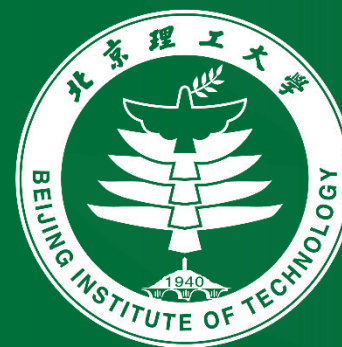


关键词提取

指导教师：张华平

汇报人：王轩、赵亚祥、张墨言、陈宇飞、许子逸



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY



目录

- 一、简介与发展过程
- 二、原理与关键技术
- 三、经典算法详解
- 四、实验及实例
- 五、未来发展方向分析



关键词提取简介及 发展历史



I'm in love with the app! It's amazing!! The mobile version works just as well as the web version. You can create pages and control how your content is displayed very easily as the app has very intuitive and simple controls

TAG	VALUE
KEYWORD	love
KEYWORD	app
KEYWORD	mobile version
KEYWORD	web version
KEYWORD	page
KEYWORD	content
KEYWORD	simple control

- 关键词通常为一个或多个能够描述文档主题信息的词语或词组

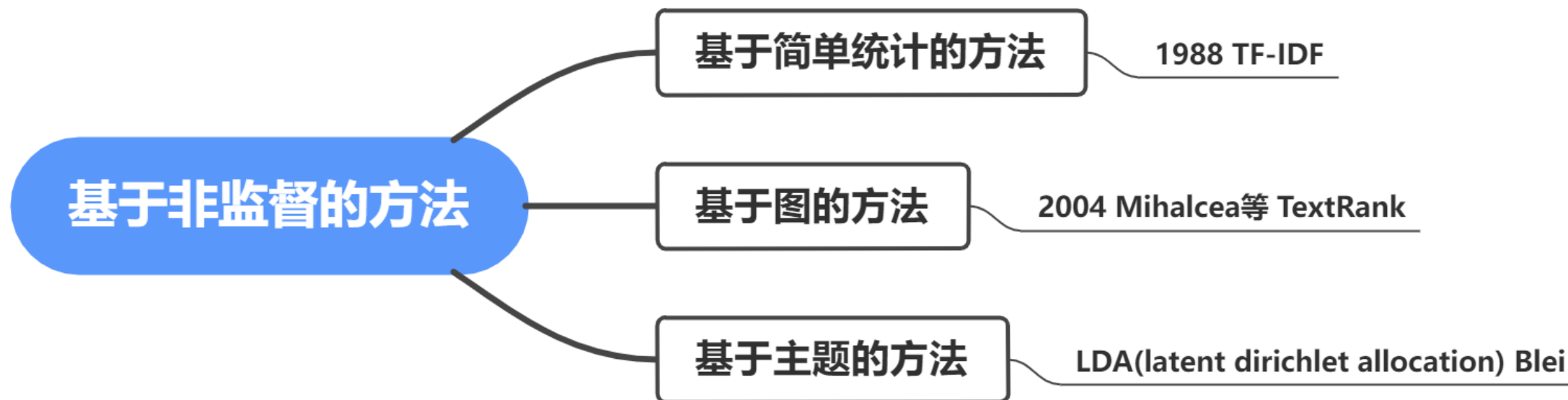


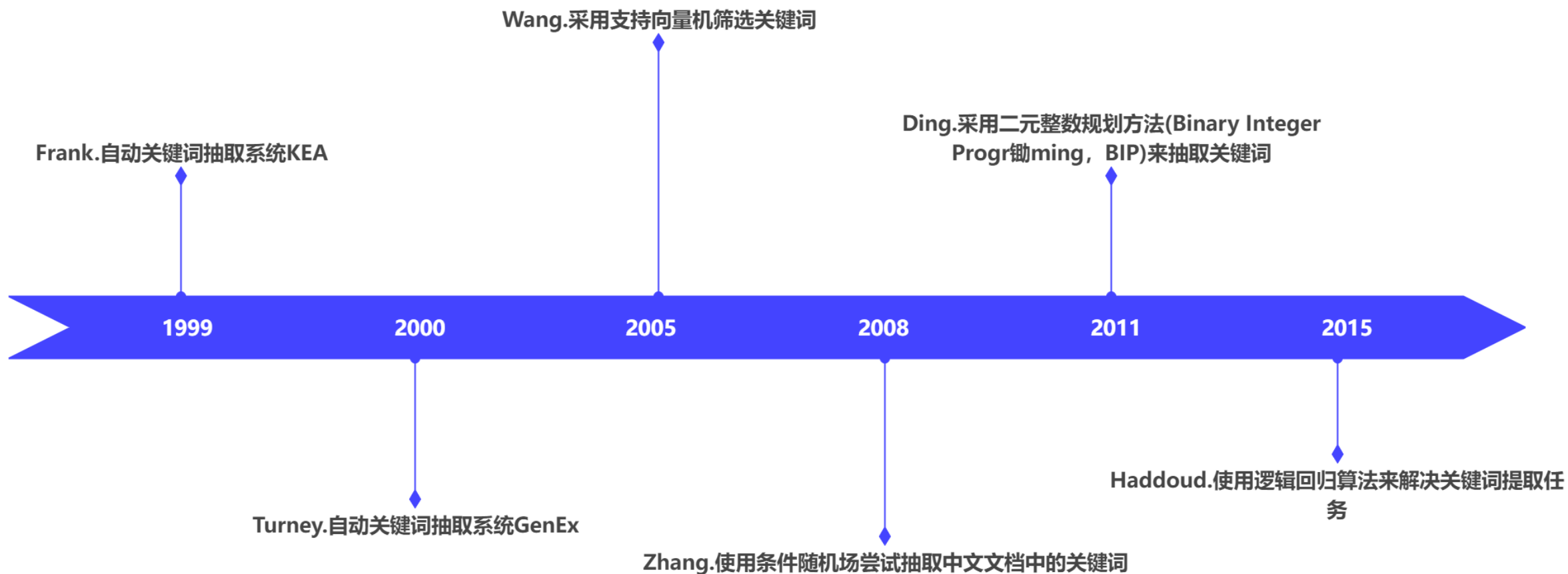
关键词抽取技术(Extmctive Keyphrase Extraction)

从文档中筛选得到能表达文档主题的单词(词组), 该关键词必然在文档中出现

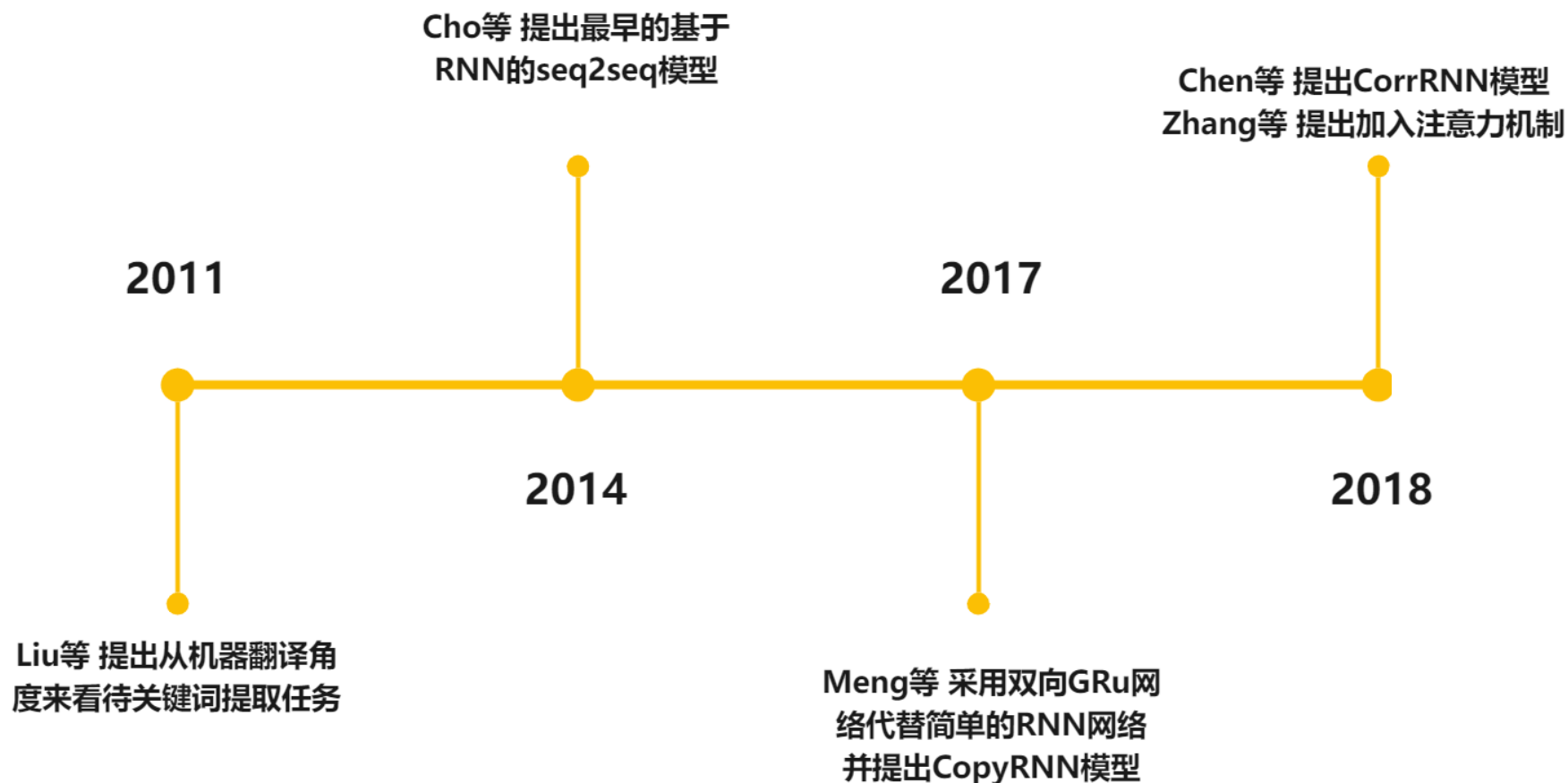
关键词生成技术(Abstractive Keyphmse Genemtion)

