利用背景知识提高 web 语音浏览中的识别精度的方法

李红莲[1],王春花[2],袁保宗[1]
(1.北方交通大学信息科学研究所,北京 100044;2.北京三星通信技术研究所,北京 100081)

摘 要: 语音识别的精度不够高一直是阻碍语音技术得以广泛应用的瓶颈,在具体的应用中充分利用背景知识是解决此问题的一种有效方法.在 web 语音浏览中,用户的语音输入为某个有限集的元素之一,本文利用这个特点,首先定义了一种文本字符串之间的相似度,利用相似度对识别引擎的识别结果进行后处理,进而给出更准确的识别结果.实验结果表明,采用这种方法,语音识别的正确率能够达到 95%以上,为真正实现语音上网提供了有力支持.

关键词: web 语音浏览; 相似度; 语音识别与理解
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2002) 12-1836-04

Improve the Accuracy of Recognition in Web Speech Browsing Using Context Knowledge

LI Hong-lian[1], WANG Chun-hua[2], YUAN Bao-zong[1]
(1. *Institute of Information Science, Northern Jiaotong University, Beijing* 100044, *China*;
2. *Beijing Samsung Communication Technology Research Institute, Beijing* 100081, *China*)

Abstract: The accuracy of speech recognition is still a bottleneck to baffle the application of speech technology. The accuracy of recognition may be improved greatly by using context knowledge efficiently. In web speech browsing, user's speech input is usually one element of a finite set. Based on these observations, this paper first defines a kind of similarity between two Chinese text strings, then processes the recognition results of engine to acquire more accurate results. Experiments show that our approach is mostly efficient: the accuracy is improved from less than 60% to more than 95%.

Key words: web speech browsing; similarity; speech recognition and understanding

1 引言

语音技术与 internet 技术的结合使语音上网成为一个新的研究热点.目前世界上许多大的公司、研究所都已经开始了这方面的研究.如 AT&T, IBM, Lucent, Motorola 在 1999 年成立了 VoiceXML 联盟,开始共同研究制定语音上网的语音表示标准,已推出了 VoiceXML2.0[1].微软中国研究院也在从事这方面的研究[2,3].语音上网的实现将给人们带来巨大的便利,人与 internet 的交流将更自然、更富有情趣.尤其是如能使用便携设备通过语音上网,人们将获得更大的自由,达到随时随地上网.

不过从目前来看,实现真正意义上的语音上网还是人类一个美丽的幻想,理想和现实之间还有很大的差距.制约语音上网实现的一个瓶颈问题是识别引擎的识别精度问题,单纯靠通用的识别引擎(如 ViaVoice)来完成识别的任务是不太现实的.实验表明,当用于 Web 语音浏览时,ViaVoice 的完全识别正确率(即对每一次语音输入,识别没有任何错误的比率)在 60%左右,这将大大影响浏览的效率.因此,我们需要找到一种更有效的办法来提高识别的正确率.

在通常的 web 语音浏览中,识别功能完全依赖于识别引擎来完成,然后将识别结果与候选项作简单匹配[1],这样很容易出现只要识别结果稍有错误就不能完成任务的情况.

例如,在访问北方交大主页时,用户语音输入"院系设置",识别引擎识别结果为"宴席设置",如果用简单的匹配方法,系统将认为用户的输入无效.而人类却很容易判断出用户的原始输入(本意)是什么,怎样让机器具有人类的这种智能呢?

我们注意到这样一个事实,在进行语音浏览时,识别引擎的识别结果即使不正确,但与正确结果总有某种相似性.本文正是利用这个特性,定义了一种相似度,然后将识别结果进行后处理来达到正确识别理解用户的语音输入.

收稿日期:2002-01-24;修回日期:2002-05-10
基金项目:国家自然科学重点基金(No.69789301);国家 973 计划(No.G19980305011)

我们的方法可以描述如下:首先定义一种文本字符串之间的相似度,对于用户的每一次语音输入,计算出识别引擎的识别结果与当前网页上的每个超文本之间的相似度,把与识别结果相似度最大的超文本看成是用户所说的话(我们假定用户每一次语音输入的内容是当前网页上的某个超文本),然后连接到该超文本对应的网页.

本文第 2 节给出相似度的定义,第 3 节介绍系统的实现,第 4 节是实验结果及分析,第 5 节指出存在的问题及以后需要做的工作.

