文章编号：92

文章编号：1003-0077（2011）00-0000-00

基于中文拼音输入法数据的汉语方言词汇自动识别*

张燕[1]，张扬[2]，孙茂松[1]

（1.清华大学计算机系，北京市 100084；2. 搜狗科技公司，北京市 100084）

摘要：方言研究领域中的语音研究、词汇研究及语法研究是方言研究的三个重要组成部分，如何识别方言词汇，是方言词汇研究首要的环节。目前，汉语方言词汇研究的语料收集与整理主要通过专家人工整理的形式进行，耗时耗力。 随着信息技术的发展，人们的交流广泛通过网络进行，而输入法数据包含海量的语料资源以及地域信息，可以帮助进行方言词汇语料的自动发现。然而，目前尚没有文献研究如何利用拼音输入法数据对方言词汇进行系统化分析，因此在本文中，我们探讨借助中文输入法的用户行为来自动发现各地域方言词汇的方法。特别的，我们归纳得到输入法数据中表征方言词汇的两类特征，并基于对特征的不同组合识别方言词汇。最后我们通过实验评价了两类特征的不同组合方法对方言词汇识别效果的影响。

关键词：方言词汇识别；中文拼音输入法；特征融合

中图分类号：TP391　　**文献标识码：**A

Automatic Identification of Chinese Dialect based on the Data from Chinese Pinyin Input Method

ZHANG Yan[1], ZHANG Yang[2] , SUN Maosong[1]

(1.Tsinghua University, Beijing 100084, China ; 2. Sogou Inc., Beijing 100084, China)

Abstract: The study of dialect is composed of voice study, vocabulary study and grammar study, of which the first step is to recognize the dialect vocabulary. By now, collection of Chinese idiom words is mainly accomplished by experts, and it is time-consuming and labor-intensive. With the development of information technology, people communicate widely through the network, and thus input method data contains vast amount of vocabulary resources as well as the geographical information, which can help automatically discover dialect words corpus. However, in literature, there have been very few studies on how to exploit the input method data to systematically investigate the dialects. Therefore this paper analyzes the user behavior of Chinese input method, and based on which we propose to automatically discover the geographical dialect vocabulary. Specifically, the paper gets the two representative features of dialects in Chinese input method, and uses different combinations of these two features to recognize dialect words. Finally, extensive experiments are performed to evaluate the impacts of the feature combinations on the dialect word recognition.

Key words: Dialect detection；Chinese Pinyin input method；Feature combination

1 引言

方言词汇研究是方言研究的一个重要方面，其中方言词汇的识别是方言词汇研究的首要环节。方言词汇研究在语言学研究、信息检索、机器翻译、刑事侦查等方面都有重要的应用价值[1]。但目前方言词汇研究的语料收集工作主要依赖于专家的人工整理[2,3]，这一工作需要耗费大量的时间和精力。信息技术的不断发展，特别是中文输入法的广泛应用，为人们

* **收稿日期：2013-06-01**　　**定稿日期：2013-07-15**

基金项目：国家自然科学基金重点项目（61133012）；国家 863 计划项目（2012AA011102）

作者简介：张燕（1981—），女，博士在读，主要研究为自然语言处理、机器学习；张扬（1981—），男，研究员，主要研究方向为自然语言处理、机器学习、输入法；孙茂松（1962—），男，教授，主要研究方向为自然语言处理、信息检索和社会计算。

日常的网络交流带来很大便捷，而输入法中所记录的用户行为，特别是带有用户地理信息的输入记录，能够反映出不同地域用户的语言使用习惯及地域相关词汇的特征。基于此，本文中我们主要考虑借助中文拼音输入法的记录来自动发现汉语方言词汇，为汉语方言词汇研究提供语料库。

