基于 word2vec 与 K-means 算法食品安全事件自动聚类研究

沈 思[1]，梁晓静[2]
（1.南京理工大学经济管理学院，江苏 南京 210094；2.金陵科技学院软件工程学院，江苏 南京 210000）

摘要：类别知识在食品安全事件的应急处理中不仅保证了相应应急措施推送的精准性而且可以提高食品安全事件应急处理的效率。为了更加精准的获取食品安全事件中的类别知识，结合统计、对比和分析的系统方法，通过食品安全事件的微博和信息数据，基于 word2vec 和 K-means 算法，文章探究了类别知识挖掘的具体流程、方法和相应的类别知识挖掘结果的评价等方面的相应问题。通过具体的实验，在所使用的数据上，确定了最优的聚类模型。

关键词：食品安全事件；word2vec；K-means；轮廓系数 TP393
中图分类号： **文献标识码：**A **文章编号：**1673-1131（2018）11-0008-03

Research of Automatic Clustering of Food Safety Events Based on Word2vec and K-means Algorithm

Si Shen[1],XiaoJing Liang[2]
（1.School of economics and management, Nanjing University of Science and Technology ,Nanjing, 210094;
2. School of software engineering, Jinling institute of technology,Nanjing, 210000）

Abstract: The categories of knowledge not only ensures the precision of pushing the emergency measures but also can improve the efficiency of handling the food safety event emergency in handling the food safety event emergency. By using the method of statistics, comparison and analysis of system in the food safety incidents Weibo and information data, this paper explores the specific categories of knowledge mining process, method and the corresponding problems of evaluation of category knowledge mining results in order to more accurately get the category knowledge of food safety event. Through specific experiments, the best cluster model with category knowledge is determined based on the data used.

Keywords: Food safety event; Word2vec; K - means; Outline of the coefficient

0 引言

在食品安全事件的应急处理过程中，有效地知识是应对食品安全事件的关键因素，而再把有效的知识推送给用户的过程中，为了高效和精准的完成对应急知识的推送，需要获取有关食品安全事件的类别知识。目前获取类别知识最有效和主要的途径是通过聚类的方法对食品安全事件文本进行类别知识的获取。在上述这一背景下，针对相应的食品安全事件文本，通过把 word2vec 与 K-means 算法进行相融合的策略，针对领域食品安全事件的类别知识进行了相应的获取。

有关基于食品安全事件数据的探究目前主要停留在数据库的构建阶段，比较有代表性的研究如下。目前对食品安全事件的信息处理还停留在构建数据库的阶段，比较有代表性的研究如下。为了更好的完成对食品安全事件的预警，提高食品安全事件监管的科学性和高效性，张星联和吴云红分别提出[1-2]提出构建食品安全事件动态数据库，并确保该数据库可以及时的更新，从而一方面可以有效地消除食品安全事件应对过程中的信息不对称问题，另一方面确保食品安全事件应对的透明度。在涉及到食品安全事件发生的整个过程中，食

情况提供了更有效的解决办法。通过多次实验证明，井下瞬变电磁仪在含水异常体探测中效果明显，结果准确，值得大量推广使用。该探测方法主要有以下几个优势：

（1）施工简单、仪器便于携带，工作时间短，不影响矿井正常生产作业，即能满足防治水工作需要，又不影响生产需要。

（2）井下由于特殊的支护环境，金属质地的网片、锚杆等难免会对仪器造成影响，故一定要将影响因素降到最低，孔中瞬变恰巧在一定程度上解决了这个问题。

（3）在不需要特殊布钻的情况下，将仪器接收探头送入现有钻孔，不仅不受周围环境的影响，而且使得仪器离异常更近，解释更清晰，更容易确定异常位置。

（4）降低了矿方盲目打钻的工作量，为矿方节省了成本。

参考文献：

[1] 肖杰,焦当云.瞬变电磁法在煤矿防治水工作中的应用[J].江西煤炭科技,2018(01):113-114+119.
[2] 王卫光.瞬变电磁仪在煤矿水文地质探测方面的应用研究[J].机电工程技术,2015,44(07):225
[3] 薛国强，于景邨.瞬变电磁法在煤炭领域的研究与应用新进展[J].地球物理学进展,2017,32(01):319-326.
[4] 储韬玉.矿井孔中瞬变电磁测量方法及其应用研究[D].中国矿业大学,2015.
[5] 陈丁. 矿井全空间巷道孔中瞬变电磁波场特征数值模拟研究[D].中国矿业大学(北京),2016.
[6] 张青. 瞬变电磁法在瓦斯隧道隧底采空区探测中的应用[J].科技风,2018(16):91.
[7] 王玺凯,王福生.瞬变电磁法在某水源井勘测中的应用[J].科技创新与应用,2018(15):171-172.
[8] 李健军.瞬变电磁仪法与超前水平钻孔法在隧道施工中的综合运用[J].价值工程,2017,36(32):118-120.
[9] 谢成梁.瞬变电磁仪在工作面底板水害探测中的应用[J].自动化与仪器仪表,2017(10):210-211+214.
[10] 孟宪杨.巷道探测中瞬变电磁仪的应用[J].矿业装备,2017(05):52-53.

