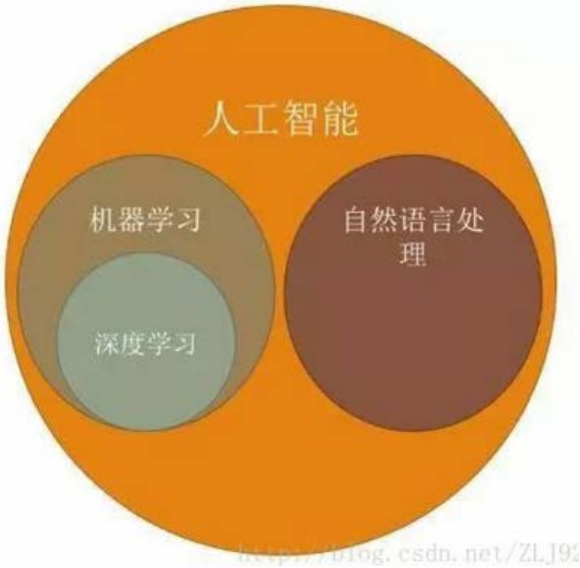
关键词提取算法介绍

欧俊杰　2018/12/20

自然语言处理（NLP）是人工智能（AI）的一个子领域。它是指**机器理解并解释人类写作、说话方式的能力。**

NLP 的目标：让计算机 / 机器在理解语言上像人类一样智能。最终目标是弥补人类交流（自然语言）和计算机理解（机器语言）之间的差距。

NLP成功应用于搜索引擎、机器翻译、语音识别和问答系统。



关键词提取是NLP领域的一个重要的子任务。

- 信息检索
- 对话系统
- 文本分类
- 。 。 。

定义：关键词是指能反映文本主题或者主要内容的词语。

文档的关键词集合应该具备以下几个性质：
①完备性
②确定性
③独立性

分词（tokenization）

① 根据空格拆分单词（Split）

② 排除停止词（Stop Word）

③ 提取词干（Stemming）

NLTK's list of english stopwords

```
i
me
my
myself
we
our
ours
ourselves
you
your
yours
yourself
yourselves
he
him
```

Compatibility of systems of linear constraints over the set of natural numbers. Criteria of compatibility of a system of linear Diophantine equations, strict inequations, and nonstrict inequations are considered. Upper bounds for components of a minimal set of solutions and algorithms of construction of minimal generating sets of solutions for all types of systems are given. These criteria and the corresponding algorithms for constructing a minimal supporting set of solutions can be used in solving all the considered types systems and systems of mixed types.

Compatibility – systems – linear constraints – set – natural numbers – Criteria – compatibility – system – linear Diophantine equations – strict inequations – nonstrict inequations – Upper bounds – components – minimal set – solutions – algorithms – minimal generating sets – solutions – systems – criteria – corresponding algorithms – constructing – minimal supporting set – solving – systems – systems

Compatibility
linear constraint
set ……
natural number
Criteria
algorithm
system

分词工具

1. Stanford NLP
2. NLTK：Python编写的开源的文本处理库
3. AutoPhrase: □ingbo Shang□□ialu □iu□□eng □iang□□iang □en□Clare □ □oss□□ia□ei □an□ □□utomated □hrase □ining from □assi□e □e□t Corpora□□□□D□ 2□1 □□
4. Jieba: "结巴"中文分词：做最好的Python中文分词组件

.....

