

TF-IDF

TF-IDF is a measure of originality of a word by comparing the number of times a word appears in a doc with the number of docs the word appears in.

$$\text{TF-IDF} = \text{TF}(t, d) \times \text{IDF}(t)$$

Term frequency

Inverse document frequency

Number of times term t appears in a doc, d

$\frac{1}{1 + \log \left(\frac{\# \text{ of documents}}{n} \right)}$

Dong-Qing Wei, Yifeng Liu et al, *Front. Bioeng. Biotechnol*, 2020. | IF=5.125



Dipeptide Frequency of Word Frequency and Graph Convolutional Networks for DTA Prediction

Xianfang Wang^{1,2}, Yifeng Liu², Fan Lu², Hongfei Li², Peng Gao² and Dongqing Wei³*

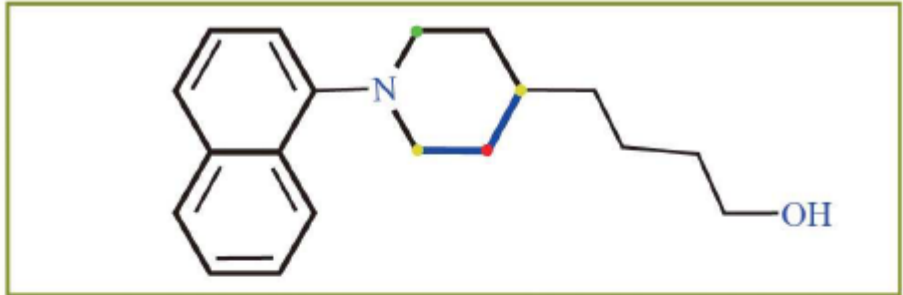
¹ School of Computer Science and Technology, Henan Institute of Technology, Xinxiang, China, ² School of Computer and Information Engineering, Henan Normal University, Xinxiang, China, ³ School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China

Outline

- Drug Molecular Feature Extraction
- Protein Sequence Feature Extraction
- Network Model Construction
- Results and Discussion

Drug Molecular Feature Extraction

Input: Drug Molecule



Feature Extraction

Node V_i	Edge E_i
<ul style="list-style-type: none">1. Atomic class2. Atomic rank3. The total number of hydrogen atoms4. Implied value of atoms5. The existence or absence of aromatic group	Whether there is a bonding bonds with adjacent atoms

Output: Graph Structure

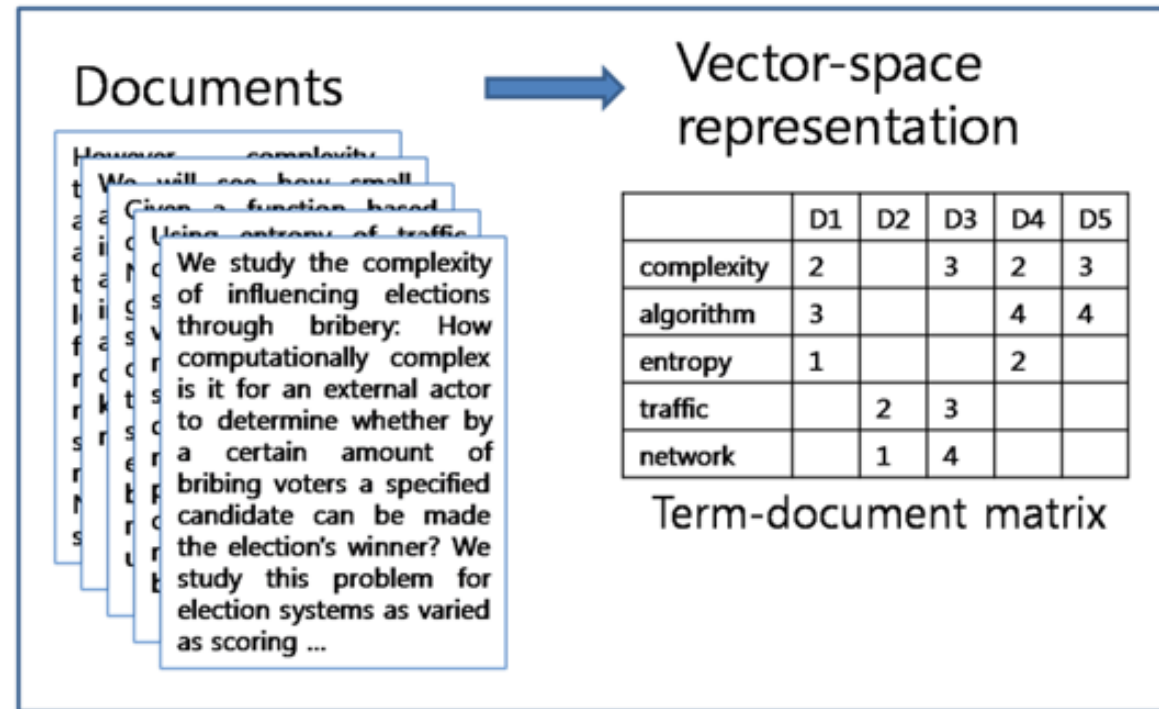
Graph $G=[V,E]$

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots \\ x_{81} & x_{82} & x_{83} \end{bmatrix}$$

(8, 8) (8, 3)

Protein Sequence Feature Extraction

- term frequency - inverse document frequency (TF-IDF) algorithm is employed.
- This algorithm plays an important role in natural language process (NLP)
- polypeptide frequency is similar to the calculation process of TF in bioinformatics.



term	df _t	idf _t
car	18,165	1.65
auto	6723	2.08
insurance	19,241	1.62
best	25,235	1.5

$$idf_j = \log \left[\frac{n}{df_j} \right]$$

Protein Sequence Feature Extraction

- term frequency - inverse document frequency (TF-IDF) algorithm is employed.
- This algorithm plays an important role in natural language process (NLP)
- polypeptide frequency is similar to the calculation process of TF in bioinformatics.

