“智慧政务”中的文本挖掘应用分析

摘要

随着物联网技术的普及，大数据、 云计算、人工智能等技术的发展，从大量的政务文本数据中挖掘出有用信息，建立基于自然语言处理技术的智慧政务系统已经是社会治理创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用

针对问题一，建立了 GCN 图卷积网络模型进行根据内容分类三级标签体系对留言分类，基于图卷积网络模型的处理的标签分类，并且对比于分类前的标签留言，建立了较为清晰的关于留言内容的一级标签分类模型。随后对通常使用 F-Score 对分类方法进行评价。

针对问题二，本文利用了关键词分布距离法以及关键词共现分析法建立了相关的数据对热点问题进行了综合的分析，并且将给出的数据中的某一时段内反映特定地点或特定人群问题的留言进行归类，定义合理的热度评价指标。随后根据构建的指标计算得出了排名前 5 的热点问题和热点明细表。

针对问题三，选取热点问题中的时间范围，用户数，热度指数等指标，基于 AuthorValue 模型给出一套评价方案。分析发现了政府在对留言的答复中的相关性和可解释性有明显的关联，基于任务数据未能发现留言答复中的完整性的显著关系。

关键词：GCN 图卷积网络模型, 关键词, 关键词分布距离法, 关键词共现分析法, AuthorValue 模型

Abstract

With the popularization of Internet of things technology, the development of big data, cloud computing, artificial intelligence and other technologies, mining useful information from a large number of government text data, establishing intelligent government system based on natural language processing technology has become a new trend of social governance innovation and development, which has a great role in promoting the management level and efficiency of the government

Aiming at problem one, GCN graph convolution network model is established to classify messages according to three-level label system of content classification, and label classification based on graph convolution network model is established. Compared with label messages before classification, a more clear one-level label classification model about message content is established. Then, we evaluate the classification method with F-score.

In view of the second problem, this paper uses the key words distribution distance method and the key words co-occurrence analysis method to establish the relevant data to carry on the comprehensive analysis to the hot spot problem, and classifies the message which reflects the specific place or the specific crowd problem in a certain period of time in the given data, defines the reasonable heat evaluation index. Then according to the index calculation, the top 5 hot issues and hot list are obtained.

Aiming at the third problem, a set of evaluation scheme is given based on the authorvalue model by selecting time range, number of users, heat index and other indicators. It is found that there is a significant correlation between the relevance and the explainability of the government's response to the message. Based on the task data, it is not found that there is a significant relationship between the integrity of the response to the message.

Keywords: GCN convolution network model, Keywords, keywords distribution distance method, keywords co-occurrence analysis method, authorvalue model

目　录

一、问题分析

近年来，随着微信、微博、市长信箱、阳光热线等网络问政平台逐步成为政府了解民意、 汇聚民智、凝聚民气的重要渠道，各类社情民意相关的文本数据量不断攀升，给以往主要依 靠人工来进行留言划分和热点整理的相关部门的工作带来了极大挑战。同时，随着大数据、 云计算、人工智能等技术的发展，建立基于自然语言处理技术的智慧政务系统已经是社会治 理创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用。附件给出了收集自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见。请利用自然语言处理和文本挖掘的方法解决下面的问题。

问题给出四个附件数据，附件一提供的内容分类三级标签体系，附件二提供 9211 条的政务留言信息，其中包括用户编号，留言用户，留言时间，留言详情等 6 个指标，附件三校对于附件二采集多了点赞数和反对数等 8 个指标，附件四收集了政府部门对于政务留言的答复。

问题一是参考附件 1 和附件 2，建立关于留言内容的以及标签分类模型。在建立模型的过程中，要把无关的指标筛选以便影响数据模型的准确性。

问题二是参考 C 题中的热点问题挖掘以及根据附件 3 的重要指标来给出排名前 5 的热点问题和响应的热点问题对应的留言信息。根据附件 3 的相关指标利用关键词分布距离法和关键词共现分析法能整理分析出排名前 5 的热点问题和热点问题对应的留言信息。

