# 第一章 绪论

## 研究背景与意义

近年来，互联网日益成为人们获取资讯的主要途径，与传统的信息媒介相比，互联网上传播的信息类型多种多样，包括文本、图像、音频、视频等，随着Web2.0、社交媒体和移动互联网等技术的发展，每个网民都成为了互联网上信息的创造者与传播者，促使互联网网上文本信息爆炸式增长。公开的知识日益以新闻、博客、网页、科学论文、书籍、图像、声音、视频和社交网络的形式被数字化存储，巨大的信息量同时也增加了人们寻找和发现自己所需要的知识的难度。

根据CNNIC《第37次中国互联网络发展状况统计报告》【1】的数据显示：截至 2015 年 12 月，中国网页数量为 2123 亿个，年增长 11.8% 。仅仅中国的互联网领域就拥有千亿级别的网页数量，迅速增长的信息体量让人们的信息分析能力与信息量之间的鸿沟不断扩大，人们需要新的计算工具以组织、搜索和理解这些庞大的信息量。

搜索引擎的出现让互联网时代的信息检索更为方便，使用Google（谷歌）和百度等搜索引擎进行信息检索已成为当代网民们必不可少的一项技能。据《第38次中国互联网络发展状况统计报告》【2】显示：截至 2016 年 6 月，我国搜索引擎用户规模达 5.93 亿，使用率为 83.5%，用户规模较2015年底增加 2635万，增长率为 4.7%；手机搜索用户数达 5.24 亿，使用率为 79.8%，用户规模较 2015 年底增加 4625 万，增长率为 9.7%。其中搜索引擎行业的一大变化趋势是“人工智能技术成为发展重点，搜索精准度持续提升”，搜索引擎算法与人工智能技术进行深度融合，并逐渐应用于用户日常使用的搜索产品中。

尽管互联网上拥有包括声音、图像、视频、文本等多种信息类别，文本信息无论从数据体量还是从内涵信息量的角度，都是一个极其重要的信息类别，如何分析、挖掘和利用这些信息，充分发挥其潜藏的巨大价值，是当前研究热点之一。文本信息（或者称为文本数据）在互联网上大多以半结构化或非结构化的形式存在，其挖掘和检索方式与现有成熟的结构化信息检索技术（如关系型数据库等）有所区别，故需要新型的技术手段对文本数据进行挖掘与检索，这即是文本挖掘（Text Mining）技术。

互联网上的文本信息，从内容的创造者角度可以分为专业创造内容（Professionally Generated Content，PGC）、职业创造内容（Occupationally Generated Content，OGC）和用户创造内容（User Generated Content, UGC）。其中，PGC是指由具有专业学识、资质的人员进行的内容创造（如资深用户的点评），OGC是指以职业为前提的内容创造（如部分新闻网站雇佣的内容编辑），而UGC是指普通用户创造的内容。

UGC作为Web2.0最大的特点，这种人人均可创造内容的特点，使得互联网信息飞速增长，形成了一种“多、广、专”的局面，对人类知识的积累与传播起到了极大的促进作用，是当前互联网繁荣的一大原因。但是，UGC却从另一个角度增长了互联网上虚假、片面信息的体量，由于人人均可创造内容，其内容的质量也需要用户自行判断，因此也催生了对于智能审核、过滤、筛选机制的需求。

值得注意的是，在UGC方面，有很大一部分来自于社交网络与电子商务网站（一个典型的例子即是在线评论），这两种媒介中的文本信息拥有“数据体量巨大”和“涵盖内容丰富”、“能反映用户真实想法”等特点，具有很大的挖掘价值。因此，如何从大量的用户评论中抽取出最具代表性的评论，在网络舆情监测、商品营销策略设计、用户行为决策等方面的具有很大价值。

## 国内外研究现状

1.2.1 文本挖掘概述

文本挖掘（Text Mining）是数据挖掘（Data Mining）和自然语言处理（Natural Language Process）相结合的产物，并且同时涉及多个学科领域，包括但不限于：信息检索、机器学习、概率统计、图论等。

文本挖掘的概念是由Feldman首次提出的【】，由于涵盖的范围很广，文本挖掘也没有一个统一、确切的定义。通常认为，文本挖掘的目的是把文本的信息转化为人们可以利用、理解的知识，以这一个目的作为出发点，有如下5种关于文本挖掘的定义：

**定义1**【3】：文本挖掘指的是对于大规模文档的处理和从大量文本数据中提取出隐含的知识信息。

**定义2**【4】：文本挖掘是基于语义学并使用概率统计理论、统计模型、贝叶斯模型、向量空间模型、图论等方法，从大量文档中挖掘出知识模式的技术。

**定义 3**【5】：文本挖掘是一种使用计算语言学规则从文本数据中提取出知识信息的研究和应用方法，其关键领域包括特征提取、聚类、摘要和主题索引。

**定义 4** 【6】：文本挖掘是指一种对自然语言文本数据进行模式提取的技术，其以特定目标对文本进行信息提取。

定义 5【7】：文本挖掘是一种结合数据挖掘的规则、信息提取技术、信息检索技术、计算语言学、文本分类、机器学习等多项理论与技术的，用于发现文本数据中的结构、模式和知识的技术。

综上所述，可以总结出如下对于文本挖掘的定义【8】：

文本挖掘（Text Mining）是以计算语言学、数理统计分析为理论基础，同时结合机器学习与信息检索的一种技术，用于从文本数据（Text Data）中发掘和提取出隐含的知识信息。文本挖掘是一个从文本数据中选取、提取出用户可以理解的知识信息的过程。

文本挖掘是数据挖掘的一个分支，因此也拥有和数据挖掘类似的处理流程。下图即是Usama M.Fayyad等人提出的处理流程：TODO 加图

文本数据的形式也区别于传统关系型数据库的结构化形式，大多数是以半结构化或非结构化的形式存在的。同时，由于文本数据是自然语言的载体，文本数据往往也包含了不同层次上的多种歧义（如语法、语义、词汇上的歧义等），因此文本数据具有区别于结构化数据的一些特征。一般地，可以总结出文本数据拥有以下特征：1. 半结构化或非结构化（如半结构化的HTML文本与非结构化的纯文本）；2. 类型多样复杂（例如，不同类型的网站所获取的文本数据类型也不尽相同）；3. 维度高（文本数据经过向量化后往往会有成千上万维，如不加以预处理，会造成极大的计算存储资源损耗）；4. 语义相关（文本数据作为自然语言的载体，包含多个语义层次的相关性，若在挖掘时忽视其自然语言的特点，会影响效果）。因此在对文本数据进行挖掘时，往往需要采取不同于结构化数据挖掘的技术手段。

文本挖掘是一个很宽泛的概念，其分支包括了信息提取（Information Exraction）、主题跟踪（Topic Tracking）、文本摘要（Text Summarization）、文本分类（Text Categorization）、文本聚类（Text Clustering）、实体链接（Entity Linkage）、信息可视化（Information Visualization）等。根据具体的应用需求，本文主要涉及文本挖掘内的文本摘要领域。

自动文本摘要（Automatic Text Summarization）是由Luhn在1958年首先提出的。自动文本摘要是文本挖掘的一个重要领域，其主要的目的是从一篇文档或多篇文档中提取、总结出能够概括该文档（集）内容的句子。

依据文摘内容的组成成分进行划分，目前有两种主流的文本自动摘要方法：抽取式（extractive）和摘要式（abstractive）。

抽取式的文本自动摘要方法主要是在现有的文本中选取词汇、词组或句子，将它们组合起来形成总结摘要。抽取式的方法总体上可以分为三个步骤：首先是对输入文本进行特征表示，利用诸如主题词(Topic Word)、词频、TF-IDF(Term Frequency - Inverse Documents Frequency，词频-倒排文档频率)值等特征进行文本建模；然后依据不同的规则对句子进行打分；最后根据评分进行句子的选取以组成最终的文本摘要内容。最初的抽取式方法由Luhn首先提出，他把文本词汇分为通用词和内容词，其中通用词又称为功能词，一般包括连词、助动词、助词、代词、介词、冠词、一部分形容词和副词等，除了通用词以外的词即为内容词。Luhn认为功能词的重要度为0，故对文本进行词频统计分析时只考虑内容词，词频超过某个阈值的词汇被认为是可以代表文章主题内容的有效词汇（Significant Term），依据某一个句子中的有效词汇数量来分析句子的重要性，并选取最重要的前k个句子作为文章的摘要。近年来兴起的TextRank运用图论的方法来计算句子的重要性，它们把每一个句子当做图结构中的一个节点，利用句子之间的重叠度来衡量节点间的相似度，相似度大于某个阈值的两个节点之间会构成一条边，最后利用图论中的连通性来衡量句子的重要度，获得了很好的抽取式摘要效果。

相对的，摘要式的方法则主要为文本建立一个语义上的表达，然后利用自然语言生成技术生成一段总结性的摘要文字，并尽可能地接近人类所概括出的文字，摘要式的方法生成的摘要文本中有可能包含原始文本中不存在的词汇。Radev和McKeown在1998年提出一种使用先验知识生成摘要的方法；Saggion等人使用自然语言生成系统（Natural Language Generation System）生成摘要；Ganesan等人提出了一种基于图的摘要式方法，利用图来为高度冗余的观点生成摘要。摘要式的方法是当前研究领域的一大趋势，然而因为复杂性等原因，目前的研究成果大部分着眼于抽取式的方法上。

