《Rumor Detection on Social Media with Event Augmentations》

本文主要是面向谣言检测，做了三种数据增强方式，每种增强方式由不同的问题考虑。同时使用了对比自监督学习来克服对labeled数据的依赖。

代码连接：https://github.com/hzy-hzy/RDEA

abstract

网络数据的快速发展，谣言检测很重要；

深度学习的高层特征表示提取能力很强，但是需要大量的标注数据用于训练，**意味着时间消耗和数据不高效**。

数据增强的方式：三种通过修改响应特征和时间结构以提取传播模式，同时学习用户（root用户）参与的内在表示。

使用对比自监督学习方式以实现事件增强的高效执行，同时缓解数据有限的问题。

引言

传播渠道的广泛应用的同时，也帮助了谣言的传播；

谣言检测的常见方法是尝试利用**谣言的内容和传播结构**，因为通常认为这两者能够**反映谣言本身的特征和传播模式**；

当前方法的不足是：严重依赖监督学习，这就意味着对标注数据的依赖很严重。

同时，【4】提出了一个数据增强技术：通过使用语境词汇表示ELMo，但是**它却忽视了事件的传播特征**。

Motivated by this（对标注数据的依赖和数据增强技术的可行性），提出了自监督学习框架RDEA：final predictation。

- 三个事件增强策略：节点掩码，子图和图边丢失；
- permute内容的特征和传播结构以生成事件的正例；**由此，数据的内在联系被用来生成自监督学习的signals，同时通过增强事件的对比预训练以增强事件表示event representation**。
- 使用label以精调模型得到最终的prediction。

方法论

一个数据集由多个事件构成，一个事件由这个事件的所有post和post之间的图结构联系。而每个post的表示由一个one-hot向量构成。

事件增强

谣言事件的一个特征是：一个谣言的post和comments是内在联系，同时彼此独立。**因此，数据增强操作应该为其量身定制。而谣言的特点是：脆弱的事件结构和结构表示；同时 malicious users and naive users恶意用户和幼稚用户倾向于提升其传播范围，而造成了 the echo chamber effect 回声室效应在社交媒体中频繁发生**。

为此，提出事件增强策略：**它们是通过修改图结构和节点属性，虽然如此，但也保留了事件的关键信息**。

节点掩码

通过分析谣言传播的两个传播者 恶意传播者和幼稚传播者的传播目的的不同，**由此得出结论如果分析仅仅专注于谣言的参与者，可能会起到坏的结果**。

为了解决这个问题，**在每次epoch中随机mask图中除了root节点以外的其他节点的特征**。

子图

观察发现，当rumor的整个传播链条被考虑的时候，大多数的人会支持真实事件，同时否定错误的谣言。与此同时，如果仅仅观察rumor的早期response的情况，会发现用户对谣言的支持，却无关于其真实性的趋势这一现象。**因此如果对谣言的整个发展链条的事件都关注的话，训练的时候，可能会阻碍infer的早期阶段检测rumor的能力**。

为了解决这个问题，在整个事件图中使用随机游走，从root post开始。 The walk parallel and iteratively travels to its neighborhood with a probability 平行行走并以一定的概率迭代到其邻域。

边dropping

【14】EdgeDropping 对于GCN模型而言能够有效缓和过拟合和过平滑等问题。

具体的，它在每次epoch训练时，随机移除图中的边。这将有助于数据增强和减少信息传递。

另外，在该rumor场景中，EdgeDrop将同时减弱回声室效应（潜在影响用户的立场和观点，同时增加社会两极分化和极端主义）

Contrastive Pre-training

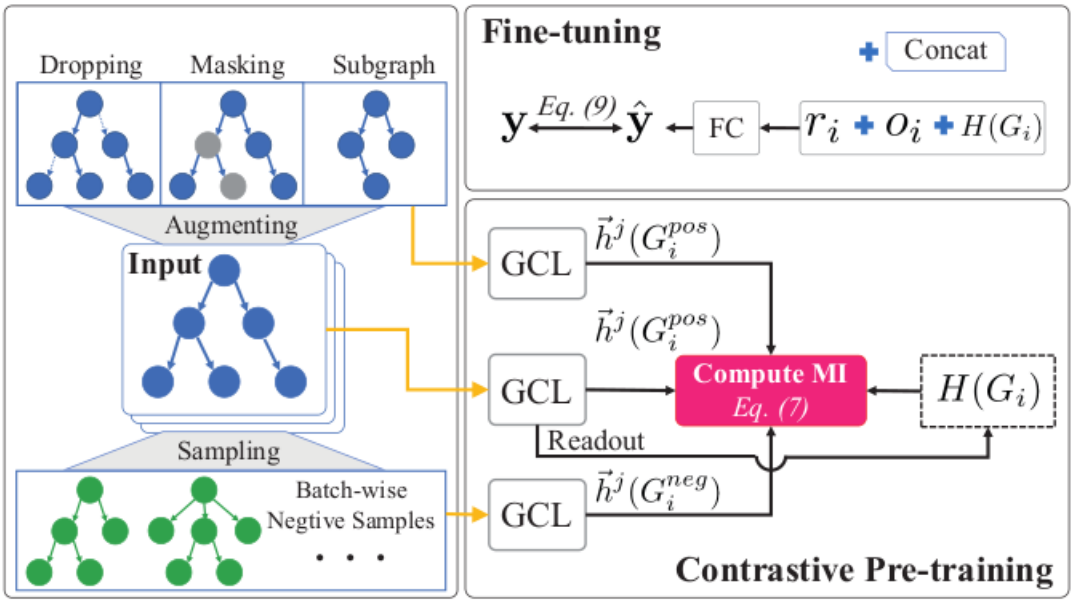


Figure 1: An overview of the proposed RDEA method. Contrastive vector of the input event graph is obtained via contrastive pre-training first. Then, we obtain the event representation by concatenating the pre-trained vector with textual graph vector and source post features. Last, we make predictions via fully connected layers and fine tune the model parameters.

（TODO：比较好奇 MI的计算公式）

GCL是图卷积编码层，一个节点的特征向量经过它将得到下一层该节点的特征向量。**可以是不同的卷积层**。

CONCAT行为是拼接加和行为，将每个节点在各层中的特征向量加和。

READOUT行为是将所有节点在各层中特征的融合向量经过某种方式得到融合，获取到事件图的全局表示。**可以是不同的池化方法**

本文中使用的卷积层为GIN，而readout池化层选择均值。

对比预训练的目标是 **最大化谣言传播图数据集的互信息值**。

计算方式是：

$$I_\psi(h^j(G);H(G)) := \mathbb{E}[-sp(-T_\psi(\vec{h}^j(G_i^{pos}),H(G_i)))] - \mathbb{E}[sp(T_\psi(\vec{h}^j(G_i^{neg}),H(G_i)))], \quad (6)$$

其中：

- ψ 表示 神经网络的参数集合；
- I表示互信息评估器，T表示判别器；G表示一个输入的事件图样本，Gpos表示G的正例，Gneg表示G的负例；
- sp 表示一个softplus函数。**它是relu函数的平滑版本**。（TODO：这里之所以使用它，是有什么考虑吗？）
- 正例事件：使用输入事件图和生成图的样本的local patch representations；负例事件：使用一个batch中其他事件图的local patch representation。

