领域术语抽取

刘秉权
智能技术与自然语言处理实验室
哈尔滨工业大学
liubq@hit.edu.cn

教学目的

- 领域术语是一种特殊的文本实体，领域术语抽取有较大难度和重要应用价值
- 体会领域术语抽取与普通生词和实体识别在方法上的区别
- 学习掌握几种方法
  - 面向领域术语抽取的文本分类（领域文本自动判别）
  - 基于局部最大算法的中文新词发现
  - 基于正规化分布熵的领域术语抽取
- 熟悉领域术语在文本分类和问答式信息检索领域的应用

主要内容

- 1. 问题描述
- 2. 领域文本自动判别
- 3. 领域术语自动抽取
- 4. 领域术语抽取的应用
- 5. 小结

1. 问题描述

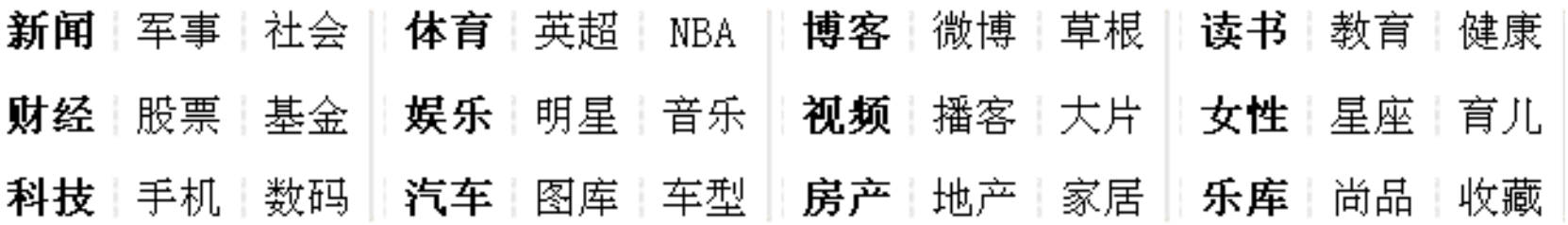
- 1.1 问题的提出
- 1.2 领域的定义
- 1.3 领域术语的定义
- 1.4 领域术语自动抽取的目的和意义
- 1.5 领域术语抽取的评价

1.1 问题的提出

- 术语作为特定专业领域中的一般概念词语，有着很强的专业性
- 它传递了专业文献尤其是技术文献中复杂领域的知识
- 在一定程度上术语的变化反映了一个学科领域的发展变化

1.2 领域的定义

- 领域（Domain）是指人类知识的一个学科类别或者一个专业范围。
- 针对不同的应用人们可以定义不同的领域分类体系。
- 中国图书馆分类法：一个领域分类体系，它包含A类（马列主义、毛泽东思想）、B类（哲学）、等38个领域类别。
- 新浪门户网站的导航栏分类体系：



1.3 领域术语的定义

- 领域术语（Domain-specific term）
  - 是指在特定领域中使用、表示该领域的概念、特征的词语。领域术语可以是词，也可以是短语。
  - 领域术语可以只属于某一个领域，也可以并存于多个领域中。
  - 领域术语也可被简称为术语。
- 领域术语自动抽取示例：

C1 马列主义	C2 法律	C3 军事	C4 体育	C5 医药卫生
无产阶级	司法	作战	比赛	患者
社会主义	人民法院	军种	首场	治疗
全党	最高人民法院	军事	球员	血管
马克思主义	案件	军队	球队	临床
共产主义	审理	战争	英格兰队	疗效
马克思列宁主义	诉讼	兵力	队友	药物
马克思	法院	事变	世界杯	病人

1.4 领域术语自动抽取的目的和意义

- 目的：
  - 为面向领域的应用提供可定制的领域文本自动判别方法和领域术语自动抽取方法。
- 意义：
  - 对所有面向领域的应用都有重要意义
    - 垂直搜索
    - 文本自动分类
    - 语言建模
    - 词义消歧

1.5 领域术语自动抽取的评价

- 1) 概述
- 2) 人工评价
- 3) 应用系统评价
- 4) 自动评价

1) 概述

- 领域术语自动抽取的评测评价比较困难，目前国际上还没有一个公开的标准数据集用于评测。
- 在不同的分类体系下或是面向不同的应用，领域术语的定义会有很大区别。

2) 人工评价

- 语言学家或者研究者通过个人判断来确定抽取的词语是否为领域术语。
- 优点：可以对术语抽取结果有一个直观感受。
- 缺点：成本大，具有主观性和不一致性。

3) 应用系统评价

- 在应用系统中评测，考察领域术语抽取模块在信息检索、文本分类等应用系统中的表现，即观测在一个应用系统中使用领域术语抽取模块前后系统性能的变化。
- 优点：不需要人们的手工标注，评测迅速，且完全面向应用。
- 缺点：不同的应用系统可能会产生不尽相同的结果。

4) 自动评价

- 正确率： $\mathrm{P} = \frac{\text{正确术语数目}}{\text{抽取的总术语数目}} \times 100\%$
- 召回率： $\mathrm{R} = \frac{\text{正确术语数目}}{\text{术语全集的术语总数}} \times 100\%$
- F-度量： $F1 = \frac{2PR}{P+R}$

