LDA 模型的应用 -以微信娱乐公众号为例

2017 年 5 月 21 日

摘要

随着智能终端的发展与普及，微信公众号推送成为人们获取信息的一个重要途径。本文从文本挖掘入手，获取热门八卦微信公众号的推送信息，应用 LDA 主题模型提取文本的主题，从所提取的主题中对相应时间段的八卦新闻进行总结概括，最后通过方差分析，找出影响微信推送点赞数的重要因素，从而为微信公众号广告营销策略提供有效参考。

关键词：文本挖掘，LDA 主题模型，方差分析，微信公众号，广告营销

目录

1 引言

1.1 研究目的

本文旨在利用文本挖掘技术分析微信娱乐公众号历史文章，探究热点话题，发现订阅读者的兴趣点，指导微信公众号的广告营销策略，优化推广效果。

1.2 背景

当手机称为人们的随身设备，微信改变人们的通讯习惯时，“微信公众号”这一种新媒体形式渗透进人们的生活。自然而然的，软文营销成为这种媒体形式的主要广告手段和创收形式。软文推广；顾名思义，它是相对于硬性广告而言，由企业的市场策划人员或广告公司的文案人员来负责撰写的“文字广告”。与硬广告相比，软文之所以叫做软文，精妙之处就在于一个“软”字，好似绵里藏针，收而不露，克敌于无形。等到你发现这是一篇软文的时候，你已经冷不盯的掉入了被精心设计过的“软文广告”陷阱。

阅读量和点赞数无疑是衡量一篇软广文效果最直接的指标，而什么样的公众号文章最容易获得人们的关注呢？本文利用文本挖掘手段，深度挖掘 3 大娱乐公众号的过去近一年的历史文章，发现最“吸睛”的话题模式，指导公众号打好软广的“组合拳”。

1.3 研究问题

1. 娱乐公众号都在聊些什么？
2. 哪些话题、哪些明星最受娱乐公众号的青睐？热门话题中谁是头条 MVP？
3. 话题与涉及明星等因素是否显著影响文章点击热度？

2 文献综述

文本挖掘技术被广泛运用在各个学科中，用于处理大批量数据。现在，对于文本挖掘，通用的定义为：文本挖掘的过程即是从文本数据中挖掘内在的，未知的，且有用的模型的过程（Tan，1999）。因此，文本挖掘技术可以在大批量数据中高效提取出有用的信息。文本挖掘分析技术包括文本结构分析、文本摘要、文本分类、文本聚类、文本关联分析、分布分析和趋势预测等（袁军鹏等，2006）。郭金龙等（2012）学者认为在人文社科领域中较为常用的文本挖掘技术是文本分类技术和文本聚类技术。文本分类技术属于有监督的机器学习应用，采用的常见技术有：决策树、朴素贝叶斯 (NB)、支持向量机 (SVM)、K-近邻等，可以用于对主题的分类、对风格的分类、对情感的分类以及文章的题材等分类。在文本分类技术中。然而文本分类技术的最终结果与所采取的停用词等有很强的关系，而且自动化程度不高。而文本聚类技术属于无监督的机器学习应用。文本聚类技术虽然也需要分词，但不需要重复复杂的训练，自动化程度较为文本分类技术更高。文本聚类技术中，用于挖掘潜在语义知识的模型有：LSA，PLSA 以及 LDA 等模型。董婧灵（2011）等通过比较研究，认为 LDA 模型有着较为突出的优点：首先，LDA 是完全概率生成模型，具有丰富的结构，成熟的算法和训练模型；其次，LDA 模型更适合在大规模语料库中构建模型。LDA 模型是由 Blei 等（2003）提出的, 是一个集合概率模型。近十几年来，LDA 模型在实际运用上都有广泛的讨论。例如：唐晓波（2014）等将 LDA 模型运用在微博热点的搜集。然而，这些文献较少涉及到微信平台数据的使用，以及将其所得模型运用到实际问题中。因此本文将运用 LDA 模型去研究微信娱乐圈公众号数据，并将其运用至广告营销中。

3 方法

3.1 LDA 模型

潜在狄利克雷分布（Latent Dirichlet allocation,LDA）主题模型，是文本挖掘中著名的生成概率模型。它由 David M. Blei、Andrew Y. Ng、Michael I. Jordan 在 2013 年提出。