TF-IDF

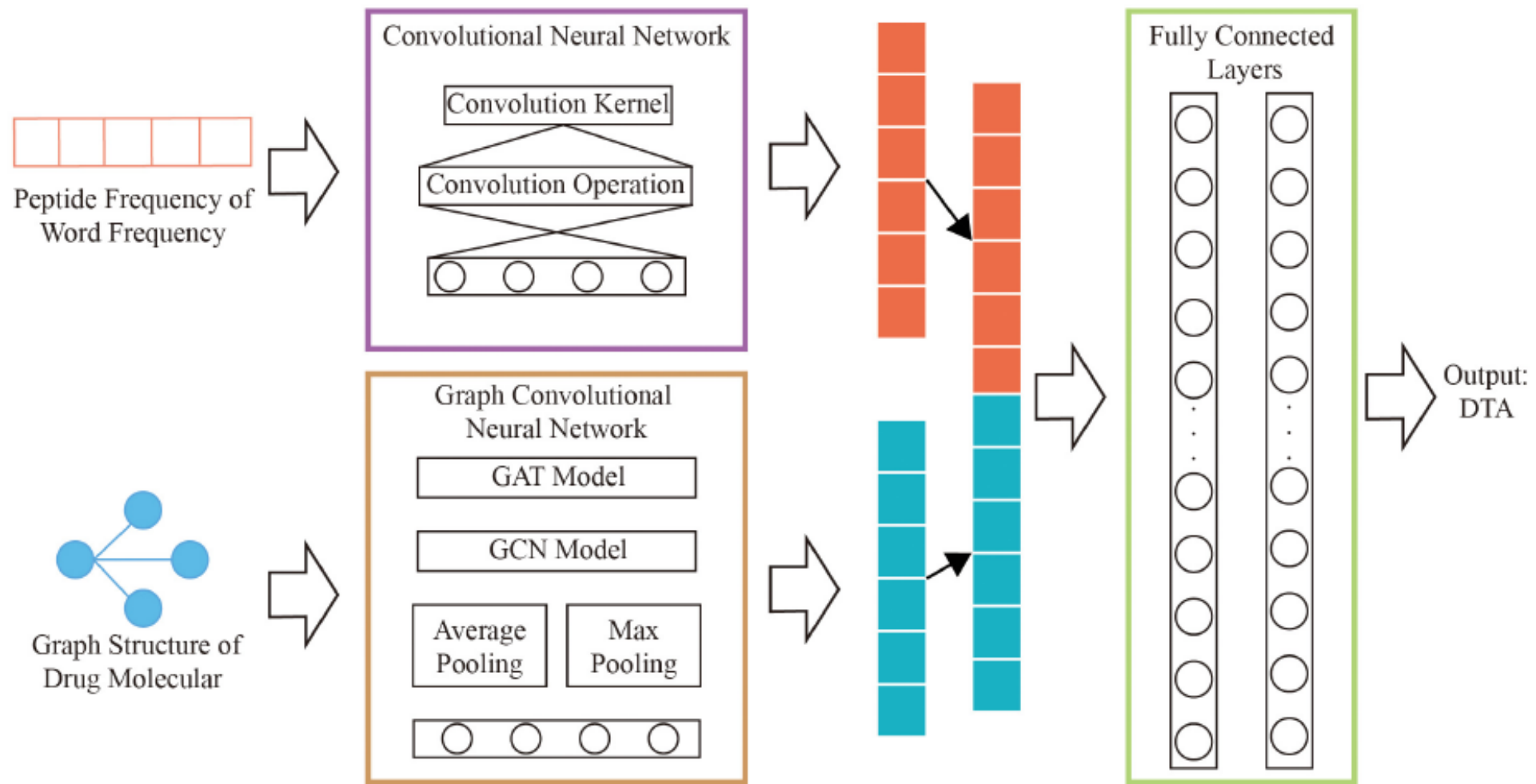
TF-IDF is a measure of originality of a word by comparing the number of times a word appears in a doc with the number of docs the word appears in.

$$\text{TF-IDF} = \text{TF}(t, d) \times \text{IDF}(t)$$

Annotations for the formula:

- TF(t, d):** Term frequency. Number of times term t appears in a doc, d .
- IDF(t):** Inverse document frequency. $\log \frac{1 + n}{1 + \text{df}(d, t)}$
 - n : # of documents
 - $\text{df}(d, t)$: Document frequency of the term t

Network Model Construction



Result and Discussion

Features	KIBA		Davis	
	MSE	CI	MSE	CI
DeepDTA	0.194	0.863	0.261	0.878
WipeDTA	0.179	0.875	0.262	0.886
[This model	0.126	0.901	0.220	0.899