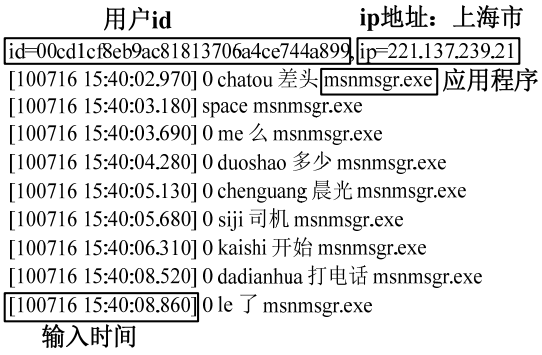


图 1 输入法用户记录示例

图 1 中的数据是某中文拼音输入法记录的一段用户输入行为。从图中我们可以看出，用户的输入记录中包含用户的地理信息（即用户的 IP 地址）、拼音选词习惯、录入习惯以及使用环境（即调用拼音输入法的应用程序）。通过用户的 IP 信息，我们可以根据 IP 地址库来确定用户所在的地理位置，而这样的地理位置是我们赖以发现汉语方言词汇的重要数据依据。

由于中文输入法可以自动记录用户的输入行为，不需要用户的主动参与即可采集到大量的方言词汇数据，从而为方言研究提供大规模语料。然而目前在研究界，少有利用中文输入法数据来进行方言研究的，只有郑亚斌等人[4]的工作在中文输入法数据的基础上研究了中文的地域相关词条，主要目的为扩充中文输入法词库，并非针对中文方言词汇进行自动发现。本文正是基于这样的考虑，试图通过输入法数据中的用户行为，来分析现代汉语在使用过程中所体现出的地域分布性质，从而自动发现汉语方言词汇并研究其时空分布。本文的主要贡献是：1）我们提出了一种通过中文拼音输入法中记录的用户行为信息来发现并分析汉语方言词汇的方法；2）其次，我们基于人工标注的方言语料分析了地理信息、使用环境等特征及其特征组合对识别汉语方言词汇的影响；3）最后我们将分析得到的有效特征组合通过特征排序的方法获得了全国各地域的方言词表。

本文的主要安排如下：首先在第 2 节根据标注语料库来确定数据的归一化等预处理，对语料库中的词条进行向量化处理，获得汉语词条的地理信息、时间信息、使用频度信息等；在第 3 节，分析并提取标注语料的特征并验证其有效性；最后，在第 4 节对有效特征进行融合通过排序的方法获得全国各地域的方言词汇。

2 问题描述

2．1 汉语方言词汇的概念及特征

汉语方言词汇是基于现代汉语词汇的横向比较研究而产生的，是语言的地域变体[3,5]。方言是一定区域内的交流工具，因此方言词汇的使用频度在地域分布上具有区内较高、区外较低的特征，且由于方言多在日常交流中使用，方言词汇的口语化程度较高[6]，所以我们主要通过地域分布以及口语化程度这两个特点来考察方言词汇。

2．2 中文拼音输入法用户记录

由于中文输入法数据可以提供词汇的使用频度，故我们可以定量分析方言词汇的地域分布特点以及口语化程度，进而根据这两个特点来识别方言词汇。图 1 中给出了一段具体的输

入法用户记录，由此记录我们可以获得以下信息：用户输入的词条，用户的录入时间，用户调用输入法的应用程序，以及用户的 IP 地址。大量的用户输入记录可以获得中文词条在各地域的输入频度，以及在不同的应用程序中使用的频度。其中词条在各地域的输入频度可用以描述词条在地域分布上的特征，而词条在不同的应用程序中的使用频度则可以描述词条的口语化程度，即在以使用口语为主的程序中出现频度较高的词条则口语化程度相对较高。由于我们的输入法数据包含了全国共 34 个省级地域的用户记录，所以每一个词条均可以获得一个 34 维的地域分布特征相关的向量，而用户调用中文输入法的应用程序数目较多，我们仅选取有代表性的频率最高的前 100 个应用程序，每一个应用程序可以根据其主要作用标注为口语型或书面语型应用程序，例如“iexplore.exe”是浏览器程序，我们将其标注为书面语型的应用程序，而“QQ.exe”是即时通讯软件，多用于用户之间的日常交流，故我们标注其为口语型应用程序。

对数据集中的每个词条，我们均可以获得一个 34 维的地域分布向量，以及一个 100 维的用以衡量词条口语化程度的向量。方便起见，我们记录该词条为 ，分别根据该词条在全国 34 个省级地域的使用频度、在 100 个应用程序中的输入频度这 2 类特征，生成一个包含 134 个特征值的向量 ，可以参考图 2 中的表示。

