第十四章　概率主题模型

概率主题模型是一系列旨在发现隐藏在大规模文档中的主题结构的算法。

在第三章（第22页）中，我们介绍了一元语言模型。给定了词的概率分布之后，我们可以通过如下的方法产生一篇长度为N的文档：

1. 根据词的概率分布，随机抽取一个词；
2. 重复1，直到产生N个词。

这种产生文档的方法也叫作**生成式模型**。在一元语言模型中，产生的每一个词概率都是独立的，和这个词出现在文档的位置无关，也和这个词的前面和后面的词无关。

通过这方法我们可以产生很多篇文档。但是，每一篇文档都是从相同的词分布中产生的，即使在不考虑词序的情况下，这样生成的文档也是杂乱无章的，这和人们"创作"一篇文章的过程是不相符的。当一个作者"创作"一篇文章时，他首先会确定一个或几个主题，然后用和文章主题相关的词来写一篇文章。比如，在一篇讨论足球的文章里，会出现很多"比赛"、"战术"等词，而很少会出现"股票"以及"银行"等词；相反在讨论经济的文章里，出现"股票"以及"银行"等词的可能性要比"比赛"、"战术"等词大很多。

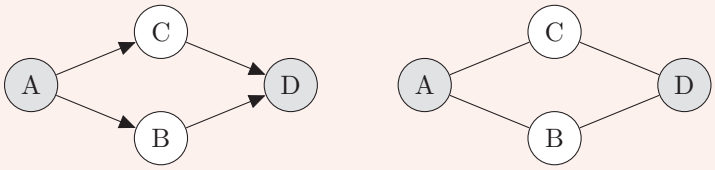
一个改进的方法是引入一个表示主题的离散随机变量，每个主题有一个不同的词分布。我们可以通过如下的方法产生一篇长度为N的文档：

1. 根据主题的概率分布，随机抽取一个主题；
2. 根据该主题对应的词分布，随机抽取一个词；
3. 重复2，直到产生N个词。

机器学习 | 概率图模型

概率图模型（Graphical Models）是用图形来表示概率模型的方法，是概率论和图论相结合的。概率图模型提供了一种简单的可视化概率模型的方法，有利于将复杂的概率模型分解为简单模型的组合，以便更好地了解概率模型的表示、推理、学习等方法[Jordan, 1998]。

下图给出了两个代表性图模型的例子：有向图和无向图，分别表示了四个变量$\{A, B, C, D\}$之间的依赖关系。图中的阴影圆圈表示可观测变量，非阴影圆圈表示潜在变量，边表示两变量间的条件依赖关系。



(a) 有向图 (b) 无向图

这个模型称为**一元混合语言模型**。一元语言模型可以看作是只有一个主题的一元混合语言模型。

这里，我们引入了**主题**（英文是topic而不是subject或theme）的概念。在主题模型中，主题是指一个词的概率分布，并不是一般意义上"中心思想"或"主要内容"。不同的主题对应不同的概率分布。

一元混合语言模型有一个缺点就是一篇文章只能有一个主题，这也不符合我们的直观感受。我们可以再改变一下文档的产生过程：

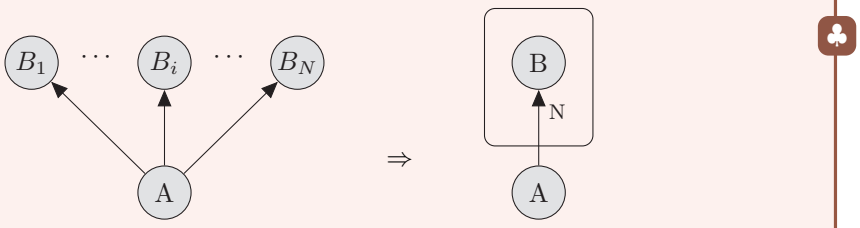
1. 根据主题的概率分布，随机抽取一个主题；
2. 根据该主题对应的词分布，随机抽取一个词；
3. 重复1、2，直到产生N个词。

