基于输入法用户词库和查询日志的若干研究

Some Research based on User Dictionary of Input Method and Query Log

（申请清华大学工学硕士学位论文）

培 养 单 位 ： 计算机科学与技术系

学　　科 ： 计算机科学与技术

研 究 生 ： 王 鹏

指 导 教 师 ： 孙 茂 松 教 授

二〇一一年四月

基于输入法用户词库和查询日志的若干研究

王鹏

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容。

本人保证遵守上述规定。

（保密的论文在解密后遵守此规定）

作者签名：_____ 导师签名：_____

日　　期：_____ 日　　期：_____

摘　要

中文输入法是中文计算机用户的重要工具，但是基于中文输入法的用户行为研究非常稀少。本文研究了用于中文输入法用户词库和搜索引擎查询日志的字、词情况。通过数据处理统计分析，本文介绍了用户词库和查询日志在用字用词上的新特点并与其他数据进行对比。结合用户词库、查询日志、Wiki、Sogout 数据，本文论述了寻找 Wiki 中文常见词条的方法并分析结果。本文还介绍了通过两个记录用户输入序列信息的输入法数据抽取拼音错误模式的方法，并对结果进行分析并试图找到错误发生的原因，总结了中文输入法输入错误的一些规律。最后，本文介绍了实现 Win32 平台下女书拼音输入法的机制和原理。

关键词：输入法　　搜索引擎　　错误模式　　用户词库
查询日志

Abstract

Chinese Input Method is an important tool of Chinese computer users. But studies on user behaviors in Chinese Input Method are rare. This paper introduces the research on Chinese characters and words used in Chinese Input Method and Search Engine. Based on data processing and statistical analysis, we introduce new features of Chinese Input Method user dictionaries and Search Engine query logs and compare them to other datasets. We also introduce and analysis some methods to find popular Wiki Chinese words based on user dictionaries, query logs, Wiki and Sogout dataset. And we introduce the method of extracting Pinyin error patterns by several methods based on two datasets of two Chinese input methods which include input sequences of users. Then we analysis these input error patterns and try to find their reasons. We summarize some rules of input errors of Chinese in input method. At last, we introduce how to implement "Female Script" Pinyin Input Method based on Win32 system.

Keywords: Input Method Search Engine error patterns
User dictionary Query log

目 录

第1章 引言

1.1 研究背景

随着网络的发展，信息传播的量和速度都显著提高。根据最新统计，中国网民数量已经达到 3.84 亿，互联网普及率为 28.9%[1][2]。这种信息的高速发展对现有的语言造成了很大影响。就汉语来说，每年都有很多新鲜词汇热门词汇诞生并传播，逐渐形成了一种特有的网络语言。网络语言也在逐渐地影响着平时生活中使用的语言。因此研究网络中的语言状况和对生活中语言的影响成为一种迫切的需求。

汉字输入法是汉语使用者在计算机中输入汉字的工具，也是网络中汉语语言的输入方式。汉字输入法包括拼音输入法、五笔输入法、联想输入法等多种。随着技术的发展，拼音输入法由于其易学易用性逐渐成为主流。输入法作为计算机上最常用的输入中文的工具，其使用情况可以视为计算机上中文使用情况的体现。

1.2 研究现状

针对汉语在网络上的使用，也已经有一些统计分析[3]，主要针对若干大型网站上的文本进行字、词的统计分析。也有针对中文搜索引擎的用户行为进行研究的工作[4]。在中文搜索引擎用户行为分析的基础之上，其他工作也得以展开和拓展[7]。相比网络文本，输入法是用户在网络上使用中文更直接的工具。而通过输入法研究中文用户行为的工作非常稀少，这可能有以下两个原因：一是中国 IT 产业发展迅猛，从较薄弱的基础迅速发展为具有巨大市场价值的产业，许多工作尚未跟进；二是关于用户输入法行为的数据非常稀少。

2006 年 6 月搜狗公司推出了搜狗拼音输入法，是第一个问世的互联网输入法。互联网输入法即是用户可以通过网络及时更新词库，并且可以将自己的词库上传到服务器中。互联网输入法的诞生促进了输入法的发展，并且通过网络收集到大量的用户输入数据（搜狗拼音输入法注册用户词库规模已经达到 100 万用户），为输入法中语言情况的研究提供了条件。搜狗拼音输入法是第一个

问世的互联网输入法，其用户词库也是第一个基于互联网的输入法用户数据集合。

用户在用输入法输入中文时会产生各种错误，对这些输入错误进行分析研究有助于提高输入法的使用效果。目前基于英文等字母语言的自动纠错研究已经有一定历史，2000 年就出现了经典的噪声信道错误模型[5]，基于大规模语料的自动纠错也得到了较好的结果[6]。由于各种原因，中文输入时的错误研究还非常稀少。

1.3 本文的主要内容与贡献

本文主要可以分为两个部分。第一部分，通过搜狗公司提供的输入法用户词库数据、搜索引擎查询日志和其他现有数据，对输入法用户使用的语言状况进行统计分析，并与普通话常用词等数据进行比较，分析网络中的语言变化发展。最后基于 Wiki 数据，进行了中文常见词条的选取排序实验，并对实验结果进行分析。第一部分主要包括第二章、第三章、第四章。第二章主要介绍了用户词库和查询日志中单字使用情况分析，并与其他数据进行对比。第三章主要介绍了用户词库和查询日志中词的使用情况。第四章论述了利用不同数据集寻找 Wiki 中文常见词条的方法并分析结果。

第二部分主要包括第五章和第六章。第五章介绍利用输入法用户输入序列的数据，尝试了若干种抽取常见拼音错误模式的方法，并对结果进行了分析。第六章介绍了在对输入法有一定了解的基础上，实现女书拼音输入法的原理。

最后第七章对之前的章节进行总结，将所得结论进行整理，并介绍了将来的研究计划。

第2章 输入法用户词库和查询日志用字情况分析

2.1 实验概述

本章主要研究了输入法用户词库和搜索引擎查询日志中的单字使用情况，以及中文常用七千字在用户词库和查询日志中的分布，对比了中文传统常用字在网络环境中使用的变化。另外对用户词库、查询日志、媒体常用字表三个数据做了比较。

2.2 数据说明

输入法用户词库数据为搜狗输入法（2006.9.5 推出正式 1.0 版）注册用户的输入数据，记录了所有用户输入的词条和次数，用户数约 90 多万。本实验使用的是截至 2008 年 8 月 21 日的用户词库。记录方法为用户使用输入法上屏时的词条，比如用户输入“中国”，则记录中国；如果用户输入整句“我在哪里”则把“我在哪里”作为一个词条记录。由于分析常用字分布情况，因此把所有词条拆成单字统计，过滤了词频过小的词条。总字频 76775392841。搜狗输入法可以选择用 sohu 账号登陆，注册的用户才会记录词库，图 2.1 是搜狗输入法登陆界面。

图2.1 搜狗输入法登陆界面

搜索引擎查询日志是 sogou 搜索引擎记录的用户查询记录，经过处理后只

保留了查询词和次数，每天分别统计，使用的是 2008 年 5 月 21 日至 2009 年 7 月 2 日的查询词。同样为了统计字频将词条拆成单字，并去掉了非中文字符。总字频 5922947983。

媒体常用字表是统计三家 Web 媒体的新闻语料得到的字频数据，共有单字 9270 个，总字频 991717782。

《现代汉语通用字表》由国家语言文字工作委员会、中华人民共和国新闻出版署 1988 年联合发布，是根据中文常用字情况对汉字按级别划分的字表，没有字频信息，包括 1 级字 2500 个，2 级字 1000 个，3 级字 2500 个。后文中以常用字

2.3 用户词库、查询日志、媒体字表单字使用情况

2.3.1 输入法用户词库

经过统计用户词库中有单字 19679 个，总字频 76775392841。一般来说，常用汉字大约有 5、6 千左右。用户词库的 19679 个单字中包含了大量繁体字、古字、异体字等，这些字一般字频较低。字频最高的 20 个字如表 2.1。

用户词库累计覆盖率曲线如图 2.2。用户词库中，前 87 个字覆盖了 50%的字频，前 431 个字覆盖了 80%的字频，前 808 个字覆盖了 90%的字频，前 1231 个字覆盖了 95%的字频，前 2287 个字覆盖了 99%的字频。常用单字占总数的小部分。

Lg（字频）关于 Lg（Rank）的曲线如图 2.3，线性相关系数-0.9583。并不是很好地符合 Zipf 定律。图 2.4 是每个单字的概率与 Rank 相乘的曲线，发现乘积变化较大，不是很符合 Zipf 定律。

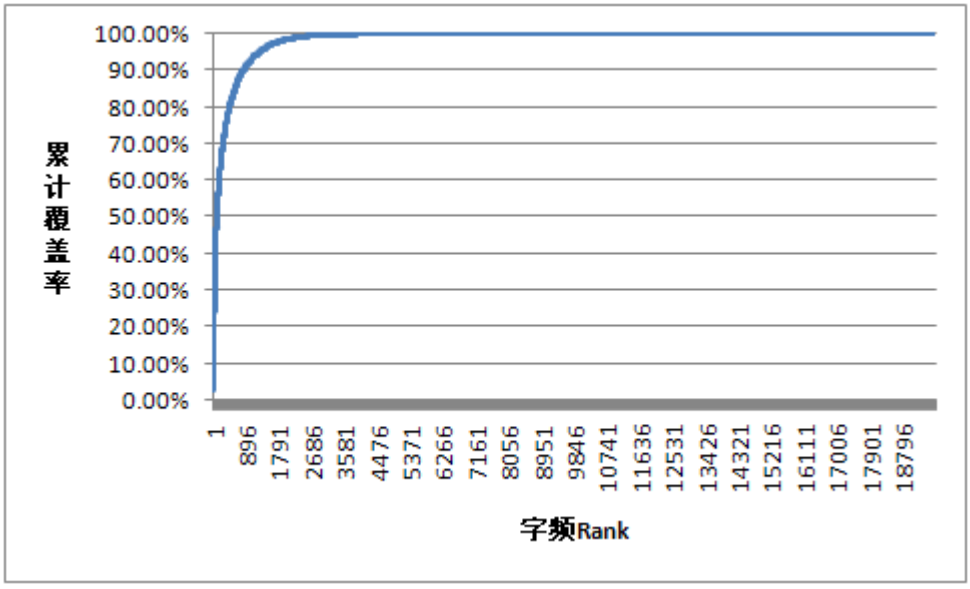

图2.2 用户词库累计覆盖率曲线

表2.1 用户词库字频前20的字

字	字频	累计字频	Rank	覆盖率	累计覆盖率
我	2044570926	2044570926	1	2.66%	2.66%
你	1957197208	4001768134	2	2.55%	5.21%
不	1849362646	5851130780	3	2.41%	7.62%
的	1729427093	7580557873	4	2.25%	9.87%
是	1443644870	9024202743	5	1.88%	11.75%
了	1366191672	10390394415	6	1.78%	13.53%
么	945551856	11335946271	7	1.23%	14.77%
有	914338373	12250284644	8	1.19%	15.96%
好	882177104	13132461748	9	1.15%	17.11%
个	794791091	13927252839	10	1.04%	18.14%
一	786687424	14713940263	11	1.02%	19.16%
没	769939061	15483879324	12	1.00%	20.17%
在	740057079	16223936403	13	0.96%	21.13%
就	692819581	16916755984	14	0.90%	22.03%
呵	660288698	17577044682	15	0.86%	22.89%
那	651385555	18228430237	16	0.85%	23.74%
要	586733172	18815163409	17	0.76%	24.51%
这	549162387	19364325796	18	0.72%	25.22%
来	546331533	19910657329	19	0.71%	25.93%
看	542732193	20453389522	20	0.71%	26.64%