从词表中选择与文档主题相近的单词(词组)作为该文档的关键词, 与该关键词是否在文档中出现无关。











原理与关键技术

汇报人：赵亚祥



从宏观上划分，关键词提取方法分为3种：

1. 有监督方法：

将关键词提取看做是二元分类问题，判断文档中的词或短语是或不是关键词。这种方法必须提供已经标注关键词的训练语料：首先，利用训练语料训练关键词抽取模型；然后，利用得到的模型对需要抽取关键词的文档进行关键词抽取；

2. 半监督方法：

这种方法不像前者需要大量的训练数据，只需要少量的训练语料，利用这些语料训练抽取模型，利用模型进行未标注文本的关键词提取，人工对提取结果进行甄别，将正确的标注加到训练语料中再训练模型。

3. 无监督方法：

这种方法不需要训练语料，也不需人工参与，利用提取系统完成文档或文档集合的关键词提取。



基于有监督学习的关键词抽取的一般步骤是:

- 1) 首先, 建立一个包含大量文本并标出关键词的训练集合;
- 2) 然后, 利用训练集合对分类或标注算法进行训练得到一个模型;
- 3) 最后, 应用训练好的模型对新文本进行关键词抽取;

有监督机器学习的分类或标注方法常借助:

决策树(DT)、朴素贝叶斯(NB)、支持向量机(SVM)、最大熵模型(ME)、隐Markov模型(HMM)、条件随机场(CRF)模型等.



在有监督方法中，主要有两个研究方向：

- 一个方向是将关键词提取看做是二分类任务，即，判断文档中的一个词是关键词或不是关键词。这个方向的研究主要是基于一些特征建立抽取关键词的分类器。例如：

- 1) 关键词提取系统**GenEx**，该系统通过使用词性的频率和词性信息作为特征，使用决策树算法作为分类器；（Turney PD. Learning algorithms for keyphrase extraction. Information Retrieval Journal, 2000, 2(4): 303–336. [[Learning Algorithms for Keyphrase Extraction | SpringerLink](#)]

- 2) 关键词提取系统**KEA**，使用朴素贝叶斯方法构造分类器；（Frank E, Paynter GW, Witten IH, Gutwin C, Nevill-Manning CG. Domain-Specific keyphrase extraction. In:Dean T, ed. Proc. of the 16th Int'l Joint Conf. on Artificial Intelligence, ACM CIKM Int'l Conf. on Information & Knowledge Management. 1999. 668-673. [[9902007.pdf \(arxiv.org\)](#)]

GenEx和KEA，奠定了自动关键词提取的有监督方法，已经成为后续改进方法和其他关键词提取系统的参照基准系统

- 另外一个方向是基于语言模型的, 研究人员从训练集中建立语言模型, 并选择出符合关键词特征的模型.