作者简介 邓立博（1990-），男，陕西渭南人，助理研究员，硕士，研究方向 地球物理探测仪器开发。

品的加工是极其重要的一环，如何确保食品加工的安全性、可跟踪性和回溯性是控制食品安全事件发生的关键，基于此，余清等[3]提出了构建食品加工风险数据库，为食品安全事件的检测奠定基础。在食品安全事件当中，生鲜农产品不仅涉及范围广泛而且在时间上对生鲜农产品也提出了更高的监管要求，鉴于此，以彭州市下面的一个镇为数据源，贾凯等[4]提出了构建生鲜农产品数据源库，从而实现对涉及到该镇的所有生鲜农产品可以实现有效的回溯。基于大数据的技术和方法，刘翠玲[5]提出了一种应对食品安全事件的大数据监管平台，并基于该平台对农药残留这一食品安全事件进行了具体的对比探究和分析。

在对食品安全事件进行聚类的过程中，所选择的聚类方法应该充分考虑到中文文本的特殊性，以便于从食品安全事件的数据中挖掘出有效和针对性比较强的类别知识。通过逐一地计算样本数据的距离，夏长辉[6]基于最大的距离样本为初始的聚类焦点，利用 k-means 聚类算法的特征逐步地完成对所选择数据的聚类，从而从文本中挖掘出相应的聚类知识，这种方法的缺陷是计算量比较大。针对微生物群落的数据分布，王侠林、贺建峰[7]基于 OTUs 数据集的 27 个样本进行了聚类的探究，并通过与 PCA 这一方法的对比，发现 K-Means 的整体聚类效果更加突出。充分利用 HowNet 当中领域语义词汇知识，结合文档集合所呈现出来的特征空间，黄建宇和周爱武[8]等给出了优化后的 k 值的 k-means 聚类策略，这在一定程度上有效提高了聚类的整体效果，但由于所依赖领域词汇库的局限性，在一定程度上限制了聚类效果的进一步提升。在云计算平台 Hadoop 下的并行 k-means 聚类算法基础上，赵卫中[9]结合不同的数据源，充分而具体的验证了在大数据平台上聚类算法的加速比、扩展率和伸缩率等不同指标的整体性能。结合大数据的技术、方法和理念，利用最小相似度文本对的理念，通过构建 MapReduce 基础上的文本聚类模型，武森和冯小东[10]提出了融合大数据的 k-means 聚类方法，所获取的聚类结果整体性能非常突出，并且具有较强的可迁移性。这一聚类的理念可以有效地拓展到以汉语为呈现形式上的非结构化文本上。

1 模型和食品安全事件文本语料介绍

为了从食品安全事件文本中获取更加有效的特征或者信息，本研究基于 word2vec[11-13]获取食品安全事件文本中词汇的向量值，该算法是由 2013 年的谷歌公司提出的，同时对该算法进行工具包式的封装。该算法主要由 CBOW（Continuous Bag-Of-Words）和 Skip-gram 这两个具体的模型构成，这两种模型均通过某一当前的词汇去预测其他词汇，从而提高词汇单位关联的精准性。word2vec 算法最为突出的一点是可以进行向量间的有效转换，从而极大地提升了对文本进行语义相似度计算的精准性。基于这一计算出来的精准度较高的相似度值，对于改进相应的文本聚类、同义词查询、检索系统的构建、相似度情感词计算等探究具有较为重要的意义和价值。该算法的具体原理如公式 1 所示。

$$p(w_t|w_{t-\frac{c}{2}},\dots,w_{t-1},w_{t+1},\dots,w_{t+\frac{c}{2}})=\prod_{i=1}^{c}p(context(w)i) \quad (1)$$

其中 CBOW 由输入层、投影层和输出层这三个层次构成，并且相互之间形成了有效的关联和嵌套。其中输入层是词向量的表示，投影层主要的功能是对输入的向量进行叠加，所使用的公式为 $\sum_{i=1}^{c}(context(w)i)$，context (w) 表示词的概率值，而输出层是通过一个树形结构进行表示的，具体为二叉树。在具体实现的时候是通过构建 Huffman 树来完成的。随着 word2vec 被广泛应用于文本聚类、信息检索和主题挖掘等研究当中，各种封装的 word2vec 版本均被大家在不同的具体探究上进行了应用。结合本文与食品安全事件信息聚类相关的研究，gensim[14]基础上的 word2vec 被选取应用于本文具体的探究当中。

