潜在语义分析
Latent Semantic Analysis

西安交通大学管理学院
信息管理与电子商务系
智能决策与机器学习研究中心
刘佳鹏

问题引出

- 文本信息处理，比如文本信息检索、文本数据挖掘的一个核心问题是对文本的语义内容进行表示，并进行文本之间的语义相似度计算
- 最简单的方法是利用单词向量空间模型（word vector space model）
  - 基本想法：给定一个文本，用一个向量表示该文本的“语义”，向量的每一维对应一个单词，其数值为该单词在该文本中出现的频数或权值
  - 基本假设：文本中所有单词的出现情况表示了文本的语义内容
  - 向量空间的度量，如内积或标准化内积表示文本之间的“语义相似度”

问题引出

- 例如，文本信息检索的任务是，用户提出查询时，帮助用户找到与查询最相关的文本，以排序的形式展示给用户
- 一个最简单的做法是采用单词向量空间模型，将查询与文本表示为单词的向量，计算查询向量与文本向量的内积作为语义相似度，以这个相似度的高低对文本进行排序
- 在这里，查询被看成是一个伪文本，查询与文本的语义相似度表示查询与文本的相关性

问题引出

- 含有n个文本的集合$D = \{d_1, d_2, \cdots, d_n\}$
- 所有文本中出现的m个单词的集合$W = \{w_1, w_2, \cdots, w_m\}$
- 将单词在文本中出现的数据用一个单词-文本矩阵（word-document matrix）表示，记作$\boldsymbol{X}$

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

  元素x_{ij}表示单词w_i在文本d_j中出现的频数或权值
- 由于单词的种类很多，而每个文本中出现单词的种类通常较少，所以单词-文本矩阵是一个稀疏矩阵。

问题引出

- 权值通常用单词频率-逆文本频率 (term frequency-inverse document frequency, TF-IDF)表示，其定义是

$$\mathrm{TFIDF}_{ij} = \frac{\mathrm{tf}_{ij}}{\mathrm{tf}_{\cdot j}} \log \frac{\mathrm{df}}{\mathrm{df}_i}, \quad i = 1, 2, \cdots, m; \quad j = 1, 2, \cdots, n$$

  式中tf_{ij}是单词w_i出现在文本d_j中的频数，$\mathrm{tf}_{\bullet j}$是文本d_j中出现的所有单词的频数之和，df_i是含有单词w_i的文本数，df是文本集合D的全部文本数
- 一个单词在一个文本中出现的频数越高，这个单词在这个文本中的重要度就越高
- 单词在整个文本集合中出现的文本数越少，这个单词就越能表示其所在文本的特点，重要度就越高
- 一个单词在一个文本的TF-IDF是两种重要度的积，表示综合重要度

问题引出

- 单词向量空间模型直接使用单词-文本矩阵$\boldsymbol{X}$的信息
- 单词-文本矩阵$\boldsymbol{X}$的第j列向量$\boldsymbol{x}_j$表示文本d_j

$$\boldsymbol{x}_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{mj} \end{bmatrix}, \quad j = 1, 2, \cdots, n$$

其中x_{ij}是单词w_i在文本d_j的权值，$i = 1, 2, \cdots, m$，权值越大，该单词在该文本中的重要度就越高

- 这时矩阵$\boldsymbol{X}$也可以写作$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_n \end{bmatrix}$

问题引出

- 两个单词向量的内积或标准化内积（余弦）表示对应的文本之间的语义相似度
- 因此，文本d_i与d_j之间的相似度为

$$\boldsymbol{x}_i \cdot \boldsymbol{x}_j, \qquad \frac{\boldsymbol{x}_i \cdot \boldsymbol{x}_j}{\|\boldsymbol{x}_i\| \, \|\boldsymbol{x}_j\|}$$

  式中$\cdot$表示向量的内积，$\|\cdot\|$表示向量的范数
- 直观上，在两个文本中共同出现的单词越多，其语义内容就越相近，这时，对应的单词向量同不为零的维度就越多，内积就越大（单词向量元素的值都是非负的），表示两个文本在语义内容上越相似

问题引出

- 单词向量空间模型虽然简单，却能很好地表示文本之间的语义相似度，与人们对语义相似度的判断接近，在一定程度上能够满足应用的需求，至今仍在文本信息检索、文本数据挖掘等领域被广泛使用，可以认为是文本信息处理的个基本原理
- 单词向量空间模型的优点是模型简单，计算效率高
  - 因为单词向量通常是稀疏的，两个向量的内积计算只需要在其同不为零的维度上进行即可，需要的计算很少，可以高效完成

