Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心


# 无监督关键词提取方法介绍

林朝坤　硕士研究生

2020年03月15日

# 内容提要

- 背景简介
- 基本概念
- 算法原理
- 优劣分析
- 应用总结
- 参考文献

# 预期收获

- 预期收获
  - 了解关键词提取相关概念
  - 了解一些无监督的关键词提取方法
  - 了解无监督关键词提取方法的应用

# 背景介绍

# 背景介绍

- 什么是关键词及关键词提取技术?
  - 关键词是表达文档主题意义的最小单位，关键词抽取是一种识别有意义且具有代表性文本片段或词汇的技术，自动关键词提取是通过计算机程序从文档中自动抽取具有重要性和主题性的词或短语的自动化技术。

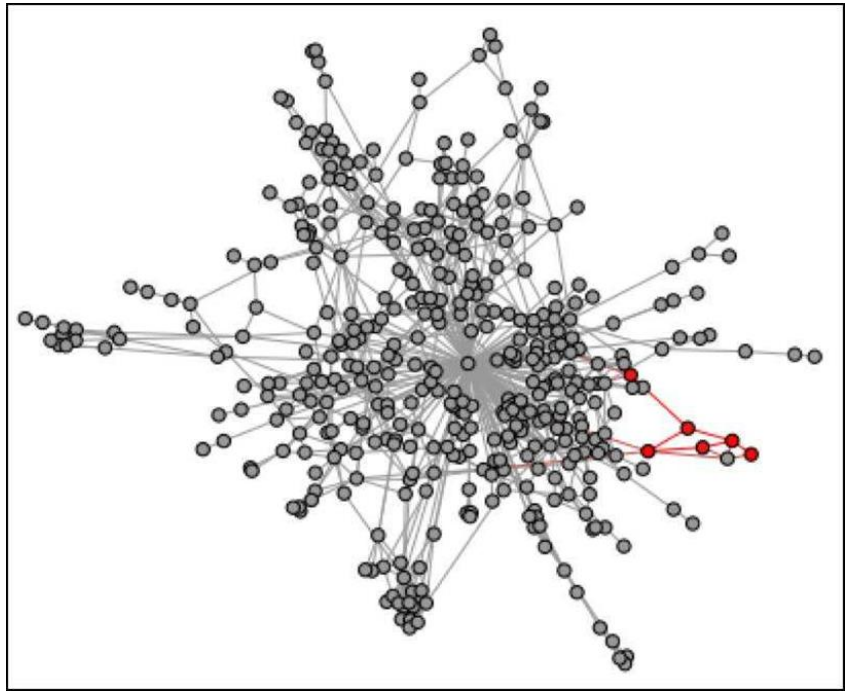
# 背景介绍

- 关键词提取技术起源
  - 美国 IBM 公司的卢恩提出了基于词频统计的文献自动标引方法，标志着关键词提取研究和应用的开始。埃德蒙森开启了用计算机程序进行文献自动检索的实用化时代，设计并实现了最早的自动关键词提取系统。
- 关键词提取技术分类
  - 有监督方法
  - 无监督方法

# 基本概念

# 基本概念

- 有监督学习：训练样本都有标签。
- 无监督学习：训练样本都没有标签。
- 随机游走：网络图上的随机游走(random walk)是指给定图和出发点，随机地选择邻居节点，移动到邻居节点上，然后把当前节点作为出发点，重复以上过程。那些被随机选出的节点序列就构成了一个在图上的随机游走过程。



# 基本概念

- 复杂网络的建模采用图论的方法，表示成 G=(V，E，W)，其中，V 代表节点，E代表边，W 代表权重。通过研究表明：人类语言也是一种复杂网络，具有复杂网络的小世界特性与无标度特性。定义单词为网络的节点，词与词之间的共现关系为边。
- 有向语言网络图反映了元素之间的某种关系，无向语言网络图不考虑这种关系，加权语言网络图量化元素之间的联系强度，无权语言网络图不考虑联系强度，实际应用中的语言网络图可能是 4 种形式的组合。

# 算法原理

# 算法原理

| | |
|---|---|
| T | 提取一篇文章或者一段文本中的关键词 |
| I | 一篇文章或者一段文本 |
| P | 关键词提取技术 |
| O | 一篇文章或者一段文本中的关键词 |