2　相似度的定义

文档集之间的相似度的定义很多[4].有 60 多种不同的相似度定义,其中经典的有[5]:

Jaccard 系数:$\frac{|A\cap B|}{|A|+|B|-|A\cap B|}=\frac{|A\cap B|}{|A\cup B|}$

Cosine 系数:$\frac{|A\cap B|}{\sqrt{|A|\cdot|B|}}$

Dice 系数:$\frac{2|A\cap B|}{|A|+|B|}$

其中 A,B 分别表示要比较的两个文档集.由于这些相似度是用于文档集的,用于字符串就不太合适,不过也有一定的参考价值.

常用的字符串之间距离的度量方法是编辑距离(从一个串变换到另一个串所需要的最少插入、删除、替换操作的数目),它有广泛的应用,如定义语音或扫描文字识别精度[6].但考虑到这种距离没有涉及到拼音,因此也不能很有效的处理我们面对的匹配问题.

Zhang Lei 等[7]通过模糊匹配来纠正键盘输入的错误,文中定义了两个中文字符串之间的距离.其距离定义考虑了字形、发音和输入码等几种因素.

我们注意到识别引擎的识别结果与语音输入的内容总有很多相似之处.例如,我们通过语音输入

"关注乔丹,关注英雄雕塑的复活"

识别引擎的识别结果为

"关注巧干,关注英雄雕塑的副和"

我们看到:"关注、关注英雄雕塑的"这些字与语音输入的本意是完全一样的,"巧(qiao)"与"乔(qiao)"、"副(fu)"与"复(fu)"虽然字形不同但拼音是完全相同的,"丹(dan)"与"干(gan)"、"活(huo)"与"和(he)"虽然字形拼音都不同,但它们或者声母相同或者韵母相同.并且两个字符串的长度是一样的(都是 14 个中文字符).

受文献[1]的启发及以上的分析,下面给出两个中文文本字符串之间的相似度的定义.

考虑下面四个因素对相似度的影响:

(1)文本串 $p1$ 与文本串 P 相同的字的个数,相同的字数越多,相似程度越大,因此设

$$e_1(p_1,p)=\frac{|p_1\cap p|}{|p|}\times w_1$$

其中 p_1 和 p 分别表示 $p1$ 与 P 中包含的字的集合,重复出现的字看成不同的元素,$|p|$ 表示 p 中元素的个数,w_1 为权重系数,以下类似.

(2)$p1$ 与 P 中拼音相同但字形不同的字的个数,这个数值越大,相似度也越大,设

$$e_2(p_1,p)=\frac{|y_1\cap y|}{|p|}\times w_2$$

其中 y_1 表示 $p_1-p_1\cap p$ 中所包含的字的拼音的集合,y 表示 $p-p_1\cap p$ 中所包含的字的拼音的集合,其中 $p_1-p_1\cap p$,$p-p_1\cap p$ 分别表示两个集合的差集.

(3)$p1$ 与 P 中只有声母或韵母之一相同的字的个数,这个数值与相似度也成正比关系,设

$$e_3(p_1,p)=\frac{|s_1\cap s|}{|p|}\times w_3$$

其中 s_1 表示 $y_1-y_1\cap y$ 中所包含的字的声母和韵母的集合,s 表示 $y-y_1\cap y$ 中所包含的字的声母和韵母的集合.

(4)$p1$ 与 P 中包含的字的个数相差的程度.当 $p1$ 与 P 的字数之比大于 1 时,$p1$ 对于 P 的相似度与这个数值成反比;当 $p1$ 与 P 的字数之比小于 1 时,$p1$ 对于 P 的相似度与这个数值成正比,因此设

$$e_4(p_1,p)=\left(-\frac{2}{\pi}\arctan\left|\ln\frac{|p_1|}{|p|}\right|\right)\times w_4$$

$p1$ 与 P 的字数之比与 e_4 的关系如图 1 所示,不难证明

$$\lim_{|p_1|/|p|\to 0}e_4(p_1,p)=-1,$$

$$\lim_{|p_1|/|p|\to\infty}e_4(p_1,p)=-1,$$

当 $|p_1|/|p|=1$ 时,$e_4(p_1,p)=0$.这些性质正好符合了我们的要求.

综合以上四种因素,我们定义 p_1 对 p 的相似度为

定义 1　$p1$ 相对于 P 的相似度为

$$e(p_1,p)=\sum_{i=1}^{4}e_i(p_1,p).$$

上述定义中权重系数如何选取是一个值得研究的问题,不过它们应该满足一个基本条件,即 $w_1>w_2>w_3$,因为字相同、拼音相同、声母或韵母相同三者对相似度的贡献显然是依次减小的.本文中取 $w_1=1$,$w_2=0.8$,$w_3=0.5$,$w_4=1$,实验表明,这样取值就能取得较好的效果.