TF-IDF算法

- **基本思想**：词语的重要性与它在文本中出现的次数成正比，但同时也会随着它在语料库中出现的频率成反比下降。

TFIDF

For a term i in document j:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

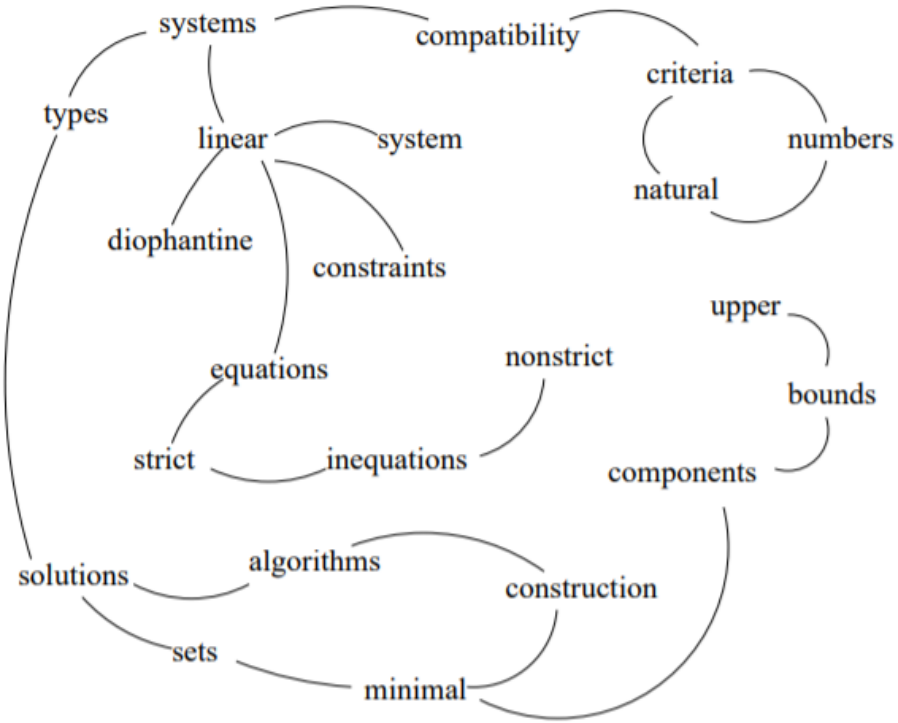
TextRank

基于排序算法**PageRank**

用关系图来表达文本、词语以及其它实体的关系，然后通过随机游走和迭代计算对顶点的重要性进行排序。

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

Compatibility of systems of linear constraints over the set of natural numbers. Criteria of compatibility of a system of linear Diophantine equations, strict inequations, and nonstrict inequations are considered. Upper bounds for components of a minimal set of solutions and algorithms of construction of minimal generating sets of solutions for all types of systems are given. These criteria and the corresponding algorithms for constructing a minimal supporting set of solutions can be used in solving all the considered types systems and systems of mixed types.

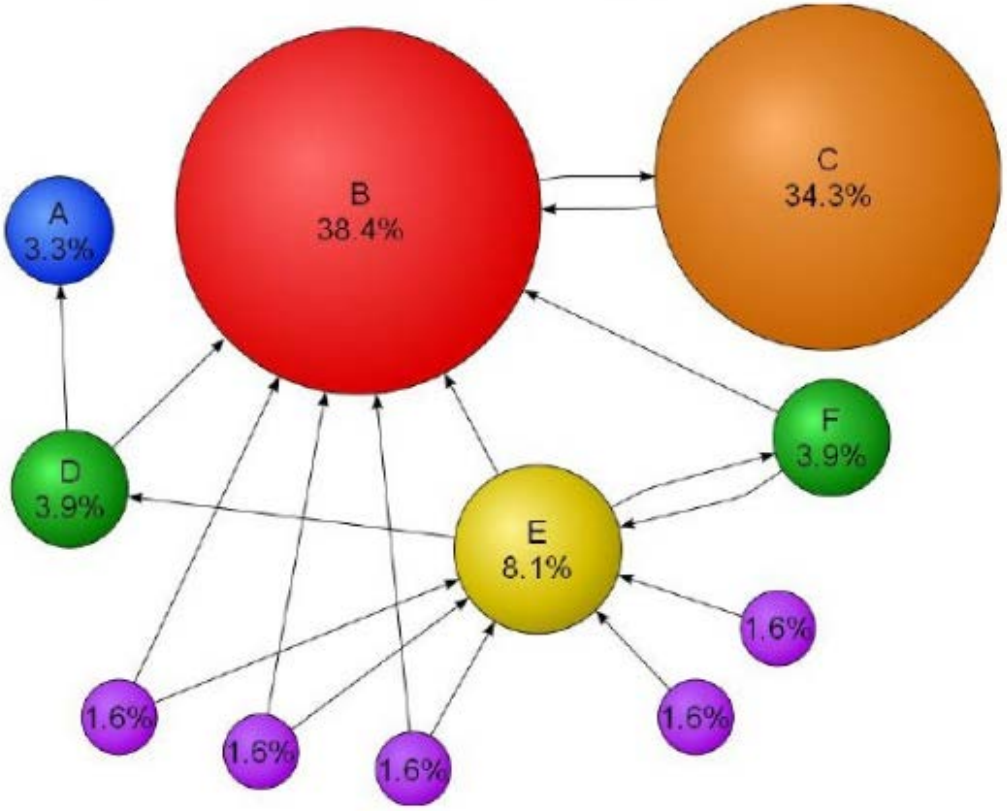


$$S(V_i) = (1-d) + d * \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j)$$