依据输入的文本数量进行划分，文本自动摘要技术又可以分为单文档自动摘要（Single-document Summarization）和多文档自动摘要(Multi-documents Summarization)。单文档自动摘要技术为单个文档生成摘要，而多文档摘要技术则为多个主题类似的文档产生摘要。单文档自动摘要技术针对单个文档，对其中的文本内容进行抽取，并针对用户或者应用的需求，将文本中最重要的内容以简化的形式呈现出来；多文档自动摘要技术主要目的是为包含了多份主题相近的文档的集合生成能够概括该集合主要内容的摘要，相比单文档自动摘要，多文档自动摘要技术除了要过滤多份文档中的冗余内容，还要识别不同文档中的独特内容，并让生成的摘要尽可能精简。多文档自动摘要的典型例子是2004年有Erkan等人提出的LexRank，它与TextRank类似，都是采用了构建句子图结构的方式，区别在于LexRank使用到了TF-IDF来衡量句子之间的相似度。除此之外，刘德荣等人提出了一种基于主题概念的多文档自动摘要方法，Cardboll等人提出了一种主题驱动的多文档摘要方法（Topic-Driven Summarization）。

综上所述，文本挖掘涉及的任务类别多种多样，本文希望将文本挖掘限定在多文档自动摘要的特定语境下，并结合其他相关理论进行实践创新。

1.2.2 图模型理论概述

图模型又称为概率图模型(Probabilistic Graphical Models)，是概率论与图论(Graph Theory)相结合的产物，它提供了概率分布(probability distribution)的图形化表示。一个概率图模型是一个由结点(nodes)和边(edges)所组成的图(graph)，其中，每个结点表示一个(组)随机变量，连接结点的边表达了随机变量之间的概率关系(probabilistic relationships)。这样的图表达了一种数学关系——所有随机变量上的联合分布(joint distribution)如何分解为若干仅依赖于部分随机变量的因子(factors)的乘积(product)形式。

概率图模型主要包括有向(directed)图模型和无向图(undirected)模型两大类。有向图模型称为贝叶斯网络(Bayesian Networks)，适合于表述随机变量之间的因果关系(causal relationships)。贝叶斯网络又称有向无环图(Directed Acyclic Graph)，其边集合E中的边有向且没有环路，同时还满足局部条件独立性假设：对于一个节点i，在给定所有父节点的前提下，条件独立于所有非子节点。无向图模型称为马尔科夫网络(Markov Networks)或马尔科夫随机场(Markov Random Fields，MRFS)，更适合于表述随机变量之间的软约束(soft constraints)。马尔科夫网络又称为马尔科夫随机场(Markov Random Field)，其边集合E中的边都是无向边，同时满足局部条件独立性：对于节点i，在给定所有邻居节点的前提下，条件独立于所有的其他节点。马尔科夫随机场中的一个重要概念是团(clique)，无向图G中任何两个节点均有边连接的节点子集称为团，若C是无向图G的一个团，并且不能再加进任何一个G的节点使其成为一个更大的团，则称此C为最大团(maximal clique)。

概率无向图模型的联合概率分布表示为其最大团上的随机变量的函数乘积形式的操作，成为概率无向图模型的因子分解，其定义为：给定一个概率无向图模型，设无向图G，C为G上的最大团，表示C对应的随机变量。那么概率无向图模型的联合概率分布可写作图中所有最大团C上的函数的乘积形式，即，其中Z是规范化因子，由式给出。其中函数称为势函数，要求势函数是严格正的，通常定义为指数函数：。

概率图模型具有以下优点：(1)用简单的方式来可视化(visualize)概率模型的结构，可用于设计和诱发新模型；(2)可从图中洞悉概率模型的特性，包括条件独立性(conditional independence)；(3)在模型上的复杂计算(推理和学习)可用图操纵(graphical manipulation)来表示，使得底层的数学表达式更为直观。为了有效解决推理问题，常常把以上图模型转换成一种称为因子图(factor graph)的通用图模型。

概率图模型能够紧凑地表示复杂的领域问题，为各种随机现象的建模和分析提供了新的理论工具，在机器学习领域，除了典型的贝叶斯网络和马尔科夫网络外， 概率图模型还有各种变种，包括：隐马尔科夫模型(Hidden Markov Models)、条件随机场(Conditional Random Fields)、高斯网络模型(Gaussian Network Models)以及各种模型的组合等。

随着概率图模型理论与方法的逐步成熟，尤其是因子图模型及其推断方法的提出，概率图模型逐步成为机器学习领域的重要方法之一。

隐马尔科夫模型（HMM）被广泛地运用于语音识别及序列标注等自然语言处理任务。条件随机场（Conditional Random Field, CRF）也被用于序列标注任务中（如中文分词和词性标注等），与隐马尔科夫模型相比，条件随机场在进行序列标注时，会考虑观测序列与标注序列的全局信息，因此往往会得到更好的效果，但是也随之带来了更大的计算复杂度。潜在狄利克雷分布（Latent Dirichlet Allocation，LDA）是由Blei和Andrew Ng等人提出的一种主题模型（Topic Model），它利用贝叶斯有向图模型来表示字词-文档-主题的概率关系，利用狄利克雷分布来估算输入文档的主题概率分布。LDA作为主题模型领域最具代表性的模型，近年来也拥有了很多种变形版本，并被广泛应用于文本聚类、文本分类、文本表示等任务上。

图模型是一项非常有用的数学工具，本文希望运用图模型的一些理论方法，与自然语言处理的相关技术相结合，进行文本挖掘任务。

1.2.3 针对在线评论的文本挖掘概述

在现有的应用层面，对在线评论进行主要观点的挖掘可以视为一种多文档自动摘要任务。以在线评论的观点挖掘为例，每一条在线评论都可以视为文档集中的一篇文档，在线评论的观点挖掘目标即是挖掘、选取出众多评论中最具有代表性的评论。

在线评论是Web2.0最重要的产物之一，作为互联网用户表达自己观点的前沿，在线评论具有很强的挖掘价值。在线评论常见于社交网路（如Facebook、Instagram、Twitter、微博、微信朋友圈等）、电商平台（如Amazon、淘宝、京东等）和新闻网站（如新浪网、网易新闻等）中。其中电商评论作为消费者对商品的评价，是其他潜在购买者的重要参考信息。消费者们倾向于在购买之前浏览已有的商品评论，这些商品评论对消费者的购买行为造成很大的影响。然而，商品评论的数量往往十分巨大，并且评论的内容包含多种信息，这使得普通用户无法完整地参考所有评论。此外，一个网站的商品评论如果能够得到很好的排序，用户的体验会得到大大提升，网站的流量也会因此增长。

国内外学者针对在线评论观点挖掘领域做了大量研究。Liu和Hu提出了一种基于商品特性的商品评论总结方法[1]，综合利用了数据挖掘和自然语言处理技术，分析大量客户对某个商品的评论，总结出客户们对商品具体某个特性的评价观点，是商品评论信息挖掘的经典；Lun-Wei Ku, Yu-Ting Liang和Hsin-Hsi Chen[2]针对新闻和博客语料集进行了观点抽取、总结和跟踪的实践；Li Zhuang、Feng Jing和Xiao-Yan Zhu[3]针对电影的评论，提出了一种基于多知识库的观点抽取方法，充分利用WordNet的信息进行观点关键词的挖掘，同时也考虑了句法信息和词性信息。哈尔滨工业大学的郝媛媛等人研究了基于影评数据的在线评论有用性[8]，从评论的情感倾向以及主客观倾向的角度分析，得出了正负情感倾向对评论有用性具有较大影响的结论。南京理工大学的刘逶迤[9]等人研究了在线商品评论的可信度，提出了一种结合多种因素的评论可信度判别的解决方案。中国人民大学的郭国庆[10]等人则着重研究了消费者在线评论可信度的影响因素，发现在线评论可信度同时受评论者的资信度以及评论内容的质量以及评论的共识性这三个重要因素所影响。Ghose等人总结出，一个重要的商品评论往往包含以下特征：1. 表达出针对商品某些特性的喜好并提供一个简单的解释，可称为“客观性评论”；2. 表达出对某个商品的个人情感，可称为“主观性评论”。Ghose等人的研究还发现，同时包含了“客观性评论”与“主观性评论”的评论，往往会被认为包含了更多的“有用信息”。

# 第二章 自然语言处理的相关理论

## 自然语言处理概述

自然语言处理（Natural Language Process，NLP）是计算机科学、人工智能与计算语言学的交叉子学科，它主要研究能实现人（Human）与计算机之间中自然语言进行有效通信交互的各种理论和方法。自然语言（Natural Language）即人们日常使用的语言，通常指自然地随着文化演进的语言，如英语、日语、汉语等。

实现人机间自然语言通信意味着要使计算机既能理解自然语言文本的意义，也能以自然语言文本来表达给定的意图、思想等。前者称为自然语言理解（Natural Language Understanding），后者称为自然语言生成（Natural Language Generation）。因此，自然语言处理大体包括了自然语言理解和自然语言生成两个部分。历史上大部分的研究都集中于自然语言理解，即把自然语言转化为计算机更易于处理的形式。无论实现自然语言理解，还是自然语言生成，都远不如人们原来想象的那么简单，而是十分困难的。从现有的理论和技术现状看，通用的、高质量的自然语言处理系统，仍然是较长期的努力目标，但是针对一定应用，具有相当自然语言处理能力的实用系统已经出现，有些已商品化，甚至开始产业化。典型的例子有：多语种数据库和专家系统的自然语言接口、各种机器翻译系统、全文信息检索系统、自动文摘系统等。

