北京郵電大學
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

计算机应用编程 实验四

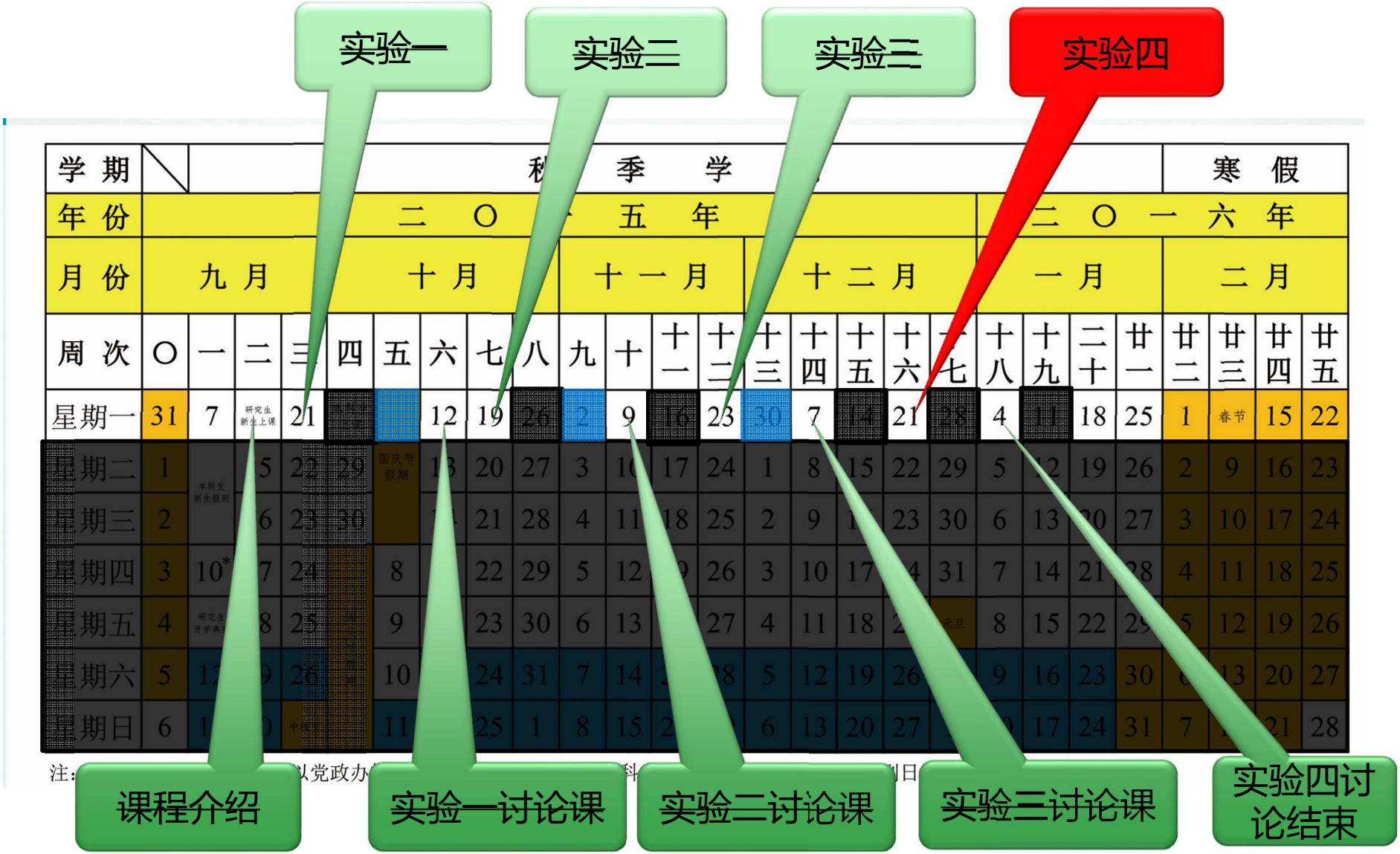
找你邻居家的妹

熊永平@网络技术研究院
ypxiong@bupt.edu.cn

周一10:10-12:00@3-134

2015.12.20

课程表


实验一
实验二
实验三
实验四
学 期
秋 季 学
寒 假
年 份
二 〇 一 五 年
二 〇 一 六 年
月 份
九 月
十 月
十 一 月
十 二 月
一 月
二 月
周 次
星期一
星期二
星期三
星期四
星期五
星期六
星期日
课程介绍
实验一讨论课
实验二讨论课
实验三讨论课
实验四讨论结束

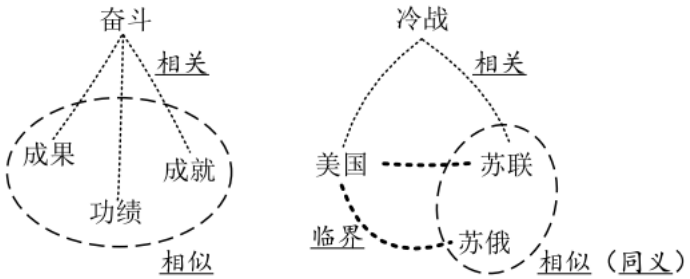
实验三：发现相关词

实验目标

- 目标
  - 设计一个分析程序，实现
  - 对一个大规模文本语料库分词
  - 计算词之间的距离
  - 寻找最相关的前20个词对
- 编程技能
  - Java语言练习
  - NLP
  - 统计距离
  - 概率模型
  - 数据挖掘

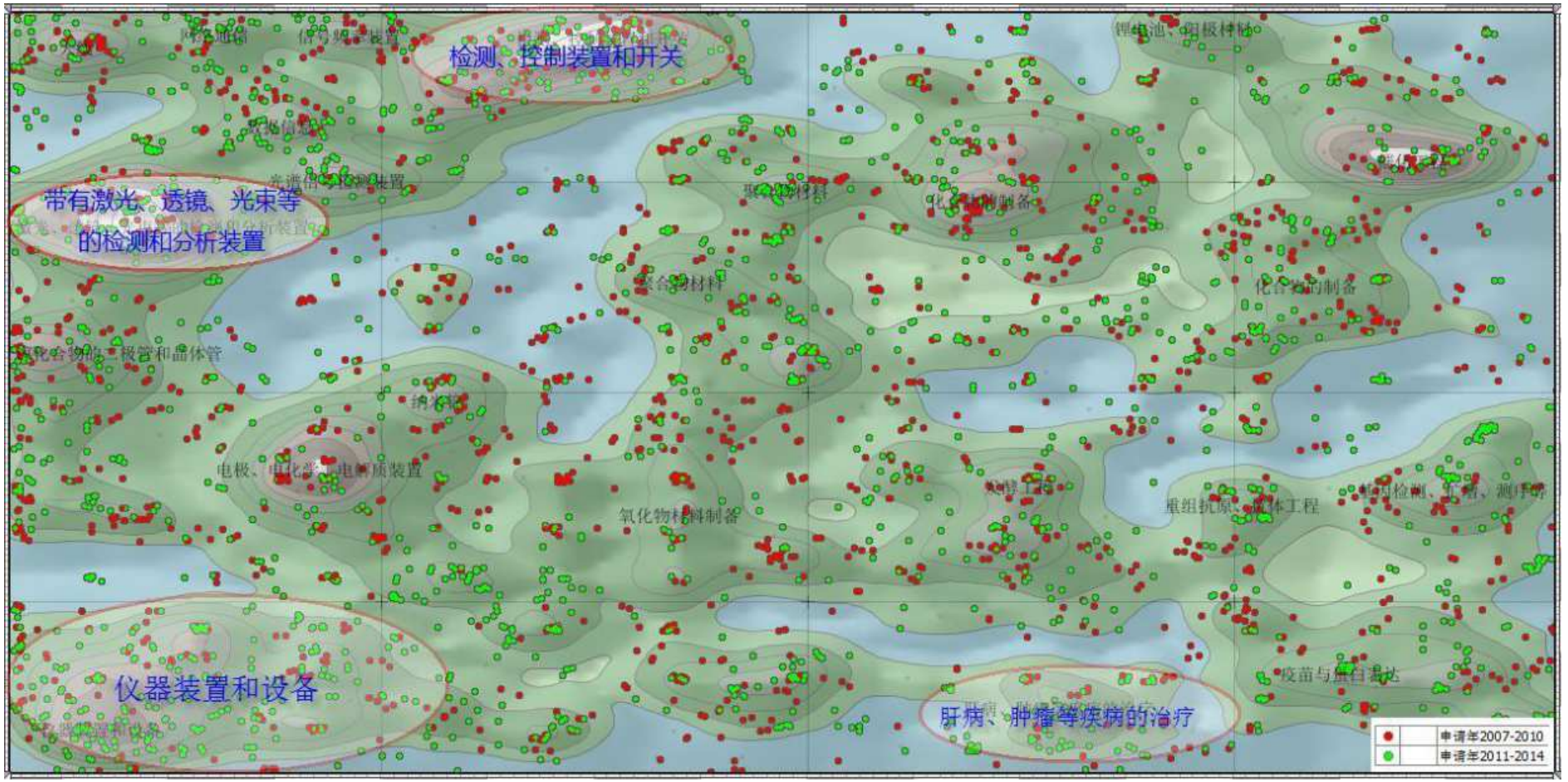
相关词是什么？

- 共同出现，co-occurrence
- 同时发生、存在或出现
- 在一个文本语料库中，两个术语超出正常频率的同时发生、存在或出现。
- 在语言学里，共现可以解释为语义上相似的指标或者是固定搭配（成语）表达。