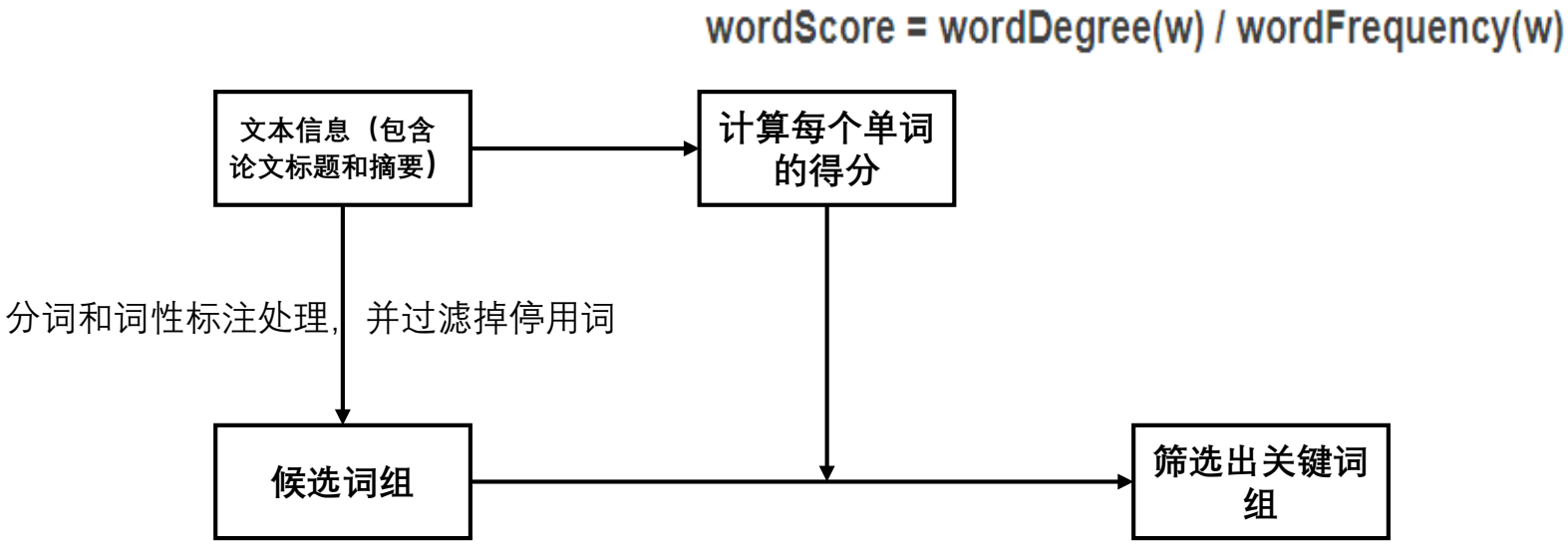
V_i: 研究的某一个网页
$In(V_i)$: 链接到V_i的网页

$S(V_i)$: 网页的PR值
Out(V_i): 网页V_i的链接出的数量

$$Pagerank(B) > PageRank(C) > PageRank(A)$$



RAKE （Rapid Automatic keyword extraction）



wordDegree = its co-occurance with other words in the text

wordFrequency = the frequency of word appearance

Compatibility – systems – linear constraints – set – natural numbers – Criteria – compatibility – system – linear Diophantine equations – strict inequations – nonstrict inequations – Upper bounds – components – minimal set – solutions – algorithms – minimal generating sets – solutions – systems – criteria – corresponding algorithms – constructing – minimal supporting set – solving – systems – systems

	algorithms	bounds	compatibility	components	constraints	constructing	corresponding	criteria	diophantine	equations	generating	inequations	linear	minimal	natural	nonstrict	numbers	set	sets	solving	strict	supporting	system	systems	upper
deg(w)	3	2	2	1	2	1	2	2	3	3	3	4	5	8	2	2	2	6	3	1	2	3	1	4	2
freq(w)	2	1	2	1	1	1	1	2	1	1	1	2	2	3	1	1	1	3	1	1	1	1	1	4	1
deg(w) / freq(w)	1.5	2	1	1	2	1	2	1	3	3	3	2	2.5	2.7	2	2	2	2	3	1	2	3	1	1	2

Figure 1.4 Word scores calculated from the word co-occurrence graph.

	algorithms	bounds	compatibility	components	constraints	constructing	corresponding	criteria	diophantine	equations	generating	inequations	linear	minimal	natural	nonstrict	numbers	set	sets	solving	strict	supporting	system	systems	upper
deg(w)	3	2	2	1	2	1	2	2	3	3	3	4	5	8	2	2	2	6	3	1	2	3	1	4	2
freq(w)	2	1	2	1	1	1	1	2	1	1	1	2	2	3	1	1	1	3	1	1	1	1	1	4	1
deg(w) / freq(w)	1.5	2	1	1	2	1	2	1	3	3	3	2	2.5	2.7	2	2	2	2	3	1	2	3	1	1	2

