题　　目 基于群众反映中热点问题的自动发现与分析

摘　要:

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。在此背景下，本文提出了对群众留言信息进行数据挖掘的课题，建立一种投诉信息热点问题自动发现与分析的系统模型，以解决目前存在的投诉业务量急剧增长与分析处理效率低下之间的矛盾。

首先，介绍了热点问题发现的研究背景发展现状，分析了当前的热点发现系统存在的问题，阐述了本系统的设计原理及工作流程。

其次，通过对投诉信息文本及热点问题特点的分析，明确了系统需求，设计了系统基础架构，提出了一种优化的 K-means 算法，实现了热点问题的挖掘和分析。

关 键 词：热点问题、数据仓库、数据挖掘、投诉信息

Abstract

In recent years, with Hatcheck, such as micro blog, mayor mailbox, sun hotline network asked Zhengzhou stage gradually become the government understand the importance of public opinion, gathering intelligence, condensed bull channels, all kinds of public opinion related text data quantity rising, leave a message to past mainly rely on artificial to divide and hot spots of relevant departments work has brought great challenge. At the same time, with the development of big data, cloud computing, artificial intelligence and other technologies, the establishment of intelligent government affairs system based on natural language processing technology has become a new trend of social governance innovation and development, which plays a great role in promoting the management level and efficiency of the government. Under this background, this paper puts forward the subject of data mining for the message information of the masses, and establishes a system model of automatic discovery and analysis of hot issues of complaint information, so as to solve the contradiction between the rapid growth of complaint business and the low efficiency of analysis and treatment.
Firstly, this paper introduces the research background and development status of hot spot problem discovery, analyzes the problems existing in the current hot spot discovery system, and expounds the design principle and work flow of the system.
Secondly, by analyzing the text of complaint information and the characteristics of hot issues, the system requirements are clarified, the system infrastructure is designed, and an optimized k-means algorithm is proposed to realize the mining and analysis of hot issues.

Keywords: hot issue、data warehouse、data mining、complaint information

目　　录

一、问题重述

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。

根据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。利用自然语言处理和文本挖掘的方法解决问题。

二、问题分析

系统首先要解决的问题是数据的采集、存储和预处理问题。由于用户数量巨大，每天会产生大量的投诉记录，并且这些投诉数据需要储存一段时间以便深入分析和挖掘，因此需要选择合适的数据采集和储存方式来适应上述处理需求，数据仓库正是一种提供辅助决策支持的数据储存和组织的技术，借助数据仓库技术就能实现对特定主题的热点问题进行多角度的分析。

处理面临的服务种类繁多，数据的格式可能完全不同，这就是要采集来的数据，先进行预处理操作才能入库储存。只要经过预处理的数据才能被用来进行下一步分析工作。其次系统要实现投诉文本中热点问题的自动发现，需要对大数量级的比较数据进行处理和运算，这就在客观上要求处理的速度应该尽量满足快速和准确的要求。

针对问题的不同特点，采用优化的 K-means 聚类算法进行文本划分，有效地克服了常规 K-means 聚类需要输入聚类数目和初始聚类中心以及结果常常陷入局部最优的缺点，再结合数据仓库技术对基于主题的热点分析进行深入分析，对于未知主题的热点问题，通过对其自身特点的分析，采用基于密度 DBSCAN 的聚类算法先进行主题发现，再用关联规则对这类热点进行挖掘。

三、模型假设

1. 假设题目所给的数据真实可靠；
2. 假设用户投诉符合当时心情；

四、定义与符号说明

$Term_i$：一条投诉文本记录通常是由字词或短语等基本语言单位组成的项集合，一般将这些语言单位称为项。

P：表示文本特征抽取中选择的特征词的总个数。

w_{ij}：是第 i 个特征词 $Term_i$ 在第 j 个文本 x_j 中的权重值。

N：代表文本库中所有的文本总个数。

$DF_{C,t}$：代表新增加的文本集合 C_t 的正向文档频率。

DF_{t-1}：代表 t-1 时刻的正向文档频率。

TF_{ij}：表示第 i 个特征项$Term_i$在整个文本集合中出现过的文本数目。

五、模型的建立与求解

文本向量空间模型[1]：

文本的向量空间模型可描述为式(1-1)：

$$x_j = (w_{1j}, w_{2j}, ..., w_{pj}) \qquad \text{式(1-1)}$$

对于特征项$Term_i$在文档x_j中的权重可以表示为式(1-2)：

$$w(Term_i, x_j) = TF(Term_i, x_j) * IDF(Term_i) \qquad \text{式(1-2)}$$

在文本的向量空间模型表示方法中，每个特征项都应该对应一个权重值，该值的大小代表着该特征项对某个文本是否重要，也就是说这个权值反映了该特征项区分所在文本与其他文本能力的大小。

