NLPIR大数据语义分析

NLPIR Big Data Semantic Analysis

张华平 博士 副教授

大数据搜索与挖掘实验室

kevinzhang@bit.edu.cn

@ICTCLAS张华平博士

2016.11

从搞笑微博开始...

恶搞明星体

最新指示

活动时间	2012.03.07-2012.03.14
活动规则	发布明星体回复 【格式：明星姓名最后一个字+字组成名词】
奖品设置	选出5名最具创意的回复者获得海尔榨汁机 
已搞明星	马云电视 范冰冰箱 苍井空调 奥巴马桶 周杰轮胎 罗玉凤爪 郑秀文胸 郭富城堡 蒋介石头 郭德纲彩 谢霆蜂窝 陈冠西饼 张柏芝士 曾轶可乐 梁朝伟哥 贝克汉母鸡 飞轮海底捞 钟楚红唇 齐达内裤 高晓松糕鞋 李开复印纸 章子怡糖 蓝心湄毛 尚雯婕毛

领导意见：可以搞

搞明星委员会 准搞章

@云南电视台

http://weibo.com/yuntelevision

，还好把问题解决了。回来经过陈冠
拿去冲水喝，路上顺道又买了点高晓
但天气还是很冷看来还是得把在郑少

转发(1) | 收藏 | 评论(2)

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

纲要

大数据
文本挖掘

I 文本挖掘基础知识综述

II NLPIR汉语分词与关键词提取

III 文本分类与聚类

IV NLPIR大数据挖掘平台与应用

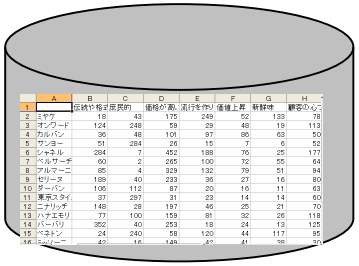
数据挖掘(DM： Data Mining)

- 结构化(Structured)数据⇒统计与数据挖掘技术，

矿山

研究对象

加工



数据收集和加工

宝

获取信息和知识

文本挖掘(TM: Text Mining)

- 文本是非结构化(Unstructured)的数据
- 文章、记号・文字的集合体

如何结构化？

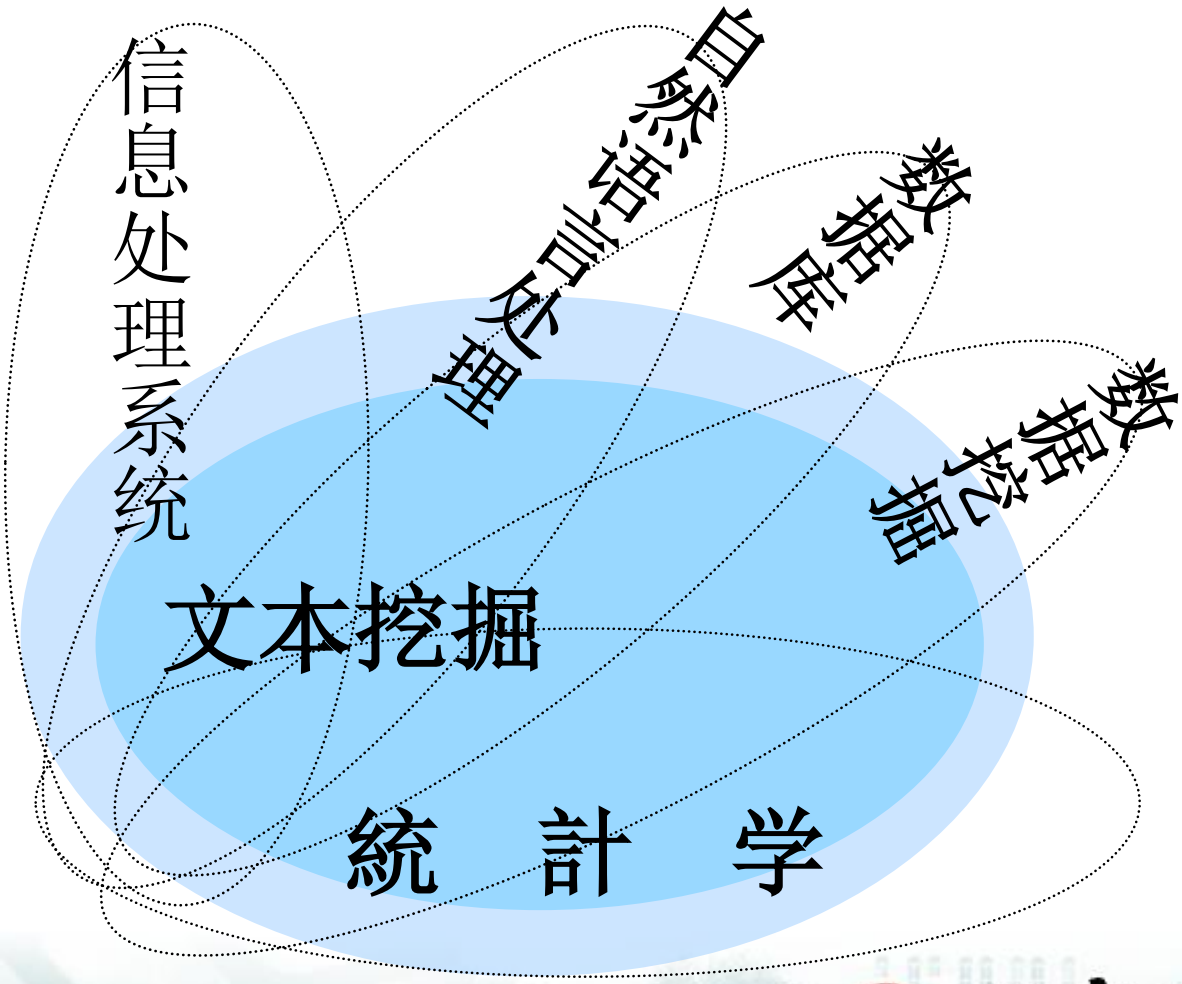
文本内的元素-->转换为向量或矩阵

文本

	A	B	C	D	E	F	G	H
1		伝統や格式	庶民的	価格が高い	流行を作り	価値上昇	新鮮味	顧客の心こ
2	ミヤケ	18	43	175	249	52	133	78
3	オンワード	124	248	59	29	48	19	113
4	カルバン	36	48	101	97	86	63	50
5	サンヨー	51	284	26	15	7	6	52
6	シャネル	284	7	452	188	76	25	177
7	ベルサーチ	60	2	265	100	72	55	64
8	アルマーニ	85	4	329	132	79	51	94
9	セリーヌ	189	40	233	36	27	16	80
10	ダーバン	106	112	87	20	16	11	63
11	東京スタイ	37	297	31	23	14	14	60
12	ニナリッチ	148	28	197	46	25	21	70
13	ハナエモリ	77	100	159	81	32	26	118
14	バーバリ	352	40	253	18	24	13	125
15	ベネトン	24	240	58	120	44	117	95
16	ミッシー	42	16	149	42	41	38	30

信息和知识

文本挖掘(TM: Text Mining)



文本挖掘的概念

- 是一个从非结构化的数据(文档)中获取用户感兴趣或者有用的模式或知识的过程
- 是一个复合学科的领域：信息技术，文本分析，模式识别，统计学，数据库技术，机器学习以及数据挖掘等技术
- 基础技术和知识：自然语言处理，数据处理(数理统计，数据挖掘，机器学习)

TM的基础

- 数理统计
- 数据挖掘
- 机器学习
- 信息处理
- 自然语言处理
- 计算语言

数据挖掘(DM)

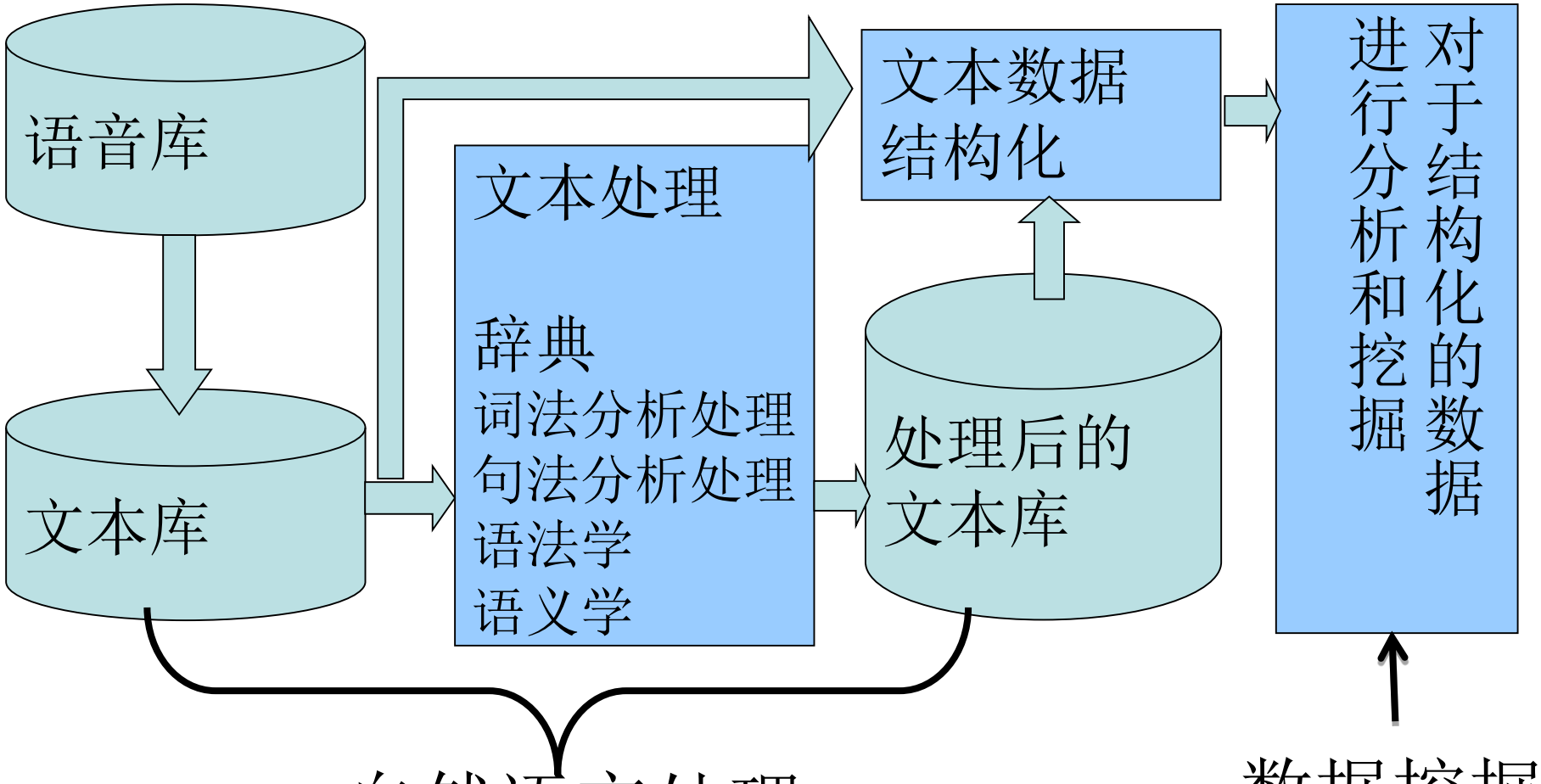
自然语言处理（NLP）

主要内容以及应用领域

- 主要内容：文本信息抽取，文本自动文摘，文本分类，文本聚类，文本数据压缩，关系抽取等

- 应用领域：企业的用户呼叫系统的内容管理与分析，企业内的日报分析，问卷调查分析，从blog中收集商机， Web挖掘，DNA信息处理，计算机网络的风险管理，文风分析，语料库语言学，医院的病志与记录的管理与分析等等

文本挖掘的框架


语音库
文本库
文本处理
辞典
词法分析处理
句法分析处理
语法学
语义学
文本数据
结构化
处理后的
文本库
对于结构化的数据
进行分析和挖掘
自然语言处理
数据挖掘

文本信息结构化的基本方法

- 字符信息→ 字频, n-gram
- 单词信息→ 词频, n-gram, co-occurrence
  同义词的聚类・・・
- 句子信息→ 句频, n-gram, co-occurrence
  词组为基础的文脉关系
- 现在的文本挖掘, 只是利用文本表面信息
- 没有真正达到语义处理

字符的n-gram

例：诺贝尔文学奖提名闹剧要闹到什么时候？

n=1 unigram

诺 贝 尔 文 学 奖 提 名 闹 剧 要 闹 到 什 么 时 候 ？

n=2 bigram

诺贝 贝尔 尔文 文学 学奖 奖提 提名 名闹 闹剧 剧要 要闹 闹到 到什 什么 ・・・

n=3 trigram

诺贝尔 贝尔文 尔文学 文学奖 学奖提 奖提名 提名闹 名闹剧 闹剧要 剧要闹 ・・・

把全文输入计算机

罗马士兵闯进阿基里德家的时候，他正在研究沙盘上的一个几何图形，他在罗马士兵的刀光戈影下张开双臂试图护住沙盘喊道：“”这是一则动人的传说，不知是记述还是虚构，反正这句“”流传千古，也值得流传千古。因为它体现了一种极为典型的西方精神，很难套在别种文化头上。

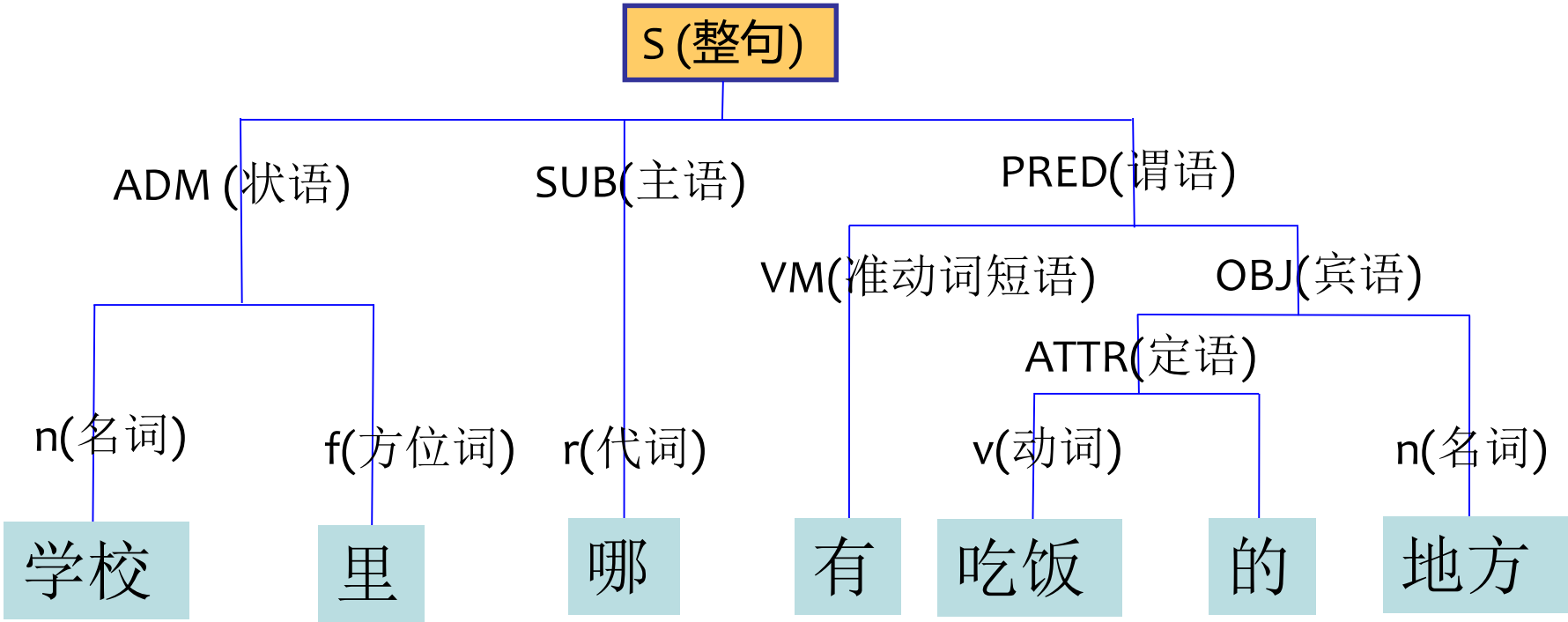
单词的切分

罗马/ns 士兵/n 闯进/v 阿基里德家/ns 的/u 时候/n ，/w 他/r 正在/d 研究/v 沙盘/n 上/f 的/b 一个/m 几何图形/l ，/w 他/r 在/p 罗马/ns 士兵/n 的/u 刀光戈/nr 影/ng 下/f 张开/v 双臂/n 试图/v 护/v 住/v 沙盘/n 喊/v 道/j ：/w “/w ”/w 这/r 是/v 一/m 则/q 动人/a 的/u 传说/n ，/w 不知/v 是/v 记述/v 还是/c 虚构/vn ，/w 反正/d 这/r 句/q “/w ”/w 流传/v 千古/n ，/w 也/d 值得/v 流传/v 千古/n 。/w

中文的计算机单词切分的精度在95 %左右

句法分析

S: 学校里哪有吃饭的地方。



中文的计算机句法分析的精度80%--90%

主要的统计分析

- 词或句子的频率分析
- 词或句子的同现分析
- 词或句子与文本的对应关系
- 聚类分析（文本的聚类，词的聚类）
- 文本的分类与识别
- 时间序列分析与预测

词项频率 tf

- 词项t的词项频率 $tf_{t,d}$ 是指t 在d中出现的次数
- 下面将介绍利用tf来计算文档评分的方法
- 第一种方法是采用原始的tf值(raw tf)
- 但是原始tf不太合适：
  - 某个词项在A文档中出现十次，即tf = 10，在B文档中 tf = 1，那么A比B更相关
  - 但是相关度不会相差10倍
- 相关度不会正比于词项频率tf

文档中的词频 vs. 文档集中的词频

- 除词项频率tf之外，我们还想利用词项在整个文档集中的频率进行权重和评分计算

罕见词项所期望的权重

- 罕见词项比常见词所蕴含的信息更多
- 考虑查询中某个词项，它在整个文档集中非常罕见 (例如 ARACHNOCENTRIC).
- 某篇包含该词项的文档很可能相关
- 于是，我们希望像ARACHNOCENTRIC一样的罕见词项将有较高权重

常见词项所期望的权重

- 常见词项的信息量不如罕见词
- 考虑一个查询词项，它频繁出现在文档集中 (如 GOOD, INCREASE, LINE等等)
- 一篇包含该词项的文档当然比不包含该词项的文档的相关度要高
- 但是，这些词对于相关度而言并不是非常强的指示词

- 于是，对于诸如GOOD、INCREASE和LINE的频繁词，会给一个正的权重，但是这个权重小于罕见词权重

文档频率(Document frequency, df

- 对于罕见词项我们希望赋予高权重
- 对于常见词我们希望赋予正的低权重
- 接下来我们使用文档频率df这个因子来计算查询-文档的匹配得分
- 文档频率指但是出现词项的文档数目

idf 权重

- df_t 是出现词项t的文档数目
- df_t 是和词项t的信息量成反比的一个值
- 于是可以定义词项t的idf权重:

$$idf_t = \log_{10} \frac{N}{df_t}$$

(其中N 是文档集中文档的数目)