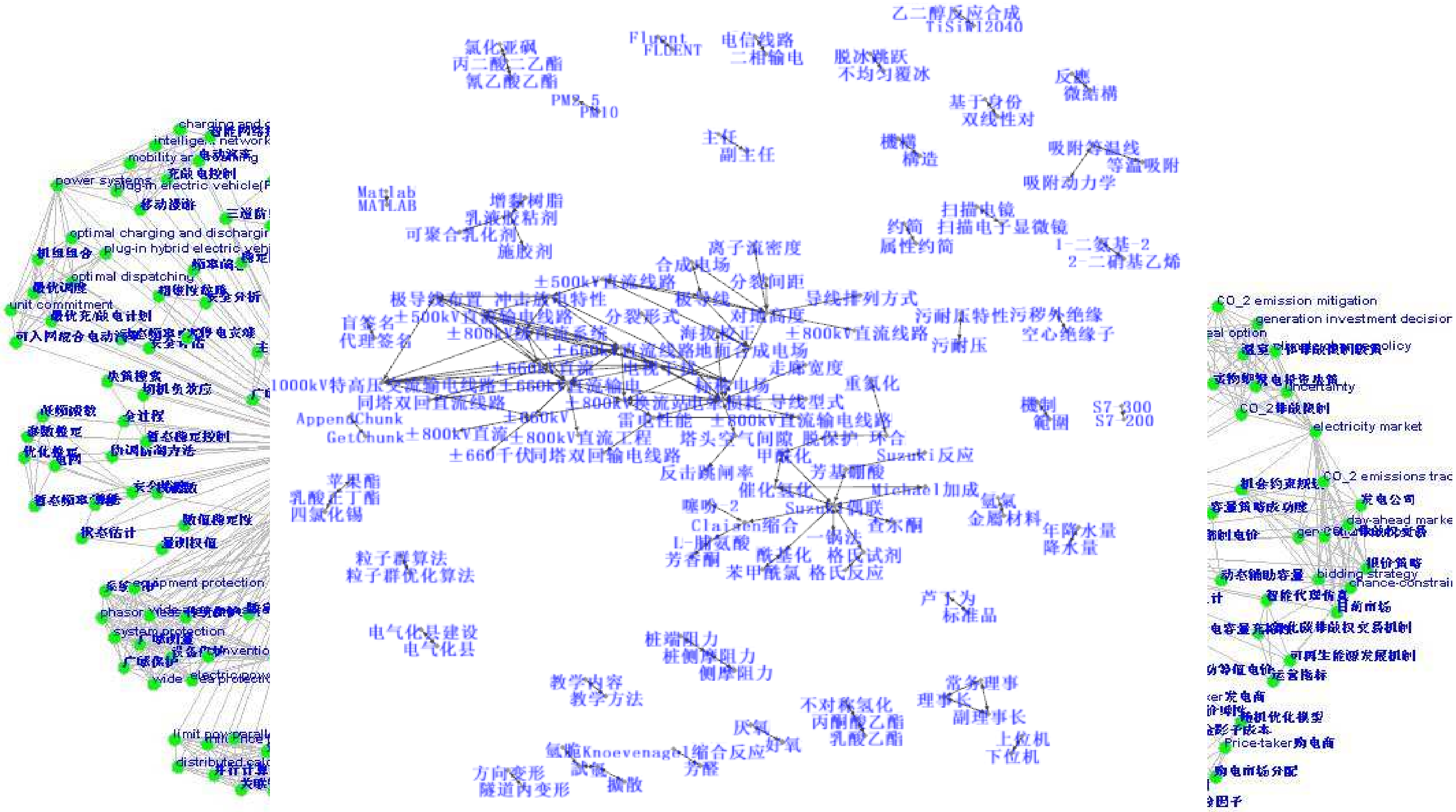


a) 普通名词　　b) 专有名词

相关词意义

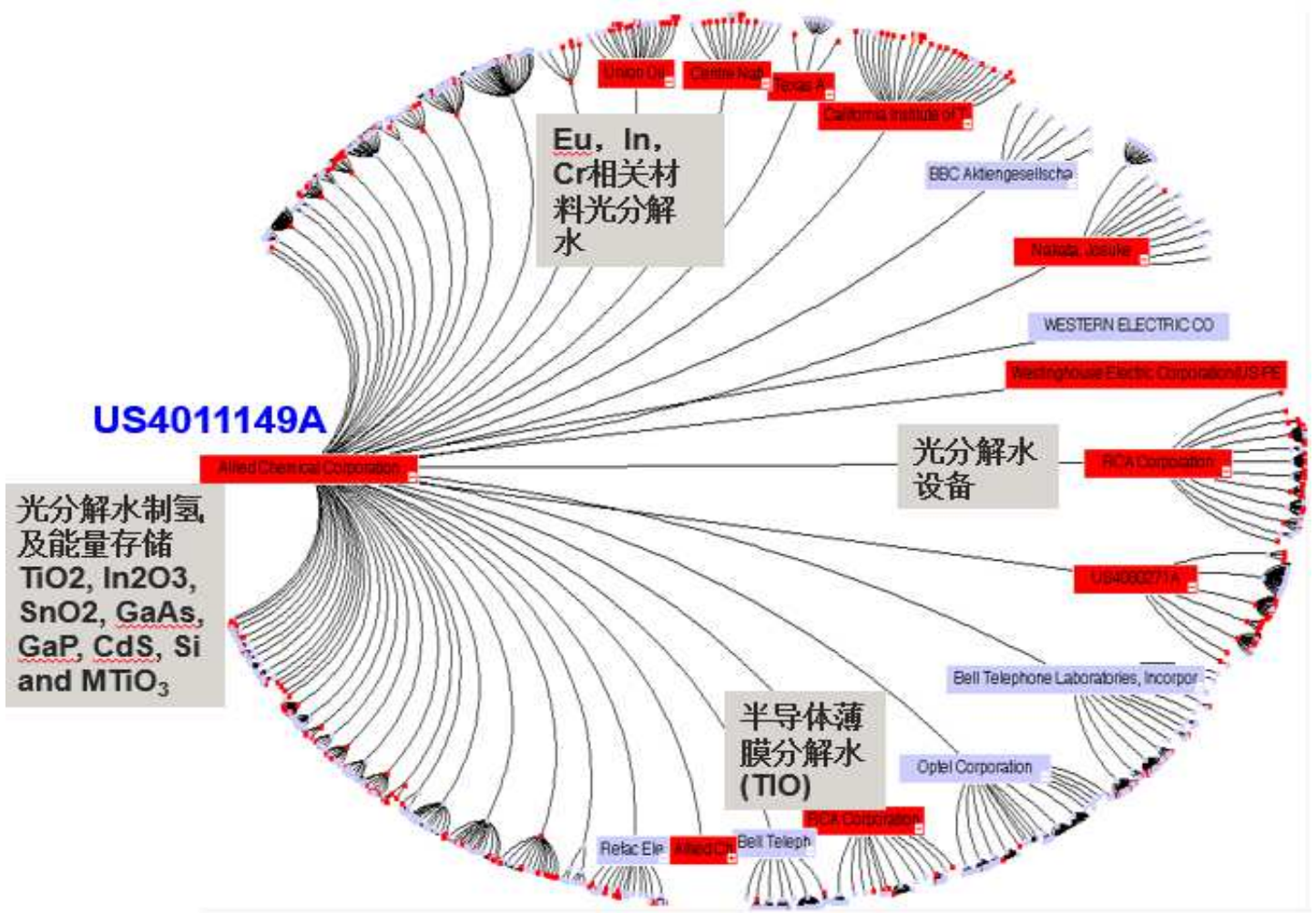

检测、控制装置和开关
带有激光、透镜、光束等
的检测和分析装置
仪器装置和设备
肝病、肿瘤等疾病的治疗
申请年2007-2010
申请年2011-2014

相关词意义

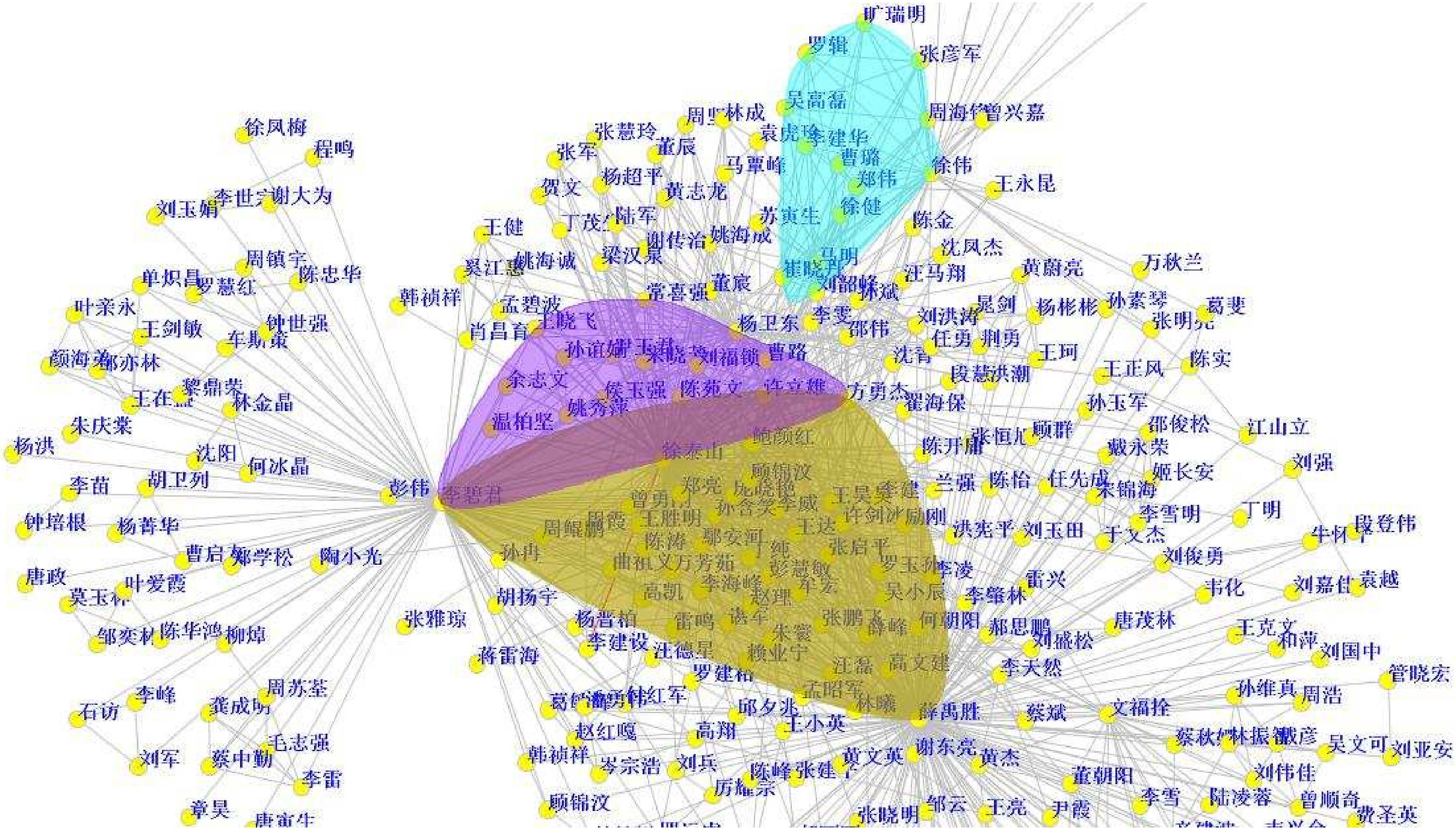


乙二醇反应合成
TiSi-1204
Fluent
FLUENT
电信线路
二相输电
脱冰跳跃
不均匀覆冰
氯化亚砜
丙二酸二乙酯
氯乙酸乙酯
反應
微結構
基于身份
双线性对
PM2.5
PM10
主任
副主任
機構
構造
吸附等温线
等温吸附
吸附动力学
Matlab
MATLAB
增黏树脂
乳液胶粘剂
可聚合乳化剂
施胶剂
扫描电镜
扫描电子显微镜
约简
属性约简
离子流密度
合成电场
分裂间距
±500kV直流线路
导线排列方式
极导线布置
冲击放电特性
极导线
污耐压特性
污秽外绝缘
±500kV直流输电线路
分裂形式
对地高度
±800kV直流系统
海拔校正
±800kV直流线路
空心绝缘子
污耐压
走廊宽度
±660kV直流
重氮化
1000kV特高压交流输电线路
±660kV直流输电
标称电场
同塔双回直流线路
导线型式
±800kV
雷电性能
±800kV直流输电线路
机制
範圍
S7-300
S7-200
AppendChunk
GetChunk
±800kV直流
±800kV直流工程
塔头空气间隙
环合
±660千伏同塔双回输电线路
甲酰化
Suzuki反应
反击跳闸率
芳基硼酸
苹果酯
乳酸正丁酯
四氯化锡
催化氢化
Michael加成
氢氧
金属材料
Suzuki偶联
查尔酮
年降水量
降水量
Claisen缩合
一锅法
L-脯氨酸
芳香酮
酰基化
格氏试剂
苯甲酰氯
格氏反应
粒子群算法
粒子群优化算法
芦丁
标准品
电气化县建设
电气化县
桩端阻力
桩侧摩阻力
侧摩阻力
教学内容
教学方法
常务理事
理事长
副理事长
不对称氢化
丙酮酸乙酯
乳酸乙酯
上位机
下位机
厌氧
好氧
Knoevenagel缩合反应
方向变形
隧道内变形
扩散
charging and
intelligent network
mobility
power system
plug-in electric vehicle
optimal charging and dischargin
plug-in hybrid electric vehi
optimal dispatching
unit commitment
equipment protection
phasor
system protection
limit
distributed
CO_2 emission mitigation
generation investment decision
electricity market
CO_2 emissions trad
day-ahead market
bidding strategy
chance-constrai
Price-taker