34个地域　　100个应用程序

词条	山东	江苏	安徽	浙江	福建	上海	广东	...	...	台湾	香港	澳门	QQ.exe	...	iexplore.exe	...	9158.exe
差头	3	5	1	5	1	591	9	...	...	0	0	0	498	...	21	...	0

图 2：词条“差头（出租车）”的向量化表示

2．3 方言词汇在中文拼音输入法数据中的特点

如上所述，方言词汇的使用频度在地域分布上具有区内较高、区外较低的特征，并且在口语中较常使用，而在书面语中出现较少。根据这一特征，我们分别列举了“差头（出租车）”、“水门汀（水泥）”、“新闸路”这 3 个词条在全国 34 个省级地域的频度分布及其在 100 个应用程序中的输入频度，参考图 3。

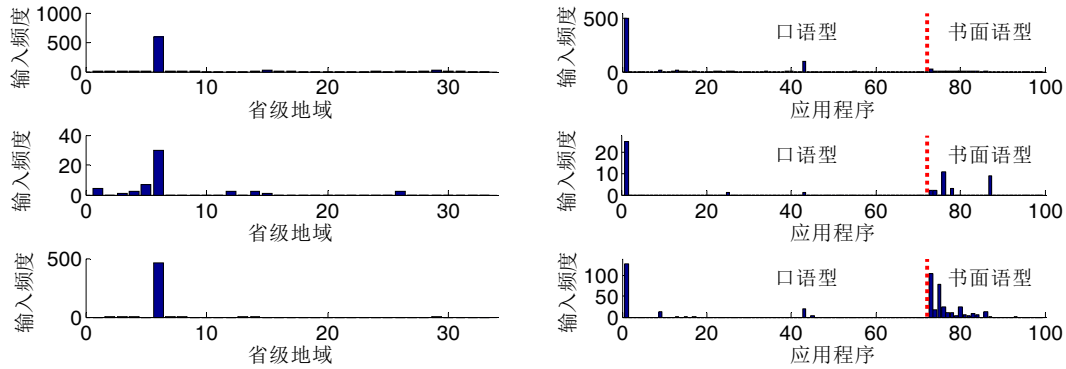


图 3：词条“差头（出租车）”、“水门汀（水泥）”、“新闸路” (a): 在各省级地域的输入频度; (b):在各应用程序中的输入频度

图 3 中的左图，自上而下分别是“差头（出租车）”、“水门汀（水泥）”、“新闸路”这 3

个词条在全国 34 个省级区域的分布，可以看出，这 3 个词条均在上海地区（横坐标为 6）时达到峰值，这与实际情况是相符的。因为“差头”与“水门汀”均属洋泾浜英语，在上海地区使用人数最多，而在其余省级地域很少使用，这两个词条可以通过地域分布特征提取出来，而“新闸路”属于上海的地名，但并不是上海地区的方言词汇，虽然其只在上海地区使用，但单纯通过地域特征计算会混淆在上海方言词汇中而被提取出来，这本质上是区分地域词汇和方言的困难所致。图 3 中的右图，自上而下分别给出了上述 3 个词条在 100 个应用程序中的输入频度，其中虚线左部为口语化应用程序，而右部为书面语型应用程序。从右图中可以看出，由于“新闸路”为地名，除去其在口语化应用程序中会出现之外，在书面语型的应用程序中也会出现，而且频度甚至比在口语化应用程序中出现的更多，故我们可以考虑通过应用程序的口语化程度来过滤部分地名。受本例启发，我们试图通过选择合适的词条特征来鉴别方言词汇。下面我们给出中文拼音输入法记录中的汉语词条地域分布特征及口语化程度的分析及计算。

3 汉语方言特征分析

3．1 特征分析及其计算

如前所述，方言词汇的使用频度在地域分布上具有区内较高、区外较低的特征，据此，我们对方言词汇在各省级地域上的频度分布以及在 100 个应用程序中的输入频度分别进行了统计，归纳得到两类特征，用以辨识词条是否属于地域 的方言词汇。表 1 是对这两类特征的说明。其中，概率比的特征 PRL 是对应“地理区域性”的，而口语化程度的计算则是在应用程序中的输入频度基础上进行的，即不同应用程序中的频率概率比 PRP。具体的每个特征的表征意义参考表 1。