在 TF-IDF 加权算法中，第一项 TF 指特征项$Term_i$在文本x_j中出现的次数，标记为TF_{ij}，TF_{ij}值越高，表示特征项$Term_i$越重要。文档频率 DF 指的是文本集合中出现了特征项$Term_i$的文本个数，标记为DF_i，DF_i值越小表示特征项$Term_i$在衡量文本相似度上发挥的作用越小。第二项 IDF（逆向文档频率）表示某个特征项在文本集合库中的表现，用该值可以判断此项在文本集合库中的分布情况，通常情况下，IDF 和 DF 的值成反比关系，如式(1-3)所示：

$$IDF_i = \log(\frac{N}{DF_i}) \qquad \text{式(1-3)}$$

IDF_i值越大，表示$Term_i$对文本的区别程度就越大。如果特征项只在一个文本中出现，即$DF_i = 1$，如式(1-4)所示关系：

$$IDF = IDF_i = \log(\frac{N}{DF_i}) = \log(\frac{N}{1}) = \log(N) \qquad \text{式(1-4)}$$

同理如果一个特征项在所有文本中都出现过了，则有如式(1-5)关系：

$$IDF = IDF_i = \log(\frac{N}{DF_i}) = \log(\frac{N}{N}) = \log(1) = 0 \qquad \text{式(1-5)}$$

在 VSM 模型中，文本特征项的权值一般使用式(1-6)：

$$w_i(x_j) = TF_{ij} * \log(\frac{N}{N_i} + 0.01) \qquad \text{式(1-6)}$$

一般要将向量归一化到单位向量，以减小文本长度差异在计算相似度时造成的影响，则最后得到的 TF-IDF 权值计算公式为式(1-7)：

$$w_i(x_j) = \frac{TF_{ij} * \log(\frac{N}{N_i} + 0.01)}{\sqrt{\sum_{k=1}^{N} (TF_{kj})^2 * [\log(\frac{N}{N_k} + 0.01)]^2}} \qquad \text{式(1-7)}$$

可见，TF-IDF 公式描述了这样一种情况，当指定文本记录中的词语出现频率高，而且该特征词在文本库中出现的频率低时，就会形成高权重的 TF-IDF 项。这样就达到了保留罕见的重要词，同时过滤常见的一般性词的目的。

在传统的 TF-IDF 算法当中，特征词在整个文本集合出现的频率大小是通过逆向文档频率(IDF)来衡量的，而这里的逆向文档频率是静态的。由于投诉数据的不断更新，获取的投诉文本集会形成新的文本流，文本集内文本总个数在不断增加，针对投诉文本集合可能发生更新的要求，文本设计了一种增量的 TF-IDF 模型，在此模型中，文档频率 DF 不再是静态不改变的，而是随时间 t 不断变化的，在 t 时刻，当一个新的测试文本集合（此处有个式子）被加入模型时就更新文档频率：

$$DF_t = DF_{t-1} + DF_{C,t} \qquad \text{式(1-8)}$$

为了表示不同类别对象间的相似程度，需要先定义一些统计量来度量，常用的相似度度量标准有距离和相似系数两种。

首先，距离是文本聚类中常用的相似度衡量标准。对于有 p 维特征向量的文本集合来说，N 条文本记录可看作是 p 维特征空间中的 N 个点，一般可用点间距离来度量文本记录间的相似度。第 i 个文本和第 j 个文本间的距离定义为 $d_{i,j}$，并且距离 $d_{i,j}$ 应满足如下的条件：

(1)距离非负，也就是对于任意下标 i 和 j，恒有 $d_{i,j} \geq 0$，而且等号仅在两个文本 p 个维度对应值都相等时才成立。

(2)距离对称，对所有下标 i，j 有 $d_{ij} = d_{ji}$。

(3)满足不等式：对所有下标 i，j，k，不等式 $d_{ij} \leq d_{ik} + d_{kj}$ 恒成立。

由上述几条性质可以得出，两个文本的距离必定在 $0 \to \infty$ 范围内，并且距离值越小，表示两个文本就越接近。在文本聚类过程中，最常见的几种距离有：

(1)明氏距离：

$$d_{ij}(q) = (\sum_{k=1}^{p} | w_{ki} - w_{kj} |)^{\frac{1}{q}}, q > 0 \qquad \text{式(1-9)}$$

