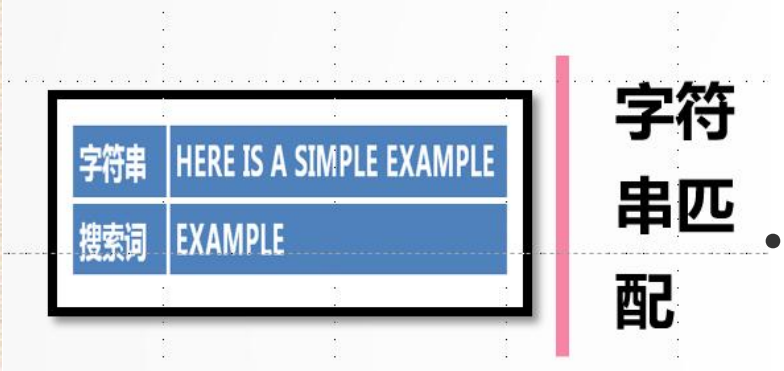
- I am a member of 519 lab in Jiangsu Normal University.

2018/年/3/月/13/日/上午/在/人民大会堂/举行/第十三届/全国人大一次会议/的/第四次全体会议。

自动分词主要分为字符串匹配、基于统计方法和基于理解的方法。

自动分词

字符串匹配方法基本思想：事先建立词库，其中包含所有可能出现的词。字符串匹配分词法又分为最大匹配法和最小匹配法。



- 优点是：分词过程是跟词典作比较，不需要大量的语料库、规则库，其算法简单、复杂性小、对算法作一定的预处理后分词速度较快；
- 缺点是：不能消除歧义、识别未登录词，对词典的依赖性比较大，若词典足够大，其效果会更加明显。

自动分词

基于统计方法基本思想：事先建立一套汉语语法规则，其中的规则不但给出某成分的结构（即它由哪些子成分构成），而且还给出它的子成分之间必须满足的约束条件。

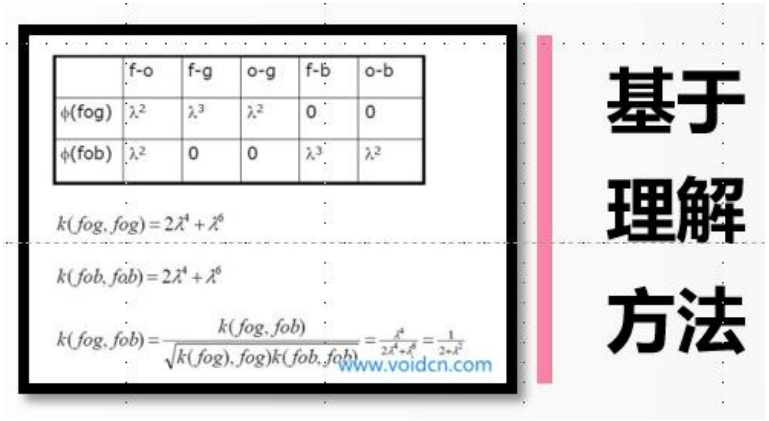
基于统计方法

- 优点是：由于是基于统计规律的，对未登录词的识别表现出了一定的优越性，不需要预设词典；
  缺点是：需要一个足够大的语料库来统计训练，其正确性很大程度上依赖训练语料库的质量好坏，算法较为复杂，计算量大，周期长，但是都较为常见，处理速度一般。

自动分词

基于理解的方法基本思想：在语法分析的基础上建立一个词库，其中包含所有可能出现的词和它们的各种语义信息对给定的待分词的汉语句子s，按照某种确定的原则取s的子串。若该子串与词库中的某词条相匹配，则从词库中取出该词的所有语义信息，然后调用语义分析程序进行语义分析。