问题三则是针对附件 4 相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对 答复意见的质量给出一套评价方案。

二、数据准备

首先根据附件 1 的提供的内容分类三级标签体系对留言分类，然后剔除附件 2 里的无关指标：用户编号，留言用户。

2.1 准备分析需要的指标

基于分析需求准备的指标：

（1）留言时间

留言时间的数据的收集是集中反映了某个峰段政务的留言信息的爆发，通过对留言时间的分析和处理，能更好促进智慧政务的发展和提高政府对政务信息处理能力。

（2）留言主题

留言主题的展现出某某政务的时常发生的地点以及问题所在，对留言主题的关键词进行提取，能够分析某种政务或者政务具体存在不足的地方。

（3）留言人数

留言人数的统计以及分布，系统的体现了政务数量的多与少，更能依据留言人数反映出某年某月的政务处理的结果。

（4）点赞数，反对数

点赞数，反对数的展出较为突出的表现了政府部门对于政务留言处理解决的效率以及评价。

三、问题一

3.1 GCN 图卷积网络模型文本分类的介绍

文本分类任务是 NLP 领域一个重要和经典的问题，但是却鲜有使用 GCN 来建模。本文提出了使用 Text GCN 来进行文本分类任务，基于词共现和关键词关系为一个语料建立一个单独的文本图，然后为语料学习 Text GCN 模型。该模型通过 One-hot 表示为词和关键词初始化，然后联合学习词和关键词的 Embedding。实验结果表明，在没有任何外部的词 Embedding 和知识的情况下，本文提出的 Text GCN 模型在多个关键词分类 benchmark 数据集上取得了 SOAT 的效果。另一方面，Text GCNT 也在学习预测词和关键词的 Embedding，在训练数据量更少的情况下，Text GCN 在关键词分类任务上比 SOAT 的优势更明显，表现出了很好的准确性。从整个语料来构造一个大图，使用词和关键词作为图的节点。然后用 GCN 对图进行建模，该模型可以捕获高阶的邻居节点的信息，两个词节点之间的边通过词共现信息来构建，词节点和关键词节点之间的边通过词频和关键词频率来构建，进而关键词分类问题就转化成了节点的分类问题。这种方法通过小部分的带标签文档可以学习强健的类别信息，学习词和文档节点之间的交互 Embedding 信息。具体措施：1.提出了一个新颖的文本分类方法 Text GCN，这是第一个采用全部的语料作为异构图的研究，使用图神经网络联合学习词和关键词的 Embedding 表示。

2.在没有使用预训练的 Word Embedding 和外部知识的情况下，在几个关键词分类 benchmark 数据集上胜过 SOAT 方法，该模型也同时在学习预测词和关键词的 Embedding 表示。

3.2 对数据文本分析及分类

本文采用了 3 个数据集：留言时间、留言人数、留言地点；进行清洗数据、分词、去停用词和去除词频小于 3 的词，预处理之后的数据集各项数据统计如下表所示：

	最多留言时间段	留言人数	留言地点
2010年	12月	4条	A8县市花明楼镇
2011年	11月	69条	西地省A7区
2012年	7月	84条	西地省M5区
2013年	3月	68条	西地省A8区
2014年	8月	111条	西地省A7区
2015年	8月	99条	西地省A7区
2016年	3月	138条	西地省A7区
2017年	7月	153条	西地省K1区
2018年	12月	188条	西地省L5区
2019年	7月	179条	西地省G5区
2020年	1月	43条	西地省G5、K2、L7、B4、J10区