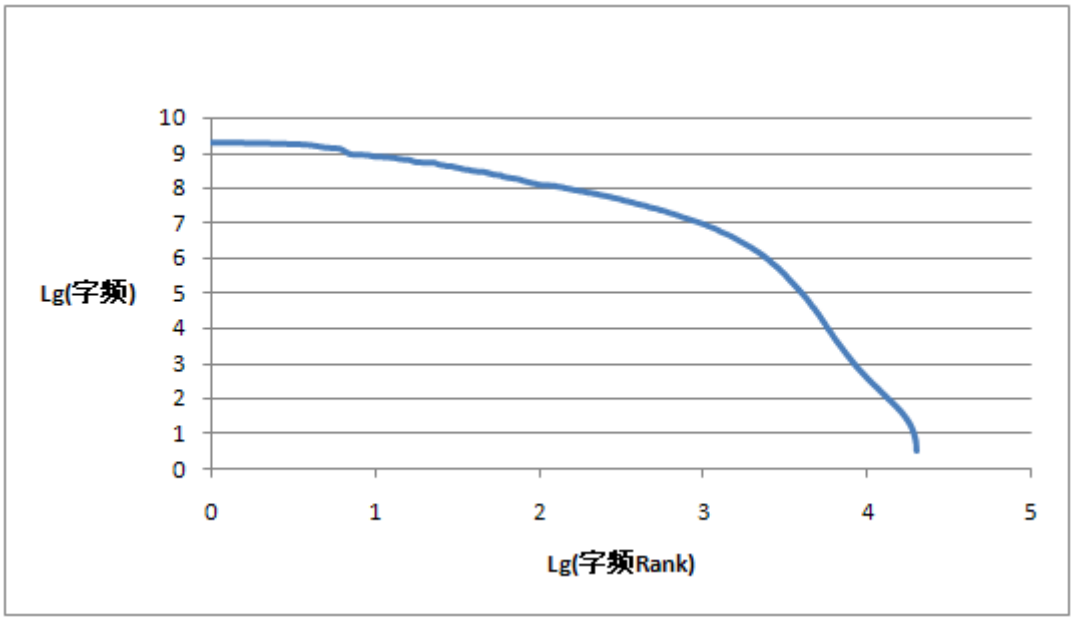

图2.3 用户词库Lg(字频)关于Lg(字频Rank)的曲线

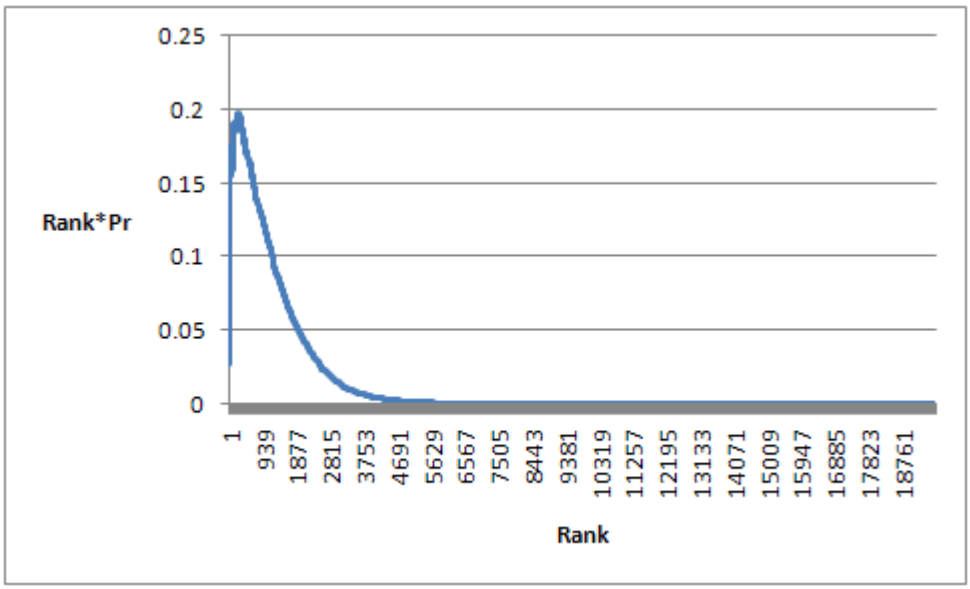

图2.4 用户词库Rank*Pr曲线

通过观察发现，用户词库中字频较高的字多为口头语中常用字，比如代词、语气词、助词等，有实际意义的较少。这与网络应用比如聊天工具、论坛成为输入法的主要应用渠道有关。

2.3.2 搜索引擎查询日志

查询日志中有单字 17715 个，总字频 5922947983，同样包含很多繁体字、古字、异体字等。字频最高的 20 个字如表 2.2。

表2.2 查询日志字频前20的单字

字	字频	累计字频	Rank	覆盖率	累计覆盖率
网	88756312	88756312	1	1.50%	1.50%
人	63484646	152240958	2	1.07%	2.57%
的	56817576	209058534	3	0.96%	3.53%
电	56496830	265555364	4	0.95%	4.48%
小	51567707	317123071	5	0.87%	5.35%
天	49460201	366583272	6	0.84%	6.19%
下	48611819	415195091	7	0.82%	7.01%
色	46581866	461776957	8	0.79%	7.80%
图	45773945	507550902	9	0.77%	8.57%
影	43526275	551077177	10	0.73%	9.30%
载	40495981	591573158	11	0.68%	9.99%
大	38612241	630185399	12	0.65%	10.64%
女	37353917	667539316	13	0.63%	11.27%
情	36121742	703661058	14	0.61%	11.88%
片	35055151	738716209	15	0.59%	12.47%
中	32249772	770965981	16	0.54%	13.02%
国	32125245	803091226	17	0.54%	13.56%
美	28335347	831426573	18	0.48%	14.04%
爱	27791941	859218514	19	0.47%	14.51%
学	27522927	886741441	20	0.46%	14.97%

查询日志累计覆盖率曲线如图 2.5，前 188 个字覆盖了 50%字频，前 658 个字覆盖了 80%字频，前 1102 个字覆盖了 90%字频，前 1577 个字覆盖了 95%的字频，前 2842 个字覆盖了 99%的字频。可见，无论是用户词库还是查询日志，都含有大量低频字，而且高频字使用比较集中，尤其是用户词库，86 个字占据了 50%的字频。

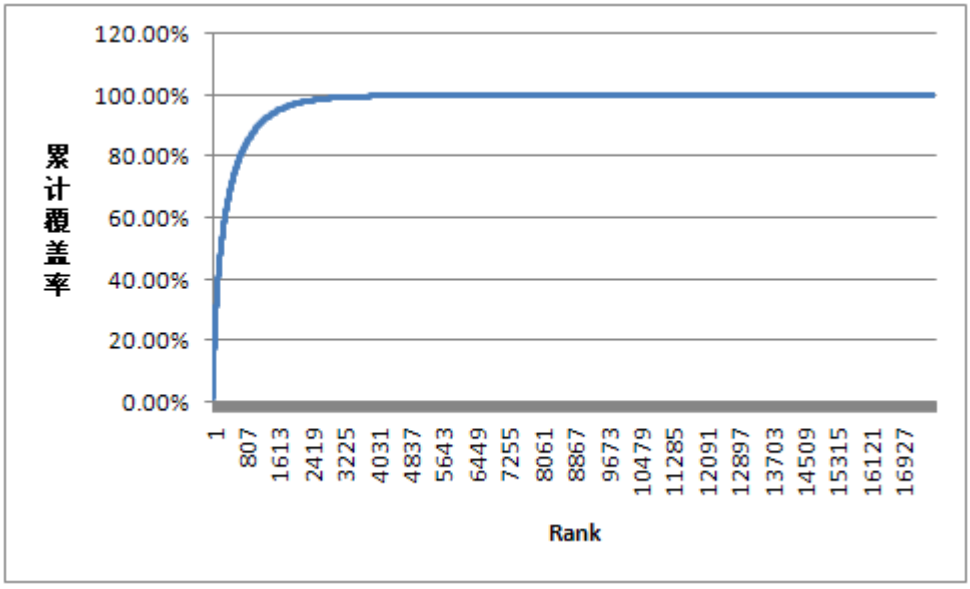

图2.5 查询日志累计覆盖率曲线

Lg(字频)关于 Lg(Rank)的曲线如图 2.6，线性相关系数-0.9395。曲线形状和用户词库类似，并不很好地符合 Zipf 定律。图 2.7 是每个单字的概率与 Rank 相乘的曲线，发现乘积变化较大，不是很符合 Zipf 定律。