- idf_t 是反映词项t的信息量的一个指标
- 实际中往往计算[log N/df_t]而不是 [N/df_t]，这可以对idf的影响有所抑制
- 值得注意的是，对于tf 和idf我们都采用了对数计算方式

idf的计算样例

- 利用右式计算idf_t:

$$idf_t = \log_{10} \frac{1{,}000{,}000}{df_t}$$

词项	df_t	idf_t
calpurnia	1	6
animal	100	4
sunday	1000	3
fly	10,000	2
under	100,000	1
the	1,000,000	0

文档集频率 vs. 文档频率

单词	文档集频率	文档频率
INSURANCE	10440	3997
TRY	10422	8760

- 词项t的文档集频率(Collection frequency)：文档集中出现的t词条的个数
- 词项t的文档频率：包含t的文档篇数
- 为什么会出现上述表格的情况？即文档集频率相差不大，但是文档频率相差很大
- 哪个词是更好的搜索词项？即应该赋予更高的权重
- 上例表明 df (和idf) 比cf (和“icf”)更适合权重计算

tf-idf权重计算

- 词项的tf-idf权重是tf权重和idf权重的乘积

$$w_{t,d} = (1 + \log \mathrm{tf}_{t,d}) \cdot \log \frac{N}{\mathrm{df}_t}$$

- 信息检索中最出名的权重计算方法
- 注意：上面的“-”是连接符，不是减号
- 其他叫法：tf.idf、tf x idf

数据的格式(变量xi是词或句子等)

text	x1	x2	x3	x4	x5	x6	x7 ·····	label
I1	37	41	25	33	10	12	12 ·····	A
I2	46	52	65	43	37	23	26 ·····	A
.								
.								
.								
M1	13	44	43	27	12	4	8 ·····	J
.								
.								
.								

$$P_{n\times m}=\left[p_{ij}=\frac{x_{ij}}{\sum_{j=1}^{m}x_{ij}}\right],\quad \sum_{j=1}^{m}p_{ij}=1$$

词或句子的频率分析

- 有可能频率高的较为重要
- 日本的首相演讲中的关键词分析
- 右表是演讲中使用的名词的次数

	安倍	福田	麻生	Fisher.p
民主党	0	0	12	1.6×10^{-7}
国	29	4	4	2.7×10^{-5}
もの	3	10	16	0.00032
不安	0	4	9	0.00034
私、わたし、わたくし	13	8	25	0.00045
立場	0	8	1	0.00048
行政	0	9	4	0.00082
安心	2	12	2	0.00117
将来	2	8	0	0.00264
私	13	8	0	0.00351
何	0	1	5	0.00426
国民生活	0	5	0	0.00465
⋮	⋮	⋮	⋮	⋮

日本的实例(为领取保险金的杀人案)

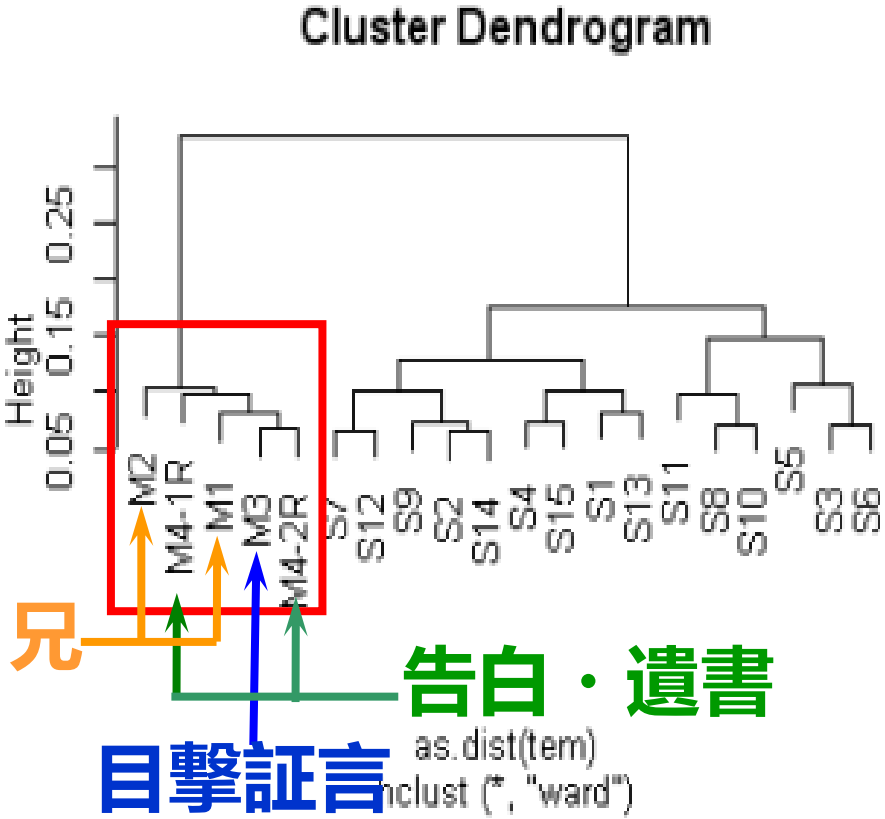
- 2003年5月日本警示厅搜查一科找金明哲教授
- 三年没有破案
- 1999年哥哥领取表弟的生命保险金问题
- 弟弟的死亡，车祸，可能是他杀
- 哥哥领取保险金
- 有哥哥写的两篇文章
- 警示厅收到两封信，一封为目击者的信，另一封为自供信兼遗书。
- 鉴定：两封信是否是哥哥写的

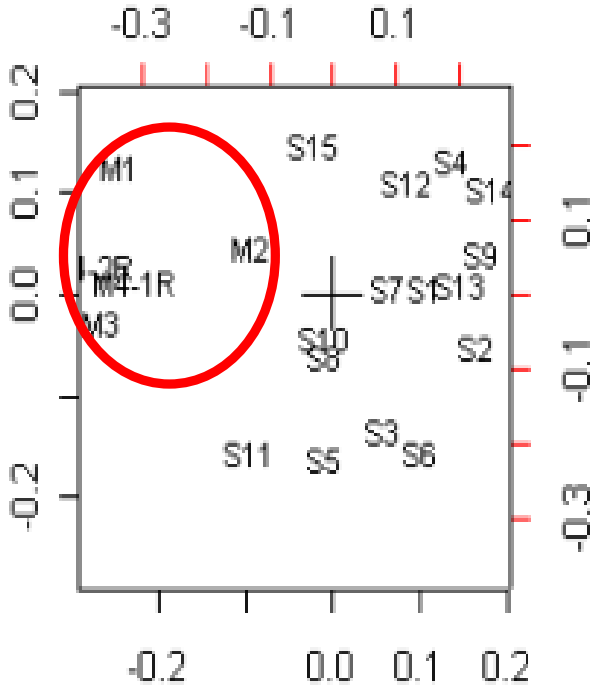
相关文档(为领取保险金的杀人案)

	字数
关于另一案件的文档（M1）	1677
关于上生命保险的文档 (M2)	1723
目击者的检举信（M3）	1636
自白兼遗书(M4)	3554

把文本M4分成2个文本。奇数文和偶数文（M4-1R, M4-2R）

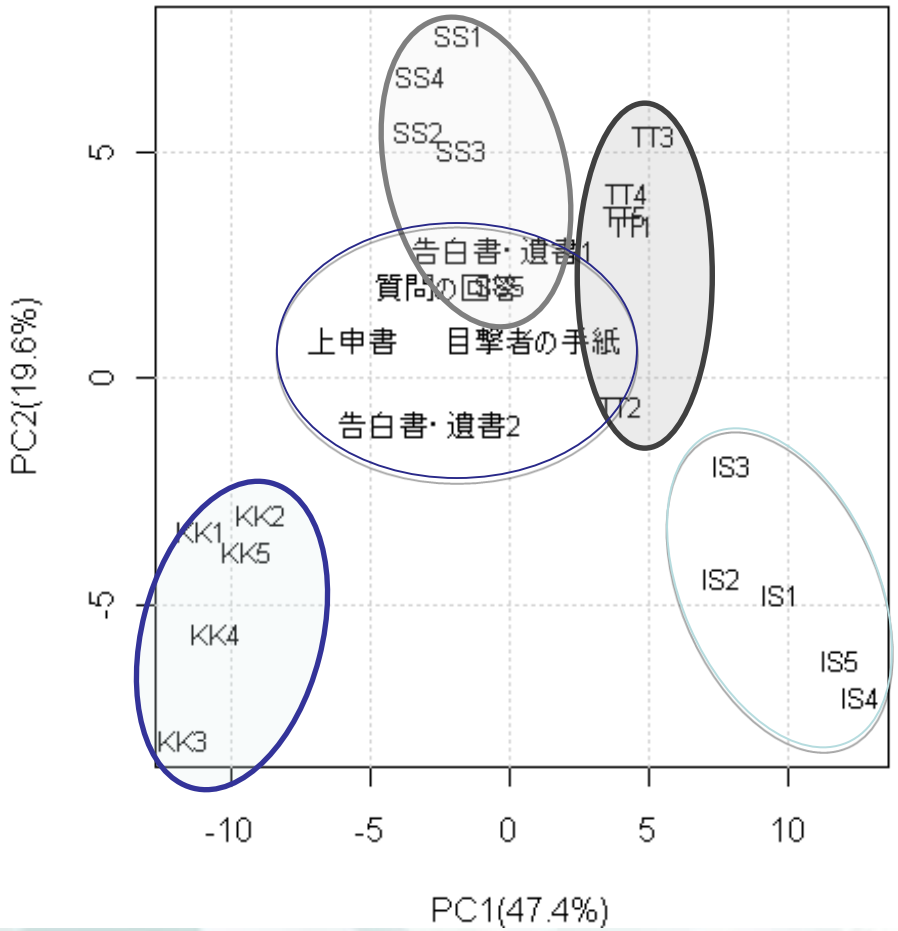
个个助词的频率

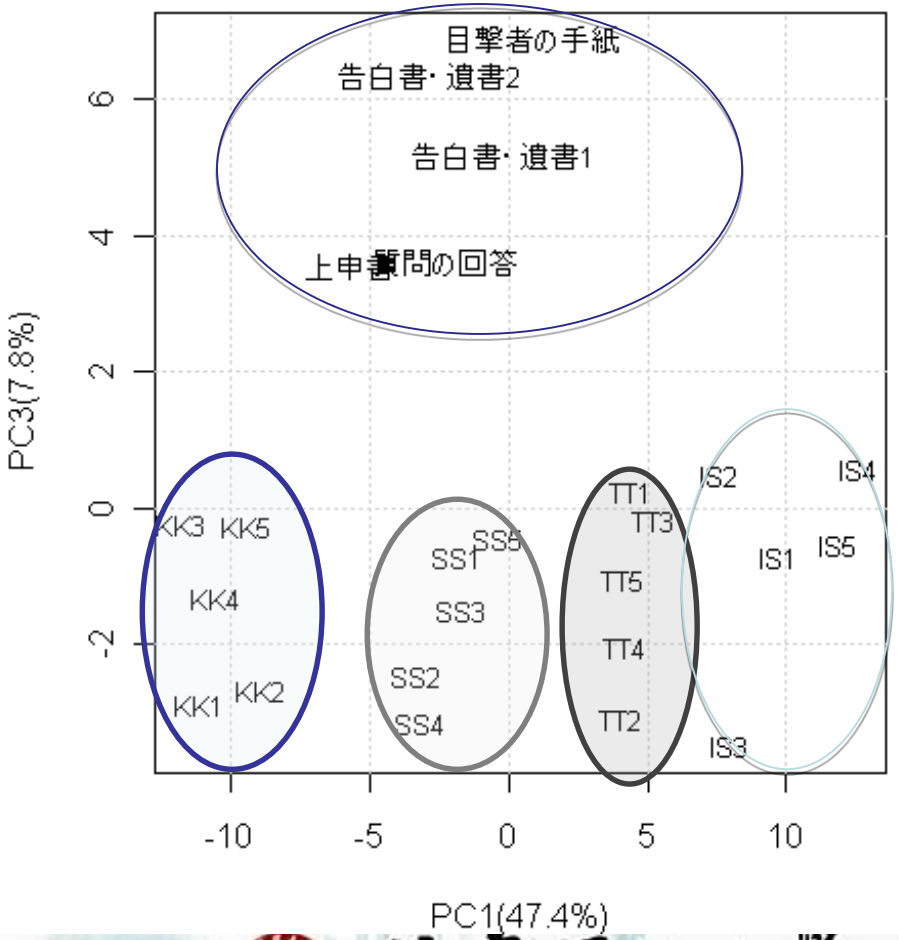




北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

五个人文本的主成分


SS1
SS4
SS2
SS3
TT3
TT4
TT5
TT1
告白書· 遺書1
質問の回答
上申書
目撃者の手紙
告白書· 遺書2
TT2
IS3
IS2
IS1
IS5
IS4
KK1
KK2
KK5
KK4
KK3
PC2(19.6%)
PC1(47.4%)
5
0
-5
-10
-5
0
5
10


目撃者の手紙
告白書· 遺書2
告白書· 遺書1
上申書
質問の回答
KK3
KK5
KK4
KK1
KK2
SS1
SS5
SS3
SS2
SS4
TT1
TT3
TT5
TT4
TT2
IS2
IS4
IS1
IS5
IS3
PC3(7.8%)
PC1(47.4%)
6
4
2
0
-2
-10
-5
0
5
10

2003年8月15日夜のニュース


事件解明のカギ
“文書鑑定”
浅草警察署
警視庁浅草警察署
ASAKUSA POLICE STATION


事件解明のカギ
"文書鑑定"
札幌学院大
金明哲教授
キム ミンチョル

日本企业的应用实例

- 佳能株式会社: 顾客服务呼叫中心的文本分析,有效利用全世界600万件数据(48%美国,22%日本)→把顾客的意见即时反应到新产品的开发
- 三井住友card株式会社:顾客服务呼叫中心的文本分析,每年1000万件以上的数据→提高服务质量
- 株式会社电通: 品牌印象分析
- KOKUYO株式会社: 新产品开发

TM的相关领域

商业
电话咨询
市场调查分析
顾客动向分析
顾客信息管理与分析
服务质量分析与服务

学术
信息检索与抽取
知识获取
语料库语言学
定量文体·文风学
作文自动判卷系统
:

频率分析
关联分析
聚类分析
分类分析
时间序列与预测

数据挖掘
机器学习
模式识别

人工智能
知识获取
机械学习

数据库
数据屋
文档库等

信息检索
信息抽取
信息摘要

自然语言处理
计算语言学
机器翻译，语音识别

纲要

大数据
文本挖掘

I 文本挖掘基础知识综述

II NLPIR汉语分词与关键词提取

III 文本分类与聚类

IV NLPIR大数据挖掘平台与应用

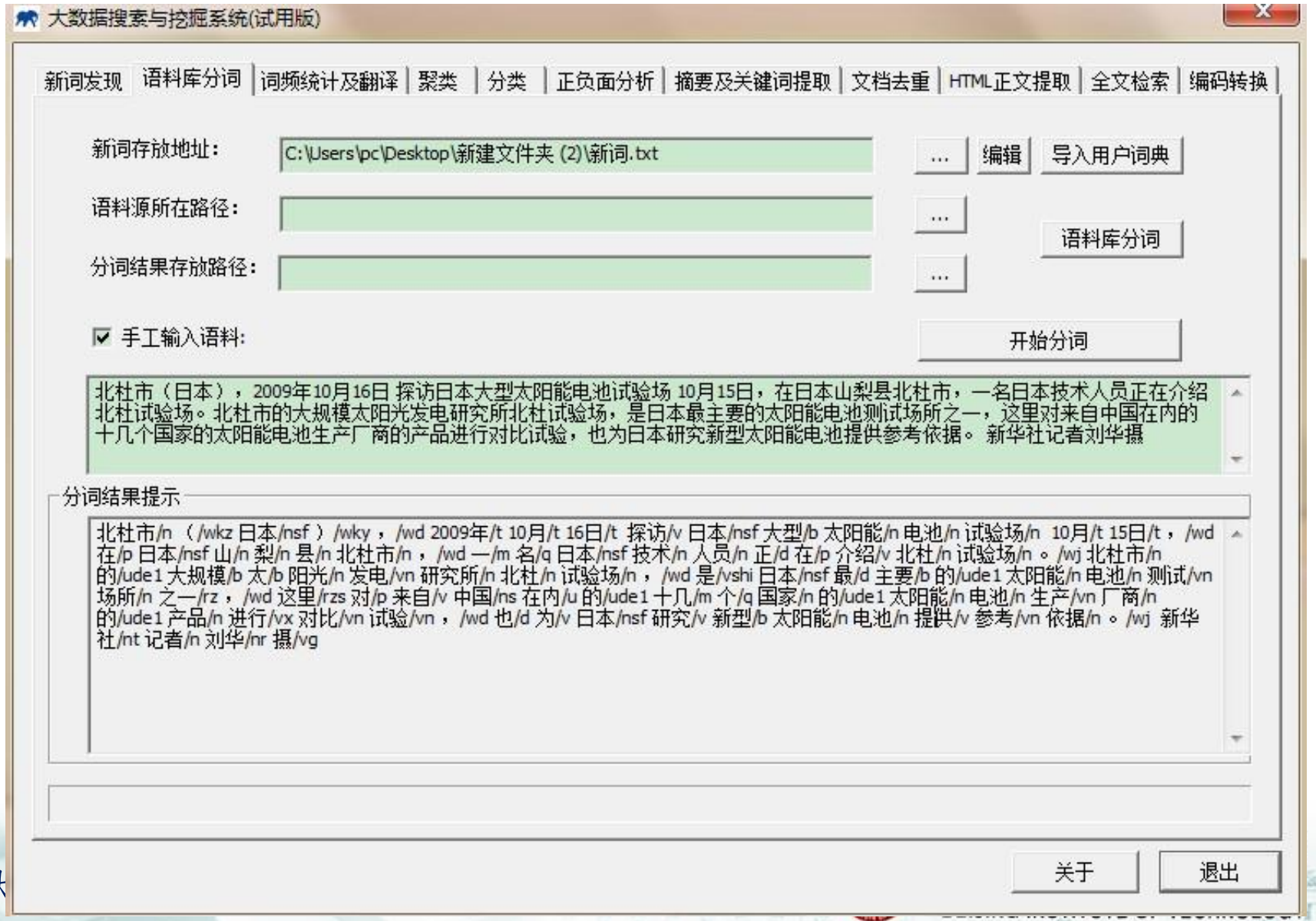
NLPIR大数据搜索与挖掘技术开发平台

- NLPIR网络搜索与挖掘共享开发平台，针对语言信息内容处理的全技术链条的共享开发平台。12年专业研究与工程积累，提供应用软件及各平台下的二次开发包，非商用永久免费。www.nlpir.org下载。

NLPIR 自然语言处理与信息检索共享平台
Natural Language Processing & Information Retrieval Sharing Platform

- 核心功能包括：
  - 搜索类：全文精准检索；
  - 语言类：新词发现，分词标注，统计分析与术语翻译；关键词提取；
  - 文档类：文本聚类及热点分析；分类过滤；自动摘要；文档去重；情感分析