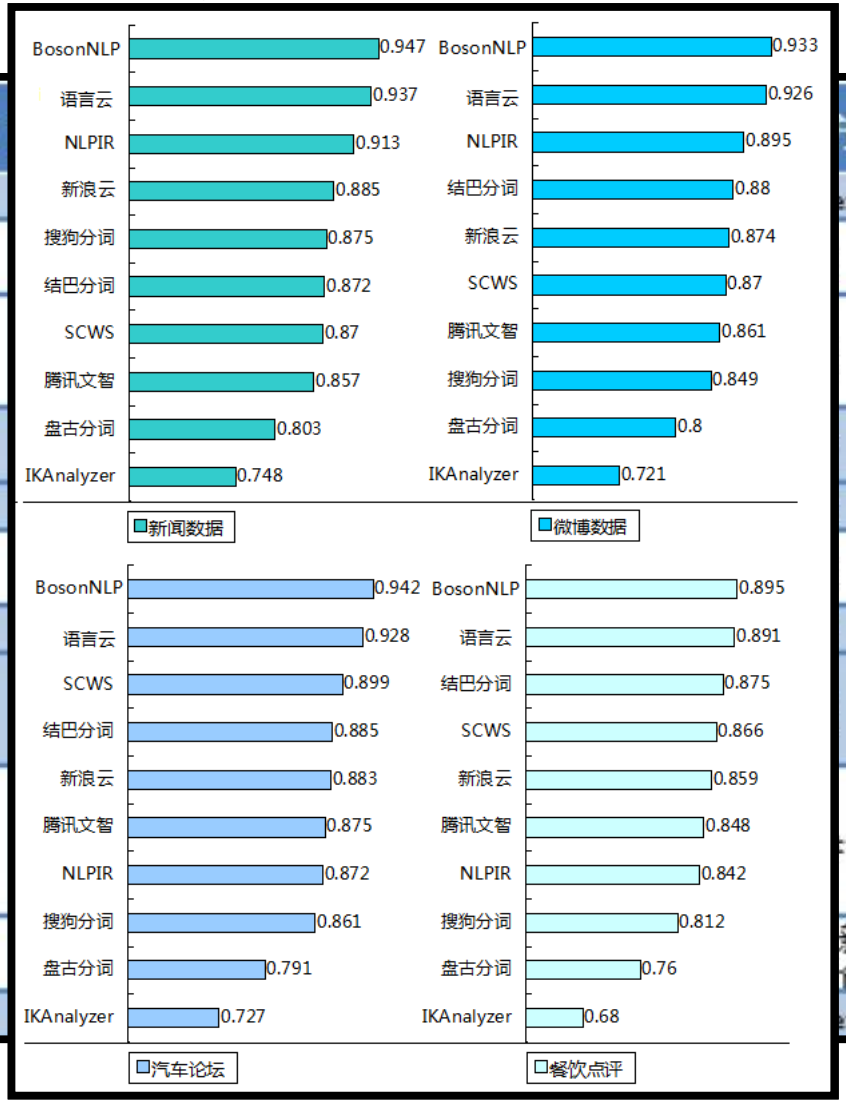
。



- 优点是：由于能理解字符串含义，对未登录词具有很强的识别能力，能很好的解决歧义问题，不需要词典及大量语料库训练；
- 缺点是：需要一个准确、完备的规则库，依赖性较强，效果好坏往往取决于规则库的完整性。算法比较复杂、实现技术难度较大，处理速度比较慢。

常用的分词系统

分词服务	分词粒度
BosonNLP	多选择
IKAnalyzer	多选择
NLPIR	多选择
SCWS	多选择
结巴分词	多选择
盘古分词	多选择
庖丁解牛	多选择
搜狗分词	小
腾讯文智	小
新浪云	大
语言云	适中



方法	接口
en	REST API
	jar包
	多语言接口
	PHP库|命令行工具
	python库
	无
	jar包
	支持上传文档，但是一直失败
ture	REST API
新浪仓库	REST API
en	REST API

3.1.2 文本表示

文本表示

文本结构化的结果称为文本表示。
常见的文本表示模型：

- 布尔模型
- 向量空间模型
- 概率模型
- 潜在语义索引模型
- … …

文本表示模型

布尔模型:

建立在集合理论和布尔运算上的一种简单的检索模型。

向量空间模型:

通过利用向量空间的数据表示和几何运算解决检索中数据表示和相似度量的问题，该模型是由Salton等人提出。

文本表示模型

概率模型:

　　又称为“二值独立检索模型”，是一种基于概率排序原则的自适应模型，其提问不是由用户直接给出，而是通过某种归类学习方法构造一个决策函数来表示提问。

向量空间模型简单、易懂并且在实际应用中非常有效，在知识表示及模型理解上具有巨大的优势。

3.1.3 文本特征提取

文本表示模型

通过构造评价函数，对特征集中的每一个特征进行独立的评估，这样，每一个特征都获得一个评估值，然后对所有的特征按照其评估值的大小进行排序，选取预定数目的最佳特征作为结果的特征子集。

通常采用的评估函数包括信息增益、文档频率、互信息、 交叉熵等。

信息增益

反映词对整个分类提供的信息量，即在获知一个特征在文本中出现或不出现时，所获得的信息的比特数。

设离散随机变量的概率分布P和Q，它们的信息增益定义为

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)}.$$

信息增益是以分布P为权重的P和Q对数差值的加权平均

文档频率

指在训练文本集中包含某一特征的文本的个数。

单词ID	单词	文档频率	倒排列表（DocID;TF;<POS>）
1	谷歌	5	(1;1;<1>),(2;1;<1>),(3;2;<1;6>),(4;1;<1>),(5;1;<1>)
2	地图	5	(1;1;<2>),(2;1;<2>),(3;1;<2>),(4;1;<2>),(5;1;<2>)
3	之父	4	(1;1;<3>),(2;1;<3>),(4;1;<3>),(5;1;<3>)
4	嫉槽	2	(1;1;<4>),(4;1;<4>)
5	Facebook	5	(1;1;<5>),(2;1;<5>),(3;1;<8>),(4;1;<5>),(5;1;<8>)
6	加盟	3	(2;1;<4>),(3;1;<7>),(5;1;<5>)
7	创始人	1	(3;1;<3>)
8	拉斯	2	(3;1;<4>),(5;1;<4>)
9	离开	1	(3;1;<5>)
10	与	1	(4;1;<6>)
11	Wave	1	(4;1;<7>)
12	项目	1	(4;1;<8>)
13	取消	1	(4;1;<9>)
14	有关	1	(4;1;<10>)
15	社交	1	(5;1;<6>)
16	网站	1	(5;1;<7>)

互信息

用来衡量特征与类别之间的统计独立关系，在词语相关性的统计语言模型中广泛使用。

可以看成是一个随机变量中包含的关于另一个随机变量的信息量，或者说是一个随机变量由于已知另一个随机变量而减少的不肯定性

$$I(X;Y)=\sum_{x\epsilon X}\sum_{y\epsilon Y}p(x,y)\,log\frac{p(x,y)}{p(x)\,p(y)}$$

交叉熵

交叉熵（Cross Entropy）是Shannon信息论中一个重要概念，主要用于度量两个概率分布间的差异性信息。语言模型的性能通常用交叉熵和复杂度（perplexity）来衡量。交叉熵的意义是用该模型对文本识别的难度，或者从压缩的角度来看，每个词平均要用几个位来编码。

$$\sum_{i} p(i) \cdot log\left(\frac{1}{q(i)}\right)$$

3.2 文本内容分析

文本内容分析

虽然可以不断提高文本表示模型的效率，但每个文本都是由大量的特征所组成这一事实，导致文本表示维数会达到数十万维的大小，对将要进行的文本内容分析可能带来灾难性的计算时间指数增长，而产生的特征子集分类结果与小得多的特征子集相近。