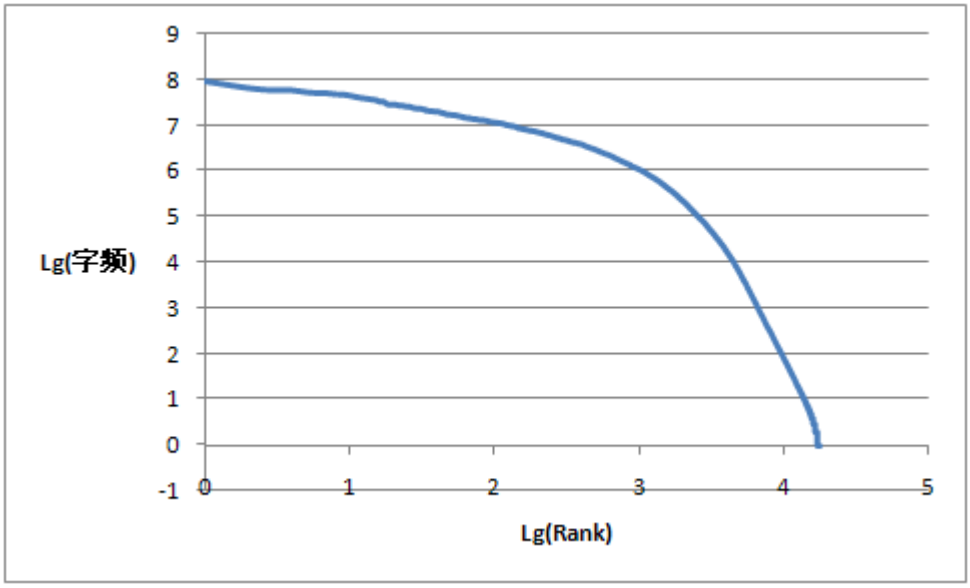

图2.6 查询日志Lg(字频)关于Lg(Rank)的曲线

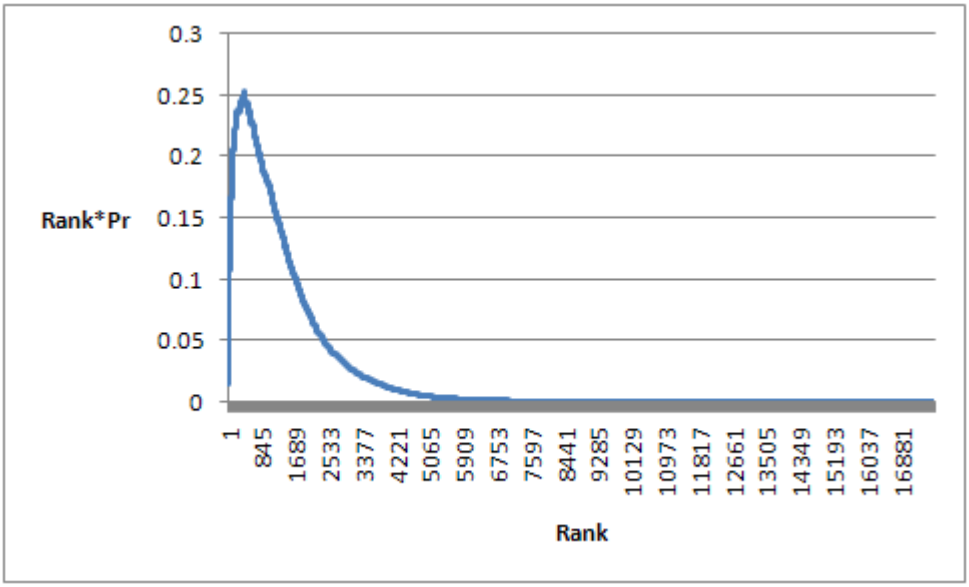

图2.7 查询日志Rank*Pr曲线

由于使用目的不同，查询日志中有实际意义的字频较高，虚词的字频相对用户词库低很多。字频较高的字多为名词、形容词的组成部分，这是与搜索引擎的功能相符的。

2.3.3 媒体常用字表

媒体字表含有单字 9270 个，总字频 991717782。字频最高的前 20 个字如表 2.3。

累计覆盖率曲线如图 2.8，前 181 个字覆盖了 50%，前 604 个字覆盖了 80%，前 970 个字覆盖了 90%，前 1372 个字覆盖了 95%，前 2381 个字覆盖了 99%。三个数据集对比来看，用户词库用字最集中。

Log(字频)关于 Log(Rank)的曲线如图 2.9，线性相关系数-0.8918，也不是很好地符合 Zipf 定律。图 2.10 是每个单字的概率与 Rank 相乘的曲线，发现乘积变化较大，不是很符合 Zipf 定律。

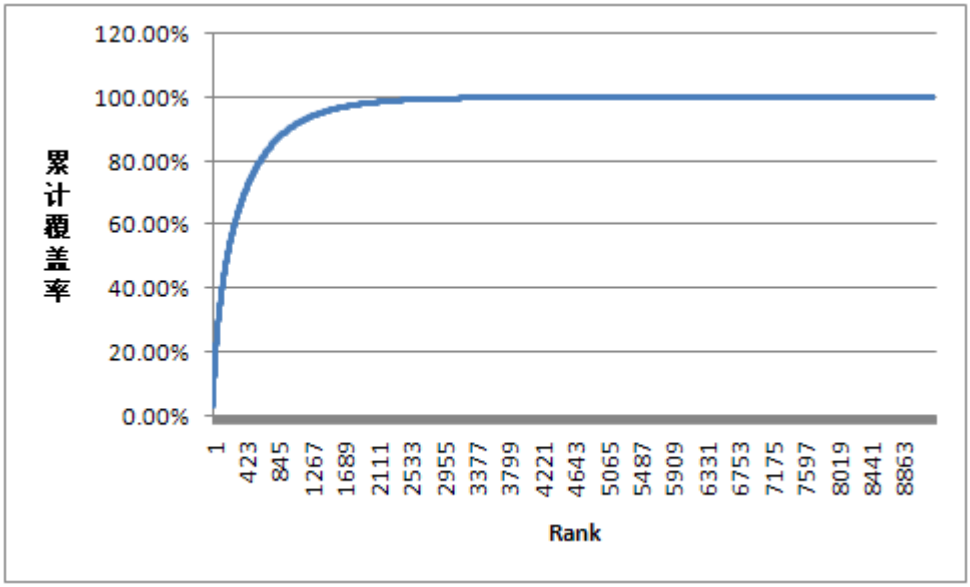

图2.8 媒体字表累计覆盖率

表2.3 媒体字表字频前20的单字

字	字频	累计字频	Rank	覆盖率	累计覆盖率
的	31651968	31651968	1	3.19%	3.19%
一	11018129	42670097	2	1.11%	4.30%
在	9270997	51941094	3	0.93%	5.24%
是	8733942	60675036	4	0.88%	6.12%
了	7937207	68612243	5	0.80%	6.92%
人	7578071	76190314	6	0.76%	7.68%
中	7545770	83736084	7	0.76%	8.44%
有	7214779	90950863	8	0.73%	9.17%
国	7037836	97988699	9	0.71%	9.88%
不	6754475	104743174	10	0.68%	10.56%
大	6493734	111236908	11	0.65%	11.22%
上	5600572	116837480	12	0.56%	11.78%
年	5402142	122239622	13	0.54%	12.33%
为	5251844	127491466	14	0.53%	12.86%
这	4857114	132348580	15	0.49%	13.35%
个	4807773	137156353	16	0.48%	13.83%
和	4752130	141908483	17	0.48%	14.31%
会	4528968	146437451	18	0.46%	14.77%
时	4432377	150869828	19	0.45%	15.21%
到	4293029	155162857	20	0.43%	15.65%

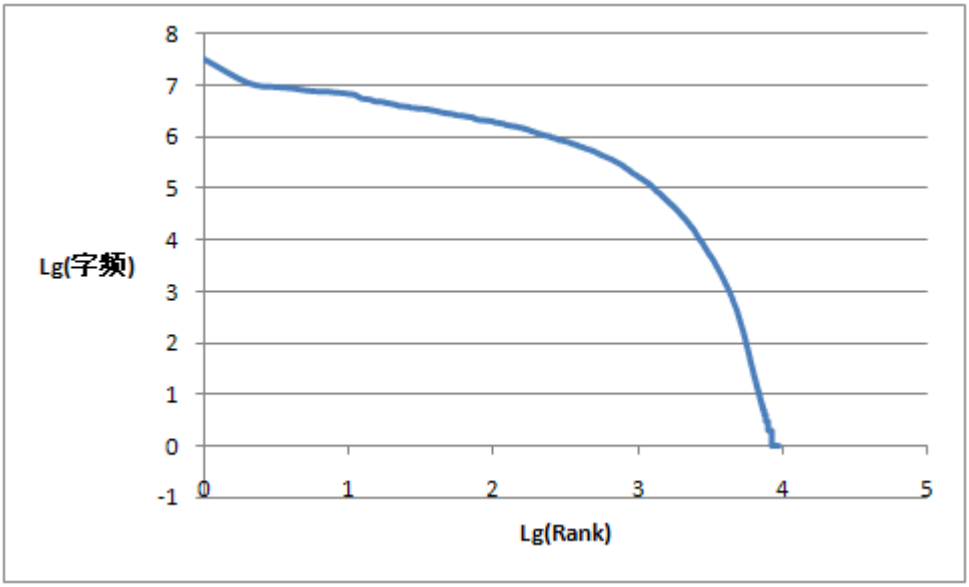

图2.9 媒体字表Lg(字频)关于Lg(Rank)的曲线

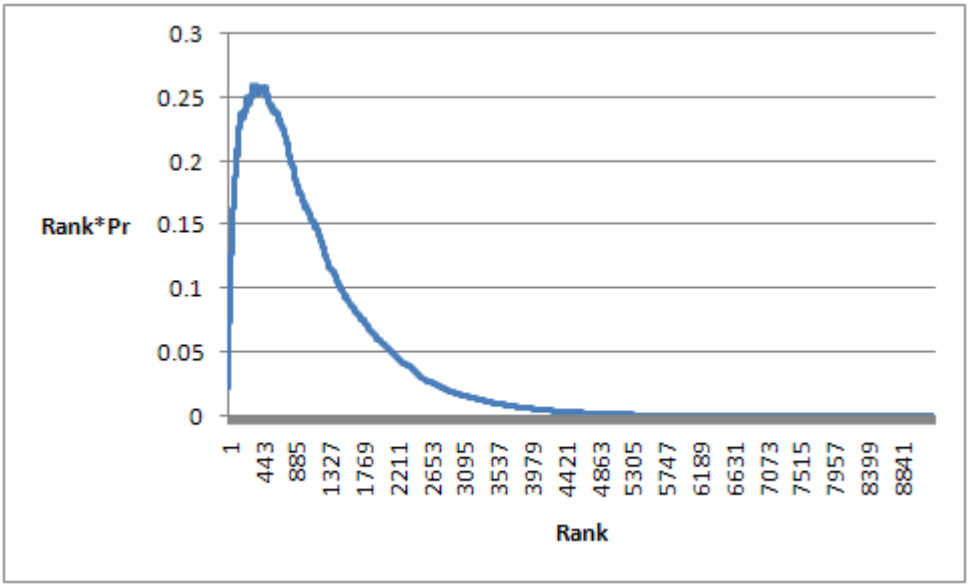

图2.10 媒体字表Rank*Pr曲线

三组数据字频覆盖情况总结如表 2.4。三组数据的共同特点是高频字使用集中。用户词库和查询日志虽然总字数很大，但是包含了大量非常用字，使用相对更集中一些。用户词库单字使用最为集中，少量的单字即可达到很高覆盖率。

表2.4 累计覆盖率所用字数对比

覆盖相应覆盖率所用字数	用户词库	查询日志	媒体字表
50%	87	188	181
80%	431	658	604
90%	808	1102	970
95%	1231	1577	1372
99%	2287	2842	2381
总字数	19679	17715	9270

2.4 常用七千字在三个数据集中的分布情况

2.4.1 常用七千字在用户词库中的分布

常用七千字中，有 6995 个在用户词库中出现，未出现的常用字如下：

轷、犮、侬、绾、鹲

5 个都是 3 级字，而且确实在日常生活中很少使用。

比较 1 级字 2500 个与用户词库中字频前 2500 个单字，共有部分有 2138 个字，非共有部分 362 个字，共有部分占 85.52%。用户词库独有的 362 个字中，有很多语气词，比如呵、嘛、嗯、嘿、哎、咯、哦、哇等；还有一些方言用字，比如冇等；还有一些网络流行字和“火星文”，比如囧、你、莪等；还有一些侮辱性、淫秽的单字。常用字独有的 362 个字中，比较多的是一些使用范围较窄、字意比较具体的字，比如坝、瓣、鞭等。

比较 1、2 级字 3500 个与用户词库前 3500 个单字，共有部分有 3048 个，非共有部分 452 个，共有部分占 87.09%。用户词库独有部分的成分与 2500 个字的情况基本一致。常用字独有的 452 个字中，1 级字 56 个，2 级字 396 个，1 级字未覆盖部分减少很多。

比较 1、2、3 级字 7000 个与用户词库前 7000 个单字，共有部分有 5821 个，非共有部分 1179 个，共有部分占 83.16%。用户词库独有的 1179 个字中，语气词大量减少，主要组成部分为“火星文”，另外有一些繁体字和粗俗淫秽字意的字。常用字的独有部分中，2 级字只有两个秕、蛉，其余都是 3 级字。可见在前 7000 常用字的范围内，用户词库和常用七千字的差别主要在于 3 级字。

