Computer Science and Application 计算机科学与应用, 2014, 4, 204-216
Published Online September 2014 in Hans. http://www.hanspub.org/journal/csa
http://dx.doi.org/10.12677/csa.2014.49029

Hans汉斯

An Approach for Algorithm of Tobacco Enterprise Archives Text Automatic Classification Based on KNN

Shifan Huang[1*], Yong Shen[1,2], Hongwei Kang[1,2], Daohong Wang[3], Jianlin Zheng[1], Bo Lang[1], Dong Wang[1], Congcong Jia[1]

[1]School of Software, Yunnan University, Kunming
[2]Key Laboratory for Software Engineering of Yunnan Province, Kunming
[3]Rural Credit Cooperatives and Settlement Center of Science and Technology of Yunnan Province, Kunming
Email: *974794674@qq.com

Received: Jul. 14th, 2014; revised Aug. 12th, 2014; accepted: Aug. 21st, 2014

Abstract

By researching historical archives text data of a cigarette factory in Yunnan province, combing with actual situation, we have detailedly designed acquisition of file text subject headings and automatic classification algorithm. Furthermore, TFIDF algorithm is introduced to acquisition algorithm of subject headings, thus the problem that algorithm can't automatically obtain subject headings when text file lack title, document number and statement items is solved. In this paper, KNN adjacent algorithm is introduced to the algorithm of automatic classification, and it solves the problem which can't be solved according to the title and approval document for automatically classifying archives text. At the same time, we also consider the problem that classifies file text according to the storage life. The experimental results show that this algorithm obviously improves the classified efficiency of archives text of the tobacco enterprise.

Keywords

TFIDF, KNN, Archives of Tobacco, Automatic Text Categorization, Storage Life

基于KNN的烟草企业档案文本自动分类算法研究

*通讯作者。

黄世反[1*]，沈　勇[1,2]，康洪炜[1,2]，王道红[3]，郑见琳[1]，郎　波[1]，王　冬[1]，贾丛丛[1]

[1]云南大学，软件学院，昆明
[2]云南省软件工程重点实验室，昆明
[3]云南省农村信用社科技结算中心，昆明
Email: *974794674@qq.com

收稿日期：2014年7月14日；修回日期：2014年8月12日；录用日期：2014年8月21日

摘　要

通过对云南某卷烟厂历史档案文本数据的分析研究，结合实际情况，对档案文本主题词的获取和自动分类算法进行了详细的设计。且在主题词获取算法中引入了TFIDF算法，解决了档案文本缺少题名、文号及责任者项时，算法无法自动获取主题词的问题。在文本自动分类算法中引入了KNN最邻近算法，解决了无法根据题名、文号进行档案文本自动分类的问题。同时，还考虑了档案文本按保存期限进行分类的问题。实验结果证明，该算法明显提高了烟草企业档案文本的分类效率。

关键词

TFIDF，KNN，烟草档案，文本自动分类，保存期限

1. 引言

随着信息技术和企业产业规模的迅速发展，企业档案资料也成倍或数倍的增加，如何对这些迅速增加的档案文本进行正确、快速、高效的分类，是目前企业档案管理面临的一个严峻问题，也是档案文本自分类的一个研究热点。

自美国 IBM 公司鲁恩(H.P. Luhn)的一系列文章[1]-[4]拉开了文献自动处理(自动标引、自动分类、自动编制文摘等)研究的帷幕之后，国外众多专家学者致力于对此项目的研究，取得了可喜的成绩。我国在汉语文本自动分类方面的研究起步较晚，特别对档案文本自动分类算法的研究不多。文献[5]提出了一种多因素加权归类算法，该算法简单、易用，在实际应用取得了很好的效果，但仍然存在不足之处：1) 没有考虑档案文本文号对分类结果的影响；2) 对没有题名和文号或者是题名和文号没有实质意义的档案文本不能进行自动分类；3) 没有考虑档案文本的保存期限。文献[6]采用档案来源原则来建立档案信息自动分类编目体系。文献[7]探讨了档案管理传统理论的发展之路，再一次肯定了“来源原则”——即本文所说的责任者原则，在档案文本分类中的重要作用。

云南某卷烟厂历史档案文本数据量庞大、种类繁多，大部分老的历史档案文本数据分类混乱，管理系统不止一个，各分厂与总部的数据互相独立，没有统一的管理机制，造成大量人力、时间及资金的浪费。基于此，该企业申报了“历史档案文本数据迁移”项目，希望能将不同年代、不同管理系统、不同数据结构及各个分厂的历史档案文本数据进行整合、归类，进行统一、有效的管理。

