# 增长归纳法的矩阵补全的Tumblr博客推荐

## 摘要

流行的微博网站，比如内容分享平台Tumblr，吸引了数以亿计的用户，在这里用户能通过推送的方式创作丰富的内容，并把这些那日容与关注他们的人分享。因为在这种服务上产生的巨大的信息量，一个重要的任务是为用户推荐获取有质量的推荐来让他们关注。除了把关注者图标作为主要信息源的传统的推荐系统设定，用户和博客的附加的边缘信息，比如用户行为（如点赞和转载）和丰富的内容（如文本和图片），也应被拿来挖掘来提高推荐性能。在这篇论文里，我们提出了一个新奇的针对博客推荐的增长归纳法的矩阵补全方法（BIMC）。BIMC是一个由两个部分组成的附加的低质的用户博客；一个部分捕获关注关系的低阶结构，另一个通过边缘信息捕获潜在的结构。我们的模型公式结合了最近提出的增长归纳法的矩阵补全（IMC）模型（针对边缘信息），和标准矩阵补全（MC）模型（针对低阶结构）。此外，我们利用最近研究的深度学习技术来获取文字和图片的语义丰富的特征代表，这一技术被应用在BIMC中。通过对从Tumblr获得大规模真实世界的数据集的实验，我们阐述了所推荐的BIMC方法的有效性。

## 1.介绍

结合了传统博客和社交网络的特征，微博服务作为一个领先的内容分享和交流的平台出现了。Tumblr是拥有超过230百万用户的最流行的微博服务之一，在这里用户可以创作并与他们的博客关注者分享他们的推送。相反地，用户通过关注他们感兴趣的博客，消费那些分享的内容，由于大量的选择，这已经成为了一件势不可挡的任务。因此，一个微博服务的核心问题是预测一个用户是否会关注一个博客。改良的博客推荐系统不会仅仅通过帮助用户发现感兴趣的内容来导向提高用户参与度，也会导向吸引更多吸引人的关注者来赞助或者推广。图1（a）展示了Tumblr的推荐模块。

推荐博客的问题在两个主要方面不同于传统的协同过滤设置，比如Netfilix评级预测问题。首先，用户与博客间的互动在关注形式下是二态的，没有关于用户表现的等级评定信息。“关注”信息代表了一个单向的无权重图，流行的基于图表结构特性的近邻测量可以被应用到这个问题中。第二，一个重要而有利的区别是，除了传统的用户信息矩阵（即关注者表），博客推荐本身需要丰富的边信息。边信息有两个主要类别：（1）用户产生的内容， 比如图片，标签和文字（例如，图1（b）），（2）用户行为，如点赞和转载。在Tumblr的例子中，合并图片特征是重要的，因为多数推送包含图片。文字信息在Tumblr中同样很丰富，因为相比Twiiter这样的其他的微博网站，Tumblr的推送在长度上没有限制。当大量的博客以这样的用户所创作内容为特征时，用户行为是一个更加直接和有益的体现用户偏好的信号，因为用户可以通过点赞和转载明确地表达他们的兴趣。这意味着点赞或转载同一篇推送的用户很可能关注着类似的博客。事实上，在许多已存在的研究中，这样的边信息不仅仅提高了推荐质量，同时也减轻了缺乏用户矩阵的问题。

另一方面，推荐系统中的合并边信息的严格途径正缺乏。考虑到标准矩阵补全（MC），这个方法是被广泛应用并且有完整理论研究的推荐方法之一，在最近这个方法已经被好几次严格担保有效了。然而，MC被曝缺乏信息并且受限于直推式设置，也就是说，预测仅仅适用于已经存在的用户或事物，因为它仅仅考虑用户食物矩阵的观察结果。最近，归纳矩阵补全（IMC）被推荐，并且通过驱动型设置来理论研究，在这种设置里，用户/事物的边信息以特征向量的形式可以得到。然而，IMC假设观察条目是被这些特征完全描述的，这在有噪点的特征中往往不适用，它们不支持用户事物矩阵。此外，IMC不能离开所有特征来为用户或事物做有意义的推荐，而这种情况在Tumblr中是常有的事。（看第三部分）。