符号

1. 一个词语是该离散数据中最基本的单位，词典中所有词语由 $1, 2, ..., V$ 索引。每个词语可由一个单位基向量表示，即词典中的第 v 个词表示为 V 维向量 w ，其中 $w^v = 1$ $w^u = 0$,$u \neq v$;

2. 一篇具有 N 个词的文档记为词序列 $\overrightarrow{w} = (w_1, w_2, ..., w_N)$ ，其中 wn 是序列中的第 n 个词语;

3. M 篇文档组成的语料库记为 $D = \overrightarrow{w_1}, \overrightarrow{w_2}, ..., \overrightarrow{w_M}$。

3.2 DGP:

LDA 是一个层次贝叶斯模型，它的基本思想是：一篇文章可能具有多个主题，而文档的主题分布服从一个潜在的狄利克雷分布，而每一个主题代表一种词语分布，即一篇文档的生成服从以下步骤：

1. 选择 $N \sim Poission(\xi)$;

2. 选择 $\theta \sim Dir(\alpha)$;

3. 对于 N 个词中的每一个词语 w_n :

- 选择其来自于哪一个主题 $z_n \sim Multinomial(\theta)$;
- 从多项式条件分布 $p(w_n|z_n, \beta)$ 中生成一个词语 w_n。

其中假设主题数 k（狄利克雷分布的维度）是已知且固定的；给定主题，词语的条件分布 kxV 维矩阵 β 其中 $\beta_{ij} = p(w^j = 1|z^i = 1)$ 是未知非随机矩阵。文档长度 N 服从泊松分布，且与过程 1.、2. 独立。

结构可表述如下图

需注意与简单狄利克雷 -多项式分布聚类模型不同，LDA 模型允许一个文档具有多个主题。具有图 [LDA 结构图] 所示结构的模型在贝叶斯方法中被称为层次模型（Gelman et al., 1995），或条件独立层次模型

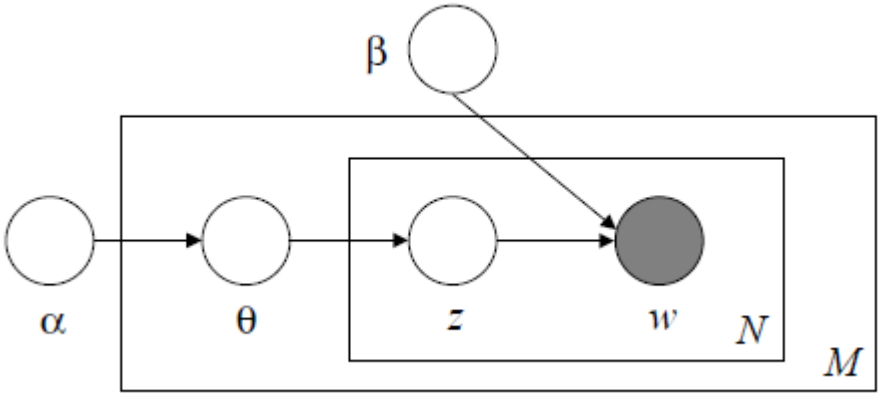

图 1: LDA 结构图

（conditionally independent hierarchical model, Kass and Steffey, 1989）。LDA 参数 α, β 可以由经验贝叶斯方法进行估计。

3.3 推断与参数估计

LDA 模型最终的目的是，给定文档，推断其潜在主题的后验分布：

$$p(\theta,\ z|\ w, \alpha, \beta) = \frac{p(\theta,\ z,\ w|\alpha, \beta)}{p(\ w|\alpha, \beta)}$$

具体过程可以采用拉普拉斯逼近、变分近似法、MCMC 方法计算。本文中我们采用 MCMC 方法，利用 gibbs 抽样近似计算后验分布。

4 数据处理与分析

4.1 数据处理

网站传送门转载了各大微信公众号的历史文章。我们从该网站上随机抽取三个关注人数较多的娱乐八卦公众号，抓取 2016 年 4 月中旬至 2017 年 2 月中旬的所有历史文章，及其阅读数与点赞数等信息，共计 1953 条记录作为我们的语料库。

