关于 GSDMM 的数学思考

GSDMM 是一种基于狄利克雷多项式混合模型的收缩型吉布斯采样算法（a collapsed Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture model）的简称，它是发表在 2014 年 KDD（数据挖掘及知识发现会议，ACM SIGKDD，数据挖掘顶级会议[1]）上的论文《A Dirichlet Multinomial Mixture Model-based Approach for Short Text Clustering》的数学模型[2]。

GSDMM 主要用于短文本聚类，短文本聚类是将大量的短文本（例如微博、评论等）根据计算某种相似度进行聚集，最终划分到几个类中的过程。GSDMM 主要具备以下优点[3]：

1. 可以在完备性和一致性之间保持平衡；
2. 可以很好的处理稀疏、高纬度的短文本；
3. 较其它的聚类算法，在性能上表现更为突出。

第 1 条优点的完备性体现在所有参入计算的短文本最终都能被聚集到某一个具体的族簇中，而一致性体现在被聚集到同一个族簇的所有短文本都具备较为强的相似性，即这些短文本在某种程度上都是跟同一事物有关的微博或者评论（如果数据集采用的是微博或者评论的文本数据）。由于短文本的特点（文本篇幅短而且用词重复率非常低）以及最终所采用的数据集能够得到一个很好的结果，故而第 2 条优点能够很好的被证明。第 3 条优点的依据如图 1 和表 1。

图 1 中的横、纵坐标分别表示评价度量方式和性能表现（基于数据可视化的考虑，该性能表现基于不同的评价度量方式进行了归一化处理），其中 NMI（Normalized Mutual Information）表示归一化互信息指数，H（Homogeneity）表示一致性指数，C（Completeness）表示完备性指数，ARI（Adjusted Rand Index）表示调整的兰德指数，AMI（Adjusted Mutual Information）调整的互信息指数。K-means 是指 K 均值聚类算法，是一种在数据挖掘与分析领域非常流行的矢量量化方法[4]；HAC 是一种层次聚类分析方法[5]；DMAFP 是一种具备去噪能力的长文本聚类方法[6]。