其中，w_{kj} 代表第 k 个特征词在第 j 个文本 x_j 中的权重值，x_j 的表示模型参见式(1-1)。

q=1 时，明氏距离演变为曼哈顿距离：

$$d_{ij}(1) = \sum_{k=1}^{p} | w_{ki} - w_{kj} | \qquad \text{式(1-10)}$$

q=2 时，明氏距离演变为欧式距离：

$$d_{ij}(2) = [\sum_{k=1}^{p} | w_{ki} - w_{kj} |^2]^{\frac{1}{2}} \qquad \text{式(1-11)}$$

$q = +\infty$ 时，明氏距离演变为切比雪夫距离：

$$d_{ij}(\infty) = \max_{1 \leq k \leq p} | w_{ki} - w_{kj} | \qquad \text{式(1-12)}$$

(2)马氏距离：明氏距离一般适用于欧式空间，通常认为欧式空间中各维度之间完全独立。但考虑到文本中各特征向量的观测值常常是随机变量，那么随机向量会呈现出一定的分布规律，文本向量的各分量之间是相当的，当需要考虑分量之间的相关性时，一般使用马氏距离，定义为：

$$d_{ij}(M) = ((x_i - x_j)^T \sum^{-1} (x_i - x_j))^{\frac{1}{2}} \qquad \text{式(1-13)}$$

其中∑是协方差矩阵。

其次，对于 VSM 空间中的两个向量，害可以用相似系数来表示其相似程度。假设第 i 个和第 j 个向量的相似系数定定义为a_{ij}，则a_{ij}有以下的性质：

(1)绝对值应小于 1，对于所有下标 i，j，必须满足$|a_{ij}|\leq 1$，且等号仅在两向量存在线性关系时，也就是可以表示$x_i = Cx_j$（其中 C 是不为零常数）时才成立。

(2)对称性质，对任意不同的 i，j，必然有$a_{ij} = a_{ji}$。

两文本向量之间的相似系数一般有两种表示形式：

(1)夹角余弦形式：p 维空间中两个向量基于相对位置会形成一个夹角，相似系数可以用这个角的余弦值来衡量，据此第 i 个和第 j 个文本的相似系数可定为：

$$sim(i,j) = \cos(\theta_{ij}) = \frac{\sum_{k=1}^{p} w_{ki} w_{kj}}{\sqrt{\sum_{k=1}^{p} w_{ki}^2 \sum_{k=1}^{p} w_{kj}^2}} \qquad \text{式(1-14)}$$

其中第 i 文本向量表示为：

$$x_i = (w_{1i},...,w_{ki},...,w_{pi}) \qquad \text{式(1-15)}$$

(2)相关系数：第 i 个向量和第 j 个向量之间的相关系数可表示为：

$$r_{ij} = \frac{\sum_{k=1}^{p} (w_{ki} - \overline{x_i})(w_{kj} - \overline{x_j})}{\sqrt{\sum_{k=1}^{p} (w_{ki} - \overline{x_i})^2 \sum_{k=1}^{p} (w_{kj} - \overline{x_j})^2}} \qquad \text{式(1-16)}$$

其中每个向量都是 p 维随机向量，$\overline{x_i}$，$\overline{x_j}$分别是第 i 个和第 j 个向量在各维度上的平均值，即：

$$\overline{x_i} = \frac{1}{p} \sum_{k=1}^{p} w_{ki} \qquad \text{式(1-17)}$$

$$\overline{x_j} = \frac{1}{p} \sum_{k=1}^{p} w_{kj} \qquad \text{式(1-18)}$$

K-means 算法：

K-means 算法是由 MacQueen 等人于 1967 年首次提出的一种划分聚类算法，它是一个不断迭代的过程，直到算法收敛到一定的结束条件才终止[2]。先给出聚类过程中涉及到的一些定义：