3.3 一级标签分类模型的建立

GCN 是一个直接在图上操作的多层神经网络，基于节点的相邻节点的属性信息引入节点的 Embedding。对于一个一层的 GCN，k-dim 的节点特征矩阵

计算公式如下：

$$L^{(1)} = \rho(\tilde{A}XW_0)$$

通过堆叠多个层来合并高阶的临近节点的信息：

$$L^{(j+1)} = \rho(\tilde{A}L^{(j)}W_j)$$

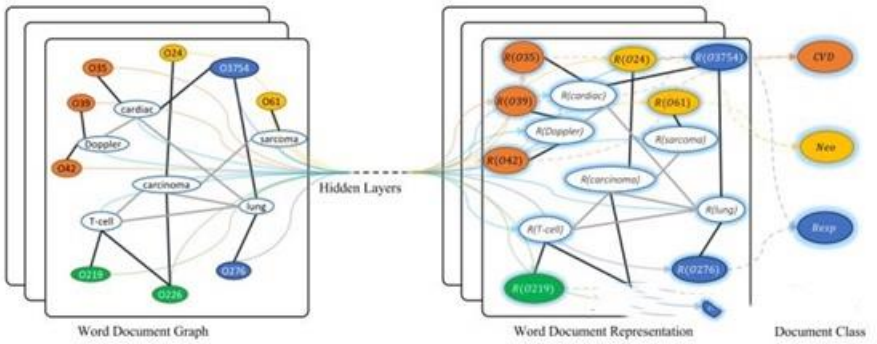
本文采用包含词节点和关键词节点的大型异构文本图，这样一来全局的词共现可以被明确的建模，图卷积可以被容易的使用。如下图所示，节点的数量是所有关键词的数量加上语料词典中所有词的数量。其中 X 单位矩阵，表示每一个词或者文档采用 One-hot 编码作为输入，边“文档-词”基于附件中的关键词出现，边“词-词”基于整个语料库的词共现，权重“文档-词”基于 TF-IDF，两个词之间的权重采用 PMI 来计算，公式如下：

$$A_{ij} = \begin{cases} \text{PMI}(i,j) & i,j \text{ are words, PMI}(i,j) > 0 \\ \text{TF-IDF}_{ij} & i \text{ is document, } j \text{ is word} \\ 1 & i = j \\ 0 & \text{otherwise} \end{cases}$$

损失函数计算如下：

$$\mathcal{L} = -\sum_{d \in \mathcal{Y}_D} \sum_{f=1}^{F} Y_{df} \ln Z_{df}$$

本文采用 2 曾的 GCN 进行训练，最大计算两阶临近节点的信息。模型的结构如下图所示：



结论

本文在 5 个被广泛使用的 benchmark 上分别进行了 10 次实验，然后得出本文提出了一个新颖的文本分类模型 Text GCN，在整个语料上构建了异构的词和关键词图，将关键词的分类问题转化为节点的分类问题，该模型可以捕获全局的词共现信息并有效的利用有限的关键词标注信息，一个简单的两层 GCN 模型就在多个 benchmark 取得了 SOAT 的效果。将来的改进方向可以放在为图增加 attention 机制或者开发无监督的 GCN 框架在大规模无标注的关键词语料上进行表示学习

四、问题二

4.1 关键词分布距离法

关键词分布距离法的主要的方法是利用计算不同的留言时间段热点问题中关键词以及关键词分布之间的相对距离来挖掘文本留言的热点的变化趋势，关键词分布中间的距离一般来说是用相对熵来量度[1]。现在我们不妨假设留言时间段 t_1 和 t_2, K 为关键词（留言主题，留言人数，点赞数，反对数）集合，留言时间段 t_1 的分布为 P（x），留言时间段 t_2 的分布为 Q（x），其中 $x \in K$，正常的情况下我们都使用相对熵来量度两个关键词热点之间的关联，计算公式如下列公式所示：

