# 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。

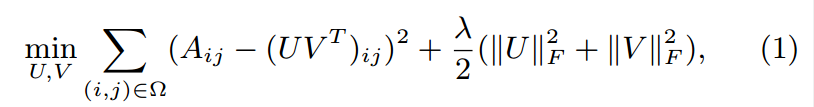
## 4.方法

在这个部分，我们描述了一种结合多种用户/博客特征和关注图像来做归纳设置的自然的方式，比如给新用户和博客做推荐。我们首次为博客推荐描述归纳矩阵补全方法，这个方法基于将低阶矩阵应用到用户博客特征的用户博客关注表现。接下来我们将展现我们所推荐的增长型的归纳矩阵补全方法。我们将再描述我们所推荐的方法前，简短证明几个所用记号。

记号：我们用G = (V1; V2; E)表示关注者图表，V1 (m = |V1|) 和 V2 (n = |V2|) 是用户和博客的集合，E = {eij|i∈V1, j∈V2}是用来描述用户i关注博客j的边的集合。让A∈Rmxn是G 的邻接矩阵，行表示用户，列表示博客，这样Aij=1就代表用户i关注博客j，0的话则相反。那就是说，我们将空值视为0。注意G是一个有向图表，既A不是对称的。用X∈Rmxf，Y∈Rnxfb分别表示用户和博客的特征矩阵。

### 4.1矩阵补全

低阶矩阵补全（MC）方法是最流行和成功的博客系统的协同过滤方法之一。目的是用观察到的A的条目恢复潜在的低阶矩阵，用公式表示为：



和中，r表示潜在特征空间的维度；是观察到的条目的集合；λ是一个调整参数。注意矩阵补全仅仅利用关注者图表A的样例，并忽略系统中可能展现的边信息。

### 4.2归纳的矩阵补全

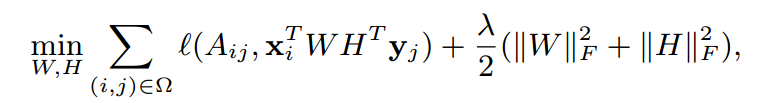
标准的矩阵补全公式受限于直推式设置，即预测仅仅由已存在的用户和内容生产，不会由新用户和内容的潜在因素训练而成。此外，标准公式会因信息的贫瘠而极大压制其表现，在Tumblr的案例里，大约50%的用户没有任何粉丝，25%的用户没有关注任何博客。一个简单的对这样的用户做预测的方式是用整体的博客排名来判断流行趋势，并推荐前几名。为了做出有意义的预测，需要更多关于用户和博客的信息。Tumblr通过丰富的内容（照片，文字）和行为（转载，点赞）信息来获取这样的信息。

最近，一个新奇的递归矩阵补全（IMC）方法被提出，P. Jain and I. S. Dhillon对这种方法进行理论研究，来减轻信息匮乏问题，并且利用用户和博客共同的边信息给新用户和内容做预测，这些边信息以特征向量的形式给出。中心思想是利用用户i的特征向量，用户j的特征向量，和一个低阶矩阵来实现Aij，表示如下：



亦即，用户i和内容j的互动关系是由将它们各自的特征向量应用到Z上生成的。对于一个新内容b，对于每个用户i的预测Aib可以用特征向量yb计算出来。

将因式分解，IMC的目标是用在A中观察到的条目将和复原。IMC表示如下：



对于某些测量观察值和预测间的不同的损失函数l，比如平方损失函数或logistic回归函数。要学习的参数是仅仅依赖于用户和事物特征的数量，然而在标准矩阵补全中，有个参数。注意矩阵补全是IMC在X=I,Y=I时的特例。

对于一个凸的损失函数l，上面的IMC目标在W或H修正时（与标准矩阵补全案例相似）变成了一个凸函数。解决最优化问题的计算量取决于损失函数l的选择。在我们的实验中，我们用目标的平方损失，并使用《缺少标签的大规模多标签学习》中的替代的最小化方法。在这种设置下，当nnz（A）是A中不为0的数，c是一个小常量时，每一步的复杂度是。在我们的实验里和r非常小（小几百），解决聚拢在10次迭代内。