（1）样本集合为：

$$X = \{x_1, x_2, ..., x_N\} \qquad \text{式（2-1）}$$

其中 N 是样本总数；

（2）k 个聚类的初始中心点为$\{z_1, z_2, ..., z_k\}$，其中n_i表示第 i 类的样本数；

$$z_i = \frac{1}{n_i} \sum_{x \in a_i} x \qquad \text{式（2-2）}$$

（3）定义 k 个类别为$\{a_1, a_2, ..., a_k\}$；

（4）定义目标收敛函数如式（2-3）所示，它实际上表示样本点到各自所属类别中心的距离的平方和，也就是最小均方误差。

$$E=\sum_{i=1}^{k}\sum_{j=1}^{n_i}d_{ij}(z_i,x_j) \qquad \text{式（2-3）}$$

K-means 类聚算法的一般步骤为：

（1）选择 k 个随机的样本点$\{z_1,z_2,...,z_k\}$为各类的初始类聚中心点；

（2）计算其每个未归类样本点到各类中心点的距离值（一般采用欧式距离），依次将其放到于中心点距离最小的那个类别中去；

（3）用式（2-2）再次计算新的类聚中心$\{z_1(y),z_2(y),...,z_k(y)\}$，其中 y 表示第 y 次迭代；

（4）重复执行（2）和（3），直到最小均方误差小于某个设定的最小阈值，表示聚类中心已经趋于平稳，此时可以结束聚类。

经过众多研究者多年来对 K-means 算法中 k 值确定方法的研究，目前大多数的研究支持一下结论[2]：

$$k_{opt}\le k_{\max}\le\sqrt{N} \qquad \text{式（2-4）}$$

其中k_{opt}是指最佳的聚类数目，$k_{\max}$是最大聚类数目。

据此，本文提出按照贝叶斯信息准则函数判断聚类数目的标准，可减少人为决定聚类数目带来的主观影响，使聚类的结果更能客观地反映数据内部特征，而且实验证明，BIC 准则估计聚类数目的误差很小，本文将通过 BIC 准则来优化 K-means 算法。

BIC 称为贝叶斯信息准则，它的理论来源是贝叶斯概率理论，通常应用于最佳模型选择的问题中，其实质是根据 BIC 值衡量待定模型复杂性和该模型描述样本数据的能力这两个指标，希望在两者之间取得最佳平衡状态，从而确定一个最优模型。如果输入的样本数据集$X=\{x_1,x_2,...,x_N\}$满足独立同分布，设其后验概率函数为$f(M_i|\omega)$,定义可供选择的模型集合$W(k)$为：

$$W(k)=\{f(M_i|\omega)|\omega=(\lambda_1,\lambda_2,...,\lambda_k),\omega\in\Omega_k\} \qquad \text{式（2-5）}$$

其中，Ω_k代表模型空间，$\lambda_1,\lambda_2,...,\lambda_k$表示空间中任一模型的 k 个参数，那么模型选择实际上是在$W(k)$中选取最能反映给定样本集特征的那个函数模型$f(M_i|\omega)|$。BIC 的定义是：

$$B(f)=\ln L_{\hat{\omega}k}(X)-k\ln N \qquad \text{式（2-6）}$$

公式中前一项$\ln L_{\hat{\omega}k}(X)$表征了函数模型$f(M_i|\omega)$在 X 上的极大后验似然概率值，也就是$f(M_i|\omega)$能够表征样本集 X 的能力，第二项$k\ln N$代表模型的复杂度，是为避免文本特征向量稀疏时可能出现维度灾难而设置的惩罚项，那么最佳函数模型$f(M_i|\omega)|$满足下式时 BIC 值达到最大。

具体算法的步骤如下：

（1）根据给定的候选模型集体，从中选择 BIC 值最大时对应的那个函数模型$f(M_i|\omega)|$。

（2）执行普通的 K-means 算法，只是在确定样本类别归属时进行以下的判断：计算当前函数模型对应的 BIC 值来测量主题的数目是否可以再进行优化，即当两个类不能确定是否需要合并时，就比较合并和不合并两种情况下整个样本集的 BIC 值，哪种情况下的 BIC 值更大，就确定那种情况下划分类的方法。

（3）当（2）中 BIC 值达到最大时，表示目前的主题数目是在 BIC 准则下最优的。

（4）根据得到的主题数目 k，由用户指定 k 个主题的关键词，并对每个主题选择出一个初始的类聚中心，执行 K-means 算法，由于聚类数目和初始中心都不再是随机指定

的，因此该算法的聚类效果优于一般的为改进的 K-means 聚类算法。

实验结果：在聚类质量评价中经常使用的评估聚类效果的指标有查准率（precision）、查全率（recall）和 F-measure，其中 F-measure 是由前两个指标运算得到的，是结合两者进行评估的综合指标，如果 F-measure 值越大表明聚类结果越好[3]。分别定义为：

$$precision(i,j)=\frac{N_{ij}}{N_i} \qquad \text{式（2-7）}$$

$$recall(i,j)=\frac{N_{ij}}{N_i} \qquad \text{式（2-8）}$$

$$F(i)=\frac{2\times precision(i,j)\times recall(i,j)}{precision(i,j)+recall(i,j)} \qquad \text{式（2-9）}$$

其中，N_{ij} 代表第 j 个结果簇中属于第 i 个原始分类的样本个数，N_j 为第 j 个结果簇的样本总数，N_i 表示第 i 个原始分类的样本总数。

由于热点问题的形成需要一定的时间，所以此处引入时间度量，本文标记时间集合为 $T=\{t_1,t_2,...,t_i...\}$，代表时间轴上的有序集合，在本文中用“天”作为时间的基本单位。