相关词意义


US4011149A
光分解水制氢及能量存储
TiO2, In2O3, SnO2, GaAs, GaP, CdS, Si and MTiO3
Eu，In，Cr相关材料光分解水
光分解水设备
半导体薄膜分解水(TIO)
BBC Aktiengesellscha
WESTERN ELECTRIC CO
RCA Corporation
Bell Telephone Laboratories, Incorpor
Optel Corporation
RCA Corporation
Bell Teleph

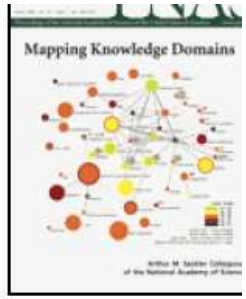
相关词意义


旷瑞明
罗辑
张彦军
吴高磊
徐伟
王永昆
陈金
沈凤杰
黄蔚亮
万秋兰
徐凤梅
程鸣
刘玉娟
周镇宇
陈忠华
单炽昌
罗慧红
叶亲永
王剑敏
钟世强
颜海涛
黎鼎荣
林金晶
朱庆棠
杨洪
沈阳
何冰晶
李苗
胡卫列
彭伟
李碧君
钟培根
杨菁华
陶小光
唐政
叶爱霞
柳婷
周苏荃
李峰
石访
刘军
毛志强
李雷
章昊
张军
贺文
王健
韩祯祥
孟碧波
肖昌育
余志文
温柏坚
徐泰山
鲍颜红
孙冉
胡扬宇
张雅琼
蒋雷海
李建设
赵红嘎
韩祯祥
顾锦汶
高翔
刘兵
王小英
薛禹胜
谢东亮
黄杰
王亮
尹霞
张晓明
蔡斌
文福拴
李雪
曾顺奇
费圣英
刘伟佳
吴文可
刘亚安
周浩
管晓宏
孙维真
刘国中
王克文
唐茂林
韦化
刘嘉佳
袁越
刘俊勇
丁明
段登伟
刘强
江山立
邵俊松
姬长安
李雪明
于文杰
洪宪平
刘玉田
雷兴
李天然
方勇杰
翟海保
孙玉军
陈实
王正风
葛斐
张明光
杨彬彬
孙素琴
汪马翔
曹路
杨卫东

Mapping Topic Bursts

Co-word space of the top 50 highly frequent and bursty words used in the top 10% most highly cited PNAS publications in 1982-2001.

Mane & Börner. (2004) PNAS, 101(Suppl. 1): 5287-5290.



中文分词

- 作用
  - 找出单个词汇
- 中文文本处理的第一步
  - 我国科学家近日研制出一套水下反恐监控系统
  - 我国/科学家/近日/研制/出/一套/水下/反恐/监控/系统
- 分词方法
  - **基于词典的方法**：给出一部词典，根据这部词典进行匹配
  - 无词典的方法：不需要词典，根据某种人工构词规则或者统计规则从字生成词

中文分词实际挑战

- 中文分词歧义
  - 交集型
    - “部分居民生活水平”：分居、居民、民生、生活、
    - “我们小组合成氢气”：我们/小组/合成/氢气或我们/小/组合/成/氢气
  - 组合型
    - “老人家”：老人、老人家
    - 他/从/马/上/下/来；我/马上/就/来/了
- 未登录词
  - 专有名词（人名、地名、机构名、译名、术语等）、新词

本实验不考虑

中文分词方法1

- **正向最大匹配(基于词典的方法)**

最大匹配法的分词实现很简单，并且可以满足一些对分词准确率要求不高的检索系统，该方法在早期的分词系统中被广泛使用

0	1	2	3	4	5	6
他	说	的	确	实	在	理

指针位置	剩余词串	首字	最大匹配词条
0	他说的确实在理	他	他
1	说的确实在理	说	说
2	的确实在理	的	的确
4	实在理	实	实在
6	理	理	理

中文分词方法2

- 逆向最大匹配(基于词典的方法)

0	1	2	3	4	5	6
他	说	的	确	实	在	理

指针位置	剩余词串	尾字	最大匹配词条
6	他说的确实在理	理	在理
4	他说的确实	实	确实
2	他说的	的	的
1	他说	说	说
0	他	他	他

几个著名的中文分词器

- Paoding Analysis（庖丁解牛）
  - Lucene御用中文分词
  - 48个java文件，6895 行，使用不用的 Knife 切不同类型的流
- Imdict
  - imdict智能词典，使用ICTCLAS HHMM隐马尔科夫模型中文分词
  - 20个java文件，2399行
- mmseg4j
  - 用 Chih-Hao Tsai 的 MMSeg 算法 实现的中文分词器
  - 23个java文件，2089行
- ik
  - 采用了特有的“正向迭代最细粒度切分算法”
  - 多子处理器分析模式
  - 22个java文件，4217行

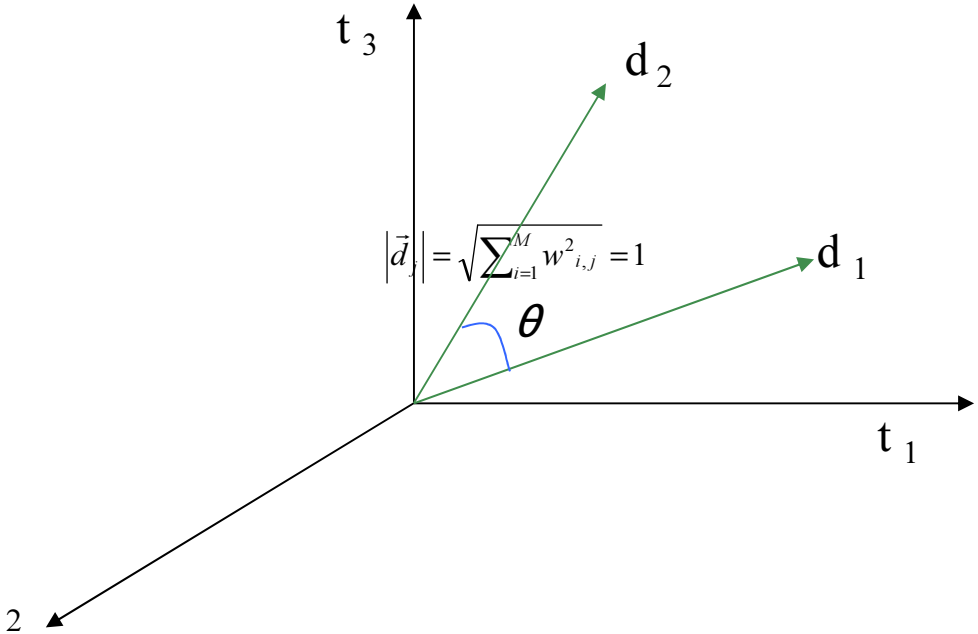
建立统计窗口

同时，除了解发文量较高的研究机构分布之外，还需进一步分析相对活跃的机构之间的合作关系。研究人员可以据此开展访问交流，或合作研究，从而进一步推动研究的进展。如图 7 所示，东京大学不仅在动画相关领域中发文量较多，在国际合作中也表现得相对活跃。另外，斯坦福大学及多伦

以标点符号分隔的一句话作为统计窗口

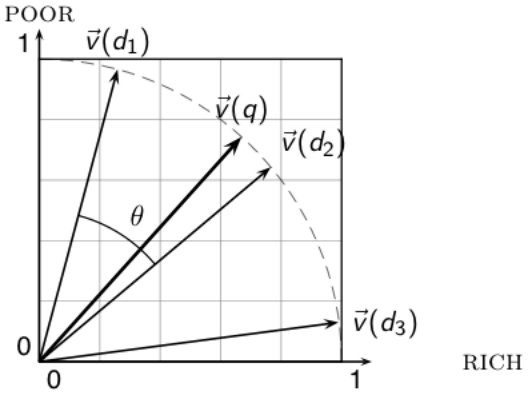
普通方法（1）：
向量空间模型

Cosine similarity余弦相似度



- 向量d_1和d_2的 “closeness”可以用它们之间的夹角大小来度量
- 具体的，可用cosine of the angle *x来计算向量相似度*.
- 向量按长度归一化 Normalization

$$sim(d_j, d_k) = \frac{\vec{d}_j \cdot \vec{d}_k}{\left|\vec{d}_j\right|\left|\vec{d}_k\right|} = \frac{\sum_{i=1}^{M} w_{i,j} w_{i,k}}{\sqrt{\sum_{i=1}^{M} w_{i,j}^2}\sqrt{\sum_{i=1}^{M} w_{i,k}^2}}$$