2.4.2 常用七千字在查询日志中的分布

常用七千字中，有 6995 个在查询日志中出现，未出现的 5 个字如下：

拚、伋、阖、侬、馕

5 个都是 3 级字，其中侬也未在用户词库中出现。

比较 1 级字 2500 个与查询日志中前 2500 个字，共有部分有 2038 个，非共有部分 462 个，共有部分占 81.52%。略低于用户词库。查询日志的独有部分中，与用户词库不同，多是有实际意义的字，可能是新兴高频字，比如伦、聊、婷、炫等，另外还有一些低俗下流字意的字。常用字独有部分主要还是使用面较窄的字。

比较 1、2 级字 3500 个与查询日志前 3500 个字，共有部分有 2883 个，非共有部分 617 个，共有部分占 82.37%，低于用户词库。查询日志独有部分情况与前 2500 字相似。常用字独有部分中，1 级字 154 个，2 级字 463 个，1 级字比例相比用户词库较大。

比较 1、2、3 级字 7000 个与查询日志前 7000 个字，共有部分有 5682 个，非共有部分 1318 个，共有部分占 81.17%，低于用户词库。查询日志的独有部分中，与用户词库类似，有较多低俗下流字意的字和“火星文”、繁体字。常用字独有部分中，2 级字 11 个：秕、盹、踱、饺、譬、嘁、迄、秫、锨、舀、谆。其余都是 3 级字。

与用户词库类似，在前 7000 常用字的范围内，查询日志和常用七千字的差别主要在于 3 级字。与用户词库相比，查询日志由于其使用目的的倾向性，多为有具体意义的查询词，因此与常用七千字的相关度低于用户词库。

2.4.3 常用七千字在媒体字表中的分布

常用七千字中，有 6803 个在查询日志中出现，未出现的 197 个字全部为 3 级字。出现的常用字较少可能和媒体字表字数较少只有 9270 个有关。

比较 1 级字 2500 个与媒体字表中前 2500 个字，共有部分有 2156 个，非共有部分 344 个，共有部分占 86.24%，高于用户词库和查询日志。媒体字表的独有部分中，与用户词库不同，多是有实际意义的字，其中有不少是中外人名、地名的组成汉字，比如萨、诺、迪、菲、廖等。

比较 1、2 级字 3500 个与媒体字表前 3500 个字，共有部分有 3103 个，非共有部分 397 个，共有部分占 88.66%，高于用户词库和查询日志。媒体字表独

有部分情况与前 2500 字相似。常用字独有部分中，1 级字 49 个，2 级字 348 个，1 级字比例相比用户词库和查询日志较小。

比较 1、2、3 级字 7000 个与媒体字表前 7000 个字，共有部分有 6391 个，非共有部分 609 个，共有部分占 91.30%，高于用户词库和查询日志。媒体字表的独有部分中，有较多繁体字。常用字独有部分中，都是 3 级字。

媒体字表由于来源较正式，因此和常用七千字的相关程度较高。

常用七千字在三个数据集中的分布情况总结对比如表 2.5。

表2.5 常用七千字分布情况对比

各部分对应共有部分字数	用户词库	查询日志	媒体字表
前 2500	2138	2038	2156
前 3500	3048	2883	3103
前 7000	5821	5682	6391
全部	6995	6995	6803

2.5 不同数据集之间单字分布比较

2.5.1 用户词库与查询日志单字分布比较

用户词库与查询日志的共有部分共有 16694 个，可见大部分用字相同。共有部分 Rank 的 Spearman 相关系数为 0.8214。共有部分在两个数据集中 Rank 相差大的几乎都是非常用字。在用户词库的独有部分中，主要包含各种繁体字。在查询日志的独有部分中，主要包含生僻字。

比较两个数据前 2500 个字，共有部分 2141 个，非共有部分 359 个。共有部分 Rank 的 Spearman 相关系数为 0.6593。用户词库的独有部分中，主要包含语气字和较口语化的字，以及少量网络流行字等。查询日志的独有部分中，有实际意义的字较多，还有不少低级下流字意的字。

比较两个数据前 3500 个字，共有部分有 3038 个。共有部分 Rank 的 Spearman 相关系数为 0.7568。用户词库独有部分中，还是主要包含语气字、网络流行字等。查询日志独有部分中多为有实际意义的字。

比较前 7000 个字，共有部分有 5782 个。共有部分 Rank 的 Spearman 相关系数为 0.8763。用户词库独有部分中，主要是火星文和繁体字。查询日志独有部分中，主要是繁体字等。

比较两组数据中覆盖 80%字频的集合，用户词库 431 字，查询日志 658 字。共有部分 294 个字，用户词库独有部分 137 字，查询日志独有部分 364 字。共有部分 Rank 的 Spearman 相关系数为 0.1028，说明共有部分差异较大。与之前类似，用户词库独有部分中包含较多口语化的字，比如语气词。

比较覆盖 90%字频的集合，用户词库 808 字，查询日志 1102 字。共有部分 628 字，用户词库独有部分 181 字，查询日志独有部分 475 字。共有部分 Rank 的 Spearman 相关系数为 0.3300，共有部分差异较大。独有部分的分布情况和 80%时类似。

比较覆盖 99%字频的集合，用户词库 2287 字，查询日志 2842 字。共有部分 2095 字，用户词库独有部分 192 字，查询日志独有部分 747 字。共有部分 Rank 的 Spearman 相关系数为 0.6305，相关性有所提高。独有部分分布情况和之前仍然类似。

用户词库中高频字有较多网络流行因素，语气字、流行词使用的字、火星文等较多。查询日志中有实际字意，有助于查询的字较多。两者在高频字段的相关性不高。

2.5.2 用户词库与媒体字表比较

媒体字表与用户词库共有部分 9215 字，占媒体字表绝大部分，说明大部分用字相同。一些非常用字在两者中的 Rank 相差较大。共有部分 Rank 的 Spearman 相关系数为 0.9254，高于查询日志与用户词库。

比较用户词库前 2500 个字与媒体字表，共有部分有 2216 个，独有部分 284 个，共有部分数量高于查询日志与用户词库。共有部分 Rank 的 Spearman 相关系数为 0.7882，高于用户词库与查询日志前 2500 的相关系数。用户词库的独有部分中，与之前类似，主要是语气字以及繁体字、火星文等。媒体字表的独有部分中的字相对比较有实际意义。

比较两者的前 3500 个字，共有部分有 3162 个，独有部分 338 个，共有部分数量高于查询日志与用户词库。共有部分 Rank 的 Spearman 相关系数为 0.8437，高于用户词库与查询日志前 3500 的相关系数。独有部分的组成与前 2500 字比较中的情况类似。

比较两者的前 7000 个字，共有部分有 6013 个，独有部分 987 个，共有部分数量高于查询日志与用户词库。共有部分 Rank 的 Spearman 相关系数为

0.9156，高于用户词库与查询日志前 7000 的相关系数。独有部分用户词库主要包含火星文等，媒体字表主要是非常用的生僻字。

比较两者的前 80%部分，用户词库 431 字，媒体字表 604 字。共有部分有 343 个，用户词库独有部分 88 个，媒体字表独有部分 261 个。共有部分数量高于查询日志与用户词库。共有部分 Rank 的 Spearman 相关系数为 0.4656，高于用户词库与查询日志的相关系数，但是共有部分差异仍然较大。

比较两者的前 90%部分，用户词库 808 字，媒体字表 970 字。共有部分有 669 个，用户词库独有部分 139 个，媒体字表独有部分 301 个。共有部分数量高于查询日志与用户词库。共有部分 Rank 的 Spearman 相关系数为 0.5644，高于用户词库与查询日志的相关系数。

比较两者的前 99%部分，用户词库 2287 字，媒体字表 2381 字。共有部分有 2039 个，用户词库独有部分 248 个，媒体字表独有部分 342 个。共有部分数量略低于查询日志与用户词库。共有部分 Rank 的 Spearman 相关系数为 0.7787，高于用户词库与查询日志的相关系数。

由于媒体字表来源比较正式，相比用户词库，有实际字意的字频度较高。高频字段两者的相关性不高，相比用户词库与查询日志的相关性要高。

2.5.3 查询日志与媒体字表比较

媒体字表与查询日志共有部分 8453 字，占媒体字表大部分，说明大部分用字相同。一些非常用字在两者中的 Rank 相差较大。共有部分 Rank 的 Spearman 相关系数为 0.9145，高于查询日志与用户词库。

比较查询日志前 2500 个字与媒体字表，共有部分有 2147 个，非共有部分 353 个，共有部分数量略高于查询日志与用户词库。共有部分 Rank 的 Spearman 相关系数为 0.6579，略低于查询日志与用户词库前 2500 的相关系数。查询日志的独有部分中，与之前类似，包含一些低俗下流的字等。媒体字表的独有部分中的字相对比较有实际意义。

比较两者的前 3500 个字，共有部分有 3023 个，非共有部分 477 个，共有部分数量略低于查询日志与用户词库。共有部分 Rank 的 Spearman 相关系数为 0.7547，略低于查询日志与用户词库前 3500 的相关系数。非共有部分部分的组成与前 2500 字比较中的情况类似。

比较两者的前 7000 个字，共有部分有 5781 个，非共有部分 1219 个，共有

部分数量略低于查询日志与用户词库。共有部分 Rank 的 Spearman 相关系数为 0.8760，略低于查询日志与用户词库前 7000 的相关系数。独有部分查询日志主要包含繁体字和脏话等。

比较两者的前 80%部分，查询日志 658 字，媒体字表 604 字。共有部分有 390 个，查询日志独有部分 268 个，媒体字表独有部分 214 个。共有部分数量高于查询日志与用户词库。共有部分 Rank 的 Spearman 相关系数为 0.3129，高于查询日志与用户词库的相关系数，但是共有部分差异仍然较大。

比较两者的前 90%部分，查询日志 1102 字，媒体字表 970 字。共有部分有 751 个，查询日志独有部分 351 个，媒体字表独有部分 219 个。共有部分数量高于查询日志与用户词库。共有部分 Rank 的 Spearman 相关系数为 0.4160，高于查询日志与用户词库的相关系数，但是共有部分差异仍然较大。

比较两者的前 99%部分，查询日志 2842 字，媒体字表 2381 字。共有部分有 2177 个，查询日志独有部分 665 个，媒体字表独有部分 204 个。共有部分数量高于查询日志与用户词库。共有部分 Rank 的 Spearman 相关系数为 0.6561，高于查询日志与用户词库的相关系数。

查询日志与媒体字表的相关程度也不高，比用户词库与媒体字表的相关程度要低。

将三个数据集比较的结果整理如表 2.6。

表2.6　三个数据集比较结果整理

共有部分字数和相关系数	用户词库与查询日志		用户词库与媒体字表		查询日志与媒体字表	
前 80%	294	0.1028	343	0.4656	390	0.3129
前 90%	628	0.3300	669	0.5644	751	0.4160
前 99%	2095	0.6305	2039	0.7787	2177	0.6561
前 2500	2141	0.6593	2216	0.7882	2147	0.6579
前 3500	3038	0.7568	3162	0.8437	3023	0.7547
前 7000	5782	0.8763	6013	0.9156	5781	0.8760
全部	16694	0.8214	9215	0.9254	8453	0.9145

2.6　小结

输入法用户词库可以看成中文用户在电脑中使用语言的较直接体现，用字

比较生活化口语化，差别最大的就是语气字的词频往往远高于其他数据。一些在网络流行的新词包含的字在输入法中的字频也较高。用户词库与七千常用字的相关程度要高于查询日志。