经过之前的聚类分析之后，将输入的投诉文本集合为 k 个主题类别，先从中抽取出 m 个可能为热点问题的主题类别，命名为热点问题初始集合，标记为 $CSet=\{C_1,C_2,...,C_m\}$，其中 C_j，$1\le j\le m\le k$ 表示第 j 个可能的热点问题类别，每个类别包含一定数目的样本记录，表示为 $C_j=\{x_{j1},...,x_{jy}\}$，其中 y 表示该类别包含的样本数目，单个样本表示为 $X_j=(w_{1j},w_{2j},...,w_{pj})$。加入时间度量之后第 t_i 日的热点问题初始集合为 $CSet(t_i)$，$m(t_i)$ 为初始集合包含的样本记录总数，$mC_j(t_i)$ 表示 Cj 中包含的样本数。

通过考察热点问题初始集合随时间的增长速度来进行热点问题的识别，转化为可定量分析的数学模型如下：

（1）首先 C_j 需要满足一段时间内的连续可现性，这里的重现是指一段时间内存在与 C_j 紧密关联的相似类别，本文后面将使用关联规则来确定两个类别之间是否相关。

（2）在此基础上定义热点问题初始集合中某一个类别 Cj 的增长率为：

$$Ratio=\frac{mC_j(t_i)-mC_j(t_{i-1})}{mC_j(t_{i-1})},t_i,t_{i-1}\in T \qquad \text{式（2-10）}$$

经过连续时期内对同一个类别 C_j 进行考察，若增长率均大于规定的最小阈值 $Ratio_{\min}$ 时，可以认为 C_j 类别就是一个热点问题。根据上面的定义，可以将这个增长率指标 Ratio 定义为热度指标，因为正是对这一指标的考察确定了热度问题所属的类别。

假设 C_j 的样本记录包含的特征词集合中，挖掘到热点相关的特征词集合为 $\{Term_1,Term_2,...,Term_i,...,Term_y\}$，则可以用此集合代表该类热点问题。当已知一个热点问题类别 C_j，如何判断其他的类别是否与 C_j 描述的热点问题相关联，本文采用关联规则来判断两个集合的关联性。

（1）对于两个类别的文本集合 A 和 B，支持度 Support 定义为同时包含 A 和 B 文本的概率，如式（2-11）所示，其中 A 称为规则的前件，B 称为后件。

$$Support(A\Rightarrow B)=P(A\cup B) \qquad \text{式（2-11）}$$

（2）置信度表示 A、B 相关联的可信度有多大，用公式表示是

$$Confidence(A \Rightarrow B) = \frac{P(A \cup B)}{P(A)} \quad \text{式（2-12）}$$

定义最小支持度 $Support_{\min}$、最大支持度 $Support_{\max}$，最小置信度 $Confidence_{\min}$，若支持度符合式（2-13），置信度满足时（2-14），增长率满足（2-15）

$$Support_{\min} \le Support(A \Rightarrow B) \le Support_{\max} \quad \text{式（2-13）}$$

$$Confidence(A \Rightarrow B) \ge Confidence_{\min} \quad \text{式（2-14）}$$

$$Ratio \ge Ratio_{\min} \quad \text{式（2-15）}$$

那么 A 和 B 同属于热点问题的集合 C_j，并且 A、B 包含的特征词集可以代表该类热点问题，此外可以根据导出的关联规则 $A \Rightarrow B$，来判断其他文本集合是否属于该类热点问题。

通过对热点问题初始集合中的类别进行增长率的比较分析，从中可以发现符合特定热点问题的类别集合，之后可以进行热点问题展示以及进一步的预警。

六、模型评价与推广

（一）模型的优点

使用了基于 BIC 准则的 K-means 聚类算法来对样本库进行主题归类处理，这种算法克服了一般 K-means 不易确定聚类数目和初始中心点的缺点。

（二）模型的缺点

主观性过强、有很多存在不确定性，同时有一定的偏差。

（三）模型的改进

在此次实验的基础上，可以考虑降低计算复杂度，比如信息增益、文本证据权等，通过效果对比，选择最优的方法。

七、参考文献

[1] 时志芳 移动投诉信息中热点问题的自动发现与分析,北京邮电大学专业学位硕士毕业论文.2013

[2] 顾洪博 基于k-means算法的k值优化的研究与应用,海南大学学报自然科学版.29(4) 2007 p132-p135

[3] 齐海风 网络舆情热点发现与事件跟踪技术研究[学位论文]，哈尔滨工程大学，2008