本文在文献[5]的研究基础上，从实用、易于实现、自动化的宗旨出发，结合项目的实际情况，对文献[5]中所用到的算法进行了改进：针对问题(1)，本文在算法中引入了文号项；针对问题(2)，本文在分类特征词获取算法中引入了 TFIDF 算法，在归类算法中引入了 KNN 最邻近算法，对原算法的不足进行了改进与优化；针对问题(3)，本文在每一个大类下，再一次对档案文本进行保存期限的分类(本文将保存期

限分为：永久、长期和短期三种，其分别用字母 Y、C、D 来表示)，其中短期又分为：D30(30 年)、D20(20 年)、D10(10 年)，其分类类别结构图如图 1 所示：

2. TFIDF 算法和 KNN 最邻近算法介绍

目前，中文文本特征词的提取算法主要有：特征频率方法(Term Frequency: TF)、文档频率方法(Document Frequency: DF)、反文档频率方法(Inverse Document Frequency: IDF)、信息增益方法(Information Gain: IG)、互信息方法(Mutual Information: MI)、期望交叉熵(Expected Cross Entropy: ECE)及 $\chi 2$ 统计量(Chi-square: CHI)等。

其中，因为 TFIDF 算法相对简单、且有较高的查全率和查准率，一直受到众多应用领域的青睐。此外，在本文中因领域的特殊性，大部分档案文本我们可以根据题名和文号信息确定特征词，不需要挖掘文本内容信息；只有少部分没有题名和文号或者题名和文号模糊不清的文本需要从文本内容中提取特征词。所以本文可以选择算法简单且高效的算法。因此，本文采用 TFIDF 算法作为特征词提取算法。

2.1. TFIDF 算法

Salton 在文献[8]中提出了 TFIDF(Term Frequency & Inverse Documentation Frequency)算法。TFIDF 算法是由：词频算法(Term Frequency, TF)与反文档词频算法(Inverse Documentation Frequency, IDF)互补结合而成。

TFIDF 算法是中文文本分类中一种常用的加权算法，用以评估特征词对一个文本的贡献度。值越大，则特征词对文本的贡献度越大，越能代表该文本。TFIDF 算法的基本思想是：如果某一个词或短语在某一文档中出现的频率高，并且在其他整个文档集合中很少出现，则认为该词或短语具有很好的区分类别的能力，能代表此类文本。TFIDF 的计算公式如下：

$$W_{ik} = tf_{ik} \times \lg\left(N/n_i + L\right) \tag{1}$$

其中，W_{ik} 表示特征项 i 在文档中的权重，tf_{ik} 表示特征项 i 在文档中出现的频率，N 为文档集合中总文档数，n_i 为出现特征项 i 的文档数，L 为一个常数，通常凭经验取值为：0.01 或 0.1，其存在否和取值对特征词提取结果影响非常小，在很多实际应用中都省略了 L，本文中取 L = 0.01，用来修正当 N = ni 时，出现 W_{ik} = O 的情况。

2.2. 改进的 KNN 最邻近算法

最初的 KNN 算法由 Cover 和 Hart 于 1968 年提出，该算法的主要思想为：根据传统的向量空间模型，文本内容被形式化为特征空间中的加权特征向量，即 D = D (T1, W1; T2, W2; ……, Tn, Wn)。对于一个给定的测试文本，计算它与训练样本集中每个文本的相似度，找出 K 个最邻近(最相似)的文本，根据加权

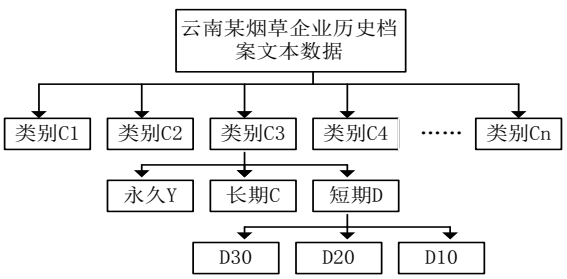

Figure 1. Structure of file text classification category
图 1. 档案文本分类类别结构图

距离和，判断测试文本所属的类别[9]。

KNN 算法的主要优点：1) 思路简单，易于实现；2) 当有新文本加入训练集时，无需重新训练；3) 该算法还能解决出现多峰值的情况；4) KNN 算法是分类效果最好的分类算法之一。