更好地预测谣言的真实性

融合了对比预训练得到的互信息，想要强调的source post和textual features文字特征，以得到事件图的representations。

其中textual features 是本事件的所有post的features的均值得到的一个特征向量。

$$o_i = \frac{1}{n_i}(\sum_{j=1}^{|\mathcal{V}_i|-1} x_j^i + r_i).$$

$$\mathbf{S}_i = \mathbf{CONCAT}(H(G_i), o_i, r_i). \quad (7)$$

Fine tuning

使用预训练的参数作为该阶段模型的初始化参数，接着用labeled data训练模型。

预测的输出通过多个全连接层和一个softmax层；通过交叉熵损失函数和L2正则项得到损失值；

实验细节

数据集：Twitter15/16两个数据集中，节点表示用户，边表示响应关系。其中特征根据TF-IDF值排序选择前5000个word。其中每个source post 被标注为四个类别：Nonrumor (N), False rumor (F), True rumor (T), and Unverified rumor(U),（**谣言不一定是假的，是中性词**。）

Baseline：

- DTC：使用了决策树；

- SVM-TS：基于SVM的线性时间序列模型，使用手工特征做预测；
- RvNN：结合了GRU单元的递归树结构模型，通过树结构学习谣言表示；
- PPC_RNN+CNN：结合了RNN和CNN的谣言检测模型，特别面向早期的谣言检测；
- BI-GCN：直接使用GCN，通过双向传播结构学习谣言特征表示；

Metrics：acc和F1；

parameters settings：

- 数据集分成五份，进行五折交叉验证；
- SGD+Adam
- 隐藏层特征向量的维度是64；
- 掩码率0.2，子图率0.4，drop率0.4.
- 自监督预训练epoch25；监督fine-tuning100epcoh；同时使用早停法，如果验证集的acc**停止上升连续10次**，将停止训练。

实验结果与分析

两个数据集上的acc和F1结果：

Table 1: Rumor detection results on *Twitter15* and *Twitter16* datasets with 100% label fraction data. Abbrev.: Non-Rumor (N), False Rumor (F), True Rumor (T), Unverified Rumor (U).

Method	Acc.	F_1			
		N	F	T	U
Dataset: *Twitter15*					
SVM-TS	0.544	0.796	0.472	0.404	0.483
DTC	0.454	0.733	0.355	0.317	0.415
RvNN	0.723	0.682	0.758	0.821	0.654
PPRC_RNN+CNN	0.697	0.689	0.760	0.696	0.645
Bi-GCN	0.836	0.791	0.842	0.887	0.801
RDEA	**0.855**	**0.831**	**0.857**	**0.903**	**0.816**
Dataset: *Twitter16*					
SVM-TS	0.574	0.755	0.420	0.571	0.526
DTC	0.465	0.643	0.393	0.419	0.403
RvNN	0.737	0.662	0.743	0.835	0.708
PPRC_RNN + CNN	0.702	0.608	0.711	0.816	0.664
Bi-GCN	0.864	0.788	0.859	0.932	0.864
RDEA	**0.880**	**0.823**	**0.878**	**0.937**	**0.875**

- 深度学习方法效果都好于手工fueatures；
- proposed method 优于其他的深度学习方法；

其他模型的不足的分析：

- RvNN：仅仅使用了所有叶子节点的特征向量，受传播链越往后的post的影响越大，丢失了对former post的信息。
- PPC_RNN+CNN：将传播结构视为平坦的时间序列，而丢失了较多的结构性信息；

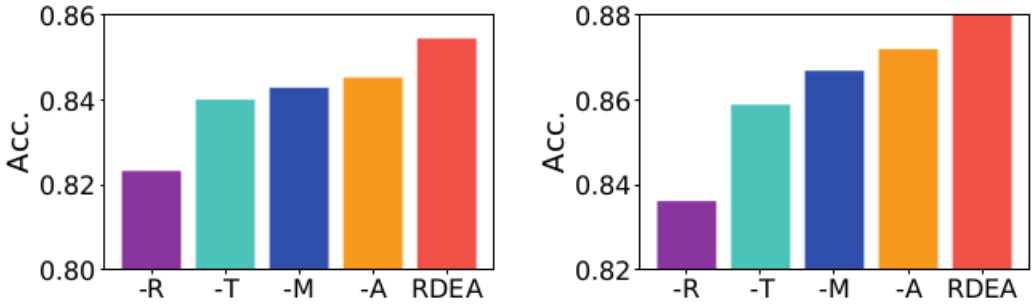
- Bi-GCN：特征表示比较容易受到噪声的干扰，同时需要大量的标注数据用于训练。

RDEA模型：

- 通过对比预训练得到最大化的互信息值，使得模型能够捕捉到谣言传播过程中的内在联系；
- 通过强调root post，模型能够更重视root post中的信息；

消融实验

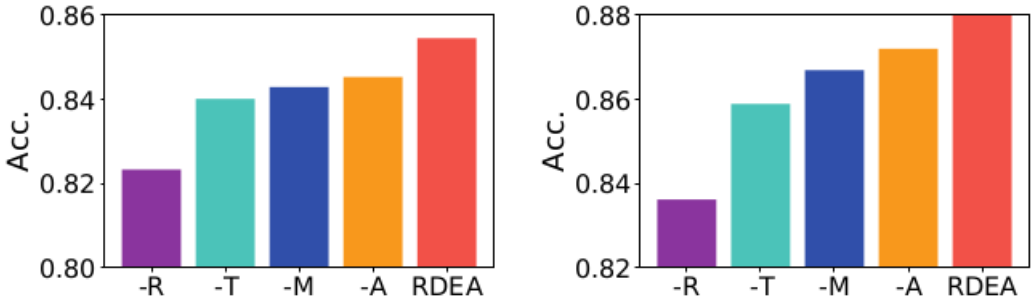
- root feature enhancement：indispensible 必不可少。
- textual graph：
- event augmentation：
- mutual information：



(a) *Twitter15* (b) *Twitter16*

有限标注数据

标注越少，提升提升越大，由此呈现出模型的鲁棒性；



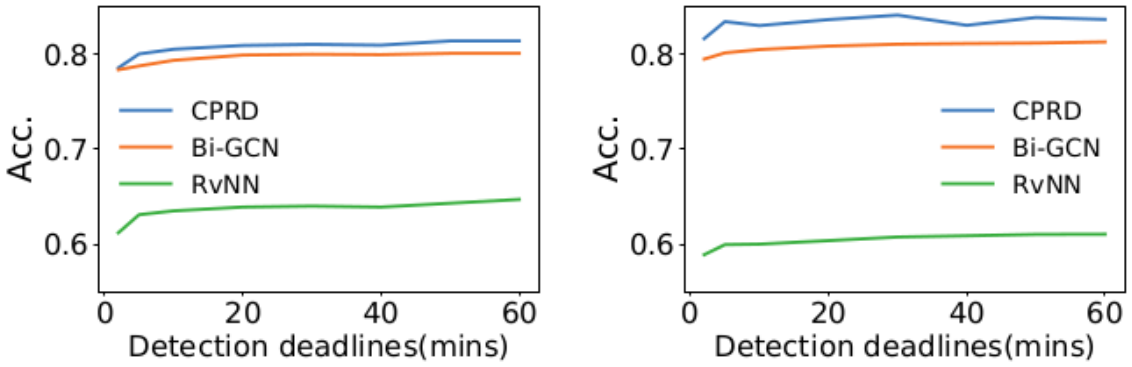
(a) *Twitter15* (b) *Twitter16*

（TODO：这一布是怎么做的？）

提早进行谣言检测的价值

在谣言出现的早期进行谣言检测，能有效地阻止谣言的传播和影响。

实验：选择一系列的检测deadlines，只使用deadlines之前的post用于测试实验的acc；



(a) *Twitter15* (b) *Twitter16*

Figure 4: Result of rumor early detection on two datasets

结果显示：We can also observe that the performance of all three methods is almost fixed at an early time (slightly improves over time).