自然语言处理的历史能追溯到二十世纪50年代，1950年Alan Turing发表了一篇名为《Computing Machinery and Intelligence》的文章，提出了著名的“图灵测试”。之后人们开始热衷于研究机器翻译系统，1954年的“乔治城大学-IBM实验（Georgetown experiment）”成功地利用计算机自动翻译系统将60多句俄文翻译成了英文，该系统以语言学规则作为翻译的基础。以此为开端，语言学规则占据了翻译系统的核心，但是实际的翻译效果不尽人意，研究的热度也一度低迷，直至1980年代统计语言学的兴起。1980年起，NLP领域迎来了一个重要的变革——统计学习方法或者称机器学习方法开始被大量运用于语言处理：隐马尔科夫模型（Hidden Markov Model）被应用于词性标注（Part Of Speech Tagging）任务中，决策树（Decision Tree）等早期的机器学习分类方法被用于代替传统自然语言处理中的手写“if-then”规则，“语料库方法”的引入使得自然语言处理开始转向统计模型与数据驱动。

若以处理的数据类型作为区分点，自然语言处理又可以分为两大类：语音相关的自然语言处理与文本相关的自然语言处理。语音相关的自然语言处理主要包括了语音识别、文本朗读等子学科，文本相关的自然语言处理则围绕着文本数据的处理理论与技术。

发展到现在，自然语言处理的范畴包括：文本朗读/语音合成（Text to Speech）、语音识别（Speech Recognition）、中文分词（Chinese Word Segmentation）、词性标注（Part-of-Speech Tagging）、句法分析（Parsing）、自然语言生成（Natural Language Generation）、文本分类（Text Categorization）、信息检索（Information Retrieval）、信息抽取（Information Extraction）、文字校对（Text-Proofing）、问答系统（Question Answering）、机器翻译(Machine Translation)、自动摘要（Automatic Summarization）、文字蕴涵（Textual Entailment）等。

在本文的研究中，会涉及到自然语言处理中的中文分词、词性标注、自动摘要等内容，与此同时也会涉及到数据挖掘与机器学习的基本处理流程。

## 词性标注的相关理论与技术

词性（Part-Of-Speech）是语言学中的一个概念，又可以称为词类（Word Category），它是指一类拥有共同的语法特性的词。被指定为同一个词性的词汇通常来说都会在句法表现上具有相似性——即在同一种语法结构下扮演着相似的语法角色，不同的语言拥有不同的词性分类。通常来说，现代汉语的词汇可以分为两大类共12种词性，第一大类是实词，包括：名词、动词、形容词、量词、代词和数词；第二大类是虚词，包括：副词、介词、连词、助词、叹词和拟声词。类似地，现代英语的词汇也可以大致分为10类： noun（名词），verb（动词），adjective（形容词），adverb（副词），pronoun（代词），preposition（介词），conjunction（连词），interjection（感叹词），determiner（定冠词），numeral（数词）等。

词性标注（Part-Of-Speech Tagging），顾名思义是一个将文本中的每个词标注上相应词性的过程，词性标注需要同时考虑词汇本身的定义以及在句子中的上下文环境（Context）。

词性标注并非仅仅依靠一个“词汇-词性”的列表就可以完美完成的，因为有些词汇会同时拥有两个以上的词性，并且这些词汇的词性会随着复杂的上下文环境而有所不同。

词性标注的任务可以简单地解释为：输入一个句子，如“我，是，北京邮电大学，的，学生”，返回相应的词性标注序列“我/代词，是/动词，北京邮电大学/名词，的/助词，学生/名词”，属于序列标注问题。基于概率图模型的HMM（隐马尔科夫模型）和CRF（条件随机场）等模型可以很好地解决序列标注问题。

* + 1. 隐马尔科夫模型概述

隐马尔科夫模型是一种可用于标注问题的统计学习模型（Statistic Model），它被广泛地运用于语音识别、自然语言处理、生物信息、模式识别等领域。

隐马尔科夫模型是关于时序的概率模型，主要是描述了由一个隐藏的马尔科夫链随机生成不可预测的状态随机序列，再由各个状态随机生成一个相应的观测结果，以此形成一个随机的观测序列的过程。由马尔科夫链随机生成的状态组成的序列又称为“状态序列”（State Sequence）；由每个随机状态生成的观测结果组成的序列又称为“观测序列”（Observation Sequence）。马尔科夫模型中序列的每个位置都可以看做是一个“时刻”。因此，对于序列标注问题，在隐马尔科夫模型中即是在已知观测序列的前提下，如何找出最可能的状态序列。

隐马尔科夫模型中最重要的三个参数分别是：状态初始概率分布、状态转移概率分布以及观测概率分布，由此可以引出隐马尔科夫模型的形式定义：

令$S={s\_1,s\_2,…,s\_N}$是所有可能的状态集合，其中$N$是可能的状态数；$V={v\_1,v\_2,…,v\_M}$是所有可能的观测集合，其中$M$是可能的观测数。设$I=(i\_1,i\_2,…,i\_T)$和$O=(o\_1,o\_2,…,o\_T)$分别是长度为T的状态序列和观测序列。

因此，令$A$为状态转移概率矩阵：

$$A=[a\_{ij}]\_{NXN}$$

其中，

$$a\_{ij}=P(i\_{t+1}=s\_j|i\_t=s\_i), i=1,2,…,N; j=1,2,…,N$$

表示t时刻处于状态$s\_i$，并在t+1时刻转移到状态$s\_j$的概率。

令$B$为观测概率矩阵：

$$B=[b\_j(k)]\_{NXM}$$

其中，$$b\_j(k)=P(o\_t=v\_k|i\_t=s\_j), k=1,2,…,M; j=1,2,…,N$$

是在t时刻处于状态$s\_j$下生成观测$v\_k$的概率。

又令$\pi$是初始状态概率向量：

$$\pi=(\pi\_i)$$

其中，$$\pi\_i=P(i\_1=s\_i)$$是$t=1$时刻处于状态$s\_i$的概率。

由以上定义可知，隐马尔科夫模型拥有两个基本假设：1. 齐次马尔科夫性假设，即假设了隐藏马尔科夫链在任意时刻t的状态只依赖于前一时刻t-1的状态，与具体的时刻t无关，也与其他时刻的状态及观测无关。2. 观测独立性假设，即假设了任意时刻的观测只与该时刻的马尔科夫链状态有关，与其他观测及状态无关。

综上所述，每个隐马尔科夫模型$\lambda$都可以用三元符号表示：$$\lambda=(A,B,\pi)$$，$A,B,\pi$可以称为隐马尔科夫模型的三要素，其中：

状态转移概率矩阵A与初始状态概率分布向量$\pi$可以确定生成状态序列的隐藏马尔科夫链；观测概率矩阵B则确定了如何利用状态生成观测，B与生成的状态序列相结合，可以确定如何产生观测序列。

在实际运用中，人们常常关注隐马尔科夫模型的三个基本问题，即：概率计算问题、学习问题、预测问题。

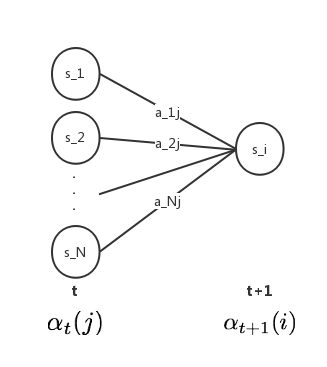
概率计算问题，即给定了隐马尔科夫模型$\lambda=(A,B,\pi)$和观测序列$O=(o\_1,o\_2,…,o\_T)$，计算出在模型$\lambda$条件下观测序列O出现的概率。概率计算的问题采用前向算法（Forward Algorithm）和后向算法（Backward Algorithm）进行计算。前向算法与后向算法都是为了减少概率计算时的计算量，它们具有相似的形式与思想，因此下面仅简单介绍前向算法：

前向算法中一个重要的概念即是前向概率，它的定义是：给定一个隐马尔科夫模型$\lambda$，到时刻t部分的观测序列为$o\_1,o\_2,…,o\_t$且状态为$s\_i$的概率为前向概率，记为：

$$\alpha\_t(i)=P(o\_1,o\_2,…,o\_t,i\_t=s\_i|\lambda)$$

由此可以得到前向概率与观测序列概率的递推公式：

1. 初始值：$$\alpha\_1(i)=\pi\_ib\_i(o\_i), i=1,2,…,N$$
2. 前向概率递推：对于$t=1,2,…,T-1$，有: $$\alpha\_{t+1}(i)=[\sum\_{j=1}^{N}{\alpha\_t(j)a\_{ij}}]b\_i(o\_{t+1}), i=1,2,...,N$$
3. 观测序列概率：$$P(O|\lambda)=\sum\_{i=1}^N{\alpha\_T(i)}$$

 如图是前向概率的递推公式的示意图：

前向概率存储了之前所有状态的概率和，利用前向概率与后向概率可以方便地计算出观测序列的概率，在此不再赘述。

隐马尔科夫模型的学习问题即是模型参数$\lambda=(A,B,\pi)$的训练估计问题，根据训练数据中是否同时包含了观测序列和对应的状态序列，该问题可以分别使用监督学习与非监督学习的方法解决。