这个生成模型就我们本章所讨论的**概率主题模型**（Probabilistic Topic Model）。

本章先介绍概率主题模型的基本概念，然后详细介绍下使用非常广泛的Latent Dirichlet Allocation （LDA）模型。

机器学习 | 概率图模型 | 盘子表示法

盘子表示法（plate notation）是图模型中表示重复变量的方法。对于重复变量，我们可以把他们都单独画出来，但是这样不够简洁。盘子表示法是用一个盘子（矩形）把重复的变量设为一组，作为一个子图，然后用一个数来表示子图重复的次数。



(a) 普通表示 (b) 盘子表示

14.1 概率主题模型

概率主题模型一个文档集的生成式概率模型，其基本思想是每一篇文档是一个或多个不可观察的主题来生成的，每个主题都有一个相应的词的分布。

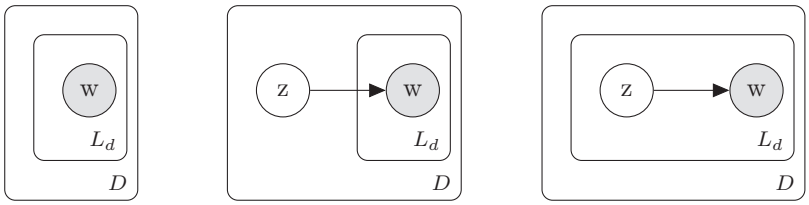
在介绍概率主题模型之前，我们首先定义几个概念：

- **词典**：所有词的集合，每个词在词典中有一个索引 $\mathcal{V} = \{1, \cdots, V\}$
- **文档**：一个文档是有 N 个词组成的序列，表示为 $W = (w_1, w_2, \cdots, w_N)$，其中 $w_i \in \mathcal{V}$ 为离散变量。
- **文档集**：$\mathcal{D} = \{W^1, \cdots, W^D\}$。第 d 篇文档可以表示为 $W^d = (w_1^d, w_2^d, \cdots, w_{L_d}^d)$，其中 L_d 为第 d 篇文档的长度。
- **主题**：一个主题 z 是一个离散随机变量，取值为 $\{1, \cdots, T\}$。每个主题都决定了一个词概率分布。

图14.1给出了一元语言模型、一元混合语言模型和概率主题模型的对比。这些模型对应的概率分布为：

对于一篇文档 d，其的概率分布为：一元语言模型：

$$P(W^d) = \prod_{i=1}^{L_d} p(w_i^d) \tag{14.1}$$



(a) 一元语言模型 (b) 一元混合语言模型 (c) 概率主题模型

图 14.1: 不同语言模型的图表示

一元混合语言模型：

$$P(W^d) = \sum_{z^d} p(z^d) \prod_{i=1}^{L_d} p(w_i^d | z^d), \tag{14.2}$$

其中，z^d 为文档 d 的主题，为离散随机变量。

概率主题模型:

$$P(W^d) = \prod_{i=1}^{L_d} \sum_{z_i^d} p(z_i^d) p(w_i^d | z_i^d), \tag{14.3}$$

其中，z_i^d 为文档 d 中第 i 个词的主题，为离散随机变量。

从上面三个公式可以看出，每个模型都假设了每个词是独立产生的，和它的上下文无关，其本质上都是一元模型。

在公式14.3中，并没有给 $p(z_i^d)$ 和 $p(w_i^d|z_i^d)$ 赋予任何的函数形式或某个特定的分布。假设文档 d 对应一个主题分布 θ^d，θ^d 是在主题集合 $\{1, \cdots, T\}$ 上的分布函数。θ_t^d 是文档 d 属于主题 t 的概率，并满足 $\theta_t^d \geq 0, t = 1, \cdots, T$，并且 $\sum_t \theta_t^d = 1$。