构建向量

假定生成了N个统计窗口
每个词对应一个N个元素的向量
计算所有两两词对之间的Cos距离

	窗口1	窗口2	窗口3	窗口4	……
词I	0	1	0	1	
词j	0	1	1	0	

普通方法（2）：互信息排序

信源

- 信源：信息来源。
  - 离散信源
  - 连续信源
- 通常用随机变量x来表示一个离散信源

- 思考：当信源发出某个信号x时，它提供了多少信息？
- 进一步思考：什么样的事件信息更多？

自信息

- 事件的不确定性决定了信息的多少
- 我们用自信息I(x)来表示事件x所具有的信息，那么I(x)应该是p(x)的一个函数。
- 思考：I(x)应该有哪些性质？
  - 非负性：I(x) ≥ 0
  - P(x)=0,I(x)=+∞;P(x)=1,I(x)=0
  - 单调性：若p(x) > p(y) ，那么I(x) < I(y)
  - 若p(x,y) = p(x)p(y)，那么I(xy) = I(x) + I(y)

$$I(x) = C * \log \frac{1}{p(x)}$$

最优编码

- 实例：假定有一个房间中有时没有人，有时甲在房间中，有时乙在房间中，有时甲乙都在房间中，房间状态服从下面的概率分布

房间状态	房间没有人	甲在房间	乙在房间	甲乙均在房间
概率	0.5	0.125	0.125	0.25

- 某人受命监视房间，每五分钟记录一次房间状态，并经一个通讯设备将房间状态发送出去。
- 问题：状态如何编码？

最优编码

- 一种可行的定长编码方案：用00 表示没有人在房间中，01 表示甲在房间中，10 表示乙在房间中，11 表示甲乙两人均在在房间中。
- 按照这样的编码，发送一个消息所需要的码的长度为2，平均发送一个消息需要2个二进制位。
- 思考：有没有一种更短的编码方式？

消息	编码
房间没有人	0
甲在房间	110
乙在房间	111
甲乙均在房间	10

最优编码

- 如果消息x 的概率为p(x)，则给其分配一个长度为⌈−log2p(x)⌉个二进制位的编码
- 平均发送一个消息所需要的编码的长度（以二进制位衡量）：

$$-\sum p(x)\log_2 p(x)$$

熵

定义1 熵 设 X 是取有限个值的随机变量，它的分布密度为

$p(x) = P\{X=x\}$， 且$x \in X$

则，X的熵定义为

$$H(X) = -\sum_{x \in X} p(x) \log_a p(x)$$

规定 $0\log_a 0 = 0$

通常a=2，此时熵的单位为比特。

熵的基本性质:

1. $H(X) \geqslant 0$，等号表明确定场(无随机性)的熵最小。
2. $H(X) \leqslant \log|X|$，等号表明等概场的熵最大。

- 熵描述了随机变量的不确定性。

熵

- 假定一种语言P有6个字母p、t、k、a、i、u，每个字母的概率为：

P	*p*	*t*	*k*	*a*	*i*	*u*
概率	1/8	1/4	1/8	1/4	1/8	1/8

则随机变量 P 的熵为：

$$H(P) = -\sum_{i\in\{p,t,k,a,i,u\}} p(i)\log p(i)$$

$$= -[4\times\frac{1}{8}\log\frac{1}{8} + 2\times\frac{1}{4}\log\frac{1}{4}]$$

$$= 2\frac{1}{2}\text{ bit}$$

语言的字母熵

联合熵与条件熵

- 定义2 联合熵 设X、Y是两个离散型随机变量，它们的联合分布密度为$p(x,y)$，则X,Y的联合熵定义为：

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y)$$

- 定义3 条件熵 设X、Y是两个离散型随机变量，它们的联合分布密度为$p(x,y)$，则给定X时Y的条件熵定义为：

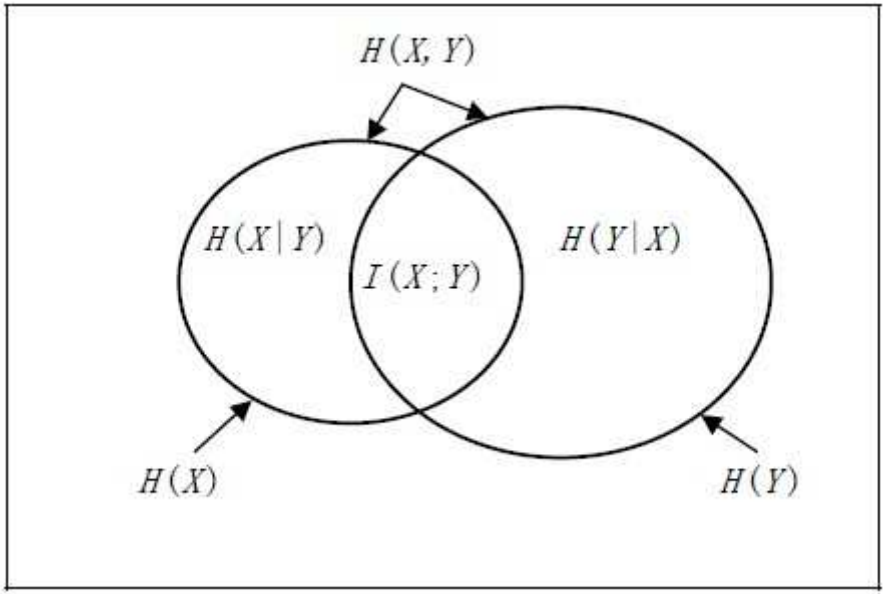
$$H(Y|X) = -\sum_{x\in X} p(x)H(Y|X=x)$$

$$= \sum_{x\in X} p(x)\left[-\sum_{y\in Y} p(y|x)\log p(y|x)\right]$$

$$= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y|x)$$

- 链式规则 $H(X,Y) = H(X) + H(Y|X)$

- H(X)和H(X|Y) 的差称为互信息，一般记作I(X;Y)。 I(X;Y)描述了包含在X 中的有关Y 的信息量，或包含在Y 中的有关X 的信息量。



互信息

- 如何量化互信息？

$$I(X;Y) = H(X) - H(X|Y)$$

$$= H(X) + H(Y) - H(X,Y)$$

$$= \sum_{x} p(x) \log \frac{1}{p(x)} + \sum_{x} p(y) \log \frac{1}{p(y)} + \sum_{x,y} p(x,y) \log p(x,y)$$

$$= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

点间互信息

- 更为常用的是两个具体事件之间的
  - 我们常说的“互信息”，一般称之为点间互信息。
- 点间互信息：事件x, y之间的互信息定义为:

$$I(x,y)=\log\frac{p(x,y)}{p(x)p(y)}$$

- 点间互信息度量两个具体事件之间的相关程度
  - 当时I(x,y)>>0，x和y高度相关。
  - 当时I(x,y)=0，x和y高度相互独立。
  - 当时I(x,y)<<0，x和y呈互补分布

计算方式

同时，除了解发文量较高的研究机构分布之外，还需进一步分析相对活跃的机构之间的合作关系。研究人员可以据此开展访问交流，或合作研究，从而进一步推动研究的进展。如图 7 所示，东京大学不仅在动画相关领域中发文量较多，在国际合作中也表现得相对活跃。另外，斯坦福大学及多伦

P(x,y)计算所有两两词对同时出现在一个滑动窗口内的概率
P(x)计算单个词出现的概率

高阶方法（1）：频繁模式挖掘

关联规则简介

- 关联规则反映一个事物与其他事物之间的相互依存性和关联性。如果两个或者多个事物之间存在一定的关联关系，那么，其中一个事物就能够通过其他事物预测到。
- 典型的关联规则发现问题是对超市中的货篮数据（Market Basket）进行分析。通过发现顾客放入货篮中的不同商品之间的关系来分析顾客的购买习惯。

什么是关联规则挖掘

- 关联规则挖掘
  - 首先被Agrawal, Imielinski and Swami在1993年的SIGMOD会议上提出
  - 在事务、关系数据库中的项集和对象中发现频繁模式、关联规则、相关性或者因果结构
  - 频繁模式: 数据库中频繁出现的项集
- 目的: 发现数据中的规律
  - 超市数据中的什么产品会一起购买？— 啤酒和尿布
  - 在买了一台PC之后下一步会购买?
  - 哪种DNA对这种药物敏感?
  - 我们如何自动对Web文档进行分类?

频繁模式挖掘的重要性

- 许多重要数据挖掘任务的基础
  - 关联、相关性、因果性
  - 序列模式、空间模式、时间模式、多维
  - 关联分类、聚类分析
- 更加广泛的用处
  - 购物篮分析、交叉销售、直销
  - 点击流分析、DNA序列分析等等