NLPIR大数据搜索与挖掘技术开发平台

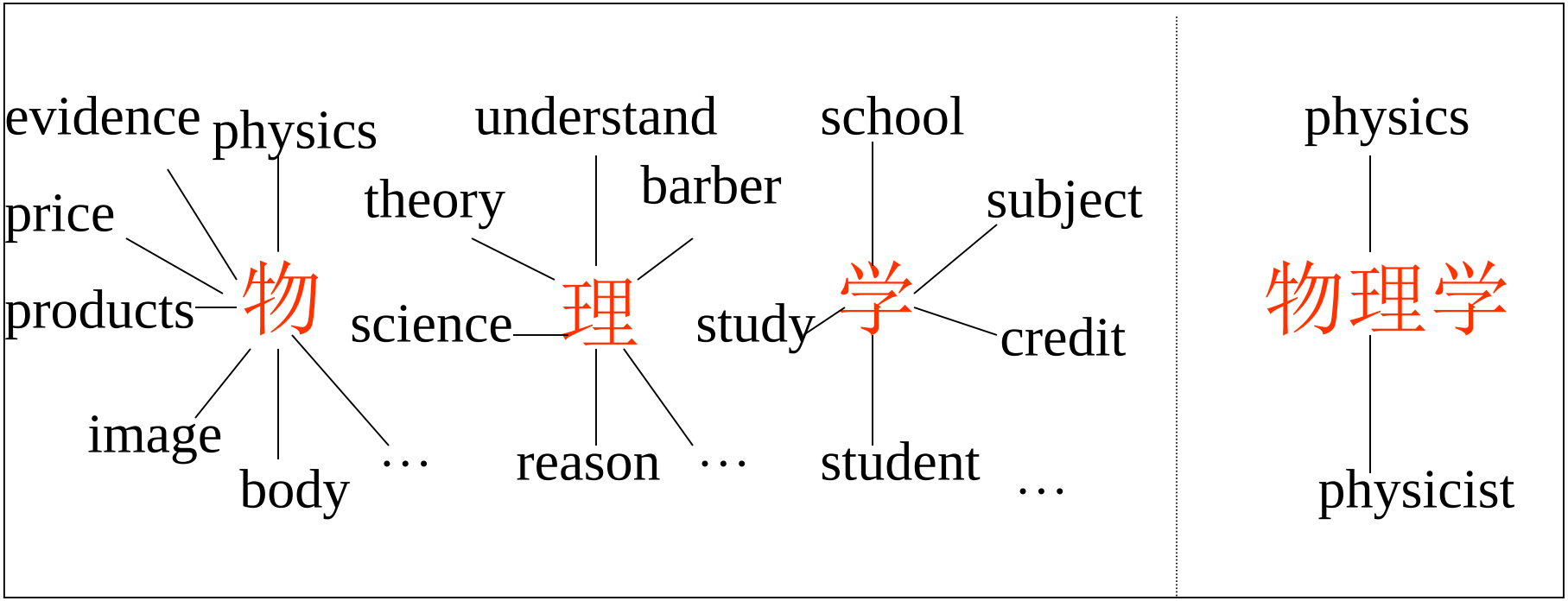


问题背景

- 汉语的书面语是按句分开的,词与词之间没有明确的分隔标记。
- 词是最小的能够独立活动的有意义的语言成分 。
- 中文信息处理只要涉及句法、语义(如检索、翻译、文摘、校对等应用),就需要以词为基本单位。句法分析、语句理解、自动文摘、自动分类和机器翻译等，更是少不了词的详细信息 。

分词的必要性：

词语信息熵大，计算速度更快



6 × 5 × 5 = 150 : 2

主要困难

重叠词、离合词、词缀

- 高高兴兴，高兴高兴，糊里糊涂，白花花，研究研究，个个，回回，工作工作（错误）
- 洗了一个澡，担什么心，发理了没有
- 学术性、花儿，盆儿

主要困难2：汉语的切分歧义

- 交集型歧义（交叉型歧义）：如果字串abc既可切分为ab/c，又可切分为a/bc。其中a，ab，c和bc是词；占86%.
  - 有意见： 我 对 他 有 意见。 总统 有意 见 他。
- 组合型歧义（覆盖型歧义）：若ab为词，而a和b在句子中又可分别单独成词，占14%.
  - 马上： 我 马上 就 来。 他 从 马 上 下来。
  - 将来： 我 将来 要 上 大学。 我 将 来 上海。
- 混合型歧义：由交集型歧义和组合型歧义自身嵌套或两者交叉组合而产生的歧义
  - 人才能：这样 的 人才 能 经受 住 考验。
  - 人才能：这样 的 人 才能 经受 住 考验。
  - 人才能：这样 的 人 才 能 经受 住 考验。

主要困难2续：歧义问题

- 歧义全局歧义与局部歧义：

  乒乓球拍/卖/完了；

  乒乓球/拍卖/完了；

  [护士对喝酒的病人说：]“小心/肝”

  [爱人对你说：]“小/心肝”

主要困难3：未登录词问题

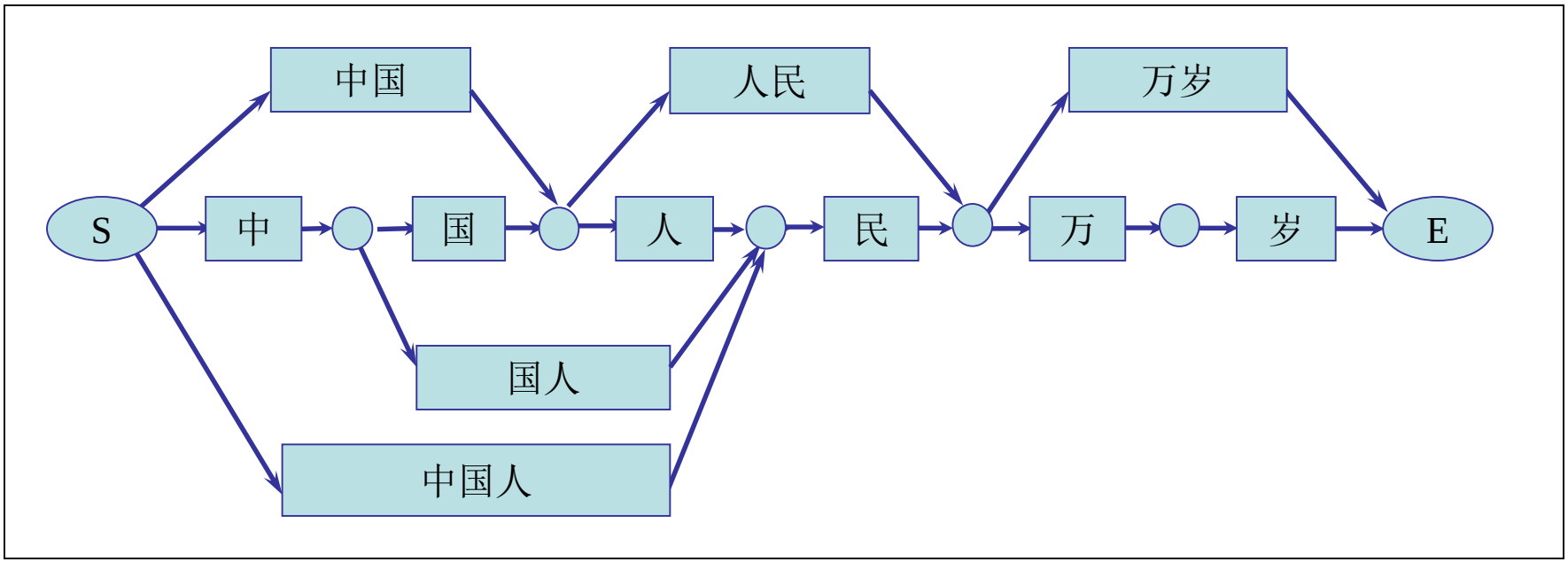
- 命名实体、新词术语往往不能全部收录到分词词典中，一般分词系统的词典是静态的，对未登录词的处理
- 干扰作用

克林顿**对内**塔尼亚**胡说**

龚学**平等**领导

- 根据我们的实验，未登录词和歧义问题大约占所有词语中的1.73%,但是导致了3.76%的切分错误。

汉语切分的数据结构—词图



根据这个数据结构，我们可以把词法分析中的几种操作转化为：

- 给词图上添加边（查词典，处理重叠词、离合词、前后缀和未定义词）；
- 寻找一条起点S到终点E的最优路径（切分排歧）；
- 给路径上的边加上标记（词性标注）；

汉语切分算法综述

- 规则方法
  - 全切分
  - 最大匹配方法
  - 最短路径方法
- 统计方法
  - N元语言模型；
  - 互信息、
  - 最大熵方法、条件随机场；
- 规则统计结合方法
  - N元语法

全切分方法

- 给出所有的切分结果
- 算法（略）
- 算法的时间复杂度随着句子长度的增加呈指数增长

最大匹配方法 1

- 正向最大匹配（MM）
  - 自左往右
  - 每次取最长词
- 逆向最大匹配（RMM)
  - 自右往左
  - 每次取最长词
- 双向最大匹配
  - 依次采用正向和逆向最大匹配
  - 如果结果一致则输出
  - 如果结果不一致再用其他方法排歧

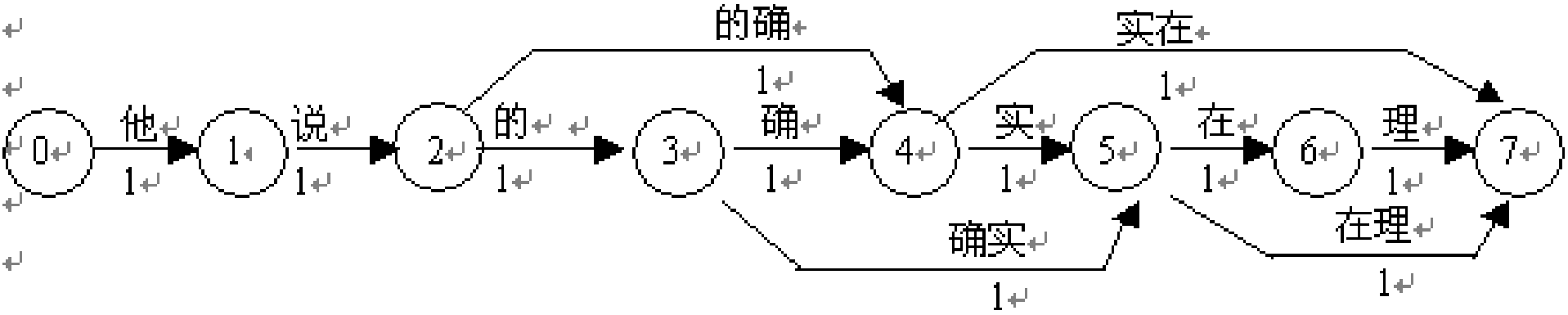
最大匹配方法 II

- 优点
  - 简单、快速
  - 在某些应用场合已经足够
- 缺点
  - 单向最大匹配会忽略交集型歧义和组合型歧义

    幼儿园 地 节目 / 独立自主 和平 等 互利 的 原则
  - 双向最大匹配会忽略链长为偶数的交集型歧义和组合型歧义

    原子 结合 成分 子时 / 他 从 马上 下来

最短路径方法



最短路径方法 II

- 基本思想：
  - 在词图上选择一条词数最少的路径
- 算法：
  - 动态规划算法
- 优点：好于单向的最大匹配方法
  - 最大匹配：独立自主 和平 等 互利 的 原则(6)
  - 最短路径：独立自主 和 平等互利 的 原则(5)
- 缺点：忽略了所有覆盖歧义，也无法解决大部分交叉歧义
  - 结合 成分 子时

N元语法分词方法

- 句子的出现概率用P(W)

$$P(W) = P(w_1 w_2 ... w_k) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1 w_2)...P(w_n \mid w_1 w_2 ... w_{n-1})$$
$$= \prod_{i=1}^{n} P(w_i \mid w_1 w_2 ... w_{i-1})$$

- 将分词问题转化为求概率最大的词语序列问题。
- 引入三元模型，不考虑未登录词问题，精度可以达到98%以上；
- 常用的模型为二元(一阶马尔科夫模型)和三元模型(二阶马尔科夫模型)

互信息与双字耦合度方法

- 互信息(MI，Mutual Information)用来表示两个字之间结合的强度

$$MI(x,y)=\log_2\frac{p(x,y)}{p(x)p(y)}=\log_2\frac{p(y\mid x)}{p(y)}$$

- 双字耦合度

$$Coup(<c_i,c_{i+1}>)=\frac{N(*c_m\ldots c_ic_{i+1}\ldots c_n*)}{N(*c_m\ldots c_ic_{i+1}\ldots c_n*)+N(*c_m\ldots c_i*c_{i+1}\ldots c_n*)}$$

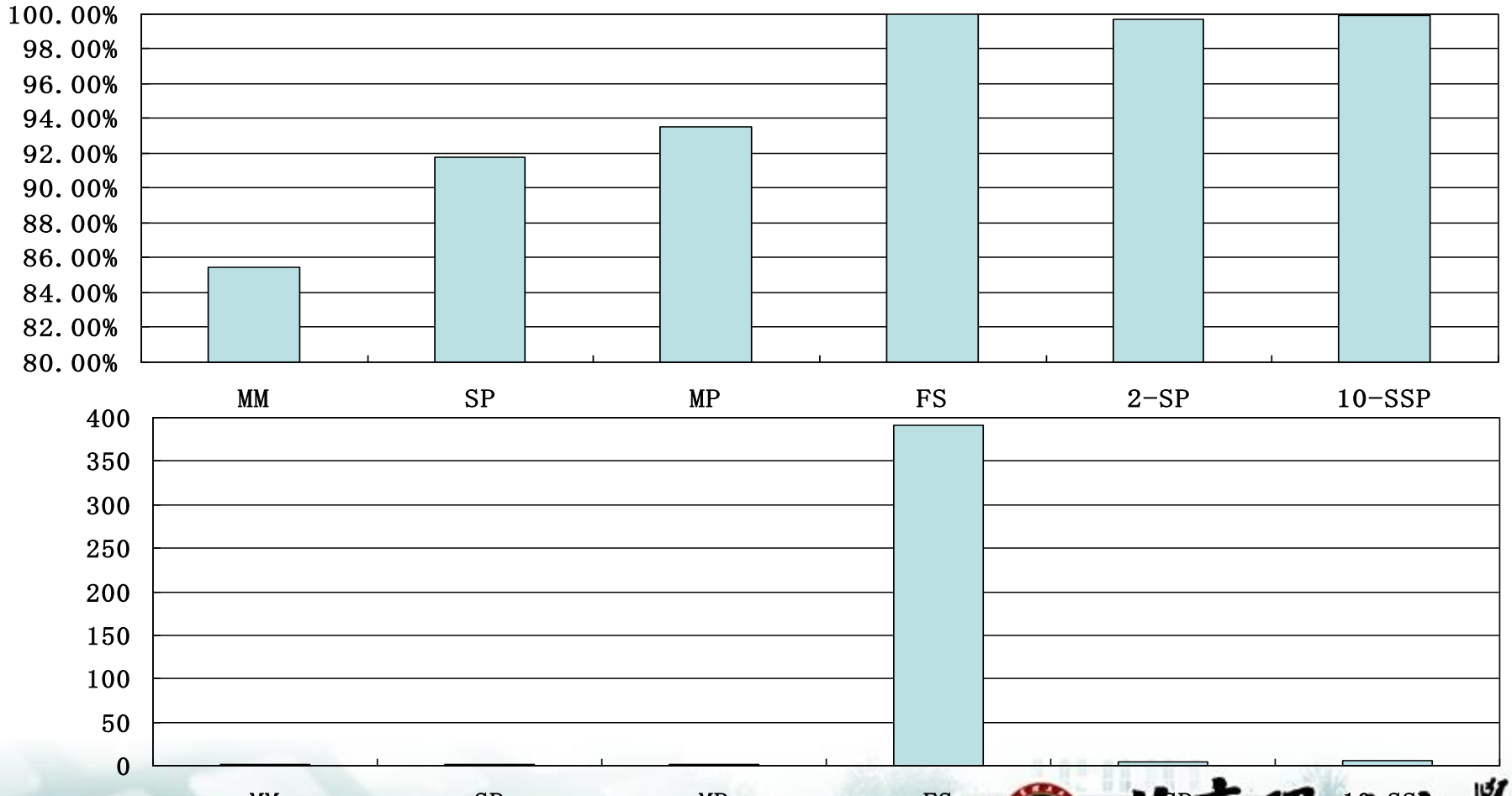
- “过目”这一双字对在出现16次，其中出现在“过目不忘”，“一一过目”这样的词中12次，而在“超过/目前”这样的语境中出现了4次，所以Coup（<过,目>）=12/(12+4)=0.75。
- 研究结果表明：随机字对总数超过3600万，但只有10万左右的字会相邻构词，规律性极强，可以通过这一规律进行分词。

其他方法

- 决策树方法：
  - 将分词问题转化为决策判断问题
- 最大熵方法：
  - 将字分为单字词、词首、词中、词尾，训练信息熵，最后将分词问题转化为求解信息熵最大的标注方法（类似与词性标注）。
  - 他/SS 说/SS 的/WF 确/WE 在/WF理/WE。/DELIM
- 最大压缩方法：
  - 将词语作为一个信息单元，最后对文本进行压缩，压缩比最好的信息单元就是最佳的分词结果。

相关切分算法的对比测试实验

[召回率/结果数]



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

Word class definition

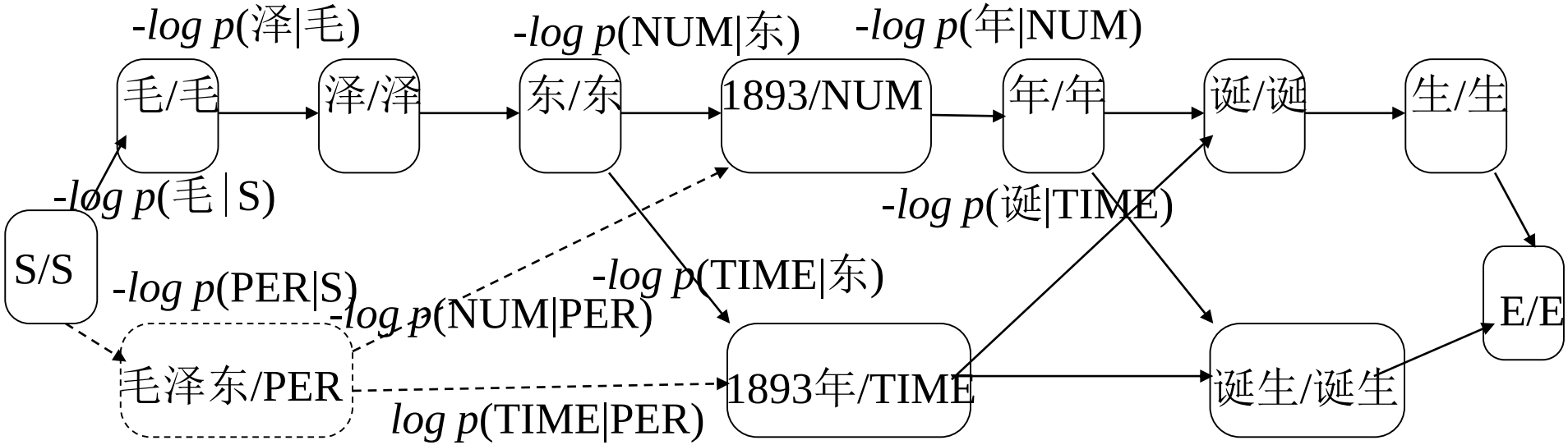
$c_i =$

- w_i iff w_i is listed in the segmentation lexicon;
- PER, LOC, ORG, TIME or NUM iff w_i is an unknown named entity;
- STR iff w_i is an unknown symbol string;
- BEG iff beginning of a sentence
- END iff ending of a sentence
- OTHER otherwise.

