面向卷烟质量评价的自动文摘系统设计

王 强[1],丁香乾[2],王 涛[3],周照艳[1]
(1.中国海洋大学 信息科学与工程学院,山东 青岛 266071;
2.中国海洋大学 信息工程中心,山东 青岛 266071;
3.山东中烟工业有限责任公司青岛卷烟厂信息处,山东 青岛 266100)

摘 要:基于.NET 平台,结合 SQL Server2005 数据库,设计了一个针对日常卷烟产品质量反馈意见汇总的自动文摘系统。系统的运行可以极大提高烟厂分析人员的工作质量和效益,减少差错,减轻劳动强度,提高市场测试评价人员的工作效率。因此,面向卷烟质量评价的自动文摘系统,可以作为分析市场测试评价信息的有效工具,在实际中也得到了良好的应用。

关键词:卷烟质量评价;.NET 平台;自动文摘系统
中图分类号:TP311 **文献标识码**:A **文章编号**:1674-7720(2013)23-0010-03

The design of cigarette-oriented quality evaluation system of automatic summarization

Wang Qiang[1], Ding Xiangqian[2], Wang Tao[3], Zhou Zhaoyan[1]
(1.Department of Information Science and Engineering, Ocean University of China, Qingdao 266071, China;
2.Center of Information Engineering, Ocean University, Qingdao 266071, China;
3.Qingdao Cigarette Factory Information Service, China Tobacco Shandong Industrial Co., Ltd., Qingdao 266100, China)

Abstract: The paper based on .NET platform, combine with SQL Server2005 database to design the system of automatic summarization for daily cigarette product quality feedback and summary. The operation of the system, can greatly improve work quality and efficiency of the tobacco industry analysts, reducing errors, reducing labor intensity, so as to improve the efficiency of market test evaluation workers.Therefore, the cigarette-oriented quality evaluation system of automatic summarization, can be used as effective tools for analyzing market test evaluation information, also got good application in practice.

Key words: cigarette quality evaluation; .NET platform; automatic summarization system

随着经济的飞速发展,企业的竞争也越来越激烈,烟厂的卷烟产品质量的及时反馈与汇总对烟草企业的发展有着重要的作用。目前,卷烟产品的质量评价主要依靠评吸专家经验通过品尝来完成,其评价结果受个人的主观爱好、生理差异和心理状态等因素的制约,由于评吸人员比较多,描述比较复杂,人为的分析和整理耗时比较长,无法及时地对日常卷烟产品质量反馈意见进行汇总。因此,研究一种快速从消费者对卷烟质量综合评价数据中挖掘质量优劣信息的方法,是卷烟数字化设计发展的需要。

为了满足烟厂企业的需求,本文基于.NET 平台,结合 SQL Server2005 数据库,基于三层结构的思想,设计了一个针对日常卷烟产品质量反馈意见汇总的自动文摘系统。与传统的人为分析和整理相比,该系统功能完善、性能稳定、可移植性高,可以极大提高企业市场测试评价人员的工作效率,减少差错,减轻劳动强度。因此,面向卷烟质量评价的自动文摘系统,可以作为分析市场测试评价信息的有效工具,在实际中也得到了良好的应用。

1 系统结构

1.1 Visual Studio.NET 技术

面向卷烟质量评价的自动文摘系统采用.NET 技术

架构 C# 设计。现阶段.NET 平台主要由以下几部分组成：Windows.NET、.NET 框架、Visual Studio.NET、.NET 企业服务器、Web 服务和.NET 应用以及模块构建服务[1]。Windows.NET 是指 Windows 操作系统的下一代产品，.NET 框架运行于该系统之上，提供对.NET 框架应用的运行支持。Visual Studio.NET 则是开发.NET 框架应用的集成开发环境。在.NET 框架的更上一层，是具体的应用和微软公司为.NET 平台提供的服务，包括 Web 服务、企业服务器和模块构建服务等。

.NET Framework 的诞生解决了许多开发人员多年来一直困扰的问题，并提供了这些问题的解决方案。每一种编程语言都有自己独特的地方，如它们可能是强类型的，有垃圾回收机制、基于例外的错误处理，或是以虚拟机方式运行，以及拥有强大的类库[2]。Visual Basic、Powerbuilder 以及 C++标准模板库（STL）或是其他语言都有一些这样的特性。然而，Java 语言以及基于 Java 的 J2SE 和 J2EE 框架表现得最为出色，以至于常常有人将 Java 和微软的.NET Framework 相提并论。现在微软正在将最好的特性融入自己的产品中，这其中包括支持多种语言的.NET Framework。微软所做的一切，将在它未来的开发语言和工具中得到体现[1]。

