1. **介绍**

线上社会网络允许世界范围内的海量用户生产和消费内容。他们为海量的不同种类的信息资源提供了入口。线上社会网络在信息传播中也扮演了重要角色，例如：加快新奇的信息和各不相同的观点的扩散速度。他们被证明在很多领域都是有力武器，比如：Facebook在2010年阿拉伯之春和twitter在2008年美国总统大选中起到的作用。鉴于线上社会网络给这个世界带来了如此大的影响，现在的研究重点主要集中在如何从海量的信息中提取出真正有价值的信息。事件、社会问题、令人感兴趣的事物等，都在社会网络中和其捕获中迅速地发生、演化。对这些捕获信息的理解、可视化和预测，成为了终端用户和研究人员都期望的事情。了解这些网络的动态可以对处理事务有帮助（如：分析革命浪潮），解决问题（如预防恐怖袭击、预测自然灾害），优化业务绩效（如优化社会营销活动）等等。因此，研究人员们近年来提出了许多技术和模型，用以在线上社会网络中捕获信息传播、分析它、从它们中提取有用信息并试着去预测它。

信息传播是一个非常宽泛的研究领域，并且已经引起了很多其他领域的兴趣，如物理学、生物学等等。网络中的传播行为是研究网络的最初原因之一，正如疾病在人群中的传播现象已经被研究了数个世纪。作为计算机科学家，我们我们将注意力放在某几种信息在线上社会网络中的传播情况，这引起了下面几个问题：（1）哪些信息和话题是受欢迎且传播最快的 （2）信息具体是如何传播的？沿着哪些路径传播传播？为何沿着这些路径传播？将来又会沿着哪些路径传播？ （3）网络中的哪些成员在传播过程中起重要作用？

这篇文章的主要目的在于回顾与这些问题相关的研究进展，给读者展示该领域。我们在文章中指出现存的技术的优缺点，并将他们分门别类。····

在第二部分，我们我们主要讨论线上社会网络的基本特点和信息传播的性质。在第三部分，我们提出了一些方法去发现社会网络中引起人们兴趣的话题（通过使用信息传播属性）。接着我们讨论如何去对信息传播进行建模。我们将在第四部分介绍那些解释性、预测性的模型。接下来，在第五部分，我们提出方法去定位那些有很大影响力的信息传播者。在最后一个部分我们总结并将这些模型分类，展示他们的不足之处和显示出来的问题。

1. 线上社会网络和信息传播基本概念

线上社会网络（OSN）的形成是由专用网络服务直接引起的，这一概念经常和社会网络地点（SNS）相关联，SNS允许他的使用者去：（1）做一个简介页，并且可以发布信息 （2）直接去与其他用户取得联系，建立社会关系。 事实上，一个OSN可以被描述为一个用户自行生成的、允许它的用户之间进行沟通和信息共享的内容系统。

一个OSN形式上由图形表示，其中节点是用户、边表示关系（直接或间接，取决于SNS中如何管理关系）。更确切地说，这取决于边的类型是单向的（如twitter社会模型中的following关系）还是双向的（如facebook社会模型中的friendship关系）。在这种类型的服务中，Message是主要的信息传播媒介。用户们通过发送消息来向他人分享不同种类的消息，例如提出建议、发表政治见解、提出想法等等。一条Message经常被描述为：

（1）一种文本数据

（2）有一个作者

（3）有一个时间戳（视需要而定，可以没有）

（4）有一个信息目标受众的集合（在社会网络术语中称为“mentioned users”）。

图1中展示了一个用有向图表示的OSN，它的四个成员发出的消息扩充了这个图。一条弧e=（ux，uy）表示ux收到了uy发出的消息。

定义1：（Topic）一个由语义相关的元素（集合中的元素表达某种观点）构成的集合。在实践中，我们发现了三种对该定义的解释。（1）集合大小为1的集合S 如{“obama”} （2）集合大小大于1的集合S如{“obama”，“visit”，“china”} （3）一个集合S中的元素的概率分布

每种信息都可以（使用在定义1中定义的形式中的其中一种）被转化成topic。在全球范围内，OSN成员生产的内容是一种message流。图2表示在前一个例子中，网络中成员生产的“流”。这个流可以被视为一些连续的决策（是否去接受某个topic），后来的人可以看到先前的人的动作。因此，个体们被他人采取的行为所影响，这个影响也被称为社会影响（social influence），被如下定义：

