**Collaborative topic regression for predicting topic-based social influence**

**Abstract**

信息对不同用户的影响程度是不一样的，它不仅在不同的用户之间有所不同，而且在不同的主题中对同一用户也有所不同。社交网络的结构和用户生成的内容可以揭示关于用户及其基于主题的影响力的大量信息。虽然许多研究都考虑过测量全局用户的影响力，但是基于主题的用户影响力的测量和估计还没有得到充分的研究。本文提出了一种基于协同主题的社会影响模型，该模型融合了网络结构和用户生成内容，用于基于主题的影响度量和预测。我们预测基于主题的用户对未观察到的主题的影响，基于所观察到的基于主题的用户对其在社交网络中生成的内容和活动的影响，们对Twitter数据进行了实验分析，结果表明，在预测基于主题的用户影响方面，我们的模型在召回率、精确度、精确度和F-score方面优于最先进的方法。

**1.introduction**

社交网络如今无处不在，是一种重要的沟通方式，人们与广泛的在线用户分享他们的观点，讨论他们感兴趣的各种话题。社交网络的用户共享信息、想法和观点，这些可能会触发其他用户，从而影响他们。然而，不同用户的影响程度差别很大，对于一个用户来说，她的影响力可能会因话题的不同而明显不同。

在影响度量和预测方面，有基于主题和全局的用户影响预测，基于主题的影响研究与全局用户影响研究的关键区别之一是，除了考虑网络结构外，还考虑了用户生成的内容。另一方面，全局用户影响力是建立在网络结构的基础上的，在度量中忽略了主题。此外，还有影响最大化和传播，其目的是找到一组具有最高传播能力的用户。然而，影响最大化并不是本文研究的重点。据我们所知，目前还没有可用的研究来预测基于主题的用户对新主题的影响，而这方面还没有观察到的数据。

主题建模是基于主题的影响预测的主要阶段。LDA作为一个标准的主题模型，将文档看作主题的混合。每个主题都由一组单词表示，这些单词由单词包上的单词分布组成，这些主题是从文档中单词同时出现的模式中提取的。由于LDA在短文本语料库中出现的频率较低，单词的共现性较弱，如果在短文本语料库中使用LDA，将会产生很多问题。LDA很难在短文本语料库中推断出文档/词之间的相关性，词义变得更加模糊，因此，考虑到这些限制，需要专门的主题模型来从短文本语料库推断主题。

在本文中，我们反复使用观察到的主题和未观察到的主题术语。一个主题是否被观察到，定义如下。观察到的主题是我们的模型所针对的历史数据中发现的主题，而未观察到的主题是我们的模型预测其用户影响的那些主题。例如，如果我们用用户对US 2012 election topic的影响来训练我们的模型，那么这个topic就是observe topic，如果我们想预测用户对即将到来的话题的影响，比如2016年美国大选，这是一个未被观察到的话题。

基于主题的用户影响预测是本文的一个重要挑战和研究重点。这个任务对于不同的应用程序(如市场营销和竞选活动)非常重要。在这项工作中，我们预测基于主题的用户对未观察到的主题的影响，基于用户对他们在社交网络上发布的历史主题的影响的观察。提出了一种社会影响协同主题回归(SICTR)学习用户、主题和社会因素潜在空间的方法，用于估计用户对社交网络中未被发现的主题的社会影响。SICTR使用的是生成biterm主题模型，在该模型中，单词的共现性在整个语料库中被显式地建模。

我们的方法代表用户的主题兴趣和他们对每个观察到的主题的社会影响。更详细地说，我们的贡献是:

1.提出了一种基于主题的影响预测方法，将用户与主题的关系、主题内容信息和用户之间的社会关系整合到同一原则模型中。

2. 该模型没有考虑用户对用户的影响和全局用户的影响，而是考虑了单个用户对一个话题的影响和兴趣，从而能够预测用户对一个新话题的影响

3. 通过基于主题的用户影响预测实验，验证了考虑主题内容和用户主题矩阵共现的有效性:基于主题的方法比其他只考虑一般项目而忽略其特征的方法表现出更好的性能

**2.related work**

**2.1 topic-based influence**

在社交网络影响研究(总体用户影响)之后，基于主题的影响研究较少。最近，基于主题的影响力研究将用户帖子的内容与基于链接的指标结合起来。Haveliwala(2002)提出了PageRank的一个主题敏感扩展，用于对查询主题的查询结果进行排序。主题敏感的PageRank概念后来被用于Twitter等社交网络，并对其进行了调整，以对基于主题的用户影响力进行排名。Pal和count也对局部权威进行了研究数量(2011)。他们提出了一个基于高斯的排名来有效地对用户进行排名。他们使用概率聚类来过滤特征空间异常值，并表明提及和主题信号是排名权威中更重要的特征。Kong和Feng(2011)的目的是识别和排名那些发布高质量推文的用户。他们定义了一个基于主题的高质量推文，具有作者的主题特异性影响，主题相关作者的行为。他们在跟随者和转发的图表上应用了他们提出的度量。Hu等人(2013)以转发传播话题权威为假设，致力于检测话题权威。Sung等(2013)提出了PageRank的另一种扩展，与翁等(2010)不同，它不需要预定义的主题来实现基于主题的用户影响力。Liu等人(2014)基于话题的影响框架考虑转发频率和链接强度。采用基于泊松回归的潜在变量模型对用户之间的转发频率进行估计，得到了链路强度。