这样，每个字的主题 z_i^d 对应的概率 $p(z_i^d)$ 可以重新表示为 $p(z_i^d|\theta^d)$。

$p(w_i^d|z_i^d)$ 表示给定主题 z_i^d 条件下词 w_i^d 后对应的概率。主题 z_i^d 的取值为 $\{1, \cdots, T\}$。每个主题 t 都决定了一个词的分布 ϕ_v^t，满足 $\phi_v^t \geq 0, v = 1, \cdots, V$，并且 $\sum_v \phi_v^t = 1$。

因此，$p(w_i^d|z_i^d)$ 可以重新表示为 $p(w_i^d|z_i^d, \phi^{z_i^d})$ 或 $p(w_i^d|z_i^d, \phi^1, \cdots, \phi^T)$。

要注意的是这里涉及到两类概率分布：

- 一类是 d 个主题分布 θ^d，记为 Θ，$\Theta = \{\theta^1, \cdots, \theta^D\}$。
- 另一类是 T 个词分布 ϕ^t，记为 Φ，$\Phi = \{\phi^1, \cdots, \phi^T\}$。

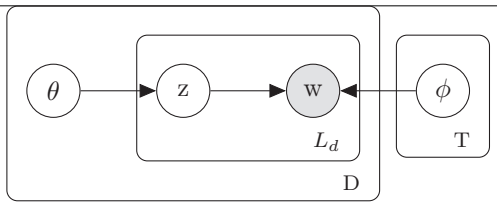


图 14.2: 概率主题模型的图模型表示

这样，完整的概率主题模型的联合概率分布为：

$$P(\Theta, \Phi, W, Z) = \prod_d p(\theta^d) \prod_t p(\phi^t) \prod_d \prod_i p(z_i^d|\theta^d) p(w_i^d|z_i^d, \Phi), \tag{14.4}$$

其中，$Z = \{z_i^d\}$，$W = \{w_i^d\}$，$i = \{1, \cdots, L_d\}$，$d = \{1, \cdots, D\}$。公式14.4中，W 是可观测变量，Θ, Φ, Z 为不可观测变量。图14.3给出了对应的图模型表示。

$$\prod_i p(z_i^d|\theta^d) = \prod_i \theta_{z_i^d}^d \tag{14.5}$$

$$= \prod_t (\theta_t^d)^{m_t^d}, \tag{14.6}$$

其中，m_t^d 为文档 d 中属于主题 t 的词的个数。

$$\prod_d \prod_i p(w_i^d|z_i^d, \Phi) = \prod_d \prod_i p(w_i^d|z_i^d, \phi^{z_i^d}) \tag{14.7}$$

$$= \prod_d \prod_i p(w_i^d|\phi^{z_i^d}), \tag{14.8}$$

$$= \prod_d \prod_i \phi_{w_i^d}^{z_i^d}, \tag{14.9}$$

$$= \prod_t \prod_v (\phi_v^t)^{n_v^t}, \tag{14.10}$$

其中，n_v^t 为整个文档集中属于主题 t 的词的个数。

将公式14.6和14.10代入到14.4，得到

$$P(\Theta, \Phi, W, Z) = \prod_d p(\theta^d) \prod_t p(\phi^t) \prod_d \prod_t (\theta_t^d)^{m_t^d} \prod_t \prod_v (\phi_v^t)^{n_v^t}, \tag{14.11}$$

现在给定一个文档集，我们希望估计公式14.11中参数 Θ, Φ, Z，这些参数的后验概率为：

$$P(\Theta, \Phi, Z|W) = \frac{P(\Theta, \Phi, Z, W)}{P(W)} \tag{14.12}$$

为了计算参数Θ, Φ, Z的后验概率，需要首先给出$p(\theta^d)$以及$\prod_t p(\phi^t)$的具体形式，即Θ, Φ的先验概率。从公式14.6和14.10看出，主题分布和给定主题下词的分布都为多项分布，因此，我们设$p(\theta^d)$以及$\prod_t p(\phi^t)$的具体形式为多项分布的共轭分布：Dirichlet分布。这时的概率主题模型就叫做**潜在狄利克雷分配**（Latent Dirichlet Allocation）模型。