KNN 算法的缺点：1) K 值的确定；2) 时间和空间复杂度随着规模增加。

在本文的档案文本分类算法中，并不是每一个文本的分类都需要采用 KNN 算法来计算，大部分档案文本可以在不需要 KNN 的情况下进行分类，从而大大降低了 KNN 算法时间和空间复杂度，巧妙的克服了 KNN 算法的一个最主要的缺陷。

在传统的 KNN 算法中一般先设定一个初始的 K 值，然后根据实验测试的结果调整 K 值的大小。由于 K 的取值不能自动调整，且 K 取值不当或者训练文本分布不均匀都会影响分类性能，进而影响分类结果。为了降低 K 值对分类效果的影响，本文采用文献[10]中的改进算法——类内均值算法，解决了 K 值难以确定问题。

综上所述，将 KNN 算法巧妙的应用于本文所设分类计算法中，是最后的选择。

类内均值 KNN 算法的基本思想是：根据训练文本集中每个分类 Cj 所包含文本的数量，来确定不同的 K 值，即每个分类都有一个专属于自己的 K 值 Kj。从每个分类中选择 Kj(即该分类的 K 值)个与待测文本最相似的文本，计算它们之间的相似度并取其算数平均值，将待测文本分配到平均值最大的类别中。具体算法步骤如下：

1) 根据特征项集合重新形成文本向量。

2) 对于一个测试文本，根据特征词形成测试文本向量。

3) 计算该测试文本与训练集中每个文本的文本相似度，选出 Kj 个文本，Kj 的值根据每个类别的文本数量来确定，计算公式为：

$$Sim\left(d_i,d_j\right)=\frac{\sum_{k=1}^{M}W_{ik}\times W_{jk}}{\sqrt{\sum_{k=1}^{M}W_{ik}^2}\sqrt{\sum_{k=1}^{M}W_{jk}^2}} \tag{2}$$

其中：d_i 为测试文本的特征向量，d_j 为第 j 类的中心向量；M 为特征向量的维数；W_k 为向量的第 k 维。

4) 依次计算待分文本属于每类的权重平均值，计算公式如下：

$$p\left(x,c_j\right)=\frac{\sum_{i=1}^{K}sim\left(x,d_i\right)}{2} \tag{3}$$

5) 比较待分类档案文本与每个类别的权重平均值，将档案文本分到平均值最大的那个类别中。

3. 档案文本分类所用词典库的构造

为了实现计算机对档案文本的自动分类，笔者在对云南某卷烟厂历史档案文本数据的分析研究的基础上，慎重考虑了以下几个方面的问题，并根据这些思路结合实际情况，构造了相应的词典结构。

1) 主题信息的提取，即主题信息源的确定。我们通过对云南某卷烟厂大量历史档案文本数据的分析研究得出，本单位的档案文本数据与其他中文文本数据相比，具有其独特的性质：①绝大多数档案文本的题名都能反映档案文本的主题内容，同时，大部分档案文本的文号也能反映档案文本的主题内容，若再加上对档案责任者项的判断，则更能确保大部分档案文本都能得到的正确归类；②对于那些题名非常模糊不清或直接没有题名或既没题名也没文号或题名、责任者项和文号都没有的档案文本，为了确保档案文本的分类正确性，本文引入了 TFIDF 算法，通过该算法直接对文本内容进行分析、计算，提取最能

代表该档案文本内容的特征项，用于代表该档案文本。因此，本文将档案文本的题名、档案文号、档案责任者及从文本中提取的特征项信息确定为计算机分类的信息源。

2) 建立主题词与类号之间的关系。经对云南某卷烟厂历史档案文本数据分析研究发现，主题词与分类号之间可能存在一对一、一对多或多对多的关系，即一个主题词可能分属不同的类，一个类也可能拥有不同的主题词。

①一对一的关系，即一个主题词只属于一个类，如图 2 所示：

②一对多的关系，即一个主题词属于多个类别，如图 3 所示：

③多对多关系(一种交叉关系)，即一个主题词属于多个类别，一个类别也包含多个主题词，如图 4 所示：

因，主题词与分类号之间存在一对多或多对多的关系，所以，每一组关系中必然存在主次强弱之分，而如何正确表达这种主次强弱关系，成为能否将档案文本正确分类的一个关键性因素。

3) 主次强弱关系权重系数的计算方法。为了表明同一主题词与不同类号之间的主次强弱关系，本文规定了对主要关系给予较大的权值，次要关系给予较小的权值，即：

$$P(k,c)=p \quad (1\le p\le g) \tag{4}$$

其中，$P(k, g)$为主题词 k 与类别 c 的关系权重值，P 值的大小与主题词 k 和分类号 c 之间的主次强弱关系成正比。这样我们就可以在文本分类过程中，以不同类号的权值之和的大小来选择类号。本文所用算法权值分配尺度规定如下：