1.2 体系结构

本系统使用三层架构，通过.NET 平台可以快速方便地部署。显示层放在 ADO.NET 页面中，数据库操作和逻辑层用组件来实现，如图 1 所示。

图 1 三层架构结构图

2 系统设计

2.1 数据库结构设计

系统数据操作采用 SQL Server 2005 数据库系统。根据系统需求分析，设计了 4 个表结构，具体如表 1、表 2、表 3、表 4 所示。

表 1 分词词表

名称	数据类型	作用
W_ID	bigint	编号
Words	Varchar	词语

表 2 名词指标词词表

名称	数据类型	作用
N_ID	bigint	编号
Nounword	varchar	名词指标词
Freqs	int	词频

表 3 情感词词表

名称	数据类型	作用
E_ID	bigint	编号
EmotionalVocabulary	varchar	情感词
Freqs	int	词频
Polarity	int	情感极性

表 4 停用词词表

名称	数据类型	作用
_ID	bigint	编号
StopWord	varchar	停用词

其中，分词词表是词汇量足够大的一个中文词典，系统使用参考文献[4]提供的一个中文分词词表[5]。名词性指标词词表用来存储日常卷烟产品质量反馈意见汇总表中出现的名词信息，如刺激性、香味、余味、烟气、口感、杂气、外观质量等，这些名词是重点评价分析卷烟质量的指标项。情感词词表用来存储评吸人员对卷烟产品质量描述的情感评价信息，包含一篇文档中出现的情感形容词、词频及情感极性。其中，形容词为主关键字，包括较好、适中、较差、较小、较浓等。每个形容词在对卷烟的名称指标词的评价情感下都对应了一个极性。其中，0 代表中性，1 代表褒义，-1 代表贬义。停用词表里存放着一些虚词、连词等无实际意义的词，以便在进行分词操作时将文档中含有的停用词表中的字、词去掉，以减少不必要的资源浪费，提高分词速度。这些基础信息为提取和生成评价信息的摘要提供了有效的数据。

2.2 系统的功能设计

随着传统的面向通用领域的自动文摘技术[6]的日趋成熟，越来越多的目光转向了针对特定领域的、更加个性化的自动摘要技术，以满足更加丰富的需求。本文设计了面向卷烟质量评价这一特定领域的自动文摘系统。该系统主要包括 5 个模块：文本预处理模块、文档分词模块、加载词库模块、词频统计分析模块及摘要生成模块。系统的结构模型如图 2 所示。

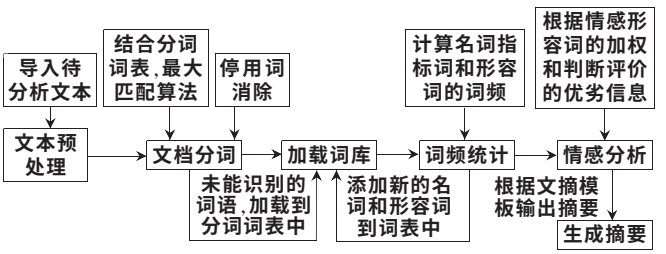

图 2 自动文摘系统结构模型图

系统各模块具体的功能如下：

（1）文本预处理模块：将待分析文本信息按照一定的标准格式导入，即按照计算机能够识别的形式导入文本信息。为保证文本标识的准确性，在进行文档处理时，统一使用全角的标点符号。

（2）文档分词模块：此模块为摘要系统的首页所显示的内容，导入的待分析文档信息将在窗体中显示出

来。首先对待分析的文档信息进行分词处理，并把文档中的停用词等一些非重要词剔除，分词结果中每个词中间用空格隔开，在分词结果中提供了未能识别的候选新词，在加载词库模块中可以把未能识别的候选新词添加到分词词库中，从而提高文档分词的效率。

(3)加载词库模块：由于文档分词模块不可能达到100%的分词准确率，不可避免地会出现不能识别的新词，在此模块中，可以把这些未能识别的新词加载到分词词库中，可以随时添加及更新分词词库，从而提高分词的准确率；在该模块中，还可以添加评价卷烟质量的名词指标词和形容词性的情感词。

(4)词频统计分析模块：在该模块中，根据标点符号将待分析的文档划分成一个个的句子，对每个句子进行分词，根据名词库统计待分析文档中出现的名词频率，并把与名词相关的句子显示在单独的列表中，然后统计每个名词对应的句子中形容词的词频，显示词频统计的相关信息。

(5)摘要生成模块：通过对文档的统计分析，系统可以自动计算分析出每个包含名词指标词的句子中，各名词指标词的情感值的加权和，系统自动组合各名词指标词和其对应的情感形容词，从而得到评价摘要。

3 系统实现与应用

3.1 最大正向匹配算法的C#实现

本系统采用最大正向匹配分词算法[7]，该算法复杂度比较小，技术实现比较容易，分词效率高，以下是程序中实现正向最大匹配算法的部分关键代码：

```
int maxlen=8;          //最大词长为8字符(即4个汉字)
string Separator="";                    //分词结果以空格隔开
private ISegmentDictionary SegmentDictionary=ne
        w ForwardSegmentDictionary();//定义分词词典
public void Segment(string text,StringBuilder result)
                        //对text进行正向最大匹配分词
{while(!string.IsNullOrEmpty(text))
                        //文本text不空则循环分词
{int len=text.Length;                   //文本的长度
int subLen=maxlen;                      //从最大词长开始匹配
string strWord="";
while(len>2)                            //文本中有词则循环
{strWord=text.Substring(0,subLen);
subLen=subLen-2;
if(SegmentDictionary.Contains(strWord))         //匹配词典
{result.Append(strWord);
result.Append(Separator);
                        //把匹配的词语添加到分词结果中
break;}}
text.Remove(0,subLen);//把匹配的词语从文本中除去
return result;}}
```