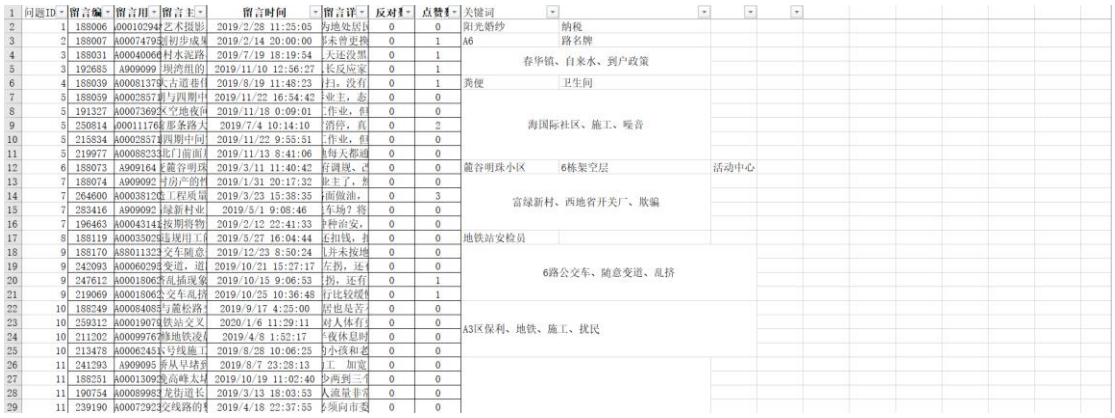
$$Dis(P \mid Q) = \sum_{x \in K} P(x) log[P(x) \mid Q(x)]$$

4.2 关键词共现分析法

关键词共现分析法是由法国的科学家 M.Callon、J.Law 和 A.Rip 提出的，它主要是源于一个假设：附件里的主要内容是用关键词来表述，倘若多个不同的关键词共同出现在同一个附件中，那么则认为这多个关键词它们之间有一定的某种特殊的联系[1]。利用计算附件中所有关键词之间两两共现的频次以及出现次数，便得到了关键词的之间的共现矩阵，

再由聚类分析将相互联系的多个关键词合聚为一类，得到研究主题的关键词簇。最后对每一个关键词簇来进行分析研究，就可以发现该附件里面关键词的研究点的密度和热点，还可以根据关键词簇来对本附件的知识结构进行构建，让用户对该热点问题有一个清晰的认识。

热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
1	34		2019/07/02至2019/09/01	A市武广新城广铁集团	A市伊景园滨河苑车位捆绑销售
2	35		2019/6/25至2020/01/26	A市丽发新城小区	小区附近建搅拌站噪音扰民、环境污染
3	226		2019/01/01至2019/07/08	西地省A市A4区	58车贷案件进展情况
4	1347		2019/05/05至2019/9/19	A5区五矿万境K9县	交房后仍存在诸多问题
5	156		2019/01/06至2019/05/22	A1区辉煌国际城二期商铺	辉煌国际城二期物业提供虚假场地证明，居民楼下商铺非法取证