当主题词可以直接由档案题名、档案文号和档案责任者三者的独立或组合关系确定时，若某一个主题词特指为某类号，且只要该主题词出现，该档案文本就一定为这一分类号时，$P(k, g) = g$；若一个关键词和多个类号存在关系，根据主次强弱程度，分别给予 1，2，3 权重值；对于比较专指或词的长度较长的词，给予较高的权重值。

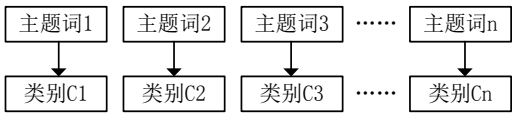

Figure 2. Keywords and categories one-to-one relationship diagram
图 2. 主题词与类别一对一关系图

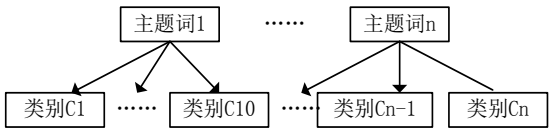

Figure 3. A one-to-many relationship between keywords and categories
图 3. 主题词与类别之间的一对多关系图

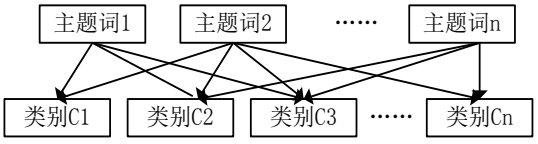

Figure 4. A many-to-many relationship between keywords and categories
图 4. 主题词与类别之间的多对多关系图

当主题词不能由档案题名、档案文号和档案责任者三者的独立或组合关系确定时，

由(1)式计算提取档案文本主题词，再由(2)式计算出待分配档案文本与训练集中每个文本的相似度，最后由(3)式计算出待分配档案文本属于每一类的平均权重值，进而得出主题词与类别之间的主次强弱关系。

4) 主题分类控制关系的建立。由于本文所用算法选择的分类信息源为档案文本题名、档案文本文号和档案文本责任者。而题名中的用词没有规范性，若将题名中的词均保存至主题词与类号的关系中，必然造成分类词典的庞大，增加了分类所用时间，对分类效果也会造成很大的不良影响。因此，我们所选择的分类用词均为规范词。例如，我们可以对“任职”、“任免”、“免职”、“聘任”、“免去”及“任命”等词赋予一个规范的主题词“任命”。这样分类词典库中就只存在“任命”一词来代替以上多个词。这样处理既能减小分类词典容量，又能减小分类用时，还能提高分类效果。但，在实际的档案文本题名中可能会出现与“任命”异词同义的词汇，如上面提到的“任职”等词，如果我们不能将“任命”与“任职”等词建立关系，必然导致分类的错误或失败。基于此问题，我们构造了一个分类前控词典，即将题名词与规范词建立联系。

5) 停用词典的构造。A) 档案文本文件的题名中存在大量没有分类价值的词汇。如，关于、通知、报告、汇报等。我们构造停用词典库将这类词聚集，以供分类时剔除这些词汇之用，进而提高分类效率。B) 在档案文本文件的文号中也存在大量对分类没有价值的词汇。如，表示日期的“某年某月”、表示文件号的“几号”文件等等词汇，同样我们将这些词聚集到停用词典库中。C) 根据实际情况，在停用词典库中还收集了其他一些对分类没有实际意义，但可能导致分类错误的词汇。停用词典库的建立，既加快了分类速度，又减小了分类错误的概率。

由以上分析，我们构造了 4 个分类用词词典库，见表 1 到表 4。

Table 1. Structure of weights of subject classification dictionary library table
表 1. 主题分类权值词典库表结构

名称	类型	长度	索引	说明
主题词	Varchar	30	Y	索引词可重复
分类号	Varchar	2		
权重值	Int	1		

Table 2. The author dictionary table structure
表 2. 责任者词典库表结构

名称	类型	长度	索引	说明
责任者	Varchar	30	Y	索引词可重复
分类号	Varchar	2		
权重值	Int	1		

Table 3. Structure of a control specification word dictionary library table
表 3. 非规范词控制词典库表结构