3.2.1 语义特征抽取

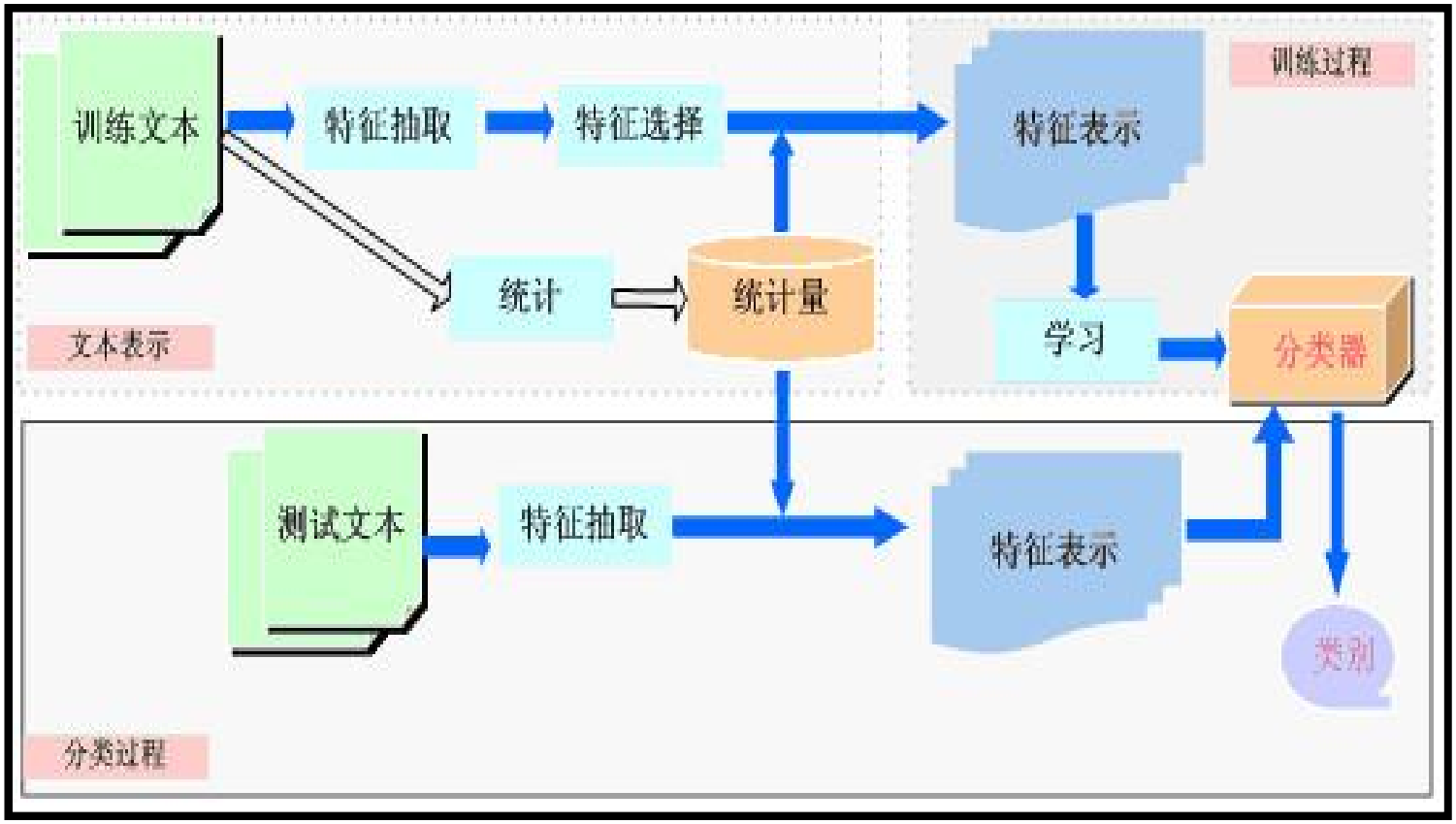
语义特征

语义特征需具备如下特征：

- 特征项要能确实标识文本内容
- 具有将目标文本与其他文本相区分的能力
- 特征项的个数不能太多
- 特征项分离要比较容易实现

- **根据语义级别由低到高来分，文本语义特征可分为：亚词级别、词级别、多词级别、语义级别和语用级别。其中，应用最为广泛的是词级别。**

语义特征



语义特征

有一篇很长的文章，我要用计算机提取它的关键词（Automatic Keyphrase extraction），完全不加以人工干预，怎样才能正确做到?

语义特征

- 假定现在有一篇长文《中国的蜜蜂养殖》，我们准备用计算机提取它的关键词；
- 首先思路，如果某个词很重要，它应该在这篇文章中多次出现。于是，我们进行“词频”统计；
- 我们可能发现"中国"、"蜜蜂"、"养殖"这三个词的出现次数一样多。这是不是意味着，作为关键词，它们的重要性是一样的？

如果某个词比较少见，但是它在这篇文章中多次出现，那么它很可能就反映了这篇文章的特性，正是我们所需要的关键词

语义特征

如下的向量来表示某第二篇，以便于计算机理解和处理。

w2=（文本，5，统计学习，4，模型，0，……）

这个向量表示在w2所代表的文本中，“文本”这个词出现了5次（这个信息就叫做词频），“统计学习”这个词出现了4次，而“模型”这个词出现了0次，依此类推，后面的词没有列出。