问题引出

- 单词向量空间模型也有一定的局限性，体现在内积相似度未必能够准确表达两个文本的语义相似度上
- 因为自然语言的单词具有一词多义性（polysemy）及多词一义性（synonymy），即同一个单词可以表示多个语义，多个单词可以表示同个语义，所以基于单词向量的相似度计算存在不精确的问题

问题引出

	d_1	d_2	d_3	d_4
airplane	2			
aircraft		2		
computer			1	
apple			2	3
fruit				1
produce	1	2	2	1

单词-文本矩阵例

- 单词向量空间模型认为文本d_1与d_2相似度并不高，尽管两个文本的内容相似，这是因为同义词“airplane”与“aircraft”被当作了两个独立的单词
  - 单词向量空间模型不考虑单词的同义性，在此情况下无法进行准确的相似度计算

问题引出

	d_1	d_2	d_3	d_4
airplane	2			
aircraft		2		
computer			1	
apple			2	3
fruit				1
produce	1	2	2	1

单词-文本矩阵例

- 单词向量空间模型认为文本d_3与d_4有一定的相似度，尽管两个文本的内容并不相似，这是因为这是因为单词“apple”具有多义，可以表示“apple computer”和“fruit”
  - 单词向量空间模型不考虑单词的多义性，在此情况下也无法进行准确的相似度计算

话题向量空间

- 两个文本的语义相似度可以体现在两者的话题相似度上
- 所谓话题（topic），并没有严格的定义，就是指文本所讨论的内容或主题
- 一个文本一般含有若干个话题。如果两个文本的话题相似，那么两者的语义应该也相似
- 话题可以由若干个语义相关的单词表示，同义词（如“airplane”与“aircraft”）可以表示同一个话题，而多义词（如“apple”）可以表示不同的话题
- 这样，基于话题的模型就可以解决上述基于单词的模型存在的问题

话题向量空间

- **话题向量空间：** 给定一个文本，用话题空间的一个向量表示该文本，该向量的每一分量对应一个话题，其数值为该话题在该文本中出现的权值
- 用两个向量的内积或标准化内积表示对应的两个文本的语义相似度
- 注意话题的个数通常远远小于单词的个数，话题向量空间模型更加抽象

话题向量空间

- 三个矩阵：（1）**单词-文本矩阵$\boldsymbol{X}$**

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

  $\boldsymbol{X}$构成原始的单词向量空间，每一列是一个文本在单词向量空间中的表示
- 矩阵$\boldsymbol{X}$也可以写作$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_n \end{bmatrix}$

话题向量空间

- 假设所有文本共含有k个话题，每个话题由一个定义在单词集合W上的m维向量表示，称为话题向量，即

$$\boldsymbol{t}_l = \begin{bmatrix} t_{1l} \\ t_{2l} \\ \vdots \\ t_{ml} \end{bmatrix}, \quad l = 1, 2, \cdots, k$$

  其中t_{il}是单词w_i在话题t_l的权值，$i = 1, 2, \cdots, m$，权值越大，该单词在该话题中的重要度就越高
- 这k个话题向量$t_1, t_2, \cdots, t_k$张成一个话题向量空间（topic vector space），维数为k

话题向量空间

- 三个矩阵：（2）单词-话题矩阵 $\boldsymbol{T}$

$$\boldsymbol{T}=\begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1k} \\ t_{21} & t_{22} & \cdots & t_{2k} \\ \vdots & \vdots & & \vdots \\ t_{m1} & t_{m2} & \cdots & t_{mk} \end{bmatrix}$$

矩阵 $\boldsymbol{T}$ 也可以写作 $\boldsymbol{T}=\begin{bmatrix} t_1 & t_2 & \cdots & t_k \end{bmatrix}$

话题向量空间

- 考虑文本集合D的文本d_j，在单词向量空间中由一个向量$\boldsymbol{x}_j$表示，将$\boldsymbol{x}_j$投影到话题向量空间T中，得到在话题向量空间的一个向量$\boldsymbol{y}_j$
- $\boldsymbol{y}_j$是一个k维向量，其表达式为

$$\boldsymbol{y}_j = \begin{bmatrix} y_{1j} \\ y_{2j} \\ \vdots \\ y_{kj} \end{bmatrix}, \quad j = 1, 2, \cdots, n$$