例如:

- 1) 利用词性标注、名词短语块等作为特征进行关键词抽取的方法; (Hulth A. Improved automatic keyword extraction given more linguistic knowledge. In:Collins M, ed. Proc. of the. Conf. on Empirical Methods in Natural Language Processing (EMNLP). Sapporo, 2003. 216-223.[doi:[10.3115/1119355.1119383](https://doi.org/10.3115/1119355.1119383)])
- 2) KPSpotter系统, 利用词性标注、信息增益、词位置等作为特征进行自动关键词抽取; (Song M, Song IY, Hu X. KPSpotter:A flexible information gain-based keyphrase extraction system. In:Chiang R, Laender AHF, Lim EP, eds. Proc. of the 5th ACM Int'l Workshop on Web Information and Data Management. New Orleans, 2003. 50-53.[[KPSpotter | Proceedings of the 5th ACM international workshop on Web information and data management](#)])



一般情况下, 基于有监督学习的算法往往需要建立训练集合, 称为语料库, 主要指由大量实际使用的语言信息组成, 专供分析、描述和研究的语言资料库;

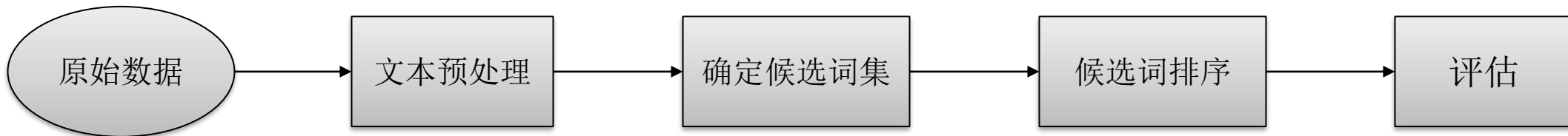
- 20世纪60年代的布朗语料库(Brown corpus)是第一个具有代表性的通用平衡语料库;
- 80年代后, 英国制作了标注语料库(LOB corpus)与语音语料库(birmingham corpus)
- 20世纪90年代出现了一批实用的语料库, 包括日本的EDR语料、NHK新闻稿语、英国国家语料库(BNC);
- 典型的中文语料包括国际语言资源联盟LDC的Chinese Gigaword新闻语料、搜狗实验室(Sogo labs)和数据堂(datatang)的网络语料;
- 特定应用领域的小型语料库;

训练语料的质量往往会直接影响到模型的准确性, 从而影响着关键词抽取的结果. 现已标注关键词的文本有限, 训练集要自己去标注, 人工标注带有一定的主观因素, 会造成实验数据具有不真实性;

关键词提取的半监督方法研究和应用较少

- 1) 构建文本语义网、超图, 利用文本标题中出现的词汇进行迭代, 利用维基百科知识进行推理实现关键词抽取的方法; (Li DC, Li SJ. Hypergraph-Based inductive learning for generating implicit key phrases. In:Simpson S, ed. Proc. of the Int'l Conf. on Companion on World Wide Web. New York:ACM Press, 2011. 77-78.[doi:[10.1145/1963192.1963232](https://doi.org/10.1145/1963192.1963232)])
- 2) 构建TPDG(transition probability distribution generator)基准系统, 借助马尔可夫条件转移矩阵进行关键词抽取的方法; (Lynn HM, Choi C, Choi J, Shin J, Kim P. The method of semi-supervised automatic keyword extraction for Web documents using transition probability distribution generator. In:Kim J, ed. Proc. of the Int'l Conf. on Research in Adaptive and Convergent Systems. Odense:ACM Press, 2016. 1-6.[doi:[10.1145/2987386.2987399](https://doi.org/10.1145/2987386.2987399)])

- 出现的较早，种类也最多
- 不需要人工标注
- 近几年研究和应用的重点



无监督学习经常采用的技术手段有：

- 统计法 (TF-IDF)
- 基于主题的方法 (LDA)
- 基于网络图法 (PageRank, TextRank)