14.2 预备知识：Dirichlet 分布

在介绍LDA模型前，我们必须先理解两个分布：多项式分布和Dirichlet分布。关于多项式分布以及相关的概率论知识在第三章（第22 页）中已经介绍过了。这里我们只重点介绍下Dirichlet分布。

Γ函数和Γ分布

Γ函数：

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx \tag{14.13}$$

性质：

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha) \tag{14.14}$$

特殊值：

$$\Gamma(1) = 1 \tag{14.15}$$

$$\Gamma(n+1) = n! \tag{14.16}$$

$$\Gamma(\frac{1}{2}) = \pi \tag{14.17}$$

Γ分布：

$$\Gamma(x|\alpha) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}, x > 0, \alpha > 0 \tag{14.18}$$

均值：

$$\mu(x) = \alpha \tag{14.19}$$

方差：

$$\sigma^2(x) = \alpha^2 \tag{14.20}$$

B 函数和 B 分布

B 函数：

$$\mathbf{B}(\alpha,\beta)=\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx,\alpha>0,\beta>0 \tag{14.21}$$

$$\mathbf{B}(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{14.22}$$

B 分布：

$$\mathbf{B}(x|\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}, \tag{14.23}$$

均值：

$$\mu(x)=\frac{\alpha}{\alpha+\beta} \tag{14.24}$$

方差：

$$\sigma^2(x)=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{14.25}$$

性质：如果 $x_1\sim\Gamma(\alpha)$ 和 $x_2\sim\Gamma(\beta)$，那么

$$Y=\frac{x_1}{x_1+x_2}\sim\mathbf{B}(x|\alpha,\beta) \tag{14.26}$$

Dirichlet 函数和 Dirichlet 分布

Dirichlet 函数

$$\mathbf{Dir}(\alpha_1,\cdots,\alpha_K)=\int_{x_1,\cdots,x_K}\cdots\int x_1^{\alpha_1-1}\cdots x_K^{\alpha_K-1}dx_1\cdots dx_K, \tag{14.27}$$

其中，$0<x_1,\cdots,x_K<1$，$\sum_{i=1}^K x_i=1$。

Dirichlet 函数是多参数的 **B** 函数。$K=2$ 时，$\mathbf{Dir}=\mathbf{B}$.

性质：

$$\mathbf{Dir}(\alpha_1,\cdots,\alpha_K)=\frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}{\Gamma(\alpha_1+\cdots+\alpha_K)} \tag{14.28}$$

Dirichlet 分布

$$\mathbf{Dir}(x_1,\cdots,x_K|\alpha_1,\cdots,\alpha_K)=\frac{\Gamma(\alpha_1+\cdots+\alpha_K)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}x_1^{\alpha_1-1}\cdots x_K^{\alpha_K-1}, \tag{14.29}$$

$$= \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}x_k^{\alpha_k-1}, \tag{14.30}$$

$$\mu(x_i) = \frac{\alpha_i}{\alpha_1+\cdots+\alpha_K} \tag{14.31}$$

$$\sigma^2(x_i) = \frac{\alpha_i(\alpha_1+\cdots+\alpha_K-alpha_i)}{(\alpha_1+\cdots+\alpha_K)^2(\alpha_1+\cdots+\alpha_K+1)} \tag{14.32}$$

$$\sigma(x_i,x_j) = \frac{\alpha_i\alpha_j}{(\alpha_1+\cdots+\alpha_K)^2(\alpha_1+\cdots+\alpha_K+1)} \tag{14.33}$$

性质 1：

$x_1 \sim \Gamma(\alpha_1),\cdots,x_K \sim \Gamma(\alpha_K)$，那么

$$y_i = \frac{x_i}{x_1+\cdots+x_K} \sim \mathbf{Dir}() \tag{14.34}$$

性质 2：边际分布

性质 3：分组

14.3 潜在狄利克雷分配模型

潜在狄利克雷分配（Latent Dirichlet Allocation）模型，简称 **LDA 模型**，

$$P(\Theta,\Phi,W,Z) = \prod_d p(\theta^d)\prod_t p(\phi^t)\prod_d\prod_t(\theta_t^d)^{m_t^d}\prod_t\prod_v(\phi_v^t)^{n_v^t}, \tag{14.35}$$