系列的第三篇文章可以表示为

w3=（文本，9，统计学习，4，模型，10，……）

其含义同上。如果还有更多的文档需要表示，我们都可以使用这种方式。例如我们的问题就可以抽离出一个词典向量

D=（文本，统计学习，模型，……）

所有的文档向量均可在参考这个词典向量的基础上简化成诸如

w2=（5，4，0，……）

w3=（9，4，10，……）

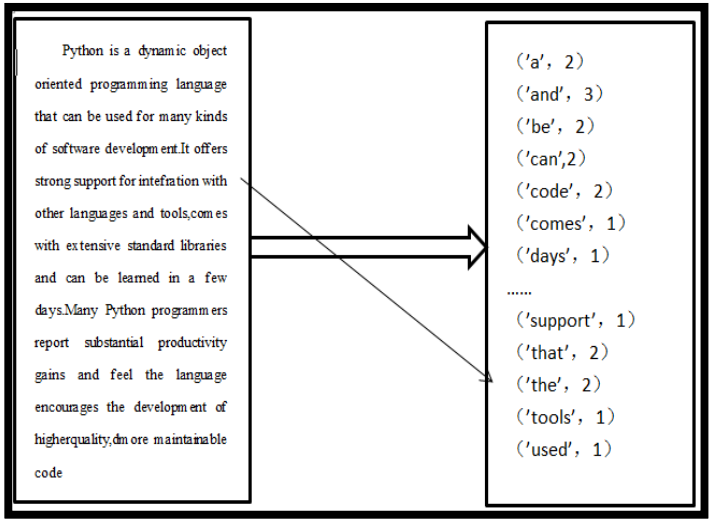
的形式，其含义没有改变。

5，4，10这些数字分别叫做各个词在某个文档中的权重，实际上单单使用词频作为权重并不多见，也不十分有用，更常见的做法是使用地球人都知道的TF/IDF值作为权重。

词级别语义特征

词特征可进行计算的因素有很多，最常用的有词频、词性：

- 词频
- 词性
- 文档、词语长度
- 词语直径
- 首次出现位置



词袋模型

- **词级别(WordLevel)以词作为基本语义特征。以单词作为基本语义特征在文本分类、信息检索系统中工作良好，是最常见的基本语义特征**

亚词级别语义特征

n元模型

- n元模型将文本表示为重叠的n个连续字母(对应汉语情况为单字)的序列作为特征项;
- 采用 n元模型时，需要考虑数值n的选择问题。

多词级别语义特征

- 多词级别 (Multi-Word Leve1) 指用多个词作为文本的特征项，多词可以比词级别表示更多的语义信息;
- 经常从统计角度根据词之间较高的同现频率(Co-OccurFrequency) 来选取特征项

- **亚词级别**(Sub-Word Level)**也称为字素级别**(GraphemicLevel)。**在英文中比词级别更低的文字组成单位是字母，在汉语中则是单字。**

汉语语义特征抽取

表 3-1 文档及特征项各参数含义

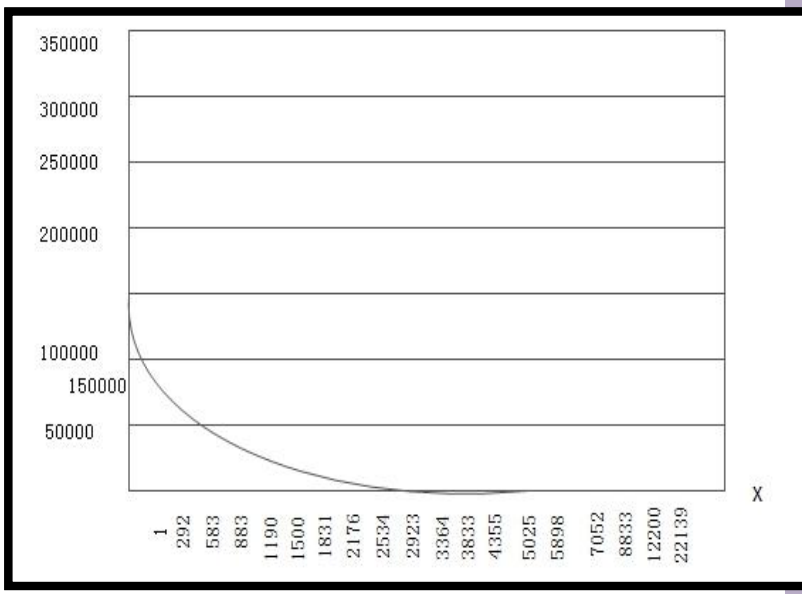
N	训练样本数
n_{c_i}	c_i 类别包含的训练样本数
n(t)	包含特征项 t 至少一次的训练样本数
$\overline{n}(t)$	不包含特征项 t 的训练样本数
$n_{c_i}(t)$	类别包含特征项 t 至少一次的训练样本数
$\overline{n}_{c_i}(t)$	类别不包含特征项 t 的训练样本数
tf	所有训练样本中所有特征项出现的总次数
tf(t)	特征项 t 在所有训练样本中出现的次数
$tf_{d_j}(t)$	特征项 t 在文档中出现的次数

机器学习领域存在多种特征选择方法，Guyon**等人对特征子集选择进行了详尽讨论，分析比较了目前常用的3种特征选择方式：过滤**(filter)**、组合**(wrappers)**与嵌入**(embedded)

文档频率阈值法与齐夫定律

文档频率阈值法

- 用于去除训练样本集中出现频率较低的特征项
- 词频对标识文本类别的重要性
- 齐夫定律
- 单词出现的频率rank(t)与其序号n(t)存在近似反比关系



一个中文预料的齐夫定律现象验证

TF-IDF(特征项频率——逆文本频率指数

第一部分可以用 TF(t)来表示：第二部分采用逆文本频率指数来表示，一个特征项 t 的逆文本频率指数 IDF(t)由样本总数与包含该特征项文档数决定，可得：

$$IDF(t) = \log \frac{n}{n(t)} \tag{3-8}$$

第一部分和第二部分都满足取值越大时，该特征对类别区分能力越强，取两者乘积作为该特征项 TF-IDF 值，可得：

$$TF - IDF(t) = TF(t) \cdot IDF(t) = n(t) \cdot \log \frac{n}{n(t)} \tag{3-9}$$

一般停用词第一部分取值较高，而第二部分取值较低，因此 TF-IDF 等价于停用词和文档频率阈值法两者的综合。

- TF-IDF方法则结合考虑两个部分，第一部分认为，出现次数较多的特征项对分类贡献较大；第二部分认为，如果一个特征项在训练样本集中的大多数样本中都出现，则该特征项对分类贡献不大，应当去除