4.2 LDA 提取文档主题

4.2.1 分词处理和词典构建

初步形成语料库之后，需要对语料进行分词处理和词典构建。

我们首先进行预分词处理，随机抽取 10% 的文档分词结果，筛选关键词语作为用户词典；从网上抓取明星名库，与中文停止词典合并作为新的停止词。如此，在最终分词中我们筛去了无意义的语气词和明星姓名，选取复现频率大于 50 的词语组成字典。

4.2.2 主题数 k 的确定

由于主题数 k 未确定，我们将 k 作为超参数，计算 k 从 5~45 共 40 个模型的五折交叉验证困惑度，这里的困惑度（perplexity）是 LDA 模型的误差指标，用来衡量模型在预测测试集样本的效果。

理想情况下主题数应选取使困惑度最小化的 k 或明显的拐点。但这里呈现较平滑的凸曲线，故此考虑选取 k = 20，k 20 时困惑度下降较快，随后平缓。

4.2.3 LDA 模型其它参数选择

模型初始参数 α 影响主题的集中度，α 越大最终所有文档倾向于集中在某几个主题；β 则影响词语的集中度，体现在 β 越大，每个主题更集中在几个词汇上面，或者而每个词汇都尽可能的百分百概率转移到一个主题上。我们选取 $\alpha = 0.1, \beta = 0.1$ 这是较为常用的选择。最终提取出的主题与主题关键词表 1 所示：

表 1: LDA 提取主题及相应关键词

主题	关键词	主题名
1	微博 -网友 -爆料 -粉丝 -事件	微博爆料
2	结婚 -恋情 -分手 -感情 -拍	恋情绯闻
3	孩子 -妈妈 -女儿 -爸爸 -儿子	家庭
4	孙杨 -比赛 -宁泽涛 -奥运会 -冠军	关注奥运
5	吃 -买 -生活 -爱 -喜欢	私人生活
6	穿 -时尚 -红毯 -衣服 -造型	造型
7	电影 -导演 -中国 -作品 -票房	电影票房
8	男神 -颜值 -表情 -爱 -脸	男神颜值
9	媒体 -香港 -离婚 -群众 -八卦	离婚八卦
10	粉丝 -直播 -韩国 -明星 -偶像	直播热度
11	女主 -男主 -剧 -剧情 -主角	电视剧话题
12	节目 -老师 -综艺 -嘉宾 -主持人	综艺话题
13	角色 -演员 -演 -演技 -戏	演技评论
14	脸 -照片 -拍 -颜值 -好看	颜值
15	相声 -春晚 -喜剧 -安吉 -徒弟	相声春晚
16	王思聪 -网红 -冯轲 -公众 -男人	网红
17	喜欢 -生活 -女人 -男人 -爱	感情生活
18	小主 -拍 -姑娘 -妹子 -吃	吃吃吃拍拍拍
19	音乐 -歌手 -唱 -演唱会 -专辑	演唱会
20	钱 -公司 -买 -投资 -老板	投资经济

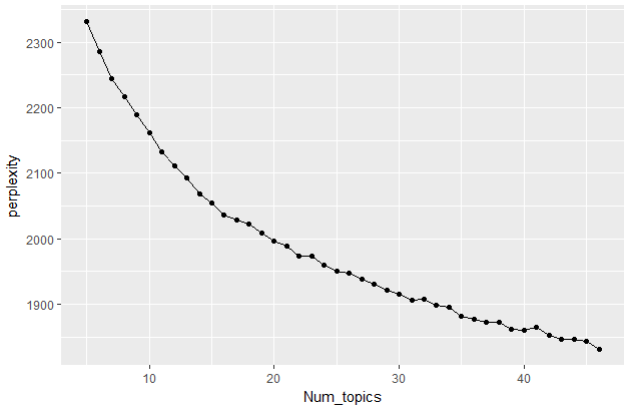

图 2: 困惑度 -标题数关系图

4.2.4 数据说明

根据 LDA 主题模型的运行结果，我们对所得的 1953 数据进行进一步处理，整理后的数据结构如表 2 所示：

4.3 描述性统计

4.3.1 公众号都在讨论哪些话题

通过对以上数据整理分析后，可以看出自 2016 年 1 月 22 日至 2017 年 2 月 18 日之间各个时间段，公众号对于各个话题均有不同数量的报导内容（如图 3 所示）。根据每条公众号所匹配的各个主题，其数量从多到少依次为话题 13、2、1、8、17、14、11、12、18、5、9、19、6、16、7、3、10、4、20、15（如图 4 所示）。可以看出，公众号最喜欢讨论的主题是演技（话题 13）、感情绯闻（话题 2）、微博话题爆料（话题 1）以及颜值（话题 8）。这四个话题的讨论度占据了总话题讨论度的 34%（如图 4 所示）。