查询日志记录查询词，用字也是查询词的组成部分，虚词部分较少，有实际意义字的字频要高于其他数据。查询日志与其他数据的相关程度也较差。

媒体字表主要是书面语构成，书面语尤其是新闻报道常用字的频度较高，与前两者的相关程度几乎相当。

第3章 输入法用户词库和查询日志用词情况分析

3.1 实验概述

本章通过搜狗输入法用户词库数据、搜索引擎查询日志和其他现有数据，对输入法用户使用词的语言状况进行统计分析。

3.2 数据说明

3.2.1 输入法用户词库

本实验使用的是搜狗输入法 2010 年 3 月 15 日的用户词库数据，该数据统计了所有注册用户使用的词条的用户数和词频。和第二章使用数据类似，该词库将用户输入的上屏词条整体作为一个词条保存，保存的词条不一定是语言学意义上的词。

用户词库记录了词条长度不超过 7 的词条，2010 年 3 月 15 日的数据共 111659347 个词条，总词频 327029776076，平均词频 2928.817。用户数达到百万级。

3.2.2 搜索引擎查询日志

本实验使用的查询日志是搜狗搜索引擎的日志数据，每日统计在搜狗搜索引擎上的查询词和查询次数。日志数据中只保留了 2 字词至 7 字词，对英文、数字等非汉字字符进行了全角化处理。查询日志记录了 2009 年全年的查询情况，有词条 81970629 个，总词频 1766113757。

3.2.3 其他数据

本实验还使用了普通话常用三千词词表[8]，包括 3815 个词。该词表发表于 1992 年，可视为传统汉语常用词，可以用来与用户词库等较新数据对比。该词条包含 1009 个单字、2571 个双字词、204 个三字词、26 个四字词、3 个五字词、2 个七字词。该词表没有词频信息。

3.3 输入法用户词库用词情况分析

3.3.1 总体情况

用户词库词频前 20 的词条如表 3.1。

表3.1 输入法用户词库词频前20的词条

词条	词频	用户数	Rank
啊	3990088189	3578525	1
了	3848708636	7437262	2
就	2940184432	4862211	3
在	2564723683	5028618	4
好	2474029080	4981193	5
的	2414096285	5279846	6
我	2382157145	5038798	7
有	2361758812	4993551	8
呵呵	2234793465	4683200	9
没	2225026282	4937635	10
吧	2168180403	4866926	11
去	2075258235	4857055	12
要	1974892907	4877510	13
都	1964884870	5648914	14
那	1934999163	5344280	15
什么	1773690404	5271040	16
恩	1762991445	4409577	17
说	1737688923	5107365	18
呢	1701677361	5851196	19
也	1694687752	4835222	20

将用户词库的词条按词频排序并计算累计覆盖率，前 32000 个词条的累计覆盖率曲线如图 3.1。覆盖总词频百分比所需要的词条数和词条数所占比例如表 3.2。可见输入法用户用词相当集中，不到 1%的词条就可以覆盖绝大多数词频。

图 3.2 是用户词库 Log（词频）关于 Log（Rank）的曲线，线性相关系数 r = -0.9998，Log（词频）与 Log（Rank）负线性相关，符合 Zipf 定律。

表3.2 累计覆盖率与所需词条数

累计覆盖率	所需词条数	词条数百分比
50%	523	0.0005%
60%	1284	0.0011%
70%	3532	0.0032%
80%	12634	0.0113%
90%	93670	0.0839%
95%	600340	0.5377%

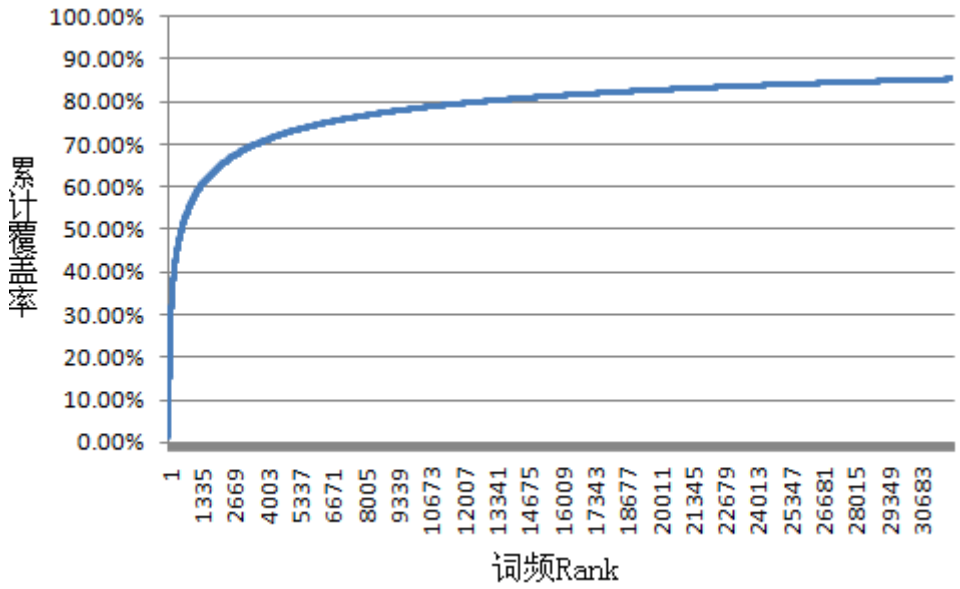

图3.1 用户词库累计覆盖率曲线

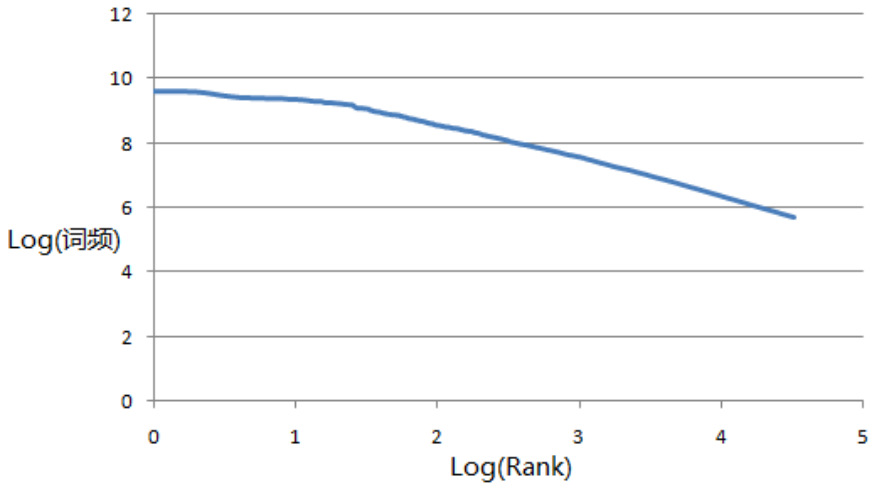

图3.2 用户词库Log(词频)关于Log(Rank)的曲线

表 3.3 给出了 2010 年 3 月 15 日的用户词库按不同词长统计的分布情况。其中单字有 2 万多个，远远超出常用汉字个数（常用汉字大约 2000-7000 个）。这是因为用户词库中的词条并不标准，包含大量繁体字、异体字等。

表3.3　2010-3-15用户词库长短词分布

词长	词条数	总词频
1	25775	129393411999
2	4061157	146987428400
3	26841194	34609501910
4	41409430	12332722427
5	24144590	2688079719
6	11761215	820903823
7	3415986	197727798

图 3.3 是词条数与词条长度的关系图。从图中可以看出，单字词条数最少，因为只是单字个数；随着词条长度上升，单字的组合增多，词条数也相应增多，但是不同组合能成为词的概率也减少。3 字时词条数增长迅速，在词条长度为 4 时达到峰值，词条长度大于 4 时词条总数开始下降。