表 1：特征说明

表示符号	说明	特点
PR_L	不同地域内的频度概率比	词条在该地域多而其余地域少
PR_P	不同应用程序中的频度概率比	词条口语中使用较多而书面语中较少

表 1 中所列的特征，计算如下：

1) 特征 PR_L 主要用以判断词条是否属于某地域 l 的相关词条，假设给定词条 w 的归一化特征向量为 v，该特征的计算如下：

$$PR_L = PR(w,l) = \frac{p(w|l)}{P(w|l\_)} = \frac{Freq(w,l)/Freq(l)}{Freq(w,l\_)/Freq(l\_)} \quad (1)$$

其中，PR_L 表示的是词条 w 在地域 l 中的分布概率与其在地域 l 之外的地域（即公式中的 $l\_$）分布的概率之比，此值越大，则表明词条 w 属于地域 l 的方言词汇的可能性越大。

2) 特征 PR_P 主要用以判断词条是否属于口语化词汇，因为方言词汇在口语化的应用程序中使用较多，而在书面语的应用程序中使用较少，所以我们通过计算词条在不同类型的应用程序中的频度分布概率比来度量词条属于方言词汇的可能性，其计算方法如下：

$$PR_P = PR(w,p) = \frac{p(w|p_+)}{p(w|p\_)} = \frac{Freq(w,p_+)/Freq(p_+)}{Freq(w,p\_)/Freq(p\_)} \quad (2)$$

其中，PR_P 描述的是词条 w 在口语化应用程序（即公式中的 p_+）中使用的概率与书面语应用程序（即公式中的 p_-）中的概率之比，比值越大，越能说明词条 w 是方言词汇的可能性较大。

3．2 特征组合

由上面的计算公式可以看出，两部分的特征可以统一看作词条 w 属于某地域 l 的概率比，以及属于口语化词汇的概率比，这两类特征可以看作是概率比公式的统一计算，而且两者的取值范围均在 $[0,1]$ 之间，故我们考虑通过加权调和平均的特征组合方式来考察两种特征对于方言词汇自动识别的贡献。我们分别假设两种特征的权重为 α 和 $1-\alpha$，对上述 2 种特征进行组合，参见公式(3)。在下面的实验部分我们考察了权重参数 α 对实验结果的影响。

$$P(w) = \left(\frac{\alpha}{PR_L} + \frac{1-\alpha}{PR_P}\right)^{-1} \tag{3}$$

公式(3)中，参数 $\alpha \in [0,1]$，用以调整特征 PR_L 与特征 PR_P 的权重，$P(w)$ 则用以表示 w 属于方言词汇的概率。我们通过对 $P(w)$ 的排序来确定方言词汇。

4 实验结果和分析

4．1 数据描述

我们的输入法数据是从搜狗拼音输入法中获得的 2010.7.1-2010.7.7 之间共 7 天的用户输入记录，共约 262GByte 的数据，过滤掉总频度低于 50 的低频词条后，可以获得输入记录的词条数目为 2,478,039，这些词条作向量化处理后，最后得到 2,478,039 个 134 维向量的集合。

为了对比汉语词语的地域性，我们根据语言学专家提供的数据集，选取标注了 3 个语料库作为观察数据，包括：北京方言[6]、上海方言[7]以及常用词条[8]。针对这三个观察数据集，我们可以获得数据集中的词条在搜狗拼音输入法中的记录。去除了总频度低于 50 的低频词条之后的上海方言词汇为 169 条，北京方言以及现代汉语常用三千词在搜狗拼音输入法中出现的词条数则分别为 3010 和 2565。由于上海方言的词汇集合较小，而北京方言及常用词的数目较多，为了实验的可比性，我们最终选择上海方言 169 条，随机选取北京方言、常用词条各 200 条，作为我们的标注数据集，以观测权重参数对实验结果的影响，从而指导未标注集合上的方言词汇识别。