名称	类型	长度	索引	说明
非规范词	Varchar	30	Y	索引词不重复
规范词	Varchar	30		必须为规范主题词

Table 4. Structure of stopping using dictionary library table
表 4. 停用词典库表结构

名称	类型	长度	索引	说明
停用词	Varchar	30	Y	对分类无实际意义的词、短语

4. 档案文本分类用词切分算法设计

要实现档案文本的自动分类，必须要能够从档案文本相关信息(如：题名、文号、责任者及文本内容)中正确获取能够代表档案文本主题内容的特征项。那么，如何正确的从档案文本题名、文号及文本内容中提取分类用词，是我们在实现档案文本自动分类前遇到的一主要难题。为了获得正确的分词，我们的主要思路是：1) 对于题名或文号存在的档案文本，利用停用词库剔除停用词，再利用非规范词控制词库切分题名和文号中关键词，最后，以控制词库中的规范词作为切分结果；2) 对于题名和文号非常模糊或者不存在的档案文本，首先用式(1)计算并提取出能代表此文本的特征项，其次用停用词库剔除停用词，在利用非规范控制词库切分特征项中的非规范词，最后，以控制词库中的规范词作为切分结果。由于汉语词汇越长，其专指度越高。例如，“人事档案管理”就比“档案管理”具有更高的专指度，而且获得的主题词与分类号更为准确。因此，本文所采用的切分词原则为“词汇最长匹配原则”。具体分类用词提取算法步骤如下：

1) 对于一个给定的待分类档案文本，获取题名信息：

A、题名信息获取成功，题名字符串指针置 1；

B、题名信息获取失败，转到(3)。

2) 定位于题名指针位置，并取一个汉字，转到(8)。

3) 获取文号信息：

A、文号信息获取成功，文号字符串指针置 1；

B、文号信息获取失败，转到(5)。

4) 定位于文号指针位置，并取一个字符，转到(8)。

5) 档案文本预处理(进行文本分词等处理)。

6) 利用公式(1)计算并获取能代表档案文本主题内容的特征词，并将特征词字符串指针置 1。

7) 定位于特征词指针位置，并取一个汉字。

8) 将取出的字符以前方一致的要求搜索停用词典库索引：

A、有命中的停用词，将停用词与题名或文号或特征词匹配(以停用词的长度，从题名或文号或特征词指针位置开始与停用词比较)，匹配成功，取最长停用词从题名或文号或特征词中删除，指针位置移动到所剔除的停用词长度，回到(2)或(4)或(7)；匹配不成功，进入(9)。

B、无命中停用词，进入(9)。

9) 根据从题名或文号或特征词中取出的字搜索非规范词控制词典库(前方一致)：

A、搜索失败，则将指针移到两个字节，转到(2)或(4)或(7)。

B、搜索成功，则以匹配停用词的方法匹配关键词。

匹配成功：取最长匹配成功的关键词作为切分结果，当前指针位置移动关键词的一个长度，进入(10)。

匹配不成功：当前指针位置移动两个字节，进入(2)或(4)或(7)。

10) 通过切分出的关键词获取分类规范用词，并记下供分类时采用。转到(2)或(4)或(7)。

注：搜索成功以后的匹配方法为，只要词库欲匹配记录的首字与搜索值相同，均继续向下匹配，直

到两者不同或到文件尾部。获取主题词的切分的整个过程如图 5 所示：

5. 基于 KNN 最邻近算法的档案文本自动分类算法设计

针对档案文本的分类，主要包括两个过程：一、档案文本主题词的分析，获取能代表档案文本内容的主题词；二、档案文本的归类，即为给定的档案文本确定其所属的类号。在计算机分类的过程中，这两个步骤基本交融，没有明显界限。第一个过程已在上一节中完成，接下来将对档案文本的自动分类算法进行分析、设计研究。本文在档案文本的自动分类算法中巧妙的结合了 KNN 最邻近算法，提高了该分类算法自动、智能的性能，提高了档案文本分类的查全率和查准率，提高了档案文本分类的效率。算法具体步骤如下所示：

1) 获取分类号。经过上一节的档案文本分类用词的切分算法获得文本主题词后，首先利用文本主题词搜索分类权重词典索引，获取分类号及相应的权值。由第 3 节分析可知，主题词与分类号之间可能存在一对多和多对多的关系，因此，每一个分类用词都要扫描到搜索词与词典词不相等时，才能终止。