语义特征提取

假定该文长度为1000个词，“中国”、“蜜蜂”、“养殖”各出现20次，则这三个词的“词频”（TF）都为0.02。假设中文网页有250亿

	包含该词的文档数（亿）	IDF	TF-IDF
中国	62.3	0.603	0.0121
蜜蜂	0.484	2.713	0.0543
养殖	0.973	2.410	0.0482

"蜜蜂"的TF-IDF值最高，"养殖"其次，"中国"最低。（如果还计算"的"字的TF-IDF，那将是一个极其接近0的值。）所以，如果只选择一个词，"蜜蜂"就是这篇文章的关键词。

3.2.2 文本语法分析方法

文本语法分析

如何减少文本特征的维数，避免维度诅咒是一个至关重要问题，主要包含三个方面。

- 1.**词义**
  指通过语言模型或语法模型来处理文本的过程，包括隐马尔科夫（Hidden Markov Model，HMM）词性标注、最大熵（Maximum Entropy，ME）等

- 2.**文本语义分析**
  将句子转化为某种可以表达句子意义的形式化表示，即将人类能够理解的自然语言转化为计算机能够理解的形式语言，做到人与机器相互沟通。

- 3.**文本语用分析**
  利用语用学进行文本分析，针对句子群（又称话题，Topic）开展高端分析，获取对文本内涵的掌握。

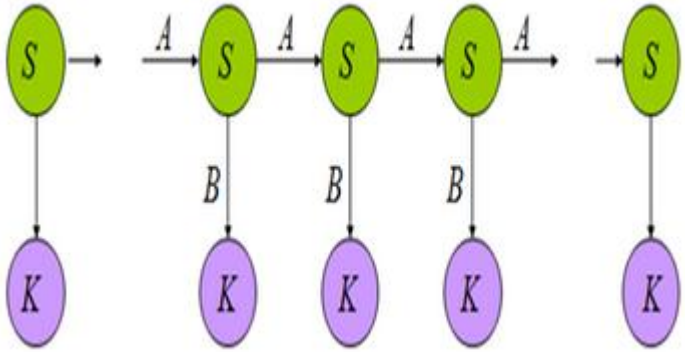
文本语法分析-HMM

对 HMM 来说，有三个重要假设：

假设1：马尔可夫假设（状态构成一阶马尔可夫链）

假设2：不动性假设（状态与具体时间无关）　（状态转移矩阵A）

假设3：输出独立性假设（输出仅与当前状态有关）



绿色：隐藏状态（天气）
紫色：可观察状态（海藻潮湿度）
A：隐藏状态转移矩阵
B：混淆矩阵

文本语法分析-HMM

应用隐马尔可夫模型主要解决三个方面的问题

（一）评估问题：对于给定的模型λ=（A,B,π），和给定的观察序列O=（O1O2O3...OT），如何有效的计算观察值序列O的**概率**P(O|**λ**)。

（二）解码问题：对于给定的模型λ=（A,B,π），和给定的观察序列O=（O1O2O3...OT），如何寻找一个**状态转换序列**Q=（q1q2...qT）,使得该状态转换序列**最有可能**产生上述观察序列。

（三）学习问题或训练问题：在模型参数未知或不准确的情况下，如何根据观察序列O=（O1O2O3...OT）求得模型参数或调整模型参数，即如何**确定一组模型参数**，使得P（O|λ）最大。

文本语法分析-ME模型

最大熵模型是通过求解一个有条件约束的最优化问题来得到概率分布表达式。用条件熵作为衡量一致的标准,如何求最优值p（y|x）问题。

$$H(p) \equiv -\sum_{x,y} \tilde{p}(x)\, p(y \mid x) \log p(y \mid x)$$

上式H(P)满足以下三个限制条件:

$$(1)\ p(y|x) \geq 0 \ for\ all\ x,\ y$$

$$(2)\ \sum_{y} p(y|x) = 1 \ for\ all\ x$$

$$(3)\ \sum_{x,y} \tilde{p}(x) p(y|x) f(x,y) = \sum_{x,y} \tilde{p}(x,y) f(x,y) \ for\ i \in \{1,2,\ldots,n\}$$

文本语法分析

语义分析方法包括词义消歧、信息抽取和感情倾向性分析内容。

- 1.**词义消歧**
  对多义词根据上下文给出它所对应的语义编码，该编码可以是词典释义文本中该词所对应的某个义项号，也可以是义类词典中相应的义类编码

- 2. **信息抽取**
  其研究内容包括实体识别、术语自动识别和关系抽取。命名实体识别包括中国姓名、中国地名、组织机构、英译名的自动辨识。

- 3. **情感倾向性分析**
  对一篇文章进行情感色彩判断。具体来说，就是对说话人的态度（或称观点、情感）进行分析，即对文本中的主观性信息进行分析。

文本语法分析-词义消歧

基于词条语法属性的词义消歧的基本思路如下

算法 WSD：词义消歧算法

输入：待消歧的词条

输出：消歧后的词条

①依据《现代汉语语法信息词典》，对每一个多义词 W，比较不同同形的属性特征进而找出相互之间的肯定性区别特征，对每一个同形 S_i，以 $f_k = v_{ki}$ 的形式列出其肯定区别特征，对每一个多义词 W 生成一个属性特征文件 W_Lex_Rule（如上文“保管.txt”）；