2. 领域文本自动判别

- 2.1 研究现状
- 2.2 有监督文本分类方法

2.1 研究现状

- 1) 概述
- 2) 有监督文本分类方法
- 3) 半监督文本分类方法
- 4) 基于正例的文本分类方法

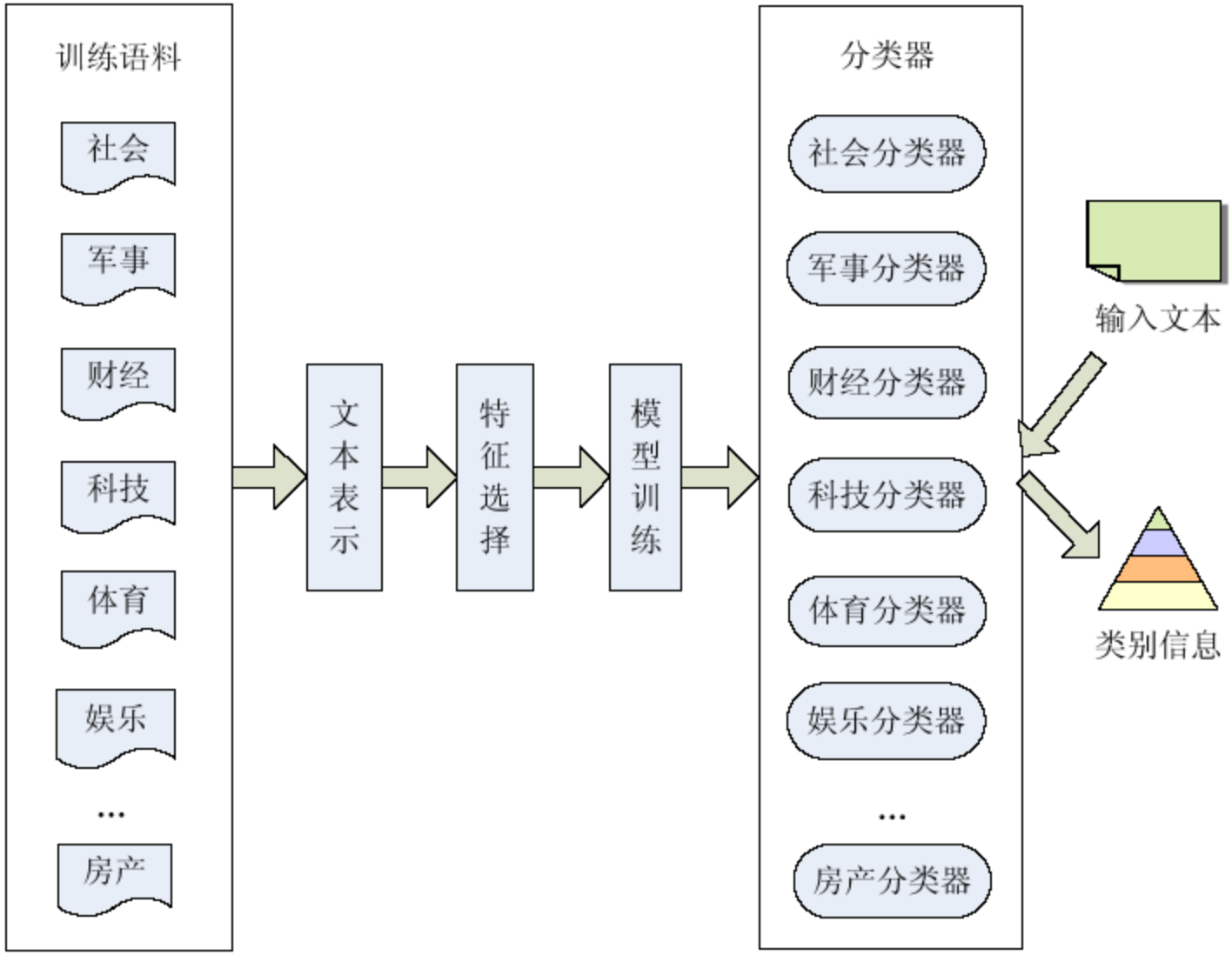
1) 概述

- 领域文本自动判别需要用到文本分类的基本技术，同时根据自身应用特点，采用一些特殊的处理技术。
- 根据不同的前提条件与应用场景，领域文本自动判别需要采取不同技术。
- 文本自动分类（Text Automatic Classification）:基本任务是对一篇文档，根据其内容，从预先定义好的类别标记集中找出一个或者多个最适合于该文档的标记。

2) 有监督文本分类方法

- 当事先给定一个包含多个领域类别的相对完整的领域分类体系，并给定每个类别下的一定规模的已标注语料作为训练文本时，有监督文本分类技术可以直接被应用于多个类别的领域文本获取。
- 有监督文本分类方法
  - Rocchio法、贝页斯分类法、K最近邻分类法、支持向量机、决策树、神经网络

2) 有监督文本分类方法


训练语料
社会
军事
财经
科技
体育
娱乐
...
房产
文本表示
特征选择
模型训练
分类器
社会分类器
军事分类器
财经分类器
科技分类器
体育分类器
娱乐分类器
...
房产分类器
输入文本
类别信息

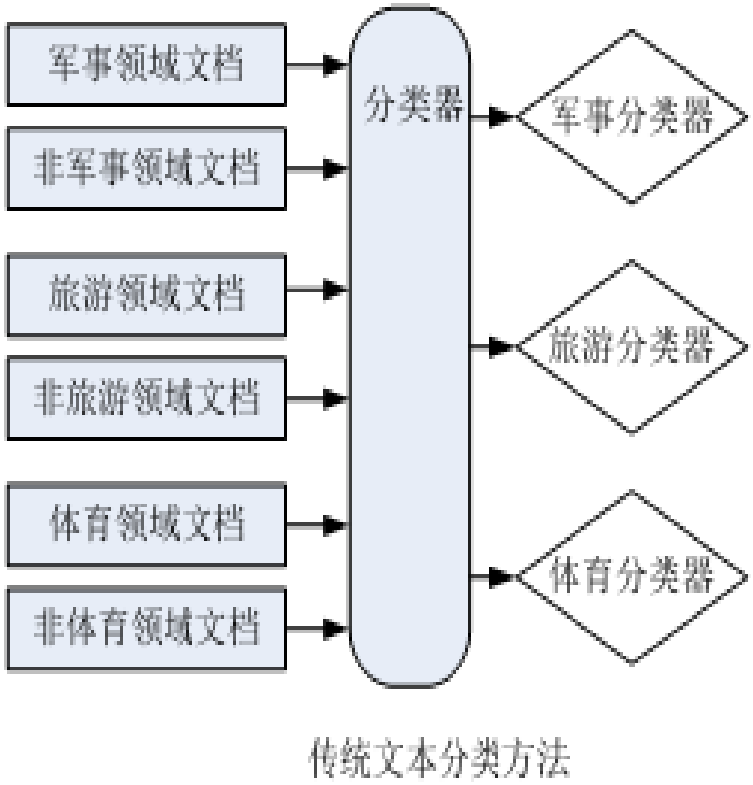
3) 半监督文本分类方法

- 标注样本的获取比较困难，通常需要人工标注，费时费力。半监督文本分类方法使用大量的未标注数据和少量已标注数据来构建分类器。
- 半监督文本分类方法
  - 产生式模型和期望最大化、直推式支持向量机、自助学习法、协同训练法、主动学习法

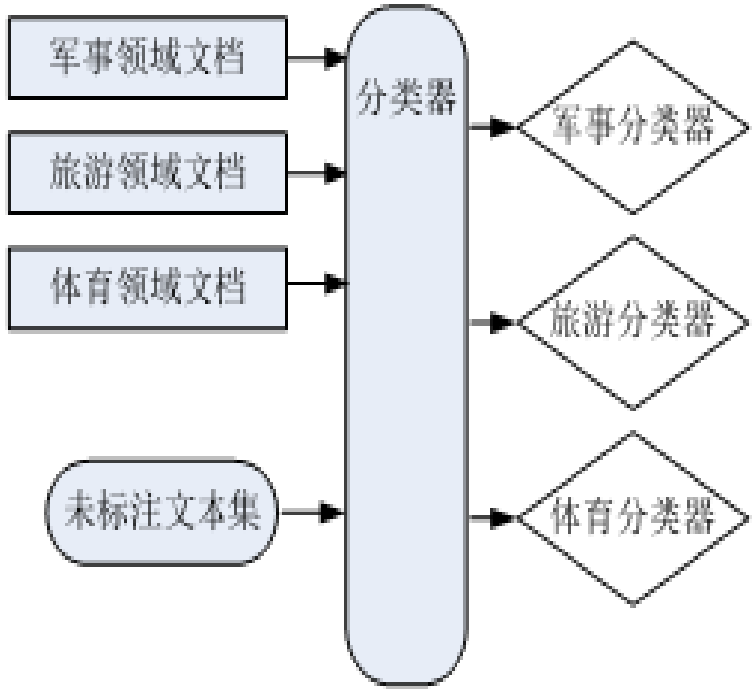
4) 基于正例的文本分类方法

- 当待获取文本的领域类别不能构成一个完整的分类体系时，比如只获取某一特定领域类别的文本，只给定少量已标注好的该领域文本作为正例，并没有给定标注好的不属于该领域的文本作为分类器的反例。要使用基于正例的文档类别判别方法。
- 基于正例的文本分类方法
  - PEBL、 Spy_EM、Roc_SVM、类别约束SVM（Biased_SVM）、 PNLH方法

4) 基于正例的文本分类方法



传统文本分类方法



基于正例的分类方法

2.2 有监督文本自动分类方法

- 1) 文本自动分类问题描述
- 2) 文本自动分类过程
- 3) 文本表示
- 4) 特征选择
- 5) 领域术语作为文本分类的特征
- 6) 分类算法

1) 文本自动分类问题描述

问题描述：给一个决策矩阵中的每个元素赋值，并且在{0,1} 的实数范围内取值，如下表所示：

	d_1	…	…	d_j	…	…	d_n
c_1	a_{11}	…	…	a_{1j}	…	…	a_{1n}
…	…	…	…	…	…	…	
c_i	a_{i1}	…	…	a_{ij}	…	…	a_{in}
…	…	…	…	…	…	…	…
c_m	a_{m1}	…	…	a_{mj}	…	…	a_{mn}

其中，$C=\{c_1,\cdots,c_m\}$ 是预先定义好的类别集合或者标记集合，$D=\{d_1,\cdots,d_n\}$ 是待分类文档集合，$a_{i,j}$ 是一个文档和类别标记的隶属度，$a_{i,j}\in[0,1]$ 。