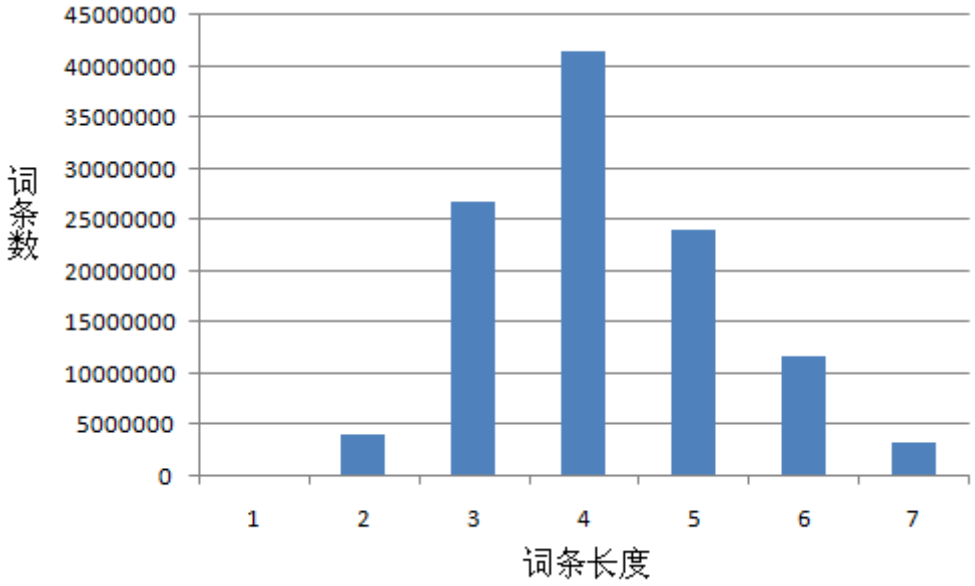

图3.3　用户词库词条数关于词条长度的分布

图 3.4 是词频与词条长度的关系图。双字词词频最高，单字其次。单字词和双字词的词频远大于多字词。

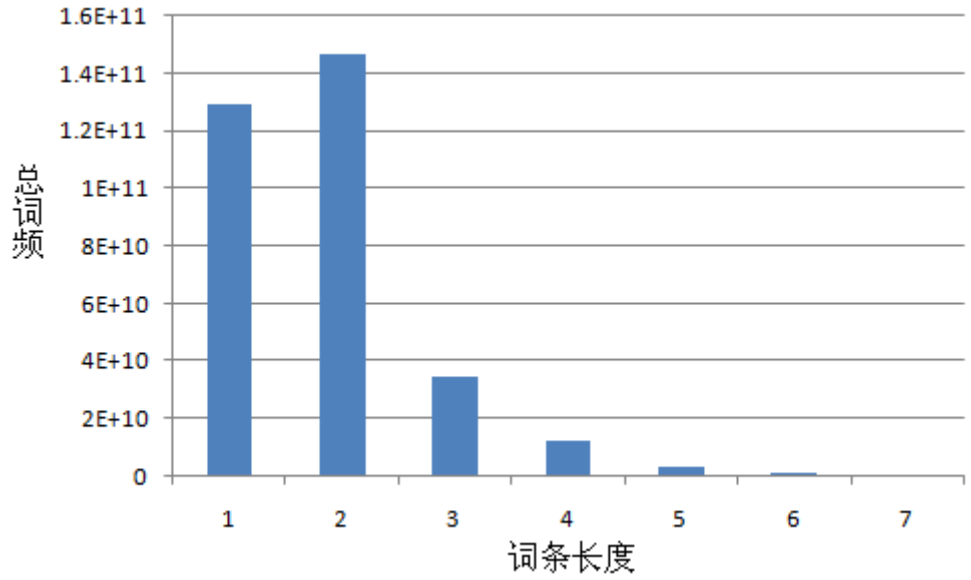

图3.4 用户词库词频关于词条长度的分布

图 3.5 是评价词频与词长的关系图。单字的平均输入频度远远大于双字词和多字词。双字词和多字词的平均输入频度几乎不可视。

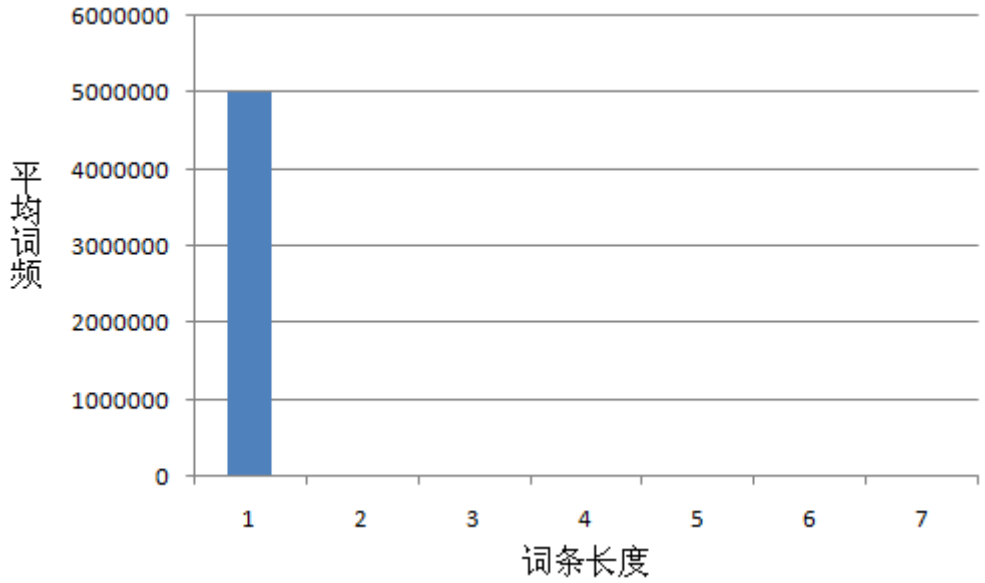

图3.5 用户词库平均词条输入频度关于词条长度的分布

以上数据表明，输入法用户在输入时更偏向于每次输入较短词条而不是多字的词条，尤其以单字、双字词为主。这可能因为输入时以词为单位的语言习

惯，另外也有输入法在长词条时的组词效果不如短词条的因素影响。

3.3.2 三千常用词分布情况

普通话常用三千词共 3815 个词条中，在用户词库中出现了 3814 个，总词频 177003381336，覆盖了用户词库 54.12%。如表 3.4 所示。

不在用户词库中出现的词条只有“留声片”一词。“留声片”由于时代发展已经不是很常用的词。

表3.4 普通话三千常用词在用户词库的分布

词条长度	词频	词条数	平均词频
1	114064595361	1009	113047171
2	62000330464	2571	24115258.8
3	930639369	203	4584430.39
4	7367559	26	283367.654
5	423340	3	141113.333
6	0	0	0
7	25243	2	12621.5

图 3.6 是用户词库中出现的三千常用词词频与词条长度的关系图。可以看出，单字词频最高，双字词其次，多字词词频较少。与用户词库结果不同的是，双字词词频不再是最高，多字词词频比例也下降，这与三千常用词中收录的双字词和多字词不像用户词库数量庞大有关。

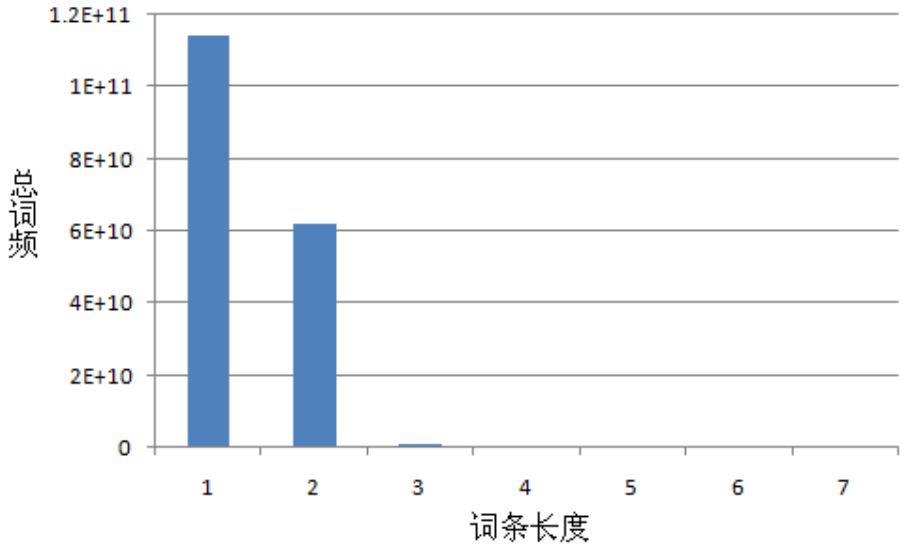

图3.6 三千常用词词频关于词条长度的分布

图 3.7 是用户词库中出现的三千常用词与词长的关系图。单字的平均词频要高于其他。由于收录的双字词和多字词数量较少且为常用词，双字词和多字词的评论词频相对用户词库也有所提高。

在用户词库中，词频较高的前 20 个三千常用词如表 3.5，与用户词库基本相同。

词频小于 1000 的三千常用词词条如表 3.6。这些词在现在确实不常用，如“理发员”一词现在多用其他词代替。

表3.5　在用户词库中词频较高的三千常用词词条

词条	词频	用户数
啊	3990086631	3578495
了	3775294971	4888705
就	2940163200	4862158
在	2564717059	5028565
好	2474022104	4981139
的	2412153469	5047074
我	2382109773	5038736
有	2361752018	4993452
没	2224023575	4849208
吧	2168168227	4866854
去	2075251508	4856995
要	1974888092	4877470
都	1932930570	4743350
那	1931709047	4883441
什么	1749444318	4867118
说	1732959146	4832550
也	1694679714	4835188
呢	1669596557	4741654
个	1621582548	4842391
还	1582125638	4773274

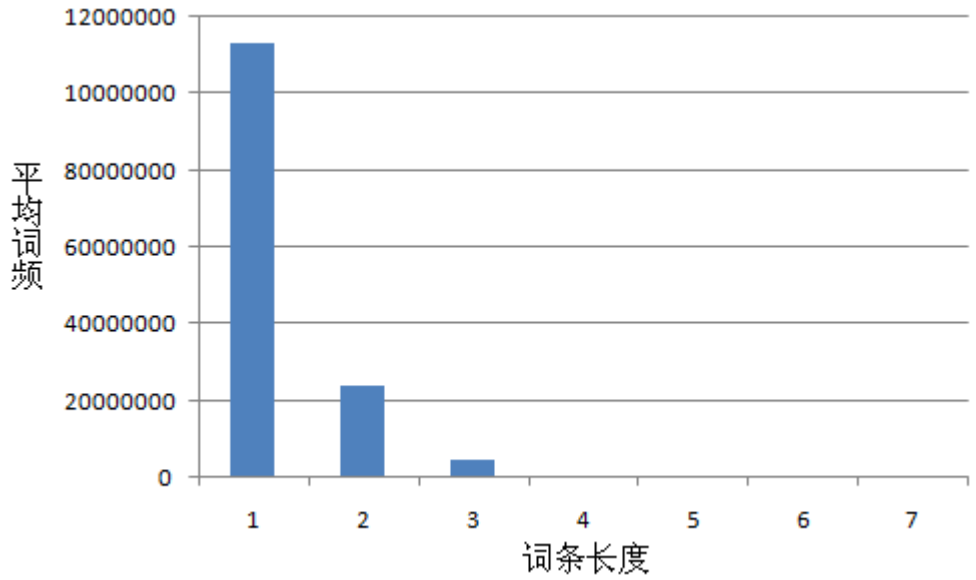

图3.7　三千常用词平均词频关于词条长度的分布

表3.6　三千常用词中词频<1000的词条

词条	词频	用户数
理发员	813	628
铁甲车	433	271
人家儿	275	245
双铧犁	223	87
枝对	50	43

三千常用词的累计覆盖率曲线如图 3.8，覆盖三千常用词词频相应覆盖率所需要的词条数和词条数百分比如表 3.7。

表3.7　覆盖三千常用词词频覆盖率所需要的词条数和词条数所占比例

累计覆盖率	词条数	词条百分比
50%	74	1.94%
60%	124	3.25%
70%	208	5.45%
80%	374	9.81%
90%	753	19.74%
95%	1206	31.62%

三千常用词 Log（词频）关于 Log（Rank）的曲线如图 3.9，高频部分基本呈线性。经过计算得线性相关系数 r = -0.8797，整条曲线并不符合 Zipf 定律。