### 4.3 增长递归矩阵补全

接下来我们展示我们称为增长递归矩阵补全（BIMC）的方法。一个与IMC相关的问题是，模型太死板，因为它严重依赖于用户特征X和事物特征Y。也就是说从不同来源得到的用户和事物特征应该支持潜在的关注者模型A的结构，使得能够做出好的预测。使，其中是X的SVD。近似地，使，是Y的SVD。从IMC公式中，我们得到



其中。因此，占据的子空间必须与A所占据的子空间有重大重叠，来获得小差错。比如，点赞和转载行为特征为直接的用户兴趣反应，会非常有帮助。相似的参数也适用于Y。

然而，多种来源的特征可能不是总是支持矩阵A的，IMC在这样的案例里可能会表现大减。举个例子，Tumblr中的文本信息极端稀少和杂乱，因此不能直接揭露用户偏好。此外，并不是所有用户和事物都有特征（就如第三部分展示的那样），在这种情况下IMC不能够做任何预测。

为了解决这些问题，我们提出结合标准矩阵补全和递归矩阵补全，从而更好地利用两种方法的力量。亦即，我们结合了MC的理论来减少输入数据的噪点等级，同时利用IMC的优势来合并用户和事物的边信息。我们塑造Aij模型的公式为：

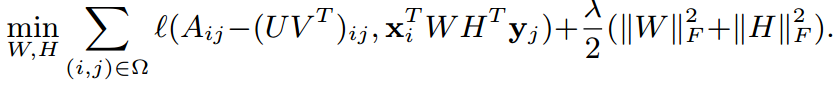


其中参数调整了最终预测中特征的权重。选择一个好的参数对性能和解决最优化问题都至关重要，而这个参数很难调整。此外，同时解决所有四个潜在因素矩阵U,V,W,H，由于增长的参数数量，将导致较慢的聚合。

因此，我们的策略是首先学习MC模型的潜在矩阵U和V。结果的近似误差，或者说剩余矩阵代表了关注者矩阵的链接，这些链接MC不能完全捕获。然后我们用IMC构造Rij模型：



换句话说，我们首先通过潜在因素U和V试着找到关注者图表A的支持，并集中于IMC无法精确塑造模型的部分。这在剩余的标准很大时很有用，它暗示着与A的低阶结构的巨大偏差。我们的目标是：



当l成为平方损失并修正H，上面的矩阵中的目标的斜率为：



注意我们不一定要明确地形成，它会是一个浓厚的矩阵，并且无法存在内存中。

我们的方法消除了找到合适参数的必要。此外，现存的有理论支持MC和IMC的高效解决方法可以被直接应用。图3对我们推荐的BIMC模型有个综述，我们能看见BIMC可以处理A的贫乏和没有特征的用户/事物。假设用户i有特征xi，博客j有特征yj，我们用已学得的因素和设置为1的参数来获取预测。最好，我们注意到一个相反的方法也能够使用，即首先学习IMC模型，然后对结果剩余矩阵进行MC模型的训练；我们发现结果是可以比较的，所以我们仅仅展现前者的结果。

我们阐述在对MovieLens-100K 数据集做实验，特征很杂乱的情况下，BIMC超出IMC的优势。我们计算用户-电影矩阵和集合，以及的排名20的SVD，左边和右边的A的单一向量分别是用户和电影特征。然后我们通过增加干扰扰乱X的列，并测量相关的IMC的近似误差：和BIMC的近似误差：。结果在图4给出，证实了BIMC由于给剩余矩阵建模，比IMC获取了更低的近似误差比。这说明了BIMC对于杂乱特征更加健壮，反之IMC这在用户特征X中的干扰等级增加时表现大大下降。