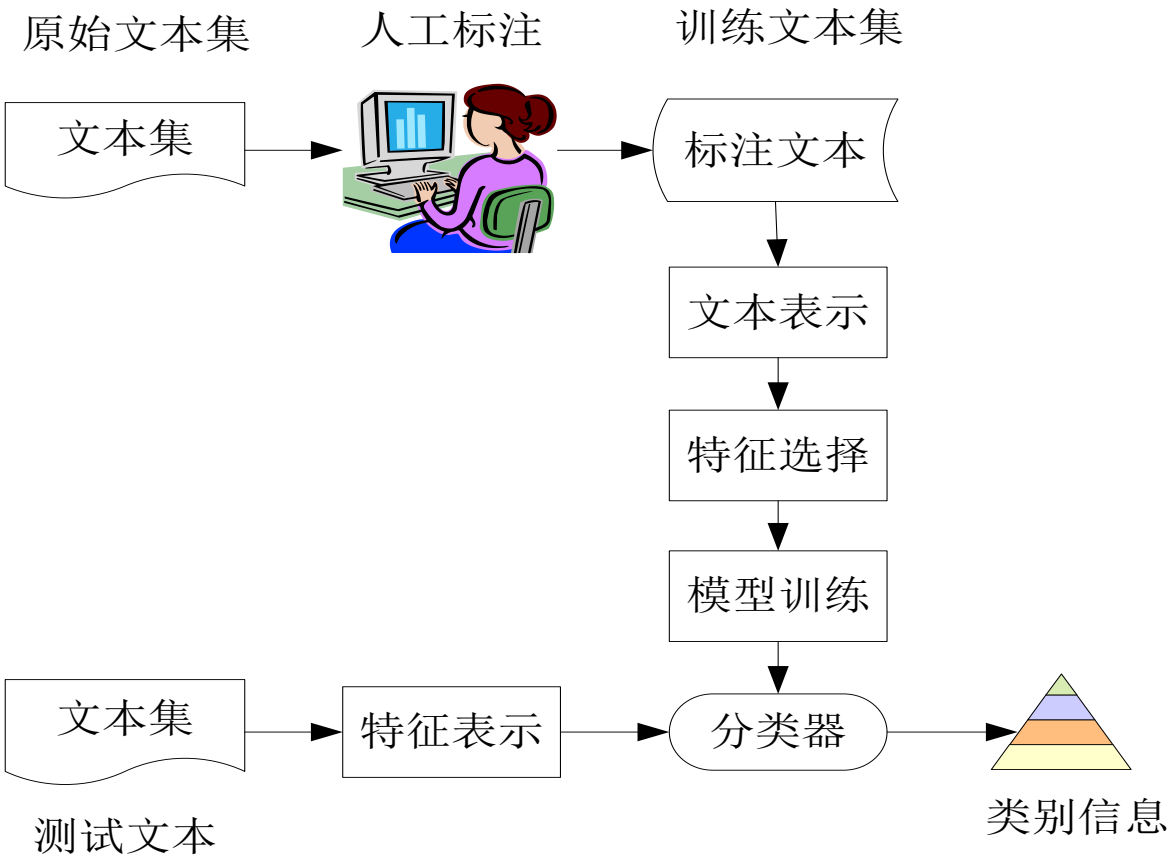
说明

类别标记是一组符号。例如，大多数搜索引擎使用的类别体系，包括科学技术、社会文化、政治军事、医疗健康、体育健身等；

文档对类别的隶属度应该是基于文档的内容，而不是基于描述文档的元数据（Metadata）（例如文档出版日期、文档类型等）；

$a_{i,j}$ 为一条件概率 $a_{i,j} = p(c_i \mid d_j)$，如果 $a_{i,j} = 1$，表示第 j 个文档完全属于第 i 个类别（或者说完全相关）；$a_{i,j} = 0$，表示文档 j 和类别 i 完全无关。

2) 文本自动分类过程


原始文本集
人工标注
训练文本集
文本集
标注文本
文本表示
特征选择
模型训练
文本集
特征表示
分类器
类别信息
测试文本

3) 文本表示

- 要判断一个文档的语义类别， 首先要将文档表示成一种通用计算形式。
- 向量空间模型（Vector Space Model，VSM）：
  - 将文档看作为是由相互独立的词条构成，采用词袋（Bag of Word，BOW）文档表示方法，$d=(t_1, t_2, \ldots, t_k)$。
  - 文档d映射为一个特征向量$V(d)=(w_1, w_2, \ldots, w_k)$，其中$w_i$为索引词语$t_i$在文档d中的权重。
  - 包含n个文档具有k个索引词语的文档集被表示成一个矩阵$A=[a_{ij}]$，其中a_{ij}表示词语t_i在文档d_j中的 权重，第j列向量表示文档d_j。
  - 当权重仅用出现与否标志0和1来表示时，向量空间模型转换为布尔模型。

4）特征选择

- 如果使用所有的在训练文档集合中出现的词语作为二维空间的横坐标，那么表示一个文档的向量的维数会过大。通常都要对向量空间进行约简。
- 不是所有的词都能表达领域特性，需要选择与领域关系密切的词作为特征。

主要特征选择方法

1. 词频/倒排文档频度（Term Frequency Inverse Document Frequency，TFIDF）

2. 基于信息增益度（Information Gain）的特征选择

3. 基于 χ^2 分布的特征选择

4. 基于矩阵分解的特征选择

TFIDF

$$tfidf_d(w) = tf \cdot idf = \frac{f(w)}{\sum f(w)} \cdot \log\left(\frac{|D|}{|\sum d|} + 0.01\right)$$

其中，$f(w)$为词w在文档d中出现的次数，$\sum f(w)$为一篇文档总词数，$|\sum d|$为语料库中的文件总数，$|D|$ 为包含词语w的文件数目。

5) 领域术语作为文本分类的特征

- 领域术语抽取方法可以用作文本分类的特征选择方法。
- *陈文亮*基于词语共现和自助法从未标注大规模语料中发现领域术语，然后将领域术语作为特征进行文本分类，与使用小规模训练集的简单贝叶斯分类器相比，获得了更好的效果。

6) 分类算法：Rocchio算法

- 文档分类的经典方法。
- 基本思想：
  - 为每一个类别c_i建立原型向量
  - 根据文档向量和类别原型向量的距离，确定文档的类别
  - 类别i的原型向量是通过计算属于该类别的所有文档向量的平均值而得到的
- 特点：速度快，但是精度较低。

6) 分类算法：Naïve Bayes算法

通过对训练数据的学习，得到在一个文档出现的条件下类别i出现的条件概率；我们用Bayes方法来估计这一概率: $P(c_i \mid d) = \frac{P(c_i)P(d \mid c_i)}{p(d)}$，简化为： $P(c_i \mid d) = P(c_i)P(d \mid c_i)$

1. 估计 $P(c_i)$。 $\hat{P}(c_i)$ 为 $P(c_i)$ 的估计： $\hat{P}(c_i) = \frac{N_i}{N}$

2. 估计 $P(d \mid c_i)$。

假设：文档中词语的出现是相互独立的，所以： $P(d \mid c_i) = \prod_{j=1}^{K} P(t_j \mid c_i)$

$\hat{P}(t_j \mid c_i)$ 为 $P(t_j \mid c_i)$ 的估计： $\hat{P}(t_j \mid c_i) = \frac{1 + N_{j,i}}{M + \sum_{m=1}^{M} N_{j,m}}$

其中， $N_{j,i}$ 是训练集中词语（特征）j在类别 c_i 中出现的次数。

这种方法是一种基于最小错误率的贝叶斯决策理论的分类方法。

3. 领域术语自动抽取

- 3.1 领域术语的特点
- 3.2 研究现状
- 3.3 基于局部最大算法的中文新词发现
- 3.4 基于正规化分布熵的领域术语抽取

3.1 领域术语的特点

- 领域术语一般只在一个或几个特定的领域流通，只有该特定领域的人使用。
- 如果用流通度表示一个语言单位流行通用的程度，则领域术语仅在本领域具有高流通度。
- 领域术语是各个专门领域独用的词语，因此各个领域具有互不相同的领域术语数目。