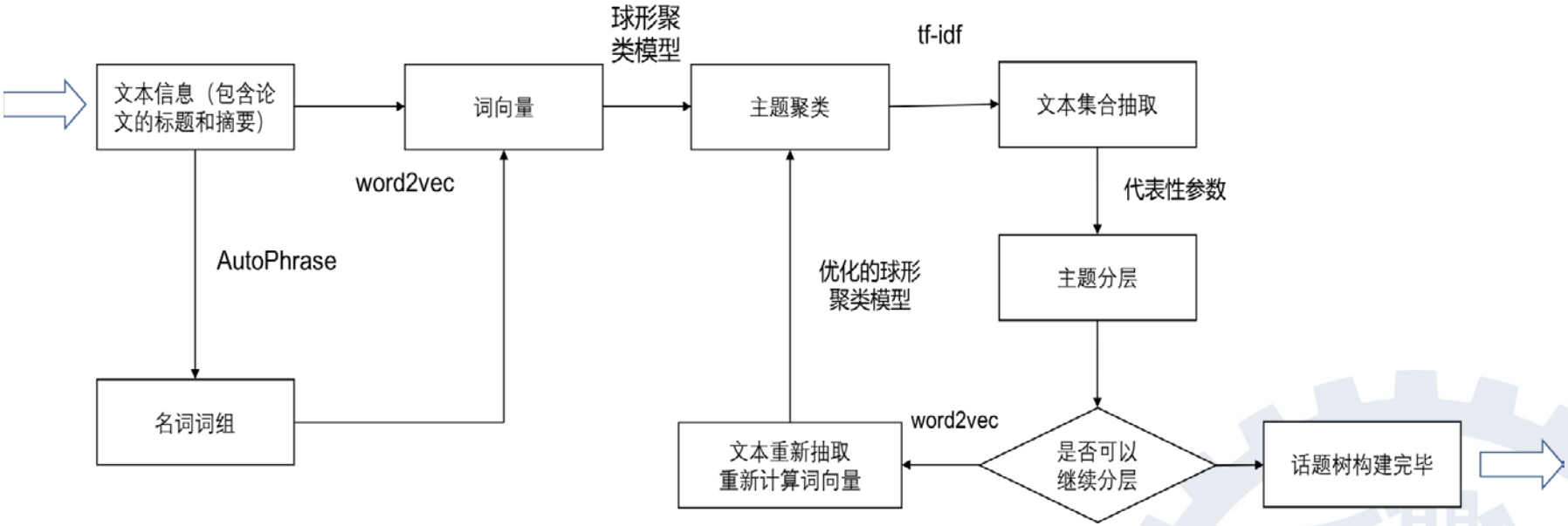
Figure 1.4 Word scores calculated from the word co-occurrence graph.

Each candidate keyword score = the sum of the member scores

minimal generating sets (8.7), linear diophantine equations (8.5), minimal supporting set (7.7), minimal set (4.7), linear constraints (4.5), natural numbers (4), strict inequations (4), nonstrict inequations (4), upper bounds (4), corresponding algorithms (3.5), set (2), algorithms (1.5), compatibility (1), systems (1), criteria (1), system (1), components (1),constructing (1), solving (1)

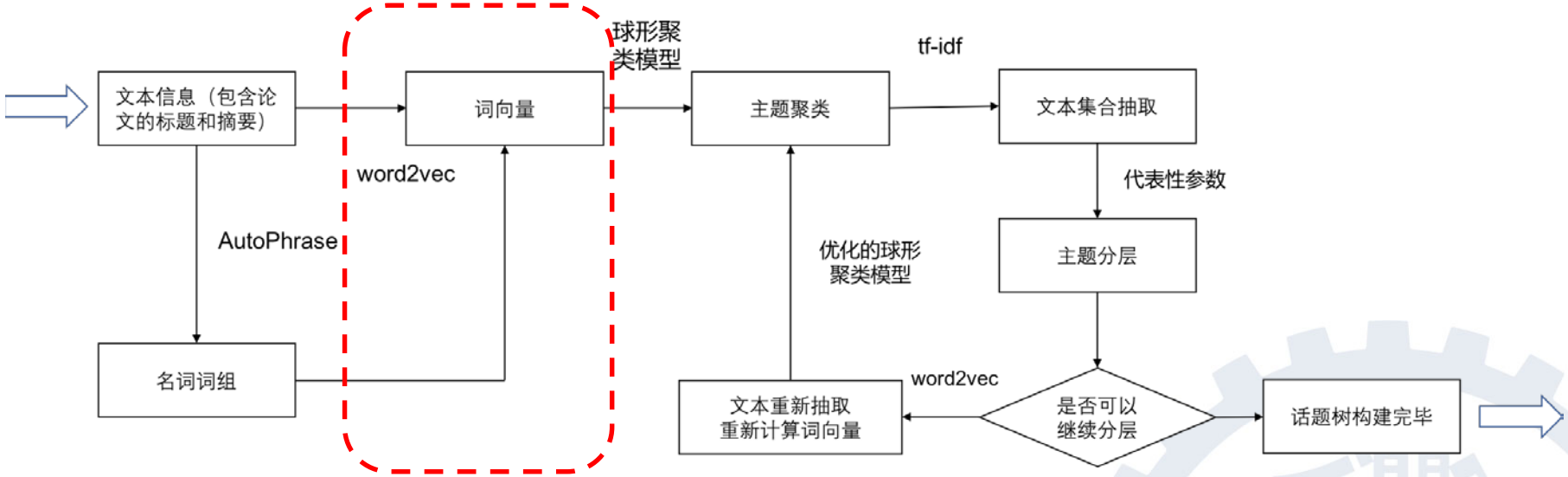
Figure 1.5 Candidate keywords and their calculated scores.