**2.2 Collaborative topic regression**

用户和项目信息被认为对改进性能很有价值。Wang和Blei(2011)提出了协同主题回归(Collaborative Topic Regression, CTR)，该方法利用用户和项目信息建立基于主题建模的协同过滤(CF)模型，进一步提高推荐性能。最近，HU等人(2018)结合了评分、项目评论和社会关系三个信息源，通过潜在因素和隐藏话题来预测评分。赵等人(2018)开发了一个异构的社会感知电影推荐系统，其中包括电影海报图像作为输入，以及文本数据、评分和社会关系。此外，Chen等(2017)提出了一种基于神经网络的推荐系统，并试图解决基于神经网络的算法计算开销大的效率问题。

与其他研究相比，本文在基于主题的影响度量方面又向前迈进了一步，提出了一种基于主题的用户影响度量方法，并预测用户对未观察到的top的影响

**3. background**

**3.1 Biterm topic model**

概率主题建模，如潜在狄利克雷分配(LDA)，通过检测一组潜在主题来表示语料库的低维空间。主题概率模型的基本思想是主为文档集合中每个单词的共现设置一个t隐藏变量。t可以包含k个主题，其中每个主题都是固定词汇表上的一个分布。给定一个语料库，一个文档可能包含多个主题，并且假定单词是由这些主题生成的。

给定一组由表示的文档，主题模型生成主题集合，表示为，文档的主题可以用多项式分布表示，每个主题都与m个单词的加权表示有关，表示为，其中wm是表示单词m属于主题tk的概率的权重。单词在所有主题上的分布P(w|t)可以表示为一个矩阵*φ*，其中每行*φk*与主题k相关，矩阵的条目为

BTM可以通过以下过程生成:

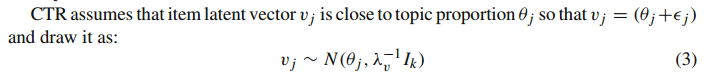
1.获得一个cita，生成文档的主题上的分布。这是从狄里克莱分布与一个特殊语料库中获得的。

2.对于每个主题生成：从狄里克莱分布中抽取*φk*

3.然后为每个biterm生成：通过利用主题上的分布分配主题，最后，在字典中通过biterms上的主题分布生成wm、1、wm、2，这意味着每个文档的biterms都来自于主题的混合。

**3.2 Matrix factorization (MF) and collaborative topic regression (CTR)**

CF通过历史用户行为分析用户之间的关系及其与项的关联，不需要显式的用户配置文件。CF的一个基本方法是基于邻域的方法，它分析了两者之间的关系。对于基于项目的方法，用户对项目j的评分是基于其对相似项目的评分来估计的，而基于用户的方法是通过查看其他具有相似兴趣的用户的评分行为来估计项目j的评分。CTR (Wang and Blei 2011)是在矩阵因子分解的基础上提出的，利用概率主题建模。它假设条目(如新闻和电影评论等文档)由主题模型生成，该模型表示为主题潜在向量vj。在CTR，用户由主题兴趣表示。与矩阵分解相似，CTR计算用户界面和项目vj的潜在参数，潜在变量ej根据用户对项的评分为用户捕捉主题之间的差异，表示为。其中N是高斯分布的概率密度函数平均值为零,方差等于λ−1 v和I = 1的指标函数,如果用户我评分项j和等于0。



**4. Topic-based social influence prediction**

**4.1 Problem definition**

假设G=(V,E)代表一个社交网络，假设用户发布了一组文本D=[d1,d2,…d3]，并且讨论不同主题，每个用户文本(post) dq包含一个或多个主题，并通过回复、点赞或重新发布从其他用户那里获得参与。一篇文章中其他用户的参与可以揭示这篇文章对读者的影响，基于社交关系和其他用户在社交网络中的参与度来量化每个用户基于话题的影响力，我们可以确定用户ui对话题tj的影响力，表示为Fij。所以我们有矩阵表示所有用户在不同主题下的影响。

给定用户列表、主题和每个用户对这些主题的社会影响，我们感兴趣的是预测一个未知的，用户ui对新主题tj的社会影响，其中t为主题集。具体来说，我们的目标是评估用户对未观察到的主题的影响，基于用户从其活动历史和生成内容中观察到的对主题的影响。