方法类型	方法描述	优点	缺点	
无监督学习	简单统计	基于N-gram、TF-IDF、词频、词共线等统计指标抽取关键词	操作简单易行	准确率不高,但在不同数据集上的表现不稳定
	图结构	通过图结构对候选词进行排序,如TextRank、SingleRank、SGRank等	可以体现候选词间的联系	准确率有限,且不适用于短文本
	主题模型	通过主题模型计算候选词的信息量,并以此作为单词重要程度的依据	操作简单	带有较强的主观性,缺少严谨的评价指标
有监督学习	分类模型	传统机器学习 选择特征表示单词并通过模型将其进行区分 深度学习 利用深度学习模型对关键词加以区分	抽取准确率较高	忽略了上下文的语境对候选词的影响
	序列标注模型	传统机器学习 在判断当前单词的标签时会考虑上下文的信息,如CRF 深度学习 利用循环神经网络实现对序列的标注	抽取准确率较高	需要大规模的标注语料支持



与大多数自然语言处理任务一样, 关键词提取过程中对文档进行分析分为两个层次:

- 一种是浅层的语言分析, 包括文档的组成元素和结构信息分析, 组成元素主要涉及语言中的词性、词频等基础语言信息, 结构信息主要指共现、语法等组合语言信息, 通过分析获取候选关键词, 进而得到文档关键词;
- 另一种是深层的语义分析, 因为现实中的实体和概念通常表达成词汇, 如果能让系统理解词义, 分析出其表达的现实实体和概念, 则关键词抽取就会变得非常简单;

虽然近年来提出的方法有上百种, 但是有些关键词抽取系统不只是用单一的方法, 可能是两种或多种方法的组合, 有的系统可能是多个简单系统的融合



方法	技术特点	代表性成果	发展趋势
语言模型法	浅层词法、语法分析、深层语义分析	H.P.Luhn	应用较多，长期研究方向
机器学习法	利用学习算法训练模型、测试、应用	GenEx、KEY	应用较多，长期研究方向
神经网络法	深度学习网络模型	word2word	研究热点
数据挖掘法	利用数据挖掘算法,如聚类、关联规则等	SPSS/clementine	领域性、小范围应用
在线众包法	众包、验证		领域性、小范围应用
查询日志法	结合信息检索的日志（关键词）		领域性、小范围应用
机器翻译法	不同语言之间的对应		领域性、小范围应用
超网络法	文本网络的网络		领域性、小范围应用
协同功能法	与其他文本挖掘任务协同，如文本摘要		长期研究方向
本体法	结合现实世界的实体、概念（本土）		长期研究方向



经典算法介绍



TF-IDF（词频-逆文档频率）是一种用于信息检索与文本挖掘的常用加权技术。

字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。

TF越大越重要

IDF越小越重要

TF

X

IDF

词频 (Term Frequency) 表示词条在文本中出现的频率。

逆文档频率 (Inverse Document Frequency) 表示词条在文本中出现的频率。

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

$$idf_i = \log \frac{|D|}{|\{j: t_i \in d_j\}|}$$

十分简单!



说到TextRank，就不得不提PageRank

情景引入

- 数量假设：**在Web图模型中，如果一个页面节点接收到的其他网页指向的入链数量越多，那么这个页面越重要。
- 质量假设：**指向页面A的入链质量不同，质量高的页面会通过链接向其他页面传递更多的权重。所以越是质量高的页面指向页面A，则页面A越重要。



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

其中, $WS(V_i)$ 表示句子 i 的权重, 右侧的求和表示每个相邻句子对本句子的贡献程度,

在单文档中, 我们可以粗略的认为所有句子都是相邻的, 不需要像多文档一样进行多个窗口

的生成和抽取, 仅需单一文档窗口即可, w_{ji} 表示两个句子的相似度 $WS(V_j)$ 代表上次迭代出

的句子 j 的权重。 d 是阻尼系数, 一般为 0.85。 http://blog.csdn.net/leiqian_bird

The PageRank Citation Ranking: Bringing Order to the Web

隐含狄利克雷分布（Latent Dirichlet Allocation）是一种主题模型，它可以将文档集中每篇文档的主题按照概率分布的形式给出。首先由David M. Blei、Andrew Ng和Michael I. Jordan于2003年提出。