4.3.2 公众号最青睐的明星

当从数据中提取公众号最频繁提到的明星时，发现无论是从上面最热的 4 个话题中提取（选取被提到次数大于 40 次的明星），还是从所有话题中提取（选取被提到次数大于 90 次的明星），最受公众号青睐的 10 个明星都分别为：范冰冰（FBB）、胡歌（HG）、黄晓明（HXM）、霍建华（HJH）、李易峰（LYF）、林心如（LXR）、文章（WZ）、杨幂（YM）、杨

表 2: 数据说明

变量名	取值范围	详细说明
topic	1-20	主题编号，定性变量
content	eg. 角色 -演员 -演 -演技 -戏	主题关键词，定性变量
starnumber	正整数	涉及明星数
starname	eg. 范冰冰	涉及明星
tme	2016.1.22-2017.2.18	发布时间
read	1-100000，10 万 +	阅读量
like	正整数	点赞数
original	0,1	是否原创，原创取 1
top10	0,1	是否包含 10 个最热明星
account	eg.shenyebagua818	公众号账号，定性变量

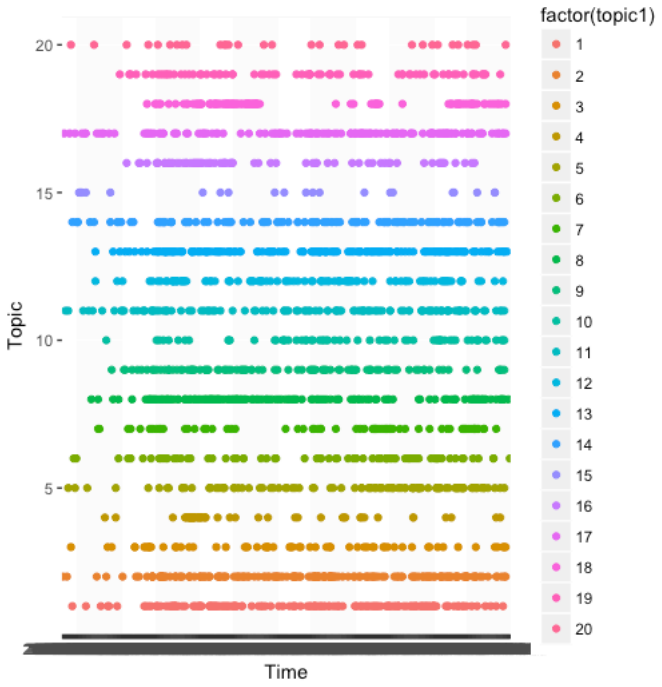

图 3: 时间 -话题关系图

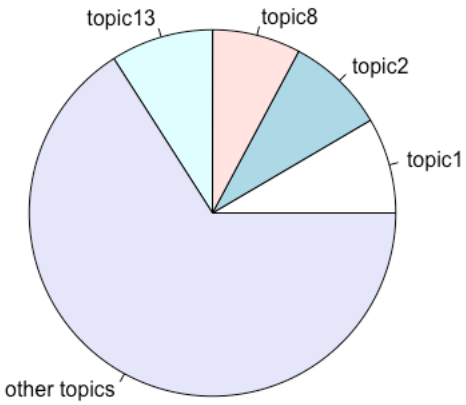

图 4: 话题 -文章占比

洋（YY）与赵丽颖（ZLY）（如图 5 与图 6 所示）。包括与其他明星一起被提到的话题，这十个明星就占据了总话题量的 68.56%。

4.3.3 最热的十位明星一般出现在哪些话题中

在案例所包括的时间内，娱乐圈出了许多爆炸新闻，例如：“霍建华和林心如公布恋情、大婚”等。因此，霍建华与林心如频繁出现感情绯闻类的话题中（话题 2）。图 9 显示了最热的十位明星最频繁出现在的话题。大致可以总结出，最经常出现在微博话题爆料（话题 1）的明星是文章、杨幂和刘亦菲，大约为 59.12%；占据感情绯闻话题（话题 2）大的明星是霍建华与林心如，大约为 41.76%；胡歌、霍建华和杨洋占据男神颜值榜话题（话题 8）的 42.5%；而频繁出现在演技话题（话题 13）中的明星是胡歌、刘亦菲和杨幂，大约为 42.55%。

