



语音大数据信息处理架构及关键技术研究

杨震,徐敏捷,刘璋峰,秦达,姚晓辉

(中国电信股份有限公司上海研究院 上海 200122)

摘要:针对如何从音频信息中获得有商业价值的信息这一大数据研究的重要方向,首先分析了大数据的基础技术特性、涉及的技术,然后分析了语音识别的基本技术环节,最后结合大数据信息处理架构,给出了语音识别技术与大数据相结合的应用开发技术架构及应用流程。

关键词:自然语言处理架构;大数据;自动语音识别

doi: 10.3969/j.issn.1000-0801.2013.11.001

Study of Audio Frequency Big Data Processing Architecture and Key Technology

Yang Zhen, Xu Minjie, Liu Zhangfeng, Qin Da, Yao Xiaohui

(Shanghai Research Institute of China Telecom Co., Ltd., Shanghai 200122, China)

Abstract: As the major part of big data research, the audio frequency's value has not been discovered. The features of big data, and the automatic speech recognition (ASR) technology were studied, and the value was mined through the method that combined with the two technologies. Finally, the application development technical architecture and application process of audio frequency which combined with big data were given.

Key words: natural language processing architecture, big data, automatic speech recognition

1 引言

信息主要存在3种表达方式:文本(文字)、音频(声音)、视频(图像),对于这3种信息表达方式的理解和应用,是信息处理系统的基本研究内容。信息从产生起就具备各种属性,对这些属性进行理解、开发和利用成为各类成功企业的基础。例如,Google利用互联网信息的链接关联性模型完成了搜索引擎的开发应用;Facebook、新浪微博研究信息的社会属性,构建了类似人类社会人际交流的信息流动平台。这些企业从信息的理解、认知、表达、应用过程中找到了具体理论模式的实际应用方式,因此获得了成功。随着计算机硬件计算能力、网络通信能力的提高,传统学术界一直致力于研究的大规模文本处理及挖掘、复杂

系统、语音识别、本体论等理论及技术开始获得在实际中应用的物质基础。

随着3G技术的推广应用,运营商的竞争更加激烈,产品同质化更加严重,运营商的竞争由单纯的业务的竞争,转变为依托技术和服务发展的深层次竞争。从某种意义上来说,这种深层次的竞争,就是对新技术的掌握和灵活运用的竞争;是依靠运营商规模优势,在现有产品、技术架构上引入新技术的竞争。这种竞争性产品或技术革新可以从以下几个方面评价。

- 依托新技术开展基于新技术特性的新产品开发。
- 在原有产品及技术架构上的新技术应用具有原有产品、营销方式所没有的效果,为原有产品增加了新的特性。
- 新技术节约了人工、服务成本,提升了劳动生产率。



现今,大数据的研究逐渐引起学术界和运营商的重视。如何从新技术的开发和应用角度在现有业务系统中引入大数据处理技术,使学术研究的成果转化成为实际的商业价值,是值得探索的问题。而音频作为信息表达的基础方式之一,如何从音频信息中获得有商业价值的信息,是大数据研究的重要方向。

2 语音大数据及应用研究

目前大数据的研究及处理基本涉及以下两个方面。

- 大数据处理技术架构,如起源于 Google 的 Hadoop 数据处理架构,解决的是超大规模数据集面向需求的计算处理问题;
- 面向业务及应用本身的特征提取及分析技术,研究这些特征提取及分析技术如何在超大数据集上按照预设的模型实现。

相较而言,根据业务需求如何处理信息,使其符合业务系统的要求,进而产生商业上的价值,是大数据研究和应用的关键。

2.1 大数据处理技术

大数据的特点可以总结为 4 个 V,即 volume(体量浩大)、variety(模式繁多)、velocity(生成快速)和 value(价值巨大但密度很低)。大数据处理技术是面向海量数据信息处理和应用而提出的,处理技术需满足以下几点需求^[1-3]。

- 高度可扩展性。要求横向大规模可扩展,具备大规模并行处理能力,处理系统可以方便地根据计算需求进行横向扩展,以支持快速增加的业务数据处理要求。
- 高性能。即快速响应复杂查询与分析,除了需要在数据库体系架构、数据处理模型上进行深入研究外,还需在处理机制上引入自然语言处理的模型和算法,使处理更加面向应用和自然语言状态下的信息检索等。
- 其他要求还包括:高度容错性、支持异构环境、较低的分析时延、易用且接口开放等。