[中英字幕]吴恩达机器学习系列课程

312.5万播放 · 总弹幕数4.7万 2019-04-28 18:08:23



岁月殇 发消息

+ 关注 2.9万

弹幕列表

展开



新服震撼开启，注册领取海量福利！

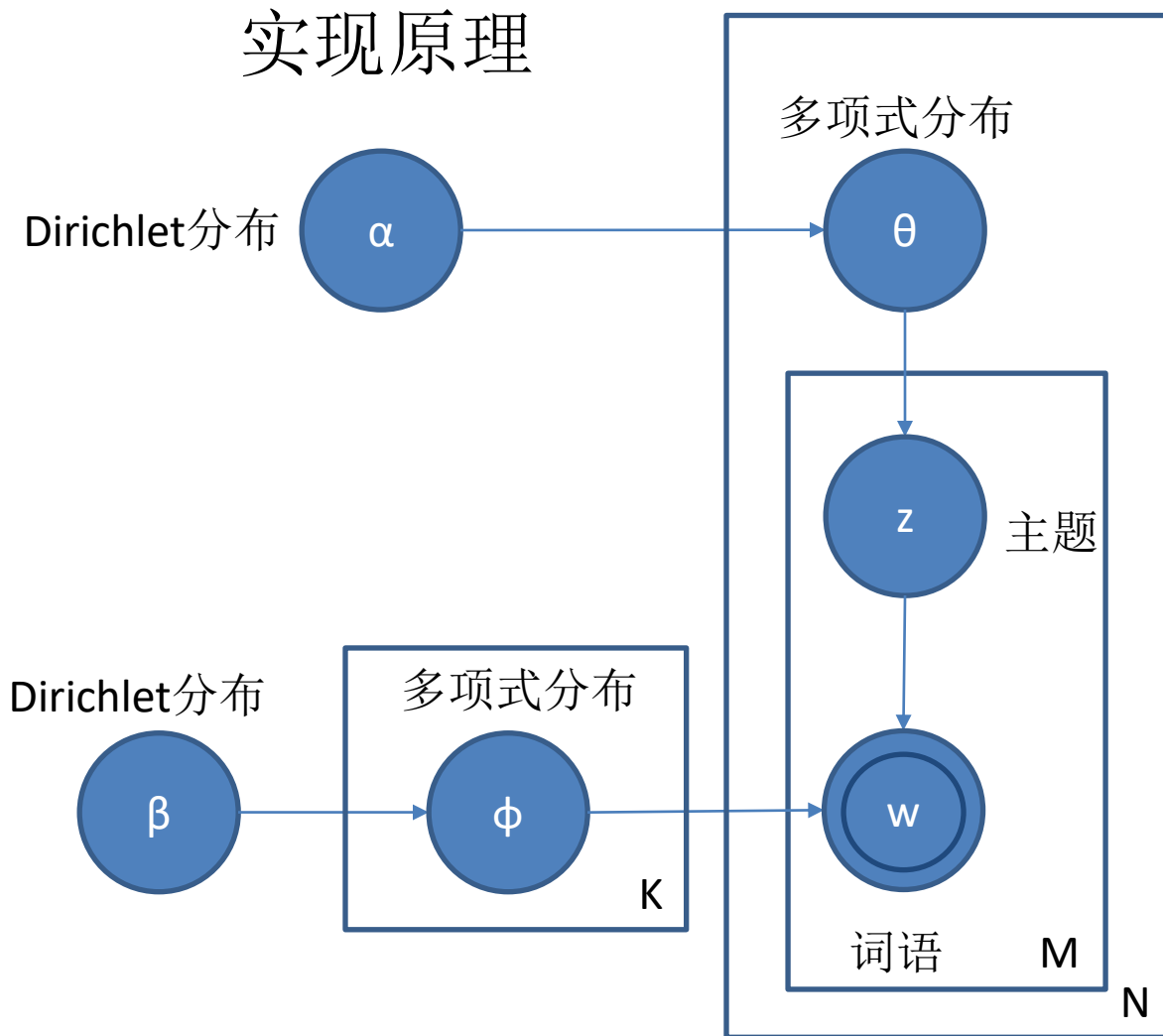
广告 超激斗梦境

视频选集 (1/112)