使用监督学习进行参数估计时，采用的是极大似然估计法，具体方法如下：

1. 状态转移概率$a\_{ij}$的估计：

设样本中t时刻处于状态i，而t+1时刻处于状态j的样本频数为$A\_{ij}$，那么状态转移概率$a\_{ij}$则可以写成：

$a\_{ij}=\frac{A\_{ij}}{\sum\_{j=1}^N{A\_{ij}}}$

1. 观测概率$b\_j(k)$的估计：

设样本中状态为j且观测为k的样本频数为$B\_{jk}$，则状态为j且观测为k的观测概率$b\_j(k)$可以写成：

$b\_j(k)= \frac{ B\_{jk}}{\sum\_{k=1}^M{ B\_{jk}}}$

1. 初始状态概率$\pi\_i$的估计可以表示为样本中初始状态为$s\_i$的频率。

使用无监督学习进行参数估计时，训练数据中只有S个观测序列${O\_1,O\_2,…,O\_S}$，此时可以采用基于期望最大化（Expectation Maximization，EM）思想的Baum-Welch算法进行参数估计。

针对隐马尔科夫模型的预测问题，通常采用维特比算法（Verterbi）进行状态译码。维特比算法实际上是使用了动态规划的方法进行隐马尔科夫模型的预测问题求解，即利用动态规划方法（Dynamic Programming）求概率最大的路径，每一条路径都对应着一个状态序列。由动态规划的原理可以推知，最优路径应该具有如下特性：如果最优路径在t时刻通过了节点$i\_t^\*$，那么这一路径中从节点$i\_t^\*$到终点$i\_T^\*$的路径，一定是优于从$i\_t^\*$到$i\_T^\*$的所有可能的路径。因为假设上述特性不成立，则从$i\_t^\*$到$i\_T^\*$就会有另一条更好的路径，这另一条路径与之前从$i\_1^\*$到$i\_t^\*$的路径合并起来，肯定会组成一条优于原来路径的新路径，于是便出现了矛盾。因此，我们根据该特性，只需从t=1时刻开始递推地计算在t时刻状态为i的各部分路径的最大概率即可，直至得到t=T时刻状态为i的各路径最大概率。然后从终点开始，由后往前地逐步得到最终的最优路径。

为了说明维特比算法，需要定义两个变量$\delta$和$\psi$，其中$\theta$定义为在t时刻状态为i的所有路径中的概率最大值，即：

$\delta\_t(i)=\max\_{i\_1,i\_2,...,i\_t-1}{P(i\_t=i,i\_{t-1},...,i\_1,o\_t,...,o\_1|\lambda)}, i=1,2,...,N$

可得$\delta$的递推公式：

$\delta\_t(i)=\max\_{1<=j<=N}[\delta\_t(j)a\_{ji}]b\_i(o\_{t+1}) ,\ \ i=1,2,...,N$

则t时刻状态为i的所有路径中概率最大的路径的第t-1个节点值为：

$\psi\_t(i)=\arg\max\_{1<=j<=N}[\delta\_{t-1}(j)a\_{ji}],\ \ i=1,2,...,N $

由此可以倒推出最优路径中每一个时刻的节点状态，此为维特比算法进行预测问题的解决方法。

* + 1. 基于隐马尔科夫模型的词性标注

在进行基于隐马尔科夫模型的词性标注时，主要关注的是隐马尔科夫模型中的学习问题和预测问题。通常情况下，为了获得更好的标注效果，会采用含人工标注训练数据的有监督式学习方法，即最大似然估计法。

标注好的训练数据格式应该有如下形式：包含S句形如“我/代词，是/动词，北京邮电大学/名词，的/助词，学生/名词”的句子，以词汇序列的形式，将词汇与相应的词性做好一一对应。以上述句子为例，观测序列为“我，是，北京邮电大学，的，学生”，隐藏状态序列为“代词，动词，名词，助词，名词”，在训练时只需要统计相应的频数即可近似估计出隐马尔科夫模型中的三个重要的模型参数。特别地，观测的集合数量为标注样本集中的“词典”大小，即有多少个不同的词汇，通常情况下会有成千上万个观测状态；隐藏状态的集合数量为标注样本集中的词性种类数，通常只有十几个隐藏状态。

利用有监督的方式进行隐马尔科夫模型参数估计后，就可以通过维特比算法对新来的测试观测序列进行译码，得到相应的隐藏状态序列（即词性标注序列）。

以上即是基于隐马尔科夫模型的词性标注方法。

* + 1. 条件随机场概述

条件随机场（Conditional Random Field，CRF）是一种马尔科夫随机场（Markov Random Field，MRF）拥有无向图和MRF所共有的特性，如全局马尔科夫性（Global Markov property）、局部马尔科夫性（local Markov property）和成对马尔科夫性（pairwise Markov property）等，因而可以使用MRF的基本理论进行概率推断。

马尔科夫随机场或概率无向图模型中的一个重要定理是关于因子分解的Hammersley-Clifford定理，它用来描述概率无向图模型的联合概率分布$P(Y)$：

$P(Y)=\frac{1}{Z}\prod\_C{\Psi\_C(Y\_C)}$

$Z=\sum\_Y\prod\_C{\Psi\_C(Y\_C)}$

其中C是指该概率无向图中的最大团，$Y\_C$是C中的节点所对应的随机变量，$\Psi\_C(Y\_C)$是在最大团C上所定义的严格正函数，或者称为势函数。乘积是在该概率无向图中的所有最大团上进行的。

Hammersley-Clifford定理描述了如何对概率无向图进行因子分解，这一定理被广泛地应用在概率无向图的模型计算当中。

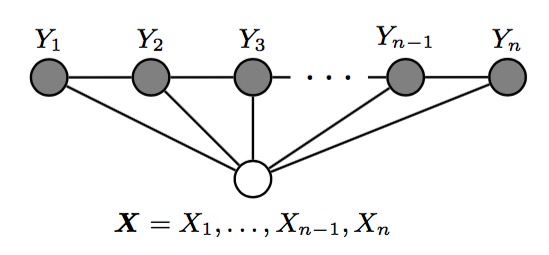
条件随机场是一种判别式的无向图模型，相对于生成式的模型（如隐马尔科夫模型和MRF等）直接对联合分布进行建模，判别式的模型则是对条件分布进行建模。条件随机场是给定一组随机输入变量的条件下，求另一组随机输出变量的条件概率分布模型，它可以用于不同的预测问题中，如图像标注、序列标注问题等。在应用于标注问题时，会对条件随机场的结构加上一个限制，形成线性链（linear chain）条件随机场，则问题变成根据输入序列对输出序列进行预测的判别式模型，并且通常使用最大似然估计法或正则化的最大似然估计法进行参数估计。

条件随机场的定义是：设X、Y分别是两个随机变量，若随机变量Y构成的一个由无向图$G=(V,E)$表示的马尔科夫随机场有：

$ P(Y\_v|X,Y\_w,w\not=v)=P(Y\_v|X,Y\_w,w~v) $

且对于任意节点v都成立，则称条件概率分布$P(Y|X)$为条件随机场，其中$w\not=v$指的是在$G=(V,E)$中除节点v以外的所有节点w，$w~v$表示所有的与节点v有边连接的节点w，$Y\_v$、$Y\_w$则表示与$v$、$w$相对应的随机变量。

在条件随机场的定义中，X和Y并不要求具有相同的结构，然而在线性链条件随机场中，一般假设XY具有如下结构：



在此结构下，该图模型的最大团（Clique）是向量的两个节点集合，在标注问题中，X为观测序列，Y为对应的标记序列。线性链条件随机场有如下定义：

$P(Y\_i|X,Y\_1,...,Y\_{i-1},Y\_{i+1},...,Y\_n)=P(Y\_i|X,Y\_{i-1},Y\_{i+1})$

即相应的变量状态只与相邻的随机变量有关。

利用Hammersley-Clifford定理可以得到线性链条件随机场$P(Y|X)$的因子分解表达式，由此可以得到线性链条件随机场的参数化形式。设线性链条件随机场$P(Y|X)$，则在随机变量X的取值为x的条件下，随机变量Y的取值为y的条件概率为：

$P(y|x)=\frac{1}{Z(x)}exp(\sum\_{i,k}{\lambda\_kt\_k(y\_{i-1},y\_i,x,i)+\sum\_{i,l}{\mu\_ls\_l(y\_i,x,i)}})$

其中$Z(x)$是规范化因子：

$$Z(x)=\sum\_y{exp(\sum\_{i,k}{\lambda\_kt\_k(y\_{i-1},y\_i,x,i)+\sum\_{i,l}{\mu\_ls\_l(y\_i,x,i)}})}$$

在上述两个式子中，$t\_k$和$s\_l$是特征函数，$\lambda\_k$和$\mu\_l$则是相应的权重。$t\_k$是定义在节点边上的特征函数，称为“转移特征函数”，只与当前位置和前一个位置相关；$s\_l$是定义在节点上的特征函数，又称为“状态特征函数”，只与当前的位置有关。通常情况下，特征函数都是二值函数，只有满足特征条件时为1，否则为0。每一个条件随机场完全可以由特征函数$t\_k$和$s\_l$与对应的权重$\lambda\_k$和$\mu\_l$所确定。

与隐马尔科夫模型类似，线性链条件随机场也拥有三个基本问题，即：概率计算问题、参数学习问题和序列预测问题。概率计算问题同样使用了前向-后向算法进行计算，大大减少了计算复杂度；在参数学习问题中，使用了最大似然估计法进行参数估计；而序列预测问题则同样使用维特比算法进行序列预测。