4.4 方差分析

我们对于不同类别的数据进行多因素方差分析，希望观察公众推送的热度与不同的主题以及公众推送是否是原创是否存在明显差异，我们采用每篇公众文章的点赞数量作为对于推送热度的度量。自变量我们采用了三

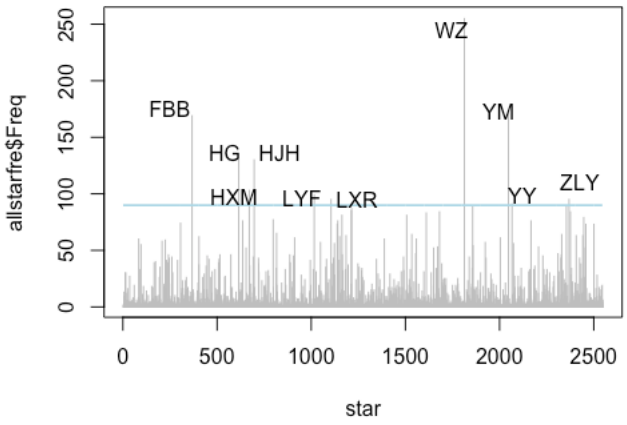

图 5: 明星被提及总次数

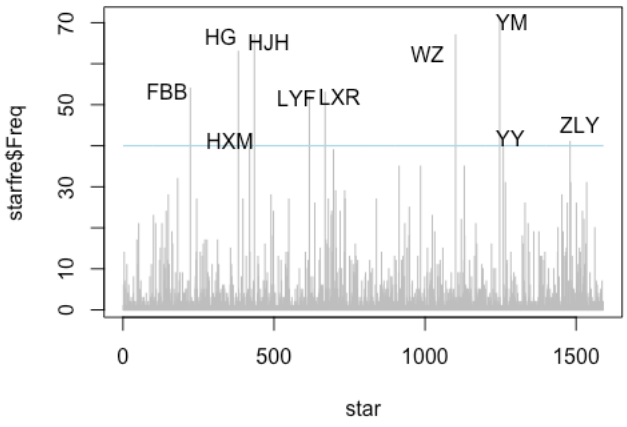

图 6: 热门话题明星被提及次数

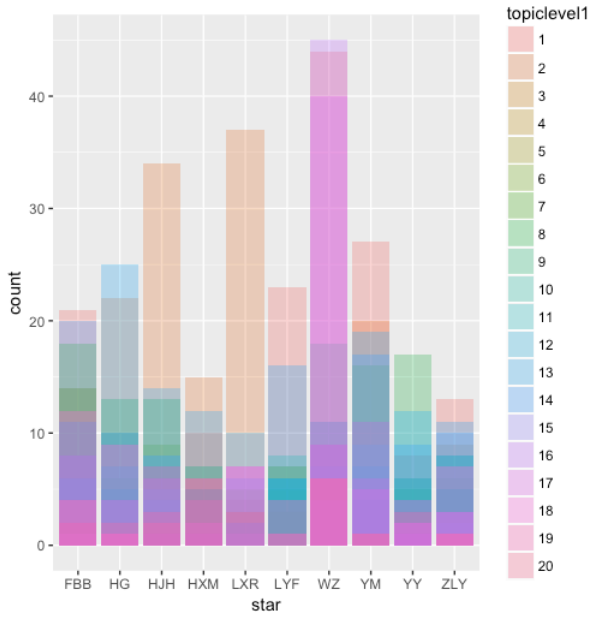

图 7: 明星 -话题数

种不同的因子型数据，分别为是否为原创、是否包含热门明星以及不同的主题。

4.4.1 数据预处理

我们首先观察对于是否为原创以及是否包含排名前 10 位的明星做分组的箱型线，通过下图我们发现原创以及不同的主题是具有明显区别的。

然后，我们观察不同主题箱型线，我们发现 16、17 两个类别中的异常值现象比较严重，观察这两个主题，我们发现这两类主题对应的主题词类别并不明显，因此我们选择删掉这两个主题。同时在异常值处理中我们删掉样本主题的样本数量小于 10 的样本。完成异常值处理。