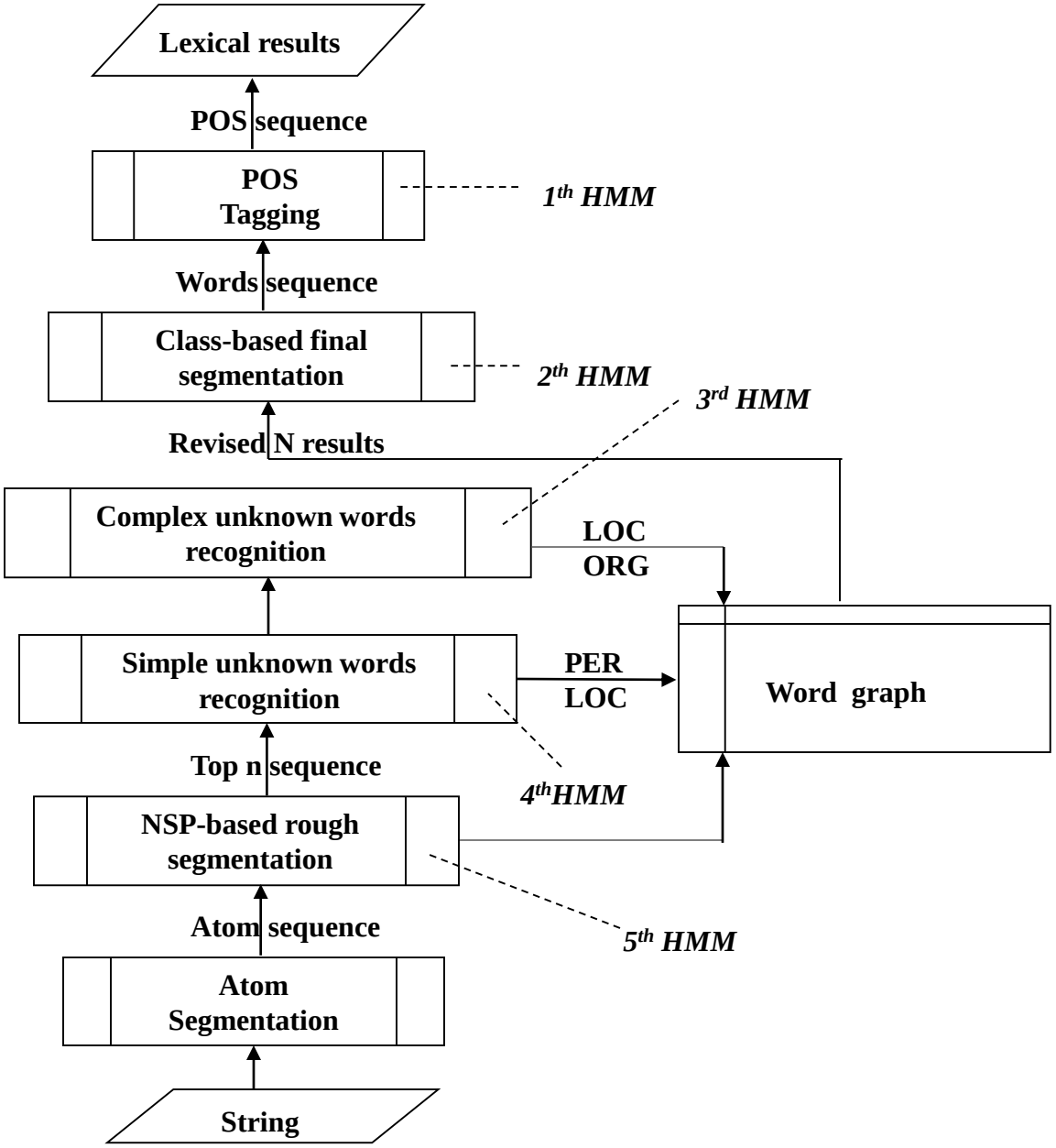
Class-based segmentation model

$$W^{*} = \arg\max_{W} P(W|C)P(C) \approx \arg\max_{w_1 w_2 \ldots w_m} \prod_{i=1}^{m} p(w_i \mid c_i)p(c_i \mid c_{i-1})$$

NLPIR/ICTCLAS2014分词

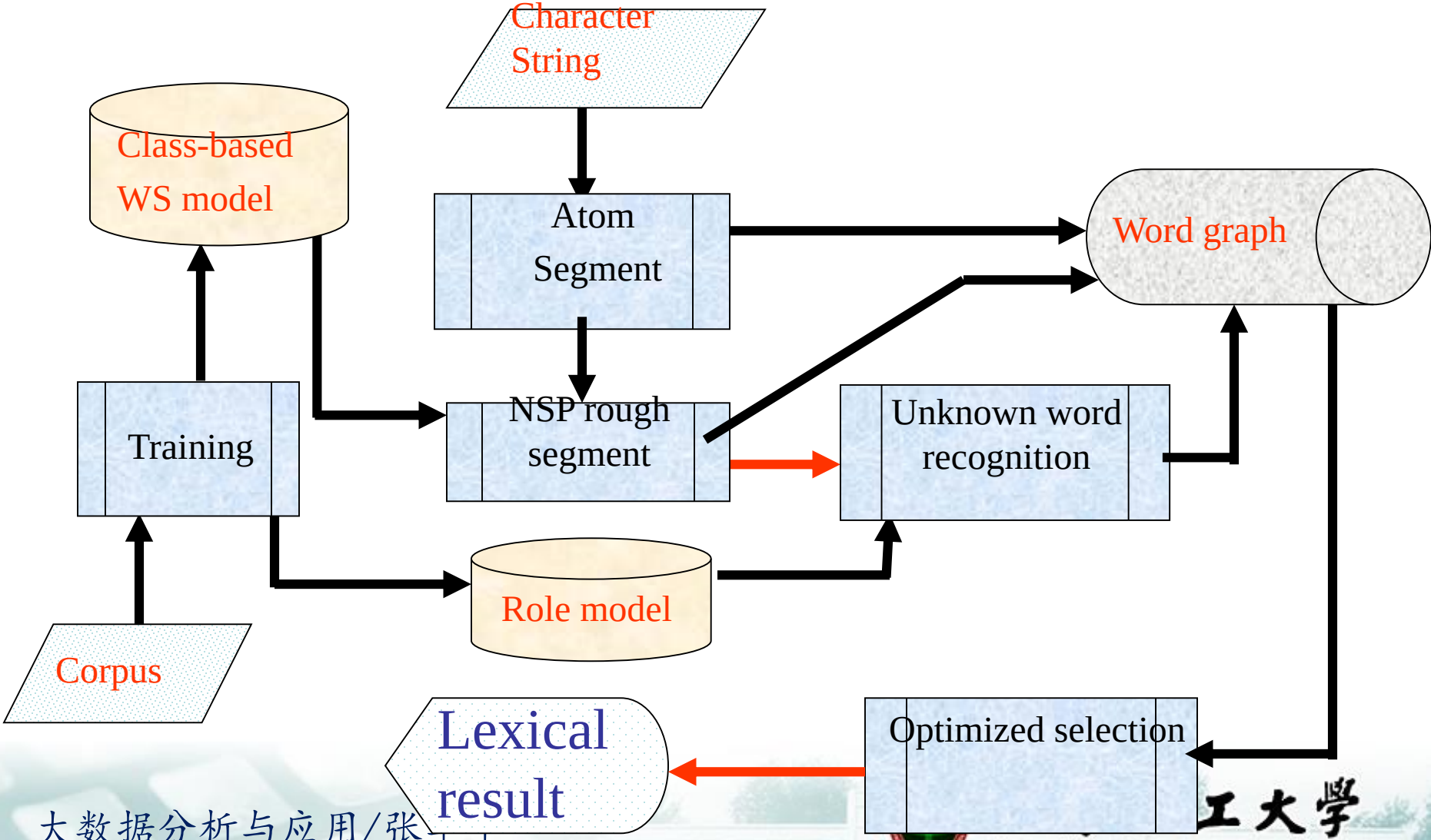


HHMM Architecture: Trace



HHMM-based Chinese lexical analysis

NLPIR/ICTCLAS2014分词



NLPIR之关键词提取

NLPIR汉语分词系统（又名：ICTCLAS2013版）张华平博士出品,新增新词发现、关键词识别与微博分词

NLPIR分词

分词

用户词典

关键词提取

指纹提取

相关介绍

，正确处理一致性和多样性的关系。坚持长期共存、互相监督、肝胆相照、荣辱与共的方针，加强同民主党派和无党派人士团结合作，促进思想上同心同德、目标上同心同向、行动上同心同行，加强党外代表人士队伍建设，选拔和推荐更多优秀党外人士担任各级国家机关领导职务。全面正确贯彻落实党的民族政策，坚持和完善民族区域自治制度，牢牢把握各民族共同团结奋斗、共同繁荣发展的主题，深入开展民族团结进步教育，加快民族地区发展，保障少数民族合法权益，巩固和发展平等团结互助和谐的社会主义民族关系，促进各民族和睦相处、和衷共济、和谐发展。全面贯彻党的宗教工作基本方针，发挥宗教界人士和信教群众在促

打开文件　分析　清空

关键词	权重
中国特色社会主义	20.21
改革开放	11.43
科学发展观	10.36
人民生活水平	9.99
经济发展方式	9.28
社会公平正义	8.74
收入分配差距	7.92
中华民族伟大复兴	7.91
城乡发展一体化	7.91
十年	7.74
基础设施	7.73
当代中国	7.73
发展	7.52
基本公共服务	7.49
生态文明建设	7.49
和平发展	7.07
开放型经济	7.06

主题特征词提取的交叉熵原理

- 词汇化：
  - 汉语分词与词性标注：采用NLPIR(ICTCLAS2014)；
  - 英语：几乎不需要做深度分析
  - 维语：同意适用
- 利用交叉信息熵计算有代表性的关键词w，权重$f(w) = \sum_l -p_l \ln p_l + \sum_r -p_r \ln p_r$；

主题特征词提取的交叉熵原理

- word=非公有制经济　　pos=n_new freq=7 LV=7 RV=7　unit_count=2　　weight=9.34
- Inverted List
  - (1397,1453,1458,1502,2062,2067,2099,)
- LV
  - (，(1),。(1),和(1),对(1),支持(1),引导(1),激发(1),)
- RV
  - (在(1),发展(1),活力(1),健康(1),都(1),财产权(1),各种(1),)

主题特征词提取的交叉熵原理

- word=非公有制　　pos=b　　freq=10 LV=9 RV=3 unit_count=1 weight=0.89
- Inverted List
  - (1397,1453,1458,1502,2062,2067,2099,2114,2125,8255,)
- LV
  - (，(1),。(1),和(1),对(1),鼓励(2),支持(1),引导(1),激发(1),制定(1),)
- RV
  - (经济(7),文化(1),企业(2),)

纲要

大数据
文本挖掘

I 文本挖掘基础知识综述

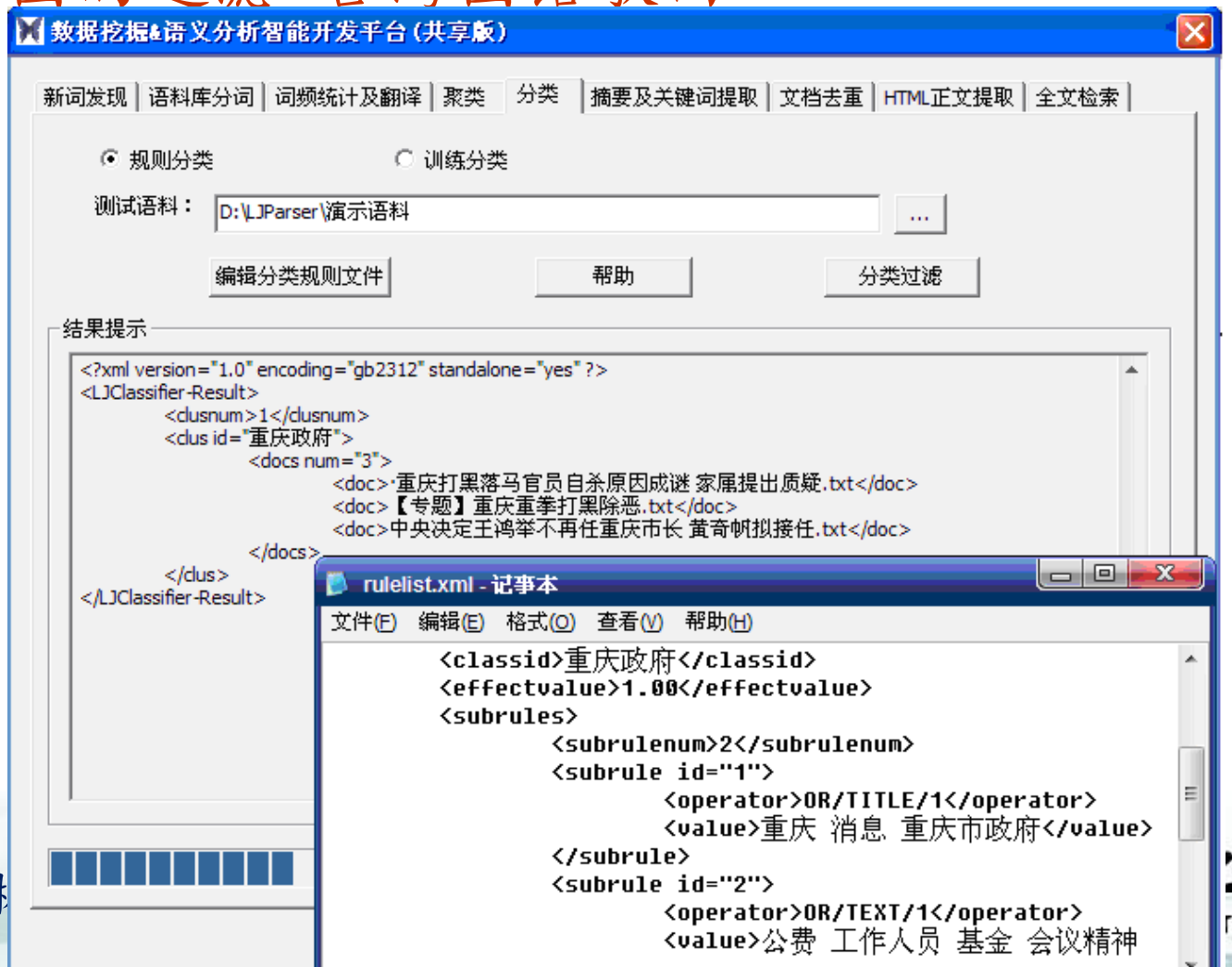
II NLPIR汉语分词与关键词提取

III 文本分类与聚类

IV NLPIR大数据挖掘平台与应用

NLPIR之大数据过滤分类

- A片的识别-世博A片区内，人们欢声雷动；
- 台湾国的过滤-台湾国语歌曲



NLPIR之大数据聚类

```
<?xml version="1.0" encoding="gb2312" standalone="yes" ?>
- <LJCluster-Result>
    <clusnum>200</clusnum>
  + <clus id="0">
  + <clus id="1">
  - <clus id="2">
      <feature>甲型 流感疫苗 流感病毒 流感确诊病例 流感疫情 流感病例 重症病例 接种疫苗 季节性流感</feature>
    - <docs num="150">
        <doc>多国H1N1流感疫情呈严重态势 美国疫苗供不应求</doc>
        <doc>卫生部就甲型H1N1流感疫情防控及疫苗预防接种工作答问</doc>
        <doc>甲型流感疫情在多国呈现严重态势</doc>
        <doc>中国报告4例接种甲流疫苗后死亡病例(实录)</doc>
        <doc>日本拟撤销对孕妇接种流感疫苗限制</doc>
        <doc>卫生部召开甲型流感防控疫苗接种通气会</doc>
        <doc>全球甲型流感疫苗接种进入高潮</doc>
        <doc>沪上近7万人已接种疫苗 天气转冷甲流危险性将累加</doc>
        <doc>对鸡蛋过敏者不宜接种甲流疫苗</doc>
        <doc>钟南山：甲流疫苗副作用小 学校应优先接种</doc>
        <doc>甲流疫苗答问公布:对鸡蛋过敏者不宜接种疫苗</doc>
        <doc>撩开甲型H1N1流感神秘的面纱:寻踪 变异 应对</doc>
        <doc>世卫公布甲型H1N1病毒调查：新病毒 人类无免疫</doc>
        <doc>福建、山东发生聚集性甲型Ｈ１Ｎ１流感疫情</doc>
        <doc>北京首现甲流死亡病例：北航军训团1名新生死亡</doc>
        <doc>流感疫苗对孕妇及胎儿免疫率达90%</doc>
        <doc>甲流可控可防可治</doc>
        <doc>首批甲流疫苗已运抵澳门 高危族23日起优先接种</doc>
        <doc>新疆报告1例甲流死亡病例 3类人群禁止接种疫苗</doc>
        <doc>法专家称接种甲型流感疫苗有助长期防疫</doc>
        <doc>接种疫苗仍是最有效手段</doc>
        <doc>美国甲流疫苗严重短缺 感染人数以百万计</doc>
        <doc>专家提醒:注射甲流疫苗后也应注重个人防护</doc>
        <doc>欧洲卫生专家称：甲流可能导致欧洲4万人丧命 </doc>
        <doc>成都严禁以防控甲型流感为由推销其它疫苗</doc>
```

引言

- 分类/聚类是大自然的固有现象：物以类聚、人以群分
  - 相似的对象往往聚集在一起
  - (相对而言)不相似的对象往往分开



什么是分类?

- 简单地说，分类(Categorization or Classification)就是按照某种标准给对象贴标签(label)

男

女

分类非常普遍

- 性别、籍贯、民族、学历、年龄等等，我们每个人身上贴满了“标签”
- 我们从孩提开始就具有分类能力：爸爸、妈妈；好阿姨、坏阿姨；电影中的好人、坏人等等。
- 分类无处不在，从现在开始，我们可以以分类的眼光看世界☺

文本分类的定义
Text Categorization/Classification

- 事先给定分类体系和训练样例(标注好类别信息的文本)，将文本分到某个或者某几个类别中。
  - 计算机自动分类，就是根据已经标注好类别信息的训练集合进行学习，将学习到的规律用于新样本(也叫测试样本)的类别判定。
  - 分类是有监督/指导学习(Supervised Learning)的一种。

文本分类的模式

- 从类别数目来分
  - 2类(binary)问题，类别体系由两个互补类构成，一篇文本属于或不属于某一类。
  - 多类(multi-class)问题，类别体系由三个或者以上的类别构成，一篇文本可以属于某一个或者多个类别，通常可以通过拆分成多个2类问题来实现，也有直接面对多类问题的分类方法
- 从是否兼类看分
  - 单标签(single label)问题：一个文本只属于一个类
  - 多标签(multi-label)问题：一个文本可以属于多类，即出现兼类现象

关于分类体系

- 分类体系的构建标准可以是按照语义(如：政治、经济、军事...)，也可以是按照其他标准(如：垃圾 vs. 非垃圾；游戏网站 vs. 非游戏网站)，完全取决于目标应用的需求。
- 分类体系一般由人工构造，可以是层次结构。一些分类体系: Reuters语料分类体系、中图分类、Yahoo！分类目录。
- 对于计算机而言，分类体系就是一棵目录树，训练样例文本就是最后的叶子节点。而且对于计算机处理而言，只需要训练样例文本及其对应类别信息，整个过程通常并不会考虑类别标签的意义。也就是说：几篇文档合在一起表示某个类别。

应用

- 垃圾邮件的判定
  - 类别 {spam, not-spam}
- 新闻出版按照栏目分类
  - 类别 {政治,体育,军事,…}
- 词性标注
  - 类别 {名词,动词,形容词,…}
- 词义排歧
  - 类别 {词义1,词义2,…}
- 计算机论文的领域
  - 类别 ACM system
    - H: information systems
    - H.3: information retrieval and storage

p://www.yahoo.com/?r37=1196390177

Make Y! your home page | Get Internet Explorer 7

Recent movie releases: Enchanted, August Rush, Hitman

YAHOO!