## 5.特征抽取

如同在第三部分描述的那样，Tumblr数据包含了三种用户和博客的边信息的主要来源：（1）点赞和转载，（2）标签和文本，（3）图像。从这些来源抽取有用的特征是推荐系统的关键步骤。我们下面会讨论如何抽取这些用户和博客特征的细节。

**用户行为**：对于用户行为信息，我们使用从Tuumblr抽取的1百万用户和博客的点赞和转载信息。博客点赞和转载信息可以代表一个类似于关注者图A的有权图，用户i和博客j的边信息权重被设置为用户i点赞和转载博客j的数量。点赞和转载的边权重都符合幂次定律分布。平均而言，每个用户从27个博客中点赞或者转载大约340条推送，每个博客得到总共410个点赞或者转载。在实验里，我们从训练数据里将点赞和转载图表合计到一个单一的行为图表，并用边权重的对数。一种获取有用并且鲁棒特征的方法是考虑与行为图表一致的邻接矩阵的主要成分。也就是说，我们计算p个主要成分，并把它们当作IMC潜在用户和博客的特征。因此，我们有用户和博客的用户行为特征，我们安装经验在实验中设置p=500。

**标签和文本：**用在Tumblr里的推送的标签和文本非常稀疏和杂乱。平均每个推送有28.7个单词，4.8个标签。此外，标签在Tumblr里是不受限制的，一个用户可以放任意文本进去。已存在的基于单词包（如，LSA,LDA）的模型在这样的问题上会表现不佳。因此，我们利用word2vec，一个最近的受神经网络启发的方法，它学习植入向量空间的单词。word2vec利用叫做skip-gram的技术和阴性采样，它们试图用向量表示每个单词以致相似文本中的单词离彼此很近。这个表示通过将每个出现在训练集中单词的预测可能性最大化来完成。在我们的工作中，我们首先计算每个用word2vec的的单词的d维向量，并用k中算法将这些单词聚合到c个群集中。用群集信息，我们最终创作出每个推送的单词群集直方图来作为简洁的用在博客中的标签和文本表示。设d=300，c=1000，并为每月的训练数据的文本特征做加工。

**图像：**就如第三部分展示的那样，图像是Tumblr数据的重要部分。我们每月从出现在训练集中的博客里随机取样80万图片。由于Tummblr数据集的图像标签不好用，我们对1.5M带标签的网络相册图片训练卷积神经网络（CNN）。CNN由7个隐藏层组成，其中有5个连续的卷积层，其后跟着两个完全连接的层，最后加上一个soft-max层。在CNN里的每个神经元非线性性是由修正线性单元（ReLUs）f(x) = max(0,x)建模的，它比起tanh单元这样的饱和非线性加速了学习速度。CNN使用一个224x224像素的RGB图像作为输入。每个卷积层，用一系列学习好的核心，ReLU非线性，两个可选层，本地响应标准化和最大合并，来缠绕它前一层的输出结果。本地响应标准化层应用在了特征频道上，最大合并层应用于邻接神经元。第七层的输出注入最后的soft-max层，soft-max层能对输入图片的预定义的958个分类输出自信分数。使用神经网络，我们从抽样Tumblr图像中抽出深度学习特征。对于用户来说，我们给用户推送的所有图像的特征向量，点赞和转载，给出平均值。对于博客，仅有被推送和转载的图像被认为是转载的推送也是博客的推送。

## 6.实验评估

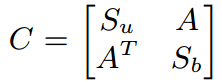
在这个部分，我们展示与IMC相比，用来评估我们推荐的BIMC模型的实验的设置，以及几个其他的Tumblr数据集的博客推荐的基本方法，这些方法有额外的用户和博客的边信息。从Tumblr关注者图像，我们随机取样了1百万用户和有约1.2亿关注的博客，即A中的非0元素。从Tumblr推送数据中收集了超过5个月的用户行为和用户生成的内容信息。