假设 θ^d 是 T 维离散随机向量，服从 Dirichlet 分布。

$$p(\theta^d) = \frac{\Gamma(\sum_{t=1}^{T}\alpha_t)}{\prod_{t=1}^{T}\Gamma(\alpha_t)}\prod_{t=1}^{T}(\theta_t^d)^{\alpha_t-1}, \tag{14.36}$$

这里 $\alpha_t, t=1,cdots,T$ 是 Dirichlet 分布的参数。为了减少模型复杂度，这里对所有 α_t 取相同值

$$p(\theta^d) = \frac{\Gamma(T\alpha)}{(\Gamma(\alpha))^T}\prod_{t=1}^{T}(\theta_t^d)^{\alpha-1}, \tag{14.37}$$

假设 ϕ^t 是 V 维离散随机向量，也服从 Dirichlet 分布，并设分布中所有的参数都等于 β，得到

$$p(\phi^t) = \frac{\Gamma(V\beta)}{(\Gamma(\beta))^V}\prod_{v=1}^{V}(\phi_v^t)^{\beta-1}, \tag{14.38}$$

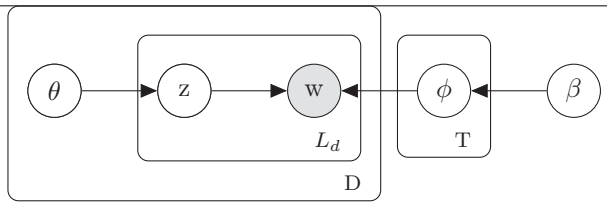


图 14.3: LDA 模型的图模型表示

将公式14.38和14.37代入14.36，得到

$$P(\Theta,\Phi,W,Z|\alpha,\beta)=\left(\prod_d\frac{\Gamma(T\alpha)}{(\Gamma(\alpha))^T}\prod_{t=1}^{T}(\theta_t^d)^{\alpha-1}\right)\left(\prod_t\frac{\Gamma(V\beta)}{(\Gamma(\beta))^V}\prod_{v=1}^{V}(\phi_v^t)^{\beta-1}\right)\cdot\left(\prod_d\prod_t(\theta_t^d)^{m_t^d}\right)\left(\prod_t\prod_v(\phi_v^t)^{n_v^t}\right), \tag{14.39}$$

$$=\left(\prod_d\frac{\Gamma(T\alpha)}{(\Gamma(\alpha))^T}\prod_{t=1}^{T}(\theta_t^d)^{m_t^d+\alpha-1}\right)\left(\prod_t\frac{\Gamma(V\beta)}{(\Gamma(\beta))^V}\prod_{v=1}^{V}(\phi_v^t)^{n_v^t+\beta-1}\right). \tag{14.40}$$

公式14.40就是LDA 模型的联合概率的最终表示形式，α,β 为超参数。因为假设 Θ 和 Φ 的先验分布都是Dirichlet 分布，最终的联合概率形式就变得非常简单。

算法14.1给出 LDA 生成模型的产生过程。

算法 14.1: LDA 生成模型的产生过程

输入: 一组是 D 个"文档-主题"分布 Θ，另一组是 T 个"主题-单词"分布 Φ

输出: W

```
for d = 1 ··· D do
    采样 θ^d ~ Dir(α) ;
    for i = 1 ··· L_d do
        采样主题 z ~ Multi(θ^d) ;
        采样词 w_i^d ~ Multi(ϕ_z) ;
    end
end
return W ;
```