Web | Images | Video | Local | Shopping | more

modern information retrieval

Web Search

Beta!

Yahoo! Home | My Yahoo! | Y! China

Nov 30, 2007 | Page Options

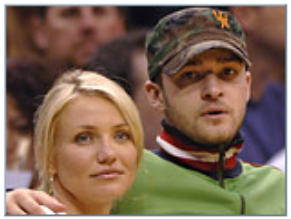
- Answers
- Autos
- Finance
- Games
- Groups
- HotJobs
- Maps
- Mobile Web
- Movies
- Music
- Personals
- Real Estate
- Shopping
- Sports
- Tech
- Travel
- TV
- Yellow Pages

More Yahoo! Services

Small Business

- Get a Web Site
- Domain Names
- Sell Online
- Search Ads

Featured | Entertainment | Sports | Video



Big celeb splits of 2007

Justin and Cameron are on the list. Find out who else ended their relationships this year. » Surprises

- Jessica Simpson, Romo set up by dad
- Latest scoop on celebrity couples


Biggest celebrity breakups of the year


College football's best players named


World's most intriguing billionaire heiresses


Top 10 consumer tech wrecks of 2007

» More Featured

In the News | World | Local | Finance

As of 10:36 a.m.

- Bush pushes Democrats to approve Iraq war funds | Bush
- Australia wants Iraq troops home by mid 2008 | Drawdown
- Clinton woos evangelicals at AIDS conference | '08 race
- FBI investigating use of stun gun on pregnant woman in Ohio
- Bernanke hints at another interest rate cut to bolster economy
- Study: Aggressive female antelopes more likely to get a mate
- Texas prison guards to get uniform makeover next year

NFL • NBA • NCAA Hoops • NCAA Football • NHL • NASCAR

» More: News | Popular | Election '08

Markets: Dow: +0.2% Nasdaq: +0.2% Sponsored by: Scottrade

Marketplace

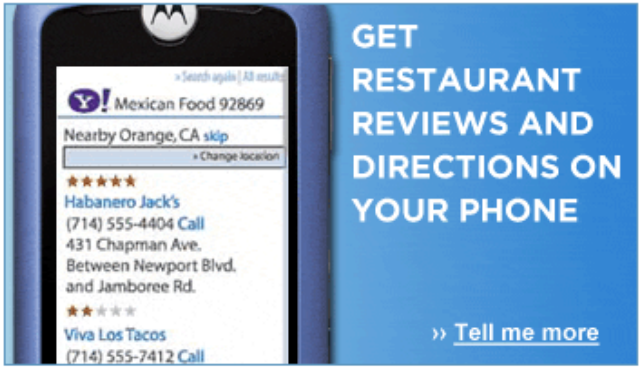

Research cars on Yahoo! Autos

We have new and used vehicle pricing, pictures and user

Check your mail status: Sign In Free mail: Sign Up

Mail | Messenger | Radio

Weather | Local | Horoscopes



Be a Better Online Destination

Get a great looking web site

Building a site is easier than ever with new tools from Yahoo! Web Hosting.

» Sign up today and save 25%

Pulse - What Yahoos Are Into

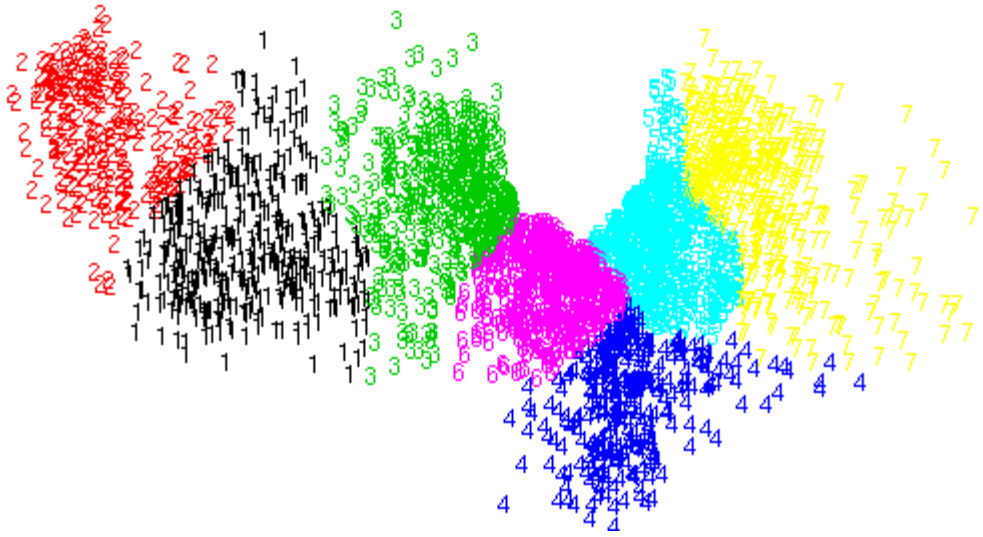
Popular Celebrity Mobile Searches

- Salma Hayek
- Rihanna
- Angelina Jolie
- Eva Mendes

什么是聚类

- 简单地说，聚类是指事先没有“标签”而通过某种成团分析找出事物之间存在聚集性原因的过程。
  - 去研究生院一个大教室上自习，往往发现大家三三两两扎推地坐，一打听，原来坐在一块的大都是一个班的。
  - 事先不知道“标签”，根据对象之间的相似情况进行成团分析。
  - Exploratory analysis(探索性数据分析)的一种

一个聚类的例子



信息处理中分类和聚类的原因

- 分类/聚类的根本原因就是因为对象数目太多，处理困难
  - 一些信息处理部门，一个工作人员一天要看上千份信息
  - 分门别类将会大大减少处理难度，提高处理效率和效果

分类/聚类的过程

- 对对象进行表示
  - 表示方法
  - 特征选择
- 根据某种算法进行相似度计算
  - 相似度计算方法
  - 分类/聚类方法

人工方法和自动方法

- 人工方法：人工总结规则
  - 优点：
    - 结果容易理解：如 足球 and 联赛→体育类
  - 缺点：
    - 费时费力
    - 难以保证一致性和准确性(40%左右的准确率)
    - 专家有时候凭空想象，没有基于真实语料的分布
  - 代表方法：人们曾经通过知识工程的方法建立专家系统(80年代末期)用于分类。
- 自动的方法(学习)：从训练语料中学习规则
  - 优点：
    - 快速
    - 准确率相对高(准确率可达60%或者更高)
    - 来源于真实文本，可信度高
  - 缺点：
    - 结果可能不易理解 (比如有时是一个复杂的数学表达式)

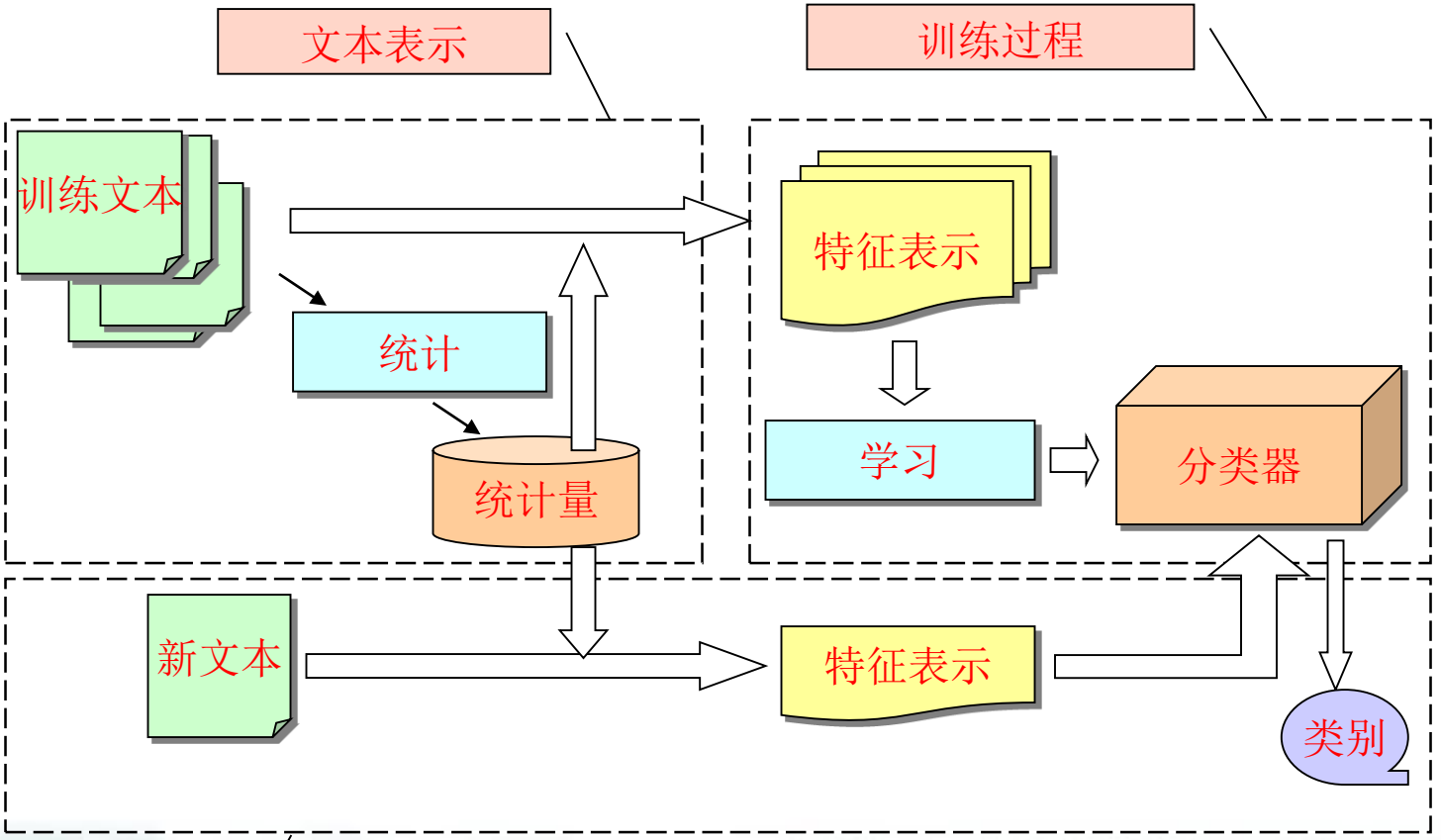
规则方法和统计方法

- 规则方法通过得到某些规则来指导分类，而这些规则往往是人可以理解的。
- 统计方法通过计算得到一些数学表达式来指导分类。
- 规则方法和统计方法没有本质的区别，它们都是想得到某种规律性的东西来指导分类，统计方法得到的数学表达式可以认为是某种隐式规则。
- 在目前的文本分类当中，统计方法占据了主流地位。

文本分类的过程

- 两个步骤：
  - 训练(training)：即从训练样本中学习分类的规律。
  - 测试(test或分类classification)：根据学习到的规律对新来的文本进行类别判定。
- 文本表示(text representation)：
  - 不管是训练还是测试，都要先分析出文本的某些特征(feature，也称为标引项 term)，然后把文本变成这些特征的某种适宜处理的表示形式，通常都采用向量表示形式或者直接使用某些统计量。

文本分类系统的组成框架



特征抽取(Feature Extraction)

- 预处理
  - 去掉html一些tag标记
  - 禁用词(stop words)去除、词根还原(stemming)
  - (中文)分词、词性标注、短语识别、…
  - 标引项频率统计
    - $TF_{i,j}$: 特征i在文档j中出现次数，标引项频率(Term Frequency)
    - DF_i: 所有文档集合中出现特征i的文档数目，文档频率(Document Frequency)
  - 数据清洗：去掉不合适的噪声文档或文档内垃圾数据
- 文本表示
  - 向量空间模型
- 降维技术
  - 特征选择(Feature Selection)
  - 特征重构(Re-parameterisation，如LSI)

文本表示

- 向量空间模型(Vector Space Model，VSM)
  - m个无序标引项t_i(特征)，可以采用词根/词/短语/其他等单位
  - n个训练文档
  - 每个文档d_j可以用标引项向量(每个a_{ij}是权重)来表示
    - $(a_{1j}, a_{2j}, \ldots, a_{mj})$
  - 通过向量的距离可以计算文档之间的相似度(分类的主要计算目标就是度量两篇文档之间的距离)

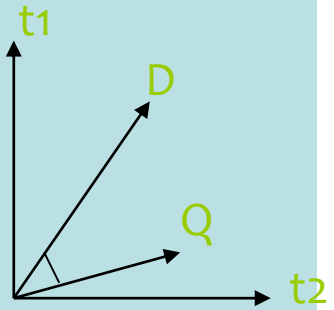
文档—标引项矩阵(Doc-Term Matrix)

$$A = \begin{array}{c} \\ t_1 \\ t_2 \\ \vdots \\ t_m \end{array} \begin{array}{c} \begin{array}{cccc} d_1 & d_2 & \cdots & d_n \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \end{array}$$

文档之间的相似度计算

$$\text{Dot}: \quad Sim(D,Q) = D \bullet Q = \sum_i (a_i \times b_i)$$

$$\text{Cosine}: \quad Sim(D,Q) = \frac{D \bullet Q}{\| D \| \times \| Q \|} = \frac{\sum_i (a_i \times b_i)}{\sqrt{\sum_i a_i^2 \times \sum_i b_i^2}}$$



$$\text{Dice}: \quad Sim(D,Q) = \frac{2 \times D \bullet Q}{\| D \|^2 + \| Q \|^2} = \frac{2\sum_i (a_i \times b_i)}{\sum_i a_i^2 + \sum_i b_i^2}$$

$$\text{Jaccard}: Sim(D,Q) = \frac{D \bullet Q}{\| D \|^2 + \| Q \|^2 - D \bullet Q} = \frac{\sum_i (a_i * b_i)}{\sum_i a_i^2 + \sum_i b_i^2 - \sum_i (a_i * b_i)}$$

Term的粒度

- Character，字：中
- Word，词：中国
- Phrase，短语：中国人民银行
- Concept，概念
  - 同义词：开心 高兴 兴奋
  - 相关词cluster，word cluster：葛非/顾俊
- N-gram，N元组：中国 国人 人民 民银 银行
- 某种规律性模式：比如某个window中出现的固定模式
- David Lewis等一致地认为：(英文分类中)使用优化合并后的 Words比较合适

中文分词

- 中文文本没有间隔，通常需要进行分词处理。
- 方法：
  - 基于词典的方法：
    - 正向、反向、双向
    - 最大匹配、最小匹配
  - 无词典的方法：
    - 转化为分类问题进行解决，如对每个字标出它是“词头”，“词中”还是“词尾”。或在每对字间标出“断”或“连”
  - 词典和统计相结合
    - 未定义词问题
    - 分词歧义问题

权重计算方法(1)

- (Term i在文档j中的)布尔权重
  - a_{ij}=1(TF_{ij}>0) or 0 (TF_{ij}=0)
- TFIDF型权重
  - *TF*: $a_{ij}=TF_{ij}$
  - *TF*IDF*: $a_{ij}=TF_{ij}*\log(N/DF_i)$
  - *TFC*: 对上面进行归一化

$$a_{ij}=\frac{TF_{ij}*\log(N/DF_i)}{\sqrt{\sum_k[TF_{kj}*\log(N/DF_k)]^2}}$$

  - *LTC*: 降低*TF*的作用

$$a_{ij}=\frac{\log(TF_{ij}+1.0)*\log(N/DF_i)}{\sqrt{\sum_k[\log(TF_{kj}+1.0)*\log(N/DF_k)]^2}}$$

权重计算方法(2)

基于熵概念的权重(Entropy weighting)*

- n_i是term i在整个文档集合中出现的总次数($\neq DF_i$)
- Entropy(i)称为term i的某种熵
  - 如果term i分布极度均匀：Entropy(i)等于-1
  - 只在一个文档中出现： Entropy(i)等于0

$$a_{ij} = \log(TF_{ij} + 1.0) * (1 + Entropy(i))$$

其中

$$Entropy(i) = \frac{1}{\log N} \sum_{j=1}^{N} [\frac{TF_{ij}}{n_i} \log(\frac{TF_{ij}}{n_i})]$$

特征选择Feature selection(1)

- 基于*DF*的选择方法 (DF Thresholding)
  - Term的*DF*小于某个阈值去掉(太少，没有代表性)
- 信息增益(Information Gain, *IG*)：该term为整个分类所能提供的信息量(不考虑任何特征的熵和考虑该特征后的熵的差值)

$$\mathrm{IG(t)} = \mathrm{Entropy}(S) - \mathrm{Expected\ Entropy}(S_t) = -\sum_{i=1}^{M} P(c_i)\log P(c_i)$$

$$-[P(t)(-\sum_{i=1}^{M} P(c_i \mid t)\log P(c_i \mid t)) + P(\bar{t})(-\sum_{i=1}^{M} P(c_i \mid \bar{t})\log P(c_i \mid \bar{t}))]$$

特征选择(2)

- term的某种熵：该值越大，说明分布越均匀，越有可能出现在较多的类别中；该值越小，说明分布越倾斜，词可能出现在较少的类别中

$$Entropy(t) = -\sum_i P(c_i \mid t) \log P(c_i \mid t)$$

- 相对熵(not 交叉熵)：也称为KL距离(Kullback-Leibler divergence)，反映了在出现了某个特定词的条件下的文本类别的概率分布和无任何条件下的文本类别的概率分布之间的距离，该值越大，词对文本类别分布的影响也大。

$$CE(t) = \sum_i P(c_i \mid t) \log \frac{P(c_i \mid t)}{P(c_i)}$$

特征选择(3)

- χ^2 统计量(念xi, chi)：度量两者(term和类别)独立性的缺乏程度，χ^2 越大，独立性越小，相关性越大($N=A+B+C+D$)

$$\chi^2(t,c)=\frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)}$$

	c	~c
t	A	B
~t	C	D

$$\chi^2_{AVG}(t)=\sum_{i=1}^{m}P(c_i)\chi^2(t,c_i)$$

$$\chi^2_{MAX}(t)=\max_{i=1}^{m}\{\chi^2(t,c_i)\}$$

- 互信息(Mutual Information，*MI*)：*MI*越大*t*和*c*共现程度越大

$$I(t,c)=\log\frac{P(t\wedge c)}{P(t)P(c)}=\log\frac{P(t|c)}{P(t)}=\log\frac{A\times N}{(A+C)(A+B)}$$

$$I_{AVG}(t)=\sum_{i=1}^{m}P(c_i)I(t,c_i)$$

$$I_{MAX}(t)=\max_{i=1}^{m}P(c_i)I(t,c_i)$$

特征选择(4)

- Robertson & Sparck Jones公式

$$RSJ(t,c_j)=\frac{c_j\text{中出现}t\text{的概率}}{\text{非}c_j\text{中出现}t\text{的概率}}=\log\frac{P(t|c_j)}{P(t|\bar{c}_j)}$$

$$TSV(t,c_j)=r*\log\frac{P(t|c_j)}{P(t|\bar{c}_j)},r\text{为出现}t\text{的}c_j\text{类文档个数}$$