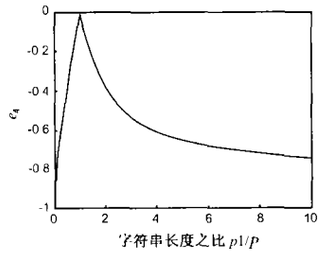

图 1　字符串长度之比 $p1/P$ 与 e_4 之间的关系

下面我们以一个例子来说明上述相似度定义.

设 $P=\{$关注乔丹,关注英雄雕塑的复活$\}$,$p1=\{$关注巧干,关注英雄雕塑的副和$\}$.因为 P 与 $p1$ 中含有 10 个相同的字符(标点符号也包括在内),P 中总共有 14 个字符,故 $e_1(p1,P)=10/14$;这时,$y=\{$qiao dan fu huo$\}$,$y_1=\{$qiao gan fu he$\}$,则 $e_2(p1,P)=2/4*0.8$;这时,$s=\{$d an h uo$\}$,$s_1=\{$g an h e$\}$,$e_3(p1,P)=2/4*0.5$;由于 $p1$ 与 P 的长度相等,因此 $e_4(p1,P)=\left(-\frac{2}{\pi}\arctan\left|\ln\frac{|p_1|}{|p|}\right|\right)\times 1=0$.所以,$e(p_1,p)=$

$$\sum_{i=1}^{4} e_i(p_1, p) = 10/14 + 2/4 * 0.8 + 2/4 * 0.5 + 0 = 1.3643.$$

3 系统的实现

(1)拼音库的建立

由于计算相似度时,涉及到拼音的提取,因此我们建立了一个拼音库,其中包含了全部的一级汉字及部分常用二级汉字,共六千余字.对于多音字,我们取它的常用的音,如"参"(can,cen),其拼音取 can.对于零声母的字,用不同的数字来代表其声母,如"阿"(a)的声母规定为 1,欧(ou)的声母规定为 3,等等,从而避免把所有零声母的字的声母都看成相同的.

(2)搜索算法

如果简单的从头到尾搜索拼音库以获取每个字的拼音,那么每个字一般需要几千次的搜索.假设每个网页中超文本所含的字数为一千左右,那么,进行一次匹配需要上百万次的搜索,试验表明,这样的反应速度用户是无法接受的.因此我们采取了如下的策略:将拼音库按每个字的机器内码的大小排序,这样每个字与一个整数值一一对应,然后采用二分法进行搜索,每个字需要的搜索次数锐减为 $\log_2 6000$,即最多需要 13 次,每一次匹配需要的搜索次数仅仅一万次左右,这几乎在瞬间就可以完成.

(3)在实际计算相似度时,考虑到相同字、拼音、声母或韵母在两个字符串中出现的位置不同,对相似度的贡献也会不一样,如果出现在同一位置,贡献应该是最大的,出现的位置相差越远,贡献越小,如"我上学堂"与"网上学堂"中的"学"字和"我上学堂"与"学生会"中的"学"字对各自相似度的贡献应该是不一样的.因此,我们给它们一个权重系数 $1/(|i-j|+1)$(其中 i,j 分别表示两个字符在各自字符串中所处的位置),以便有所区分.

(4)系统执行的基本过程

系统的具体执行过程如下:首先进入到某一网站的主页(如北方交通大学主页),通过网页分析,提取出当前网页上的所有超文本及每一个超文本对应的 URL,这时用户可以用语音读某个超文本,如"院系设置".识别引擎识别的结果为"游戏是指",系统把这个识别结果分别与提取的超文本作比较,计算出它与每个超文本之间的相似度.其中与"院系设置"的相似度最大(这当然也是我们所期望的),这样系统判定用户输入的本意为"院系设置",然后连接到"院系设置"所链接的网页.

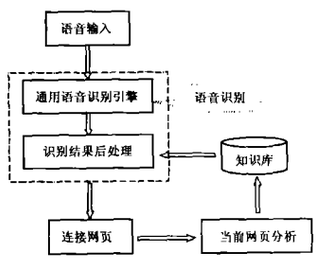

图 2 系统的基本结构

当进入到一个新的网页以后,重复进行上述过程.

系统的基本框架如图 2 所示.

4 实验结果

我们以数十所大学及常用网站的主页为例进行了实验,实验结果见表 1.我们看到,如果单纯用识别引擎来识别语音输入,正确率仅为 60%左右,而采用本文的方法进行二次处理以后,总正确率达到了 95%以上.这表明本文的方法是非常有效的.