* + 1. 基于条件随机场的词性标注

在2001年，Lafferty等人首次提出了运用线性链条件随机场进行标注问题求解的方法。

在运用条件随机场进行词性标注任务时，词性标注状态序列与观测序列的形式和隐马尔科夫模型中的完全一致。由于条件随机场中的特征函数具有很大的灵活性，不同特征函数的选取会极大地影响条件随机场词性标注的效果。

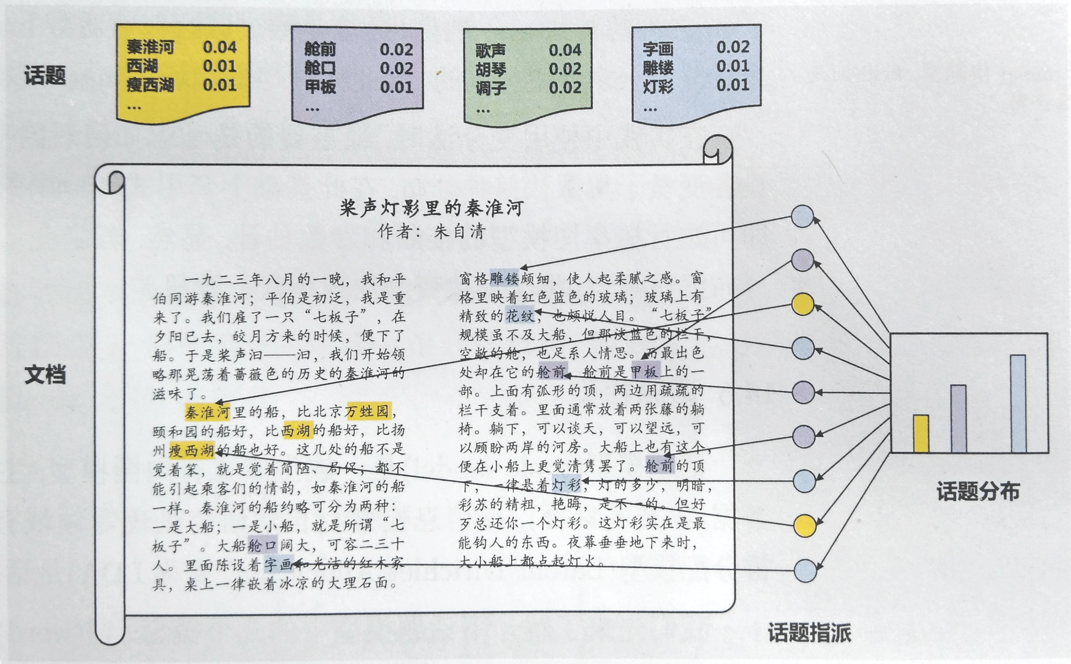
在为条件随机场准备训练数据时，除了准备观测序列集以及对应的隐藏状态集之外，还需要定义一些特征函数模板。在此仍旧以“我/代词，是/动词，北京邮电大学/名词，的/助词，学生/名词”为例，假设模板的形式为“POS[-1]|POS[0]|POS[1]=代词|动词|名词”，该模板的意思是“若前一个位置的词性是代词，当前位置词性是动词，下一个位置的词性是名词，则该定义的特征函数值为1，否则为0”，那么在训练过程中，CRF算法会不断观察相邻三个位置的标记值，并更新相应的特征函数权重。

由于特征函数的存在，条件随机场在进行序列标注时，不仅考虑了观测序列的值，还同时考虑了全局的隐藏状态序列标记值。相对于隐马尔科夫模型，条件随机场参考了更多的全局信息，并且拥有更加灵活的特征函数定义，这使得基于条件随机场的词性标注效果更加令人满意。

* 1. LDA主题模型概述

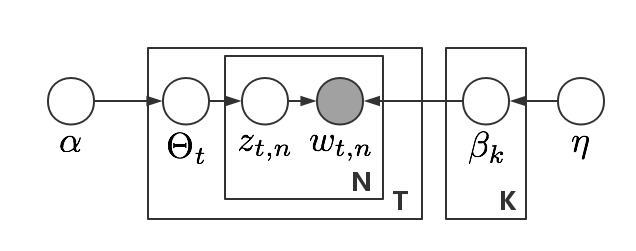
主题模型是一类生成式的有向图模型，主要用于处理离散型数据（典型的如文本数据等）。主题模型被广泛地运用于自然语言处理、信息检索等领域。LDA（Latent Dirichlet Allocation，潜在狄利克雷分配）模型是由Blei、Andrew Ng、Micheal I. Jordan等人提出的一种典型的主题模型。

在主题模型中，有几个重要的概念：词（word）、文档（document）、话题（topic）。具体而言，一个“词”即是主题模型中待处理的基本离散数据单元，例如在文本处理中的一个词就是一个英文单词或有独立意义的中文词。“文档”则是待处理的数据对象，由一组“词”组成。在“文档”中的“词”并不考虑顺序，只考虑数量与是否存在，类似一个装满不同类型的球的袋子，通常将这种表示方式称为“词袋（bag-of-words）”。词袋模型是一种统计语言学中的语言模型，它可以推广到其他领域。任何能够运用词袋模型表示的数据对象，都可以使用主题模型进行相关的分析，例如若把图像中的区域小块看做“词”，则可以把图像表示为词袋，进而运用主题模型进行分析。而话题（topic）则比较抽象，可以把“话题”视作一个箱子，每个箱子里都装着在该“话题概念”下出现概率较高的“词”，“话题”下包含了若干个拥有某种联系和相关性的“词”。

LDA主题模型提出了一种关于“文档”的生成过程。假设在样本数据集D中共有K个话题（Topic）和T篇文档（Document），文档中的所有词都来自一个包含了N个词的词典（Dictionary）。一般地，我们会使用T个N维的向量$W={w\_1,w\_2,…,w\_T}$来表示文档集，其中$w\_t\in\mathbb{R}^N$的第n个分量$w\_{t,n}$表示文档t中词n的词频；用K个N维向量$\beta\_k,\ k=1,2,…,K$来表示话题，其中$\beta\_k \in\mathbb{R}^N$的第n个分量$\beta\_{k,n}$表示话题k中词n的词频。在LDA主题模型中，认为每个文档都可能包含多个话题，因此可以用一个向量$\Theta\_{t}\in\mathbb{R}^K$来表示文档t中所包含的话题比例，其中该向量的每一个分量$\Theta\_{t,k}$表示文档t中包含话题k的比例。

上图显示了文档中“字词-话题-文档”的关系和LDA中文档生成的过程：

1. 首先从话题集合中，根据参数为$\alpha$的狄利克雷分布随机采样一个话题分布$\Theta\_t$。
2. 重复以下步骤生成接下来的N个词：
3. 根据$\Theta\_t$进行话题指派（topic assignment），进而得到文档t中第n个词的话题$z\_{t,n}$；
4. 根据指派的话题所对应的词频分布$\beta\_k$，随机采样得到文档t中的第n个词$w\_{t,n}$

根据以上的生成过程可知，所生成的文档一定以不同的比例包含了多个话题（因为在步骤1中进行了话题指派），同时文档中的每个词属于不同的一个话题（步骤2b中，根据话题进行词的采样），而这些话题是根据话题比例所产生的（步骤2a中根据话题指派而得到的具体话题）。

由此可以得到LDA模型的随机变量关系，利用盘式表示法得到上图：盘式记法的每个圆圈代表一个随机变量，相互独立的、由相同机制产生的多个变量被放在一个方框里，方框的右下角标出变量的个数，通常用阴影会灰色表示能够被观测到的变量，变量之间箭头的指向表示了变量之间的决定关系。由上图可知，文档中的词频$w\_{t,n}$是LDA模型中唯一可以观测到的变量，它依赖于对该词所进行的话题指派（topic assignment）$z\_{t,n}$以及话题所对应的词频分布$\beta\_k$，其中$\beta\_k$依赖于狄利克雷分布的参数$\eta$；而话题指派$\z\_{t,n}$则依赖于话题分布$\Theta\_t$，且$\Theta\_t$依赖于狄利克雷分布的参数$\alpha$。

因此可以得到LDA模型所对应的概率分布公式：

$p(W,z,\beta,\Theta|\alpha,\eta)=\prod\_{t=1}^Tp(\Theta\_t|\alpha)\prod\_{i=1}^Kp(\beta\_k|\eta)(\prod\_{n=1}^NP(w\_{t,n}|z\_{t,n},\beta\_k)P(z\_{t,n}|\Theta\_t))$

通常，上式中的$p(\Theta\_t|\alpha)$和$p(\beta\_k|\eta)$分别是以$\alpha$和$\eta$为参数的狄利克雷分布，如：

$p(\Theta\_t|\alpha)=\frac{\Gamma(\sum\_k\alpha\_k)}{\prod\_k{\Gamma(\alpha\_k)}}\prod\_k{\Theta\_{t,k}^{\alpha\_k-1}}$

其中$\Gamma(.)$是Gamma函数。在LDA概率分布公式中，需要确定的即是参数$\alpha$和$\eta$。在实践中，对于LDA主题模型的训练需要采用变分法来近似求解$\alpha$和$\eta$的最大对数似然函数。