而大数据处理与信息处理及信息获取相结合的应用,主要包括:大数据聚合、分布式存储、计算技术;大数据的实时索引和实时检索技术;大数据的自动分类、内容聚类、主题抽取、热点事件发现、倾向性分析技术;大数据面向业务的融合应用技术。这些应用的完成也必须应用自然语言处理技术的最新成果。

2.2 语音大数据的价值

语音大数据指个人或企业在生产经营活动中产生以音

频为载体的信息资源,广泛存在于各类传统呼叫中心、互联网、移动互联网等各类业务系统中。相比以文本为载体的信息,这类信息目前的应用研究还不充分。而在各种语音大数据中,呼叫中心存储的语音数据最具备研究和挖掘价值,可以为企业生产经营活动提供有价值的帮助,本文即以语音大数据为例进行分析,经过归纳,其具备以下优点。

- 价值密度高。呼叫中心语音大数据的价值密度高于目前所有已知的大数据资源。因为呼叫中心解决企业在产品运营中的服务问题,包含用户对企业生产经营活动的所有看法、用户在使用企业产品过程中的所有问题,从中可以挖掘出大量有用的信息。
- 使用方便。由于国家政策法规的要求,呼叫中心语音大数据都是以一定的格式进行保存,在具体的应用研究中,不存在来源、格式不统一的情况。
- 存在一定的信息标注。呼叫中心语音大数据除音频本身外,还包含其产生的时间、大概主题(来源于呼叫中心的电话小结)、产生者标记(如拨打者和座席服务者)、大概质量评价(如服务完成后用户的评价)等。
- 存在对应的以文本为载体的知识内容对应关系。呼叫中心语音大数据基本都是围绕呼叫中心知识库中存储的服务内容产生的。虽然没有明确定义,但通过记录座席在服务过程中的浏览轨迹,基本能获得其与用户对话过程中的音频与其正在浏览信息之间的一个对应关系,而对这个对应关系的研究还没有开展。

2.3 语音大数据需解决的问题

通过对这些以音频形式存在的大数据进行分析和挖掘,可以形成各类新的应用。以呼叫中心语音大数据作为具体的实例分析,通过语音大数据分析技术分析语音文件中的关键词、情绪、情感等,通过对这些特征进行统计及专业化分析可以完成以下功能。

- 座席预质检:可用于呼叫中心服务质量提升。传统的呼叫中心质检由人工质检完成,具备高级技能的质检人员对呼叫中心每天产生的大量录音进行规制抽取,之后评价每个抽取录音的服务情况,对服务人员提出改进建议。但是由于成本的限制,一般只能做到 0.5%~1% 的抽检率。通过语音大数据挖掘的方法,可获得服务质量不高的服务录音模型,通过这个模型对语音大数据进行预处理,使抽检的准确程度更高,抽检率更高,进而提高呼叫中心的整体服务水平。

- 热点信息挖掘:通过对呼叫中心一段时间内的录音文件进行分析和挖掘,可以获得某一个时间段内出现频次最高的关键词或信息概念,得到当前用户所关注的热点问题。
- 新产品市场评价:通过对呼叫中心一段时间内的录音文件进行分析和挖掘,可以分析某一个主题下用户关注的内容、反馈,进而得到企业推出新产品的市场评价报告。
- 企业形象用户评价分析:通过对企业产品相关音频大数据的分析,可以获得企业所推出产品、整体形象、市场认可、用户评价等统计指标。
- 营销机会:呼叫中心在对用户进行服务的过程中,针对用户的需求,可以发现企业经营产品的潜在用户,并可以通过与 CRM 相结合,发现潜在的、新的营销机会。
- 竞争情报:呼叫中心语音大数据中,通过有针对性的分析整理,还可以挖掘出有关竞争对手的信息,如用户提到竞争对手的产品功能更完备、费用更加低廉等。

对于语音大数据的处理技术发展,在业界也处于刚起步的阶段。以上信息的整理、统计、提炼,传统上需要耗费大量的人工时间及经济成本,如果能自动地在录音数据中进行挖掘,哪怕并不十分完备,都将对企业的生产经营活动产生有益影响。目前该领域主要关注的技术有语音大数据信息的实时处理、基于大数据集的语音识别、模型训练、语音文件热点信息感知和知识提取、基于内容理解的音频挖掘等关键技术。如果要达到较好的分析效果,各种统计分析所对应知识体系表达及分析体系也需要建立,面向应用的知识本体表达和研究也需要建立,并进行应用完善。