3.2 研究现状

- 1) 基于语言学知识的方法
- 2) 基于统计量度的方法
- 3) 规则和统计相结合方法

1) 基于语言学知识的方法

- 主要思想：

  领域术语经常以特定的语言结构和模式出现，通过发现并抽取一些符合术语模式的字串来实现术语的自动抽取。

- 比如：

  利用拼写和词汇线索为特定类别的术语构建可以体现术语命名结构的规则。

1) 基于语言学知识的方法

- 这种方法在术语消歧、准确率上有非常明显的优点。
- 但对于不同的语言、不同的领域要建立不同的术语规则，因此这种方法的可移植性差。
- 并且当已有许多规则时，还需要解决多个规则之间的冲突。

2) 基于统计量度的方法

- 常用统计量度

  词频/倒排文档频（TFIDF）

  文档频/倒排类别频（KFIDF）

  互信息（MI）

  构词力

  构词模式

  对数似然比（log-likelihood）

  左右信息熵（entropy）

  C-Value/NC-Value

2) 基于统计量度的方法

- 各种度量的作用：

（1）MI确定词语之间的搭配关系

（2）Log-likelihood参数能避免一些低频词的遗漏。

（3）C-Value/NC-Value考虑简单术语与复杂术语之间的关系。

（4）……

- 难以筛选合适的全局阈值统一抽取新词
- 对术语抽取标准描述不够细致和全面

3) 规则和统计相结合方法

- 最近的研究大部分结合了统计学和语言学的方法，使用统计学的方法获取候选术语，再利用语言学的特点来筛选、过滤或修正。
- 统计学和语言学相结合的方法是术语自动抽取的主要发展方向。

3) 规则和统计相结合方法

- Thuy VU先根据规则抽取候选术语，然后组合C/NC-value和T_Score计算术语的权值，最后得到真正术语。
- 张峰的基于互信息的中文术语抽取系统首先用互信息得到术语候选集，然后使用规则进行判别,进而得到真正的术语。

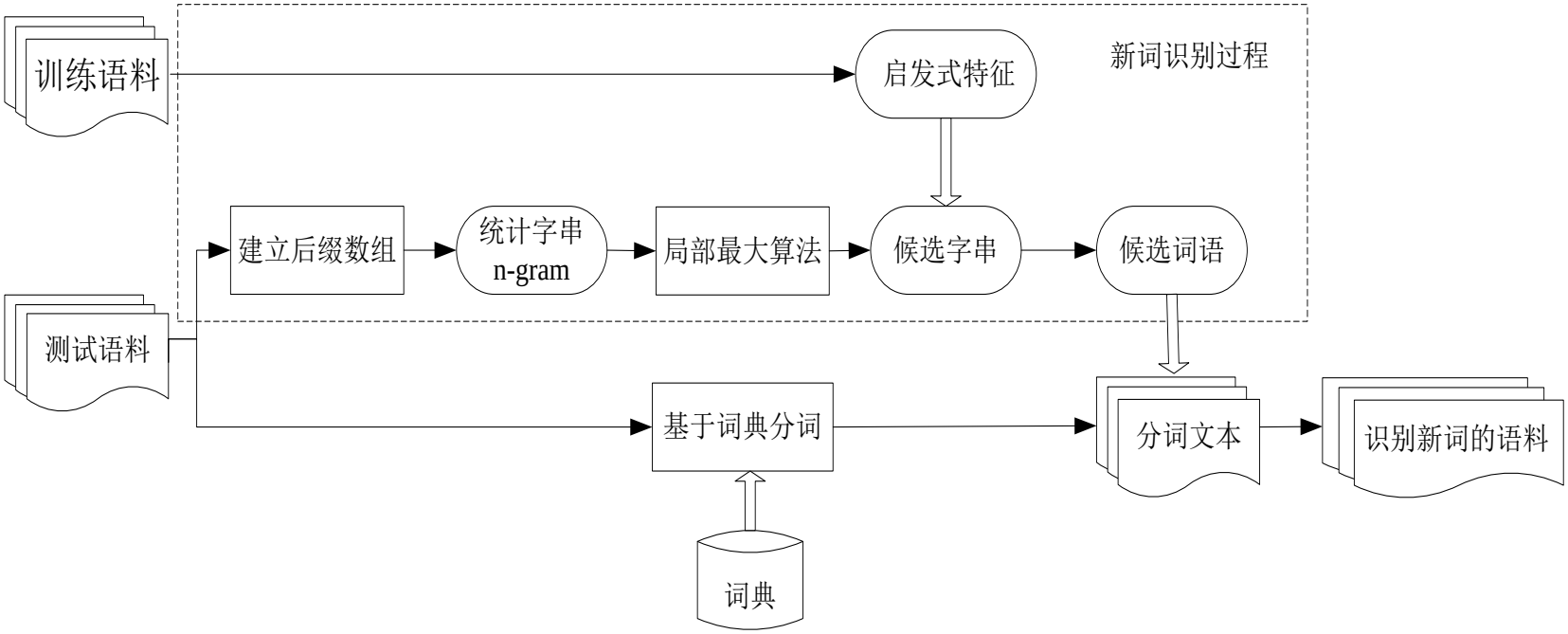
3.3 基于局部最大算法的中文新词发现

- 1) 简介
- 2) 系统框架
- 3) 词语的内聚性和上下文独立性
- 4) 字串关联强度计算
- 5) 局部最大算法的思想描述
- 6) 启发式特征过滤
- 7) 基于后缀数组的汉字ngram频次统计
- 8) 实验结果

1) 简介

- 英文：词序列关联强度越大（极少改变词汇和语法），那么它就越有可能成为多词单元
- 中文：字串关联强度越大，结合越紧密，越有可能成为多字词
- 问题：词语边界难以确定
- 本方法主要思想：取具有局部最大关联强度的字串（非词典词）作为候选新词
- 该方法充分利用词语的内聚性和上下文独立性，可以鲁棒地识别中文新词，克服了传统方法中最优经验域值选择的困难。

2) 系统框架


训练语料
启发式特征
新词识别过程
建立后缀数组
统计字串
n-gram
局部最大算法
候选字串
候选词语
测试语料
基于词典分词
分词文本
识别新词的语料
词典

解释

- 1. 为测试语料建立后缀数组用以统计基于字串ngram，在此基础上用局部最大算法从语料中得到所有可能成词的字串；
- 2. 利用从训练语料中得到的启发式特征，对局部最大算法得到的字串进行过滤；
- 3. 将新词识别模块作为原有基于词典的分词系统的后处理模块，利用候选词语合并原始分词文本中被切成单字序列的散串，进而得到新词识别后的语料。

3) 词语的内聚性和上下文独立性

- 中文词语是汉语中能够独立运用的最小语言单位。
- 内聚性：指组成词语的每个汉字之间具有很强的关联，同时表明该词语的任何子字串均具有上下文依赖性。
- 上下文独立性：指组成词语的每个汉字与其在句子上下文中的其他汉字关联强度较弱。

3) 词语的内聚性和上下文独立性

若s表示由汉字序列...$c_{i-1}c_ic_{i+1}...c_{i+n-1}c_{i+n}$ ...构成的一个句子，且字序列$c_ic_{i+1}...c_{i+n-1}$构成词语W，则

词语W的内聚性表示了如下含义：

(1) 字序列$c_ic_{i+1}...c_{i+n-1}$的关联强度值较大；

(2) 词语包含的左子字串($c_ic_{i+1}...c_{i+n-2}$)和右子字串($c_{i+1}...c_{i+n-2}c_{i+n-1}$均具有上下文依赖性。

词语W的上下文独立性表示了如下含义：

(1) 包含词语W的字串($c_{i-1}c_i\cdots c_{i+n-1}$)的关联强度值较小；

(2) 包含词语W的字串($c_ic_{i+1}\cdots c_{i+n-1}c_{i+n}$)的关联强度值较小。

4) 字串关联强度计算1: Bigram关联强度计算

- 对称条件概率（SCP）

$$SCP([x,y]) = p(x \mid y) \times p(y \mid x) = \frac{p(x,y)^2}{P(x) \times p(y)}$$

其中，p(x, y), p(x), p(y)分别是bigram(x,y)、unigram x、unigram y在语料中出现的概率；p(x|y)是y出现在bigram的第二个位置（右侧）的条件下，x出现在bigram的第一个位置（左侧）的条件概率；类似地，可以定义p(y|x)