问题ID	留言编	留言用	留言主	留言时间	留言详	反对数	点赞数	关键词		
1	188006	A00010294	艺术摄影	2019/2/28 11:25:05	为地处居民	0	0	阳光碧纱	纳税	
2	188007	A00074795	至初步成果	2019/2/14 20:00:00	未曾更换	0	1	A6	路名牌	
3	188031	A00040066	村水泥路	2019/7/19 18:19:54	天还没黑	0	1	春华镇、自来水、到户政策		
3	192685	A909099	现浩组的	2019/11/10 12:56:27	长反应家	0	1			
4	188039	A00081378	古道巷住	2019/8/19 11:48:23	扫，没有	0	1	美便	卫生间	
5	188059	A00028574	期与四期中	2019/11/22 16:54:42	业主，去	0	0			
5	191327	A00073692	空地夜间	2019/11/18 0:09:01	作业，但	0	0	海国际社区、施工、噪音		
5	250814	A00011176	那条路大	2019/7/4 10:14:10	消停，真	0	2			
5	215834	A00028574	四期中间	2019/11/22 9:55:51	作业，但	0	0			
5	219977	A00088233	北门前面	2019/11/13 8:41:06	每天都通	0	0			
6	188073	A909164	麓谷明珠	2019/3/11 11:40:42	有调规、占	0	0	麓谷明珠小区	6栋架空层	活动中心
7	188074	A909092	对房产的性	2019/1/31 20:17:32	业主了，基	0	0			
7	264600	A00038126	工程质量	2019/3/23 15:38:35	面做油，	0	3	富绿新村、西地省开关厂、欺骗		
7	283416	A909092	绿新村业	2019/5/1 9:08:46	车场？将	0	0			
7	196463	A00043141	按期将物	2019/2/12 22:41:33	种治安，	0	0			
8	188119	A00035029	违规用工	2019/5/27 16:04:44	扣钱，并	0	0	地铁站安检员		
9	188170	A88011323	交车随意	2019/12/23 8:50:24	并未按地	0	0			
9	242093	A00060293	变道，道	2019/10/21 15:27:17	左拐，还	0	0	6路公交车、随意变道、乱挤		
9	247612	A00018067	乱插现象	2019/10/15 9:06:53	拐，还有	0	1			
9	219069	A00018067	交车乱挤	2019/10/25 10:36:48	行比较缓	0	1			
10	188249	A00084085	寺麓松路	2019/9/17 4:25:00	居也是苦	0	0			
10	259312	A00019078	铁站交叉	2020/1/6 11:29:11	对人体有	0	0	A3区保利、地铁、施工、扰民		
10	211202	A00099767	地铁凌	2019/4/8 1:52:17	夜休息时	0	0			
10	213478	A00062457	号线施工	2019/8/28 10:06:25	小孩和老	0	0			
11	241293	A909095	开从早堵到	2019/8/7 23:28:13	工 加宽	0	0			
11	188251	A00013092	高峰太堵	2019/10/19 11:02:40	两到三个	0	0			
11	190754	A00089983	龙街道长	2019/3/13 18:03:53	流量非常	0	0			
11	239190	A00072923	线路的	2019/4/18 22:37:55	须向市委	0	0			

五、问题三

5.1 选取热点问题中指标

首先根据附件 4 中的关键词找出选取热点问题的指标，从附件 4 中得到留言时间，留言的人数，热度指数，以及留言的答复的 4 项。根据时间和人数等可以反映出相关部门对待政务的处理能力，依照热度指数和留言答复可以得出答复的相关性，完整性和可解释性的关联。注：热点指数=（问题数）*50%+（点赞数-反对数）*50%

如果在相关部门的答复中，有许多同样的问题需要解决和处理方式相同的情况下，一般情况下大家会认为此这个答复在这个方案上是有实际意义和价值的。基于这个前提，本文定义了从答复的相关性、完整性、可解释性等角度对 答复意见的质量给出一套评价方案：

$$AuthorValue(p_i) = \frac{\sum_{j=i}^{k} d_{ij}}{max_i^n \sum_{j=1}^{k} d_{ij}}$$

其中，n 表示搜索到的相关部门答复的相关性，k 表示相关部门答复的可解释性，dij 表示第 i 个时间段的第 j 个用户的答复数。然后将第 i 时间段中的所有答复在搜索到的 n 个相关部门中的答复都加起来，这样对每一个相关部门的答复都进行上述的计算后，取一个最大的作为分母。这样做的主要目的是为了方便计算，对公式进行了缩放操作，将结果区间又[0,∞]缩放到[0,1]【2】。从结果中得到的[0,1]的区间范围内，留言答复的评价方案的相关性，完整性和可解释性有着明显的关联。

参考文献

【1】刘金花.科技文献智能挖掘若干技术研究【A】.东北大学信息科学与工程学院，2013，【1】： 9.

【2】梁昌明，李冬强.基于新浪热门平台的微博热度评价指标体系实证研究【J】.山东师范大学历史与社会发展学院，2015，【2】：10.