4.4.2 假设检验

首先，我们需要观察数据是否符合方差分析的假设：方差齐性检验是指需要检验不同水平下的数据方差是否相等。我们采用 Bartlett 方法进行检验，得出这三种不同的分组下 P 值均大于 0.05，得出方差具有齐次性的结论。

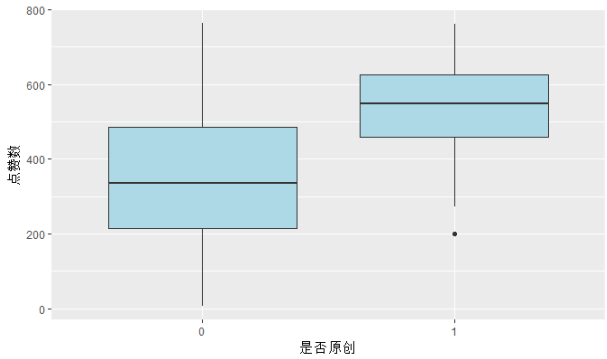

图 8: 是否原创 -热度关系图

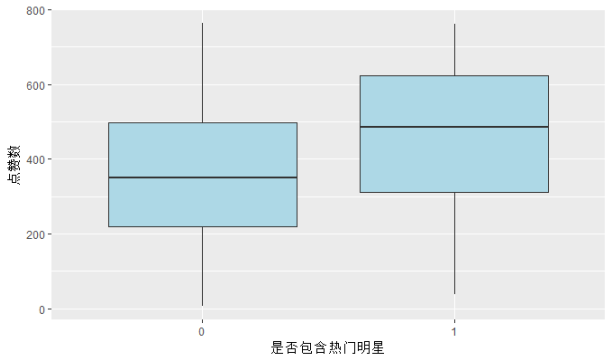

图 9: 是否包含热门明星 -热度关系图

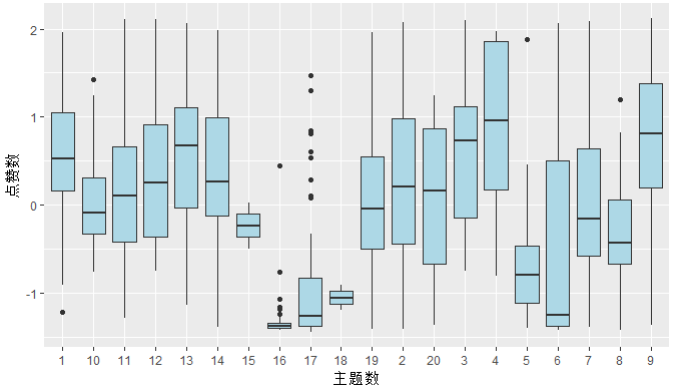

图 10: 异常值处理后的各主题热度

表 3: 方差分析运行结果

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
topic	14	3063176	218798	8.011	2.14e-15	***
top10stars	1	240616	240616	8.810	0.00315	**
original	1	1278350	1278350	46.807	2.46e-11	***
topic:top10stars	11	638890	58081	2.127	0.01742	*
topic:original	9	440093	48899	1.790	0.06771	.
top10stars:original	1	4277	4277	0.157	0.69250	
topic:top10stars:original	1	8337	8337	0.305	0.58087	
Residuals	468	12781681	27311			

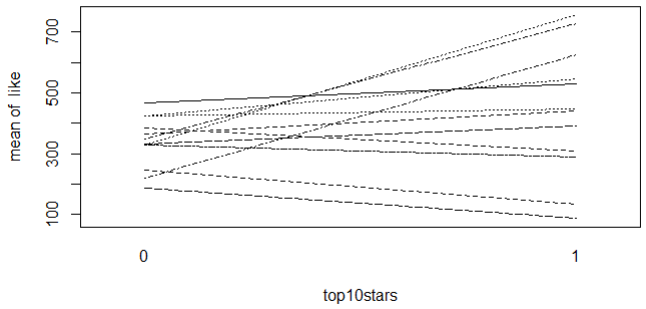

图 11: 热门明星与主题交叉项

4.4.3 三因素方差分析

我们对于不同的主题、是否为原创、以及是否包含最热门的 10 位明星这三个变量进行三因素方差分析，同时引入交叉项：

从结果中我们可以发现点赞数量与主题、原创性以及是否包含前 10 位明星这三个因子均有显著的相关关系，并且是否包含热门明星与主题，以及主题与原创性对结果也具有交叉效应。

