潜在语义分析(LSA)

①无监督学习方法：一种文本降维方法

②通过矩阵分解发现：文本-单词-话题关系

③在1990年提出，在推荐系统、文本搜索、图像处理中都有应用

1. 单词向量空间与话题向量空间

1.1. 单词向量空间

word vector space model：给定一个文本，用一个向量表示文本"语义"向量每一维对应一个单词，取值是频率或权重，向量内积表示文本相似度。

单词-文本矩阵：

(1) 文本集合 $D=\{d_1,d_2,\cdots,d_n\}$，单词集合 $W=\{w_1,w_2,\cdots,w_m\}$

(2) 单词-文本矩阵可表示成 $X=(x_{ij})\in R^{m\times n}$，其中 x_{ij} 表示单词 w_i 在文本 d_j 中出现的频数或权重

权重通常用TF-IDF(单词频率-逆文本频率)表示

$$\mathrm{TF\text{-}IDF}_{ij}=\frac{tf_{ij}}{tf_j}\log\frac{df}{df_i}$$

其中，tf_{ij} 为单词 w_i 在文本 d_j 中出现的频数，$tf_j=\sum_i tf_{ij}$

df_i 为包含单词 w_i 的文本数，$df=n$ 是总文本数

单词文本矩阵的j列 $x_j=(x_{1j},x_{2j},\cdots x_{nj})^T$ 表示 d_j 的信息，d_i，d_j 之间的相似度为 $x_i\cdot x_j$ 或 $\frac{x_i\cdot x_j}{\|x_i\|\|x_j\|}$ (余弦)

(1) 模型优点：模型简单，计算效率高

(2) 模型局限：不能准确刻画语义相似度，自然语言有一词多义和多词一义。

1.2 话题向量空间

文本的语义相似度可以用两者之间的"话题"相似度表示

假设文本共有K个话题 ($K<<m$) 每个话题由定义在单词集合W上的n维向量表示：$t_l=(t_{1l},t_{2l},\cdots,t_{ml})^T$，$t_{il}$表示单词$w_i$在话题$t_l$上的权重

得到K个话题 $t_1,t_2,\cdots,t_k$ 张成一个K维话题向量空间.

文本d_j在单词向量空间是x_j，将x_j投影到话题向量空间T中得到

$y_j=(y_{lj},\ 1\le l\le k)^T\in R^k$，$y_{ij}$代表文本$d_j$在话题$t_i$上的权重

话题文本矩阵 $Y=(y_1,y_2,\cdots,y_n)\in R^{k\times n}$

$x_j\approx y_{1j}\cdot t_1+y_{2j}\cdot t_2+\cdots+y_{kj}\cdot t_k$ (即x_j可以通过k个话题进行线性近似)

用单词-话题矩阵T以及话题-文本矩阵Y近似单词-文本矩阵X：

$X\approx TY \quad T\in R^{m\times k},\ Y\in R^{k\times n}$

文本相似度 ① 单词 $x_i\cdot x_j$

② 话题空间 $y_i\cdot y_j$

2. 潜在语义分析算法

2.1 矩阵奇异值分解

$X=U\Sigma V^T \quad X\in R^{m\times n} \quad n\le m$

$=\sum_{i=1}^n \sigma_i u_i v_i^T$

U 左特征向量矩阵

$\Sigma=diag\{\sigma_1,\sigma_2,\cdots,\sigma_n\}$

V 右特征向量矩阵

根据话题个数K对X进行截断的奇异值分解

$X\approx U^{(k)}\Sigma^{(k)}V^{(k)}$，则 $T=U^{(k)}$，$Y=\Sigma^{(k)}V^{(k)T}$

2.2 非负矩阵分解算法

给定一个非负$X\ge0$(所有元素非负) 找到两个非负矩阵 $W\in R^{m\times k}_{\ge0}$ 与 $H\in R^{k\times n}_{\ge0}$

$X\approx WH$，由于$K<\min\{m,n\}$，非负矩阵分解是对原数据的压缩

1. 损失函数

(1) 平方损失 $\|A-B\|_F^2=\sum_{ij}(a_{ij}-b_{ij})^2$

(2) 散度(divergence) 散度损失函数

$$D(A\|B)=\sum_{ij}\left(a_{ij}\log\frac{a_{ij}}{b_{ij}}-a_{ij}+b_{ij}\right)$$

① 不对称，在 $A=B$ 时取下界

② $0\log 0=0$

当 $\sum_{ij}a_{ij}=\sum_{ij}b_{ij}=1$ 时，即K-L散度，$D(A\|B)=\sum_{ij}a_{ij}\log\frac{a_{ij}}{b_{ij}}$

注：K-L散度用于度量两个概率分布之间的差异 $D_{KL}(P\|Q)=E_{x\sim p}\left(\log\frac{P(x)}{Q(x)}\right)$

非负矩阵分解转换为带约束的优化问题 $\min\limits_{W,H}\|X-WH\|_F^2$ or $\min\limits_{W,H}D(X\|WH)$

2. 算法

非负矩阵分解的优化算法，迭代对W与H分别进行优化

定理：平方损失对以下乘法更新规则 $H_{lj}\leftarrow H_{lj}\frac{(W^TX)_{lj}}{(W^TWH)_{lj}}$ $W_{il}\leftarrow W_{il}\frac{(XH^T)_{il}}{(WHH^T)_{il}}$

是非增的，当且仅当W和H是损失的稳定点时函数损失不变

$$J(W,H)=\frac{1}{2}\|X-WH\|_F^2=\frac{1}{2}\sum_{ij}(X_{ij}-(WH)_{ij})^2$$

$$\frac{\partial J(W,H)}{\partial W_{il}}=-\sum_j\{X_{ij}-(WH)_{ij}\}H_{lj}=-\{(XH^T)_{il}-(WHH^T)_{il}\}$$

$$\frac{\partial J(W,H)}{\partial H_{lj}}=-\{(W^TX)_{lj}-(W^TWH)_{lj}\}$$

$$W_{il}\leftarrow W_{il}+\lambda_{il}\{(XH^T)_{il}-(WHH^T)_{il}\},\quad \lambda_{il}=\frac{W_{il}}{(WHH^T)_{il}}$$

$$H_{lj}\leftarrow H_{lj}+u_{lj}\{(W^TX)_{lj}-(W^TWH)_{lj}\},\quad u_{lj}=\frac{H_{lj}}{(W^TWH)_{lj}}$$

(1) 在算法更新时，选取初始矩阵非负，更新中可保证W.H非负

(2) 每次迭代对W的列进行归一化为单位向量