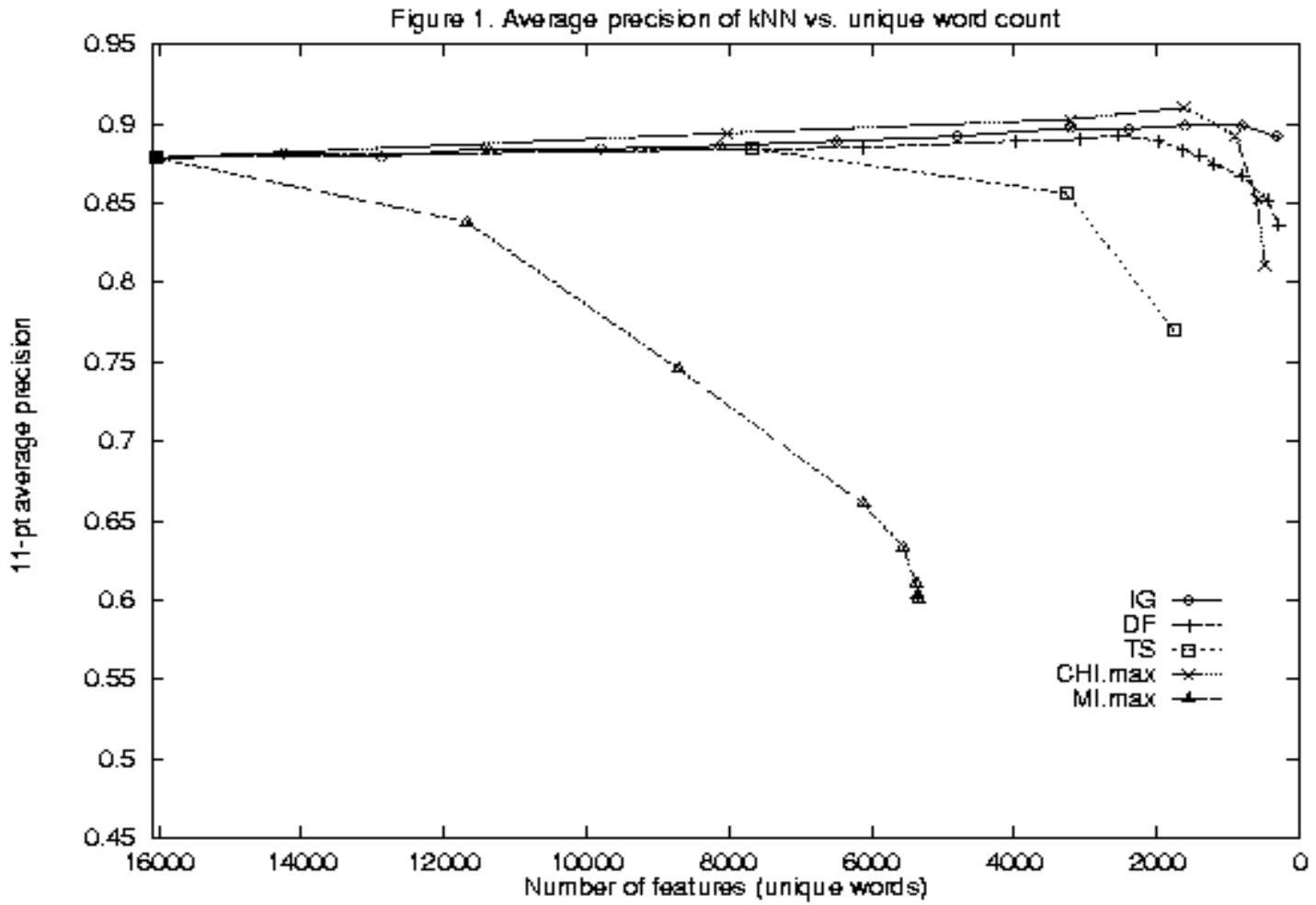
- 其他
  - Odds:

$$\frac{\log P(t|c_j)\log(1-P(t|\bar{c}_j))}{\log(1-P(t|c_j))\log P(t|\bar{c}_j)}$$

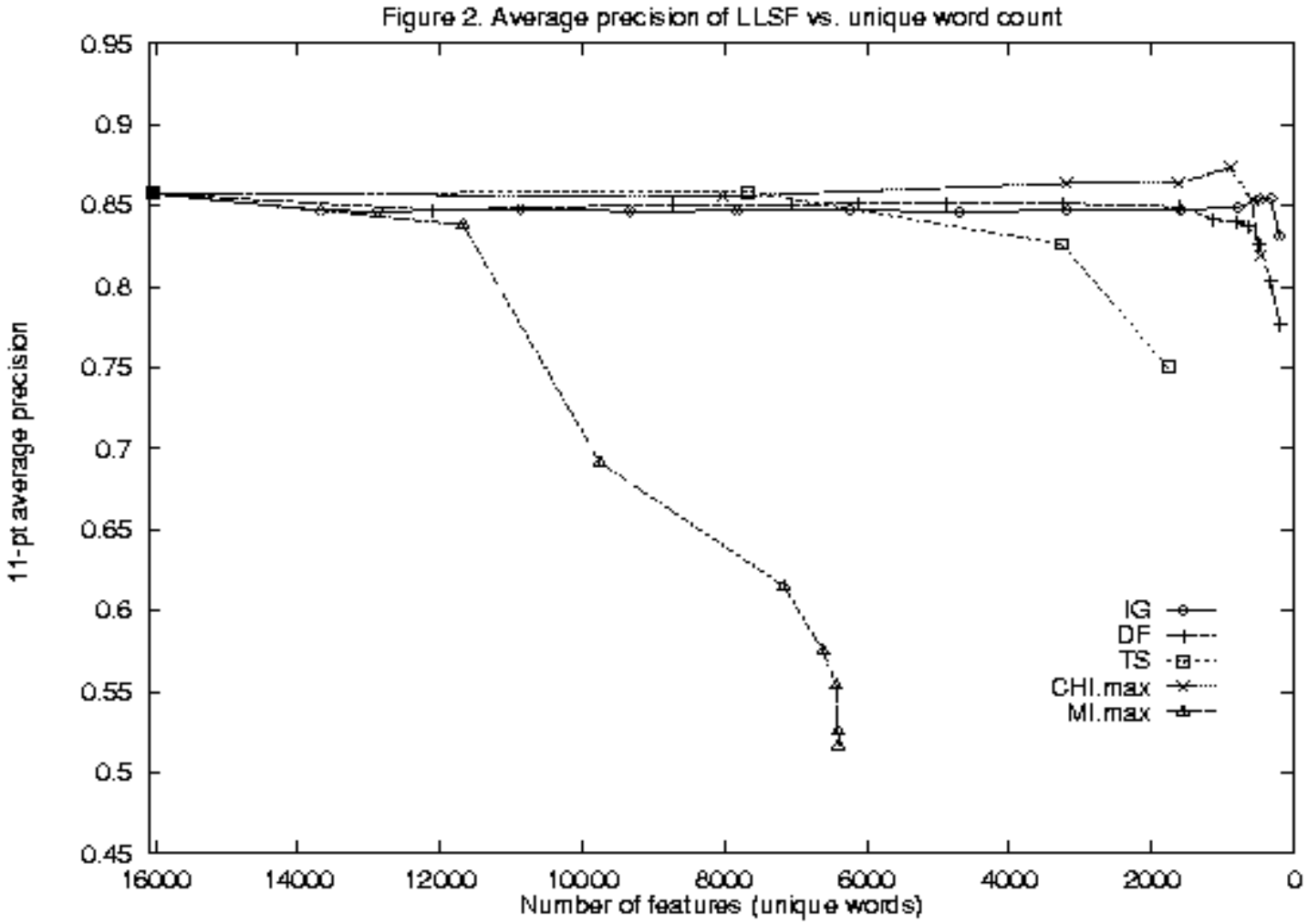
  - Term Strength(TS):

$$P(t\in y|t\in x),x,y\text{是相关的两篇文档}$$

特征选择方法的性能比较(1)



特征选择方法的性能比较(2)



特征选择方法的性能比较(3)

Yang Yi-ming 的实验结论

Method	DF	IG	CHI	MI	TS
favoring common terms	Y	Y	Y	N	Y/N
using categories	N	Y	Y	Y	N
using term absence	N	Y	Y	N	N
performance in kNN/LLSF	excellent	excellent	excellent	poor	ok

特征重构

- 特征重构的目的是将现有的特征空间映射到其他更合适的特征空间当中去，以便获得更好的特征表示。
- 隐性语义索引(Latent Semantic Index)是其中最有代表性的方法(参见第三章)。
- 另外，PCA(主成份分析)也可以用于特征重构。

自动文本分类方法

- 决策树方法Decision Tree
- Decision Rule Classifiers

规则方法

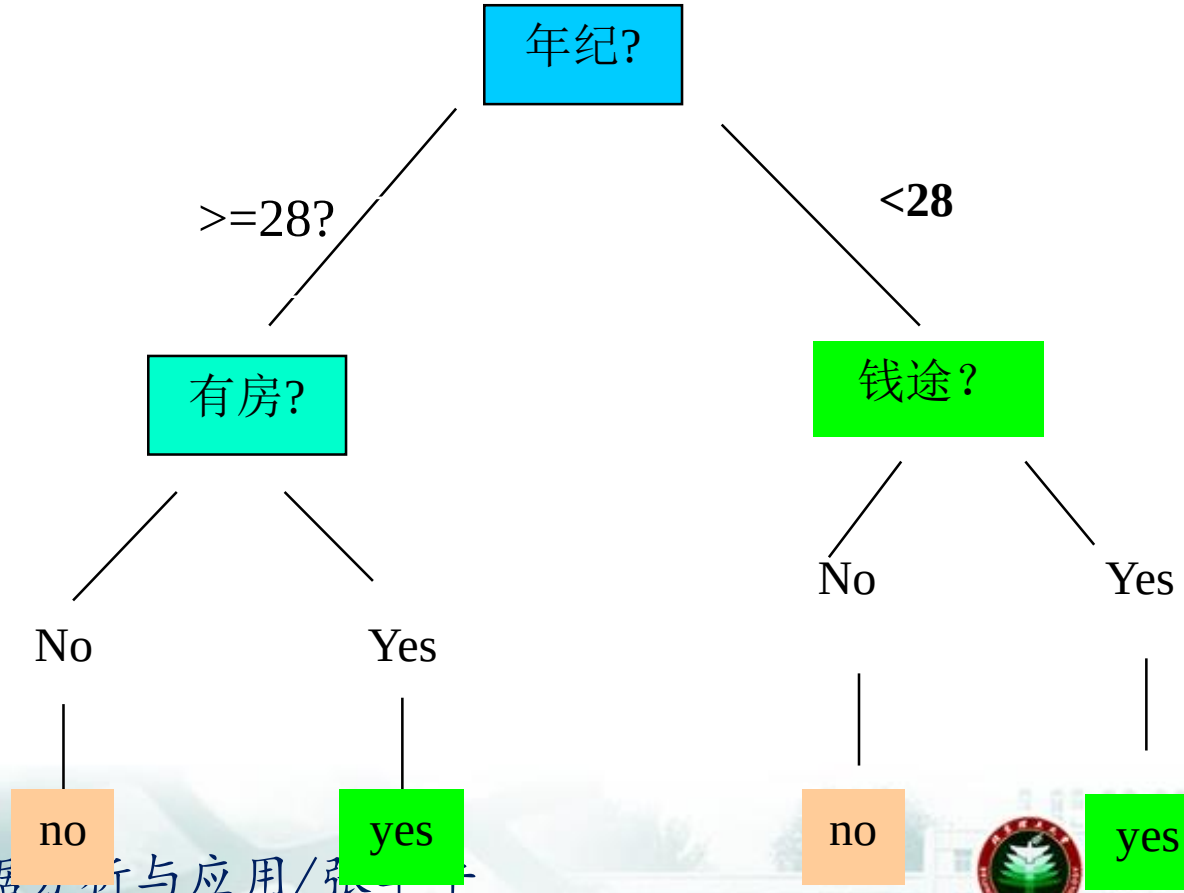
- 回归(Regression)方法
- Rocchio方法
- kNN方法
- Naïve Bayes
- Online Linear Classifiers
- 多重神经网络方法Neural Networks
- 支持向量机SVM
- 基于投票的方法(Voting methods)

统计方法

决策树(decision tree)方法(1)

- 思想：将文本的特征进行优先度排序，并将每次的特征作为判定条件(子树的根节点)进行扩展，最后生成一颗树。
- 训练：
  - 构造决策树：使用某个函数(如IG)来判断特征优先级
    - CART
    - C4.5 (由ID3发展而来)
    - CHAID
  - 决策树的剪枝(pruning)
- 分类：按照决策树的条件进行判定。

决策树方法(2)—例子(某女生见面标准)



决策树方法(3)

- 决策树方法是一种规则方法，可以生成可以理解的规则(if… then…)
- 决策树方法会遇到过学习问题(Overfitting)：训练集合的样例都满足得较好，一推广则性能马上下降。
- 效果一般，但是有时很好。

其他决策规则学习方法
Decision Rule Learning

通过学习得到类似的如下规则

wheat & form → WHEAT
wheat & commodity → WHEAT
bushels & export → WHEAT
wheat & agriculture → WHEAT
wheat & tonnes → WHEAT
wheat & winter & ~soft → WHEAT

做法：

(粗糙集)RoughSet

逻辑表达式(AQ11算法)

回归方法(1)

- 回归：用一条直线(线性回归)或者曲线去拟合已有的例子。
- LLSF(Linear Least Square Fit) 方法：
  - |FA-B|, A是所有例子构成的矩阵，F是要求的权重矩阵，B是布尔矩阵，1表示属于该类，0表示不属于。
  - F求出以后，对新来的文本D，计算结果，哪个最大属于哪个类。

回归方法(2)-LLSF

举例：2个类别c1,c2，4篇文档d1,d2,d3,d4分别属于c2,c1,c1,c2，3个特征t1,t2,t3，利用LLSF计算如下：

$$\underset{}{} \quad \underset{F}{} \quad \underset{A}{} \quad \underset{B}{}$$

$$\begin{matrix} c_1 \\ c_2 \end{bmatrix} \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{bmatrix} - \begin{matrix} c_1 \\ c_2 \end{bmatrix} \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

使|FA-B|最小，可以解得F矩阵。对一篇新文档d=[a1,a2,a3]^T，Fd会得到一个2行一列的矩阵(实际是个向量)，哪个分量大则属于哪类。显然，LLSF可以直接处理多类问题，也能处理兼类问题。

回归方法(3)-LLSF

- 训练：计算类别权重矩阵F的过程，比较耗时。
- 分类：类别权重矩阵和文档向量相乘。
- Yang Yiming通过实验证实LLSF的效果和kNN、SVM类似。

Rocchio方法(1)

- 训练时：每个类别由一个类中心向量来表示；
  分类时：对于某个文本，它和哪个类中心最近，则认为它属于那类。
  - Rocchio公式

$$w'_{jc} = \alpha w_{jc} + \beta \frac{\sum_{i \in C} x_{ij}}{n_C} - \gamma \frac{\sum_{i \notin C} x_{ij}}{n - n_C}$$

- w'_{jc}：类C中心向量的权重
- n_C：训练样本中正例个数
- x_{ij}：文档向量的权重

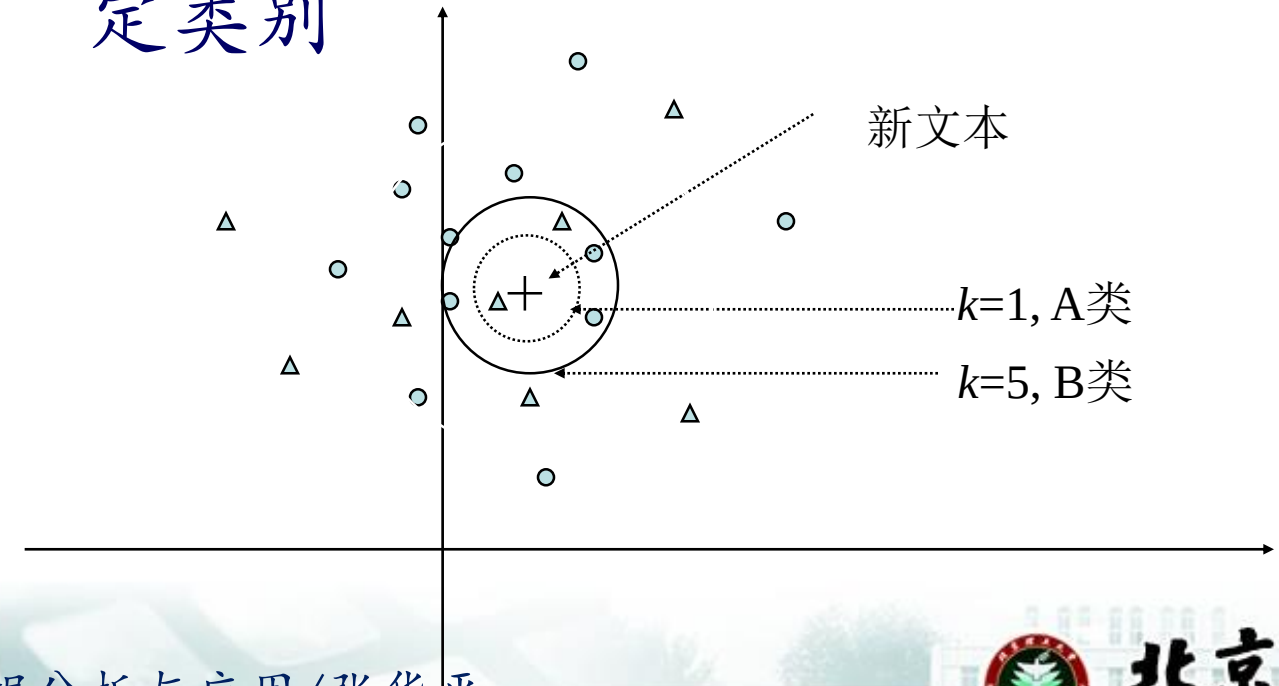
$$CSV_c(d_i) = \mathrm{w}_c \cdot \mathrm{x}_i = \frac{\sum w_{cj} \cdot x_{ij}}{\sqrt{\sum w_{cj}^2}\sqrt{\sum x_{ij}^2}}$$

Rocchio方法(2)

- 训练和分类都较快，易于实现，效果中等，可以作为Base line。
- 实现时，只考虑正例，则称为类中心向量法。有时正例过少，则将和这些正例相近的例子(并不一定是真正的反例，可以是其他未标注文本)加入作为“伪正例”进行计算。
- 在考虑反例时，也有很多做法，比如只考虑一些代表性反例。

kNN方法(1)

- 没有训练过程；分类时：一篇文本根据和它最近的k篇训练文本的归属来确定类别



kNN方法(2)

- 实现时可以有多种方式:
  - 方式1：只考虑top *k*文档的类别多少，选择出现最多的类别作为最终选择。问题是倾向于大类。
  - 方式2：选出*k*篇文档后，将相似度融入进行加权计算。
- kNN也称为Lazy learning 或 Case-based learning 方法。没有训练，分类时计算量较大。分类效果较好。

朴素贝叶斯(Naïve Bayes)方法(1)

- 思想：对于文本d_i，求条件概率$P(c_j|d_i)$，条件概率最大的那个类别作为最终选择类别。计算时，引入Term独立性假设，故称为Naïve Bayes方法。
- 训练：计算概率分布参数；分类：进行概率计算。
- Naïve Bayes是较快的一种分类方法，效果也较好。理论上错误率最低。

朴素贝叶斯方法(2)