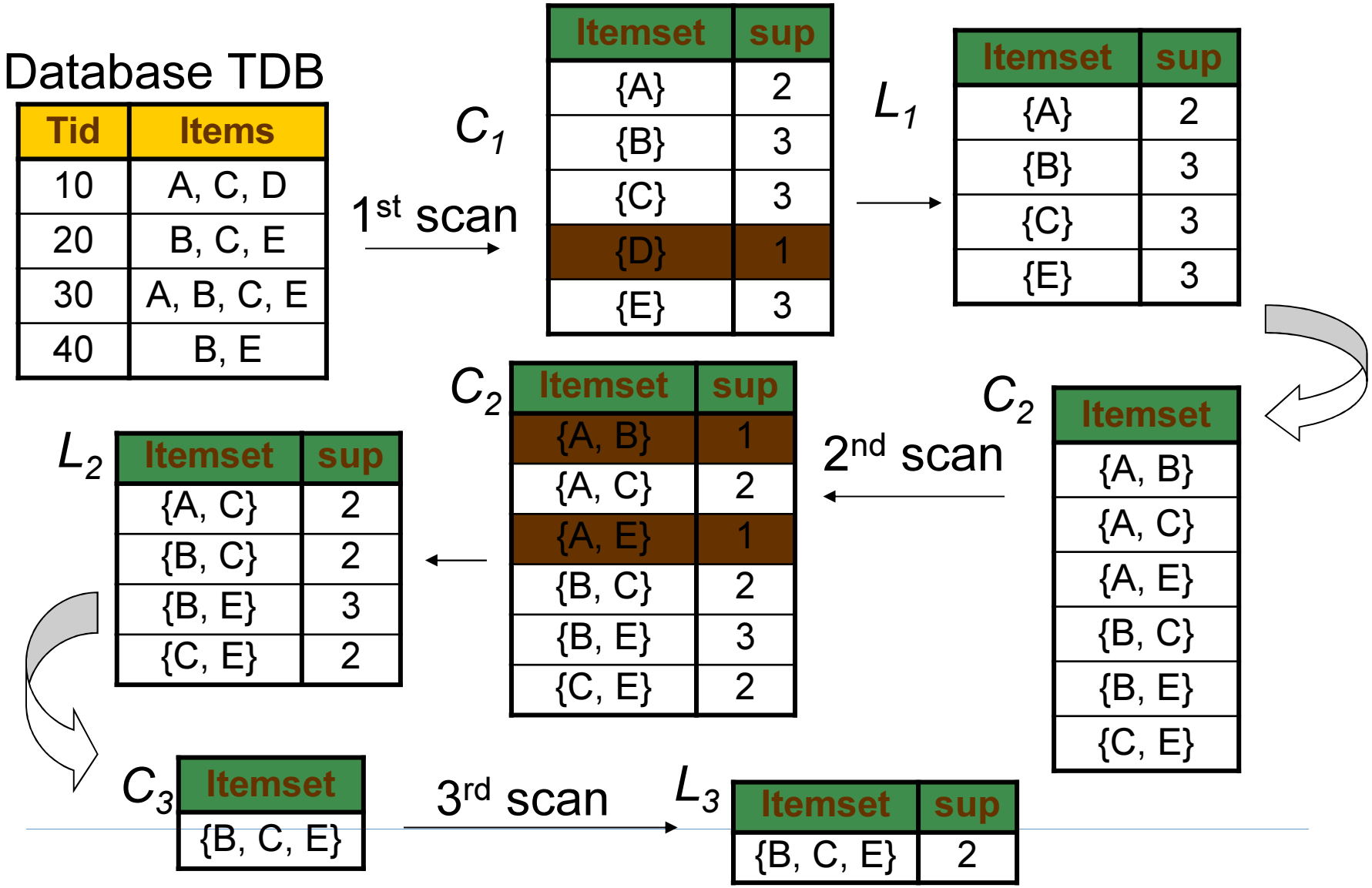
关联规则基本模型

- 关联规则基本模型
- Apriori算法
- Fp-Tree算法

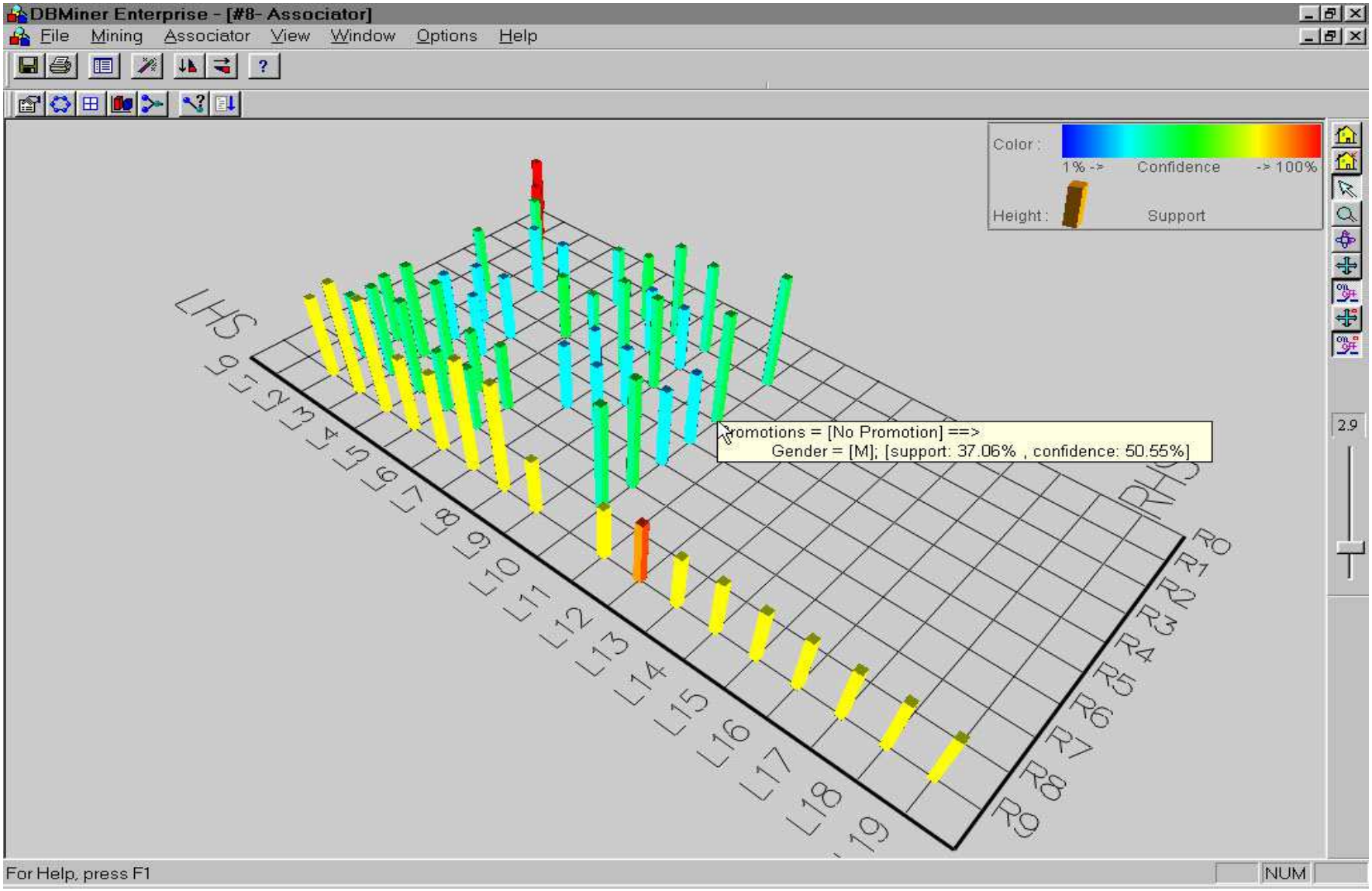
Apriori算法

- (1) L_1={频繁1项集};
- (2) for(k=2;$L_{k-1}\neq\varnothing$;k++) do begin
- (3) C_k=apriori_gen(L_{k-1}); //新的潜在频繁项集
- (4) for all *transactions* $t\in D$ do begin
- (5) C_t=subset(C_k,t); //t中包含的潜在频繁项集
- (6) for all *candidates* $c\in C_t$ do
- (7) $c.count$++;
- (8) end;
- (9) $L_k=\{c\in C_k|c.count\geq minsup\}$
- (10) end;
- (11) Answer=$\bigcup_k L_k$

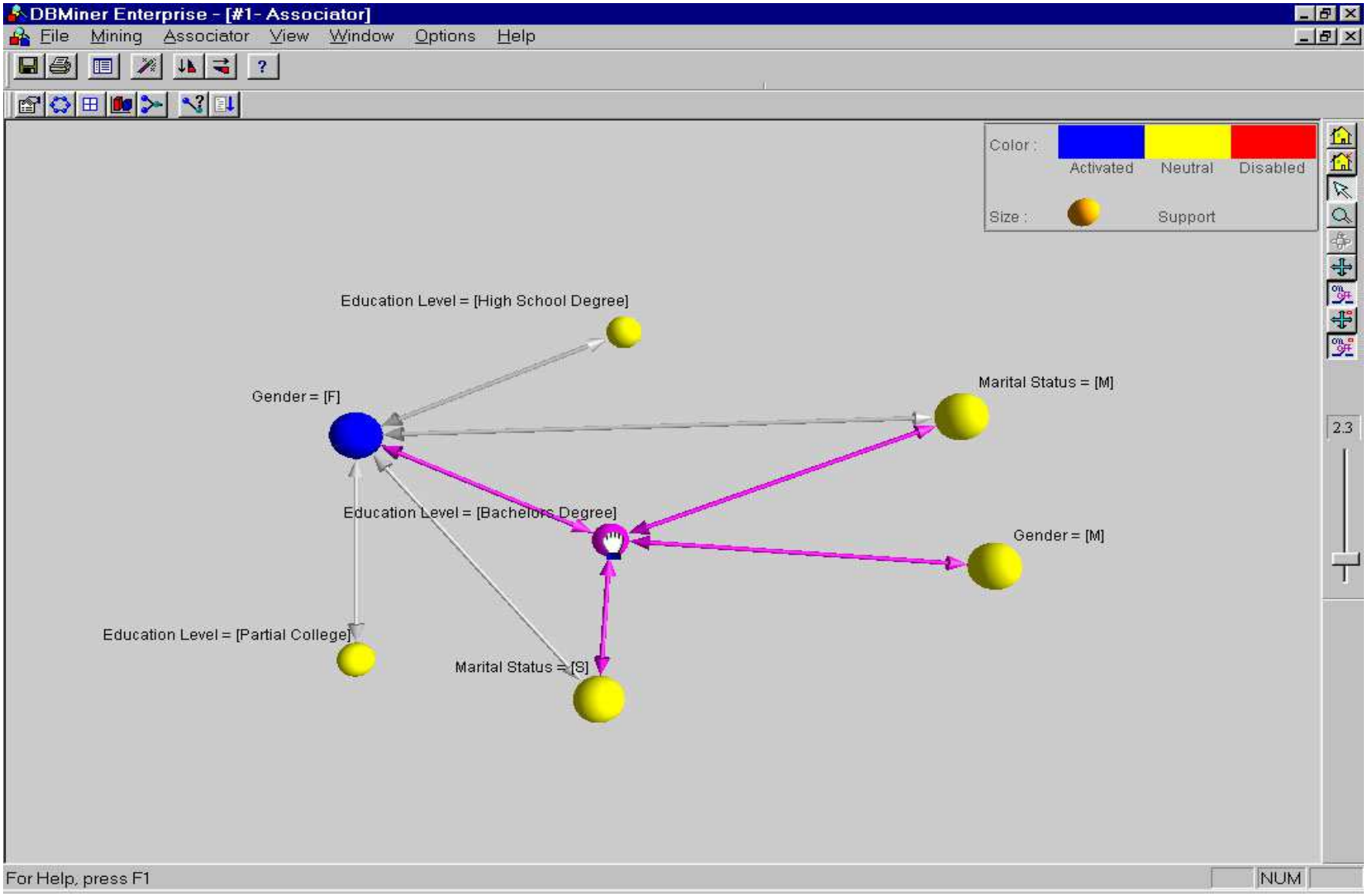
实例


Database TDB
Tid
Items
10
A, C, D
20
B, C, E
30
A, B, C, E
40
B, E
C_1
1st scan
Itemset
sup
{A}
2
{B}
3
{C}
3
{D}
1
{E}
3
L_1
Itemset
sup
{A}
2
{B}
3
{C}
3
{E}
3
C_2
Itemset
sup
{A, B}
1
{A, C}
2
{A, E}
1
{B, C}
2
{B, E}
3
{C, E}
2
2nd scan
C_2
Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}
L_2
Itemset
sup
{A, C}
2
{B, C}
2
{B, E}
3
{C, E}
2
C_3
Itemset
{B, C, E}
3rd scan
L_3
Itemset
sup
{B, C, E}
2

Visualization of Association Rules: Pane Graph



Visualization of Association Rules: Rule Graph



挖掘频繁集 不用生成候选集

- 用Frequent-Pattern tree (FP-tree) 结构压缩数据库,
  - 高度浓缩，同时对频繁集的挖掘又完备的
  - 避免代价较高的数据库扫描
- 开发一种高效的基于FP-tree的频繁集挖掘算法
  - 采用分而治之的方法学：分解数据挖掘任务为小任务
  - 避免生成关联规则: 只使用部分数据库!