# 算法原理

- **无监督关键词提取方法不需要人工标注的语料，利用某些方法发现文本中比较重要的词作为关键词，进行关键词抽取。**

# 算法原理

- 无监督关键词提取算法可以分为三大类：
  - 基于统计特征的关键词提取
  - 基于主题模型的关键词提取
  - 基于词图模型的关键词提取

# 算法原理

- 基于统计特征的关键词提取算法的思想是利用文档中词语的统计信息抽取文档的关键词。基于统计特征的关键词抽取方法的关键是采用什么样的特征值量化指标，目前常用的有三类：
  - 基于词权重(term weight）
  - 基于词的文档位置(term location)
  - 基于词的关联信息(term collocation)

# 算法原理

- 基于词权重的特征量化：主要包括词性、词频、逆向文档频率、相对词频、词长等。
- 基于词的文档位置的特征量化：这种特征量化方式是根据文章不同位置的句子对文档的重要性不同的假设来进行的。
- 基于词的关联信息的特征量化：词的关联信息是指词与词、词与文档的关联程度信息。
- 优点：基于统计特征的关键词抽取算法方法具有模型可泛化性，易于实现，具有语言和领域的独立性，关注统计特性。主流的简单统计方法是TFIDF(term frequency-inverse document frequency)。

# 算法原理

- TFIDF 算法是由 Salton在提出的，用于评估一个词对一个文档集或一个语料库中的其中一份文档的重要程度。
  - TF 称为词频，用于计算该词描述文档内容的能力。
  - IDF 称为逆文档频率，用于计算该词区分文档的能力。

$$tf_i = \frac{n_{ij}}{\sum_k n_{kj}} \tag{1}$$

$$idf_i = \log \frac{|D|}{|D_i|+1} \tag{2}$$

$$w_{ij}=tfidf_{ij}=tf_{ij}\times idf_i \tag{3}$$

$$w_{ij} = \frac{w_{ij}}{\sqrt{\sum_k w_{kj}}} \tag{4}$$

其中，$n_{ij}$ 是词$t_i$ 在文档 $d_j$中的出现次数，|D|表示语料库中的文档总数，|$D_i$|表示语料库中包含词 $t_i$ 的文档总数， $w_{ij}$表示词 $t_i$ 在文档 $d_j$ 中的权重(归一化)。

# 算法原理

- TFIDF 算法的优点：简单快速，结果比较符合实际情况。
- TFIDF 算法的缺点：
  - 单纯以词频衡量一个词的重要性，不够全面，有时重要的词可能出现次数并不多。
  - 这种算法无法体现词的位置、词性和关联信息等特征，更无法反映词汇的语义信息。
  - 本质上，IDF 是一种试图抑制噪音的加权，并且单纯地认为文档频率小的单词就越重要，文档频率大的单词就越无用。这对于大部分文本信息，并不是完全正确的。
- TFIDF 改进算法： TF*IWF*IWF、 TFIPNDF、 NEG-TFIDF等。

# 算法原理

- 基于主题的关键词提取方法-PLSA模型认为：一篇文档可以由多个主题混合而成，而每个主题都是词汇上的概率分布。每篇文档中的每个词都是这样一个过程得到的：“通过概率选取出某个主题，并从该主题中通过概率选取某个词语”。一篇文档的生成，其中每个词语出现的概率计算公式如下：

$$P(\text{词语}/\text{文档}) = \sum_{\text{主题}} P(\text{词语}/\text{主题}) \times P(\text{主题}/\text{文档})$$

## PLSA模型的文档生成过程：

**Game 11** PLSA Topic Model

1: 上帝有两种类型的骰子，一类是doc-topic 骰子,每个doc-topic 骰子有$K$ 个面，每个面是一个topic 的编号；一类是topic-word 骰子，每个topic-word 骰子有$V$ 个面，每个面对应一个词；