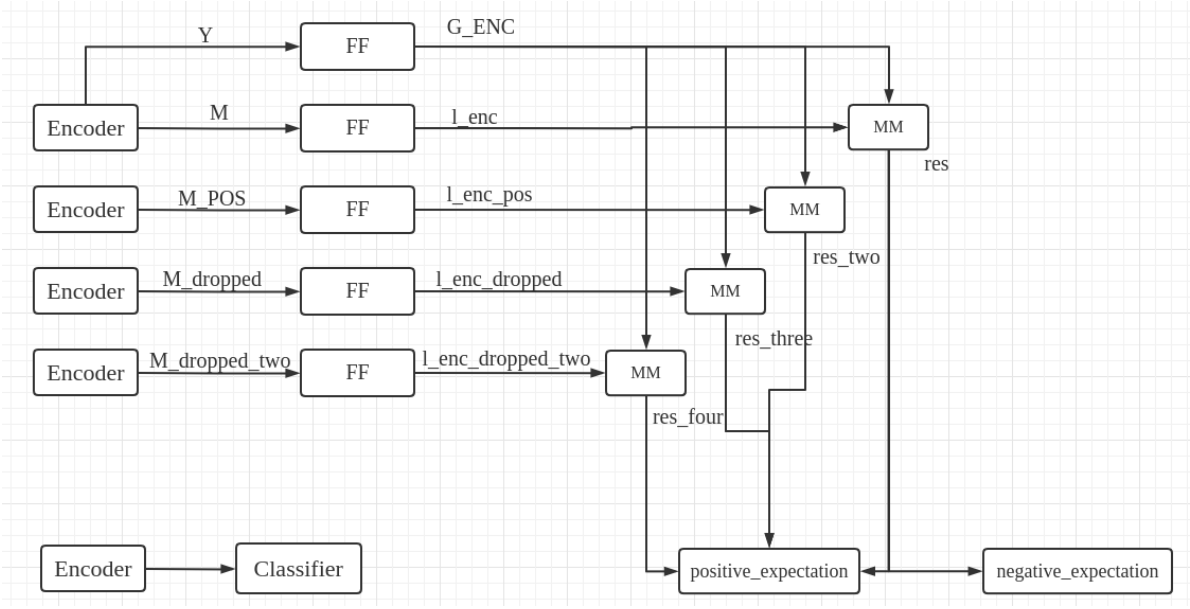
在早期的时候，acc就已经基本上不变化了，这种随着时间伴随的轻微变动，**验证了早期检测的有效性和意义**。

模型结构

- Encoder：len(num_gc_layers)个子模块：linear > ReLU > Linear > GINConv；

  计算过程中采集每一个GCL层的输出。
- FF：feed forward Layer： **block**(Linear>ReLU>Linear>ReLU>Linear>ReLU) + **linear_shortcut**(Linear)
- Classifier：**1**(Linear>dropout>prelu) > **2**(Linear>dropout>prelu) > **3** (Linear>dropout>prelu) > Linear > softmax

整体流程图：



《Rumor Detection on Social Media with Bi-Directional Graph Convolutional Networks》

本文介绍了propagation 和 dispersion 是 谣言的两个重要特征，本文引入了一种新的双向图模型，Bi-GCN。它通过上下传播和下上传播路经以捕捉这两个方面的特征。同时该方法注重source post 的重要性。（这里对于propagation和dispersion的区别是，前者的描述是深浅，后者的描述是宽窄）

引言

以往的深度学习方法以及传统方法对于谣言的检测，大都只停留在对propagation特征的学习上，而忽视谣言的扩散特征，即dispersion。而dispersion特征并非是CNN可以学习到的，**适合的方法应该是GCN，至少它面向的是这种非欧几里德空结构的数据**。

但是如果简单将数据套用到GCN中，虽然会获取到相关节点之间的关系特征，但是却无法获取到节点之间的顺序特征。

TODO：**地铁站数据是否可以使用这种GCN网络，同时考虑到节点之间的联系以及顺序特征；其次使用掩码行为，以克服过分专注于某个节点的信息。获取到比较平稳的正常的非异常的信息特征**。

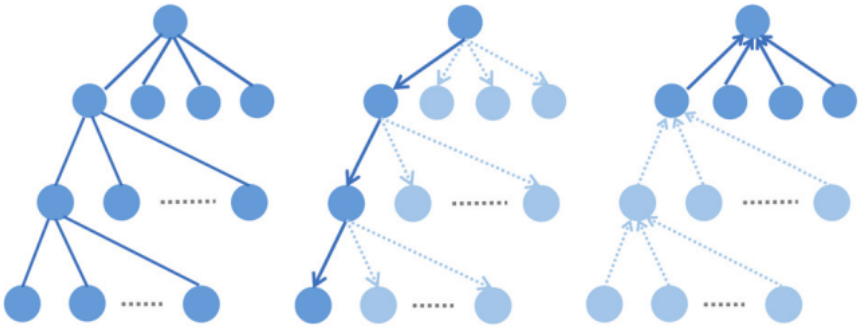


Figure 1: (a) the undirected graph with only node relationships; (b) the deep propagation along a relationship chain from top to down; (c) the aggregation of the wide dispersion within a community to an upper node.

由此，作者提出了双向GCN，TD-GCN用于构建propagation特征，后者用于表示dispersion特征。最后两个特征通过全连接层得到融合。**期间**，为了充分利用root post的信息，克服回声室效应，作者将root post的特征和每一个GCN层的隐藏层特征进行concatenate融合。训练过程中，为了避免模型过拟合，使用了DropEdge。**这就是作者所有的创新点。从这一篇文章看出，RDEA框架的一些想法在本文中已经有提及，而RDEA模型的最大优点在于其克服标注数据的对比自监督预训练行为**。

相关工作

作者简述了谣言检测经历了传统方法、RNN、CNN和GAN等方法，并且介绍了一些使用trick，比如SVM中的随机游走核、融合了注意力机制的RNN和增加了额外的特征（比如propagation 结构和文字内容的融合），最后引出了GCN，不过这里的GCN使用的是一阶切比雪夫网络。

Preliminaries 预知

这里提及事件event的代表是由三部分组成：text contents 文字内容、user information 和 propagation structure.