②定位目标多义词 W，以句子范围作为上下文语境 C；

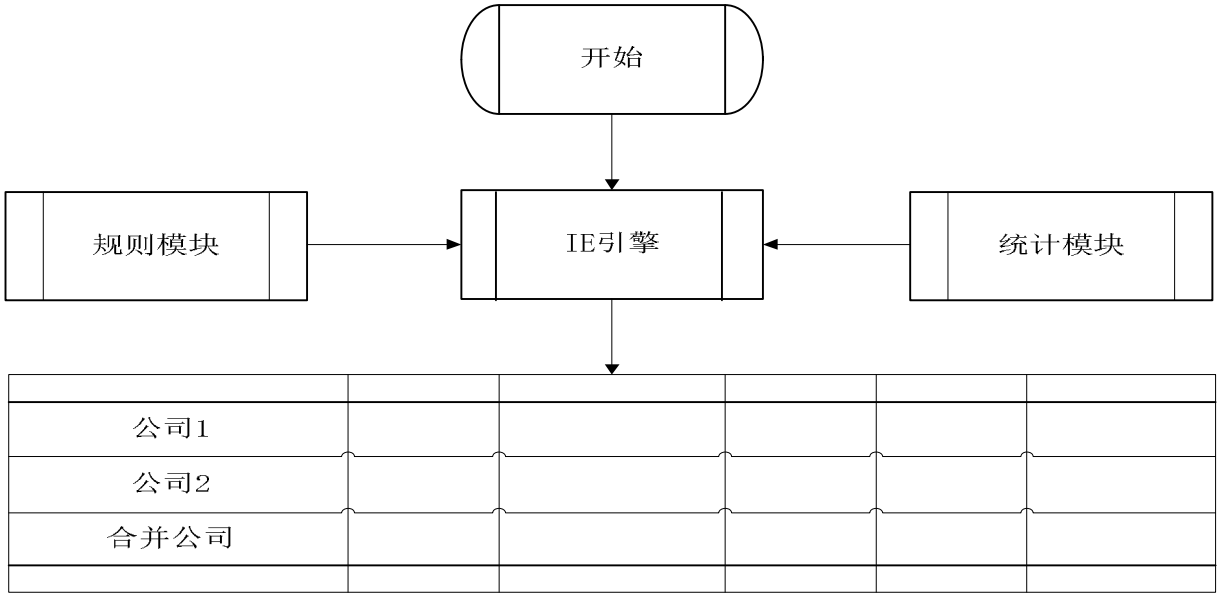
③对 W 的不同同形赋值 $S_i.Score = 0$；

④检索文件 W_Lex_Rule，提取同形 S_i 的肯定性区别特征，判断 W 所在的上下文 C 是否满足约束条件，若满足，则 $S_i.Score = S_i.Score + 1$；

⑤若文件 W_Lex_Rule 中属性特征列表非空，重复④；

⑥Score 取最大的同形 S_i 为标注结果。

文本语法分析-信息抽取



其中，信息抽取引擎的输入是一组文本，引擎通过使用一个统计模块、一个规则模块或者两个的混合进行信息抽取。IE引擎的输出是一组从文本中抽取的标注过的框架，即填好了的一张表。

文本语法分析-情感倾向性分析

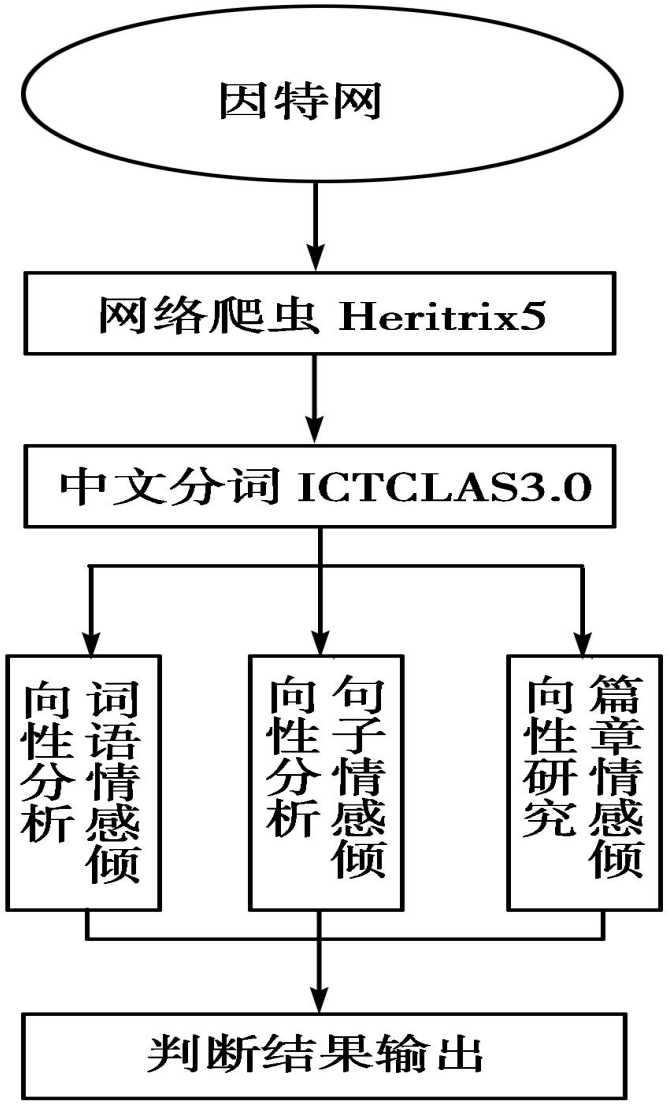
网络舆情倾向性分析模块分解为词语情感倾向性分析、句子情感倾向性分析和篇章情感倾向性研究三个子模块

- 1.**词语情感倾向性分析子模块**
  词语情感倾向性研究是倾向性研究工作的前提。具有情感倾向的词语以名词、动词、形容词和副词为主，也包括人名、机构名、产品名、事件名等命名实体