Machine Learning



分词——词向量——聚类

词向量表示


文本信息（包含论文的标题和摘要）
词向量
球形聚类模型
主题聚类
tf-idf
文本集合抽取
代表性参数
主题分层
word2vec
AutoPhrase
名词词组
优化的球形聚类模型
文本重新抽取
重新计算词向量
word2vec
是否可以继续分层
话题树构建完毕

Word2Vec

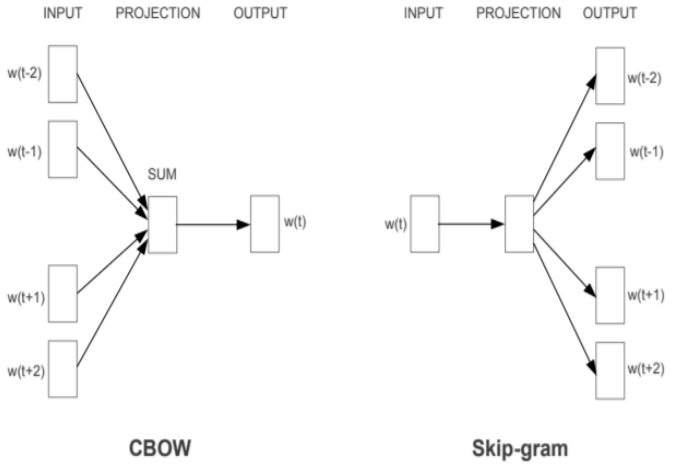
用神经网络模型训练得到词向量空间

$$\prod_{w \in C} p(w|Context(w))$$

- C是语料库(Corpus)
- $Context(w)$是词w的上下文(Context)

Input: 语料库

Output: 词向量空间

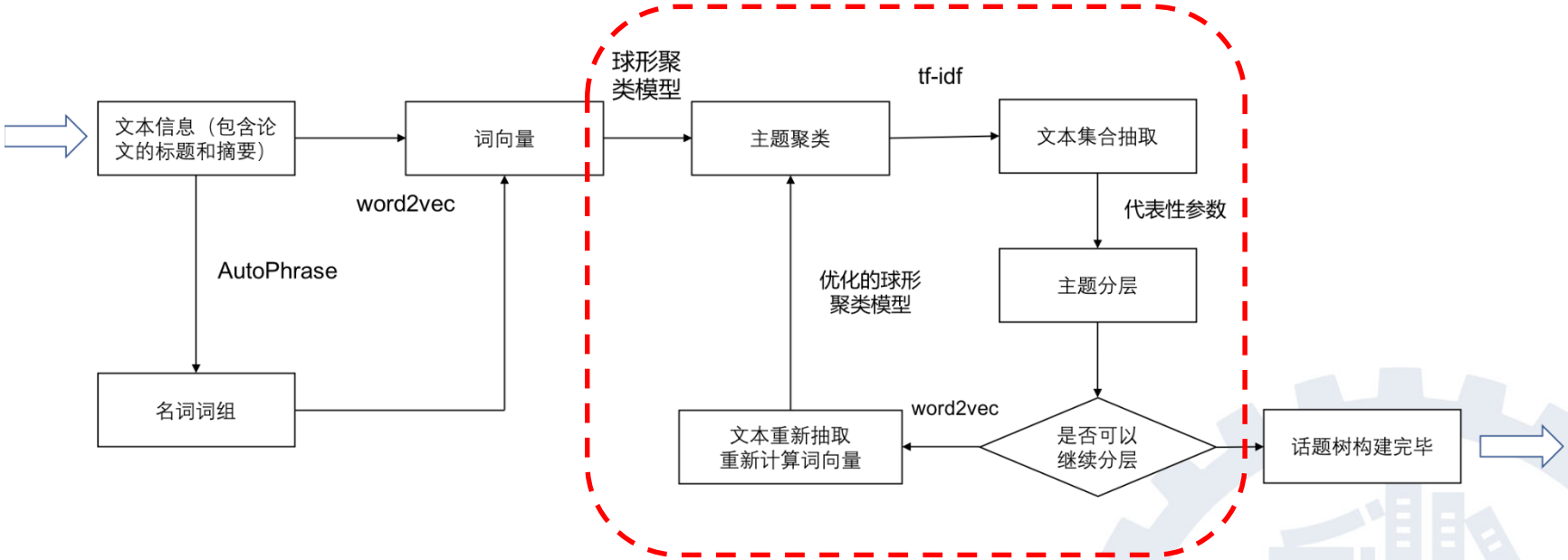


Word2vec常用的两种模型

□i□olo□T, et al. “Efficient estimation of word representations in vector space” IC□□ 2□13□