4) 字串关联强度计算2: Ngram（n>2）关联强度计算

- 平摊对称条件概率（FSCP）

$$FSCP([c_1 \ldots c_n]) = \frac{p(c_1 \ldots c_n)^2}{Avp}$$

其中：

$$Avp = \frac{1}{n-1} \sum_{i=1}^{i=n-1} p(c_1 \ldots c_i) \times p(c_{i+1} \ldots c_n)$$

5) 局部最大算法的思想描述

- 如果ngram *W*的关联强度大于所有包含该*W*的(n+1)gram的关联强度，并且ngram *W*的关联强度不小于所有*W*包含的(n-1)gram的关联强度，则ngram *W*具有上下文独立性和内聚性，它属于中文词语的概率较大。

实例说明

- 句子“薰衣草可以舒缓镇定”中不同ngram所对应的FSCP值

n-gram	FSCP
薰衣草可	0.0199
薰衣草	0.3636
薰衣	0.0615
衣草	0.0513

薰衣草

n-gram	FSCP
草可以	0.0067
可以	0.1042
可以舒	0.0090

可以

n-gram	FSCP
以舒缓	0.0040
舒缓镇	0.0952
舒缓	0.1777

舒缓

n-gram	FSCP
缓镇定	0.0303
镇定	0.0333

镇定

6) 启发式特征过滤1-构词力

- 构词力：描述一个字形成词语的概率

$$WFP(c) = \frac{Count(MCW\ Contains\ c)}{Count(c)}$$

分母是汉字c在训练语料中出现的次数，分子是包含汉字c的多字词出现的次数。

- 字序列C= *c1c2...cn*构成多字词W的概率

$$P_{WFP}(W) = \prod_{i=1}^{n} WFP(c_i)$$

6) 启发式特征过滤2-构词模式

- 构词模式：对于某个单字c，它在构成多字词时有三种模式，这三种模式对应了它在多字词中的三种位置：词首、词中和词尾
- 单字c在某个位置模式的概率

$$P(pos(c)\,|\,c) = \frac{Count(pos(c))}{Count(c\ is\ in\ MCW)}$$

- 字序列*C= c1c2...cn*构成多字词W的构词模式概率：

$$P_{pttn}(W) = \prod_{i=1}^{n} P(pos(c_i)\,|\,c_i)$$

6) 启发式特征过滤3

- 采用如下公式对局部最大算法抽取的候选字串赋予权重：

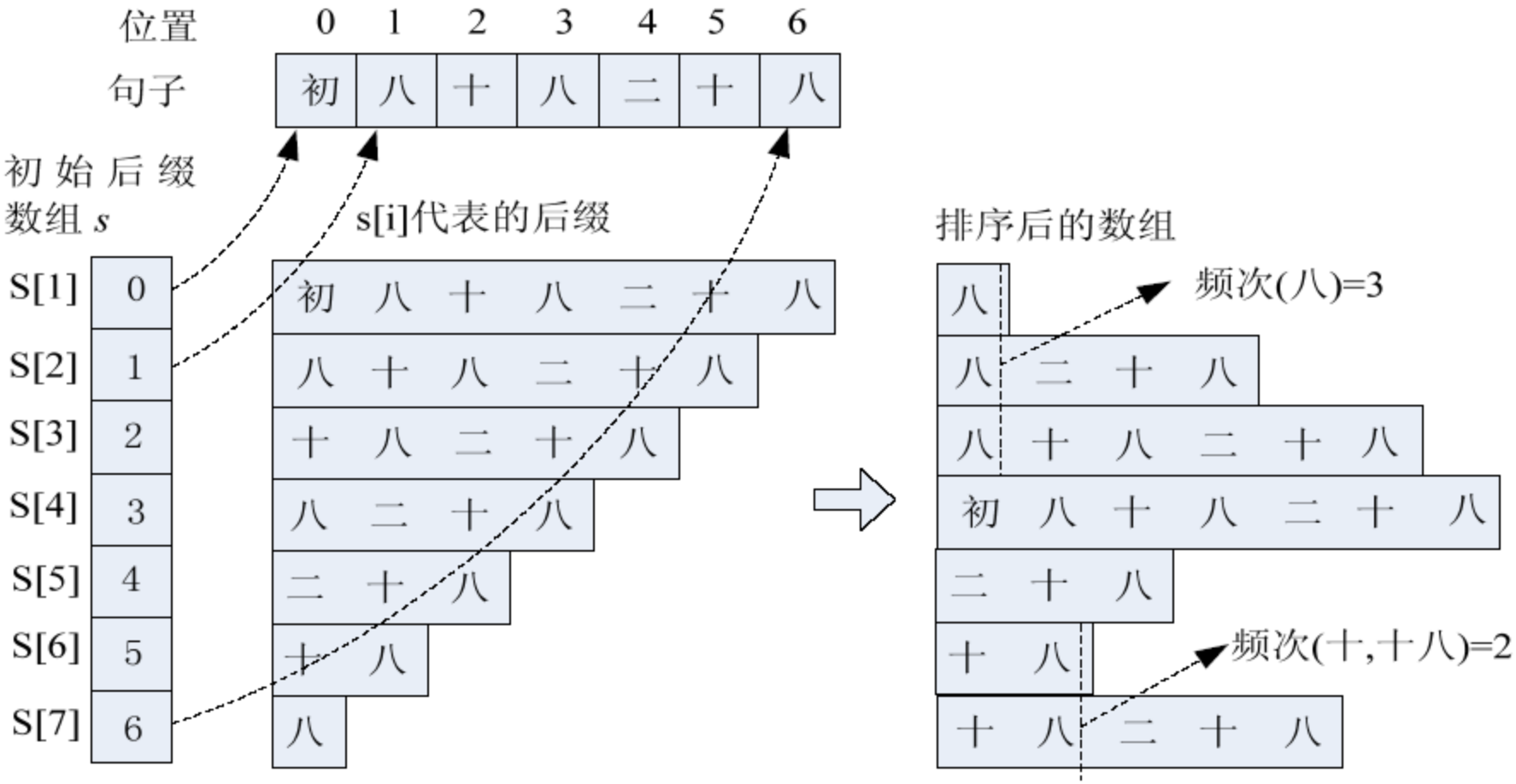
$$weight = \log(P_{WFP}(W) \times P_{pttn}(W))$$

- 通过分析实验结果，发现局部最大的关联强度与局部最大的权重相比，能够更加准确地反映测试语料中有关词语构成的统计信息，因此将启发式特征过滤作为局部最大算法的后处理模块。

7) 基于后缀数组的汉字ngram频次统计

- 基于后缀数组的实现方案
  - 传统方法在计算n>3的ngram时，速度非常慢，不能满足互联网环境下大规模网络文本的实时新词发现需求。
  - 采用基于后缀数组的数据结构。
  - 采用基于等价类划分的思想降低存储空间
    - 只需要存储每个等价类中的一个最长序列

后缀数组的存储和排序


位置
0 1 2 3 4 5 6
句子
初 八 十 八 二 十 八
初始后缀数组 s
s[i]代表的后缀
S[1] 0
S[2] 1
S[3] 2
S[4] 3
S[5] 4
S[6] 5
S[7] 6
初 八 十 八 二 十 八
八 十 八 二 十 八
十 八 二 十 八
八 二 十 八
二 十 八
十 八
八
排序后的数组
频次(八)=3
频次(十,十八)=2
八
八 二 十 八
八 十 八 二 十 八
初 八 十 八 二 十 八
二 十 八
十 八
十 八 二 十 八

8) 实验结果1

在MSR语料上的新词发现前后对比实验结果

MSR	OOV 召回率	精确率	召回率	IV 召回率	F1量度
Baseline	0.361	0.936	0.971	0.988	0.953
Baseline+NE	0.497	0.943	0.968	0.981	0.955
Baseline+NWI	0.581	0.951	0.967	0.977	0.959
Baseline+NWI+NE	0.629	0.957	0.97	0.98	0.963

在PKU语料上的新词发现前后对比实验结果

PKU	OOV 召回率	精确率	召回率	IV 召回率	F1量度
Baseline	0.234	0.891	0.937	0.98	0.913
Baseline+NE	0.42	0.914	0.943	0.975	0.928
Baseline+NWI	0.534	0.931	0.949	0.975	0.94
Baseline+NWI+NE	0.63	0.944	0.958	0.978	0.951