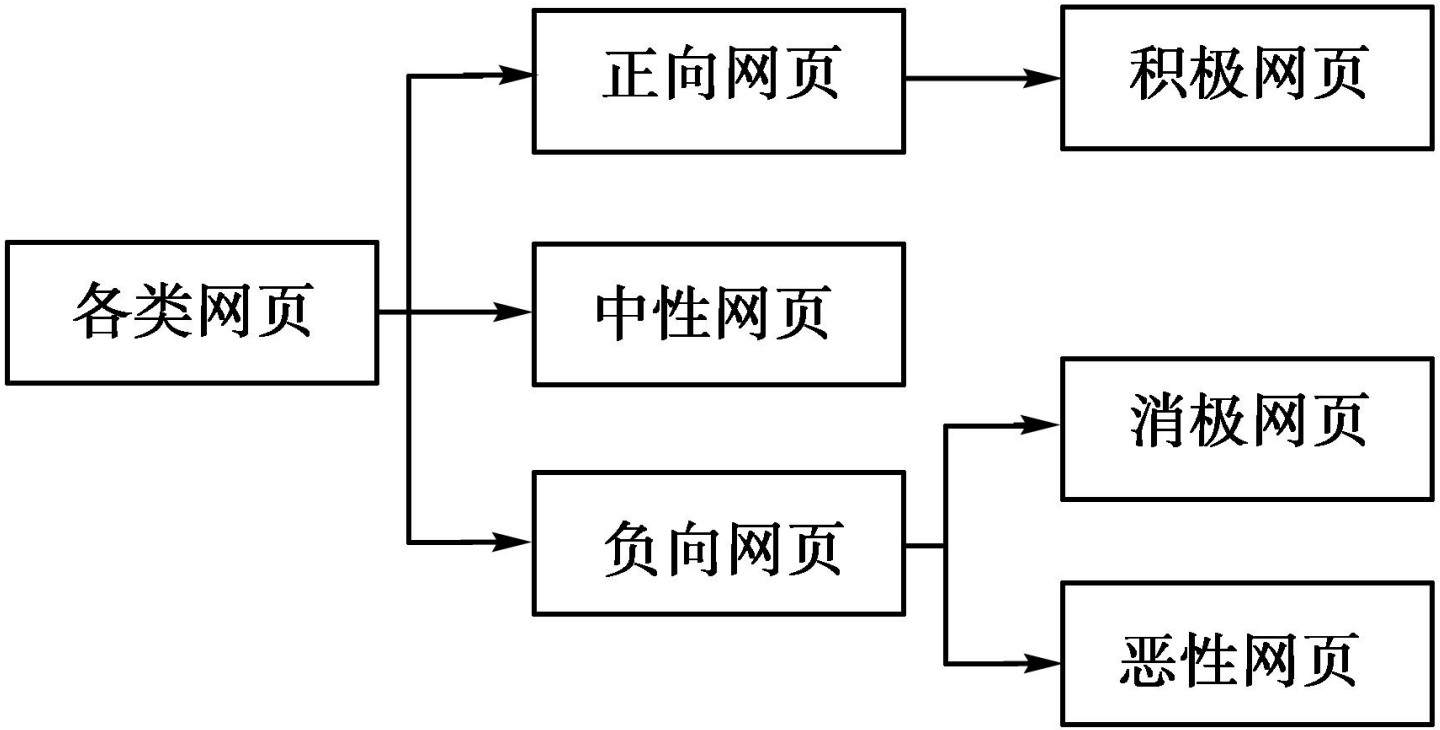
- 2.**句子情感倾向性分析子模块**
  其任务是对句子中的各种主观性信息进行分析和提取，包括对句子情感倾向性的判断，以及从中提取出与情感倾向性论述相关联的各个要素。

- 3.**篇章情感倾向性研究子模块**
  设定一定的阈值，并对含有情感的句子值综合相加，得出篇章的情感色彩，完成文本倾向性分析。根据得出的网页文本情感值与设定的阀值相比较的结果，将网页分为四级：恶性网页、消极网页、中性网页和积极网页。

网络舆情倾向性分析模块结构：



网页情感倾向性分类



3.2.2 文本语用分析方法

文本语用分析

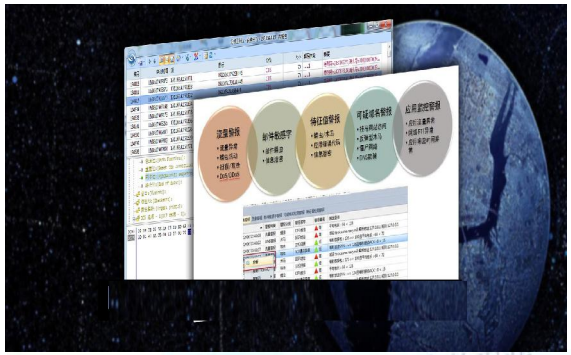
语用学是一门研究如何用语言来达成一定目的的学科，即，利用语用学进行文本分析，针对句子群开展高端分析，获取对文本内涵的掌握。