A、获取失败，进入(2)；

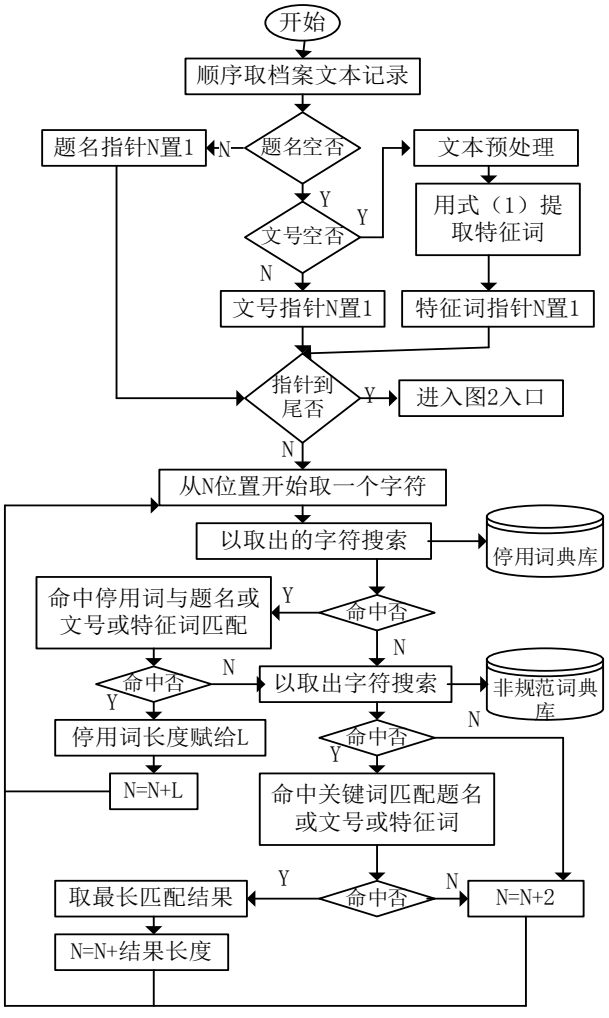

Figure 5. File word segmentation algorithm flow chart of text categorization
图 5. 档案文本分类用词切分算法流程图

B、获取成功，进入(4)

2) 根据式(2)计算待分类档案文本与训练样本集的相似度。

3) 根据式(3)计算待分类档案文本与训练样本集中最邻近的 K 个类别的平均权重值。

4) 进行分类号的合并与权值的求和计算。不同的分类用词所对应的分类号可能相同，也可能不同，同时一个分类主题词还可能对应多个分类号。所以，获取分类号终止后，我们需要将同一主题词所对应的所有类号进行合并，同时对其权重求和并排序(从大到小排序，方便分类号确定)。

5) 确定档案文本所属大类(这里用大类，是因为每个类别下还要对其所包含档案文本进行按保存期限的分类，故用大类来区分)。将已经取出的所有分类号进行分析比较，并借助于责任者项，确定档案文本所属类号。此过程主要包括以下几个方面：

A、获取的分类号唯一，则无需进行其他判断，直接分类，进入(7)；

B、类号不唯一，但仅存在一个权值最大的类号，这时通过该档案文本的责任者项搜索责任者分类词典库：

a、若取出的责任者类号与已有的类号相同，则进行定类，进入(7)；

b、若取出的责任者类号与已有的类号不相同，则将权值相加，若原有的权值大小顺序没变，则进行定类，进入(7)；若原有的权值大小顺序有变化，则将档案文本定类到当前权值最大的类中，进入(7)；

6) 如果最大权值相等的类号有多个，则删除权值较小的类号，再获取责任者项，并搜索责任者分类词典库，并求权重之和，再取最大值类号，进行定类，进入(7)；

7) 根据分类用词中获取的档案文本的特征向量，由式(2)计算档案文本与在大类下按保存期限正确分类训练集文本相似度；

8) 由式(3)计算档案文本所属期限类别权重平均值，并按权重值大小降序排序；

9) 将档案文本分配到权重值最大的保存期限类别中。

具体档案文本自动分类过程如图 6 所示：

6. 实验及结果分析

谈到算法的设计，则对算法是否高效或者是否优于其他同类算法的验证，是必不可少的过程。对本文所设计算法是否高效、是否优于其他同类算法的验证，主要是看用本文所设计算法，进行档案文本自动分类结果与人工进行分类结果进行比较，与人工分类的结果越接近，则分类的正确率越高。

文本分类的评价方法主要有以下几种：查准率(Precision，简写为：P)和查全率(Recall，简写为：R)、宏平均和微平均及 F 测度值，共三种。本文主要选用查准率和查全率评价标准来对本文所设计的分类算法进行评价，以查全率和查准率值百分值的高低，来判断本文所设计文本自动分类算法的优劣，值越高则算法越优。对查准率和查全率评价标准值的计算，一般采用二值列联表，其结构如表 5 所示：