dropedge方法是一种GCN模型中缓解过拟合的方法2019，它可以增加数输入数据的随机性和丰富性；就像旋转、水平翻转。

Bi-GCN Rumor Detection Model

作者提及一个两层的一阶cheb网络是该模型的基础。

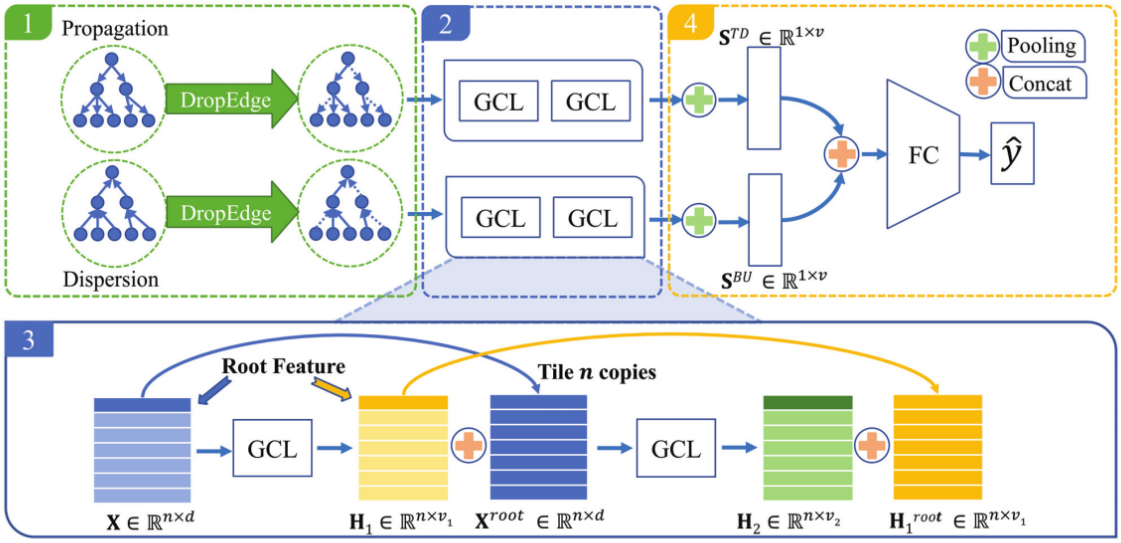


Figure 2: Our Bi-GCN rumor detection model. $\mathbf{X}$ denotes the original feature matrix input to the Bi-GCN model, and $\mathbf{H}_k$ is the hidden features matrix generated from the $k-th$ GCL. $\mathbf{X}^{root}$ and $\mathbf{H}_1^{root}$ represents the matrix extended by the features of source post.

1 Construct Propagation and Dispersion Graph

Bi-GCN中对于propagation和dispersion特征的获取是独立的，同时两个特征的获取至少从形式上看，差别在于post顺序：从上到下，从下到上。而这一点是通过单向链接的邻接矩阵实现的。**而邻接矩阵的性质是，矩阵的转置实现节点传播的方向**。因此两者代码组织的差别就在于邻接矩阵。而X是共用的。

todo：但是有个巨大的问题是：为什么方向的转置，可以获取这两个特征，为什么从高到低可以获取propagation而不是dispersion特征。

2 Calculate the High-level Node Representations

一阶chebNet>ReLU激活函数>Dropout

3 Root Feature Enhancement

在谣言传播中，root post 正是因为具有丰富的信息才引起广泛的影响。**因此有必要通过某种策略来充分利用root post的信息**：当前层的输入是上一层输入的root node 的特征向量和上一层输出的隐藏层vectors的concate

4 Representations of Propagation and Dispersion for Rumor Classification

TD和BU-GCN获取到的特征首先各自做平均池化操作 > 池化操作后得到的特征进行concate操作 > FCs > Softmax；

训练的时候使用交叉熵损失函数，同时应用L2正则惩罚项；

Experiments

1 Settings and Datasets

使用了微博2016年、推特15和推特16三个数据集。三个数据集的节点指代user，边指代响应关系，特征features are the extracted top-5000 words in
terms of the TF-IDF values。

Table 1: Statistics of the datasets

Statistic	*Weibo*	*Twitter15*	*Twitter16*
# of posts	3,805,656	331,612	204,820
# of Users	2,746,818	276,663	173,487
# of events	4664	1490	818
# of True rumors	2351	374	205
# of False rumors	2313	370	205
# of Unverified rumors	0	374	203
# of Non-rumors	0	372	205
Avg. time length / event	2,460.7 Hours	1,337 Hours	848 Hours
Avg. # of posts / event	816	223	251
Max # of posts / event	59,318	1,768	2,765
Min # of posts / event	10	55	81

2 Experimental Setup

- DTC 2011： 决策树+手工特征；
- SVM-RBF 2012：SVM+RBF核+手工特征；
- SVM-TS 2015：SVM+手工特征+时间序列；
- SVM-TK 2017：SVM+传播树kernel；
- RvNN 2018： 树结构+GRU单元；
- PPC RNN+CNN 2018：RNN+CNN+传播链中users的特征；

特别地：

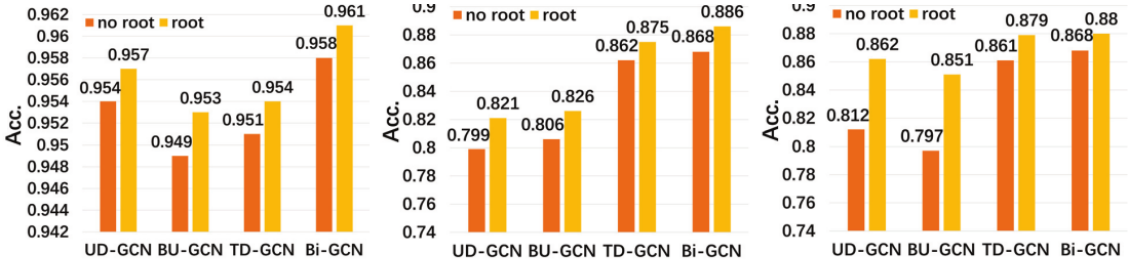
- **作者对于以上几种方法的执行 是在不同的框架中执行的**。
- 为了公平比较，数据集分成五部分，进行五折交叉验证；
- BI-GCN的模型参数的update使用SGD，optimize使用Adam；
- 特征维度是64，egdedrop 0.2，dropout 0.5，训练epoch 200；early stopping 10；
- Note that we do not employ SVM-TK on the Weibo dataset due to its exponential complexity on large datasets.（**头一次见，大家对这个进行解释**。）

3 Overall Performance

- 首先，深度学习方法要好于传统方法；
- 其次，Bi-GCN优于PPC RNN+CNN，主要是后者的RNN和CNN不擅长处理图结构，因此无法获取重要的传播结构特征表示；
- 最后，# 由于RvNN，因为后者仅仅使用叶子节点，以至于被后面的post所严重影响。与RvNN的对比，进一步强调了本方法中的root enhancement的价值。