有鉴于此，我们推荐一个新奇的针对博客推荐的增长归纳法的矩阵补全方法（BIMC），这个方法结合了归纳法的矩阵补全模型和推进型的标准矩阵补全模型。特别地，BIMC首先将MC模型应用于将矩阵平滑化，并且通过低阶近似值削减噪点级别，然后进一步用IMC模型将剩下的近似值模型化。换言之，BIMC既捕获了关注关系的低阶结构，也用附加的方法捕获MC没有获取的关注者图中的条目，通过用户和事物的边信息捕获了潜在结构。

通过合并用户/博客特征，BIMC也能在归纳设置中进行推荐，也就是说，为训练时不存在的用户和博客做预测，这些预测包括冷启动案例。这对Tumblr特别重要，因为用户和博客在关注者图中常常有很少甚至没有链接，像在第三部分所示的那样。从Tumblr实验得到的大规模真实世界的信息揭示了，我们推荐的BIMC显著超过MC，IMC和几个其他标准的博客推荐方法。

最后，一个重要的问题是，如何有效地展示三种边信息源（图像，文字和行为）作为特征。最近，深度学习方法作为一类强有力的模型浮现，它能理解图片上的语义内容，让图像识别具备最先进的表现。这也是文字信息的例子，捕获单词间关系的单词的向量表示能够从神经网络中学习到。被这些结果所激励，深度学习作为一个对用户和博客有效并有鲁棒性的表现，我们对图片和标签/文字都采用深度学习特征。对于行为特征，我们表达点赞和转载来作为有权表，这与关注表相似，我们计算行为表的主要成分作为特征。据我们所知，我们是第一个同时考虑图片和行为特征的。此外，在推荐系统中采用深度学习过的特征仍未被勘探过。

我们的贡献主要有如下几点：

* 我们提高了基于增长型归纳的矩阵补全的博客推荐系统，它结合了归纳法的矩阵补全模型和的标准矩阵补全模型。
* 我们用一个大量边信息源的集合，比如用户行为，文字/标签，和图片，来代表用户和博客；并且利用先进的深度学习方式提取综合特征集合。
* 我们展示了，推荐的BIMC模型有效地结合了从各种源获得的不同用户和博客特征，来得到更准确的推荐。
* 我们实施广泛实验，并且从Tumblr上获取大规模真实世界数据进行详细分析，证明了推荐的BIMC是比基准先进几倍的方法。

论文的剩余部分如下安排。在第二部分，我们回顾一些紧密相关的工作。然后我们在第三部分中分析了Tumblr数据，并且研究了它的一些重要特征。接着我们在第四部分展现我们推荐的博客推荐方法，并在第五部分给出提炼出的用户和博客的细节。实验结果在第六部分给出，接着在第七部分给出总结。

## 2.相关工作

一般而言，除了传统用户事务矩阵外的多种信息源可以促进推荐性能。有边信息的推荐系统绝不是新有的，许多方法已经基于他们所用的边信息的类型而被提议，比如用户生产的内容，用户/信息轮廓和属性，社交网络和上下文信息。一个最近的关于先进方法的综合调查可以被找到。

在MC的基础上扩展边信息的方法之一是共同矩阵补全（CMC）模型，它的目标是用共享的低阶结构共同恢复矩阵的集合。用户事物矩阵和基于标签信息的用户相似性矩阵共同分解，来帮助更好的推荐。最近的CMC的工作提供了某些假设下的一致性保证，这些假设由于一个普遍的结构是受限制的。有社交网络的推荐系统大部分基于客观上有附加限制的潜在因素模型，比如相连用户间的潜在因素。另一个方法是基于回归的潜在因素模型，它的属性信息被综合进了模型里。然而，这个被推荐的方法不能很好地扩展于大数据集。基于图表的方法也已被拓展来合并边信息。举个例子，构造一个带有社交和标签信息的多部分的图表，它不能很好的扩展于附加边信息，或者当特征被用稠密的矩阵表现。最后，用户产生的内容，比如评论，被通过分析感知信息来利用。在大多数案例中，方法要么专门用于某一特点的信息源，要么不能很好的扩展到大量特征，并且缺少理论支持。