如上表所示：A 代表系统和人工分类都属于此类的档案文本数；B 代表系统分类属于此类，而人工分类不属于此类的档案文本数；C 代表人工分类属于此类，但系统分类不属于此类的档案文本数；D 代表人工和系统分类都不属于此类的档案文本数。设 N 为总的档案文本数，且 N = A + B + C + D，则有：

Table 5. Binary contingency table
表 5. 二值列联表

	人工分类属于此类	人工分类不属于此类
系统分类属于此类	A	B
系统分类不属于此类	C	D

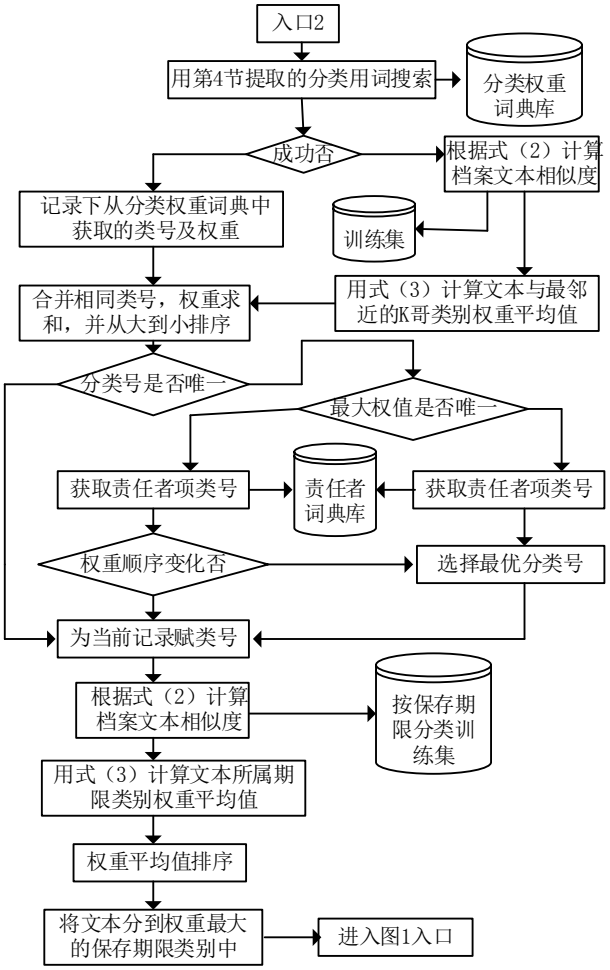

Figure 6. Flow chart of file automatic text classification
图 6. 档案文本自动分类算法流程图

查准率：$P=\frac{A}{A+B}$，指系统自动分类结果中，与人工分类结果相吻合的档案文本所占的比率。

查全率：$R=\frac{A}{A+C}$，指人工分类结果中，系统自动分类正确的文本所占的比率。

本文由于所属领域的特殊性，公用的语料库不适用。因此，文中用于实验的语料库是经过人工精心挑选出来的且将其分到正确的类别下。本文训练语料库共计 680 篇，共分为四个类别：党群 200 篇、行政 215 篇、经营 165 篇及生产 100 篇。

通过抽取 2006 年的部分档案文本共 2100 份进行测试，其中：党群 550 篇、行政 650、经营 500 篇、生产 400 篇。

由第二部分所述，Kj 的值需要根据训练文本集 Ci 中文本数量来确定，本实验中 Kj 的取值为训练集中文档总数的 10%，即：党群类中 Kj = 20、行政类中 Kj = 22、经营类中 Kj = 17、生产类中 Kj = 10。

本文所设计算法分类结果情况如表 6 到表 9 所示：

为了更好的验证本文所设计算法的高效性、优越性，本文还对传统 TFIDF 算法和传统 KNN 算法进行了实验，训练文本集同上，其结果入表 10 到 13 所示：

实验结果表明，本文所设计的档案文本分词提取算法和档案文本自动分类算法，在云南某烟草企业

Table 6. The table of party and the masses classification result
表 6. 党群分类结果表

	人工分属于党群类	人工分不属于党群类
系统分属于党群类	429	61
系统分不属于党群类	47	13
查全率	87.5%	
查准率	90.1%	
每条平均用时	约 1.5 s	

Table 7. The classification results table of administration
表 7. 行政分类结果表

	人工分属于行政类	人工分不属于行政类
系统分属于行政类	508	77
系统分不属于行政类	58	8
查全率	86.8%	
查准率	89.7%	
每条平均用时	约 1.5 s	