4 Ablation Study

- 首先，root增强好；



(a) *Weibo* dataset (b) *Twitter15* dataset (c) *Twitter16* dataset

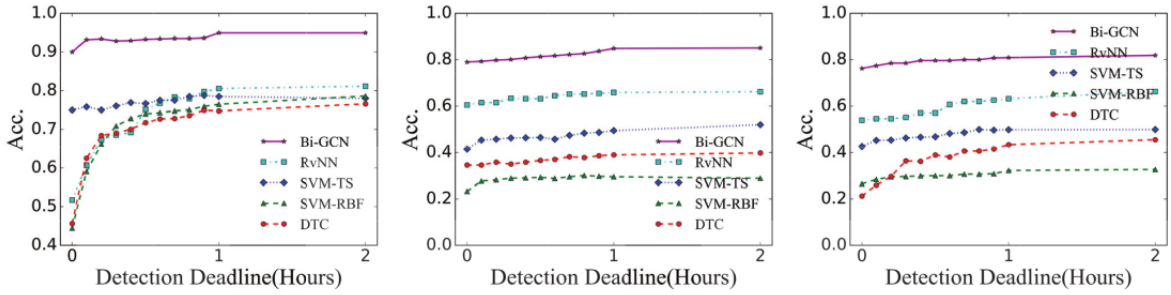
- 其次，验证了双向考虑模型会比单独考虑某个方向或者不考虑方向的效果要好；
- **最后，无论是考虑不同的数据集还是是否考虑root enhancement，基于GCN的无向、单向和双向的模型的效果都比其他的baseline效果好，因此表明GCN的优越性**。（这一点要结合上面的图表和下面的图表的数据，才能得到这样的结论。）

Twitter15					
Method	Acc.	N	F	T	U
		F_1	F_1	F_1	F_1
DTC	0.454	0.415	0.355	0.733	0.317
SVM-RBF	0.318	0.225	0.082	0.455	0.218
SVM-TS	0.544	0.796	0.472	0.404	0.483
SVM-TK	0.750	0.804	0.698	0.765	0.733
RvNN	0.723	0.682	0.758	0.821	0.654
PPC_RNN+CNN	0.477	0.359	0.507	0.300	0.640
Bi-GCN	**0.886**	**0.891**	**0.860**	**0.930**	**0.864**

Twitter16					
Method	Acc.	N	F	T	U
		F_1	F_1	F_1	F_1
DTC	0.473	0.254	0.080	0.190	0.482
SVM-RBF	0.553	0.670	0.085	0.117	0.361
SVM-TS	0.574	0.755	0.420	0.571	0.526
SVM-TK	0.732	0.740	0.709	0.836	0.686
RvNN	0.737	0.662	0.743	0.835	0.708
PPC_RNN+CNN	0.564	0.591	0.543	0.394	0.674
Bi-GCN	**0.880**	**0.847**	**0.869**	**0.937**	**0.865**

5 Early Rumor Detection

在谣言传播的早期如果可以达到跟传播很久时检测得到的效果基本一致，就表明该模型的检测性能比较卓越。



(a) *Weibo* dataset (b) *Twitter15* dataset (c) *Twitter16* dataset

Figure 4: Result of rumor early detection on three datasets

从图4看出，首先主推模型 能够在早期达到相对较高的准确度；其次，早期的表现效果优于其他模型；

因此，**Bi-GCN，不仅仅在长期检测中有价值，在早期检测中也有价值**。

《A Survey of Information Cascade Analysis: Models, Predictions, and Recent Advances》

摘要

社会的发展，使得人们对信息级联演化的轨迹和结构的探索多了些兴趣；本文对信息流行度预测方法做了综述，从feature engineering and stochastic processes特征工程和随机过程，经过图表示，到深度学习方法。

引言

理解信息如何传播，什么因素驱动着信息传播的成功，同时信息对流行度的预测等任务是比较有挑战性的，但是又很实用：病毒式营销viral marketing，广告，scientific impact quantification科学影响量化、推荐、campaign strategy竞选策略和epidemic prevention流行病预防。

作者认为，是信息传播的轨迹和结构，以及信息传播中的参与者构成了所谓的信息级联。

当前的业界和学术界都对信息级联的预测很有兴趣。而“预测”在不同的应用领域有不同的意义。

- 预测某个博客的流行度；
- Facebook中某个视频或图片的likes number。
- Youtube中某个视频的浏览量；
- 影片排名；
- 学术论文的引用量；
- 一个新闻的评论内容；

这里的prediction，根据不同的标准有不同阐释。

- 根据不同的问题，指代的是多分类、二分类或者回归问题；
    - 预测一个集群在未来某个时刻的具体大小；
    - 评估一个集群是否会grow超过某个阈值；
    - **此外，从战略上讲，基于观察到的信息范围，可以在发表之前做出预测 [138] 或通过窥视早期级联进化**
- 根据不同的分析层面，制定出不同流行度预测；
    - 宏观层面，模型可以学习一个集群的集体性行为；
    - 微观层面，模型可以学习到一个单独个体 对特定信息项的行为/响应（individual user actions/responses to specific information items）。

在方法论的角度，有比较充足多的算法选择用以建模和预测信息级联（集群）。

- 传统方法，与特定信息项相关的不同的特征（时许特征和结构性特征）可以通过特征工程feature engineering 提取。
  - 典型的机器学习方法：线性/逻辑回归，朴素贝叶斯，SVM和决策树；
  - 随机过程模型：泊松和霍克斯点过程。
- 深度学习方法。
  - Graph representation，深度walk，node2vec和图卷积网络；
  - sequential data，RNN及其变体；
  - 其他的深度学习diffusion model。

在综述方面，有很多人做过了。

- 【72，187】主要集中于不同的特征工程方法和经典的机器学习方法，用于建模预测信息项的流行度；
- 【63，146】专注于网络内容或者microblog information diffusion微博信息传播，**同时强调信息级联建模的不同方面**。

比如，【146】面向网络内容的流行度预测，综述了先验和后验的预测方法，同时综述了验证协议evaluation protocols和分类回归方法

本文综述的特点是：

- 更加广阔的视角：从对线上网络内容的建模到对信息更加泛化的定义（可以在任何网络中propagate的任何可以衡量的实体）
- 更大范围的网络和集群：涉及更多类型的网络，而非仅仅是单一的社交网络；
- 对建模方法更加广泛的更加平衡的深度调查：
  - 以不同的分析视角和层次，补充先前综述中的方法；
  - 详细分析现存方法的trade-off，优势和限制；
  - 更广范围内的特征、方法和可解释；
- 对最近综述更全面的介绍：

作者构建了一个建模方法的框架：

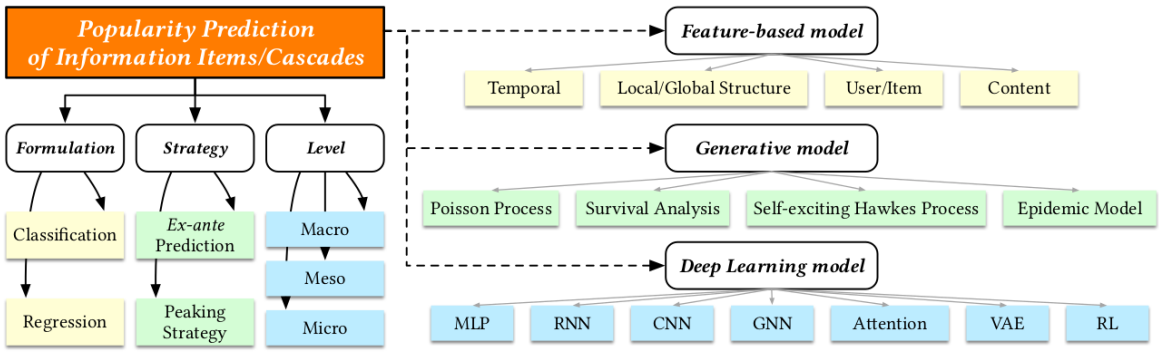


Fig. 2. Taxonomy of the information diffusion models.