Bayes公式 $$P(c_j \mid d_i) = \frac{P(d_i \mid c_j)P(c_j)}{P(d_i)} \propto P(d_i \mid c_j)P(c_j)$$

$$P(d_i \mid c_j) = \prod_{k=1}^{r} P(w_{ik} \mid c_j), \quad \text{独立性假设}$$

概率参数计算

$$P(c_j) = \frac{c_j\text{的文档个数}}{\text{总文档个数}} = \frac{N(c_j)}{\sum_k N(c_k)} \approx \frac{1+N(c_j)}{|c|+\sum_{k=1} N(c_k)}$$

$$P(w_i \mid c_j) = \frac{w_i\text{在}c_j\text{类别文档中出现的次数}}{\text{在}c_j\text{类所有文档中出现的词的次数}} \approx \frac{1+N_{ij}}{\text{不同词个数}+\sum_k N_{kj}}$$

Online Linear Classifiers

2类线性可分问题：

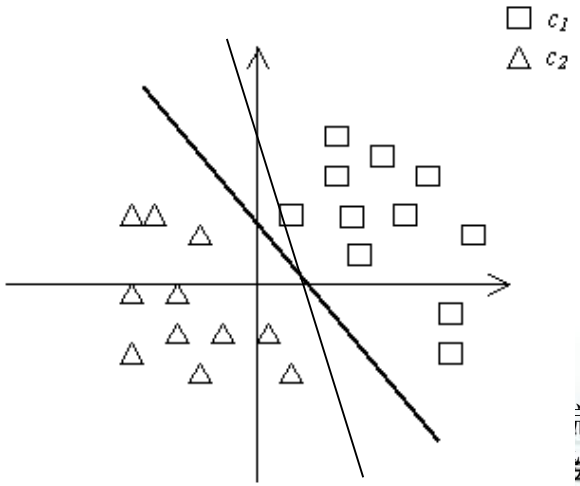
求直线$\mathbf{w}^T\mathbf{x}+b=0$，$\mathbf{w}$和$\mathbf{x}$是向量，对于下图$\mathbf{x}$就是$<x_1,x_2>$
使得在直线上面所有样例，满足 $\mathbf{w}^T\mathbf{x}+b > 0$，反之，
$\mathbf{w}^T\mathbf{x}+b < 0$

通常的解法：对$\mathbf{w}$,b赋初始值，然后通过某种方式迭代循环至收敛。

感知机Perceptron

Winnow

Widrow-Hoff



The Widrow-Hoff Classifier

Online Learning

Learning Rate

$$w_{cj}^{(t+1)} = w_{cj}^{(t)} - 2\eta(\mathrm{w_c} \cdot \mathrm{x_i} - y_i)x_{ij}$$

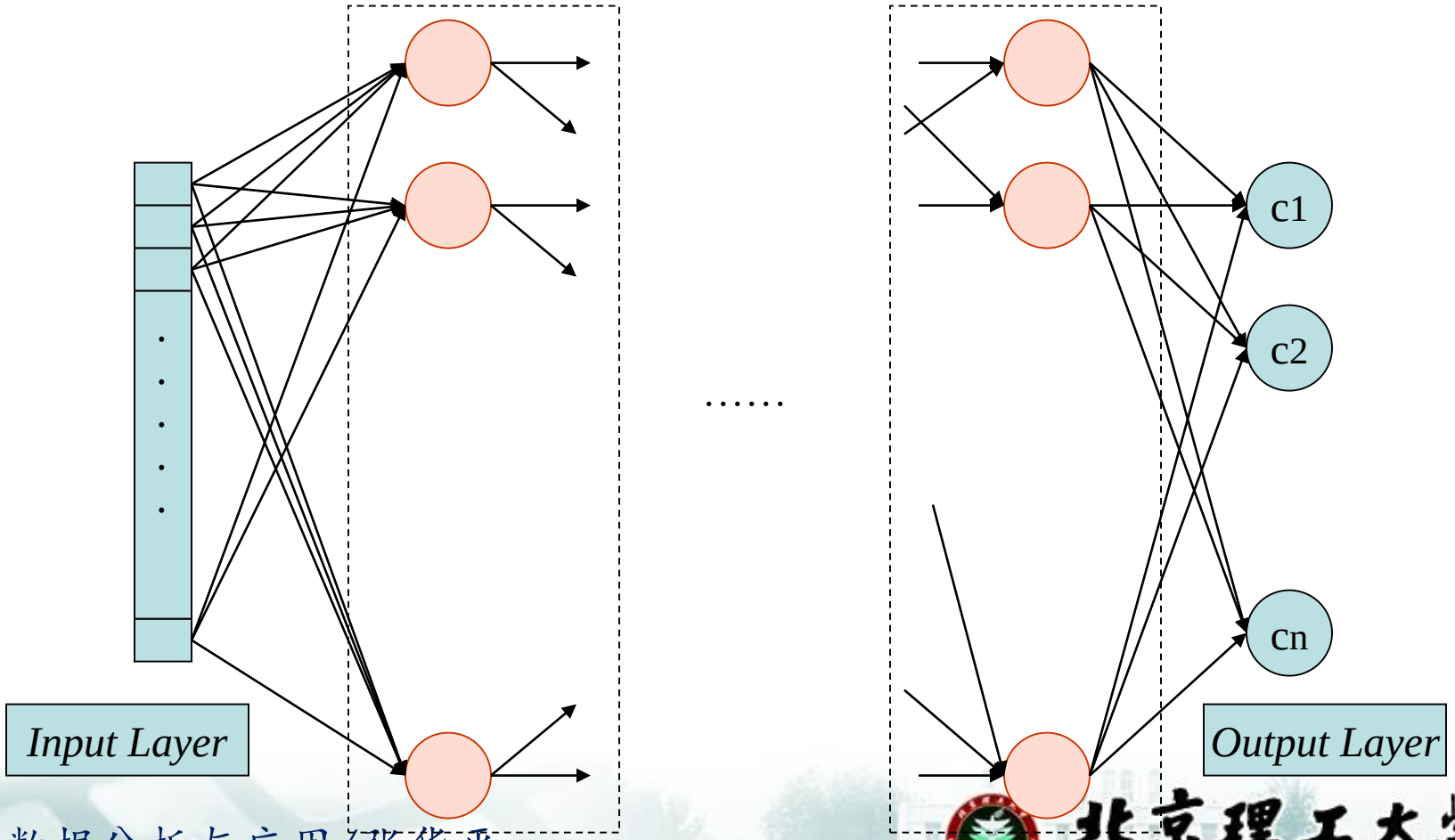
Target Value (0 or 1)

类c向量的第j个分量

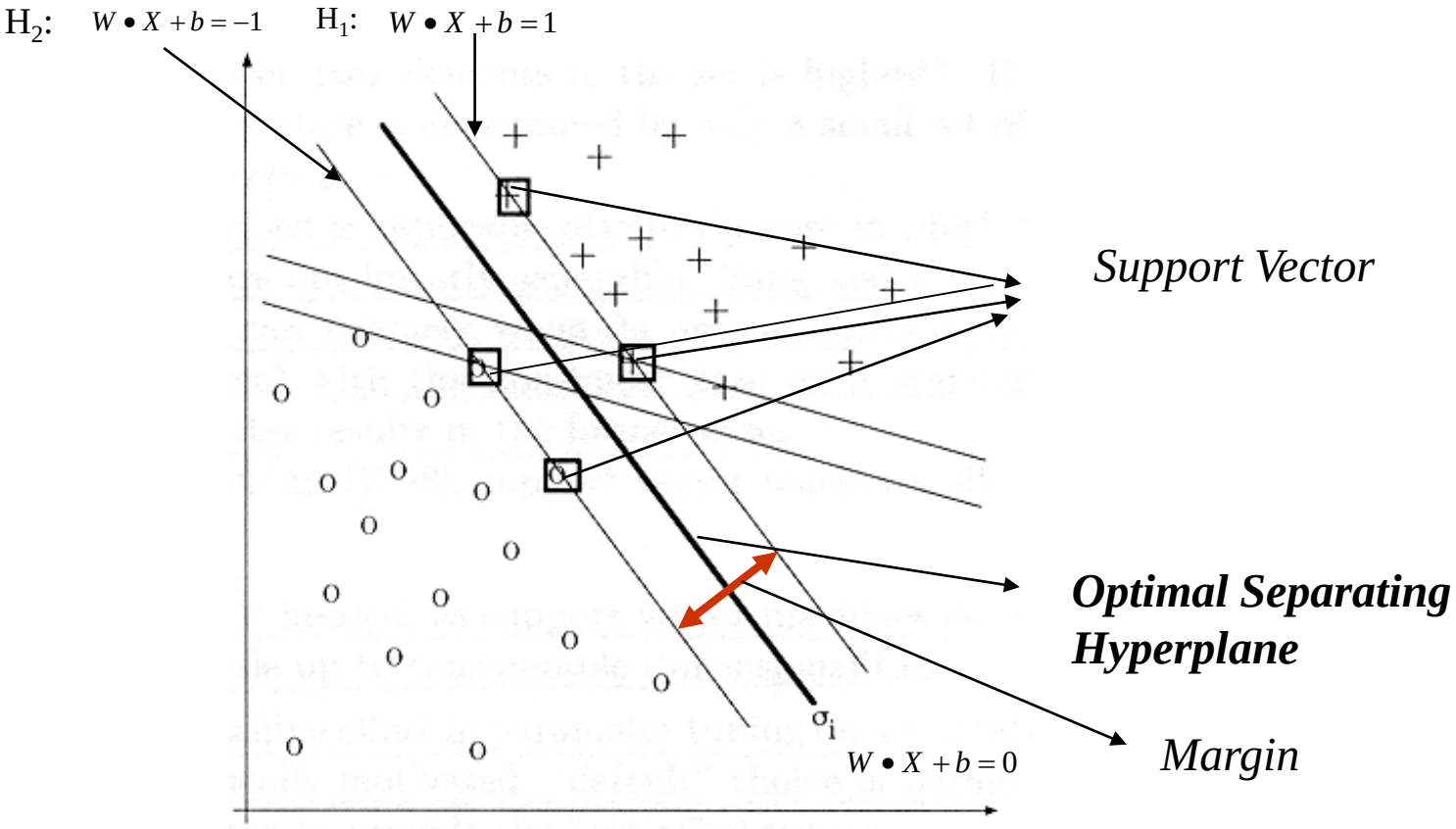
x_i的第j个分量

$$CSV_c(d_i) = \mathrm{w_c} \cdot \mathrm{x}_i = \frac{\sum w_{cj} \cdot x_{ij}}{\sqrt{\sum w_{cj}^{\ 2}}\sqrt{\sum x_{ij}^{\ 2}}}$$

多重神经网络(Neural Network)



支持向量机(1)
Support Vector Machine



线性可分情况下，不仅要区分开，而且要使得区分间隔Margin最大。

如上图的训练样本,在线性可分的情况下,存在多个超平面(Hyperplane) (如 : H1,H2….)使得这两类被无误差的完全分开。这个超平面被定义为:

$$W \bullet X + b = 0$$

其中$W \bullet X$是内积（ dot product ），b是标量。

Optimal Separating Hyperplane （最优超平面）是指两类的分类空隙(Margin)最大，即每类距离超平面最近的样本到超平面的距离之和最大。距离这个最优超平面最近的样本被称为支持向量（Support Vector）。

$$\text{Margin} = \frac{2}{\|w\|} \qquad \text{.....(1)}$$

H1平面： $W \bullet X_1 + b \geq 1$

H2平面： $W \bullet X_2 + b \leq -1$

$$y_i[(W \bullet X_i) + b] - 1 \geq 0 \qquad \text{.....(2)}$$

求解最优超平面就相当于，在(2)的约束条件下,求(1)的最大值

Minimum: $$\phi(W)=\frac{1}{2}\|W\|^2=\frac{1}{2}(W\bullet W)$$

Subject to: $$y_i[(W\bullet X_i)+b]-1\geq 0$$

- 上述二次优化问题，采用Lagrange方法求解，可得

$$f(X)=\operatorname{sgn}\left(\sum_{i=1}^{n}\alpha_i^{*}y_iX\bullet X_i+b^{*}\right)$$

支持向量(Support Vector)

非线性可分情况下的处理(方法1)

- 广义最优分类面方法：在线性不可分的情况下，就是某些训练样本不能满足式(2)的条件，因此可以在条件中增加一个松弛项ζ(也称引入Soft Margin，软边界)，约束条件变成：

$$y_i[(W \bullet X_i)+b]-1+\varsigma_i \geq 0$$

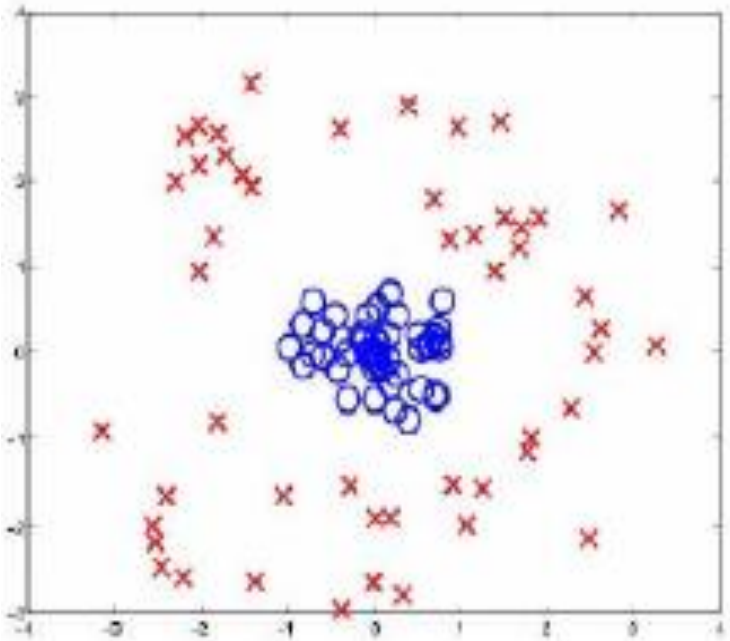
此时的目标函数是求下式的最小值:

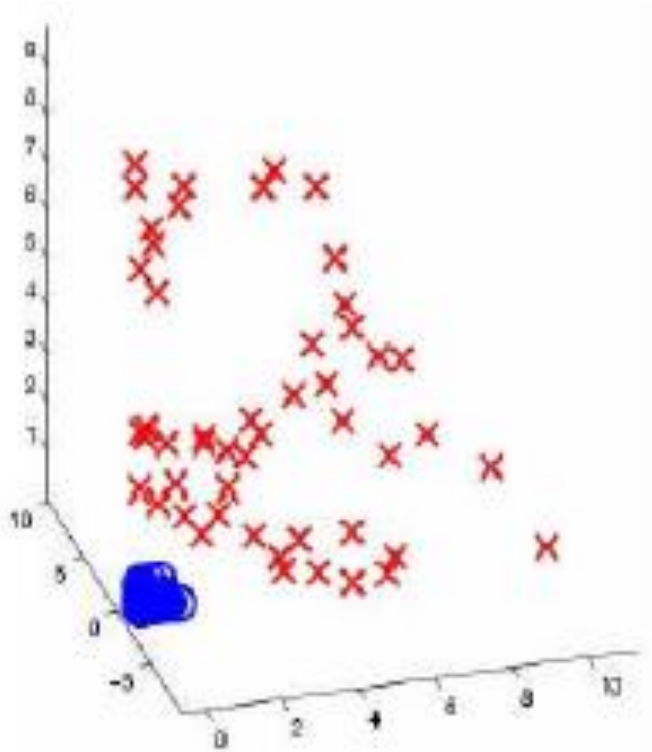
期望风险　　经验风险

$$\Phi(W,\varsigma_i)=\frac{1}{2}(W\bullet W)+C\left(\sum_{i=1}^{n}\varsigma_i\right)$$

这个二次优化，同样可以应用Lagrange方法求解

非线性可分情况下的处理(方法2)





变换到高维空间的支持向量机

- 采用如下的内积函数(核函数):

$$K(X, X_i) = [(X \bullet X_i) + 1]^q$$

$$K(X, Xi) = \exp\left\{-\frac{|X - Xi|^2}{\sigma^2}\right\}$$

$$K(X, X_i) = \tanh(\upsilon(X \bullet X_i) + c)$$

判别函数成为：

$$f(X)=\mathrm{sgn}\left(\sum_{i=1}^{n}\alpha_i^{*}y_iK(X,X_i)+b^{*}\right)$$

支持向量机(2)

- SVM训练相对较慢，分类速度一般。但是分类效果较好。
- 在面对非线性可分情况时，可以引入松弛变量进行处理或者通过空间变换到另一个线性可分空间进行处理。
- SVM有很多实现工具，SMO/SVM light/SVM torch/LibSVM等等。

基于投票的方法

- Bagging方法
  - 训练R个分类器f_i，分类器之间其他相同就是参数不同。其中f_i是通过从训练集合中(N篇文档)随机取(取后放回)N次文档构成的训练集合训练得到的。
  - 对于新文档d，用这R个分类器去分类，得到的最多的那个类别作为d的最终类别
- Boosting方法
  - 类似Bagging方法，但是训练是串行进行的，第k个分类器训练时关注对前k-1分类器中错分的文档，即不是随机取，而是加大取这些文档的概率
  - AdaBoost
  - AdaBoost MH

分类方法的比较

- 目前的实验表明，SVM/kNN/Adaboost方法一般较好。但是在具体的语料集上，一些其他的方法有时也能表现很好。
- 在实现时，特征选择、分类体系、分类算法都是很重要的因素。

文本聚类定义

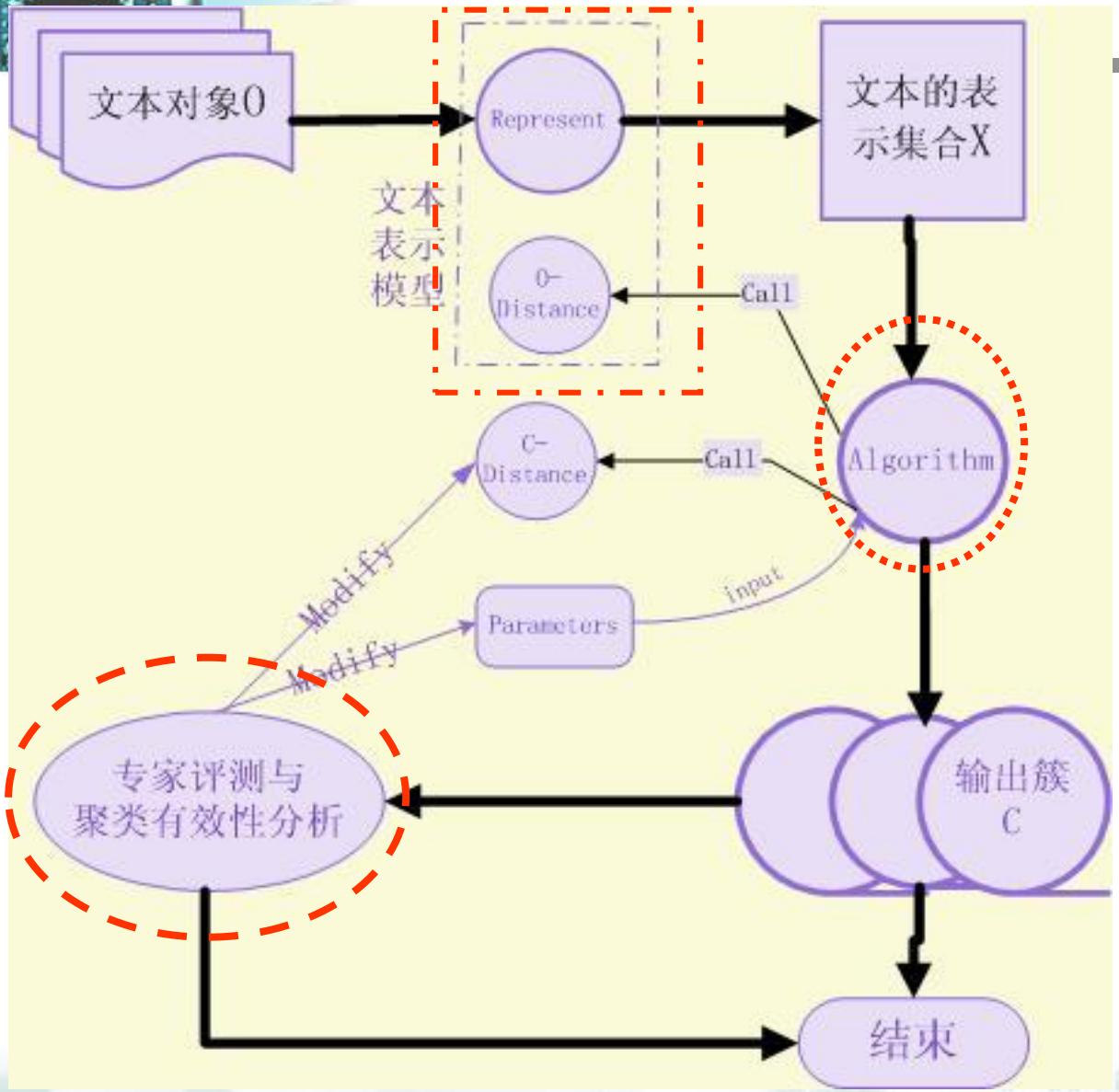
- 聚类是一个无导的学习过程，它是指根据样本之间的某种距离在无监督条件下的聚簇过程。
- 利用聚类方法可以把大量的文档划分成用户可迅速理解的簇(cluster)，从而使用户可以更快地把握大量文档中所包含的内容，加快分析速度并辅助决策。
- 大规模文档聚类是解决海量文本中数据理解和信息挖掘的有效解决手段之一。