doc-topic

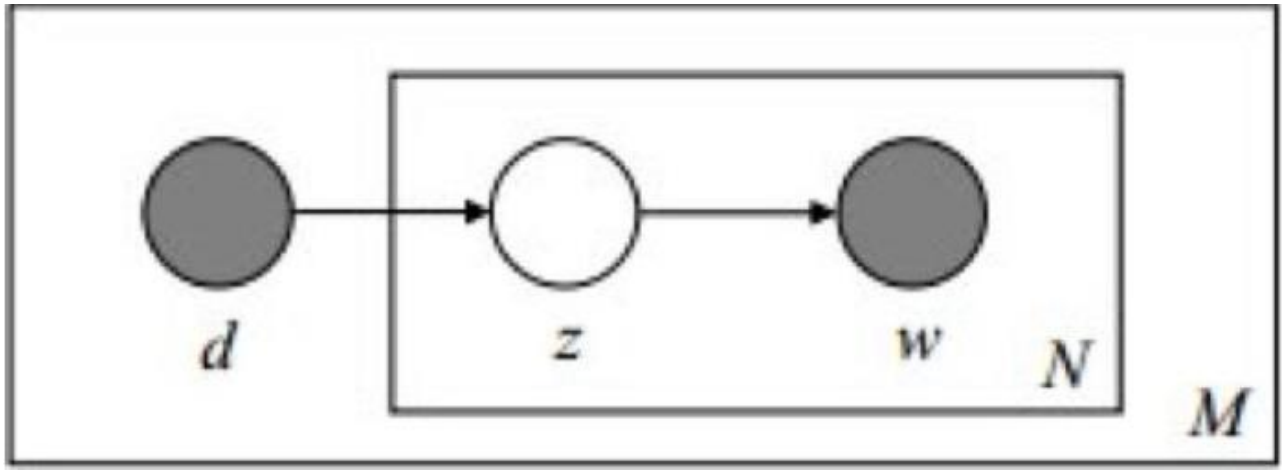
topic-word

2: 上帝一共有$K$ 个topic-word 骰子, 每个骰子有一个编号，编号从1 到$K$；

3: 生成每篇文档之前，上帝都先为这篇文章制造一个特定的doc-topic 骰子，然后重复如下过程生成文档中的词

- 投掷这个doc-topic 骰子,得到一个topic 编号$z$
- 选择$K$ 个topic-word 骰子中编号为$z$的那个，投掷这个骰子，于是得到一个词

# 算法原理




PLSA模型的参数显而易见就是： MxK个$p(z_x|d_m)$ 、KxV个$p(w_n|z_k)$ 。

$p(z_x|d_m)$ 表征的是给定文档在各个主题下的分布情况，文档在全部主题上服从多项式分布(共M个)， $p(w_n|z_k)$则表征给定主题的词语分布情况，主题在全部词语上服从多项式分布(共K个，V是主题下词汇的个数)。

# 算法原理

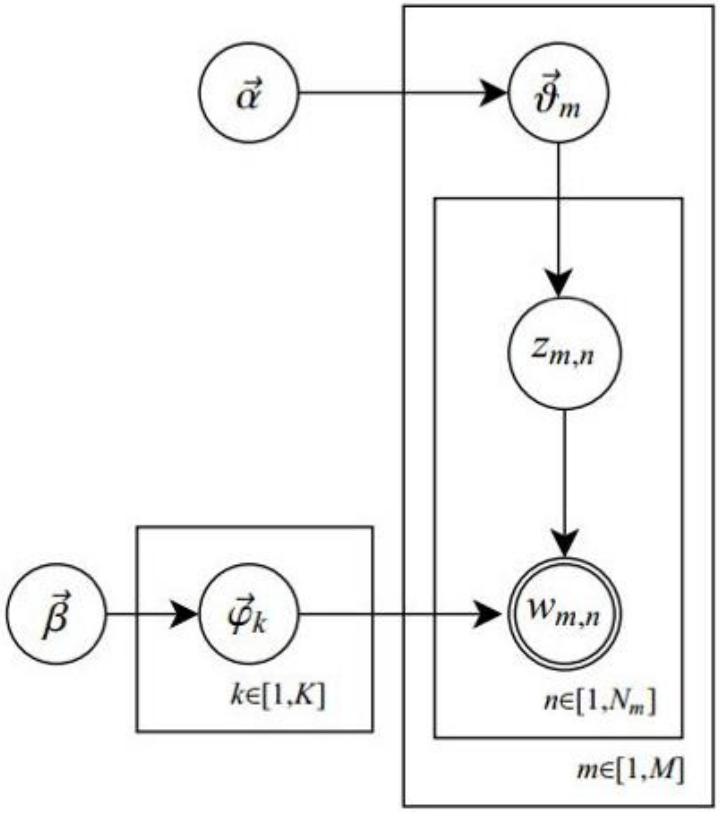
- PLSA缺点：在文档对应主题的概率分布的计算上没有使用统一的概率主题模型，参数较多会导致过拟合现象，对训练数据集以外的文档的概率分布计算会比较难。
- PLSA的改进算法：LDA(latent dirichlet allocation)模型。