□i□olo□T, et al. “Distributed representations of words and phrases and their compositionality” □I□S 2□13□

聚类


文本信息（包含论文的标题和摘要）
词向量
word2vec
AutoPhrase
名词词组
球形聚类模型
主题聚类
tf-idf
文本集合抽取
代表性参数
主题分层
优化的球形聚类模型
word2vec
文本重新抽取
重新计算词向量
是否可以继续分层
话题树构建完毕

k-means

一种基于距离的聚类算法

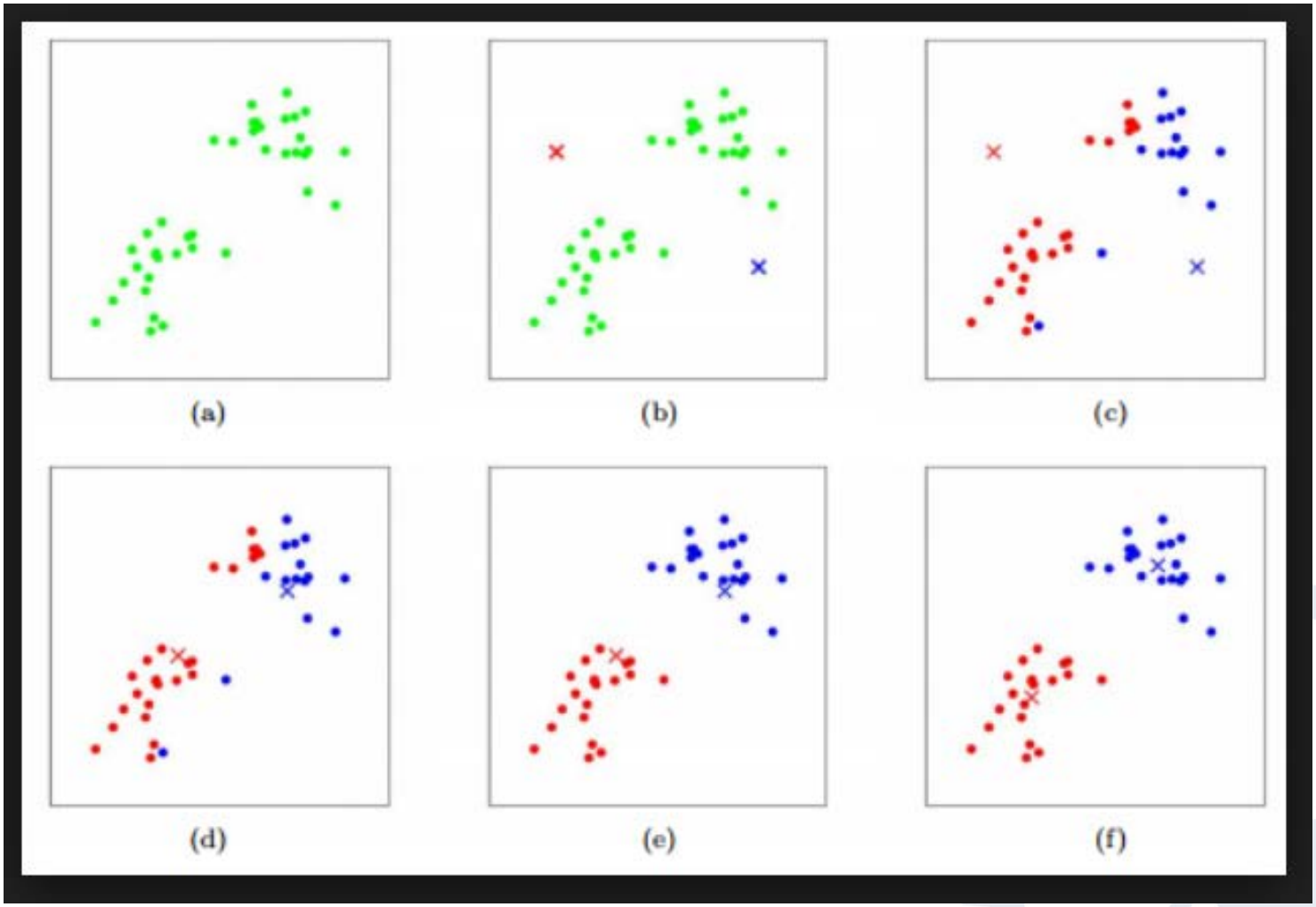
算法流程（以k = 2为例）

1. 初始化两个中心点
2. 分别计算每个分词到两个中心点之间的距离
（对于词向量，一般采用余弦距离）

$$\cos(\theta) = \frac{a^T * b}{|a||b|}$$

3.根据距离把分词集合分为两个簇，计算每个簇中样本点的均值为新的中心点。

4.反复迭代，直到达到终止条件


(a)
(b)
(c)
(d)
(e)
(f)