文本聚类的应用

- TDT(Topic Detection and Tracking)中主题事件的检测。
  - 将文档进行聚类，从聚出的类中发现新的热点主题
- 检索结果的聚类显示。
  - 检索结果聚类，以便用户浏览
- 大规模文档的组织和呈现


Vivisimo - Clustered search on 王刚 - Microsoft Internet Explorer
文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
后退 搜索 收藏夹 媒体
地址(D) http://vivisimo.com/search?tb=homepage&query=%E7%8E%8B%E5%88%9A&v%3Asourc
转到 链接 上网助手 暂停拦截 清理 修复
Vivísimo
company | products | solutions | customers | demos | press
王刚
the Web
Search
Advanced Search
Help
NEW try your query at Clusty.com
Clustered Results
王刚 (131)
张国立 (9)
英格力士 (5)
Sina.com (5)
王刚,男,汉族,1942年10月生,吉林扶余人 (5)
王刚个人档案,最佳男配角 (3)
宰相刘罗锅,铁齿铜牙纪晓岚 (3)
康熙来了,还有张铁林和王刚的加盟,而张国立则缺阵"铁三角"组合,由孙兴来顶替他的位置 (3)
独臂刀,武侠天地 - 司马紫烟作品 (3)
Spaces.msn (2)
人物信息 (2)
人们都会将他与奸猾,尽管王刚试图改变自己的银幕形象,曾扮演过其他角色,但让观众记忆犹新的还是那个狡猾的王刚 (3)
醒狮 第一章 (3)
版权所有 (3)
Top 131 results of at least 106,958 retrieved for the query 王刚 (Details)
1. 王刚 影音娱乐 新浪网 [new window] [frame] [cache] [preview] [clusters]
... 获北京电视艺术春燕奖 获第十四届中国电视金鹰奖“最佳男配角”。 >> 王刚作客新浪访谈实录(图) >> 王刚个人档案 王刚张国立犯下“欺君”之罪 “皇阿玛 ...
ent.sina.com.cn/s/m/f/wangg.html - MSN 1, MSN Search 1
2. 王刚活动报道集 [new window] [frame] [preview] [clusters]
王刚活动报道集 人民网版权所有,未经书面授权禁止使用 Copyright © 2002 by www.people.com.cn. all rights reserved ... 人民网 >> 时政 >> 时政专题 >> 王刚活动报道集 王刚同志简历 王刚活动报道集 国内活动 出访 ...
www.people.com.cn/GB/shizheng/252/9933 - MSN 2
3. 王刚任播音主持委员会副会长 狠批《康熙来了》 影 ... [new window] [frame] [cache] [preview] [clusters]
... 王刚任播音主持委员会副会长 狠批《康熙来了》 http://ent.sina.com.cn 2005年06月25日10:09 北京青年报 会议现场 (图片来源：新浪娱乐) 点击此处查看全部娱乐 ...
ent.sina.com.cn/s/m/2005-06-25/1009762313.html - MSN Search 2
4. 作家王刚作客新浪聊《英格力士》实录 读书频道 新浪网 [new window] [frame] [preview] [clusters]
作家王刚作客新浪聊《英格力士》实录 王刚虽然是实力派作家,但直到《英格力士》才引起众媒体的聚焦
book.sina.com.cn/author/subject/2005-01-05/3/148574.shtml - MSN 3
5. 张铁林：王刚张国立欺负我 “卡通皇帝”一切以 ... [new window] [frame] [cache] [preview] [clusters]
... 张铁林：王刚张国立欺负我 “卡通皇帝”一切以授课为重 05年10月27日 很多观众最初都是通过《大桥下面》这部片子认识张铁林的，不久他就去了 ...
www.chinapressusa.com/yule/200510270195.htm - MSN Search 3
6. 交友档案 - 友缘人 - 雅虎中国 [new window] [frame] [preview] [clusters]

文本聚类的流程


文本对象O
Represent
文本的表示集合X
文本表示模型
O-Distance
Call
C-Distance
Call
Algorithm
Modify
Modify
Parameters
input
专家评测与聚类有效性分析
输出簇C
结束

聚类算法(1)

- 聚类算法多种多样:
  - (1) 层次方法（Hierarchical Methods）
    - i. 凝聚算法（Agglomerative Algorithms）
    - ii. 分裂算法（Divisive Algorithms）
  - (2) 划分方法(Partitioning Methods)
    - i. Relocation Algorithms
    - ii. 概率聚类(Probabilistic Clustering)
    - iii. K-中心点算法(K-medoids Methods)
    - iv. K-平均算法(K-means Methods)
    - v. 基于密度的算法（Density-Based Algorithms）
      1. Density-Based Connectivity Clustering
      2 Density Functions Clustering

聚类算法(2)

(3) 基于网格的方法(Grid-Based Methods)

(4) Methods Based on Co-Occurrence of Categorical Data

(5) Constraint-Based Clustering

(6)Clustering Algorithms Used in Machine Learning

i. Gradient Descent and Artificial Neural Networkds

ii.Evolutionary Methods

(7) Scalable Clustering Algorithms

(8) Algorithms For High Dimensional Data

i. Subspace Clustering

ii.Projection Techniques

iii.Co-Clustering Techniques

凝聚式层次聚类(HAC)

- 算法流程:
  - Step1: 将所有的点各自单独形成一个簇;
  - Step2: 从现有所有的簇中选择最近(或者最相似的两个簇)，进行合并;
  - Step3: 如果只剩下一个簇或者达到终止条件(比如达到需要的簇的数目),聚类结束,否则返回Step2.

k-Means聚类分析

算法流程:

- Step1: 初始化*k*个簇中心;
- Step2: 对于每个文档向量，计算该文档向量与k个类中心的距离，选择距离最小(相似度最大)的簇将该文档分入该簇;
- Step3: 重新计算*k*个簇的中心，中心为该簇内所有点的算术平均。
- Step4: 如果簇变化不大或者满足某种退出条件(达到最大迭代次数、满足某种目标函数等)，那么结束聚类，否则返回Step2。

BiSecting k-Means聚类 (BiSect)

算法流程:

- Step1: 将所有的点形成一个簇;
- Step2: 从现有所有的簇中选择包含文档数最大的簇进行拆分，用*k*-Means算法(*k*=2)将该簇分成2个簇;
- Step3: 如果达到了需要的簇的数目则结束。

最近邻聚类(Nearest Neighbour)

- 算法流程:
  - Step1：随机选择一个样本，以该样本为中心建立一个新簇;
  - Step2：取下一个要分析的对象，如果没有对象需要聚类，那么聚类结束;
  - Step3：计算当前对象与当前所有簇的相似度，得到相似度最大的簇及对应的相似度d,如果d>阈值T,那么将该对象分配给选中的簇，更新簇的中心；否则以该对象为中心新建一个簇;
  - Step4：返回Step2

MaxDist算法

- 算法流程:
  - Step1：从Ds中任取一个样本，例如D1，以D1作为簇中心新建一个簇
  - Step2：在Ds中找一个与D1最远的样本并以之为中心新建一个簇，从而形成两个簇，记录该最远距离为max,同时算出阈值（可以为max的p倍，1/2<=p<1);
  - Step3: 对于剩下的点顺序扫描,计算该点与所有的簇的距离的最小值;
  - Step4: 如果最小距离大于阈值并且未达到需要的类数,则以该点新建一个簇;返回Step3，否则如果没有点了或者达到需要的类数,结束聚类;
  - Step5:返回Step3

纲要

大数据
文本挖掘

I 文本挖掘基础知识综述

II NLPIR汉语分词与关键词提取

III 文本分类与聚类

IV NLPIR大数据挖掘平台与应用

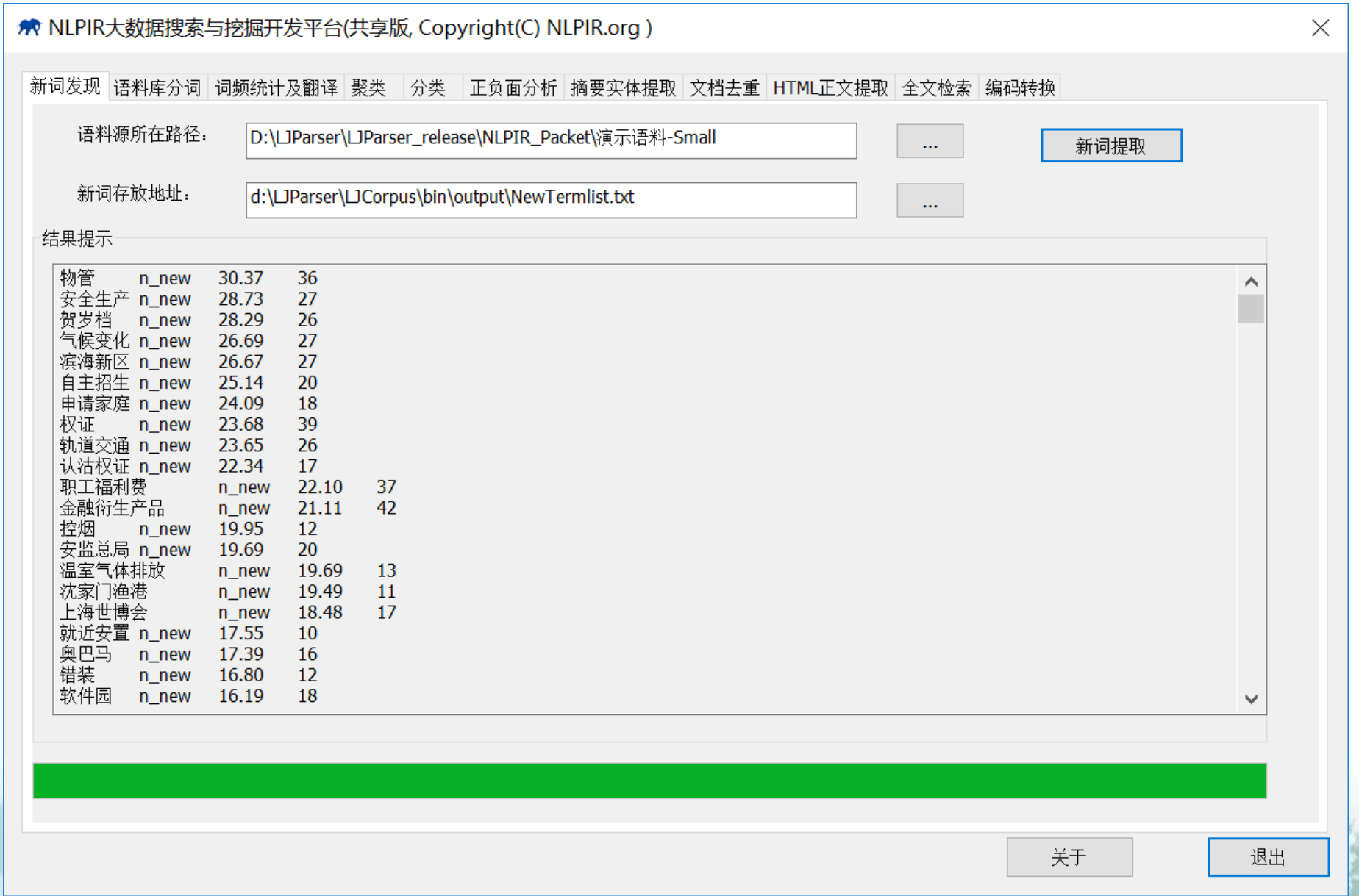
NLPIR大数据搜索与挖掘技术开发平台

- NLPIR网络搜索与挖掘共享开发平台，针对语言信息内容处理的全技术链条的共享开发平台。15年专业研究与工程积累，提供应用软件及各平台下的二次开发包，非商用永久免费。www.nlpir.org下载。

NLPIR 自然语言处理与信息检索共享平台
Natural Language Processing & Information Retrieval Sharing Platform

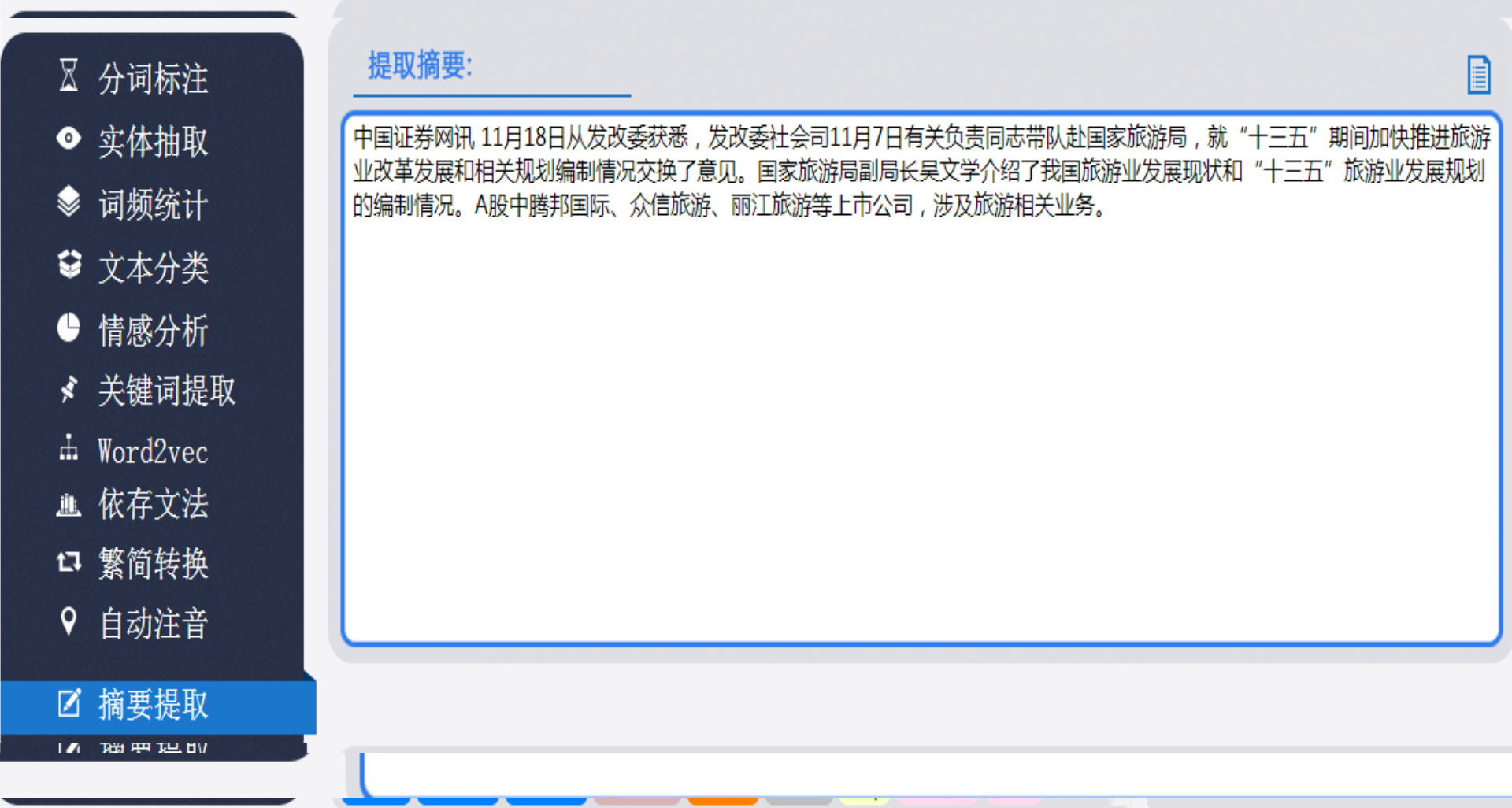
- 核心功能包括：
  - 搜索类：全文精准检索；
  - 语言类：新词发现，分词标注，统计分析与术语翻译；关键词提取；
  - 文档类：文本聚类及热点分析；分类过滤；自动摘要；文档去重；情感分析

NLPIR大数据搜索与挖掘技术开发平台



NLPIR大数据语义分析技术的在线演示

网址：http://ictclas.nlpir.org/nlpir/



产品下载试用

网址：https://github.com/NLPIR-team/NLPIR

..		
Classify	update Linux 64 bit NLPIR and JZSearch (CentOS)	3 months ago
Cluster	update sentiment and SentimentAnalysis	2 months ago
DeepClassifier	add IOS support, can be used in Macbook and iPhon3	14 days ago
DocExtractor	add IOS support, can be used in Macbook and iPhon3	14 days ago
HTMLPaser	add some archive	6 months ago
JZSearch	update JZSearch in 2015/11/5	13 days ago
JZSearchclient	update JZSearch tools	14 days ago
KeyExtract	update some demos	2 months ago
NLPIR-ICTCLAS	update some demos	2 months ago
RedupRemover	update sample projects	3 months ago
SentimentAnalysis	update sentiment and SentimentAnalysis	2 months ago
SentimentNew	Update SentimentNew for both WIndows and Linux	2 months ago
Summary	update API, fix bug with NLPIR	2 months ago

NLPIR大数据语义分析技术的在线演示-支持所有平台

NLPIR SDK

NLPIR SDK存放了13种组件包:

- Classify规则组件
- Cluster聚类组件
- DeepClassifier训练分类组件
- DocExtractor实体抽取组件
- HTMLPaser网站正文提取组件
- NLPIR-ICTCLAS分词组件
- JZsearch精准搜索组件
- JZSearch精准搜索客户端组件
- KeyExtract关键词提取组件
- RedupRemover文档去重组件
- Sentiment情感组件
- SentimentAnalysis情感分析组件
- Summary摘要组件

支持的开发语言

C语言
C++语言
C#语言
JAVA语言
等
几乎囊括了市面所有主流的编程语言

支持的操作系统

Windows 32位/64位操作系统
Linux32位/64位操作系统
Android操作系统
IOS操作系统
国产红旗等

感谢关注聆听！

张华平
Email: kevinzhang@bit.edu.cn
微博：@ICTCLAS张华平博士
实验室官网：
http://www.nlpir.org

大数据千人会