三千常用词低频词汇稀少，所以低频部分迅速下降。

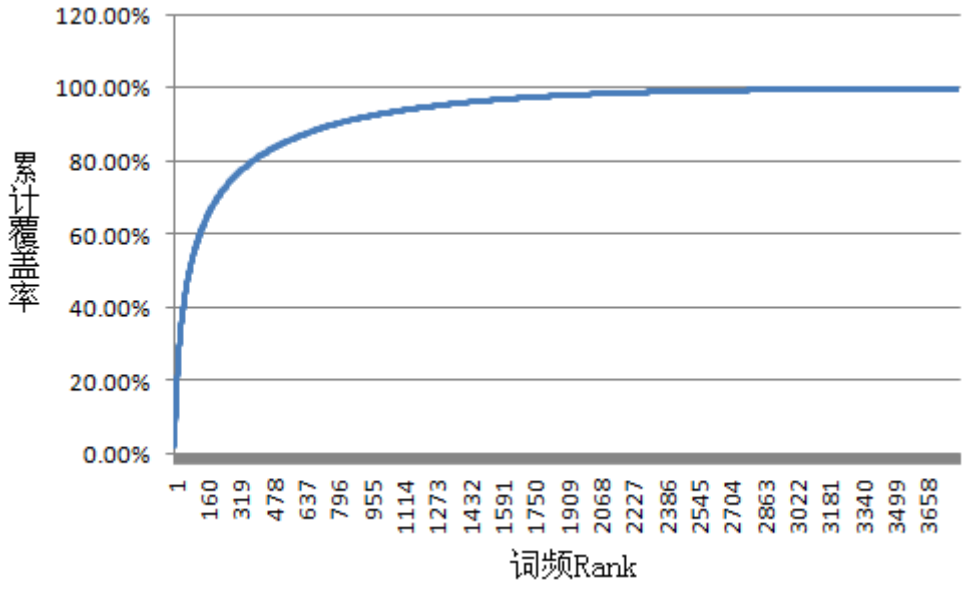

图3.8 三千常用词累计覆盖率曲线

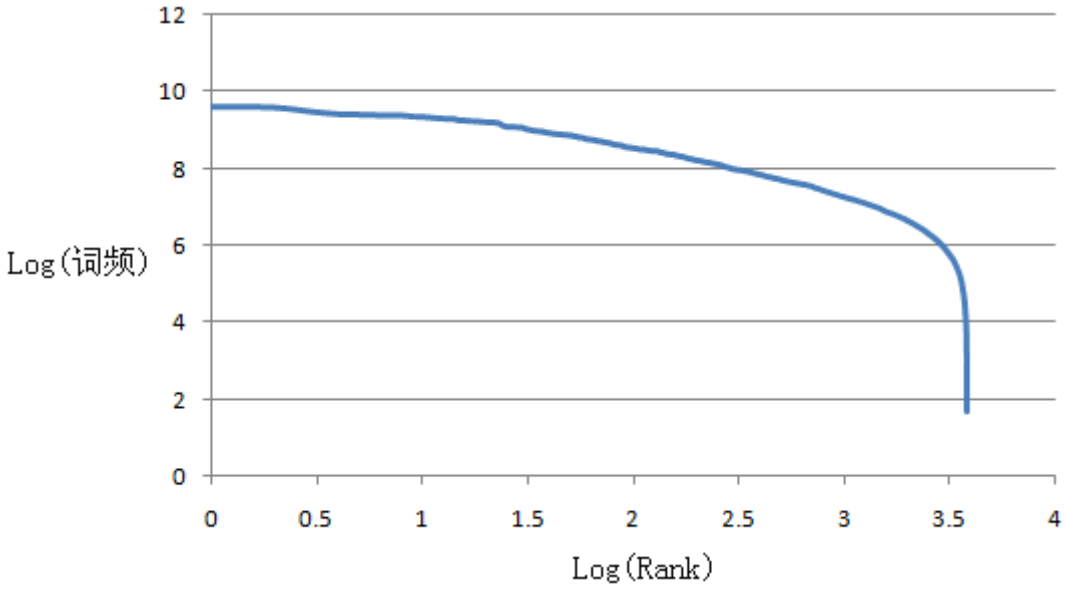

图3.9 三千常用词用于验证Zipf定律的曲线

三千常用词在用户词库中的总覆盖率为 56.39%，用户词库前 3000 词可覆盖约 70%的覆盖率。

以上数据说明三千常用词整体在用户词库中也占据了大多数词频，大多数

常用词在用户词库中也有较高词频。由于时代发展，三千常用词有些词条已经过时少有使用。

3.4 查询日志用词情况分析

3.4.1 总体情况

查询日志的词条可以分成三类：纯汉字词条（如“百度”）、纯非汉字词条（如“mp3”）、混杂词条（如“qq 空间”）。下面将就全部词条和分类词条分别进行分析。

3.4.1.1 全部词分布情况

经过计算整理，查询日志有词条 81970629 个，总词频 1766113757。排名前 20 的词如表 3.8。

表3.8 查询日志中排名前20的词条

词	词频	累计词频	累计覆盖率	Rank
m p 3	10309694	10309694	0.58%	1
b a i d u	8596338	18906032	1.07%	2
h t t p	8370998	27277030	1.54%	3
百度	6678435	33955465	1.92%	4
n b a	5772458	39727923	2.25%	5
视频	4292585	44020508	2.49%	6
下载	3688498	47709006	2.70%	7
x i a o n e i	3592199	51301205	2.90%	8
d n f	3552056	54853261	3.11%	9
电影	3339722	58192983	3.29%	10
51	3293443	61486426	3.48%	11
c o m	3144447	64630873	3.66%	12
163	3078896	67709769	3.83%	13
人体艺术	3048060	70757829	4.01%	14
小说	2886274	73644103	4.17%	15
x i x i	2829046	76473149	4.33%	16
s i t e	2727204	79200353	4.48%	17
迅雷	2718826	81919179	4.64%	18
q q	2711787	84630966	4.79%	19
开心网	2667120	87298086	4.94%	20

由于查询词的特点，对应具体网站、软件、物品等查询目标的词条频度较高，平时语言常用词汇频度较低。

覆盖率曲线如图 3.10。

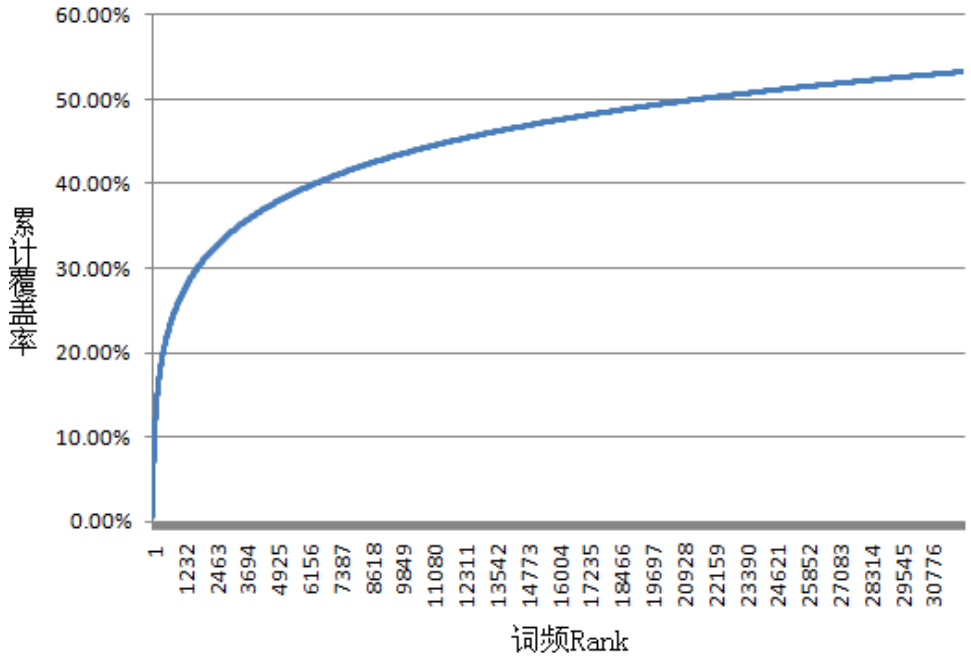

图3.10　查询日志词条累计覆盖率曲线

前 32000 个词 Log(词频)关于 Log(Rank)曲线如图 3.11，线性相关系数 -0.9997，符合 Zipf 定律：

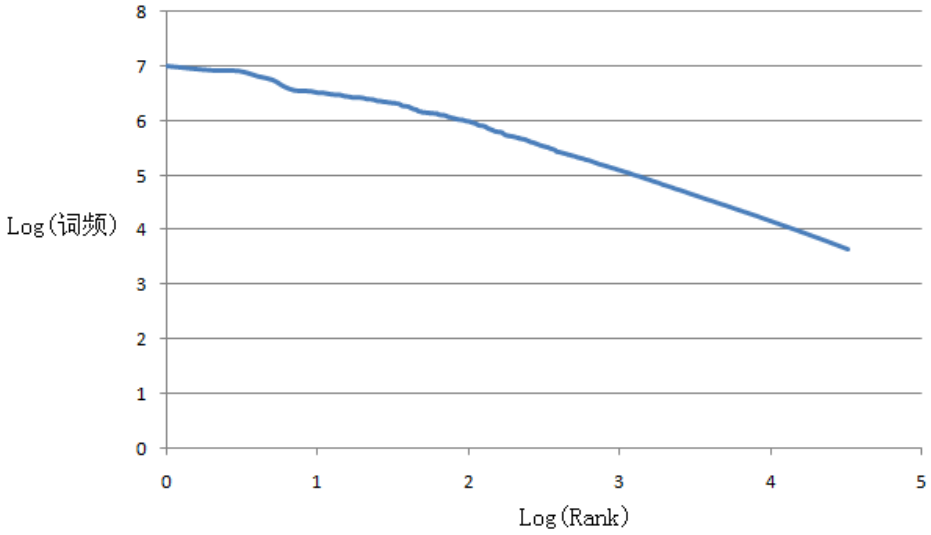

图3.11　查询日志词条Log(词频)关于Log(Rank)的曲线

不同词长分布情况如表 3.9。

表3.9 查询日志中词条在不同词长上的分布

词长	总个数	总词频	平均词频
2	1145292	287926299	251.3999
3	5484417	263844297	48.10799
4	13285774	434972152	32.73969
5	16345202	276660392	16.92609
6	25239961	284017166	11.25268
7	20469983	218693451	10.68362
全部	81970629	1766113757	21.54569

词条数随词长变化图如图 3.12，可见 6 字词之前递增，6 字词最多。

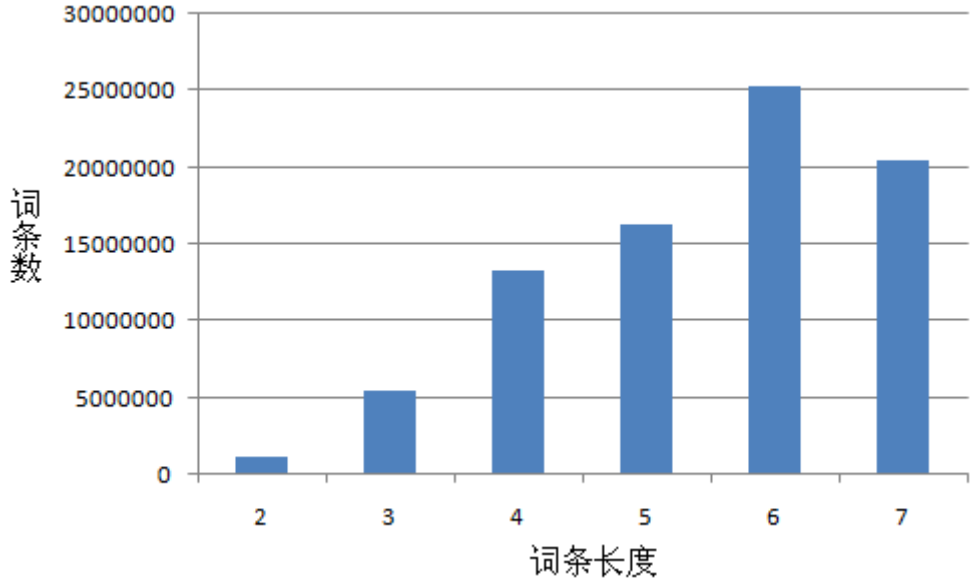

图3.12 不同词长查询日志词条数分布

总词频随词长变化图如图 3.13，整体来说比较平均，4 字词较高。并没有出现类似用户词库的双字词频较高的情况。可能因为双字的信息有限，四字词能更好地表达查询目的。