3 语音大数据研究及开发的关键技术

音频数据作为大数据重要的组成部分,亟需认真研究和挖掘。因此语音识别技术是解决语音大数据实际应用问题的重要技术。为达成语音大数据的分析目标,必须对语音识别技术的实现方式、技术架构进行分析,同时归纳整理语音大数据的分析目标,反作用于语音识别技术的研发体系,使底层的基础算法更加面向业务实现的研究和演进。

3.1 语音识别技术

科研工作者从 20 世纪 50 年代开始就进行语音识别技术的研究。AT&T-Bell 实验室实现了第一个可识别

10 个英文数字的语音识别系统(Audry);60 年代,动态规划(DP)和线性预测(LP)分析技术,实现了特定人孤立词语音识别;70 年代、80 年代语音识别研究进一步深入,HMM 模型和人工神经网络(ANN)在语音识别中成功应用;90 年代后,语音识别在细化模型的设计、参数提取和优化、系统的自适应方面取得关键进展,语音识别技术开始真正走向商业应用^[4-8]。从技术角度归结语音识别的应用有以下几类。

- 中小词汇量、孤立词识别系统。系统以词语为基元建立模板,没有次音节、音节单元,也没有上层的语句语义层,每个词条命令就是识别的最终结果。这种系统可以认为语音、语言的知识都包含在以词组为单元的模板中。电信的识别系统如 AT&T 用于电话查询的系统。
 - 以词语为识别基元、连续或连续词的语音识别系统。系统为每一词条建立模板,最终任务是按一定的语法规则将词语识别结果依次连缀成句子,这类系统往往用于特定任务(航班查询、电话查询等),具有明显的语句识别层次。
 - 以全音节为基元模型建立的识别系统。使用算法逐次获得前 N 个最好的候选单元(无调、有调音节),再按词性、句法、语法网络信息得到最后识别结果。这种方案多用于汉语大词汇量、连续语音识别系统。
- 语音识别技术架构主要由以下几部分构成。
- 物理接口层:声音进入系统的物理接口,输入语音信号。
 - 特征提取层:提取声学特征矢量,提供特征矢量序列。
 - 音节感知层:声韵母因素单元结构,提供音节候选序列及可信度,把声韵母或因素合并成为音节单元,推断合理音节,提供词语候选序列及可信度。
 - 词语识别层:音字转换,推断词语单元,提供语句候选序列及可信度。
 - 语句识别层:推断语句候选单元及可信度。
 - 语义应用层:分析语义,映射应用,由任务语法约束。

以上从逻辑层面分析了语音识别具体技术应用的几个层次,具体到与业务结合,即系统如果提供语音识别某一类业务的实例应用时,还需要针对这个业务领域的基本语料素材,以实现具体应用领域的语言模型。

3.2 基于语音识别进行语音大数据分析的关键技术

(1) 文本转写

即语音、音频信息转换文本的过程,是所有分析的基



础。语音识别文本转写的准确程度与语言模型密切相关,需要完成具体所涉及的专有名词、术语的语料素材收集,并在此基础上构建有针对性的语言模型。

(2) 关键词提取

从本质上看这项功能与文本转写十分类似,但为了提高处理速度及准确性,系统可以只完成一些配置的关键词,只针对这些关键词的出现位置(时间点)、频次进行统计,并不需要进行完整的文本转写。

(3) 声纹识别

需要完成语音大数据中不同角色的区隔,与文本转写相结合,可以在区分对话者的基础上,了解不同对话者的对话内容。声纹识别技术具体的应用还有说话者确认、说话者辨认等。

(4) 语音情绪识别

根据目前的研究结果,基音频率(pitch or fundamental frequency, F0)可以作为识别情绪的主要声学特征,其他的一些特征还包括能量、持续时间、语速等。综合来说,情绪对语音的影响主要表现在以下3个方面:基音曲线、连续声学特征、语音品质。基音曲线主要用来描述基音序列的几何分布;连续声学特征包括基因的大小、能量、说话速率、能量在频谱上的分布等。语音品质包括松紧度、粗糙度、有无带呼吸声^[9]。这3种语音品质的类型在某种程度上是相关的。在相对理想的条件下,语音情绪识别涉及的各类参数都是可测量的,可以对底层的语音识别引擎功能模块进行独立封装,这样业务系统在获得各类参数后就可以进行标准计算,获得业务系统所需的基础数据。