我们希望主题t中有影响力的用户能够对类似的主题t’也有影响力。假设基于主题的影响力权重相似的用户之间存在模式，则可以使用CF算法进行预测。然而，基于内容的方法仅使用内容信息进行推荐，例如，如果我们想预测话题tj的影响，我们可以利用主题集合T中最近邻的影响，基于主题的内容相似性。我们还可以将每个主题视为一个标签，并使用多标签方法根据内容信息训练分类器。基于共现的方法只使用用户主题矩阵F进行预测。例如，两个主题t和t’有很多共同用户，主题t中用户ui有很大的影响力，那么我们期望在t’中ui也有很大的影响力。基于内容的方法和基于共现的方法都忽略了有用的信息。因此，他们在社会影响预测方面无法取得令人满意的成绩。

**4.2 our approach**

为了测量和预测社交网络中未观察到的话题的社会影响，我们提出了SICTR，可以用来预测基于主题的影响。简而言之，我们的模型执行用户的两部分表示：根据用户的社交网络及其与其他用户的联系，对用户的潜在特征进行表征；基于用户活动主题的潜在特性表示。我们的方法采用协同主题回归(Collaborative Topic Regression, CTR)，即众所周知的将CF与主题建模相结合的方法，学习一个使用潜在主题空间来解释观察到的评分和观察到的单词的模型。我们建议使用SICTR来预测基于主题的用户在社交网络中的影响力。SICTR计算用户U和主题T的潜在参数。

**4.2.1 Social network-based representation**

我们想从社交网络g中派生出一个k维特性U来代表用户，用代表潜在用户矩阵，列向量Ui用于用户特定的潜在特征向量。将零均值球面高斯先验放在用户和因子特征向量上，得到用户和因子特征向量:



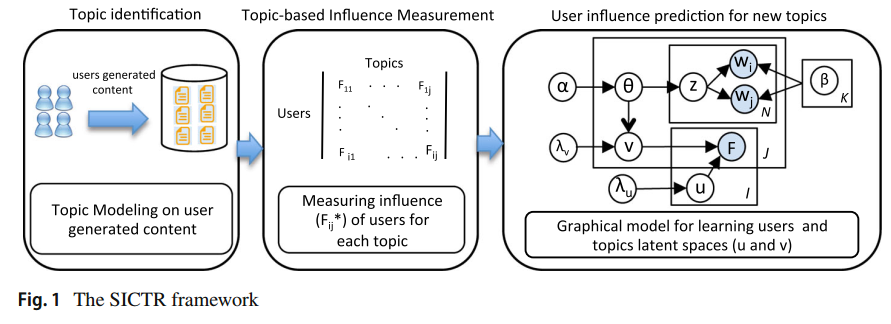
是高斯分布的概率密度函数，均值是u，方差是

**4.2.2 Topic-based representation**

在作者的模型中，条目是从社交网络用户生成内容的语料库中提取的主题，用户由他们感兴趣的主题表示。SICTR根据项目的相似性和其他用户对相似主题的影响力来预测用户某个主题的影响。我们通过对所有用户生成的文本应用概率主题建模来识别主题。每个主题都包含一组社交网络帖子及其所有相关信息和元数据，比如：内容、回复和重新发布。对于每个元组(user i,topic j)，我们测量了用户ui对主题tj的影响（4.5节中）通过这种方式,SICTR生成主题潜在空间和用户潜在空间。

SICTR的一个重要部分是生成主题潜在向量，其中ej捕捉到用户对于主题tj的兴趣，并且假设主题潜在向量vj与主题比例θj很接近。Fij的期望是θ的线性函数。

此外，在SICTR中，可以知道项目潜在向量vj与主题比例θj很接近，并且可以生成一个项目潜在向量

  
**4.5 influence measurement**

我们将社交网络中的社会影响力定义为用户在社交网络图中的重要性、用户的活动以及用户帖子中其他人的参与。通过不同的网络结构和用户在网络中的地位、用户帖子在网络中的传播规模、用户在社交网络中的活动和参与度、用户在网络中传播的信息内容等方式分析社交影响。

从网络结构上识别影响相关属性，如用户的朋友和用户在社交网络中的中心地位。从广播文本的内容中，我们可以识别出主题，从而了解用户对广播方面的影响。例如，在Twitter中，一条tweet可以包含用户的提及、接收回复和被其他用户转发。所有这些信息都可以揭示用户的社会影响力。

作者用表示收集到的文本集合，Dt是与主题t相关的文本，其中有T个主题。每个文本di包含一组属性，ui是文本的作者，ci是文本，Ri是用户列表中重新发布的文本，Mi是该文本的提及列表，fi是文本作者的追随者数量。

我们定义了四个影响指标：follower scale Ff：一个用户在网络上的朋友数；Topic Activity Fa：用户与主题有关的活动；Topic-based Attractiveness Fe：其他用户如何被userui的帖子吸引；Network centrality *Fc*：主题tj中用户在活动用户图中的中心地位。将这四种度量方式聚合为一种F\*来代表用户ui对于主题tj的影响。