在已知了模型参数$\alpha$和$\eta$以后，可以根据新文档集中的词频$w\_{t,n}$来推断文档集所对应的话题分布，即相应的$\z\_{t,n}$、$\Theta\_t$和$\beta\_k$。推断的方法是求解$p(z,\beta,\Theta|W,\alpha,\eta)$，即：

$p(z,\beta,\Theta|W,\alpha,\eta)=\frac{p(W,z,\beta,\Theta|\alpha,\eta)}{ p(W |\alpha,\eta)}$

由于上式的分母较难求解，所以通常采用吉布斯采样法进行近似概率推断。

LDA主题模型作为一个典型的概率主题模型，在文本分类、文本聚类方面占有举足轻重的一席。由于其使用的词袋模型在很多领域都十分常见，因此LDA主题模型并非仅仅运用于文档处理方面，众多的学者开始利用LDA主题模型的思想进行其他领域的应用实践。

* 1. 基于图的重要度排序

基于图的排序算法（Graph-based Ranking Algorithm）在社交网络分析、论文引用分析以及互联网链接结构分析等复杂网络领域取得了很大的成功，并被广泛应用于各大互联网搜索引擎中。简而言之，基于图的排序算法通过迭代地利用全局信息与局部信息来计算图结构中所有节点的重要度，充分利用了协同信息，以达到最好的重要度排序结果。

经典的基于图的排序算法包括了Kleinberg的HITS算法（Hypertext-Induced Topic Search）和Google的PageRank算法。

* + 1. HITS算法简介

HITS算法是链接分析中非常基础且重要的算法，Hub页面（枢纽页面）和Authority页面（权威页面）是HITS算法最基本的两个定义：

权威页面是指与某个话题或领域高度相关的高质量网页，例如在电商领域，淘宝网和京东网算是两个质量比较高的权威页面；在新闻领域，新浪网和网易新闻网是两个质量比较高的权威页面。

枢纽页面是指包含了很多指向权威页面的链接的网页，典型的如hao123.com一类的导航页面可以算是高质量的枢纽页面。

HITS算法的目的是在海量的网页中，通过技术手段找到与用户查询主题相关的高质量权威页面和枢纽页面，并作为搜索结果返回给用户。

HITS算法拥有两个基本假设，即“一个高质量的权威页面会被很多高质量的枢纽页面指向”和“一个高质量的枢纽页面会指向很多高质量的权威页面”。利用以上两个基本假设，可以迭代地计算出每个页面的“枢纽值”与“权威值”，最后按照权威值的排名由高到低将排序后的结果返回给搜索用户。

HITS算法在整体上能够取得很好的效果，但是最初版本的HITS算法却拥有计算效率低、主题漂移、易被作弊者利用和结构不稳定等问题，因此很多的后续研究都针对这些问题进行了相应的改进。

* + 1. PageRank算法简介

PageRank是Google创始人Larry Page和Sergey Brin于1997年构建早期的搜索系统原型时提出的链接分析算法，自从Google在商业上获得空前的成功后，该算法也成为其他搜索引擎和学术界十分关注的计算模型。目前很多重要的链接分析算法都是在PageRank算法基础上衍生出来的。

PageRank是Google用于用来标识网页的等级/重要性的一种方法，是Google用来衡量一个网站的好坏的唯一标准。在结合了诸如Title标识和Keywords标识等所有其它因素之后，Google通过PageRank来调整结果，使那些更具“等级/重要性”的网页在搜索结果中另网站排名获得提升，从而提高搜索结果的相关性和质量。PageRank的级别从0到10级，10级为满分。PR值越高说明该网页越受欢迎（越重要）。例如：一个PR值为1的网站表明这个网站不太具有流行度，而PR值为7到10则表明这个网站非常受欢迎（或者说极其重要）。一般PR值达到4，就算是一个不错的网站了。Google把自己的网站的PR值定到10，这说明Google这个网站是非常受欢迎的，也可以说这个网站非常重要。

PageRank之前，人们常用HITS算法计算进行网页的重要度，然而HITS算法的结果是与用户查询主题相关的，所以HITS算法必须在接收到用户的查询请求以后才能进行重要度计算，计算效率比较低。而PageRank是主题无关的，它计算出的结果是网页的重要度，与用户的查询输入没有任何关系，因此可以进行大规模的离线计算，得到各个网页的重要度，之后再配合其他相似性指标对用户的查询输入进行实时响应，计算效率更高。

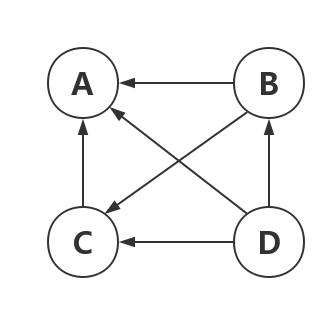
在PageRank出现以前，早期的网页重要度评价方法提出利用网页的入链数量来进行链接分析的相关计算，该种入链数量分析法假设：如果一个网页的入链数量越多，则该网页越重要。PageRank在考虑了入链数量的同时，也考虑了页面本身的质量因素，将两者结合起来以获得更好的页面重要性评估效果。

与HITS类似，PageRank的重要度计算也基于以下两个基本假设：

1. 数量假设：在Web页面图模型中，如果一个页面节点接收到其他页面节点的入链越多，则这个页面节点越重要。
2. 质量假设：本身页面质量越高的页面节点会对其他相连的页面节点产生更大的影响，因此一个页面被质量越高的其他页面所指向，该页面的重要度更大。

PageRank的计算步骤中充分体现了上述的两个基本假设：在初始阶段，网页通过链接关系构建起Web图，每个页面设置相同的PageRank值，通过若干轮的计算，会得到每个页面所获得的最终PageRank值。随着每一轮的计算进行，网页当前的PageRank值会不断得到更新。而在每一轮更新页面PageRank得分的计算中，每个页面将其当前的PageRank值平均分配到本页面包含的出链上，这样每个链接即获得了相应的权值。而每个页面将所有指向本页面的入链所传入的权值求和，即可得到新的PageRank得分。当每个页面都获得了更新后的PageRank值，就完成了一轮PageRank计算。

PageRank的基本思想即是图结构中各个节点之间利用链接进行“相互投票”，节点为自己认为重要的其他节点进行投票，最终依靠群体的力量进行节点重要度的评估。以网页Web图为例，假设一个网页T存在一个指向网页A的链接，则表明网页T的所有者认为网页A是一个重要的网页，因此把网页T的一部分重要性分给（或者说“投给”）了网页A。令T网页的PageRank值为$PR(T)$，它的出链数为$L(T)$，则具体分配出来的重要性等于$PR(T)/L(T)$。而网页A所获得的重要性评分等于一系列指向它的网页所贡献的重要性评分之和。

下面运用一个简单的例子来说明PageRank算法：

假设有A、B、C、D四个网页，他们的链接关系如图所示，则对于A网页的PageRank值可以简单记为：

$PR(A)=\frac{PR(B)}{2}+\frac{PR(C)}{1}+\frac{PR(D)}{3}$

但是通常情况下还需要考虑一些没有出链的节点，即孤立节点。因为如果有孤立节点存在，按照上面的公式，则这些孤立节点不会对其他节点贡献任何PageRank值，这会让网络上的PageRank值在迭代中不断损失。因此需要对简单版本的PageRank公式进行修正，加入一个与页面PageRank最小值相关的修正项d，这个修正项d被称为阻尼系数，通常取0.85。因此，完整版本的PageRank公式为：

$PR(p\_i)=\frac{1-d}{N}+d\sum\_{p\_j\in M(p\_i)}{\frac{PR(p\_j)}{L(p\_j)}}$

其中N为所有页面的数量，$ M(p\_i)$是链入$p\_i$的页面的集合，$L(p\_j)$是$p\_j$链出页面的数量。

PageRank算法在初始阶段为所有节点赋值（如赋值为同一个值），经过若干次跌打以后，总是可以得到一个稳定的PageRank值结果。PageRank算法作为一个经典的基于图的重要度排序算法，自其诞生以来被应用于很多领域之上，并有一系列基于PageRank的改进版本，至今仍旧焕发出旺盛的活力。

* + 1. TextRank简介

TextRank是Mihalcea等人于2004年提出的基于图的排序算法，其思想受PageRank启发。TextRank将图排序算法应用到了自然语言处理领域，分别对不同的文本单元进行建模后，可以解决关键词提取、关键短语提取、文档自动摘要等自然语言处理与文本挖掘任务。

TextRank对PageRank的公式进行了一定的修改：在PageRank的原始公式中，一个Web图是有向无权图，为了适应文本处理的相应要求，Mihalcea等人将图结构修改为无向图，并视不同的任务而确定是否为有权图。

TextRank的通用步骤分为四步：

1. 识别文本单元，并根据任务选取一个最佳的文本表示方式，将文本单元转化为图中的节点。
2. 确定文本单元之间的关系，并利用这些关系为图中的节点添加边连接，这些添加的边可以是有向的或无向的，并且可以是加权的或不加权的，视具体任务而定。
3. 迭代地使用图排序算法直至收敛。
4. 利用节点的最终得分进行排序，给出节点的排序结果。

针对关键词提取任务，TextRank设计出一个无向无权图：首先将文本中的每个句子进行分词，得到句子的集合以及单词的集合。然后将每一个单词作为图中的一个节点，并设定一个大小为k的滑动窗口，每个句子中的连续k个单词作为一个窗口。在一个窗口内的所有单词两两之间存在一个无向无权的边。最后基于PageRank算法对上述构建出来的图进行迭代计算，最终选出得分最高的若干个单词作为关键词。