4．2 评价指标

对于上文提到的两种特征，我们将计算在不同的权重参数 α 下，这两种特征在北京标注方言与上海标注方言数据集上的性能，为了评价我们提取的方言词汇的准确性，我们采用以下指标：

1） 前 N 个返回结果的准确率(记为 $P@N$)

其中，$P@N$ [9]计算在返回的前 N 个最优结果的准确率，这一标准常用在信息检索领域中来衡量检索结果的准确度。针对北京、上海方言的标注集合，对于评价系统对北京标注方言的返回性能，我们主要考虑 $P@10$， $P@20$， $P@50$， $P@100$，$P@200$ 这 5 个指标，而对于评价上海标注方言，由于我们标注的上海方言词汇在输入法数据中仅有 169 个词条有记录，所以我们采用 $P@10$， $P@20$， $P@50$， $P@100$， $P@169$ 这 5 个指标。

2） 二元偏好值(记为 *Bpref*)

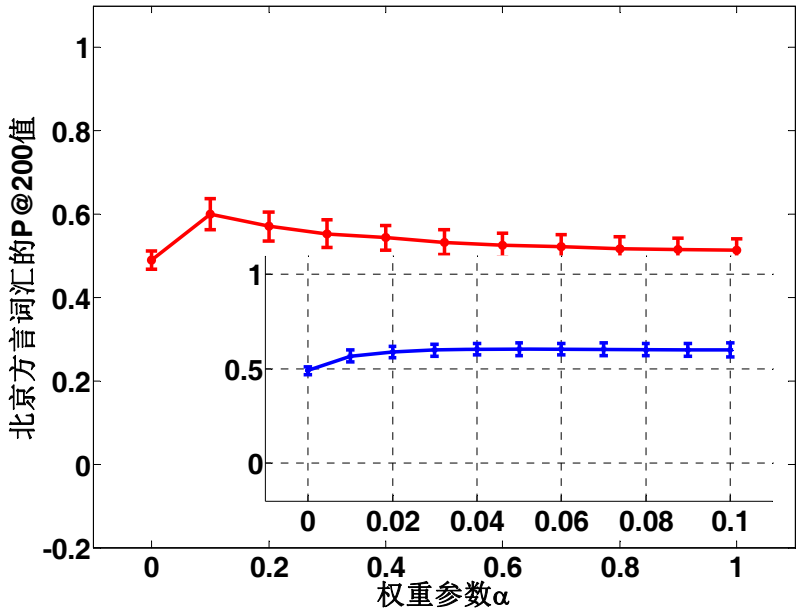
根据文献[10]，*Bpref* 用以评价返回结果中，正确词条与非正确词条的相对位置，主要用以评价系统能否将相关词条在不相关词条之前返回，其计算公式如下：

$$Bpref = \frac{1}{R}\sum_{r} 1 - \frac{|n|}{R} \tag{4}$$

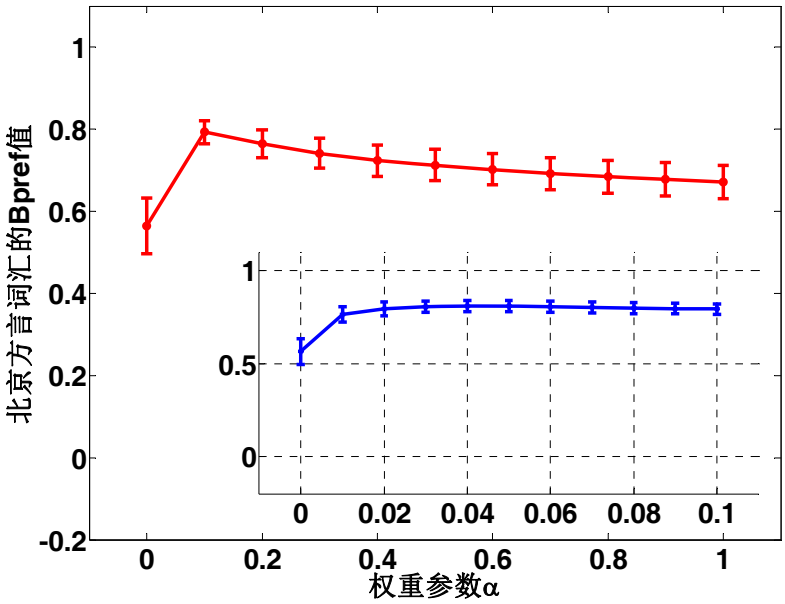
其中 n 是排在 r 之前的正确的词条的个数。对于 *Bpref* 的计算，我们选取 $R = 200$ 。