与Tumblr博客推荐很接近的是Twitter的关注谁系统.。之前的关注推荐方法包含一种概率性的模型，这种模型基于概率性的潜在语义分析。一个基于社区的方法提出，矩阵分解被独立地应用于每个发现的社区。然而，这两种方法都没有考虑到其他的明确的用户/博客特征。在一中基于近似性的协同过滤方法中，关注者/被关注者，和内容信息被用来代表用户。相近地，信息系统首次识别一组关注候选，这在关注者图表中相邻2个位置，然后用基于内容的用户档案提炼列表。基于图表的运用点之间的邻接测量方法也被应用于关注推荐中。一个主要的缺点是那些方法不能有效地处理归纳的设置。此外，没有一种已有的方法考虑图片信息和用户行为信息，这在Twitter上也是可得的。

利用深度学习方法在推荐系统上是受限制的。有一个例外，音乐推荐系统。传统的矩阵分解结合了一个深度卷积神经网络来学习一个函数，这个函数能将音乐内容特征和相应的潜在因素相匹配。另一个例外是O. U. Florez 和 L. Nachman的语义推荐系统，在这个推荐系统中，一个周期性的神经网络被训练用于捕获文档的语义。这两个研究都展示了深度学习对推荐系统是一个有希望的方法。

## 3.数据集特征

在这个部分，我们分析一些Tumblr数据的重要的特征。作为一个社交网络服务，Tumblr用户可以像Twitter一样，无需彼此认证就能关注感兴趣的博客，这与facebook不同。关注信息可以用双向图来表示，双向图的结点表示用户和博客，从结点i到结点j的边表示用户i关注博客j。我们用一个2014年6月抽样来的关注图快照，这个图包含了7686万的结点和22.7亿的边。我们在《什么是Tumblr》里发现相似的特征，包括在图2（a）里展示的出度和入度。入度遵循幂律分布，然而出度不遵循，出度在大约5000的时候会急剧下降，这是Tumblr里用户可以关注的播客的最大数。大约50%的结点没有任何粉丝（也就是说，入度为0），最大的入度是522万，然而大约25%的结点没有关注任何博客（也就是说，出度为0），最大的出度是14208。

作为一个微博平台，Tumblr提供接近于传统博客网站的有用的工具，来创建更长，更丰富，更高质量的内容。特别地，它允许用户创建8种不同的推送：照片，文字，答案，链接，引用，视频，音频和聊天。此外，Tumblr的推送没有任何长度上的限制，不像其他微博网站如Twitter,它限制每篇推送140个字符。它也支持每篇推送的标签的使用，这是和推送内容分离的。最后，用户可以给一篇推送点赞，或通过转载再次广播推送给自己的粉丝。尽管这两种行为对用户来说是有不同的用意，它们都能直接反应用户的兴趣点，这可以被利用来提高推荐质量。

我们处理了5个月的Tumblr数据，每个月包含大约1.5TB的推送创建，转载和点赞的抽样记录。注意，我们只关注每个月至少有5条动态的用户。平均而言，每个月都有超过15000万新创建的推送，25亿转载的推送和20亿点赞。我们在图2（b）展示每条推送类型的分布情况。接近80%的推送是图片推送，这意味着图片特征是分析推送重要的组成部分。图2（c）记录了每种类型的推送被转载的平均数。我们从图中可以看到，怕，引用和聊天推送比其他类型的推送都要被转载的多的多。总的来说，一条新的推送平均被转载超过15次，这阐述了Tumblr的高度分享行为。我们也可以看见，大约8.3%的用户没有任何推送，大约12.2%的用户没有任何行为信息。更多详细的Tumblr数据分析可以在《什么是Tumblr》中找到。

## 4.方法

在这个部分，我们描述了一种自然的方式