14.3.1 推断

在公式14.40中W是可观测变量，Θ、Φ和Z是不可观察的，需要从数据中推断得到。

$$P(\Theta,\Phi,Z|W,\alpha,\beta)=\frac{P(\Theta,\Phi,W,Z|\alpha,\beta)}{P(W|\alpha,\beta)} \tag{14.41}$$

$$=\frac{P(\Theta,\Phi,W,Z|\alpha,\beta)}{\sum_Z P(Z|W,\alpha,\beta)} \tag{14.42}$$

其中

$$P(Z|W,\alpha,\beta)=\int_\Theta\int_\Phi P(\Theta,\Phi,W,Z|\alpha,\beta)d\Theta d\Phi. \tag{14.43}$$

从公式14.41可以看到这里涉及到多个变量的积分，要精确推断Θ、Φ和Z的后验分布是十分困难的。因此，需要通过近似推断的方法来求解。推断方法主要有变分方法和Gibbs采样方法。变分方法是通过不断优化调整目标函数的下界来进行。因为变分方法的理论分析涉及比较多的背景知识，我们不再详细介绍，可以阅读第14.5节中的参考文献来深入了解。这里，我们详细介绍下基于Gibbs采样来LDA模型中参数推断的方法。

14.4 基于Gibbs采样的LDA模型推断方法

LDA模型有两组参数需要推断：一组是D个"文档-主题"分布Θ，另一组是T个"主题-单词"分布Φ。通过学习这两个参数，可以知道文档作者感兴趣的主题，以及每篇文档所涵盖的主题比例等。

14.5 参考文献和深入阅读

概率主题模型是一种统计语言模型，是用统计方法对自然语言进行建模，每一篇文档代表了一些主题所构成的一个概率分布，而每一个主题又代表了很多单词所构成的一个概率分布。概率主题模型属于非监督学习方法，可以用来识别大规模文档集中潜藏的主题信息。概率主题模型采用了词袋模型，本质上是一元语言模型，忽略了词与它所在上下文的关系。关于概率主题模型的深入介绍可以参考[Anthes, 2010, Blei, 2012]。

Hofmann [1999b]在潜在语义分析模型基础上，提出了**概率潜在语义分析模型**（Probabilistic Latent Semantic Analysis，PLSA），也叫做**概率潜在语义索引模型**（Probabilistic Latent Semantic Indexing）Hofmann [1999a]，引入了主题的概率来对文档进行建模。在PLSA模型中，没有考虑参数的先验概率，使得PLSA在参数估计时会存在一些问题，比如数据稀疏。Blei et al. [2003]提出了LDA模型，假设参数的先验分布为Dirichlet分

算法 14.2: 基于Gibbs采样的LDA模型推断方法

输入: 文档集: W，迭代次数：K

输出: M, N

```
for k = 1 ··· K do
    for d = 1 ··· D do
        for i = 1 ··· L_d do
            计算 t = z_i^d，v = w_i^d;
            修改 m_t^d = m_t^d − 1，n_v^t = n_v^t − 1;
            生成 s ∼;
            修改 z_i^d = s;
            修改 M, N;
        end
    end
end
return M, N ;
```

布。其初始的参数估计算法是使用变分法推导，用EM算法求解，得到一个局部最优解。Griffiths and Steyvers [2004]采用了Gibbs采样的方法来估计LDA模型中参数。因为基于Gibbs 采样的LDA 模型非常容易实现，训练过程只需要简单的计算就可以得到很好的结果，因此在信息检索、文本挖掘等领域非常流行。

当然，LDA模型也存在很多不足。一是在LDA模型中，主题之间也几乎是不相关的，这与很多实际问题并不相符，因此，提出了。

LDA模型涉及到贝叶斯理论、Dirichlet分布、多项分布、图模型、变分推断、EM算法、Gibbs采样等知识。关于概率图模型和参数估计的知识可以参考[Bishop, 2006]。关于Gibbs采样以及更基本的Markov chain Monte Carlo (MCMC)方法可以参考Andrieu et al. [2003]。

目前LDA的各种语言（C，Java，Matlab，Python等）实现都可以在网络上找到。LDA的提出者Blei的C语言版本，可以在http://www.cs.princeton.edu/~blei/lda-c/ 下载。

参考文献

王厚峰. 指代消解的基本方法和实现技术. 中文信息学报, 16(006):9–17, 2002.