针对摘要提取任务，TextRank构建出一个无向有权图：首先将文档中的每一个句子作为图中的一个节点，计算两两句子之间的相似度，并把相似度作为节点之间的边权值。TextRank使用的共现词的数量衡量句子间的相似度，计算公式为：

$Sim(S\_i,S\_j)=\frac{|{w\_k|w\_k\in S\_i\&w\_k\in S\_j}|}{log(|S\_i|)+log(|S\_j|)}$

其中$S\_i$表示文档中的第i个句子，$|S\_i|$为句子i所含的单词数，上式的分母表示句子i与句子j的共现词数。

得到有权无向图后，需要利用PageRank算法进行重要度计算。由于是有权图，所以对原始的PageRank计算公式做了一定的修改：

$WS(V\_i)=(1-d)+d\*\sum\_{V\_j\in In(V\_i)}\frac{w\_{ji}}{\sum\_{V\_k\in Out(V\_j)}w\_{jk}}WS(V\_j)$

其中$w\_{ji}$是节点j与节点i之间的权值，$ V\_j\in In(V\_i)$是对$V\_i$有入链的节点集合，$ V\_k\in Out(V\_j)$是$V\_j$出链的节点集合。

使用上述公式进行迭代计算并收敛之后，根据每个节点的得分进行排序，选取得分最大的若干个句子作为文档的摘要句。

TextRank算法进行摘要提取时，句子相似度的计算方法会影响摘要的提取效果，众多学者针对此点作出了自己的改进。本文也将在句子相似度的计算上提出的方法。

* 1. 本章小结

本章主要介绍了自然语言处理的相关理论。首先介绍了自然语言处理的基本概念，简单论述了自然语言处理的发展历史和研究趋势，阐述了自然语言处理所涉及的细分领域。然后分别介绍了本次研究中会涉及到的相关技术理论：概述了词性标注的基本理论与两种基于图模型的词性标注技术；介绍了基于图模型的概率主题模型——潜在狄利克雷分配（LDA主题模型）的基本理论和思想；讨论了基于图的重要度排序算法，包括HITS算法、PageRank算法和TextRank算法。

本章所介绍的理论与技术将会在本次研究中有所涉及。

# 第三章 基于潜层结构特征的文本表示方法

文本数据因为其半结构和非结构化特性，无法直接使用现有的数据挖掘理论技术进行处理，因此需要将这些半结构和非结构数据利用一种数学的形式表达出来。文本的表示方法（Text Representation）首先在信息检索领域被深入研究：在信息检索的任务中，人们希望根据检索词（query）查询出与检索词最为匹配或相近的文档（Document），因此寻求找到一种统一的、能被计算机所处理的文本表示形式，通过计算检索词与文档之间的相似度，返回最匹配的结果。

早期的信息检索系统采用标准布尔模型进行文档相似度的衡量，它的特点是查找那些对于某个查询词返回为“真”的文档。在该模型中，一个查询词就是一个布尔表达式，包括关键词以及逻辑运算符。通过布尔表达式，可以表达用户希望文档所具有的特征，例如必须包含哪些关键词，不能包含哪些关键词等等。例如用户希望查询那些包含了“北京邮电大学”或者“北邮”的文档，则根据布尔模型输入的查询词应为“北京邮电大学OR北邮”，最后检索系统会返回符合查询条件的所有文档。布尔模型的虽然有实现简单、查询方式易学等优点，但是由于模型仅由二值表示，存在欠缺表达上的灵活性，无法表达模糊匹配等缺点。

如今常用的文本表示模型是由Salton等人提出的向量空间模型（Vector Space Model，VSM），它是一个把文本文件表示为标识符（比如索引）向量的代数模型。向量空间模型被广泛应用于信息过滤、信息检索、索引以及相关排序。

本章将会在已有的向量空间模型及TF-IDF特征的基础上，提出一种潜层结构特征（Latent Structure Features），以期对文本的“写作风格”进行表示。

* 1. 向量空间模型

向量空间模型把对文本内容的处理简化为向量空间中的向量运算，并且它以空间上的相似度表达语义的相似度，直观易懂。当文档被表示为文档空间的向量，就可以通过计算向量之间的相似性来度量文档间的相似性。文本处理中最常用的相似性度量方式是余弦距离。

向量空间模型中有几个重要的概念，分别是“文档(document)”、“特征项(term)”和“特征项权重（term weight）”：一个“文档”通常指的是具有一定规模的片段，例如一个句子或者一篇文章，都可以当做“文档”；一个“特征项”是指向量空间模型中不可分的语言处理单元，它可以是字、词或者短语等；“特征项权重”指的是每个“特征项”在文档中根据一定规则被赋予的权重，它通常表示特征项在文档中的重要程度。

综上所述，一个文档应用了向量空间模型进行表示后，会转化为如下形式： $D\_i=(t\_1,w\_1;t\_2,w\_2;…; t\_k,w\_k;…; t\_N,w\_N)$，其中$D\_i$表示第i个文档，$t\_k$和$w\_k$分别表示文档中第k个特征项以及对应的特征项权重，N为所有特征项的总数。在某些场景中，会把特征项之间的顺序固定下来，则对文档进行向量空间表示时可以省略特征项，只保留按照顺序排列的特征项权重——由此可以直观地认为向量空间模型把每一个“文档”都投射到一个N维空间内的一个点，每个文档都可以用一个高维向量进行表示。这些经过向量表示后的文本数据可以被数据挖掘和机器学习中的一些通用算法所处理（如分类算法、聚类算法等）。

在空间向量模型中，比较重要的问题是如何选取特征项和相应的特征项权重。特征项和特征项权重的选取，在很大程度上影响着特征向量对文本内容的表达能力。在自然语言处理与文本挖掘中最为知名的一种特征项权重计算方法是TF-IDF（Term Frequency - Inverse Document Frequency，词频-倒排文档频率）法，许多其他的特征权重计算方法都建立在TF-IDF基础之上。

* 1. TF-IDF权重概述

TF-IDF权重是一种被广泛应用于信息检索（Information Retrieval）和文本挖掘的特征权重计算技术。TF-IDF作为一种统计方法，可以表示一个单词（word）对于一个文档集（Documents Set）或语料库（Corpus）中的某一个文档（Document）的重要程度。

TF-IDF的主要思想很简单：如果一个词汇或短语在某一篇文档中的出现频率很高，而同时它在其他文档中出现的次数很少，则说明该词汇或短语具备很好的类别区分能力。

TF-IDF的值其实是TF（Term Frequency，词频）乘以IDF（Inverse Document Frequency），下面分别给出TF和IDF的计算方法：

TF为某个词i在某一个文档j中出现的频率，设$n\_{i,j}$是该词i在某一个文档j中出现的次数，则有：

$tf\_{i,j}=\frac{n\_{i,j}}{\sum\_k{n\_{k,j}}}$

其中分母项为文档j中所有词的个数。

IDF与包含某个词i的文档在所有文档集中出现的频率负相关，即：

$idf\_i=log\frac{|D|}{|\{j:t\_i\in d\_j\}|+1}$

式中$|D|$表示语料库中的文档数，$|\{j:t\_i\in d\_j\}|$表示语料库中包含了词$t\_i$的文档数，分母中加1是为了防止出现分母为0的情况。由IDF的计算公式可知，如果包含某个词$t\_i$的文档数越少，说明该词具有越大的类别区分能力（或者称为“代表性”）。

IDF值是根据语料库学习到的，通常情况下它能够找到类别区分能力大的词，但是在特殊情况下会存在一些不足：如果有某一类文档中包含了词$t\_i$的文档有m个，而其他包含了词$t\_i$的文档有k个，则包含$t\_i$的文档总数应该有$n=m+k$个，假设m值很大，则n也会随之变大，又假设语料库中的文档总数$|D|$很接近n，则根据上述IDF的计算公式可以得出词$t\_i$的$idf\_i$很小，它的类别区分能力不大。但是实际上词$t\_i$很有可能是该类文档中最有代表性的词，应该赋予更高的权重。出现这种情况的原因有一部分是因为语料库的规模不够大，无法充分模拟自然语言中的情况，因而在实践中，常常需要一个规模很大的语料库进行IDF值的统计计算，以此增强TF-IDF的表达能力。

在实际运用中，首先由语料库统计出各词汇的IDF值，然后分别统计每一篇文档中每个词的TF值，最后进行乘积求出各词汇的TF-IDF值：

$tf-idf\_{i,j}=tf\_{i,j}\*idf\_i$

TF-IDF值能够给出某个词在某篇文档中的重要度，有利于找出具备类别区分度的词汇。但是，仅仅依靠词频关系来决定词汇的重要度，会存在一定的局限性：一个词在文档中的重要程度，不仅仅与出现的频率有关，有时还会与相应的上下文、词出现的位置以及词汇的具体用法（词性）等有关。因此在使用向量空间模型对文本进行建模时，特征项的权重除了考虑TF-IDF值之外，还会同时兼顾其他指标，以获取更好的表达效果。