建立 FP-tree树

TID	*Items bought*	*(ordered) frequent items*
100	{*f, a, c, d, g, i, m, p*}	{*f, c, a, m, p*}
200	{*a, b, c, f, l, m, o*}	{*f, c, a, b, m*}
300	{*b, f, h, j, o*}	{*f, b*}
400	{*b, c, k, s, p*}	{*c, b, p*}
500	{*a, f, c, e, l, p, m, n*}	{*f, c, a, m, p*}

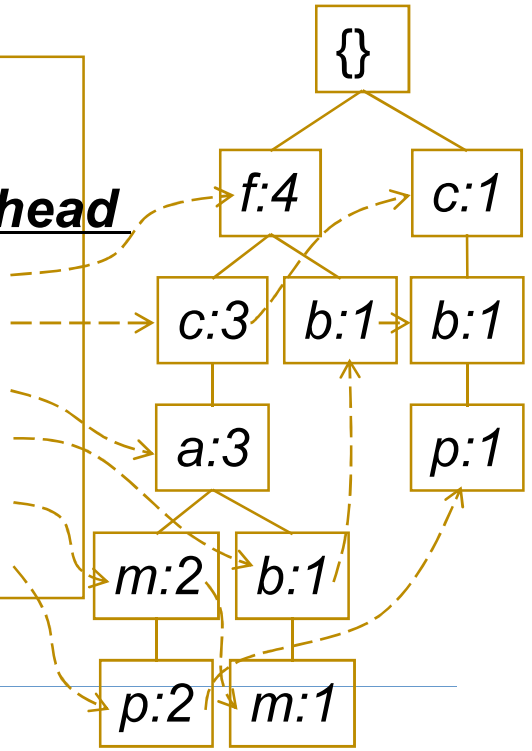
最小支持度 = 0.5

步骤:

1. 扫描数据库一次，得到频繁1-项集
2. 把项按支持度递减排序
3. 再一次扫描数据库，建立FP-tree

头表

Item	*frequency*	*head*
f	*4*	
c	*4*	
a	*3*	
b	*3*	
m	*3*	
p	*3*	



高阶方法（2）：深度学习 Word2Vector

词向量

- 自然语言中的词语在机器学习中表示符号
  - One-hot Representation

  例如：
    - “话筒”表示为 [0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 ...]
    - “麦克”表示为 [0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 ...]
    - 实现时就可以用0,1,2,3,...来表示词语进行计算，这样“话筒”就为3，“麦克”为8.
  - 存在两个问题
    - 维度比较大,尤其是用于 Deep Learning 的一些算法时
    - 词汇鸿沟：任意两个词之间都是孤立的，不能体现词和词之间的关系

词向量

- Distributional Representation
  - 词表示为：
    - [0.792, −0.177, −0.107, 0.109, 0.542, ...]，常见维度50或者100
  - 解决“词汇鸿沟”问题
    - 可以通过计算向量之间的距离（欧式距离、余弦距离等）来体现词与词的相似性
- 如何训练这样的词向量
  - 没有直接的模型可训练得到
  - 可通过训练语言模型的同时，得到词向量

语言模型

- 判断一句话是不是正常人说出来的，用数学符号描述为
  - 给定一个字符串"w1,w2,...,wt",计算它是自然语言的概率$p(w_1, w_2, ..., w_t)$，一个很简单的推论是

$$p(w_1, w_2, ..., w_t) = p(w_1) \cdot p(w_2 \mid w_1) \cdot p(w_3 \mid w_1, w_2) \cdot ... \cdot p(w_t \mid w_1, w_2, ..., w_{t-1})$$

  - 例如，有个句子"大家,喜欢,吃,苹果"
    - P(大家，喜欢，吃，苹果)=p(大家)p(喜欢|大家)p(吃|大家,喜欢)p(苹果|大家,喜欢,吃)
  - 简单表示为$p(s) = p(w_1, w_2, ..., w_T) = \prod_{i=1}^{T} p(w_i \mid Context_i)$
- 计算$p(w_i \mid Context_i)$问题

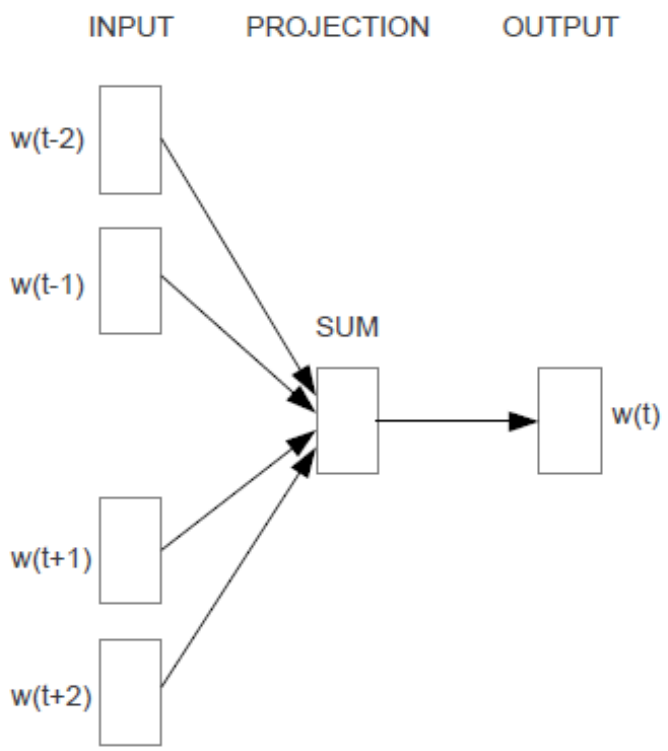
word2vec原理

- 两种模型，两种方法

模型	CBOW		Skip-Gram	
方法	Hierarchical Softmax	Negative Sampling	Hierarchical Softmax	Negative Sampling

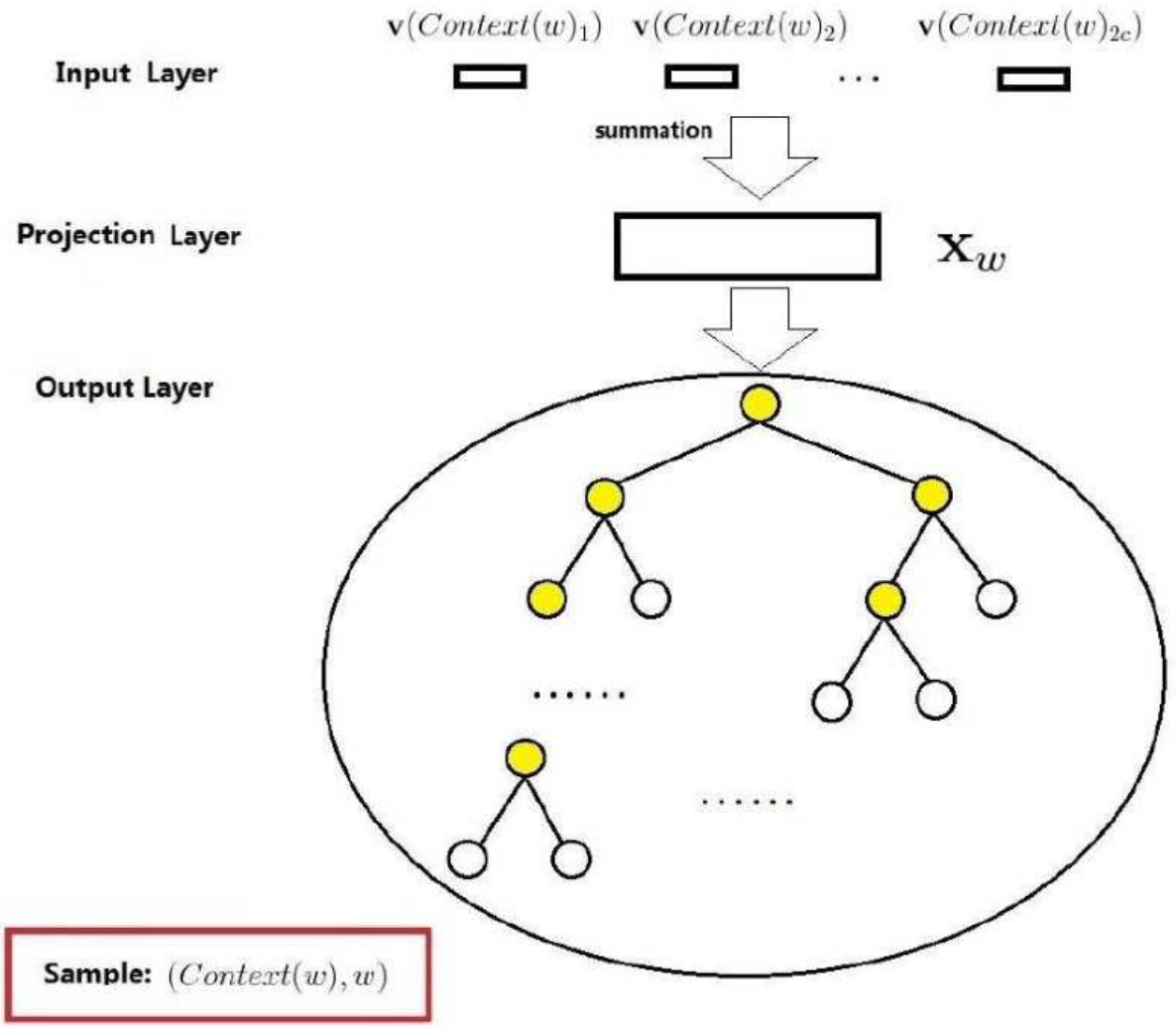
CBOW模型+Hierarchical Softmax方法

- CBOW模型
  - INPUT:输入层
  - PROJECTION:投影层
  - OUTPUT:输出层
  - w(t):当前词语（向量）
  - w(t-2),w(t-1),w(t+1),w(t+2):当前词语的上下文
  - SUM:上下文的累加和