4．3 权重参数的影响

针对北京及上海方言，我们计算了根据各地域总频度进行归一化的情况下，标注数据集合中北京、上海两地方言词汇，随特征权重参数 α 变化的识别效果。具体情况参考图 4 及图 5。

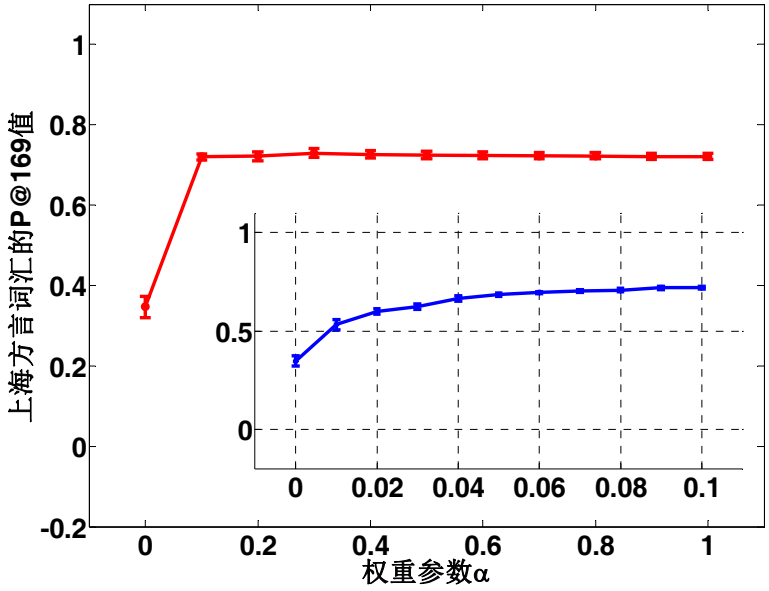


(a)

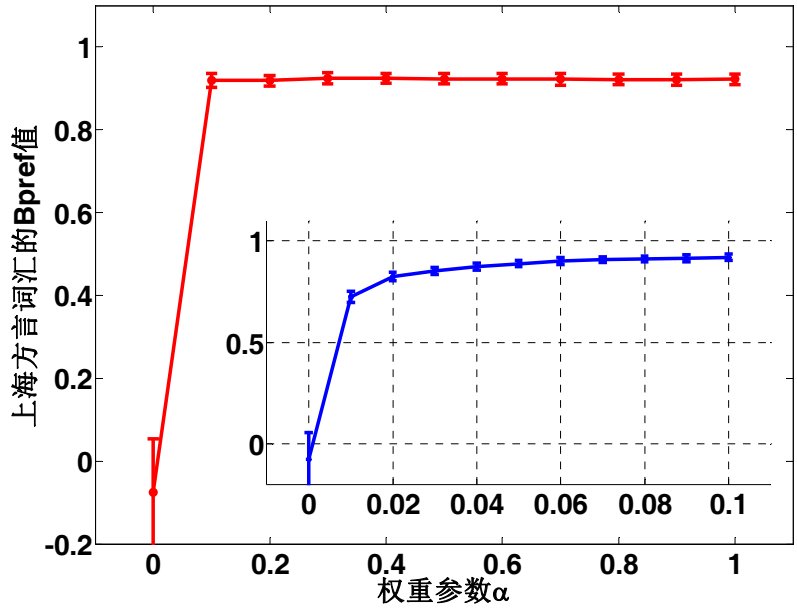


(b)

图 4：权重参数 α 对北京方言词汇识别结果的影响。(a)： $P@200$ ；(b)： $Bpref$



(a)



(b)