### 6.1基线和评估度量

我们执行离线和即时评估。对于离线评估，我们用10倍交叉生效。即时评估用来模仿联机评估，我们用之前的数据，4个月训练，其余的月份测试。

在两个案例中，我们将BIMC和标准矩阵补全公式（MC）与单值分解（SVD）作比较，SVD被认为是能够做好前N种推荐任务的模型。我们也比较吸收边信息包含归纳矩阵补全（MC）的方法。另一个流行的方法是收集矩阵补全（CMC）。CMC的目标是共同覆盖有共享低阶结构的矩阵的集合，这与IMC不同。特别地，假设用户-事物矩阵A，用户特征X和事物特征Y，CMC找到一个连接的因式分解如下：



也就是说，共享用户潜在因素矩阵U从A和X里获取（相似地V从A和Y中获取）。最近的工作提供了CMC的一致性保证，因此我们使用该算法。Katez测量，最成功的连接预测的邻近测量之一，作为一个基于图像的方法，也在对比中包含了。我们计算用户与博客间的Katz值，用了结合（对称的）矩阵，Su和Sb分别是用户和博客的相似矩阵，它们通过各自的特征得到。最后，我们报告使用简单的全局流行性排名（全局的）的结果作为基线，博客通过关注者数排名。我们在MC和SVD中使用排名r=10，在CMC，IMC，BIMC中使用排名r=100，并为所有方法设置λ = 0.1，这个值是由使用交叉确认决定的。

我们用每个方法生成的排名前十的精度(PRC@10)和回忆(RCL@10)来测量推荐性能，这是推荐系统的部分兴趣范围。为了完整性，我们也报告每种方法的AUC（ROC曲线下的区域）

### 6.2实验结果

对照基线：Global,SVD, MC, Katz, CMC,和 PRC@10、RCL@10的IMC，以及表一中离线评估的AUC和表二中的即时评估（它模拟了A/B测试条件），含有用户和博客的特征的BIMC方法的结果被展现出来。

#### 6.2.1性能比较

这是件非常有趣的事情来看看表格1，根据AUC而言，简单的全局方法胜过SVD和MC基线。这可以被大多数用户关注高度流行的博客，比如公众号或者名人的事实来解释，SVD和MC都受数据贫乏的限制。然而，我们需要注意的是，Global在精度和回忆结果上胜于SVD和MC。表1也展示了我们所推荐的方法，BIMC，在所有三种评估矩阵之中实现了最佳性能。注意两种最好的实现方法，BIMC和IMC，都利用了用户和博客的边信息。这暗示着这样的信息对提高推荐质量是重要的。

相比之下，CMC在列出前k个中表现最差，就如表1所展示的那样，关注者图和用户/事物特征之间不存在重要的共享低阶结构。此外，Tumblr中的文本特征特别贫乏和杂乱，这使得CMC公式更加不确定。这清晰证明了BIMC和IMC吸收了用户/事物特征，是更加有效的健壮的方法。尽管如此，相比IMC,CMC仍然能够实现类似于AUC的结果。Katz根据AUC，完成了类似于BIMC的实现，但不是在列出前k上，它可以被Su和Sb的近似度被杂乱特征影响的事实来解释。总之，我们可以见到BIMC通过成功合并丰富的用户和博客特征，相比其他方法有更好的性能。

#### 6.2.2即时评估

尽管交叉确认是一个广泛接受的评估方法论，当即时效果不被考虑，当将分裂信息注入训练和测试集中，它可以产生有偏见的结果。因此，我们使用修正的日期和时间将另一个数据集分开注入训练集和测试集，来评估所有的方法。特别地，我们用一个4个月周期的数据来训练，接下来的第五个月的数据来测试。这个评估更加类似于对A/B测试进行离线评估，这在工业上被广泛使用。即时评估结果在表2中给出，在那里我们能观察到和表1中的离线案例非常近似的结果。这暗示着我们所推荐的方法在产品设置中会表现的很出色。