8) 实验结果2: 被错误抽取的字串举例

被错误抽取的字串	原始句子或句子片段	在 MSR 语料中的切分
唱响	唱响国企志气歌	唱/响
憋足	职工憋足了劲	憋/足
铸宝刀	锻得新钢铸宝刀	铸/宝刀
堡岛	一向喧闹的中堡岛今夜格外宁静	中堡岛
毛乌素	穿过茫茫的毛乌素大沙漠东南边缘	毛乌素大沙漠

- “中堡岛”是一个岛的名字，被局部最大算法错误地分割成“中/堡岛”，这是由于“中”字经常单独出现。
- 局部最大算法识别较长名实体的能力（如“毛乌素大沙漠”）较弱，这是因为名实体经常包含一个或多个子词语（如“沙漠”），因此更加固定的“毛乌素”被局部最大算法抽取出来了。

8) 实验结果3: 有实际意义的字串举例

有意义的字串	原始句子或句子片段	在 MSR 语料中的切分
酬赏	即守信者得到酬赏，失信者受到惩罚	酬/赏
崇仁麻鸡	崇仁麻鸡和东乡黑鸡为主的饲养量	崇仁/麻/鸡
茶鲜菇	以茶鲜菇为主的食用菌加工企业	茶/鲜/菇
巨型机	新一代超高性能巨型机	巨型/机
贫油史	以改写中国贫油史而成为举世闻名的“功勋城”	贫油/史
早蕾	摘除７月１０日前的早蕾	早/蕾

3.4 基于正规化分布熵的领域术语抽取

- 1) 抽取原则
- 2) 类间分布熵
- 3) 类内分布熵
- 4) 词语的排序公式
- 5) 抽取算法
- 6) 实验结果

1) 抽取原则

- 传统方法:KFIDF,DR_DC
  - 对术语抽取标准描述不够细致和全面
- 基于正规化分布熵的领域术语抽取方法
- **抽取原则一**：领域术语应该在不同领域类别间分布不均匀
  - 频繁出现在某领域文档中，很少出现在其它领域文档中的词语是领域术语的可能性较大。
  - 出现该词语的领域类别数越少，该词语越有可能是领域术语。
- **抽取原则二**：领域术语在其相关领域的文档集中应尽可能分布均匀
  - 出现在某领域的大多数文档中的词语可能是该领域的术语。
- **正规化策略**：减轻不同语料规模和文档长度的影响

2) 类间分布熵

符号定义： D_i $(1 \le i \le m)$: 第i个领域类别　　WS_{Di}:类别Di的领域术语集合

d_{ij} $(1 \le j \le n_i)$: 类别D_i中的第j个文档　WS：文本中所有词语集合

l_{ij}: 文档d_{ij}的长度,即在该文档中出现的所有词语的词频之和

L_i: 类别D_i包含的所有文档长度之和

数学描述：

词语W的类间分布熵：$CD(W) = -\sum_{i=1}^{m} P(D_i \mid W) \log P(D_i \mid W)$　$P(D_i \mid W) = \frac{count(W, D_i)}{count(W)}$

词语“党性”和“知觉”在类别A(马列主义)均以0.5的概率出现,但“党性”只出现在A(马列主义)与D(政治、法律)两个类别,而“知觉”共出现在11个类别的语料

词语W的正规化的类间分布熵 NCD

$$NCD(W) = -\sum_{i=1}^{m} P'(D_i \mid W) \log P'(D_i \mid W)$$

$$P'(D_i \mid W) = \frac{P(D_i \mid W)/L_i}{\sum_{j=1}^{m} (P(D_j \mid W)/L_j)}$$

3) 类内分布熵

词语W在领域D_i的正规化的类内分布熵

NDD

$$NDD(W, D_i) = -\sum_{j=1}^{n_i} P'(d_{ij}|W) \log P'(d_{ij}|W)$$

“蛔虫”在类别G(文化、科学、教育、体育)的一篇介绍中小学生健康问题中蛔虫感染的文章中多次出现,但在该类别的其它文档中未出现,那么该词就不具有领域代表性,不能成为G类的领域术语.

$$P'(d_{ij}|W) = \frac{P(d_{ij}|W)/l_{ij}}{\sum_{j=1}^{m}(P(d_{ij}|W)/l_{ij})}$$

4) 词语的排序公式

$$RS(W, D_x) = -\lambda NCD(W) / \log m + (1-\lambda) NDD(W, D_x) / \log n_x$$

- *RS*(*W*, *Dx*)为*W*在*Dx*中的排序分值
- 将*NCD*和*NDD*均除以log(样本类别数)进行归一化，这样等式右边的每个加数的取值范围均为0到1

5) 抽取算法

- 输入：领域分类的语料库D，NCD 域值α，NDD 域值β
- 输出：每个类别的领域术语

- 说明：
  - 将类间分布熵值*NCD*小于α的词语作为领域术语候选
    - 这里假定领域术语不兼类，那么候选术语以最大正规化概率出现的领域即为其可能的相关领域
  - 将类内分布熵值*NDD*大于β的词语作为领域术语候选。

5) 抽取算法

(1) 将分类训练语料分词

(2) for i =1 to m do

(3) for j =1 to n_i do

(4) 依次读入 d_{ij} 中的每个词语 W 加入 WS 并记录频次，同时 l_{ij} 加 1

(5) end for

(6) 计算 L_i

(7) end for

(8) for all(W∈WS) do

(9) 计算 NCD(W)

(10) if NCD(W) < α then

(11) 求 $x = \arg\max_{X'}(P'(D_{x'} \mid W))$

(12) 计算 NDD(W, D_x)

(13) if NDD(W, D_x) > β then

(14) 将 W 加入 WS_{D_i} 中

(15) end for

6) 实验结果1：中图分类体系下抽取结果示例

C1 马列主义	C2 法律	C3 军事	C4 体育	C5 医药卫生	C6 轻工业
无产阶级	司法	作战	比赛	患者	包装
社会主义	人民法院	军种	首场	治疗	食品
全党	最高人民法院	军事	球员	血管	调味
马克思主义	案件	军队	球队	临床	保质期
共产主义	审理	战争	英格兰队	疗效	肉制品
马克思列宁主义	诉讼	兵力	队友	药物	玻璃瓶
马克思	法院	事变	世界杯	病人	品牌
无产者	司法机关	美军	夺冠	冠心病	肉食品
资产阶级	当事人	新军	冠军	并发症	方便化
共产主义社会	职权	战法	决赛	动脉	果汁
剥削	国家机关	我军	足协	冠状动脉	腥味
阶级	审判	武器	后卫	手术	肉类
生产资料	被告人	军兵种	主帅	症状术后	专卖店
恩格斯	民事	火力	瑞典队	口服	草莓
私有制	行使	军事科学	任意球	疗法	货架
资产者	民事诉讼	战场	小组赛	服用	糖度

6) 实验结果2：标准一过滤掉的部分词语

- P‘(D|W)值大于0.5，且NCD大于2（说明用类间熵来衡量词语的领域相关性比采用基于出现比例的方法更为合理）

B 哲学、宗教	反思 知觉 彼岸 有限性 万物 人权
E 军事	打击 变革 着眼 后勤 纵深 海湾
H 语言、文学	似的 埃及 字面 太极 听话 口腔
R 医药、卫生	狭窄 每日 医院 安慰 一例 陕西省
TD 矿业工程	考查 露天 倾角 混入 分期 花纹
TS 轻工业、手工业	粉末状 织成 居首 厂区 动物性 别致