定义2：（社会影响social influence）一种社会现象——个体们可以受到或施加影响（也被称为模仿）。该名词用以解释这种现象：用户们的行为可能会引起与他相连的其他用户也表现出相似的行为。例如当某个人转推另一个人的twitter时，这种现象表现得很清楚。

定义3：（集群行为herd behavior）一种社会行为——当一系列的个体采取完全相同的行动时发生，无视他们个人的信息信号不是必须的。

定义4：（信息级联information cascade）一种行为——社会网络上的人们决定采纳某种信息的原因是：他们无视了他们自己的信息信号，只基于先前的人们采取的行为来推断自己该如何作决定。

基于social influence的影响，信息可以在网络中通过集群行为和信息级联准则（已在之前定义）传播。在这些传播的信息中，一些topic可能变得非常受欢迎、传播得非常广并且形成新的趋势。最后，在OSN上发生的信息的传播过程可以被总结为如下几点：

1. message承载一部分信息

（2）沿着网络中的边、根据一些特定的机制进行传播

（3）依赖于节点和边的某些特定的属性。

在接下来的部分，我们讨论通过最近的一些与之相关的工作来讨论这些不同的方面。分析他们的优缺点，和对每个方面可能的提升。

1. **流行话题的发现**

研究信息传播的其中一个主要任务就是去寻找一种自动化的方法来提供一个关于某些topic（在很长一段时间内都很流行的topic，或是将会变得很流行、使网络变得活跃的topic）的全局视图。这个工作包含了从“内容表”中提取有价值的信息去总结讨论、向用户推荐热门话题、或是去预测将来的热门话题。

传统的topic发现技术主要被用于去分析静态的全局数据。因此他们不能被用来分析OSN产生的message流。为了使得从文本中发现topic更加高效，建议我们将目光放在“突发”上。Kleinberg提出了一种状态机模型用以对一个“流”中文件的到达次数进行建模，用以识别出“突发”（假设所有的文件都属于同一个topic）。Leskovec给我们展示了社交媒体上最受欢迎的topic的在一段时间内的变化是由流行度的升降交替模式形成的。

定义5（突发性话题bursty topic）一种行为——用以描述一个话题在一个时间段内突然受到广泛的关注，但是在该时间段之前和之后受关注度都很少。

接下来，我们介绍一些设计好的方法从一个描述不同topic的message流中发现那些突发性话题。

本文中涉及到的所有的方法依赖于计算一些频率（工作在离散数据上）。因此，将message流离散化（通过将横坐标（表示时间）切分成等距的时间片来实现）。这个准则在图4中描述。·······这个预处理步骤并非可有可无，因为它定义了话题发现的粒度。一个非常合适的离散化程度（例如很短的时间片）允许发现在很短一段时间的热门话题，划分较长的时间片则做不到这一点。

Shamma提出了一个简单的模型，PT（peaky topics）模型，和经典的tf-idf模型类似， 该模型也是基于一个标准化的词频度量。为了去量化全局范围内的词语使用次数，他们将每一个时间片视为一个由相关集合内所有的message组成的伪文件（pseudo-document）。标准化的词频按如下定义：

其中，tf（t，i）表示词t在第i个时间片内出现的频率，cf（t）表示词t在整个message流中的出现频率。使用这种度量标准，突发性话题可以被定义为只有某个单个时间片的频率为非零值的topic。然而单个的词可能有多种含义、再加上词本身意义可能是含糊不清的，因此单个词并不足以去唯一的辨识出一个topic。因此，需要更复杂的方法来达到这一目的。

AlSumait提出了一种更加精确的线上topic模型，该模型是一种非马尔科夫线上LDA吉布斯采集器模型，称为OLDA。LDA是一种统计学的生成模型（生成概率分布曲线），依赖于一个分层的贝叶斯网络，该网络可以将潜在话题中的词和message联系起来。该概率曲线的生成过程：将文件视为潜在的topic的混合物，而每一个topic可以用一些单词的分布来表征。OLDA的思路是去在每一个时间片中更新topic模型，更新方式是选用之前形成的模型作为先验知识，并且使用相应的message集合去引导新的模型的生成过程。这个方法基于为每个topic建立一个演化矩阵，用于去记录topic在该段时间内的变化情况，用以发现突发性话题。