#### 6.2.3不同贫乏级别下的用户和事物表现

为了更好的理解在BIMC中利用丰富的用户和事物信息的作用，我们将用户和博客基于他们关注的人的数量和关注他们的人的数量来给他们分类。换句话说，用户和博客的分割基于A中的非零元素的数量完成，来检验不同数据贫乏级别下的方法的表现。特别地，用户基于关注者数量被分为三个组：（低），（中等）和（高），每个组含有大约89.36%，7.81%和2.83%用户。类似地，我们也对博客维度分类（有相同的阈值），每个组各自有95.12%，3.26%和1.63%的博客。

对于每个用户分类，我们在图五展现Recall@k，k = 1; 2; · · · ; 20。像图5中展示的那样，BIMC按照Recall@k超过所有其他全用户组的基线。第二好的方法是IMC，紧接着是Katz。这明确地展示了利用用户和事物特征能够帮助处理不同的条件缺乏，包括冷启动。对于图5中的低级用户群体，我们能发现一个有趣的现象，SVD和MC在严峻的缺乏中表现糟糕，因此和Global基线表现的差不多。注意到，当我们从低级用户组向高级用户组移动，所有方法的性能都下架了。这可以被以及关注了许多流行博客的用户需要被推荐更多多种的博客，这通常来讲是个更加困难的任务。表3表现了所有三组用户组的AUC结果，我们可以看见BIMC在所有用户分类中取得了最大的AUC值。像在6.2.1部分中讨论的那样，CMC不能在列出前k排名上做任何好的预测，但仍然像表3所示的，拿到了合理的AUC等级。这组结果展现了BIMC通过合并丰富的用户和博客特征成功地处理数据缺乏，并且显著胜过其他方法。

接下来，我们分析所有有不同流行等级的博客群组的所有方法的性能，如图6所示，我们能观察到类似于图5的不同用户群组的趋势。在图6中，我们能发现BIMC和IMC根据Recall@k胜过所有其他博客群组的基线，然而所有其他基线受困于数据缺乏，不能做任何准确的检索。对于更加流行的博客，它们的表现更好。对于这个群组，SVD和MC比Global表现的稍微更好些，但是仍然比BIMC和IMC要差很多。另一个有趣的结果是，IMC和BIMC在中低流行度的事物上表现的差不多。这可以被BIMC中的MC步骤在这两个案例中都受困于数据缺乏，并且并不能帮助BIMC像有高流行度的事物的案例中那样表现，的事实所解释。

最后，我们共同分析所有用户群组和事物群组的所有方法的性能。特别地，图7和图8分别根据Precision@k 和 Recall@k展示了所有方法对所有用户和事物群组的共同表现。我们可以观察到，精度和回忆对有高行为/流行度的用户和事物提升显著。对于有中低行为的用户，所有除了BIMC和IMC的方法都严重受困于数据缺乏。对于有高行为的用户，BIMC比IMC表现的好，而这两者都显著超过其他基线。另一个有趣的结果是，IMC和BIMC在低行为/流行度的用户和事物上表现的差不多，这表示BIMC中的MC步骤受困于数据缺乏，不能有效地帮助IMC，所有的BIMC的预测都仅仅依赖于IMC的步骤。总的来说，这系列实验清晰地证明了BIMC优于IMC的力量，也利用了BIMC和IMC中的超过其他基线的丰富的用户和事物特征集。

## 7．总结

要关注的推荐博客是在线微博网站，如Tumblr，的核心任务之一，用来提高用户参与度，并提高广告税收。在这篇论文里，我们为归纳矩阵推荐了一种新奇的增长型归纳矩阵补全（BIMC）模型，这种模型结合了归纳矩阵补全模型和标准矩阵补全的力量。推荐的BIMC模型集中于由标准矩阵补全（MC）模型的近似矩阵计算出的剩余矩阵，并且学习了归纳矩阵补全模型（IMC）来有效利用用户和博客的丰富的边信息，来学习标准MC不能学习的关注者表中缺失的连接。我们利用最先进的深度学习方法，如word2vec和卷积神经网络来抽取特征的综合集。一个从Tumblr得出的大规模真实世界数据的广泛实验集合证明了推荐的BIMC在有效性上超过MC和IMC方法，以及其他集中基线。