宗成庆. 统计自然语言处理（第二版）. 中文信息处理丛书. 清华大学出版社, 2013.

J. Allen. Natural language understanding. Benjamin-Cummings Publishing Co., Inc. Redwood City, CA, USA, 1995.

Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to mcmc for machine learning. Machine learning, 50(1-2):5–43, 2003.

Gary Anthes. Topic models vs. unstructured data. Commun. ACM, 53(12):16–18, December 2010. ISSN 0001-0782.

C.M. Bishop. Pattern recognition and machine learning. Springer New York., 2006.

David M Blei. Probabilistic topic models. Communications of the ACM, 55(4):77–84, 2012.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. J. Mach. Learn. Res., 3:993–1022, March 2003. ISSN 1532-4435. doi: http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993. URL http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993.

Sabine Buchholz and Erwin Marsi. Conll-x shared task on multilingual dependency parsing. In Proceedings of the Tenth Conference on Computational Natural Language Learning, pages 149–164. Association for Computational Linguistics, 2006.

N. Chomsky and G.A. Miller. Introduction to the formal analysis of natural languages. Wiley, 1963.

Yoeng-Jin Chu and Tseng-Hong Liu. On shortest arborescence of a directed graph. Scientia Sinica, 14(10):1396, 1965.

Michael Collins. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, 2002.

Susan P Converse. Resolving pronominal references in chinese with the hobbs algorithm. In Proceedings of the 4th SIGHAN workshop on Chinese language processing, pages 116–122, 2005.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. Journal of Machine Learning Research, 7:551–585, 2006.

R. Duda, P. Hart, and D. Stork. Pattern Classification. Wiley, New York, 2nd edition, 2001. ISBN 0471056693.

Jack Edmonds. Optimum branchings. Journal of Research of the National Bureau of Standards B, 71(233-240):160, 1967.

Yoav Freund and Robert E Schapire. Large margin classification using the perceptron algorithm. Machine learning, 37(3):277–296, 1999.

Jeffrey Friedl. Mastering regular expressions. O'Reilly Media, Inc., 2006.

Haim Gaifman. Dependency Systems and Phrase-Structure Systems. Information and Computation/information and Control, 8:304–337, 1965. doi: 10.1016/S0019-9958(65)90232-9.

Niyu Ge, John Hale, and Eugene Charniak. A statistical approach to anaphora resolution. In Proceedings of the sixth workshop on very large corpora, pages 161–170, 1998.

Thomas L Griffiths and Mark Steyvers. Finding scientific topics. Proceedings of the National academy of Sciences of the United States of America, 101(Suppl 1):5228–5235, 2004.

T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning. Springer, New York, 2001.

Jerry R Hobbs. Resolving pronoun references. Lingua, 44(4):311–338, 1978.

T. Hofmann. Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pages 50–57, 1999a.

Thomas Hofmann. Probabilistic latent semantic analysis. In Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, pages 289–296. Morgan Kaufmann Publishers Inc., 1999b.

John E Hopcroft. 自动机理论, 语言和计算导论. 清华大学出版社, 2002.

R. Hudson. English Word Grammar. Basil Blackwell, Oxford, 1990.

T. Joachims. Text categorization with suport vector machines: Learning with many relevant features. Proc. of Euro. Conf. on Mach. Learn. (ECML), pages 137–142, 1998.

M.I. Jordan. Learning in Graphical Models. Kluwer Academic Publishers, 1998.

D. Jurafsky, J.H. Martin, A. Kehler, K. Vander Linden, and N. Ward. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition, volume 2. Prentice Hall New Jersey, 2000.

Christopher Kennedy and Branimir Boguraev. Anaphora for everyone: pronominal anaphora resoluation without a parser. In Proceedings of the 16th conference on Computational linguistics, pages 113–118. Association for Computational Linguistics, 1996.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the Eighteenth International Conference on Machine Learning, 2001.