Cataldi提出了TSET（Temporal and Social Term Evaluation）方法，该方法着眼于message流的瞬时性和社会属性。为了达到这个目的，他们提出了一个5步的过程：

1. 首先将各个message的内容形式化为由单词的相对频率构成的向量。
2. 接着，我们用他们之间的关系和PageRank算法来评估每个元素的重要程度
3. 它允许去对每个单词的生命周期进行建模（基于一个生物学隐喻），该模型主要用来计算营养和能量的值，会影响到用户的重要性
4. 使用监督或非监督的技术、基于能量来计算最高点（临界下降值），这个方法可以确定突发现象最明显的单词。
5. 最后，我们可以用共生网络图来给出突发性话题的定义

这些用于寻找特定种类的的话题的方法在过去引起了广泛的关注。Lu发现了一种方法可以预测哪一个话题将在未来引起人们的关注。作者建议我们采用类似股票交易中使用的技术分析指标去定位突发性话题，即所谓的MACD（平滑异同移动平均线），定义在单个项上。MACD的原则是。。。。。（介绍MACD可参照百度百科。） 作者给出了两种简单的描述关于一个词的出现趋势与曲线的关系：（1）当这个曲线由负变正时，topic的出现次数增长 （2）由正变负时，topic的关注度下降。

······然而，现在越来越多的OSN中的用户倾向于去发布非文本的信息，比如URL、图片和视频。为了处理这些非文本的内容，Takahashi提出使用message中包含的mention**（Here we mean by mentions links to other users of the same social network in the form of message-to, reply-to, retweet-of, or explicitly in the text. One post may contain a number of mentions.）**来定位突发性话题，而不是使用他的文本内容。Mention主要用于明确message发送的目标，使人参与讨论。他们开发了一种方法，结合了···和···，这个得分是通过计算每个用户的标准mention行为得来的，用一个概率模型进行估计。

·······

1. 信息传播的建模

信息传播过程的建模在阻止病毒传播、分析谣言如何传播等方面都受到广泛关注。

定义6（激活序列activation sequence）一个有序的节点集合，集合中的元素根据收到消息的先后顺序进行排序。

定义7（传播瀑布sperading cascade）一颗有向的树，根节点是激活序列中的第一个元素这棵树记录节点之间的相互影响（边表示节点之间的信息传播方向）、同时按照激活序列的顺序展开。

==================================================================

对于信息传播建模有两个主要方向，一类是了解信息是从哪传播、合适开始传播的；一类是信息具体是如何传播、以及为什么这样传播的。因此我们可以将描述信息传播的模型分为两类：解释性的模型和预测性的模型

* 1. 解释性的模型

解释性的模型的目标在于推断出潜在的传播瀑布，并给出一个完整的激活序列。这些模型使得根据信息传播路劲进行溯源成为可能，对于理解信息传播过程非常有用。

Gomez提出了一种根据节点间的相互影响次数来探索节点之间的相互关系，并由此推出整个传播瀑布的结构；该模型假设被激活的节点会以一定的概率分布影响他的每一个邻居节点（该影响是相互独立的）。因此，节点的激活时间不同会使得一个节点给其他节点传播信息的概率变小（？）。他们提出了NETINF，一个子模块间相互交互的优化算法，用于使得发现数据的概率最大化。

Gomez扩展了NETINF模型，并且提出将信息的传播过程建模为一个空间离散的网络 的连续的、状态各自独立的短暂过程（以不同的概率发生）。一个节点在一个给定的时间内是否会影响另一个节点，这个可能性用一个概率密度函数来描述，这个概率密度函数的生成依赖于信息的传播次数和任意两个节点间的传播可能性。扩展后的模型名为NETRATE，可以推断出节点两两间的传播可能性（传播率？），并且能够通过用公式表示并求解一个凸似然最大问题来将传播图构建出来。

这些方法都将待发现网络在相当一段时间内视为静态的。这显然不是一个令人满意的假定，因为OSN的拓扑结构是在随时间不断快速变化的，每时每刻都存在着边的生成和销毁。由于这个原因，Gomez又扩展了NETRATE模型，提出了一种随时间变化的推断算法，INFOPATH，使用随机梯度法去提供一个对时刻变化的网络的结构和短时间内的变化进行估计。