图 5：权重参数 α 对上海方言词汇识别结果的影响。(a)： $P@169$；(b)： $Bpref$

图 4 中，左图(a)以及右图(b)中的曲线分别是权重参数 α，在 $[0,1]$ 之间按照步长 0.1 取不同的值时，北京方言词汇识别的 $Bpref$ 和 $P@200$ 的结果。各个子图中的内嵌图是对大图中最高取值区间的细分，左图(a)中是在 $[0,0.1]$ 之间按照步长 0.01 取不同的值时，北京方言识别结果的 $Bpref$ 值，而右图则是 $P@200$ 的值。同样的，图 5 中的左图(a)与右图(b)则分别是上海方言识结果的 $Bpref$ 值与 $P@169$ 的值。

综合图 4、图 5 中可以看出，当 $\alpha = 0.08$，在观察数据上可以获得的北京方言及上海方言词汇的准确率以及二元偏好值均较高，而且试验效果受地域分布相关特征 PR_L 的影响较大，而口语化相关特征 PR_P 则相对而言不是非常敏感。在此后的试验中，我们均采用设置权重参数 $\alpha = 0.08$。由于试验结果受地域相关特征影响较大，所以对地域的更细划分，会更有助于我们的试验，这部分将作为我们下一步的工作继续研究。

4．3 实验结果及分析

根据标注集合的评价结果，我们确定了特征组合方式及权重系数，针对未标注数据，我们分别计算了全国 34 个地域的方言词汇，并在表 2 中给出了 6 个方言区中的 6 个有代表性的地域上的方言检测的前 10 个结果。在这里之所以没有给出客家方言区的代表区域，是因为客家方言的分布比较复杂，集中分布在某几个地区的某几个区域，由于我们目前采用的地域分区只细分到省份，所以不能确切地给出客家方言的代表区域。从表 2 中可以看到各个代表地域检测出的前 10 个方言词汇，在不同的应用程序及地域分布都是比较集中的。

表 2：方言词汇识别的结果

方言区	北方方言	吴方言区	粤方言区	闽方言区	湘鄂方言	赣方言区
代表	北京	上海	广东	福建	湖北	江西
1	有地么	测那	淫龙	收我为徒吧	么思	婉清
2	吗去了	瓦拉	马兰开花	各税	么昂	哟里
3	你吗呢	噶散户	老虎哥哥	火毛	洗了睡	逼战
4	真孙子	满叫	浪险	靠妖	静之	恰噶
5	嘛去	一炮无	喽米	加氨	搞么斯	该枪
6	有什么能	是伐啦	果到	有傻	利马	杂处
7	杀星不	乖囡	做紧乜	颠趴	屋地	小乃
8	叶卡	你比样	体紧	喝铁观音	麽办	让费
9	你是北京人吗	快来西	唔入	无语掉	莫样	瓦擦擦
10	嘛呢你	去伐	哩几日	私麦	麽比	蝙蝠魔

可以看出，我们的算法在南方的五大方言区的检测效果较好，而在以北京地区为代表的北方方言的检测效果最差。这是由于北京地域的方言与普通话的差异较小，要更好检测出北京方言，还需要引入其他的特征，由于篇幅关系，我们下一步工作将详细研究。

为了评测为标注集合上各个特征的性能，我们选取了北京、上海、广东三地各 6 组结果中的前 200 个返回结果进行人工评测，根据 4.2 节给出的 2 种评价标准统计了人工标注的结果，参考图 6。其中可以看出，综合考虑了地域相关的特征以及程序口语化特征的情况下，北京、上海、广东这 3 个地区的方言检测结果均比只考虑了地域相关特征的效果有明显改善，特别是对北京、上海地区的方言，前 200 个返回结果中，能提高 50%以上的效果。