自动连播

- P1 1-1.欢迎参加《机器学习》课程 06:56
- P2 1-2.什么是机器学习? 07:15
- P3 1-3.监督学习 12:30
- P4 1-4.无监督学习 14:14
- P5 2-1.模型描述 08:11
- P6 2-2.代价函数 08:13

实现原理



- 从Dirichlet分布 α 中取样生成文档 i 的主题分布 θ 。
- 从主题的多项式分布 θ 中取样生成文档 i 第 j 个词的主题 z 。
- 从Dirichlet分布 β 中取样生成主题 z 的词语分布 ϕ 。
- 从词语的多项式分布 ϕ 中采样最终生成词语 w 。

<https://www.icourse163.org/learn/BIT-1449601164?tid=1463297505#/learn/content?type=detail&id=1240683329>

词向量 (word to vector) :

利用浅层神经网络模型自动学习词语在语料库中的出现情况，把词语嵌入 (word embedding) 到一个高维的空间中，通常在100-500维，在新的空间词语被表示为词向量的形式。包括**CBOW模型**和**skip-gram模型**

+

K-means聚类算法



Word2Vec词聚类文本关键词抽取方法:

对于用词向量表示的文本词语，通过K-Means算法对文章中的词进行聚类，选择聚类中心作为文章的一个主要关键词，计算其他词与聚类中心的距离即相似度，选择topN个距离聚类中心最近的词作为文本关键词，而这个词间相似度可用Word2Vec生成的向量计算得到。



主要思想：把关键词提取的过程看成特征提取的过程，通过已有的语料库来计算词语的信息增益，信息增益越大的就越有可能是关键词。

❖ **Information Content (Entropy):**

$$I(P(v_1), \dots, P(v_n)) = \sum_i -P(v_i) \log_2 P(v_i)$$

□ **For a training set containing p positive examples and n negative examples:**

$$I\left(\frac{p}{p+n}, \frac{n}{p+n}\right) = -\frac{p}{p+n} \log_2 \frac{p}{p+n} - \frac{n}{p+n} \log_2 \frac{n}{p+n}$$



- ❖ A chosen attribute A divides the training set E into subsets E_1, \dots, E_v according to their values for A , where A has v distinct values.

$$\text{remainder}(A) = \sum_{i=1}^v \frac{p_i + n_i}{p + n} I\left(\frac{p_i}{p_i + n_i}, \frac{n_i}{p_i + n_i}\right)$$

- ❖ **Information Gain (IG)** or reduction in entropy from the attribute test:

$$IG(A) = I\left(\frac{p}{p + n}, \frac{n}{p + n}\right) - \text{remainder}(A)$$

- ❖ Choose the attribute with the **largest IG**

来自刘峡壁老师：Machine Learning课件



原理：卡方检验就是统计样本的实际观测值与理论推断值之间的偏离程度，**如果卡方值越大，二者偏差程度越大**

考虑皮尔逊检验统计量：

$$\chi^2 = \sum_{i=1}^n \frac{(f_i - np_i)^2}{np_i}$$

构建高频词集G，计算候选词的卡方值，认为TopK个词为关键词



实验以及应用demo



实验：无监督经典算法+同一文档→横向对比

文档：

拜登称俄罗斯“已决定入侵乌克兰” 德国、乌克兰不附和

(2022年02月21日10:00 | 来源：[新华网](#))

- 参与计算的文档无标题
- 返回前十个关键词

Tfidf: (jieba实现)

```
[('乌克兰', 0.37856925544545456),  
(('慕尼黑', 0.18258063572954547),  
(('连斯基', 0.1773250465848485),  
(('19', 0.14490627276242424),  
(('贝尔', 0.12701855690166666),  
(('俄方', 0.11615987319321211),  
(('俄罗斯', 0.11339637078672729),  
(('总统', 0.11233532558672726),  
(('臆断', 0.11173854308636363),  
(('拜登', 0.11008107438818182))]
```

textrank: (jieba实现)

```
[('俄罗斯', 1.0),  
(('俄方', 0.7732940628966294),  
(('情报', 0.606864166325317),  
(('国家', 0.5911272170527998),  
(('总统', 0.5748623035198418),  
(('经济', 0.5291050025828828),  
(('美国', 0.514431957162694),  
(('入侵', 0.4999322456156636),  
(('德国', 0.4804228589642128),  
(('记者', 0.4758626504335891))]
```