# 算法原理

- LDA模型的创始者 Blei在 PLSA 的基础上加入了 Dirichlet 先验分布，是 PLSA 的突破性的延伸。

$$P(\text{词语}/\text{文档}) = \sum_{\text{主题}} P(\text{词语}/\text{主题}) \times P(\text{主题}/\text{文档})$$

# 算法原理




- LDA 生成文档的两个阶段：

(1) $\alpha \rightarrow \theta_m \rightarrow z_{m,n}$：生成文档的主题$z_{m,n}$；

(2) $\beta \rightarrow \varphi_k \rightarrow w_m, {}_n|k=z_{m,n}$：根据文档的主题生成词$w_{m,n}$。

参数α和β是由经验给出的，$z_{m,n}$，$\theta_m$，$\varphi_k$ 是隐含变量，需要推导。

# 算法原理

- PLSA模型与LDA模型的对比分析：
  - 同：都认为文档是主题的概率分布，而主题是词的概率分布。
  - 异：
    - PLSA认为每篇文档的文档-主题分布是确定的，而LDA认为每篇文档的文档-主题分布不是确定的，并且服从Dirichlet的先验分布。
    - PLSA模型的参数为M个文档-主题分布和K个主题词汇分布（M问文本集中的文档数，K为文本集中的主题个数）；
    - LDA模型的参数为α、β以及K个主题-词汇分布，比PLSA参数少。

# 算法原理

- 主题模型的优点：
  - 可以获得文本语义相似性的关系。
  - 可以解决多义词的问题。
  - 语言无关。
- 主题模型的缺点：
  - 需要事先设置主题数K。
  - 训练人员需要根据训练出来的结果，手动调参。

# 算法原理

- 基于复杂网络的关键词抽取，首先要构建文档的语言网络图，然后对语言网络图进行分析，在整个语言网络图上寻找起重要作用和中心作用的词或短语，将这些词或短语抽取出来作为关键词。根据特征词之间的连接方式，语言网络图的主要形式如下图所示：

# 算法原理

- 自动关键词抽取过程中，文档的语言网络图构建需先将文档进行预处理，然后以特征词为节点，以特征词之间的关系作为边，可以是无权的，如果是有权语言网络图，关系程度作为边的权重，可以是有向，也可以无向的。
- 网络中节点的重要性评估方法：综合特征值法、系统科学法和随机游走法。

# 算法原理

- 综合特征法也叫社会网络中心性分析方法，这种方法的核心思想是节点中重要性等于节点的显著性，以不破坏网络的整体性为基础。此方法就是从网络的局部属性和全局属性角度去定量分析网络结构的拓扑性。
- 常用的定量计指标有：度、强度、中介性、接近性、特征向量、平均最短路径等。

# 算法原理

- 系统科学法主要基于节点(集)的删除，核心思想是重要性等价于该节点(集)被删除后对网络的破坏性，对网络中重要节点的挖掘是通过节点(集)删除前后网络连通性、性能的变化来反映的。