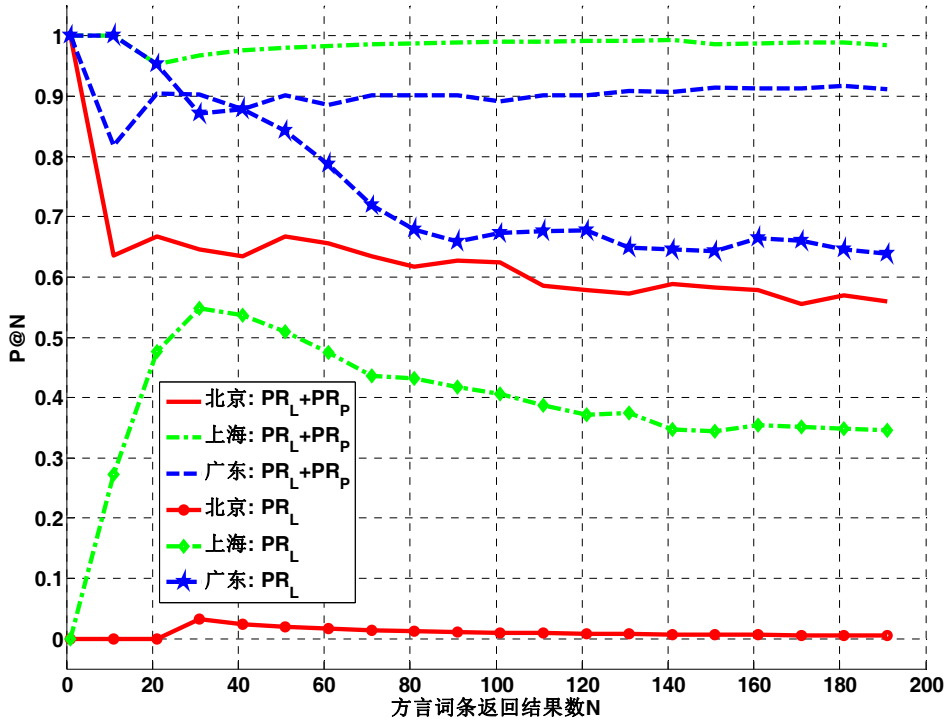


图 6：北京、上海、广东地区方言返回结果

同样的，针对评价指标 *Bpref* 的结果，表 3 中可以看出，两种特征结果之后的检测结果相比只考虑了地域相关特征的结果，在上海、广东地区，能提高 90%以上，而在北京地区则能提高 32%以上。这说明，两种特征结合的效果优于只考虑了地域相关特征的效果，因此再一次验证了引入输入法记录词条的口语化相关的特征是必要的。

表 3：北京、上海、广东地区方言返回结果： *Bpref*

$PR_L + PR_P$			PR_L		
北京	上海	广东	北京	上海	广东
0. 6564	0. 9936	0. 9418	0. 3282	0. 0144	0. 0076

5 总结及下一步工作

本文首先提出了一种利用中文拼音输入法中记录的用户行为来识别并分析汉语方言词汇的方法；基于此方法，我们对人工标注的方言语料的特性进行统计，分析了地理信息、语言特征对汉语方言识别中的影响；最后我们通过交叉验证的方法来调节有效特征的权重参数，对特征融合后通过排序的方法获得了全国各地域的方言词汇。基于本文的工作，一旦获得个地域方言词汇库，下一步我们可以对地域划分得更细，分析各地域方言的异同，从而对全国方言进行更细的分区。

参考文献

[1] 顾明亮，沈兆勇.基于语音配列的汉语方言辨识[J].中文信息学报，vol.20, No.5, 77-82.

[2] 李如龙.谈汉语方言的比较研究——兼评《汉语方言大词典》[J].辞书研究，2000，(4).

[3] 詹伯慧.汉语方言及方言调查[M].湖北教育出版社，2001.

[4] 郑亚斌.中文用户输入法用户行为分析及其应用[D].清华大学博士学位论文，2011.

[5] 邢向东.关于深化汉语方言词汇研究的思考[J].语言文字学研究, 2007, (2):117-122.

[6] 董树人.新编北京方言词典[M].商务印书馆，2010.

[7] 李庆鸿.上海话托福(常用词汇)[M].学林出版社，2010.

[8] 郑林曦.普通话三千常用词表[M].1987，文字改革出版社.

[9]Yates R, Neto B. Modern information retrieval[M]. Addison-Wesley Longman Publishing Co.,Inc., 1999.

[10] Buckley C,Voorhees E. Retrieval evaluation with incomplete information[C]. Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.ACM, 2004.25-32.