(5) 语义理解

事实上把语义理解技术作为语音识别技术的一个子集并不合适,本文为了面向业务应用语音大数据处理体系架构的完善,将其归为实现语音大数据的一个环节。另一方面,在文本转写的过程中,为了实现较高的转写准确程度,已经应用了基本的语义理解技术,实现连续语音的准确识别。在语音大数据的开发过程中,为了准确地挖掘出语音大数据的特征,必须有面向业务领域的语义理解技术,以解决针对同一对象的不同描述问题,即解决特征的归类和聚类问题。

3.3 面向语音大数据的技术处理架构

业界针对海量数据进行处理的技术架构已经进行了充分研究,并有大量实践案例。从技术特征来看主要分为两个层次,一个是面向海量数据的操作,应用系统如

何对大数据集进行面向业务应用的底层数据操作、存储、归并、清洗、转化;另一个是如何应用先进技术发现大数据的特征价值,其可以与第一个层次有限度融合,也可以在第一个层次基础上针对已经形成的数据集进行处理,处理结果是方便业务系统进行调用、查询、展现,或分析系统更有效地提取数据特征,进行相应的分析。本文主要关注第二个层次,即在语音大数据中如何发现业务系统所需的特征,挖掘大数据中的价值,如图1所示。

此构架的思路,是把语音识别技术(含语义理解及文本挖掘技术等)细分并模块化,通过定义针对语音信息的处理目标定义,使其能服务于业务需求,并适应大数据的处理架构。从体系架构上分为五大部分。

- 语料部分:分为语料资源库及服务资源库,存储语音识别的语言模型及语义理解特征提取、语义聚类、语义归类所需的行业语料。
- 基础能力层:语音识别及语义理解的细分模块,提供标准的输入输出调用接口及相应参数定义。
- 能力组合层:把能力层的语音识别、语义理解各类细分能力模块分别组合,形成不同的标准调用服务接口,针对特定的服务打包特定的能力。
- 业务封装层:适应各类调用需求、访问方式的再封装。
- 调用管理部分:整体平台对外提供能力的管理及维护。

架构的核心是把语音大数据需要处理的各类基础能力进行模块化区分,并定义各类模块化对外服务接口,使语音大数据的处理更加面向应用的软件系统、分析系统的业务需求,使大数据中蕴含的价值能被充分挖掘。需要说明的是,语义理解技术在大数据挖掘中也是核心技术,事实上单纯的语音识别技术如果不与语义理解技术进行充分融合,语音大数据挖掘及应用的效果将大打折扣。

4 结束语

本文分析了大数据和语音大数据的特点,结合目前的应用方向,重点研究了语音大数据的关键技术体系架构,提供了语音大数据的处理流程和应用方向。未来的研究方向在于语音大数据底层可检测分析变量和特征的算法方法,把其标准化,进一步结合语义处理技术,完善语音大数据识别内容面向业务层次的理解和识别。同时探索语音大数据特征提取分析技术面向大数据集的处理效率,是否可以形成分布式计算的处理架构,可以并行地处理