其中$\boldsymbol{y}_{lj}$是文本d_j在话题$\boldsymbol{t}_l$的权值，$l = 1, 2, \cdots, k$，权值越大，该话题在该文本中的重要度就越高

话题向量空间

- 三个矩阵：（3）**话题-文本矩阵**$\boldsymbol{Y}$

$$
\boldsymbol{Y}=\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & & \vdots \\ y_{k1} & y_{k2} & \cdots & y_{kn} \end{bmatrix}
$$

矩阵$\boldsymbol{Y}$也可以写作$\boldsymbol{Y}=\begin{bmatrix} \boldsymbol{y}_1 & \boldsymbol{y}_2 & \cdots & \boldsymbol{y}_n \end{bmatrix}$

话题向量空间

- 这样一来，在单词向量空间的文本向量$\boldsymbol{x}_j$可以通过它在话题空间中的向量$\boldsymbol{y}_j$近似表示，具体地由k个话题向量以$\boldsymbol{y}_j$为系数的线性组合近似表示

$$\boldsymbol{x}_j \approx y_{1j}\boldsymbol{t}_1 + y_{2j}\boldsymbol{t}_2 + \cdots + y_{kj}\boldsymbol{t}_k, \quad j = 1, 2, \cdots, n$$

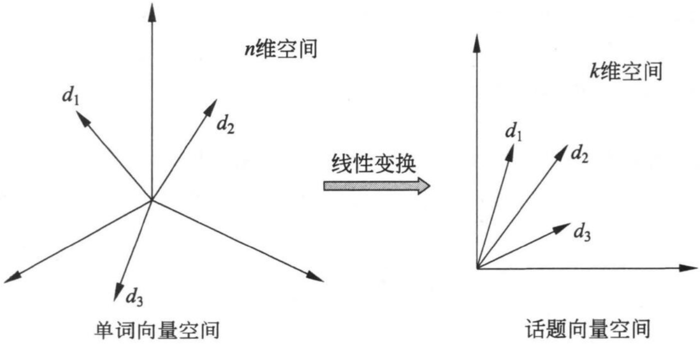
- 所以，单词-文本矩阵$\boldsymbol{X}$可以近似的表示为单词-话题矩阵$\boldsymbol{T}$与话题-文本矩阵$\boldsymbol{Y}$的乘积形式

$$\boldsymbol{X} \approx \boldsymbol{T}\boldsymbol{Y}$$

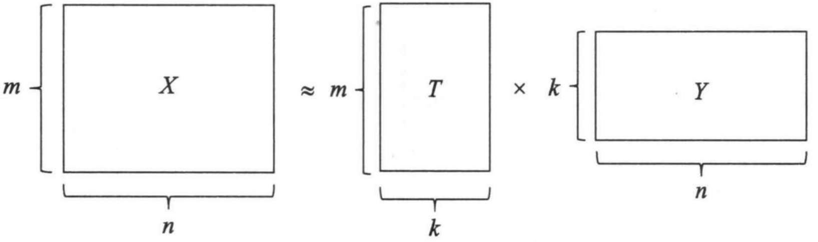
- 这就是**潜在语义分析**

话题向量空间

- 直观上潜在语义分析是将文本在单词向量空间的表示通过线性变换转换为在话题向量空间中的表示



话题向量空间



潜在语义分析通过矩阵因子分解实现，单词-文本矩阵 X 可以近似的表示为单词-话题矩阵 T 与话题-文本矩阵 Y 的乘积形式

- 在原始的单词向量空间中，两个文本d_i与d_j的相似度可以由对应的向量的内积表示，即$\boldsymbol{x}_i \cdot \boldsymbol{x}_j$
- 经过潜在语义分析之后，在话题向量空间中，两个文本d_i与d_j的相似度可以由对应的向量的内积即$\boldsymbol{y}_i \cdot \boldsymbol{y}_j$表示

潜在语义分析的矩阵奇异值分解算法

- 要进行潜在语义分析，需要同时决定两部分的内容，一是话题向量空间$\boldsymbol{T}$，二是文本在话题空间的表示$\boldsymbol{Y}$，使两者的乘积是原始矩阵数据的近似
- 潜在语义分析可以利用矩阵奇异值分解，具体地，对单词-文本矩阵进行奇异值分解，将其左矩阵作为话题向量空间，将其对角矩阵与右矩阵的乘积作为文本在话题向量空间的表示
  - 其他算法包括非负矩阵分解等（见李航《统计学习方法》(第2版)）