话题检测与跟踪的意义

问题的产生：传统的关键词检索技术，信息冗余度高，对信息简单罗列，难以全面把握

问题的解决方案：话题检测与跟踪技术能把分散的信息有效地汇集并组织起来，从整体上了解一个话题的全部细节以及该话题中事件之间的相关性。

政府安全分析人员

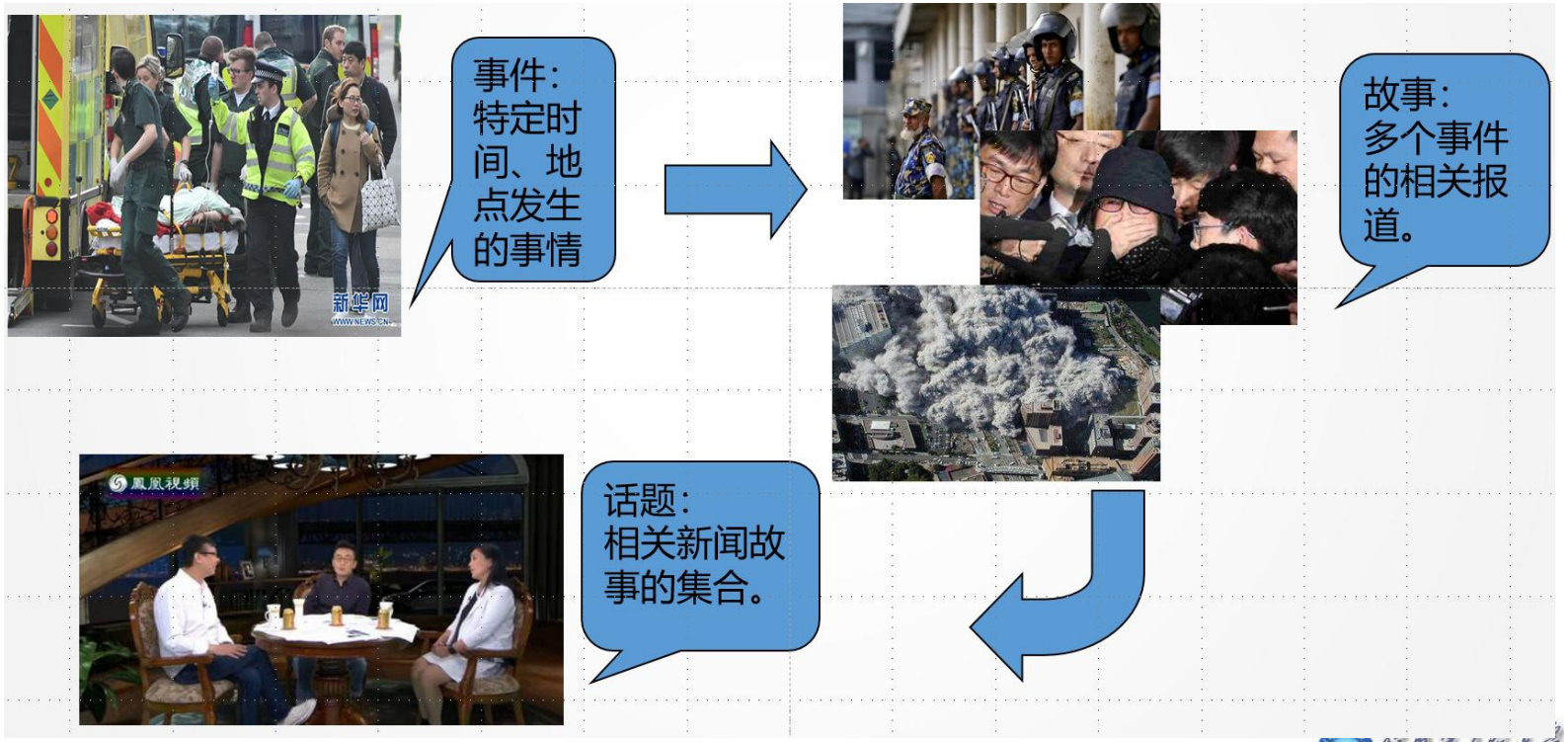


情报分析人员

社会学研究者

话题检测与跟踪方法

在信息内容安全领域，为了检测出蓄意制造混乱的报道集，需要采用话题检测与跟踪技术对其内容进行分析，将具有混乱性质的报道聚集形成话题，分析其动向，一段时间内一旦某个话题的发展超过预期数目（阈值门限），就通知有关人员采取行动加以约束。



话题检测与跟踪任务

研究任务	英文简称	含　义
报道切分	SST (Story Segmentation Task)	将原始数据流切分成具有完整结构和统一主题的报道。
首次报道检测	FSD (First-Story Detection Task)	从具有时间顺序的新闻报道流中自动检测出未知话题出现的第一篇报道。
关联检测	UDT (Link Detection Task)	对给定的两篇新闻报道做出判断，即是否讨论同一个话题。
话题检测	TD (Topic Detection Task)	检测和组织系统预先未知的话题。
话题跟踪	TT (Topic Tracking)	监测新闻信息流，找到与某已知话题有关的后续报道。

话题检测与跟踪任务

任　务	切分方式	工作方式	特　点
报道切分	基于语音识别系统的报道切分	根据语音信号的分布规律划分报道边界	可以相对准确地识别边界，但是边界之间包含的信息却不一定准确地指向一个报道，往往其中包含多个报道。
	基于内容的报道边界识别	转录为文本形式，根据报道之间主题内容的差异估计报道边界。	可以根据话题的内涵识别出不同报道，但报道与报道之间边界的划分相对模糊。

任　务	子任务	工作方式
话题检测	首次报道检测（FSD）	准确定位新话题出现的最初报道。
	在线话题检测（OTD）	不仅要求系统识别最新话题，也要收集该话题的所有相关报道。

话题检测与跟踪方法

主要特点

- 多数采用传统的文本分类、信息过滤和检索的方法，专门针对话题发现与跟踪自身特点的算法还未形成；
- 对某个用户感兴趣的特定话题，现有系统都无法保证取得满意的效果；
- 综合使用多种相对成熟的方法，在实际应用中可能效果最佳。

3.3 文本内容安全应用

基于内容的网页过滤

基于内容的网页过滤技术是一个双重概念：既要能够过滤从因特网进入终端的内容，也要能够过滤从终端出去的内容。

基于内容的网页过滤技术应用：

- 过滤用户互联网请求从而阻止用户浏览不适当的内容或站点；
- 过滤从因特网“进来”的其他内容从而阻止潜在的攻击进入用户的网络系统；
- 为了保护个人、公司、组织内部的数据安全，避免敏感数据通过互联网暴露给外界而实施的堵塞过滤。

基于内容的网络监控

网络信息内容监控技术能够增强互联网运用和驾驭能力，为传播社会主义先进文化提供新的空间，为维护国家文化安全和意识形态安全提供重要保障。

网络内容监控目前主要是对文本类型的网络信息进行搜索过滤，主要涉及两类关键技术：文本挖掘和模式匹配。前者用于将新出现的具有相同特征的文本信息挖掘出来；后者根据已知的特征码对文本信息进行分析，以便实施拦截。

基于内容的网络监控

网络内容监控产品：

第一，明确内容监控的目的，从大的方面讲是维护国家利益，从小的方面讲是维护企业和个人的利益。

第二，内容监控对我们来说越来越重要，很多单位都有安全保密条例，所有的内容监控都应在统一到单位的保密策略下，密切围绕单位的信息安全条例来展开。