# 算法原理

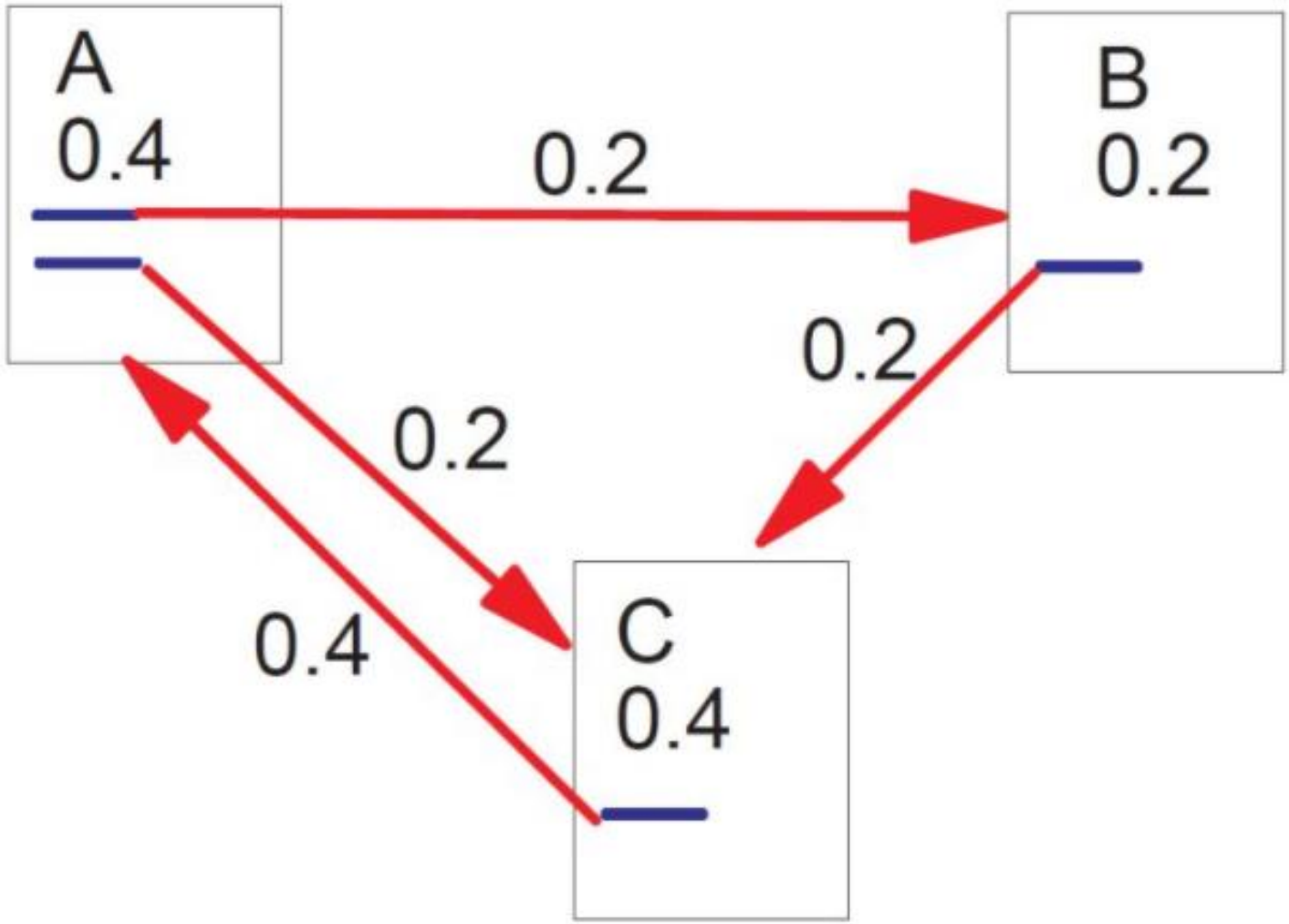
- 著名的PageRank 算法采用了随机游走思想进行互联网网页的重要性计算，该算法不仅在搜索引擎中扮演重要的作用，在文本处理领域也具有很广泛的应用。计算方法见如下公式：

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

$S(V_i)$ 表示节点 $V_i$的 PageRank 值，d 为阻尼系数，一般取值为0.85。

$In(V_i)$表示是指向网页i的网页集合， $|Out(V_j)|$是指集合中元素的个数。

# 算法原理



# 算法原理

- 基于网络图的关键词提取算法-TextRank在构建的是共现网络图，该模型认为：
  - 如果一个单词出现在很多单词后面的话，那么说明这个单词比较重要。
  - 一个TextRank值很高的单词后面跟着的一个单词，那么这个单词的TextRank值会相应地因此而提高。
- TextRank节点重要性计算公式如下:

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

公式中的$w_{ij}$为图中节点$V_i$和的边 $V_j$的权重。其他符号与PageRank公式相同。

# 算法原理

- TextRank模型优点：
  - 无需训练数据，节省了大量成本。
  - 适应性强。
  - 速度快。
- TextRank模型缺点：
  - 忽略了词汇的语义相关性，也未考虑上下文及辅助信息。
- 针对这些问题，TextRank改进的算法：TextRank-CM、TextRank+TF-IDF等。

# 应用总结

- 算法的应用领域
  - 图书馆学
  - 情报学
  - 自然语言处理领域
  - 数据挖掘领域
- 未来发展
  - 多种方法的有效融合
  - 结合语义的方法

# 参考资料


- 赵京胜, 朱巧明, 周国栋, 等. 自动关键词抽取研究综述[J]. 软件学报, 2017, 28(9): 2431-2449.
- 刘知远. 基于文档主题结构的关键词抽取方法研究[D]. 北京: 清华大学, 2011.
- 文本关键词提取：
  - https://zhuanlan.zhihu.com/p/33605700
- PankRank与TextRank：
  - https://www.cnblogs.com/motohq/p/11887420.html

# 道德经

大成若缺，其用不弊。
大盈若冲，其用不穷。
大直若屈。大巧若拙。
大辩若讷。静胜躁，寒
胜热。清静为天下正。

谢谢！