另外，因为某些技术性原因、以及爬取程序API的限制，该模型存在潜在的性能瓶颈，可能会丢失数据。为了克服这个问题，**一个方法**是想办法将爬取程序变得尽可能的高效。Choudhury分析了数据采样策略如何影响社交媒体中信息传播的发现。基于一些在twitter上的实验，他们得出结论：采样策略要同时考虑到网络的拓扑结构和用户的分布，例如活动和定位相对于简单的采样策略（随机采样或只基于活动进行采样）而言，允许我们去以更低的错误率获取到信息的传播情况。**另一个方法**是去将这部分遗失的信息考虑进去，以此构建新的模型。Sadikov提出了一种基于k树的模型，该模型仅需要一部分的激活序列便可估计出完整的传播瀑布的一些属性，如它的大小或深度。

* 1. 预测性的模型

这部分模型目标在于去预测一个特定的传播过程如何在一个特定的网络中展开。······我们将现存的模型分为两类：基于图的或是不基于图的方法。

* + 1. 基于图的方法

这个类型主要有两种模型：独立级联模型（IC）和线型阈值模型（LT），这两类模型都假设在传播过程中存在一个静态的拓扑图结构，主要将注意力放在传播过程的构建上。模型基于有向图，其中每一个节点可以被激活基于一个单调的假定，例如，已经被激活的节点不能重新回到未激活状态。IC模型对于每一条边都需要一个传播概率；而LT模型需要为每一条边定义一个影响度、为每个节点定义一个影响力阈值。对两个模型来说，传播过程沿着一个离散的时间轴同步向前进行，从一个初始被激活的节点集（一般被称为early adopters）开始。

定义8（early adopters）第一批收到消息的用户集合，是传播过程的起点。

在IC模型的情况下，对于每一次迭代，最新被激活的节点都尝试一次去激活他们的邻居节点；激活是否成功取决于这些边上定义的概率分布。在LT模型的情况下，对于每一次迭代，在与邻居节点相连的边影响力的总和超过自身的影响力阈值的情况下，未被激活的节点会被激活。成功的激活将会影响下一次的迭代。不管是哪一种模型，迭代会在无法进行新的传播时停下来，例如没有邻居节点可以联系了。可以看出这两种模型机制不同：IC模型是以发送方为中心的，而LT模型是以接收方为中心的。图6展示了一个IC模型的例子。

Galuba提出使用LT模型去预测信息传播图，······。他们的模型依赖于一些参数，例如信息病毒性、用户两两之间的影响力和用户接受某个信息的概率。LT模型被用于描述传播过程的开始阶段，具体需要使用梯度上升的方法来优化参数。然而LT模型无法处理现实中的网络的瞬时变化问题。

Saito放宽了传统的IC和LT模型的同步性假设，提出了异步的扩展模型，称为AsIC和AsLT，该模型沿着一个连续的时间轴进行迭代，需要和他们的同步的副本相同的参数，在此基础上给每条边加上一个时延参数。模型参数以参数方式定义，作者提供了一种方法，用于从节点属性中学习模型参数的功能依赖性。他们将任务表述为最大似然估计问题以及确保收敛的更新算法。 但是，他们只是试验了人造数据而没有提供实用的解决方案。

Guille等也将传播过程建模为异步独立级联。他们开发了T-BaSIC模型（即基于时间的异步独立级联），其中的参数不是固定的数值，而是根据时间起作用。 使用逻辑回归从社交，语义和时间节点的特征估计模型参数。

* + 1. 不基于图的方法

基于非图形的方法不假设已存在特定的图形结构，主要用于模拟流行病学过程。他们将节点分为几类（即状态），并专注于每个类中的节点的比例演变。 SIR和SIS是两个开创性的模型，其中S代表“易感”，I为“感染”（即采用了信息）、R为“恢复”（即有抗性，不会再被感染）。

不管是哪一种模型，S类型的节点会以一个固定的概率β转变成I类型；在SIS模型中，I类型的节点会以一个固定的概率γ转化回S类型的节点；在SIR模型中他们会有γ的概率消除感染，并永远处于R状态。每种类型的节点所占的百分数可以用一个微分方程表示。这两种模型都假设每一个节点与任意一个其他节点进行联系的概率是相同的。