预测任务

- 首先，**预测**可以根据问题设想和应用任务，被认为是分类问题或者是预测问题。
  - 用户的节点行为预测是分类任务；
  - 对未来特定时间某个item或者cascade集群的具体valume的预测是回归任务；
- 其次，**可以在信息发布之前或之后进行预测。根据预测时间的不同，现有的方法可以分为事前预测和事后预测**。
- 最后，信息预测可以根据任务的粒度the granularity of the tasks分为宏观、微观和中观层面。

建模方法

三类：基于特征的方法、generative models以及深度学习方法；

2 PROBLEM DEFINITION, EVALUATION, DATASETS, AND TAXONOMY

2.1 Types of Information Items/Cascades

这篇文章中，作者将information items 根据其popularity视为可衡量的实体，而信息级联由 **信息项的传播序列**构成。

信息项可以是UGC，其具体内容可以是post、threads，photos和videos。**这些内容彻底改变了用户与信息和其他用户的交互方式，以及信息的创建、呈现、传播dissminated和消亡方式。了解决定信息项传播的内部驱动对于许多实际应用程序（如广告、决策和缓存策略）来说是非常重要的**。

信息项可以被分类为endogenous or exogenous 内源性和外源性两种。前者的信息项来源于社交影响力，而后者来源于突发事件。

2.2 Problem Definitions

信息项或者信息级联 本身是transient, sparse, and biased data.瞬态的，稀疏的和有偏差的数据，也就不能预测，但是为了刻画信息项和信息级联同时探索在何种条件下，预测问题可以被解决。**作者提出了三个类别：分类和回归、事前和事后预测以及预测粒度**。

2.2.1 Classification versus Regression.

预测问题是鉴于其初始状态，预测信息的最终受众，注意力和影响。

- 二分类问题，相关阈值；
- 多分类问题，需要预定义几个popularity interval间隔，预测信息项将落到哪个区间；
- 预测问题，预测具体的值；

一般而言，分类问题相对比预测问题简单。**尽管公式化回归背后的直觉更加自然**，并提供了一个细粒度的范围来分析哪些因素会影响未来的流行度并导致信息项成功，但**精确的回归预测通常需要更多关于项目和用户的信息**，提取出更加复杂的关系，因此**意味着更高的复杂性**。同时，它经常会遇到overfitting, inductive bias, and prediction error accumulations。过拟合、感应偏差和预测误差累积；**同时，作者也发现许多模型可以在回归模型和分类模型之间轻松的转换**。

2.2.2 Ex-ante Prediction versus Peeking Strategy

一般而言，事前预测更加具有挑战性，因为我们**获得的信息是有限的、数据量巨大的，又是敏感的资源，很难获得**；而其价值是吸引人的：事前预测常常利于下游任务，比如广告和市场。

因为事前预测的数据不易获取，因而通常求其次，即peek into the early stage of a cascade’s evolving process窥探演化过程的早期阶段。事后分析的价值虽然没有事前价值大，但是其价值存在且不小。因为通过分析，得到三条结论：

- 对于微博推文来说，它们获得用户关注的速度越快，由于新出现的竞争对手，它们消失的速度就越快；
- 对于推特标签，其在用户中的流行持续时间要比微博上的持续时间长；
- 至于APS论文，项目的发展速度是在中间。

事后分析，表明一个项目成功传播的可能性。其基本原理是，在早期阶段成功传播的信息项往往会变得流行，即早期模式表明长期流行[183]。但是，**Previous works found that the similar, or even the same, content information may vary significantly in terms of popularity [39].以前的研究发现，相似或相同内容的信息其受欢迎程度差异也会很大**，由此人们产生了疑问：

- 是数据不足还是本来就不可预测；

- 是否是现有数据的内在影响力比较小，或是外部的一些数据因素决定了事情的最终流行程度；

这个问题没有答案，但也促使研究人员研究各种peek 策略。

2.2.3 Macro-, Micro-, and Meso-level.

简单来说，分别对应三个程度：全局整体、group community 和个体。

2.3 Evaluation Metrics and Datasets

分类指标：Accuracy, Precision, Recall, and F -measure.且常伴随一个阈值；

- 准确率：类别不均衡问题，由于受欢迎程度分布的高度倾斜。

  缓解途径：

  - 类别均衡二分类；
  - 从大量数据集中过滤出（下采样）得到一个均衡的数据集；
  - 早期的一些peek 策略的研究员，会选择忽视小样本，以产生均衡数据集。

回归指标：mean square error (MSE) and ts variants.

- Popularity也常使用logarithmic scale对数形式，MSLE or RMSLE，以防止损失函数和指标被极端值影响，同时确保数值稳定性。
- 确定/相关系数及其变体、排名[186]、k-top覆盖率[152，248]也是某些特定场景中常用的指标。

Previous works found that a model might perform well in one metric but significantly drop in another [74], making it difficult to do a fair comparison between various approaches. 先前的研究表明，一个模型可能在一个指标上表现好，但是可能在另一个上表现不好。因此很难作出一个公正的裁决在不同的模型方法之间。

数据集：如果希望一个模型能够从一个数据集泛化到另外一个有差异的数据集，是很难的，同时有时甚至是不可能的。

3 信息级联的特征和特征工程方法

信息级联的可以被分成五类：时序、级联结构、全局图、user用户或者item信息项的属性和内容特征content features。

3.1 时序特征

3.1.1 Observation Time.

3.1.2 Publication Time.

24个小时的时间点。晚间发布的内容相对比白天发的更不容易被人view。

3.1.3 First Participation Time.

首个参与者是首发者后第一个转发或者参与的人。

3.1.4 Evolving Trends.

这一点已经被证明，它具有informatively signals。

时间序列的时序模型被分成几个类别：平滑增长，突然的增长和下降等十种演化趋势。

十种演化趋势可以通过**分层聚类算法**进行聚类

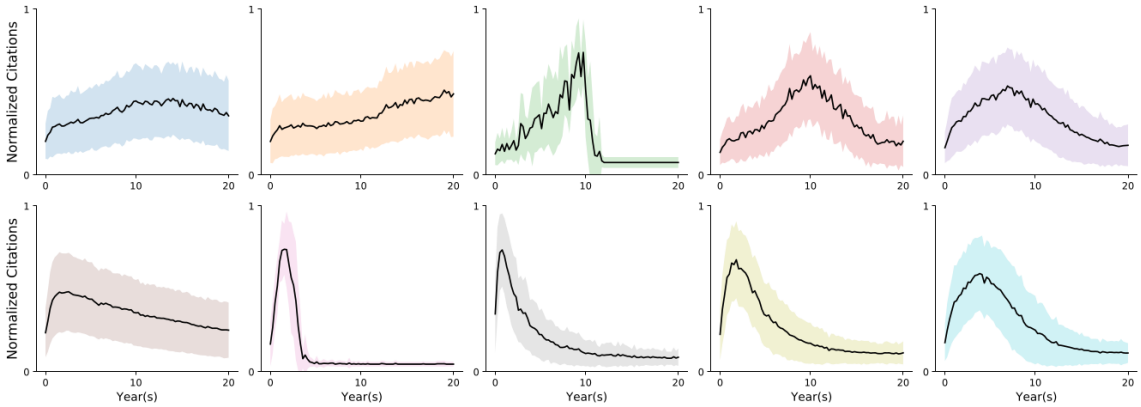


Fig. 6. Ten evolving patterns characterizing the citation cascades in the APS dataset. We use hierarchical clustering implemented by *sicpy* package in Python to cluster the APS papers. For each cluster, we show 20 years of evolving trends of citations, by use of mean values and ± standard deviations.