表 1 实验结果

Http 地址	语音输入次数	引擎识别结果		处理结果	
		正确次数	正确率	正确次数	正确率
http://www.njtu.edu.cn/	43	26	60.5%	42	97.7%
http://www.tsinghua.edu.cn/	22	13	59.1%	22	100%
http://www.pku.edu.cn/	52	31	59.6%	50	96.2%
……	……	……	……	……	……
http://www.sina.com/	172	95	55.2%	163	94.7%
http://www.sohu.com/	215	121	56.3%	203	94.4%
http://www.263.net/	156	79	50.7%	150	96.2%
……	……	……	……	……	……
总 平 均	3258	1912	58.7%	3098	95.1%

在实验过程中,我们注意到这样一个问题:由于有的词是多音字,在本来应该得到正确结果的情况却不能得到正确结果,如,语音输入"攻略",识别引擎的识别结果为"功率",按相似度的定义,"功率"与"攻略"的相似度应为 0.65,"股市"与"攻略"的相似度应为 0.5,匹配结果应为"攻略",但实际上匹配的结果却是"股市",原因在于"率"字在拼音库中的拼音为"shuai",而不是预想的"lv".这样实际计算结果为,"功率"与"攻略"的相似度为 0.4,"股市"与"攻略"的相似度为 0.5,从而造成匹配错误.

5 结束语

由于我们定义的相似度是针对中文文本的,因此,我们的方法只适用于超文本全部为中文字符或绝大部分为中文字符的情形.当超文本全部或大部分为非中文字符时(如"ICTIS'2002",此时识别引擎的识别结果已经面目全非),方法将失效.对于不同的语言,我们需要定义不同的相似度.如针对英文的语音浏览,我们可以重新定义一种英文文本之间的相似度,当然,定义与中文是类似的.同时,还有第 4 节中提到的由于多音字造成的匹配错误也是一个尚待解决的问题.

最后,还想指出一点,本文提到的方法并不是只适用于 Web 语音浏览,也可以用于对话系统,手机上网等.概括起来,此方法可应用到这样的情形:在当前状态下,能够预知语音输入的内容为某个有限集中的一个元素,就可以利用这种匹配技术极大提高识别的正确率.

参考文献:

[1] Voice Extensible Markup Language (VoiceXML) Version 2.0[Z]. http://www.w3.org/TR/2001/WD-voicexml20-20011023/.

[2] Gu H X, *et al*. Spoken query for Web search and navigation[A]. Best Student Poster Award, 10th International World Wide Web Conference,

Poster Proceedings[C]. Hong Kong,2001.2 - 3.

[3] Sami Rollins, Neel Sundaresan. AVoN calling: AXL for voice-enabled Web navigation[J]. Computer Networks,2000,33:533 - 551.

[4] Zhang J, Edie M, Rasmussen. Developing a new similarity measure from two different perspectives[J]. Information Processing and Management, 2001,37:279 - 294.

[5] Christine Michel. Ordered similarity measures taking into account the rank of documents[J]. Information Processing and Management,2001, 37:603 - 622.

[6] 丁丰.对话系统中知识获取的研究[D].北京:北方交通大学研究生院,2001.

[7] Zhang L, *et al*. Automatic detecting/correcting errors in Chinese text by an approximate word-matching algorithm[A]. ACL-2000, The 38th Annual Meeting of the Association for Computational Linguistics[C]. Hong Kong. 2000. http://reseach. microsoft. com/asia/d/oad-nlps/NLPSL/n4.pdf.

[8] Jim White. Voice Browsing[J]. IEEE INTERNET COMPUTING,2000, JANUARY-FEBRUARY:55 - 56.

[9] Voice Browsers[Z]. http://www.w3.org/TR/NOTE-voice.

[10] Dong L, *et al*. Using Chinese spoken-language access to the WWW [A]. Proceedings of WCC-ICSP 2000[C]. Beijing: Publishing House of Electronics Industry,2000,2:1321 - 1324.

作者简介:

李红莲　男,1971 年生于河北保定市,博士研究生.研究方向:语音识别与理解,语音上网及知识获取等.

王春花　女,1971 年生于山西太原市,博士.研究方向:数据挖掘,移动通信等.已发表论文 20 余篇.

袁保宗　男,1932 年生于江苏吴江市,教授,博士生导师,IEEE 北京计算机分会、信号处理分会主席,IEE 北京中心主席.研究领域:信号与信息处理,多媒体视听信息处理,计算机视觉,图象信号处理,语音信号处理,计算机图形学与虚拟现实.已发表论文 200 余篇.