Leskovec提出了一种简单的SIS模型变体，该模型需要一个参数β。它假定所有的节点都以相同的概率β接受信息，并且已经接受信息的那些节点在下一个时间片内又会变成易受感染的状态。这是一个很强的假设，因为在现实世界的社会网络中，影响不是均匀分布在所有的节点之间的，因此需要构建更加复杂的模型来描述这些特点。

基于一个假设：信息的传播被独立的节点掌管，Yang提出了一种方法。该方法专注于预测信息传播在短时间内的变化，使用一个时间序列来描述信息的传播速率。他们使用线性影响力模型（LIM），每个独立节点由影响力函数来定义传播率。影响力函数由一个无参数的形式表征。

Wang提出了一种偏微分方程模型去预测从指定节点开始的信息传播过程。更确切地说，该模型使用了描述传播行为的logistic方程模型来预测拓扑结构和瞬时变化。在这里的拓扑结构指的是每一个节点到源节点（最先得到消息的那个节点）的距离。过程的动态变化由一个logistic方程描述，该方程主要对被影响的用户（给定距离和时间）密度进行建模。这个拓扑的定义允许我们以简单的方式表达问题。这个模型的参数可以通过三次样条插值法进行估计。

1. 辨识有影响力的信息传播者

辨识出在一个网络中最有影响力的信息传播者对确保高效的信息传播具有重要意义。例如，社交媒体活动可以通过将信息投放给那些最具影响力的信息传播者来使得宣传效果最优。

定义9（K-CORE k核） G是一个网络拓扑图。设H是G的一个子图，σ（H）表示H中最小的度。则H中每一个节点都至少与σ（H）个节点相连。如果H是G的一个极大连通子图（极大指边的数量极大）且有σ（H）>=k，则称H为G的一个k核。

Kitsak提出：最好的信息传播者不一定是网络中度最高的那个节点。他们发现最高效的传播者是那些位于核中的节点。基本上，k核分解的原则是为每一个节点指定一个核心度ks，ks的值最小的节点集合位于网络的外围；最大值的节点集合位于网络的中心。最中心的节点就形成了网络的核。Brown发现经过k-shell（一个可以用来计算网络中节点影响力的方法。先找出所有度为1的节点，将其删除。然后在剩下的节点中继续查找度为1的节点，并删除。直到网络中没有度为1的节点，将之前删除的所有度为1的节点kshell值赋值为1,即KS=1。用相同方法接着查找度为2的节点，并赋值KS=2。接着是度为3、度为4……并分别赋值KS=3、KS=4……直到网络中所有节点都有kshell值。KS值越高，表示影响力越大。）分解后的twitter网络是高度偏态的。因此他提出一种改进算法，使用对数函数映射去生产更少、更有意义的k-shell值。

Cataldi提出使用众所周知的PageRank算法评估网络中的影响力分布。给定节点的PageRank值与在社交网络的随机游走中访问该节点的概率成比例。

以上的模型仅仅依赖与网络的拓扑结构，无视了其他的重要属性，比如每个节点的特性和他们传播信息的方式。基于一个发现：OSN中的成员绝大部分都是被动的消息消费者，Romero提出了一种基于图的方法（和广为人知的HITS算法、IP算法类似），为每一个用户设定一个相对影响力和消极程度评分（基于他们转发消息占比）。然而，没有个体可以是一个全局影响者，网络中有影响力的成员基本上只在某几个特定的专业领域有较大的影响力。因此，Pal提出了一种不基于图、topic敏感的方法。为此，他们定义了一组节点和主题功能，用于表征网络成员。 在此特征空间上使用概率聚类，他们使用集群内排名过程对节点进行排名，以识别给定主题的最有影响力和权威的人员。

Weng也提出了一种针对Twitter的topic敏感的PageRank算法，称为TwitterRank。

Kempe等人，采用不同的方法，并建议使用IC和LT模型（先前在4.2.1节中描述）来解决影响最大化问题。 对于参数k，该问题要求在网络中找到具有最大影响的k节点集。 给定节点集的影响对应于根据IC或LT在扩散过程结束时激活的节点的数量，使用该集合作为初始激活节点的集合。 它们使用基于子模块函数的贪婪爬山策略为此优化问题提供近似。