在具体聚类算法选择上，基于本文食品安全事件文本数据的特征，k-means 聚类算法被选为针对食品安全事件数据的聚类知识挖掘算法，该聚类算法不仅设计加单，而且时间复杂度相对较低，并且可以有效地兼容不同的类型的数据。为了更好地利用该算法，针对食品安全事件数据，轮廓系数[15-16]被用来对初始聚类值进行设定，具体轮廓系数的计算公式如下：

$$s_i=\frac{b_i-a_i}{\max(a_i,b_i)} \quad (2)$$

在这一公式中 a_i 代表某一样本到其他潜在统一类别样本的平均距离，b_i 代表某一样本表到与其最相近的类内部的平均距离，因此 a_i 值越小则说明某一文本分类越合理，而 b_i 越大，则说明某一样本不应当归到其他类比，从而 s_i 值越靠近 1，则说明某一样本聚类合理，而 s_i 若靠近 0，说明某一样本聚类不合理。通过对所有样本的轮廓系数就平均值，进而会得到全局轮廓系数，具体计算公式如下。

$$s_k=\frac{1}{n}\sum_{i=1}^{n}s_i \quad (3)$$

通过关键词‘甲醛白菜’获取了微博上相关的信息共 7901 条，具体年份如表 1 所示。首先对相应的数据进行了清洗，去除掉了编码不一致和有乱码的文本。其次，调用 pynlpir 包分词，去除‘【’和‘】’符号，建立二维列表，满足 word2vec 进行计算的需要。最后，对所获取的数据进行了统一的存储，具体经过分词后的数据样例如图 1 所示。

表 1 新浪微博中‘甲醛白菜’主题数据集合

相关微博发表年份	去重前文本记录数	去重后文本记录数
2012	6547	5090
2013	2919	1948
2014	2057	53
总和	11524	7091

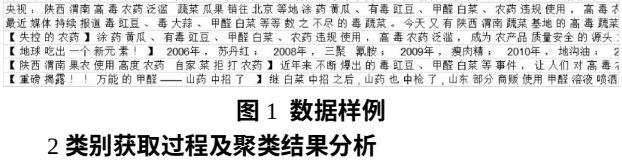

图 1 数据样例

2 类别获取过程及聚类结果分析

首先，通过 numpy 包进行矩阵运算，得到文本向量的表示，具体的向量表示如下所示：

[[7.69951130e-04 -3.15880664e-03 -9.92675225e-05 6.17912020e-03 -1.34906650e-03 -2.59567903e-03 5.11573336e-03 -4.85016753e-03.... -2.70609957e-03 -9.13482654e-03 6.09617936e-03 -3.95046254e-03 1.14068410e-03 1.87863904e-03 -4.74375231e-04 -8.93719038e-03], ...

[-6.88628220e-05 -1.19989728e-03 -2.36077557e-04 2.49695818e-03 -8.79188088e-04 -1.48803692e-03 1.81370428e-03 -1.79939355e-03.... -5.38808817e-04 -4.01521463e-03 2.63561918e-03 -1.31672230e-03 1.45825999e-04 6.91009550e-04 2.14562120e-04 -3.06116202e-03]

其次，根据“肘部法则”来确定最优“簇值”K，即当图形变化很快，出现类似“肘部”时，及 K 较为适合，具体实验过程中“肘部法则”的最优化选择过程如图 2 所示。

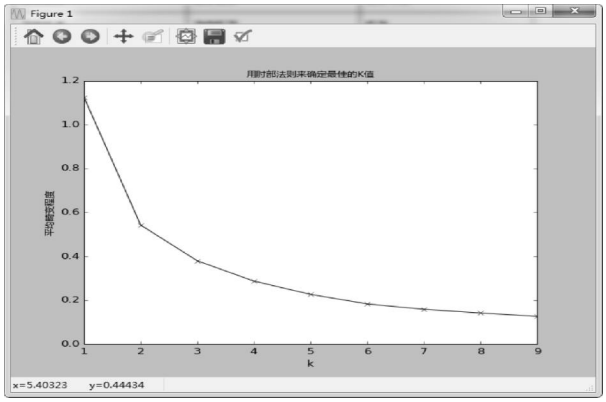

图 2 “肘部法则”来确定最优“簇值”K 过程图

最后，调用 sklearn 包来进行聚类，并得到“轮廓系数”(轮廓系数的值是介于 [-1，1]，越趋近于 1 代表内聚度和分离度都相对较优)。

在上述整个聚类的流程下，对于“甲醛白菜”微博文本当中所蕴含的类别知识进行了预估计和逐步的实验，并观察在给定的类下轮廓系数的不同。具体实验结果如表 2 所示。其中超参结果的选择为 alpha=0.001，min_count=40, size=200，最佳聚类类别 k=2。