潜在语义分析的矩阵奇异值分解算法

- 准备文本集合$D = \{d_1, d_2, \cdots, d_n\}$和$W = \{w_1, w_2, \cdots, w_m\}$
- 潜在语义分析首先将这些数据表成一个单词-文本矩阵

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

这是一个$m \times n$矩阵，元素x_{ij}表示单词w_i在文本d_j中出现的频数或权值

潜在语义分析的矩阵奇异值分解算法

- 潜在语义分析根据确定的话题个数k对单词-文本矩阵X进行截断奇异值分解

$$
\boldsymbol{X} \approx \boldsymbol{U}_k \boldsymbol{\Sigma}_k \boldsymbol{V}_k^{\mathrm{T}} = \begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \cdots & \boldsymbol{u}_k \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 \\ & & \ddots & \\ 0 & 0 & & 0 \\ 0 & 0 & 0 & \sigma_k \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1^{\mathrm{T}} \\ \boldsymbol{v}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{v}_k^{\mathrm{T}} \end{bmatrix}
$$

式中$k \leqslant n \leqslant m$，$\boldsymbol{U}_k$是$m \times k$矩阵，它的列由$\boldsymbol{X}$的前$k$个互相正交的左奇异向量组成，$\boldsymbol{\Sigma}_k$是$k$阶对角方阵，对角元素为前$k$个最大奇异值，$\boldsymbol{V}_k$是$n \times k$矩阵，它的列由$\boldsymbol{X}$的前$k$个互相正交的右奇异向量组成

潜在语义分析的矩阵奇异值分解算法

- 矩阵$\boldsymbol{U}_k$的每一个列向量$\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_k$表示一个话题，称为话题向量
- 由这k个话题向量张成一个子空间

$$\boldsymbol{U}_k = \left[\begin{array}{cccc} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \cdots & \boldsymbol{u}_k \end{array} \right]$$

称为话题向量空间

潜在语义分析的矩阵奇异值分解算法

- 有了话题向量空间，接着考虑文本在话题空间的表示

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_n \end{bmatrix} \approx \boldsymbol{U}_k \boldsymbol{\Sigma}_k \boldsymbol{V}_k^{\mathrm{T}}$$

$$= \begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \cdots & \boldsymbol{u}_k \end{bmatrix} \begin{bmatrix} \sigma_1 & & & & \\ & \sigma_2 & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \sigma_k \end{bmatrix} \begin{bmatrix} v_{11} & v_{21} & \cdots & v_{n1} \\ v_{12} & v_{22} & \cdots & v_{n2} \\ \vdots & \vdots & & \vdots \\ v_{1k} & v_{2k} & \cdots & v_{nk} \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \cdots & \boldsymbol{u}_k \end{bmatrix} \begin{bmatrix} \sigma_1 v_{11} & \sigma_1 v_{21} & \cdots & \sigma_1 v_{n1} \\ \sigma_2 v_{12} & \sigma_2 v_{22} & \cdots & \sigma_2 v_{n2} \\ \vdots & \vdots & & \vdots \\ \sigma_k v_{1k} & \sigma_k v_{2k} & \cdots & \sigma_k v_{nk} \end{bmatrix}$$

潜在语义分析的矩阵奇异值分解算法

- 由上式可知，矩阵$\boldsymbol{X}$的第j列向量$\boldsymbol{x}_j$满足

$$
\begin{aligned}
\boldsymbol{x}_j &\approx \boldsymbol{U}_k\left(\Sigma_k \boldsymbol{V}_k^{\mathrm{T}}\right)_j \\
&= \begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \cdots & \boldsymbol{u}_k \end{bmatrix} \begin{bmatrix} \sigma_1 v_{j1} \\ \sigma_2 v_{j2} \\ \vdots \\ \sigma_k v_{jk} \end{bmatrix} \\
&= \sum_{l=1}^{k} \sigma_l v_{jl} u_l, \quad j = 1, 2, \cdots, n
\end{aligned}
$$

式中$\left(\Sigma_k \boldsymbol{V}_k^{\mathrm{T}}\right)_j$是矩阵$\left(\Sigma_k \boldsymbol{V}_k^{\mathrm{T}}\right)$的第$j$列向量