Shalom Lappin and Herbert J Leass. An algorithm for pronominal anaphora resolution. Computational linguistics, 20(4):535–561, 1994.

C.D. Manning and H. Schütze. Foundations of statistical natural language processing. MIT Press, 1999.

A. McCallum, D. Freitag, and F. Pereira. Maximum entropy markov models for information extraction and segmentation. In Proceedings of the Seventeenth International Conference on Machine Learning, pages 591–598. Citeseer, 2000.

Ryan McDonald, Koby Crammer, and Fernando Pereira. Online large-margin training of dependency parsers. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL '05, pages 91–98, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1219840.1219852. URL http://dx.doi.org/10.3115/1219840.1219852.

I.A. Melčuk. Dependency syntax: theory and practice. State University of New York Press, 1988.

R. Mihalcea and P. Tarau. Textrank: Bringing order into texts. In Proceedings of EMNLP, volume 4, pages 404–411. Barcelona: ACL, 2004.

T.M. Mitchell. Machine learning. Burr Ridge, IL: McGraw Hill, 1997.

Ruslan Mitkov. Anaphora resolution: the state of the art. Citeseer, 1999.

Vincent Ng. Supervised noun phrase coreference research: The first fifteen years. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1396–1411, 2010.

H. Nikula. Dependensgrammatik. LiberFörlag, 1986.

Jens Nilsson, Sebastian Riedel, and Deniz Yuret. The conll 2007 shared task on dependency parsing. In Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL, pages 915–932. sn, 2007.

Joakim Nivre. An efficient algorithm for projective dependency parsing. In Proceedings of the 8th International Workshop on Parsing Technologies (IWPT). Citeseer, 2003.

Joakim Nivre and Jens Nilsson. Pseudo-projective dependency parsing. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05),

pages 99–106, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P/P05/P05-1013.

F. Peng, F. Feng, and A. McCallum. Chinese segmentation and new word detection using conditional random fields. Proceedings of the 20th international conference on Computational Linguistics, 2004.

L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2):257–286, 1989.

Jane J Robinson. Dependency structures and transformational rules. Language, pages 259–285, 1970.

F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological review, 65(6):386–408, 1958.

G. Salton, A. Wong, and CS Yang. A vector space model for automatic indexing. Communications of the ACM, 18(11):613–620, 1975.

F. Sebastiani. Machine learning in automated text categorization. ACM computing surveys, 34(1):1–47, 2002.

C. Sutton and A. McCallum. An introduction to conditional random fields for relational learning. Introduction to statistical relational learning, page 93, 2007.

Ben Taskar, Carlos Guestrin, and Daphne Koller. Max-margin markov networks. In In proceedings of the 17th Annual Conference on Neural Information Processing Systems, Whistler, B.C., Canada, 2003.

Lucien Tesnière. Eléments de syntaxe structurale. Librairie C. Klincksieck, 1959.

I. Tsochantaridis, T. Hofmann, T. Joachims, and Y Altun. Support vector machine learning for interdependent and structured output spaces. In Proceedings of the International Conference on Machine Learning(ICML), 2004.

F. Xia. The part-of-speech tagging guidelines for the penn Chinese treebank (3.0), 2000.

N. Xue. Chinese word segmentation as character tagging. Computational Linguistics and Chinese Language Processing, 8(1):29–48, 2003.

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. Natural language engineering, 11(2): 207–238, 2005.

H. Yamada and Y. Matsumoto. Statistical dependency analysis with support vector machines. In Proceedings of the International Workshop on Parsing Technologies (IWPT), volume 3, 2003.

Y. Yang and J.O. Pedersen. A comparative study on feature selection in text categorization. In Proc. of Int. Conf. on Mach. Learn. (ICML), volume 97, 1997.

R.B. Yates and B.R Neto. Modern Information Retrieval. New York: ACM Press Series/Addison Wesley, 1999.

索引