3.1.5 Discussion.

尽管时序特征很重要，最近的研究表明他们在一些场景下效果并不好；

时序特征的优势随着时间消失，它的影响没有其他的特征的效果好；

3.2 Structural Features

级联结构有时指代为信息传播。

对级联结构进行建模的研究被分成三个部分：

- 参与者，仅仅级联图；
- 全局图，参与者和非参与者；
- R阶可达图，一种折衷方案，将级联图扩展到全局图的范围内。

3.2.1 Cascade Graph.

3.2.2 Global Graph.

和我们常见的图很相似。

图的类别：

- 有向图，边带方向；
- 有权图，节点或者边有相应的权重；
- 超图，节点或者边具有超过一种类型的属性，一张图以作者，论文作为节点，发布时间、引用者等作为边；
- 属性图，节点或者边具有特征，推特的文本，论文的摘要；

3.2.3 r -reachable Graph.

是从Global Graph.中获取到的子图。

3.3 User/Item Features

时间和结构特征需要对早期观测进行探索，因此显得不切合实际，因而会选择从用户和信息项入手。它们具有独特的属性和innate attractiveness天生的吸引力，因此，在事前预测中比较实用。

3.3.1 User Features.

用户行为在信息传播和消费中起着至关重要的作用选择（例如，查看、评论、共享和首选）。最常见的用户功能之一是作为用户影响力代理的追随者数量，这意味着未来的受欢迎程度[240]。**拥有大量追随者（观众）的人，如名人而新闻机构比普通用户更可能产生大量的级联，因为它们的信息在网络中更加可见**。然而，大型级联不仅仅是由有影响力的用户产生，**研究由普通用户产生的大型级联很有意义而不是名人**[51]。许多其他用户特性已经被广泛研究和探索了多年分析和预测信息项/级联的流行程度，例**如简介（姓名、年龄/地区、教育、就业、账户创建日期等）[238]，历史行为（频率发布项目和与其他用户交互的数量、活动时间等[11,83,182]，用户兴趣[229]，集体[132]，相似[177]，过去的成功[11]，活动/被动[227234]，发现[142]，亲和力和响应能力[238]**，等等。

3.3.2 Item Features.

- users' interfaces affect the visibility of items 互动会影响item的可见性；
- 文章的metadata影响评论的数量等指标。

3.3.3 Discussion.

大多数的user/item特征是不言而喻的，并不难理解他们与流行度之间的关系。比如说一个拥有大量受众人群，说同一种语言，经常讨论当前流行话题的人将会有更多的机会使的信息流行。

但是一些特征需要更加复杂的算法和计算才能获得，比如用户影响力，偏好，相似度等。

先前有人研究了个人是如何影响信息的传播的。

【109】，被关注人数，Pageank，转发数目用来对用户影响力进行排名；

【19】，兴趣导向影响，社会导向影响，流行导向影响；

但是，鲜有人通过 **用户拓扑图结构**来研究社交影响力；

3.4 Content Features

内容特征被认为是内在驱动和关键因素之一。比如，突发事件、谣言、假消息、热点话题、controversial/peculiar topics争议特殊话题、disinformation and misinformation 虚假信息和误传信息。

3.4.1 Text Content.

常见的研究方法：TF-IDF，LDA，朴素贝叶斯，SVM和线性回归等；

3.4.2 Content Features of Image.

常借助计算机视觉的技术。具体的特征有：

- 基础属性：事件，季节，设备，主题色，分辨率，位置和标签。
- simple human-interpretable features, low- and high-level image features 简单的人类可解释特征、低级和高级图像特征；
  - 醒目色的图片比较受欢迎；
  - 低层次：利用局部二值模式和CNN提取图像的**基调、纹理、色块、梯度和代表性特征**；
  - 高层次：ImagNet中某个图像类别是否存在。
- 视觉特征被分为aesthetic and semantic features审美与语义特征。
  - 前者解释了图片的美丽性；dominant colors, saturation, brightness, contrast, texture, background area, region focus and focus centrality/density,**主色调、饱和度、亮度、对比度、纹理、背景区域、区域焦点和焦点中心/密度**。
  - 后者，通过不同的视觉技术进行提取：likelihood scores of categories 类别似然分数，extracting features like texture and color 提取纹理和颜色，将图片转化为一系列的文字。

3.4.3 Other Content Features.

对于视频，基础特征包含：视频长度，帧数，分辨率等；

新增加的特征：color histogram and aesthetic features, quality features, object features 色彩直方图、美学特征、质量特征和对象特征。

3.4.4 Discussion.

项目内容已被证明是合格的预测指标；

但同时研究者对内容特征的有效性的观点有不同。

- 内容特征相较于时序、结构化和个体特征而言，其有效性相对较弱；
- 当考虑到的参与者越来越多时，内容特征就越来越不重要了；
- 内容特征不能很好地解释受欢迎程度的差异。
- 基于内容特征的研究方法可能会受到性能的阻碍：尽管NLP和CV发展很好，但是效果差强人意。**进而得到结论，可能仅仅依靠单一的内容特征，本质上是无法预测的，或者无法事先预测。inherently unpredictable or cannot be predicted a priori.**

3.5 Prediction Methods

由于基于特征的模型的主要挑战在于特征工程，因此提高预测模型的性能并不是相关文献的重点。

- 参考文献[39]表明，**尽管存在时间/空间复杂性，但大多数机器学习方法具有相似的性能**。为了完整起见，我们总结了常见的机器学习方法，或作为其主要构建块采用的方法，如表6中的预测方法。除了表6中列出的方法外，还研究了一些学习范式，如**归纳/归纳学习、早期特征融合和多视角学习方法**，以预测信息项/级联的流行程度[36、80、134、212]。
- 我们建议研究人员在其特定的数据集上试验不同的预测方法，自动选择机器学习模型和超参数等技术[104]将极大地促进训练和优化过程。

Table 6. Machine Learning Methods

Method	Abbr.	Reference
Autoregressive (–moving-average)	AR(MA)	[48, 75, 76, 125, 139, 225, 239]
Decision Tree	DT	[12, 39, 49, 50, 60, 61, 63, 75, 86, 99, 102, 103, 105, 107, 108, 136–138, 140, 196, 202, 220, 232]
k-nearest Neighbors Algorithm	k-NN	[12, 48, 61, 63, 75, 86, 99, 100, 125, 136, 137, 220]
Linear Regression	LR	[1, 3, 12, 14, 24, 31, 39, 48, 50, 63, 74, 75, 79, 92, 100, 102, 103, 125, 134, 138, 158, 162, 170, 183, 186, 188, 194, 210, 214, 220, 231, 235, 236, 240, 248]
Logistic Regression Classifier	LRC	[39, 46, 49, 50, 61, 83, 87, 88, 99, 105, 137, 140, 151, 153, 160, 167, 177, 202, 217, 229, 243]
Multilayer Perceptron	MLP	[49, 63, 102, 103, 202]
naïve Bayes classifier	Bayes	[39, 49, 50, 61, 63, 75, 87, 136, 137, 174, 175, 231, 232]
Random Forests	RF	[4, 23, 39, 49, 50, 63, 73, 74, 88, 99, 105, 138, 144, 163, 177, 189, 191, 193, 202, 209, 210, 217, 232, 243]
Support Vector Machine	SVM	[12, 36, 39, 49, 50, 61, 65, 74, 75, 86, 88, 93, 98, 99, 102, 103, 105, 110, 130, 133, 134, 136, 137, 140, 160, 168, 177, 190, 194, 202, 220, 227, 229, 234, 243]