6) 实验结果3：标准二过滤掉的部分词语

- 在其对应领域满足NCD值和NDD值均为零的部分词语（领域术语抽取第二条标准是合理并有效的）

B 哲学、宗教	笃实 知命 隐者 阴德 译稿 显摆 下篇 唤回 后王 法共
E 军事	像片 圣路易斯 肩章 公安部队 大尉 长机 巴解组织 流通性 塔山 北宁
H 语言、文学	硬腭 译员 娘娘 外务部 速记 舌尖 舌根音 平声 辣子 大姨
R 医药、卫生	珍奥 药丸 血肿 维生素 D 手术组 内窥镜 磨牙 门店 纯净水 清心
TD 矿业工程	紫金山 振磨机 砚台 砚石 探矿权 台板 三叶虫 南翼 辉铜矿 导水管
TS 轻工业、手工业	云锦 维生素 A 筒子 手袋 圣雪绒 三角区 乳粉 男式 华邦 金龙鱼

6) 实验结果4

中图分类体系下随机抽取的六个领域上的领域术语抽取数目

类别编号	词语总数	抽取词语个数	
		DR+DC	NCD+NDD
B 哲学、宗教	88830	1776	881
E 军事	41030	621	677
H 语言、文字	38666	638	741
R 医药、卫生	18182	444	571
TD 矿业工程	27925	318	162
TS 轻工业、手工业	21792	257	358

DR+DC方法抽取词语个数会随着语料规模的变化产生较大变化
(DR:Domain Relevance, DC:Domain Consistence)
NCD+NDD方法抽取词语数目不完全依赖于语料规模