因此，我们做出交叉效应图，可以观察到，是否包含热门明星与不同的主题确实具有明显的交叉效应。

而对于是否包含热门明星与是否为原创这两个因子交叉效应不显著，如下图示：

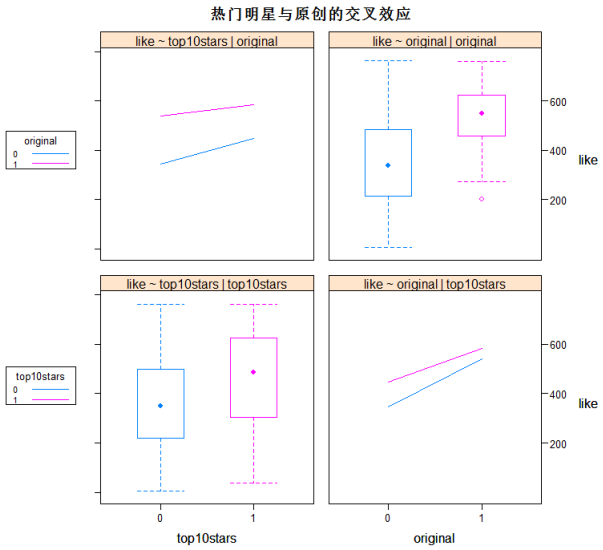

图 12: 热门明星与原创的交叉效应

从而我们得出结论：明星公众号的点赞数量与原创性及明星热度均有明显关系，原创的推送以及涉及热门明星，以及部门热门主题均能提高微信推文的热度。

5 总结及改进

5.1 总结

通过 LDA 文本聚类，我们发现娱乐公众号的话题主要可归为 20 类，涉及微博爆料、恋情绯闻、网红等等。以范冰冰、胡歌、霍建华为代表的十位明星最常成为娱乐公众号关注的焦点。进一步的，主题 -明星组合（例如：颜值 -胡歌）成为各大公众号的老生常谈。进行文本聚类后，我们对不同主题的公众号文章热度进行探究。原创性和热门明星成为点击量的保证，而不同主题之间也表现出明显的差异。主题类别与涉及明星之间存在交叉效应，热门主题 + 热门明星的组合带来了阅读量、点赞量的显著提升，最易获得订阅读者支持。 以上分析我们可以得出，成功经营公众号的软广业务，需要娱乐公众号紧追微博爆料、家庭、奥运、离婚八卦、演技评论等保证流量的话题，尤其是几位热门明星的相关娱乐新闻。除此之外，原创性也是公众号持续获得关注的重要因素。

5.2 改进

在此次微信公众号的案例分析中，尽管最终的结果符合能够观察到的真实现象，然而在以后的研究中可以再进行改进。首先，通过 LDA 模型所计算出的 20 个主题，有部分会重叠，可以对其进行再处理，使所有主题能够尽可能独立。其次，LDA 模型能够将一篇文档归类到多个主题中，但本次研究为了简单起见只将每篇文档归到概率最高的一个主题下，这样我们就损失了一部分文档归类信息，在以后的研究中，可以对原始公众号的内容再进行细化，以获得更加准确的模型和结果。

6 参考文献

[1]Tan A H. Text mining: The state of the art and the challenges[C]//Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases. 1999, 8: 65-70.

[2]Blei, D., Ng, A., and Jordan,M. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993–1022.

[3] 袁军鹏, 朱东华, 李毅, 等. 文本挖掘技术研究进展 [J]. 计算机应用研究, 2006, 2: 1-4.

[4] 郭金龙, 许鑫, 陆宇杰. 人文社会科学研究中文本挖掘技术应用进展 [J]. 图书情报工作, 2012, 56(8): 10-17.

[5] 董婧灵, 李芳, 何婷婷. 基于 LDA 模型的文本聚类研究 [J]. 孙茂松, 陈群秀. 中国计算语言学研究前沿进展 (2009r 2011). 北京: 清华大学出版社, 2011.

[6] 唐晓波, 向坤. 基于 LDA 模型和微博热度的热点挖掘 [J]. 图书情报工作, 2014, 58(5): 58-63.