CBOW

CBOW模型+Hierarchical Softmax方法（续）



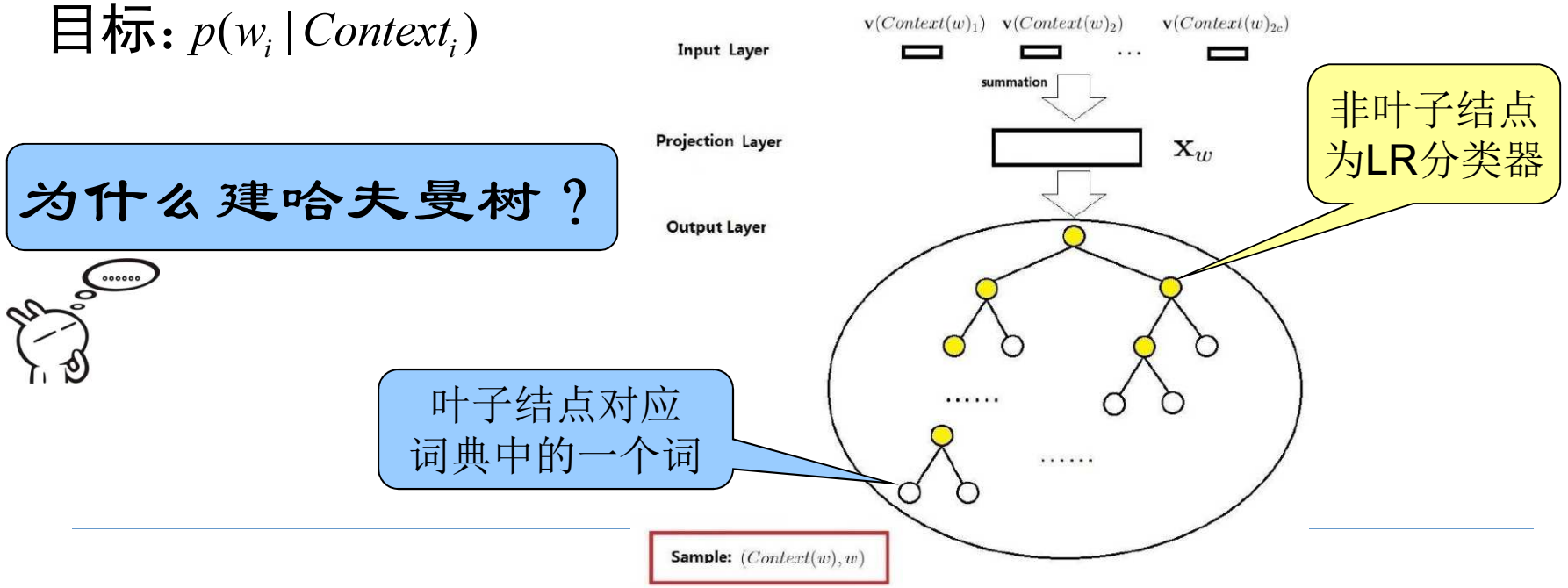
CBOW模型+Hierarchical Softmax方法（续）

1. **输入层**: 包含 $Context(w)$ 中 $2c$ 个词的词向量 $\mathbf{v}(Context(w)_1), \mathbf{v}(Context(w)_2), \cdots, \mathbf{v}(Context(w)_{2c}) \in \mathbb{R}^m$. 这里, m 的含义同上表示词向量的长度.

2. **投影层**: 将输入层的 $2c$ 个向量做求和累加, 即 $\mathbf{x}_w = \sum_{i=1}^{2c} \mathbf{v}(Context(w)_i) \in \mathbb{R}^m$.

3. **输出层**: 输出层对应一棵二叉树, 它是以语料中出现过的词当叶子结点, 以各词在语料中出现的次数当权值构造出来的 Huffman 树. 在这棵 Huffman 树中, 叶子结点共 N $(= |\mathcal{D}|)$ 个, 分别对应词典 $\mathcal{D}$ 中的词, 非叶子结点 $N-1$ 个 (图中标成黄色的那些结点).

目标：$p(w_i \mid Context_i)$



CBOW模型+Hierarchical Softmax方法（续）

- 句子：我,喜欢,观看,巴西,足球,世界杯
- w=足球

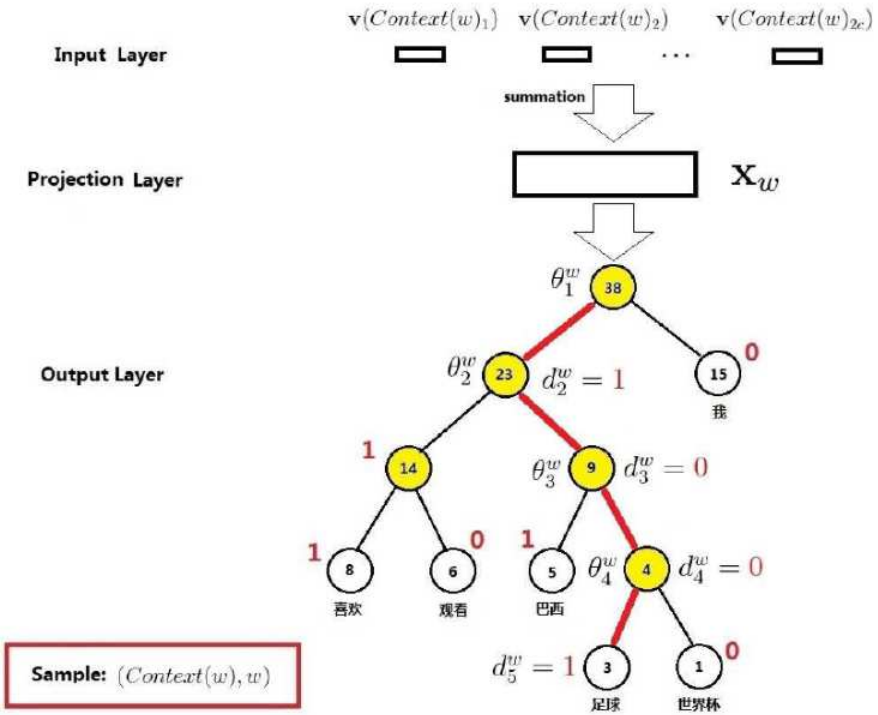
d_j^w ：编码（1或0）

θ_j^w ：非叶子节点向量

二分类

d_j^w ：正负类（1：负类,0：正类）

θ_j^w ：类别向量



CBOW模型+Hierarchical Softmax方法（续）

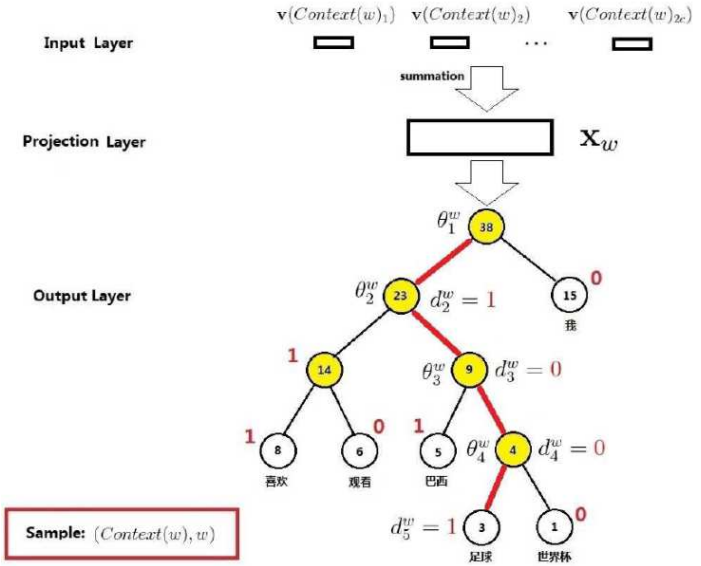
- 正类概率 $\sigma(\mathbf{x}_w^\top\theta) = \dfrac{1}{1+e^{-\mathbf{x}_w^\top\theta}}$,
- 负类概率 $1-\sigma(\mathbf{x}_w^\top\theta)$,
- "足球" 叶子节点经过4次二分类，每次分类结果对应的概率为

$$\text{第 1 次:}\quad p(d_2^w|\mathbf{x}_w,\theta_1^w) = 1-\sigma(\mathbf{x}_w^\top\theta_1^w);$$

$$\text{第 2 次:}\quad p(d_3^w|\mathbf{x}_w,\theta_2^w) = \sigma(\mathbf{x}_w^\top\theta_2^w);$$

$$\text{第 3 次:}\quad p(d_4^w|\mathbf{x}_w,\theta_3^w) = \sigma(\mathbf{x}_w^\top\theta_3^w));$$

$$\text{第 4 次:}\quad p(d_5^w|\mathbf{x}_w,\theta_4^w) = 1-\sigma(\mathbf{x}_w^\top\theta_4^w),$$



- 由Context("足球")预测"足球"出现的概率

$$p(\text{足球}|Contex(\text{足球})) = \prod_{j=2}^{5} p(d_j^w|\mathbf{x}_w,\theta_{j-1}^w).$$

CBOW模型+Hierarchical Softmax方法（续）

- 对于词典中的每个词w有 ，l^w 结点个数

$$p(w \mid Context\ (w)) = \prod_{j=2}^{l^w} p(d_j^w \mid X_w, \theta_{j-1}^w)$$

- 其中 ，$p(d_j^w \mid X_w, \theta_{j-1}^w) = \begin{cases} \sigma(X_x^T \cdot \theta_{j-1}^w), d_j^w = 0; \\ 1-\sigma(X_x^T \cdot \theta_{j-1}^w), d_j^w = 1. \end{cases}$
- 或者表示为 $p(d_j^w \mid X_w, \theta_{j-1}^w) = [\sigma(X_w^T \cdot \theta_{j-1}^w)]^{1-d_j^w} \cdot [1-\sigma(X_w^T \cdot \theta_{j-1}^w)]^{d_j^w}$
- 对于由S个句子组成的语料库C有

$$L(X,\theta) = \prod_{s \in C} \prod_{w \in s} p(w \mid Context(w)) = \prod_{s \in C} \prod_{w \in s} \prod_{j=2}^{l^w} p(d_j^w \mid X_w, \theta_{j-1}^w)$$

- 取对数似然函数

参数1

$$\log L(X,\theta) = \sum_{s \in C} \sum_{w \in s} \sum_{j=2}^{l^w} \log p(d_j^w \mid X_w, \theta_{j-1}^w)$$

参数2

$$= \sum_{s \in C} \sum_{w \in s} \sum_{j=2}^{l^w} [(1-d_j^w) \cdot \log \sigma(X_w^T \cdot \theta_{j-1}^w) + d_j^w \cdot \log(1-\sigma(X_w^T \cdot \theta_{j-1}^w))]$$

CBOW模型+Hierarchical Softmax方法（续）

- 梯度下降法进行求解
  - 令$f(w,j)=-(1-d_j^w)\cdot\log\sigma(X_w^T\cdot\theta_{j-1}^w)-d_j^w\cdot\log(1-\sigma(X_w^T\cdot\theta_{j-1}^w))$
  - f(w,j)关于θ_{j-1}^w 和X_w的梯度分别为

$$\frac{\partial f(w,j)}{\partial\theta_{j-1}^w}=-[1-d_j^w-\sigma(X_w^T\cdot\theta_{j-1}^w)]\cdot X_w$$

$$\frac{\partial f(w,j)}{\partial X_w}=-[1-d_j^w-\sigma(X_w^T\cdot\theta_{j-1}^w)]\cdot\theta_{j-1}^w$$