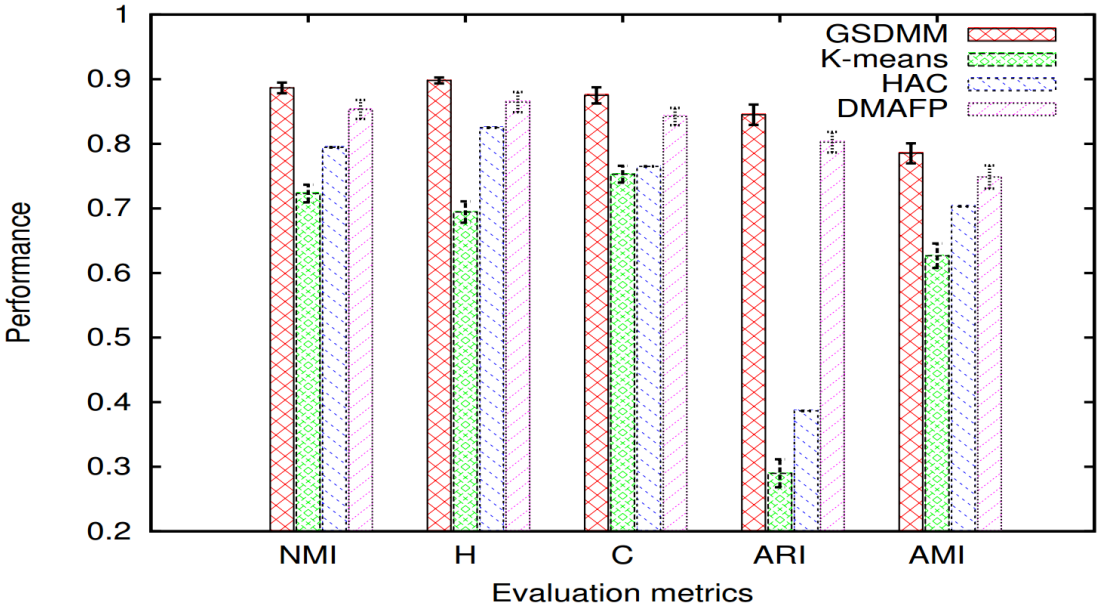

图 1 GSDMM 和其它三种聚类算法的在 TweetSet 数据集上的性能表现

表 1 GSDMM 和其它两种聚类算法的在三个数据集上的性能表现

数据集	指标	GSDMM	K-means	DMAFP
TSet	NMI	0.874±0.007	0.732±0.007	0.852±0.009
	H	0.853±0.010	0.692±0.009	0.831±0.010
	C	0.896±0.006	0.775±0.006	0.875±0.007
	ARI	0.693±0.043	0.133±0.030	0.657±0.051
	AMI	0.831±0.012	0.639±0.011	0.814±0.015
SSet	NMI	0.896±0.006	0.759±0.008	0.868±0.008
	H	0.871±0.008	0.754±0.009	0.846±0.011
	C	0.921±0.005	0.764±0.009	0.892±0.007
	ARI	0.746±0.014	0.262±0.017	0.703±0.018
	AMI	0.853±0.009	0.708±0.008	0.819±0.012
TSSet	NMI	0.928±0.004	0.834±0.005	0.901±0.008
	H	0.911±0.005	0.836±0.005	0.889±0.006
	C	0.945±0.003	0.832±0.005	0.912±0.004
	ARI	0.789±0.018	0.370±0.029	0.736±0.023
	AMI	0.897±0.006	0.800±0.006	0.847±0.009

GSDMM 采用类比的方法——通过电影分组过程（Movie Group Process，MGP）模拟 GSDMM 的聚类过程，通俗易懂地阐明了 GSDMM 聚类的全过程。MGP 的类比短文本聚类的内容如表 2，短文本聚类问题可以看作通过每个学生看过的电影清单将学生分组的问题，自然的每一组的学生看的电影是类似的，即同一组的学生的电影清单是类似的，而不同组的学生的电影清单差异性是极大的。

表 2 电影分组过程类比短文本聚类的内容

MGP	短文本聚类
所有学生	数据集、语料库
每个学生、每个电影清单	每篇文档
学生看过的电影、电影清单上的电影	文档中的单词

电影分组过程（MGP）如下：

1. 预定义 K 个组，将学生随机分配到这 K 个组中
2. 针对每一个学生，根据以下准则重新分配分组：
   a) 选择学生更多的小组
   b) 选择电影清单更相似的小组
3. 将第 2 步反复进行，直至保留下的组趋于稳定

GSDMM 的第 1 条优点的完备性和一致性分别在准则 a 和准则 b 上得到体现，准则 a 让族簇的完备性更强，即让同一个小组尽可能多的包含属于该小组的学生，而准则 b 让族簇的一致性更强，即让有着同样电影清单的学生尽可能的在一个小组中。

GSDMM 通过下面的条件概率进行每个学生的所属的小组的重新分配：

$$p(z_d = z|\vec{z}_{\neg d}, \vec{d}) \propto \frac{m_{z,\neg d} + \alpha}{D - 1 + K\alpha} \frac{\prod_{w \in d} \prod_{j=1}^{N_d^w} (n_{z,\neg d}^w + \beta + j - 1)}{\prod_{i=1}^{N_d} (n_{z,\neg d} + V\beta + i - 1)}$$

上面的条件概率公式中橙色虚线框（左边虚线框）中的部分对应准则 a，蓝色虚线框（右边虚线框）中的部分对应准则 b。公式中的符号说明见表 3。

表 3 条件概率中的符号说明

符号	说明
z_d	文档所属的族簇
z	某一个族簇
$\vec{z}_{\neg d}$	除文档 d 所属族簇外的所有族簇
$\vec{d}$	所有文档
$m_{z,\neg d}$	不包含文档 d 的族簇 z 中的文档数
α	参数 Alpha
D	数据集中的所有文档数
K	参数 K
w	某一个单词
N_d^w	文档 d 中单词 w 的出现次数
$n_{z,\neg d}^w$	不包含文档 d 的族簇 z 中单词 w 的出现次数
N_d	文档 d 的单词数
$n_{z,\neg d}$	不包含文档 d 的族簇 z 中单词数
V	数据集的所有不重复单词数
β	参数 Beta

电影分组过程（实际上是 GSDMM 算法）存在四个参数（除了表 3 中说明的三个参数，还有一个是电影分组过程的第 3 步的隐含迭代次数），这四个参数对于模型的好坏有较大影响。参数 K 对于聚类族簇数量的影响见图 2，对于数据集 TweetSet，由图可知初始族簇大小 K 值趋于 300 左右时，GSDMM 的聚类效果基本与实际相符。参数 Alpha 对于聚类族簇数量的影响见图 3，对于数据集 TweetSet，由图可知参数 Alpha 等于 0.1 时，GSDMM 的聚类效果基本与实际相符。参数 Beta 对于聚类族簇数量的影响见图 4，对于数据集 TweetSet，由图可知参数 Beta 等于 0.08 时，GSDMM 的聚类效果基本与实际相符。迭代次数对于聚类族簇数量的影响见图 5，对于数据集 TweetSet，由图可知迭代次数为 20 次时，GSDMM 的聚类结果趋于平稳且效果基本与实际相符。