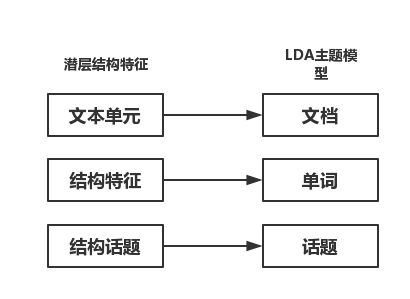
* 1. 文本的潜层结构特征

本文提出一种基于潜层结构特征（Latent Structure Feature）的文本表示方法，试图寻找一种能够表示“文本风格”的文本表达方法。

如上文所述，文本数据经过预处理后可以得到分词结果、词性标注结果、TF-IDF结果等表层的、易观察到的特征，利用向量空间模型可以将文本投射到一个高维的空间向量上。在此我们称这些表层的、易观察到的特征为“表层特征”

所谓的“文本风格”，可以认为是某一类文本单元（如句子或篇章）之间所共享的一种结构上、用词上的特征。文本单元在结构上的特征包括了：各种词性所占的比例、词性种类的丰富度、句子的长度等；文本单元在用词上的特征包括了：是否经常使用该类别的代表词、是否经常使用新颖词等。因为上述特征表达了文本“遣词造句”的水平，在某种角度上揭示了文本单元的结构特性，我们可以将上述特征称为“结构特征”。“结构特征”可以通过简单地统计“表层特征”获取，因此可以认为“结构特征”是“表层特征”的一种同层级衍生。

为了寻求更高层面上“文本风格”的表达方式，我们假设属于同一种“文本风格”的文本单元在“表层特征”上存在一定的相关度，或者存在一定的隐层关联，我们的目的就是找到这些体现“文本风格”的特征关系。回顾上文介绍的LDA主题模型，我们得知主题模型假定了一篇文档的生成过程，即：首先根据一定规则选取一个话题分布，然后重复地根据话题分布选取出一个话题，并从话题中随机选取出一个单词组成文档。单词组成了文档，每一个单词都可能属于多种话题，每个话题下都会有多个相对重要的代表性单词，话题体现了单词之间的联系。

类比LDA主题模型，假设一个文本单元（句子或篇章）是由若干个“结构特征”或“表层特征”所组成的，这些“结构特征”或“表层特征”在某个潜在层面（Latent Level）上具有联系，假设称这些“联系”为“结构话题”，则每个结构特征都可以同时属于多种结构话题（但是会有一定的偏好倾向），每种结构话题下都包含了多种结构特征（但是重要性会有所区别）。由此我们可以假设出一种文本单元结构特征的生成过程：在生成一个文本单元时，首先根据一定的规则选取一个结构话题分布（即我们需要生成一个符合哪种“文本风格”的文本单元），然后重复地根据结构话题分布选取出一个结构话题，并据此结构话题选取出一个结构特征，组成文本单元。

将上述的类比应用到现有的LDA主题模型中，我们可以根据文本单元的结构特征推断出相应的结构话题分布。假设定义的结构话题总数为k，则最终每个文本单元i的结构话题分布就是一个k维的向量$v\_i=(t\_1,t\_2,…,t\_k)$，该向量的每一个分量代表了具体某个结构话题在该文本单元内的分布比例（0到1之间的值）。

在此我们将表示文本单元i结构话题分布的k维向量称为文本单元i的潜层结构特征（Latent Structure Feature）。利用空间向量模型与潜层结构特征相结合进行文本表示时，可以认为在超空间上位置相近的两个文本具有相似的文本风格。

* 1. 文本潜层结构特征的获取

文本潜层特征结构相当于LDA主题模型中的话题分布，因此在获取方法上与原始LDA主题模型的话题分布获取方法类似，即通过大量的样本数据进行参数估计，得到与话题分布及话题-字词分布相关的两个狄利克雷分布的参数$\alpha$和$\eta$，然后根据待测试文档的字词词频，利用吉布斯采样求出相应的话题分布、话题-字词分布。

在进行类比转换时，需要解决的一大难题就是如何将结构特征转换为LDA中的字词，即利用词袋模型表示文本单元的结构特征。LDA主题模型中的字词（word）是一个不可再分的最小数据处理单元，而在词袋模型的描述中，字词（word）除了“数量”之外不具备其他的属性，并且拥有规模巨大的“词典”，保证了字词的丰富度。相反，上文提出的结构特征往往是连续值，并且特征数量很少，不适合直接套用词袋模型。

因此需要对连续值的结构特征进行离散化处理，并将离散化后的结构特征运用于词袋模型中。

* + 1. 结构特征的离散化

连续特征的离散化是数据挖掘、文本挖掘、机器学习领域经常会遇到的一项任务。常用的离散化方法包括了等值划分离散化与等量划分离散化。等值划分离散化是指将特征值按照值域进行均分，每一段的取值等同处理，离散值之间的间隔是固定的。例如某一个连续值的值域为[0,10]，现在需要划分为5段，则根据值域的上下限进行5等分，得出$[0,2], [2,4],...,[8,10]$等5段。等值划分离散化方法简单、便于实现，但是如果样本集的特征值分布不均，则有可能造成某些数据段的样本数很多而其他数据段样本数很少的情况，进而导致该特征的区分度降低。为了避免样本数据倾斜的问题，可以采用等量划分离散化的方法。等量划分离散化首先对所有样本的特征值进行排序，根据样本数量进行均分，并以划分段的特征值上下限作为该划分段的边界，划分段的边界间隔会根据样本密度而改变。等量划分离散化能够利用样本的统计分布信息，提高特征的区分度，但是在具体实现时需要记录各划分段的边界信息，在实现复杂度上比等值划分离散化更大。

为了确保离散化后结构特征的种类丰富度，我们推荐采用等量划分的方法进行特征离散化。

* + 1. 针对离散化结构特征的词袋模型

在此运用一个最简单的例子说明如何构建基于离散化结构特征的词袋模型：

假设有一个文本单元（句子）i，它的结构特征(Structure Feature)为$sf\_i =(f\_1,f\_2,f\_3)=(0.123, 0.456, 0.789)$，假设样本的特征分布很均匀，采用了等值离散化，划分间隔为0.001，则该文本单元i的离散化结构特征（Discrete Structure Feature）为$dsf\_i=(d\_1,d\_2,d\_3)=(123,1456,2789)$。可以注意到$d\_1$、$d\_2$和$d\_3$在离散化后的取值上依次加上一个1000的偏移量，这是因为离散划分间隔为0.001，即$d\_1$有1000个划分段，在“词典”中占据0到999号的位置，以此类推$d\_2$在“词典”中占据1000到1999号位置，$d\_3$占据2000到2999号位置。离散化后的结构特征词典规模与离散化的划分间隔有关，通常我们希望词典中的大部分“单词”能被所有“文档”大概率地使用，即词典中的“生僻词”越少越好，因此等量划分离散化是最好的选择。结合上面的例子，我们可以给出文本单元i相应的词袋描述——文本单元i中包含了“词典编号”为123、1456、2789的三个“单词”，并且每个“单词”只出现了一次，记为：$ D\_i=\{(123,1), (1456,1), (2789,1)\}$。

综上所述，我们根据所有样本集的离散化结构特征构建了一套“结构特征词典”，并利用离散化后的结果（分段编号）作为结构特征在“词典”中的索引编号。在离散化结构特征上运用词袋模型时，可以认为每一个文本单元包含了若干种“结构特征词典”中的“字词”，其中每种“字词”在文本单元中出现的次数为1。由此构建了离散化结构特征的词袋模型后，即可以运用LDA主题模型的相关理论及算法进行话题分布的推断。

* 1. 本章小结

本章主要提出了文本的潜层结构特征的概念，首先介绍了现阶段流行的文本表示方法——空间向量模型，并概述了空间向量模型中特征项权重的一种重要算法算法TF-IDF；然后通过类比LDA主题模型的方式提出了文本的潜层结构特征，“潜层”是相对于TF-IDF、词性分布等“表层”特征而言的，该潜层结构特征能够在一定程度上体现出“文本风格”；最后探讨了如何获取文本的潜层结构特征，涉及到结构特征的离散化以及结构特征词袋模型的建立。

文本的潜层结构特征是一种更高层次的、更为抽象的文本表示特征，有理由相信它能够包含更多的文本结构信息，因此会在文本相似计算、文本分类、文本聚类等任务上有良好的表现。接下来会利用潜层结构特征做一些文本挖掘的任务，具体而言是利用潜层结构特征进行文本表示，然后针对在线评论的重要度挖掘任务进行一定的实践探索。

# 第四章 在线评论重要度预测

本章，我们将会针对在线评论（Online Comments）这一互联网上的特殊文本形式进行文本挖掘。

* 1. 针对在线评论的文本表示

在线评论（Online Comments或Product Reviews）是互联网当中常见的一种文本形式，相比于新闻文本、微博文本等其他常见的文本形式，在线评论存在其独有的特点：话题指向性、群体相关性等，因此在对在线评论进行文本表示时，可以充分考虑其独有的特点。在线评论的话题指向性是指：同一个商品或新闻（或者微博）下的在线评论，其内容的话题都指向一个具体的事件或事物，因此其在用词（事件相关或事物相关的词）和内容（评价一个东西的好坏、评论对事件的看法等）上都十分的集中并有明确的指向；在线评论的群体相关性是指：同一个商品或新闻（或者微博）下的评论，相互之间会有紧密的联系，如存在相互佐证、相互反对等关系，在分析评论信息时应该充分利用评论之间的相关性。

4.2 基于有监督方法的评论重要度预测

4.3 基于图的评论重要度预测

4.4 实验结果对比

第五章 总结与展望