DBSCAN

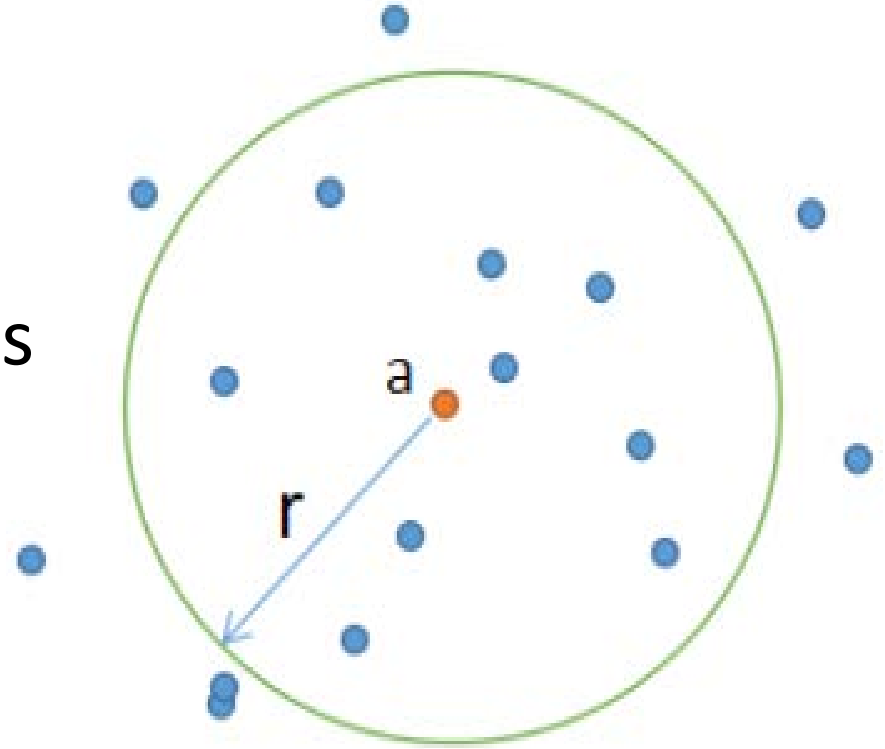
一种基于密度的聚类算法

Density defination

① core points: >MinPts

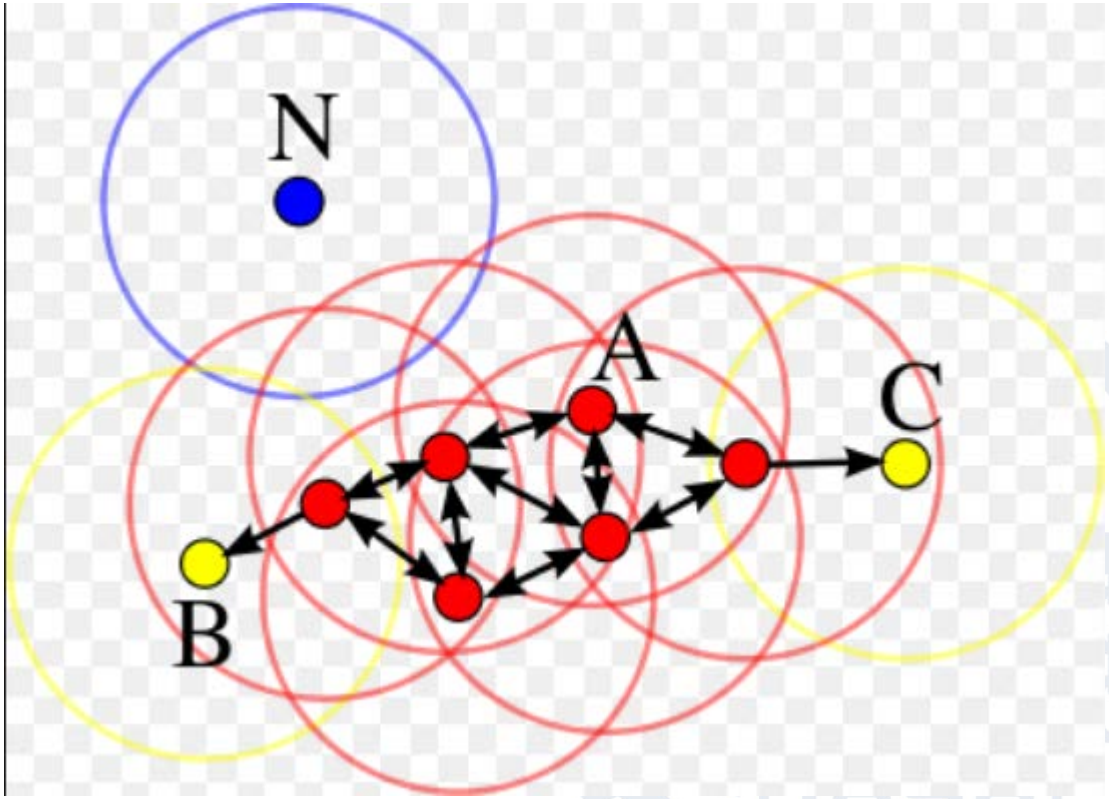
② boundary points: <MinPts