3.6 Global Overview of Pros and Cons

与其他模型相比，基于特征的模型通常被认为是具有竞争力且可解释的[74]。然而，**基于特征的模型阻碍其在实际应用中实现的主要瓶颈是手工制作的特征工程。**

由于隐私问题，有些功能很难获得，例如偏好和查看历史记录，有些功能（例如，用户分类和聚类）需要大量计算，这限制了模型的可伸缩性。

- 大多数时态特征——以及用户/项目特征——**都很容易提取和计算**。
- **结构特征**，特别是对于那些大规模图，例如全局图和r-可达图通常包含数千甚至数百万个节点和边，**需要大量的计算资源**。
- 文本、图像、音频和视频的内容特征，取决于特定的问题公式、数据规模和建模算法，**具有不同的时间/空间复杂性**。

本着这种启发，**考虑到一组详尽的特征，如何选择代表性特征中相对较小的一部分，以在有效性和效率之间最大化预测的边际效益，成为设计实用预测模型的关键考虑因素**[60、232、248]。**也就说说在分析的这几种特征中，选择效果不错的几个，做特征融合，会得到不错的效果。**

此外，现有模型需要更多的功能，如历史视图计数和扩散路径，这在大多数情况下是不可用的，最重要的是，不能概括到不同的场景。
我们回顾了四类特征和相应的预测模型。但是，不可能提及所有特性和模型，也不可能评估它们在所有特性组合中的性能。**对不同条件下的不同特征进行综合评价，将有助于特征工程和特征选择/组合的标准化**。

4 GENERATIVE MODELS

信息项的传播广泛采用概率统计生成方法；**主要是因为一些问题与时间相关，因此建模时候使用概率统计比较合适**。

4.1 Poisson Processes

基于**点过程的模型**因**其统计、概率和生成形式而区别于基于特征的模型**。

在排队论和运筹学中，在对时间序列（如客户到达率、电话和机械故障）建模时，通常使用点过程。

Pros and Cons

- **生成模型通常不需要大量的特征工程，并且具有内在的可解释性。它们主要依赖于时间序列数据，一旦准备好模型并估计出参数，就可以实时进行预测**[248]。
- 然而，它们的性能受到了质疑[29,74]，
    - **它们通常很容易受到异常值的影响**[144]。
    - 此外，生成模型通常对固定参数进行强有力的假设，这**限制了其通用性和模型表达能力**[52]。
    - 此外，**为了模拟/再现级联扩散过程，低估和简化了控制级联成功的复杂潜在机制**。
    - 最后，大多数生成模型都是**网络不可知**的——也就是说，它们**无法对有助于理解信息传播过程/路径的重要结构信息进行建模**。
- 因此，尽管生成模型具有效率和可解释性，但它们在做出**精确预测方面**的能力较弱[29]。

5 DEEP LEARNING MODELS

深度学习往往会比线性模型模型。

- 基于RNN的模型并不需要级联模型准确的假设条件，同时能够比较灵活的捕获时序依赖特征。
- 基于图表示学习的模型不需要从级联的基础图中获取到的laborious hand-crafted features费力的手工特征。

现有模型可以分成三个类别：

- 基于上述手工特征的模型，通过NLP和CV技术来学习 **用户或者items的表达性表征**；
- 基于时间序列的模型，比如社会网络和学术网络中的级联行为，他们都依赖于RNN和池化结构；
- 基于图的模型，级联图和全局图，它们处理图神经网络和图表示学习，旨在学习图结构数据中节点/边/图的embedding特征。
- 其他的技术，注意力机制、variational inference, reinforcement learning变分推理和强化学习等。同时，多模态、多尺度和多任务学习也用来提升预测的能力；

以往的技术：

- 93，预测图片的浏览量，手工特征+低级高级社会/视觉特征（过去的成功，联系人数，色彩和纹理）；CNN模型；
- DeepCas，第一个学习图表示的模型，借用了DeeWalk思想，采样得到的节点序列进入Bi-GRU，注意力机制，以获取node embedding。端到端学习。
- DCGT，在DeepCas的基础上，融合了节点的内容信息。
- DeepHawkes，整合生成式模型的优势和深度学习技术。user influence, self-exciting mechanism, and time decay effect 用户影响力，自激励即指和时间衰减影响 三个概念被转化到该模型中。**它没有直接对级联图的结构性模型进行建模，而是在信息及其传播轨迹上，使用GRU，加和池化及非参数时间核以融合早期参与者的贡献**。
- ANPP，使用GloVe以对word embedding，使用node2Vec对用户图编码，带有注意力机制的GRU用来融合embedding和时间序列特征项量。
- DTCN temporal context ，用户/图片的embedding，传播序列的时序context，多时间尺度注意力机制；其中，Reset和LSTM分别对视觉依赖和时序依赖进行建模。
- UHAN，同时考虑了图片数据的视觉和文本模态，通过对三种不同的表示进行刻画：
  - VGG与训练得到的视觉特征表示；
  - LSTM编码得到的文本表示；
  - 监督学习下的用户表示

  最后，在结合内部注意力机制，对两种模态进行联合学习。
- 诸如此类。

表7总结了基于深度学习的级联模型及其主要构建模块。通常，这些作品通过各种深度学习技术学习信息项/级联的不同方面，例如，利用RNN或其变体捕捉时间序列的长期依赖性和参与/引用的时间特征；通过深层语言和视觉模型学习文本和图像的表示；通过无监督网络嵌入[27]或（半监督）图神经网络[252]处理图结构数据。
与基于特征的模型相比，基于特征的模型依赖于手工制作的特征（特定于平台并依赖于先验知识），或者生成模型采用硬编码的扩散协议（缺乏灵活性，也依赖于人工设计），深度学习模型不需要大量的功能工程，可以捕获用户/项目内容和流行度累积趋势的非线性表示。

深度学习在其他领域的成功似乎在信息级联建模中继续，越来越多的方法采用了深度学习的技术并取得了最新的成果。

深度学习模型的主要优势在于其**相对简单的体系结构（具有少量人工设计的深层堆叠层）**和在监督下通过反向传播（具有大量数据和计算资源）实现的强大学习能力[113]。然而，尽管深度学习模型在预测性能上有所改进，但仍面临许多局限性。

深度学习模型的一个主要缺点是，由于神经网络的“黑箱”性质，**缺乏模型的可解释性**。深度学习模型的**计算成本明显高于基于特征和生成模型**。模型调整过程、超参数选择、过度拟合风险等，有时会导致工程师付出大量努力以获得令人满意的性能。