表 2 基于 word2vec 的聚类结果与分析

相关微博发表年份	对应轮廓系数值
2012	0.6607
2013	0.7153
2014	0.8808

根据三个不同的轮廓系数和图 3 所示的聚类结果示意图可以看出，利用 word2vec 与 K-means 算法进行相融合的策略，能有效对真实的与甲醛白菜相关的所有微博文本中蕴含的类别知识进行聚类。

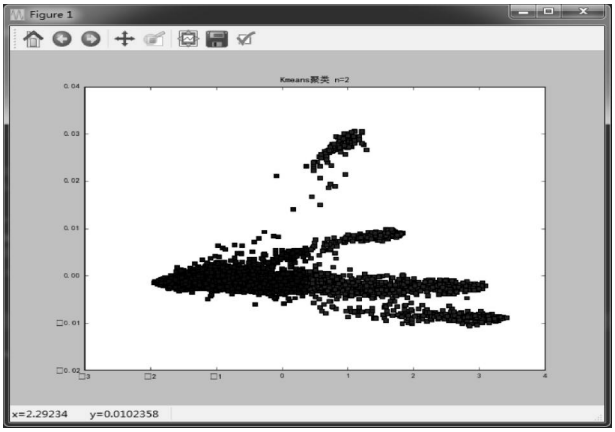

图 3 k=2 时 2012 年微博文本聚类结果图

3 结语

为了从食品安全事件微博文本中获取相应的类别知识，本文在新的文本向量表示下，结合已有的相应聚类算法，针对甲醛白菜中所蕴含的类别知识进行了系统的挖掘和分析。通过具体实验的验证，本文发现所聚出来的类为 2 的时候，可以看作最有效的聚类结果，并符合甲醛白菜微博文本中所包含的类别知识。

参考文献：

[1] 张星联. 我国食品安全预警数据库系统的建设与实现[J]. 食品科技, 2008, 33(12): 250-254.
[2] 吴云红.食品监管改革的关键--基于互联网的动态第三方数据库[J].食品工业科技, 2009(09):272-274.
[3] 余清, 洪源. 加工食品风险数据库的构建思路[J]. 价值工程, 2013(30):174-175.
[4] 贾凯, 彭培好, 阮伟玲. 四川省彭州市三界镇农民专业合作社调查研究[J]..北京农业,2014,1:247-248.
[5] 刘翠玲, 徐莹莹, 孙晓荣,等. 基于多源大数据食品安全监测预控系统的设计与实现[J]. 食品科学技术学报, 2018, 36(3): 88-94.
[6] 夏长辉. 一种改进的 K-means 聚类算法[J]. 信息与电脑(理论版),2017,(14):40-42.
[7] 王侠林, 贺建峰. 基于 K-Means 聚类的微生物群落结构研究[J]. 软件导刊, 2018(1):146-148.
[8] 黄建宇, 周爱武,肖云,谭天诚. 基于特征空间的文本聚类[J/OL]. 计算机技术与发展,2017,(08):1-3.
[9] 赵卫中, 马慧芳, 傅燕翔,等. 基于云计算平台 Hadoop 的并行 k-means 聚类算法设计研究[J]. 计算机科学, 2011, 38(10):166-168.
[10] 武森, 冯小东, 杨杰,等. 基于 MapReduce 的大规模文本聚类并行化[J]. 北京科技大学学报, 2014(10):1411-1419.
[11] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013.
[12] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and their Compositionality [J]. Advances in Neural Information Processing Systems, 2013, 26:3111-3119.
[13] 姜霖, 王东波. 采用连续词袋模型(CBOW)的领域术语自动抽取研究[J]. 现代图书情报技术, 2016, 32(2):9-15.
[14] eh ek R. models.word2vec-Deep learning with word2vec [EB/OL]. [2018-07-26]. https://radimrehurek.com/gensim/models/word2vec.html.
[15] 张冬梅. 基于轮廓系数的层次聚类算法研究[D]. 秦皇岛: 燕山大学, 2009.
[16] 朱连江, 马炳先, 赵学泉. 基于轮廓系数的聚类有效性分析[J]. 计算机应用, 2010(s2):139-141.

基金项目 本文系江苏省社会科学基金“时间感知大数据特征下的食品安全突发事件应对策略挖掘研究”(项目编号：15TQC003)的研究成果之一。

作者简介 沈思(1983-)，女，讲师，博士，主要研究方向为信息检索，学术文本挖掘，时间语义分析；梁晓静(1996-)，女，主要研究方向为学术文本检索。