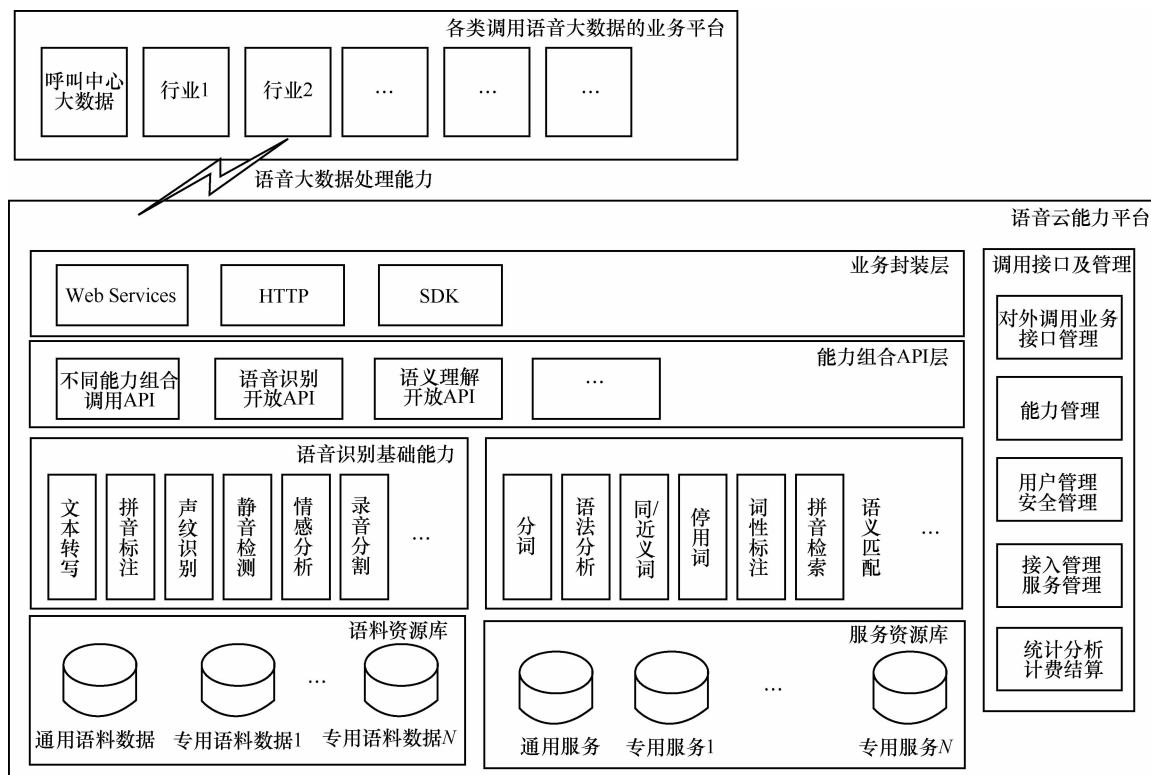


图1 语音大数据处理基础架构

超大数据集,快速提取出业务系统所属的数据特征。在业务层次,需要完善业务系统分析调用的规则,完善面向业务和挖掘需求的报表系统设计,使大数据蕴含的价值能充分体现。

参考文献

- 1 吴广君,王树鹏,陈明等.海量结构化数据存储检索系统.计算机研究与发展,2012,49(S1):1~5
- 2 元开元,赵卓峰,房俊.针对高速数据流的大规模数据实时处理方法.计算机学报,2012,35(3):477~490
- 3 Chang F, Dean J, Ghemawat S, *et al.* Bigtable: a distributed storage system for structured data. ACM Transactions on Computer System (TOCS), 2008, 26(2): 1~26
- 4 倪崇嘉,刘文举,徐波.汉语大词汇量连续语音识别系统研究进展.中文信息学报,2009,23(1):112~123,128
- 5 Jelinek F. Continuous speech recognition by statistical methods. Proceedings of the IEEE, 1976, 64(4): 532~556
- 6 颜永红.语言声学与内容理解研究进展.应用声学,2012,31(1): 35~41
- 7 邵阳,王岚.语音识别技术概述.先进技术研究通报,2009,3(2): 20~24
- 8 马俊.语音识别技术研究.哈尔滨工程大学硕士学位论文,2004
- 9 张盛.汉语语音情绪识别.中国科技大学博士学位论文,2007

[作者简介]



杨震,男,博士,中国电信股份有限公司上海研究院高级工程师,主要研究方向为大数据、人工智能、模式识别、自然语言处理、信息搜索技术、系统分析与综合建模、数据挖掘、知识管理、Agent 技术。

徐敏捷,男,中国电信股份有限公司上海研究院主任级工程师,主要研究方向为行业信息化应用、大数据。

刘璋峰,男,中国电信股份有限公司上海研究院主任级工程师,主要研究方向为移动网络平台、移动数据业务规划开发、项目管理。

秦达,男,中国电信股份有限公司上海研究院高级工程师、信息技术部主任,主要研究方向为大数据、云计算。

姚晓辉,男,中国电信股份有限公司上海研究院工程师,主要研究方向为数据规划、数据挖掘、信息管理。

(收稿日期:2013-10-29)