③ noise points: neither nor

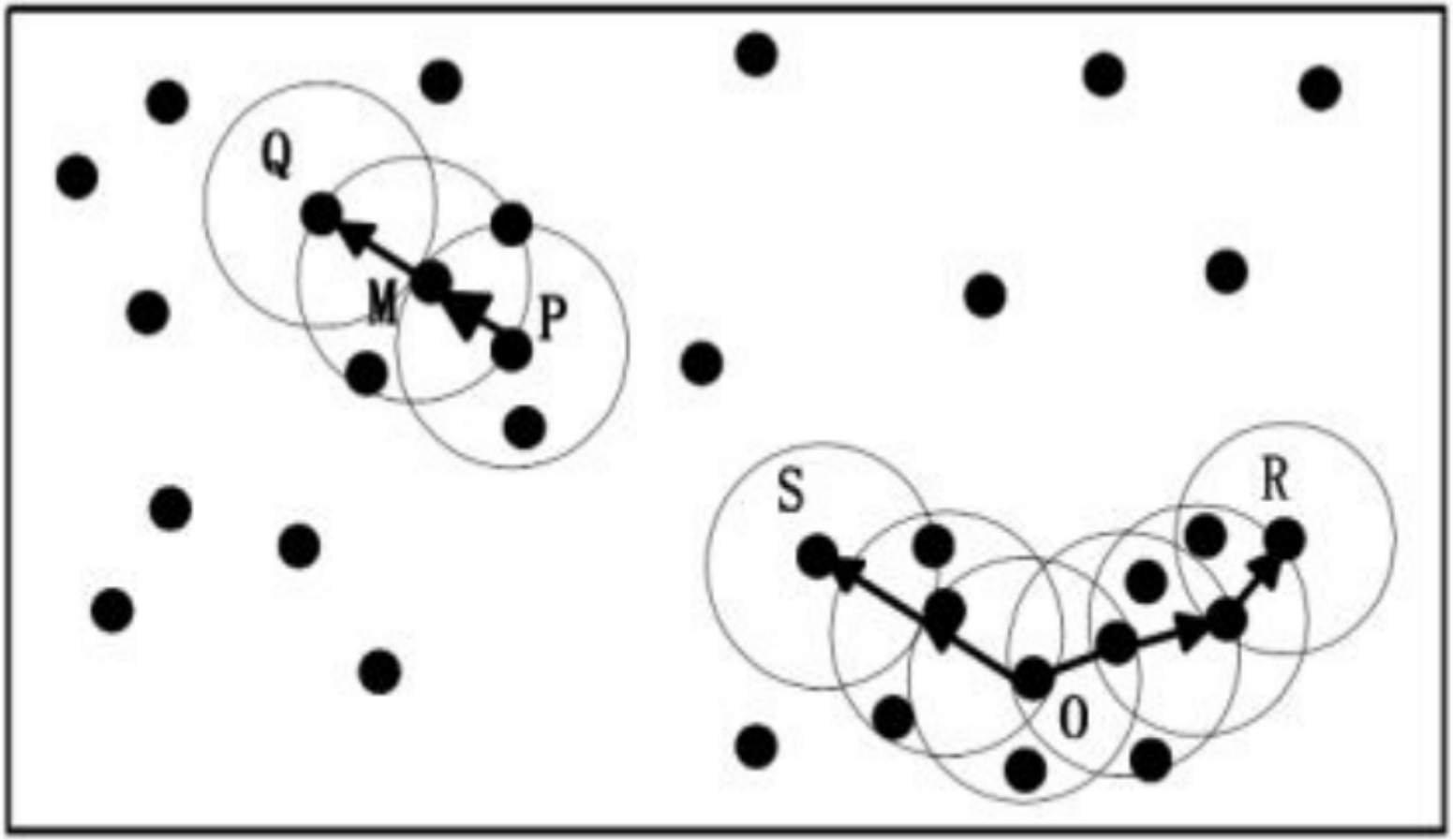


红色节点 ： core points
黄色节点： boundary points
蓝色节点： noise points

minPts = 4



直接密度可达：p在核心节点q的r邻域内



Thanks for listening!