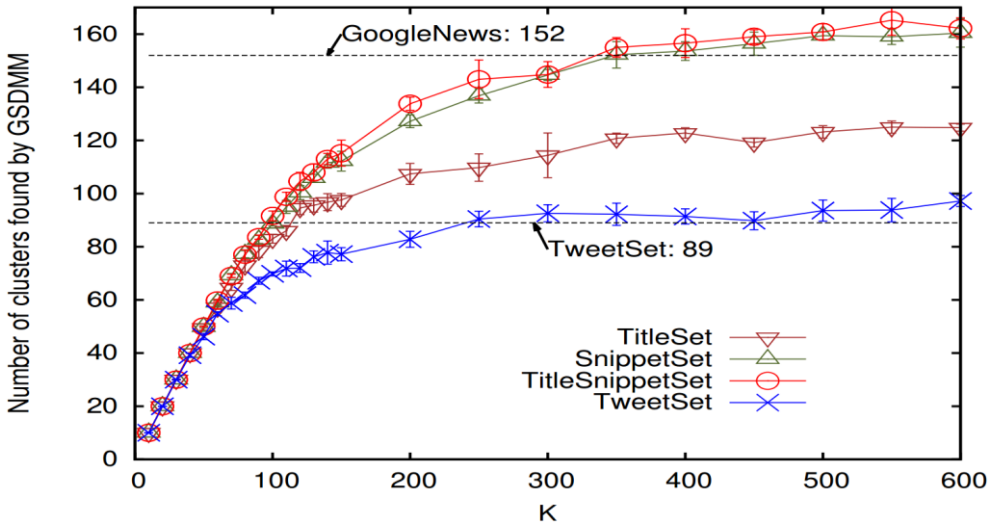

图 2 参数 K 对聚类族簇数量的影响

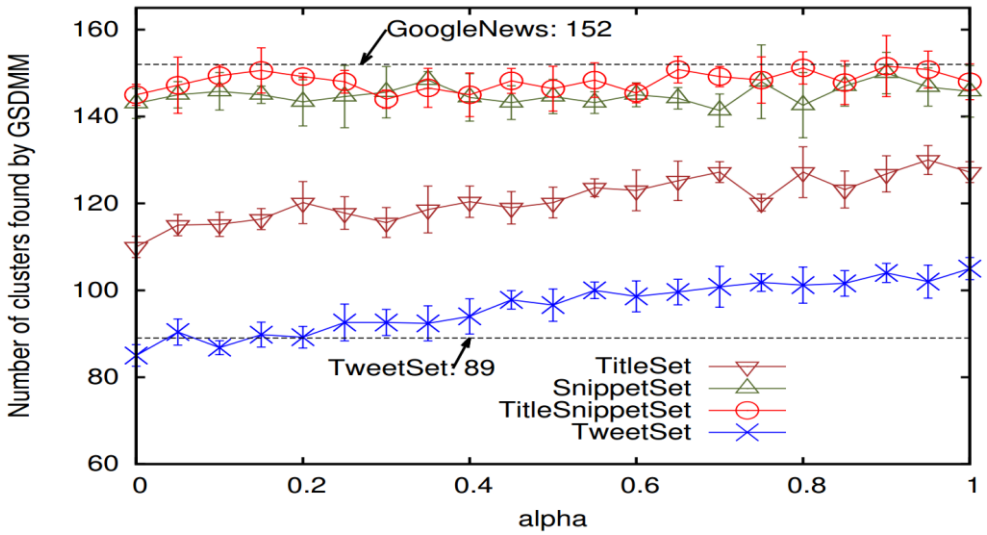

图 3 参数 Alpha 对聚类族簇数量的影响

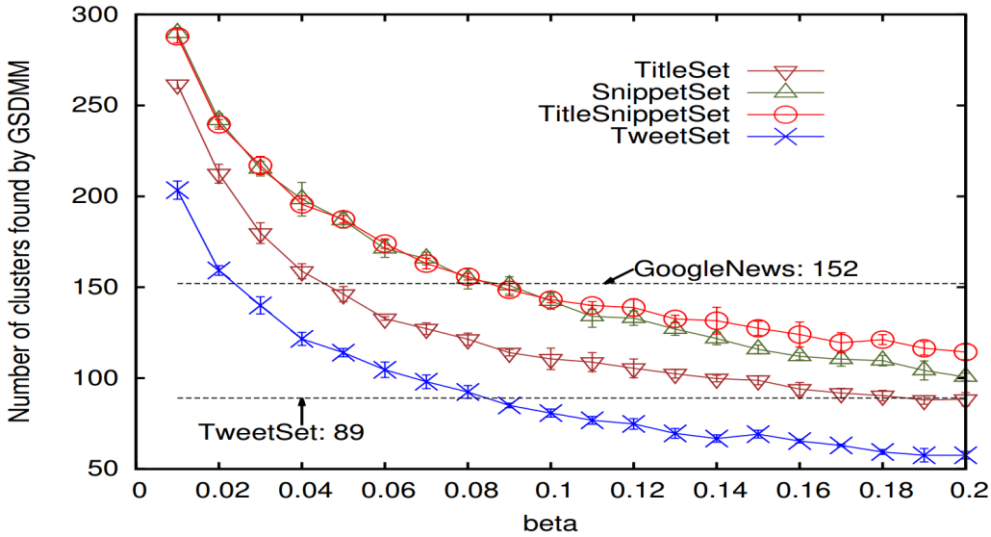

图 4 参数 Beta 对聚类族簇数量的影响

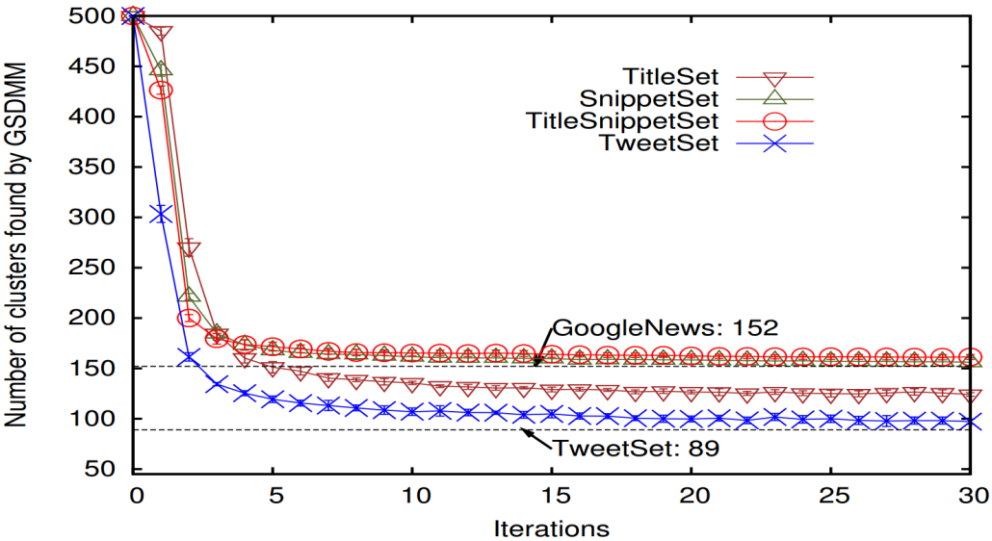

图 5 迭代次数对聚类族簇数量的影响

上述 GSDMM 的四个参数为经验参数，对于不同的数据集（各个数据集差异较大）最佳的参数取值也会不同。在实际应用中，当给定较好的经验参数，GSDMM 具备较好的聚

类效果，这使得它具备较高的应用价值。

参考文献

[1] ACM SIGKDD ——http://www.kdd.org/

[2] 《A Dirichlet Multinomial Mixture Model-based Approach for Short Text Clustering 》——http://dl.acm.org/citation.cfm?id=2623715

[3] 《KDD 2014 “A Dirichlet Multinomial Mixture Model-based Approach for Short Text Clustering”的主要思想》——http://blog.csdn.net/yueliangku/article/details/42737965

[4] K-means Clustering——https://en.wikipedia.org/wiki/K-means_clustering

[5] HAC——https://en.wikipedia.org/wiki/Hierarchical_clustering

[6] 《Dirichlet Process Mixture Model for Document Clustering with Feature Partition》——http://ieeexplore.ieee.org/document/6152106/