潜在语义分析的矩阵奇异值分解算法

- 上式是文本d_j的近似表达式，由k个话题向量u_l的线性组合构成
- 矩阵$(\Sigma_k V_k^{\mathrm{T}})$的每一个列向量

$$\begin{bmatrix} \sigma_1 v_{11} \\ \sigma_2 v_{12} \\ \vdots \\ \sigma_k v_{1k} \end{bmatrix}, \quad \begin{bmatrix} \sigma_1 v_{21} \\ \sigma_2 v_{22} \\ \vdots \\ \sigma_k v_{2k} \end{bmatrix}, \cdots, \quad \begin{bmatrix} \sigma_1 v_{n1} \\ \sigma_2 v_{n2} \\ \vdots \\ \sigma_k v_{nk} \end{bmatrix}$$

分别是各文本在话题向量空间的表示

潜在语义分析的矩阵奇异值分解算法

- 综上，可以通过对单词-文本矩阵的奇异值分解进行潜在语义分析
$$\boldsymbol{X} \approx \boldsymbol{U}_k \Sigma_k \boldsymbol{V}_k^{\mathrm{T}} = \boldsymbol{U}_k \left(\Sigma_k \boldsymbol{V}_k^{\mathrm{T}}\right)$$
得到话题空间$\boldsymbol{U}_k$以及文本在话题空间的表示$\left(\Sigma_k \boldsymbol{V}_k^{\mathrm{T}}\right)$

潜在语义分析示例

- 假设有9个文本，11 个单词，单词-文本矩阵X为11×9矩阵，矩阵的元素是单词在文本中出现的频数，表示如下：

Index Words	Titles								
	T1	T2	T3	T4	T5	T6	T7	T8	T9
book			1	1					
dads						1			1
dummies		1						1	
estate							1		1
guide	1					1			
investing	1	1	1	1	1	1	1	1	1
market	1		1						
real							1		1
rich						2			1
stock	1		1					1	
value				1	1				

潜在语义分析示例

- 进行潜在语义分析：实施对矩阵的截断奇异值分解，假设话题的个数是 3，矩阵的截断奇异值分解结果为

Book	0.15	−0.27	0.04
Dads	0.24	0.38	−0.09
Dummies	0.13	−0.17	0.07
Estate	0.18	0.19	0.45
Guide	0.22	0.09	−0.46
Investing	0.74	−0.21	0.21
Market	0.18	−0.30	−0.28
Real	0.18	0.19	0.45
Rich	0.36	0.59	−0.34
Stock	0.25	−0.42	−0.28
Value	0.12	−0.14	0.23

∗

3.91	0	0
0	2.61	0
0	0	2.00

∗

T1	T2	T3	T4	T5	T6	T7	T8	T9
0.35	0.22	0.34	0.26	0.22	0.49	0.28	0.29	0.44
−0.32	−0.15	−0.46	−0.24	−0.14	0.55	0.07	−0.31	0.44
−0.41	0.14	−0.16	0.25	0.22	−0.51	0.55	0.00	0.34

潜在语义分析示例

- 将Σ_3与$\boldsymbol{V}_3^{\mathrm{T}}$相乘，整体变成两个矩阵乘积的形式

$$
X \approx U_3(\Sigma_3 V_3^{\mathrm{T}})
$$

$$
= \begin{bmatrix} 0.15 & -0.27 & 0.04 \\ 0.24 & 0.38 & -0.09 \\ 0.13 & -0.17 & 0.07 \\ 0.18 & 0.19 & 0.45 \\ 0.22 & 0.09 & -0.46 \\ 0.74 & -0.21 & 0.21 \\ 0.18 & -0.30 & -0.28 \\ 0.18 & 0.19 & 0.45 \\ 0.36 & 0.59 & -0.34 \\ 0.25 & -0.42 & -0.28 \\ 0.12 & -0.14 & 0.23 \end{bmatrix} \begin{bmatrix} 1.37 & 0.86 & 1.33 & 1.02 & 0.86 & 1.92 & 1.09 & 1.13 & 1.72 \\ -0.84 & -0.39 & -1.20 & -0.63 & -0.37 & 1.44 & 0.18 & -0.81 & 1.15 \\ -0.82 & 0.28 & -0.32 & 0.50 & 0.44 & -1.02 & 1.10 & 0.00 & 0.68 \end{bmatrix}
$$

- 矩阵$\boldsymbol{U}_3$有3个列向量，表示 3 个话题，矩阵$\boldsymbol{U}_3$表示话题向量空间。矩阵$(\Sigma_3 \boldsymbol{V}_3^{\mathrm{T}})$有9个列向量，表示9个文本，矩阵$(\Sigma_3 \boldsymbol{V}_3^{\mathrm{T}})$是文本集合在话题向量空间的表示