6) 实验结果5

- 随机抽取的六个领域上领域术语抽取正确率

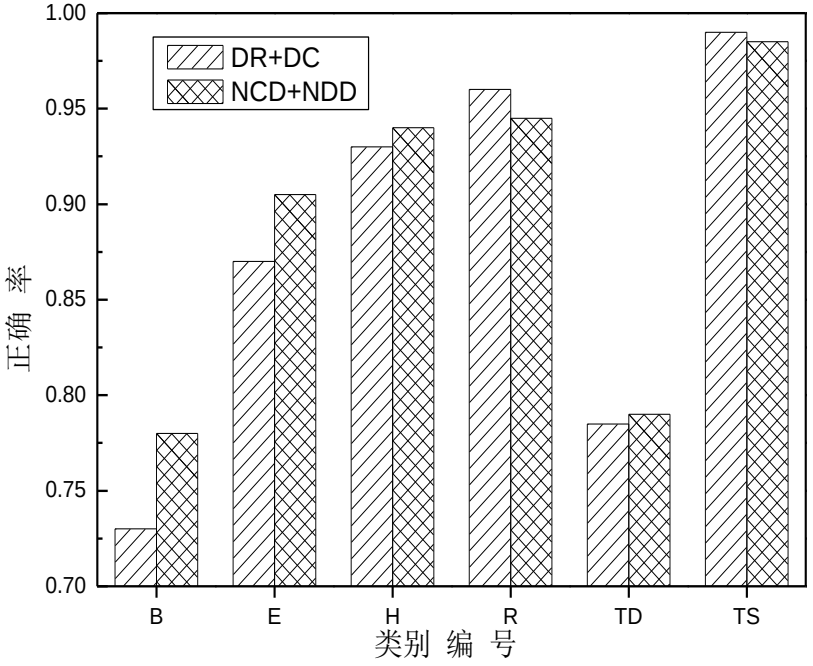


图1 前200个词语的正确率

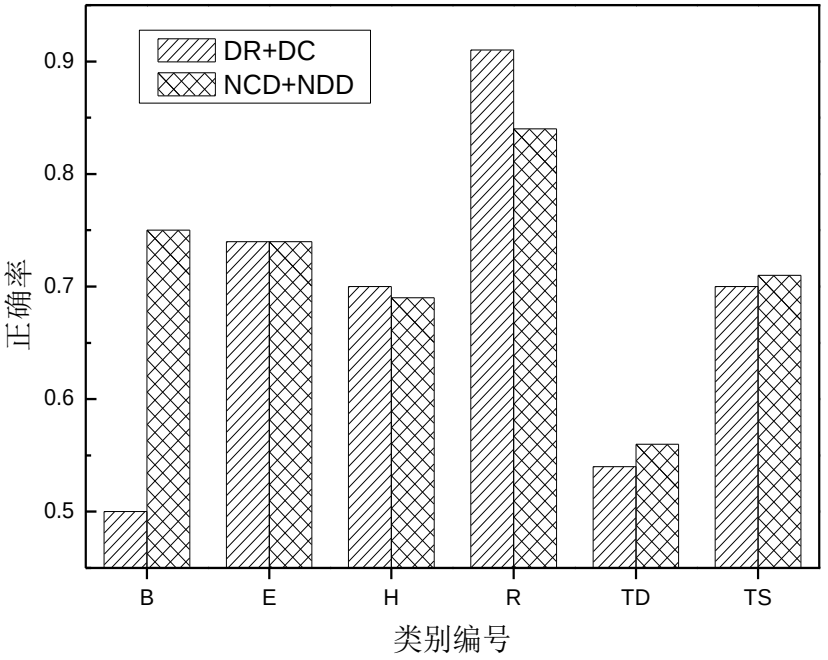


图2 其余词语的正确率

对于语料规模很大的类别, NCD+NDD法的正确率要明显高于DR+DC法
在其它正确率相当的类别中,抽取的术语数目要明显高于DR+DC法

4. 领域术语自动抽取的应用

- 4.1 在文本自动分类中的应用
- 4.2 在旅游领域问答式信息检索中的应用

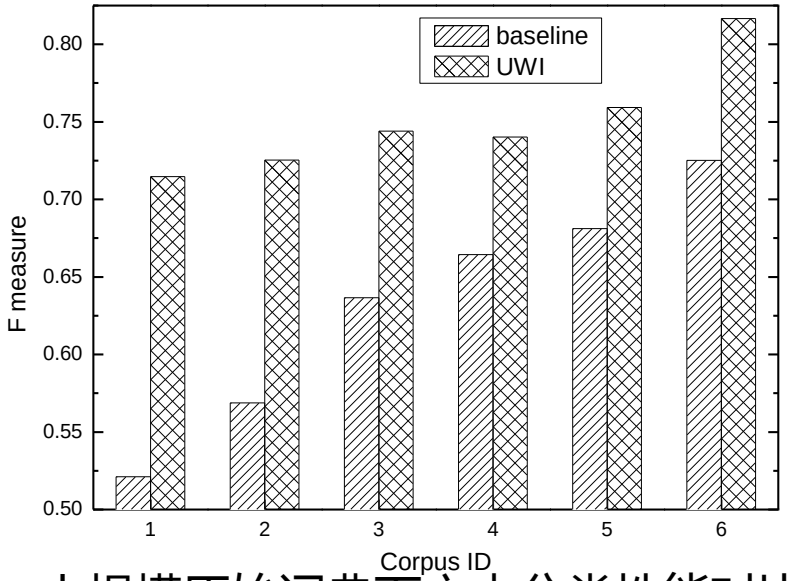
4.1 在文本自动分类中的应用

- 应用一：特征集扩展

  将基于局部最大算法的中文新词发现方法应用于文本分类语料的分词处理，从而扩大文本表示的特征集

特征集扩展前后文本分类性能对比

方法	准确率	召回率	F值
特征集扩展前	0.866	0.818	0.841
特征集扩展后	0.874	0.828	0.850



小规模原始词典下文本分类性能对比

4.1 在文本自动分类中的应用

- 应用二：特征选择

基于统计量度NCD+NDD的术语抽取 代替 传统特征选择

在中图分类数据集上的对比实验

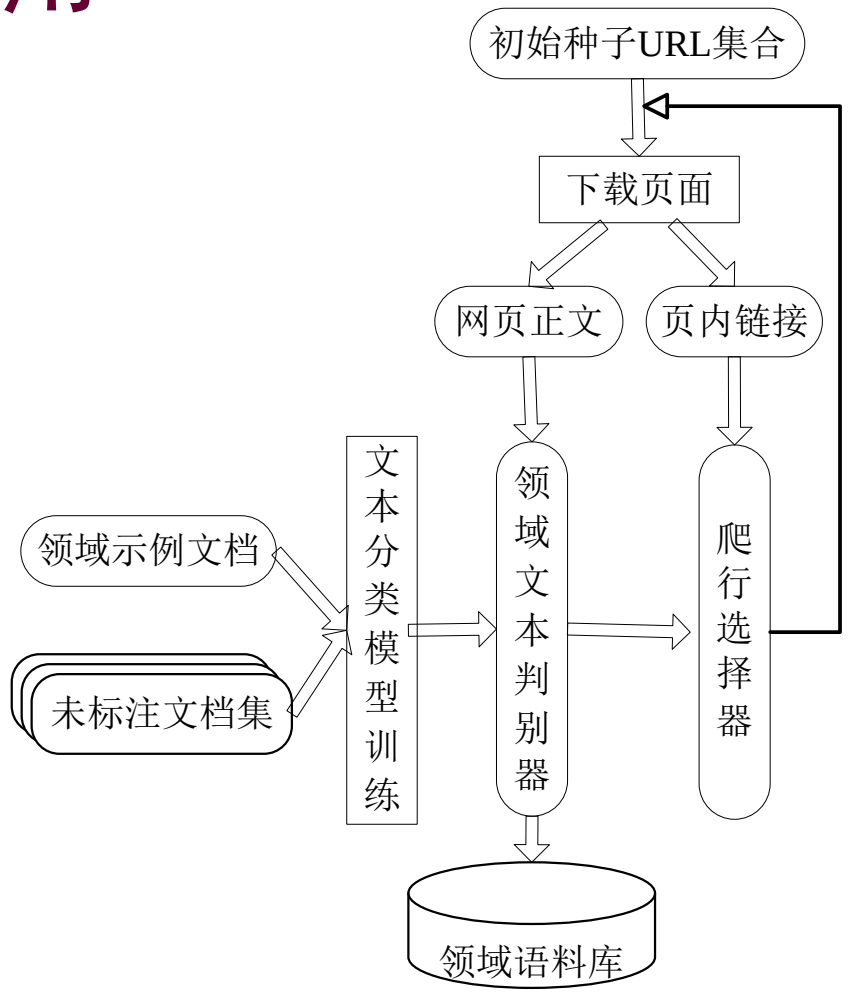
方法	准确率	召回率	F值
MI	0.419	0.409	0.414
DF	0.556	0.529	0.542
WE	0.564	0.541	0.552
IG	0.559	0.546	0.552
TFIDF	0.596	0.572	0.584
ECE	0.617	0.597	0.607
KFIDF	0.616	0.601	0.608
CHI	0.633	0.602	0.617
DR+DC	0.631	0.626	0.628
NCD+NDD	0.663	0.669	0.666

在旅游类数据集上的对比实验

方法	准确率	召回率	F值
MI	0.660	0.626	0.643
WE	0.719	0.672	0.695
DF	0.730	0.694	0.712
IG	0.751	0.701	0.725
KFIDF	0.783	0.746	0.764
TFIDF	0.785	0.750	0.767
ECE	0.790	0.776	0.783
CHI	0.802	0.782	0.792
DR+DC	0.853	0.801	0.826
NCD+NDD	0.874	0.828	0.850

4.2 在旅游领域问答式信息检索中的应用

- 应用一：专业文本采集
  - 网络文本采集是通过预先设定的种子URL集合，以各种不同的爬行策略循环迭代地访问Web下载网页
  - 当采集的信息只限定于特定的领域，出于性能上的考虑其不必也不可能对整个Web进行遍历
- 探讨了领域文本自动判别技术在专业文本采集中的应用



4.2 在旅游领域问答式信息检索中的应用

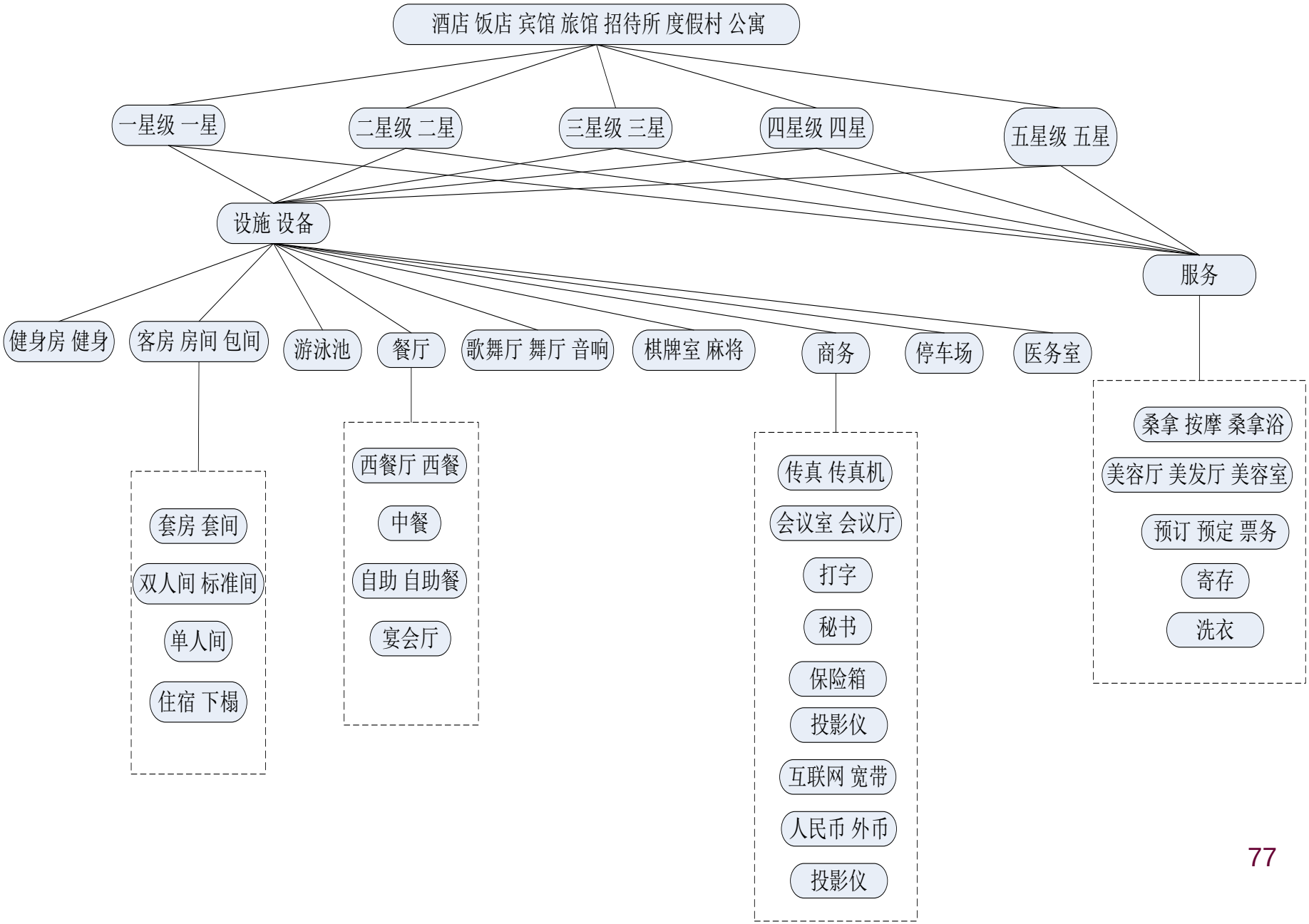
- 应用二：概念语义网络:以实现智能化的概念检索

领域术语抽取算法获取旅游领域术语 → 旅游领域概念语义网络

  - 构建了如下八个类别的概念语义网络

    宾馆饭店、城市概况、地方文化、交通指引、休闲娱乐、旅游景点、旅游服务、购物美食

宾馆饭店类别的部分概念语义网络图示



5. 小结

- 领域术语抽取的关键是对领域术语分布规律的准确把握：主观认识+客观需求+数学描述
- 领域术语抽取=(分词+)(新词发现+)领域判别？
- 领域术语抽取与名实体识别的对比
  - 目标不同，领域术语与名实体的交集很小
  - 构词规律不同
  - 通常用于名实体识别的机器学习方法较少用于领域术语抽取
- 某些问题有待深入研究（如某些方法的定量评价、基于全局上下文的领域术语抽取等）
- 某些问题需要课后了解（如基于正例的文本分类）

本章结束

谢　谢！