Ida

```
[('乌克兰', 0.008474599),  
( '俄罗斯', 0.008357868),  
( '德国', 0.008349175),  
( '泽', 0.008344788),  
( '总统', 0.008337006),  
( '慕尼黑', 0.008334773),  
( '记者', 0.008334101),  
( '美国', 0.008329632),  
( '国家', 0.008327998),  
( '伯克', 0.008327664)]
```


- Jieba. `posseg`
('n', 'nr', 'ns',
'nt', 'eng', 'v',
'd')
- 停用词词典

针对
问题
防止
附近
限制
随后
随时
随著
难道说
集中
需要
非特
非独
高兴
若果

行 1602, 列 1

人民日报 

3-10 11:14 来自 微博视频号

【欢迎回家！#乌克兰撤侨航班响起我和我的祖国#❤️】9日早上7点31分，第七架接返自乌克兰撤离中国公民临时航班安全抵达兰州。起飞前，#撤侨航班上的机长广播#令人泪目。抵达祖国上空时，乘客们挥舞起手中的国旗，一同唱响《我和我的祖国》。“祖国永远是我们最强大的后盾！”  人民日报的微博视频



Tfidf

['祖国', '后盾', '一万八', '18000', '吓一跳', '永远', '机票',
'不比', '唱歌', '免费', '强大', '巨婴', '全免费', '总做', 'ktv',
'撤侨', '只花', '战狼带', '飞机', '人家']

textrank

['祖国', '免费', '回到', '飞机', '视频', '知道', '印度', '隔离',
'需要', '配合', '社会', '自费', '环境', '瞧瞧', '人家', '动荡',
'机票', '后盾', '国土', '吓一跳']

对于这次撤侨行动，通过提取关键词，
可以看出几种观点：

(仅限微博，不代表本人观点)

1. 祖国是强大后盾
2. 18000，自费，对比印度
3. 唱歌配合录视频



未来发展方向



在目前阶段，关键词的评价标准仍有瑕疵。其原因是关键词的定义在语言学上仍存在部分争议。我们经常会用学术论文作为评价和训练集，然而不同论文的作者的评价标准往往不同，导致了现阶段关键词提取评价标准不一的问题。如何提出一套通用的关键词提取标准，会是未来我们要研究与解决的问题。



语料库基于有监督的机器学习算法往往需要大量的训练数据集，但现阶段已标注关键词的数据非常有限。如果采用论文作为训练数据，一方面存在着版权问题，另一方面也存在着关键词标准不一的问题。网络上存在着大量的未标注关键词的新闻，小说等资料。如何利用现有的素材构建合适的语料库，同样是一个需要深入研究的方向。



(1) 可读性:

尤其是对中文而言。中文的字与字之间是没有空格，需要分词工具对文本进行切分。而且中文中还存在多义词等情况，所以系统所提取出来的关键词的可读性对系统的实用性是个很大的考验。为了提升关键词可读性，我们可能需要去做专门的研究。



(2) 高速性:

系统应该具有较快的速度，能够及时处理大量的文本。比如一个针对各类新闻的关键词提取系统，当新闻产生后，应该能在数秒内提取出该新闻的关键词，才能保证新闻的实时性。这需要我们在未来尽可能优化当前的算法，使得在准确率不下降的情况下，算法的效率尽可能提高。



(3) 学习性:

实用的关键词提取系统，应该能处理非常广泛的领域的文本，而不是仅仅局限于特定领域。随着社会的高速发展，各种未登录词，网络新词频频出现，系统应具有较强的学习能力。也就是说在未来，我们还需要加强产品智能性的研究。



(4) 健壮性:

系统应该具有处理复杂文本的能力，如中、英文混杂，文本、图表、公式混杂的文本。对于复杂的文本而言，我们的关键词抽取系统需要具备挑选合适的语言、甚至概括的能力。这也意味着在未来的发展道路上，我们需要从不同的角度定义提取原则，才能使系统有足够的实用性。



谢谢观看

答辩人：王轩、赵亚祥、张墨言、陈宇飞、许子逸

时间：2022-3-22