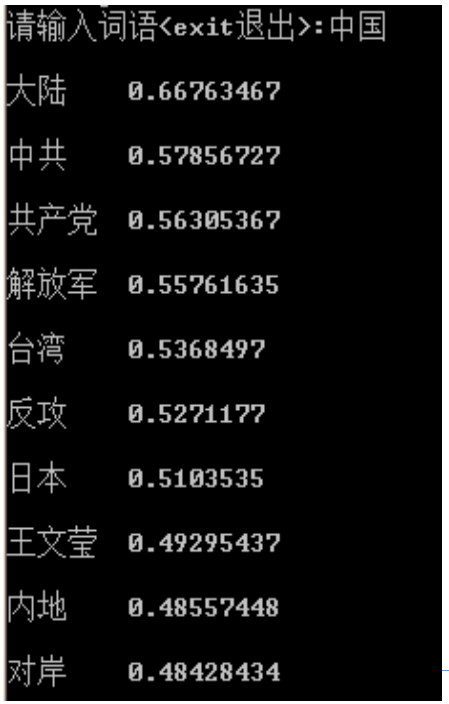
  - 更新公式

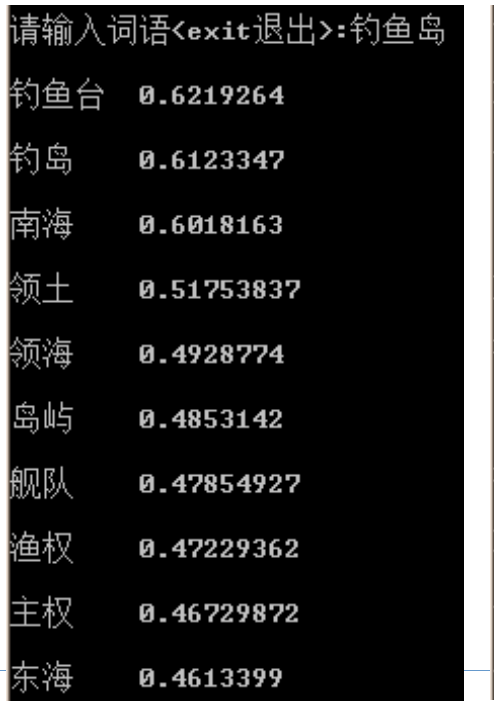
$$\theta_{j-1}^w:=\theta_{j-1}^w-\eta\cdot\frac{\partial f(w,j)}{\partial\theta_{j-1}^w}$$

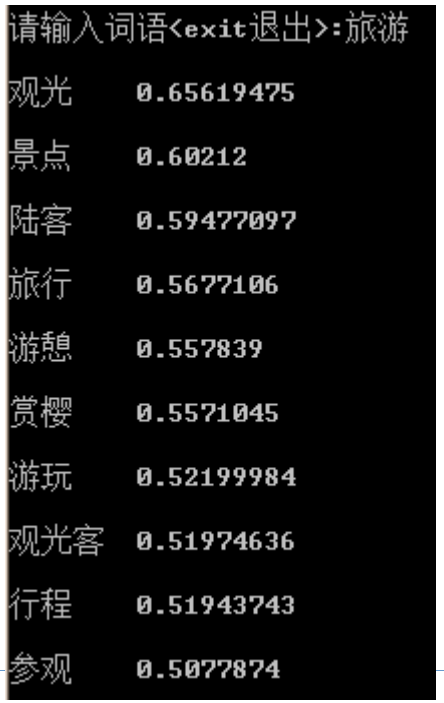
$$V(\tilde{w}):=V(\tilde{w})-\eta\cdot\sum_{j=2}^{l^w}\frac{\partial f(w,j)}{\partial X_w},\tilde{w}\in Context(w)$$

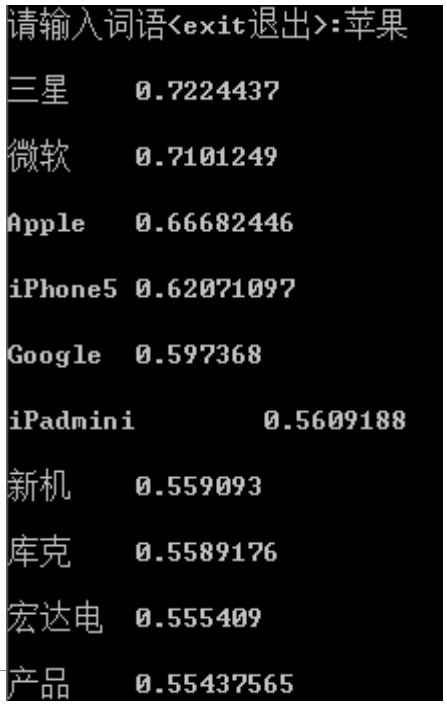
Word2vec效果

- 训练数据集：经过分词后的新闻数据，大小184MB
  - 查看"中国"，"钓鱼岛"，"旅游"，"苹果"几个词语的相似词语如下所示



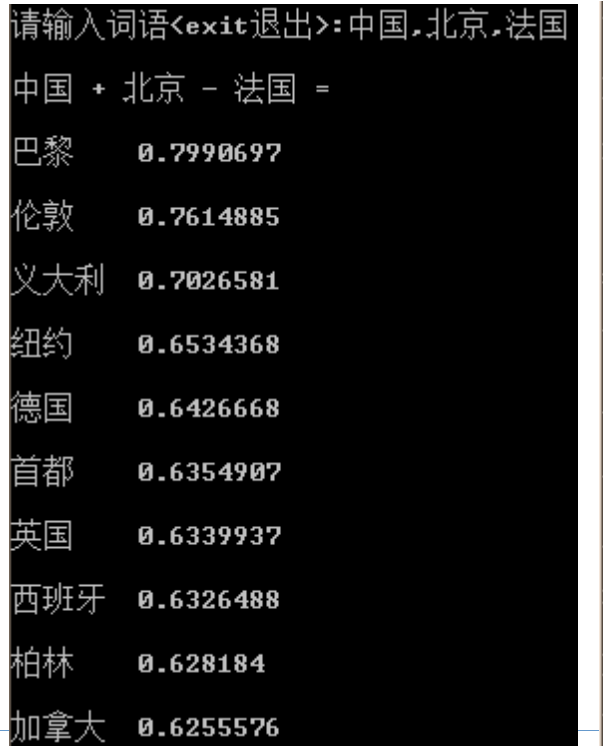


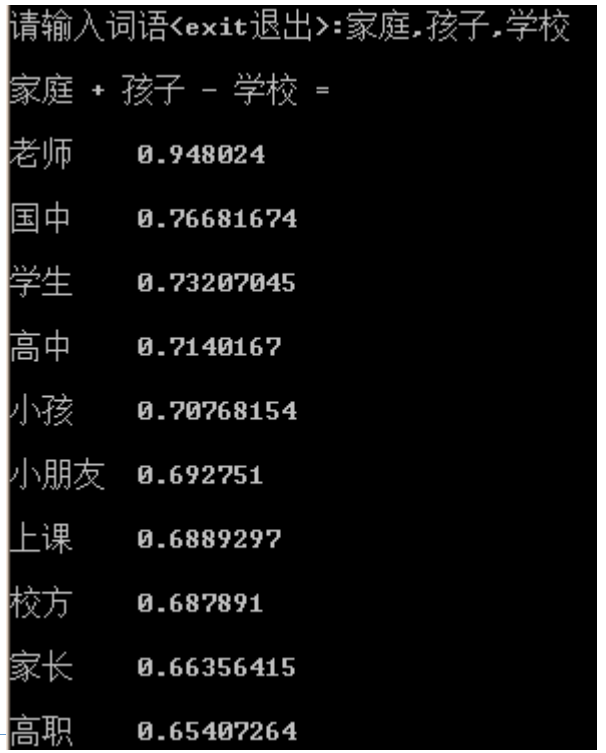




Word2vec效果

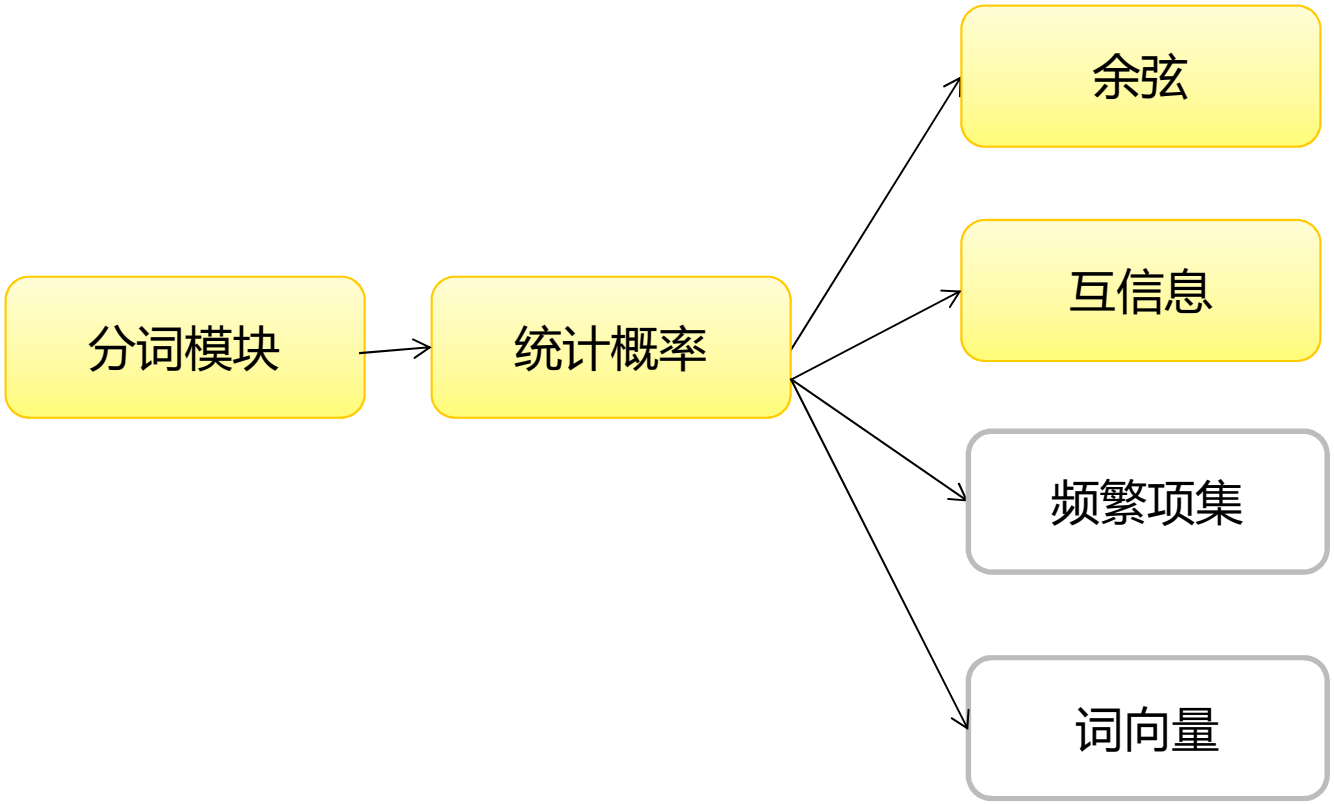
- 向量加减法
  - "中国+北京-日本"，"中国+北京-法国"，"家庭+孩子-学校"





具体工程实现

总体设计



实验说明

- 语料库
  - 大约1G电子书
- 运行要求
  - ./simword article.txt result.txt
  - Article.txt
    - 原始文本文件
  - Result.txt
    - 第1行输出：Cosine
    - 第2-21行输出：前20个相关性最高的词对
    - 第22行输出：MutualInfo
    - 第23-42行输出：前20个最相关的词对


The End