3.2 系统实现

本系统采用C/S架构，可以在Windows 2003,Windows XP,Windows 7操作系统平台上运行，本文在.NET平台下，采用Visual C#开发语言、Microsoft SQL Server2005后台数据库，ADO.NET进行开发的Windows Form应用程序。在日常卷烟产品质量市场测试评价的实际业务中，该系统包含了使用简单的图形用户界面，可以作为分析市场测试评价信息的有效工具。另外，本系统测试过山东中烟工业公司提供的1 000条左右的消费者评价数据信息。下面以部分评价数据信息为例，通过该系统得到摘要信息，系统运行界面图分别如图3、图4、图5所示。

图3 系统预处理、分词界面

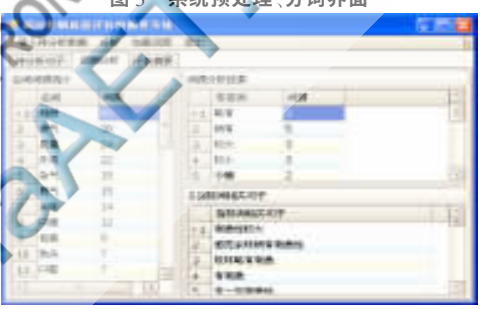

图4 词频统计分析界面

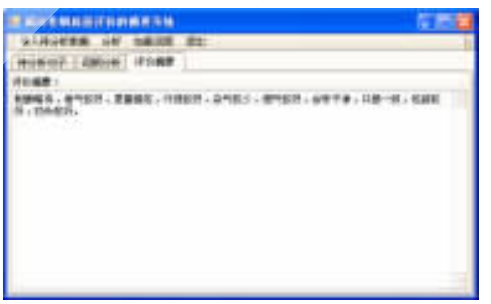

图5 摘要生成界面

在系统预处理、分词界面中，可以看到分词结果以及未能识别的候选新词，这些候选新词能够加载到分词词库中；在词频统计分析界面中，向用户展示各名词指标词和形容词的词频，并按照从大到小排列，与名词指标词有关的句子也显示在该界面中，方便用户分析判断评价结果；在摘要生成界面中，系统自动分析得到摘要结果。

在系统性能方面，人为的分析消费者对卷烟质量综

合评价数据，消耗大量的时间和精力，系统能够快速地得到数据的统计分析以及摘要结果，大大地提高了市场测试人员的工作效率。

本文从烟草企业的实际应用出发，面对人为分析和整理卷烟产品质量耗时比较长这一问题，设计了一个针对日常卷烟产品质量反馈意见汇总的自动文摘系统。该系统基于.NET 平台，采用三层架构，结合 SQL SERVER 2005 数据库技术，可以很好地实现程序员并行开发，提高程序的开发速度。为了验证所提出方法的可行性和有效性，本文采用内部评价方法对开发的文摘系统进行评估。从山东中烟工业公司提供的卷烟质量综合评价数据中获取文摘，进行评测，可以看出本文提出的面向卷烟质量评价的自动文摘系统，能够满足烟厂对日常卷烟产品质量反馈意见及时汇总的需求，与传统的人为的分析方法相比，极大地提高了烟厂分析人员的工作质量和效益，减少了差错，减轻了劳动强度。

综上所述，本文设计的针对日常卷烟产品质量反馈意见汇总的自动文摘系统，可以作为烟草行业分析卷烟产品质量反馈意见的有效工具。

参考文献

[1] 吴杉杉，宋小倩. .NET 框架介绍和 WinCE 开发环境搭建[J].中国新技术新产品，2011(6)：95.

[2] 付明柏.基于.NET Framework 的软件复用技术研究[J].软件导刊，2013(5)：15-17.

[3] http://www.mandarintools.com/segmenter.html

[4] Feng Haodi, Chen Kang, Deng Xiaotie, et al. Accessor variety criteria for chinese word extraction[J]. Computational Linguistics，2004(30)：75-93.

[5] 程娟.中文文档自动摘要技术[D].济南：山东大学，2006.

[6] 吴旭东.正向最大匹配分词算法的分析与改进[J].科技传播，2011(20)：164-165.

（收稿日期：2013-07-06）

作者简介：

王强，1988 年生，硕士研究生，主要研究方向：软件开发，计算智能。

丁香乾，1962 年生，博士生导师，教授，主要研究方向：计算智能，软件工程，数字家庭，制造业信息化。

王涛，1976 年生，工程师，本科，主要研究方向：烟草信息化。