Table 8. The table of management classification result
表 8. 经营分类结果表

	人工分属于经营类	人工分不属于经营类
系统分属于经营类	408	43
系统分不属于经营类	46	3
查全率	90.4%	
查准率	89.9%	
每条平均用时	约 1.5 s	

Table 9. The table of production classification result
表 9. 生产分类结果表

	人工分属于生产类	人工分不属于生产类
系统分属于生产类	314	37
系统分不属于生产类	40	9
查全率	89.5%	
查准率	88.7%	
每条平均用时	约 1.5 s	

Table 10. The table of party and the masses classification result
表 10. 党群分类结果表

	人工分属于党群类	人工分不属于党群类
系统分属于党群类	429	61
系统分不属于党群类	47	13
查全率	86.0	
查准率	88.2%	
每条平均用时	约 2.1 s	

Table 11. The table of administration classification result
表 11. 行政分类结果表

	人工分属于行政类	人工分不属于行政类
系统分属于行政类	489	85
系统分不属于行政类	60	16
查全率	85.2%	
查准率	89.1%	
每条平均用时	约 2.1 s	

Table 12. The table of management classification result
表 12. 经营分类结果表

	人工分属于经营类	人工分不属于经营类
系统分属于经营类	393	52
系统分不属于经营类	53	2
查全率	88.3%	
查准率	88.1%	
每条平均用时	约 2.1 s	

Table 13. The table of production classification result
表 13. 生产分类结果表

	人工分属于生产类	人工分不属于生产类
系统分属于生产类	314	37
系统分不属于生产类	40	9
查全率	88.8%	
查准率	89.1%	
每条平均用时	约 2.1 s	

的档案文本分类中具有很高的效率。本文所设计算法在查全率和查准率上比使用传统的 TFIDF 算法和 KNN 算法提高了，特别是在分类时间上，本文所设计算法平均每条所需时间明显比使用传统方法提高了很多。

这是由于本文所设计的算法考虑了该烟草企业档案文本特殊性，大部分关键词都是从题名和文号中提取的，能更好的代表文本内容；同时，因减少了 TFIDF 算法和 KNN 最邻近算法的计算量，从而减少了分类用时。

在进行档案文本保存期限分类结果验证时，由于计算方式完全同上，此处不再重复累述。实验结果表明，其查全率和查准率都在 89%左右。

7. 结束语

本文基于对云南某卷烟厂历史档案文本数据的研究，对档案文本分类的两个主要过程进行了详细分析，并设计了相应的算法，给出了算法步骤，并在档案文本分类用词获取算法中应用了 TFIDF 算法，对缺少题名、文号和责任者的档案文本，也进行了主题词获取算法的设计，这样该档案文本分类用词获取

算就能灵活的应对于所有的档案文本。在档案文本分类算法中引入了 KNN 最邻近算法，解决了在最简单方式下无法将档案文本自动分类到其所属类别下的问题，智能化、自动化了档案文本的分类过程，减少分类过程对人的依赖性，人只要对部分分类不正确的档案文本进行修正，进一步提高了分类的查全率和查准率。且，本文所设计的算法还考虑了按档案文本保存期限的分类，进一步细化了分类结果，优化了对档案文本的管理。

基金项目

云南省软件工程重点实验室面上基金项目(2012SE306; 2011SE12)。

参考文献 (References)

[1] Luhn, H.P. (1957) A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, **1**, 309-317.

[2] Luhn, H.P. (1958) The automatic creation of literature abstracts. *IBM Journal of Research and Development*, **2**, 159-165.

[3] Luhn, H.P. (1960) Key word in context index for technical literature (kwic index). *American Documentation*, **11**, 288-295.

[4] Luhn, H.P. (1961) Selective dissemination of new scientific information with the aid of electronic processing equipment. *American Documentation*, **12**, 131-138.

[5] 苏新宁, 徐进鸿 (1995) 档案自动分类算法研究. *情报学报*, **3**, 194-200.

[6] 齐菁 (2012) 基于档案来源原则建立档案信息自动分类编目体系的思考. *湖北档案*, **2**, 19-21.

[7] 陈嵩 (2013) 简述档案管理理论的新发展. *城建档案*, **8**, 55-56.

[8] Salton, G. and Yu, C.T. (1973) On the construction of effective vocabularies for information retrieval. *ACM*, **10**, 48-60.

[9] Cover, T.M. (1968) Rates of convergence for nearest neighbor procedures. *Proceedings of the Hawaii International Conference on Systems Sciences*, 413-415.

[10] 刘辉 (2010) 基于 KNN 算法的中文 Web 文本分类技术研究. 硕士论文, 辽宁工程技术大学, 阜新.