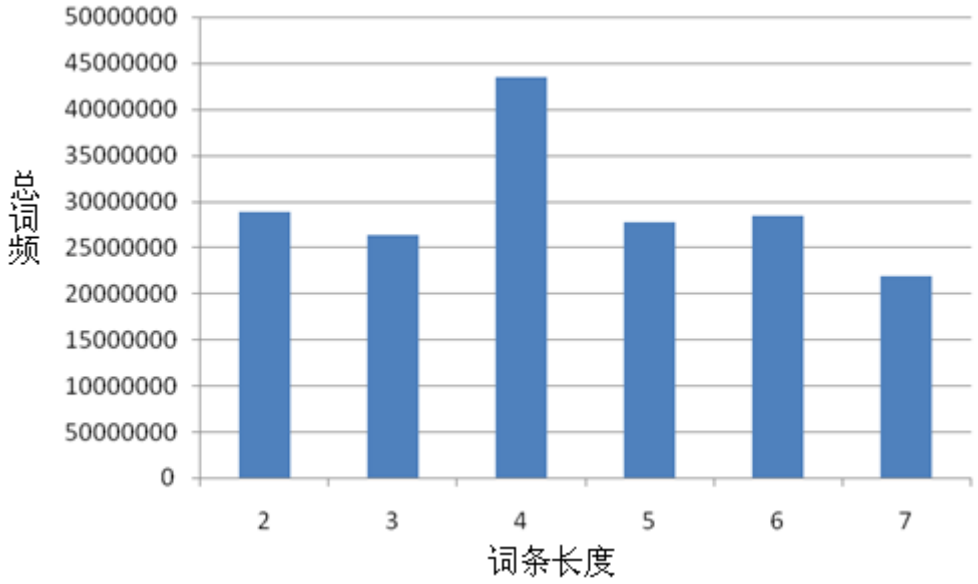

图3.13 不同词长查询日志词频分布

平均词频随词长变化如图 3.14，2 字词远高于多字词，随词长递减。这与用户词库的结果相似。

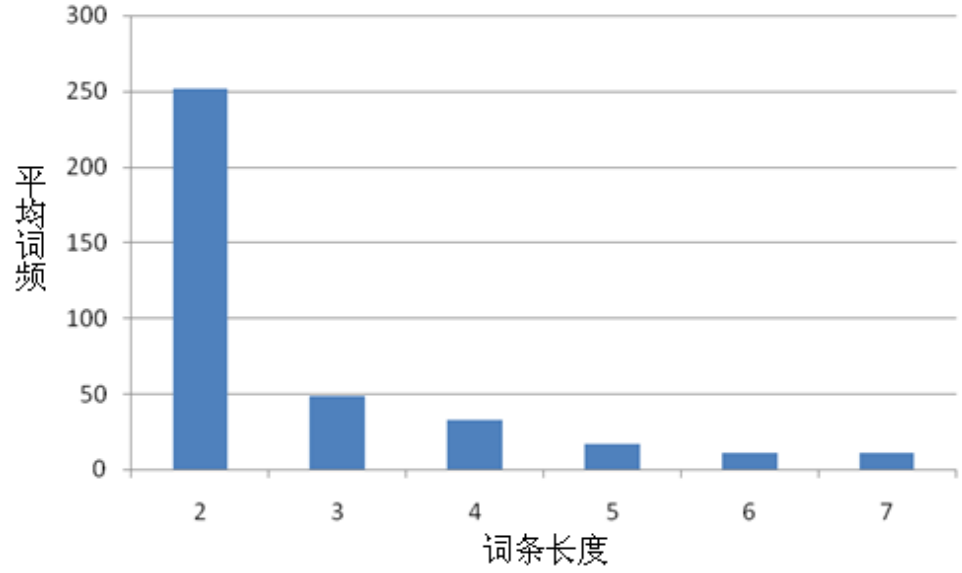

图3.14 不同词长查询日志平均词频分布

3.4.1.2 纯汉字词分布情况

经过计算整理，查询日志有纯汉字词条 67312912 个(82.12%)，总词频 1272261067(72.04%)。排名前 20 的纯汉字词如表 3.10。

表3.10 查询日志中排名前20的纯汉字词条

词	词频	累计词频	累计覆盖率	Rank
百度	6678435	6678435	0.52%	1
视频	4292585	10971020	0.86%	2
下载	3688498	14659518	1.15%	3
电影	3339722	17999240	1.41%	4
人体艺术	3048060	21047300	1.65%	5
小说	2886274	23933574	1.88%	6
迅雷	2718826	26652400	2.09%	7
开心网	2667120	29319520	2.30%	8
色情五月天	2563187	31882707	2.51%	9
海阔天空	2539768	34422475	2.71%	10
搜索	2342214	36764689	2.89%	11
盘龙	2341789	39106478	3.07%	12
优酷	2290672	41397150	3.25%	13
情色五月天	2210942	43608092	3.43%	14
校内网	2171512	45779604	3.60%	15
酷狗	2171319	47950923	3.77%	16
火影忍者	2033307	49984230	3.93%	17
妞妞基地	1893763	51877993	4.08%	18
小沈阳	1720109	53598102	4.21%	19
淘宝网	1665133	55263235	4.34%	20

累计覆盖率曲线如图 3.15。

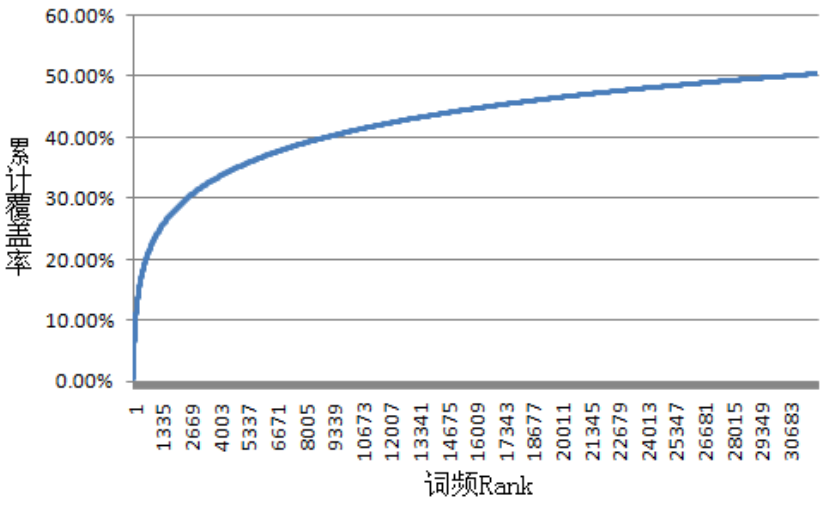

图3.15 查询日志纯汉字词条累计覆盖率曲线

前 32000 个词 Log(词频)关于 Log(Rank)曲线如图 3.16，线性相关系数 -0.9997，符合 Zipf 定律：

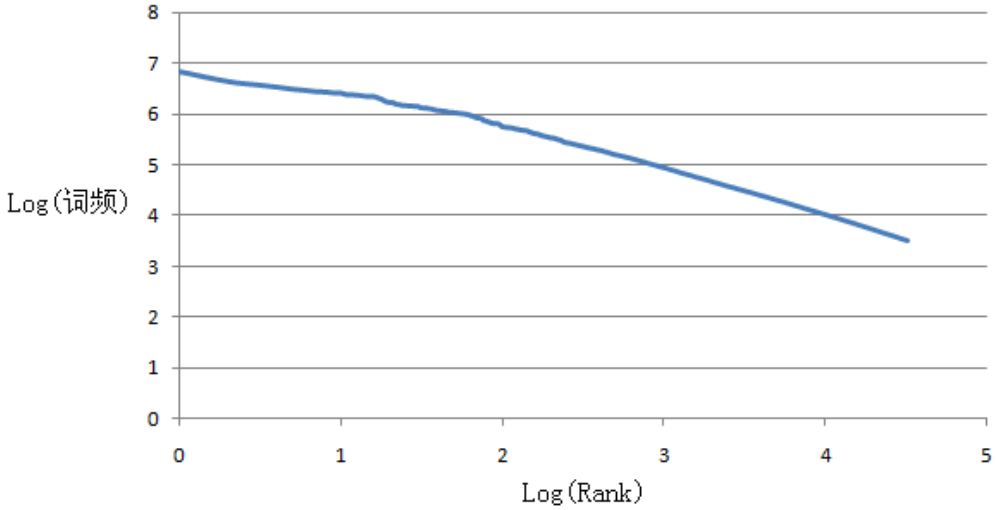

图3.16　查询日志纯汉字词条Log(词频)关于Log(Rank)的曲线

不同词长分布情况如表 3.11。

表3.11　查询日志中纯汉字词条在不同词长上的分布

词长	总个数	总词频	平均词频
2	1079211	243133760	225.2884
3	5000414	173820734	34.76127
4	11781972	336628872	28.57152
5	13270631	185674260	13.99137
6	20661170	189588074	9.176057
7	15519514	143415367	9.24097
全部	67312912	1272261067	18.9007

词条数随词长变化图如图 3.17，可见 6 字词之前递增，6 字词最多。与整体分布结果类似。

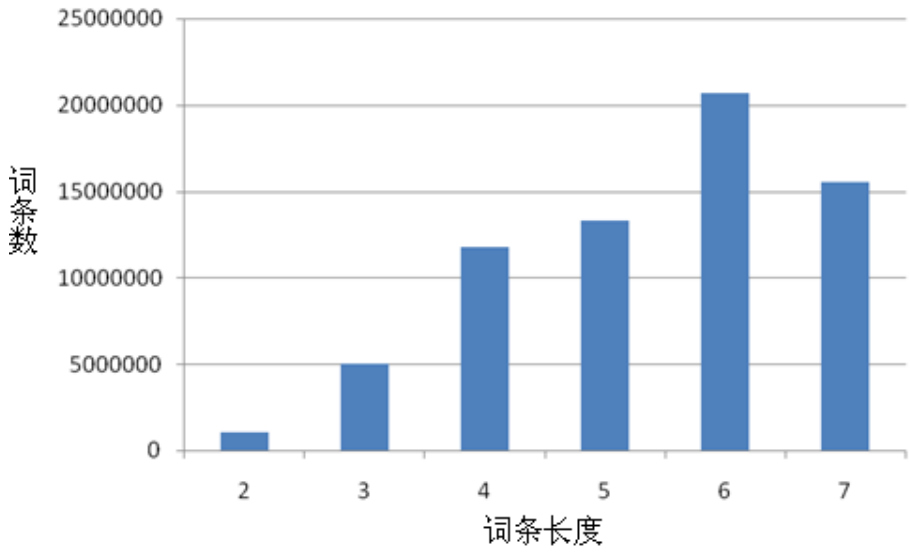

图3. 17　不同词长查询日志纯汉字词条词条数分布

总词频随词长变化图如图 3.18，整体来说比较平均，2、4 字词较高。相比整体分布，2 字词词频有所上升。

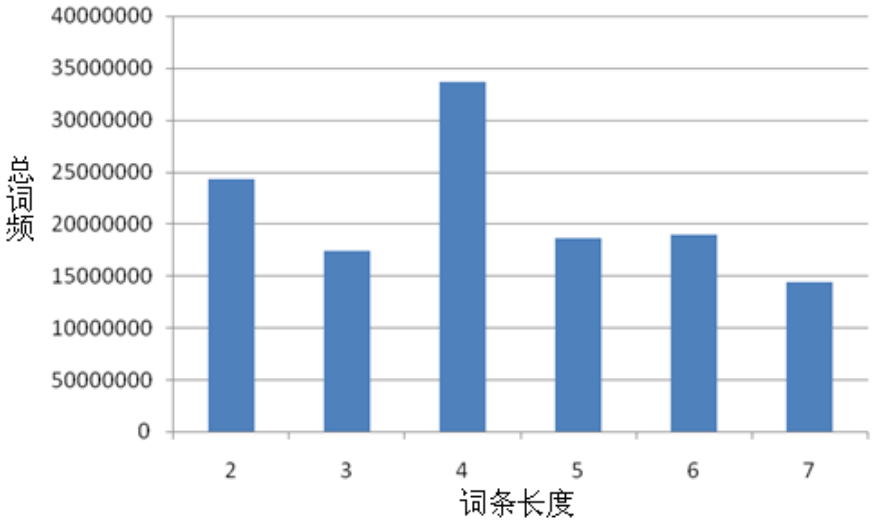

图3. 18　不同词长查询日志纯汉字词条词频分布

平均词频随词长变化图如图 3.19，2 字词远高于多字词。

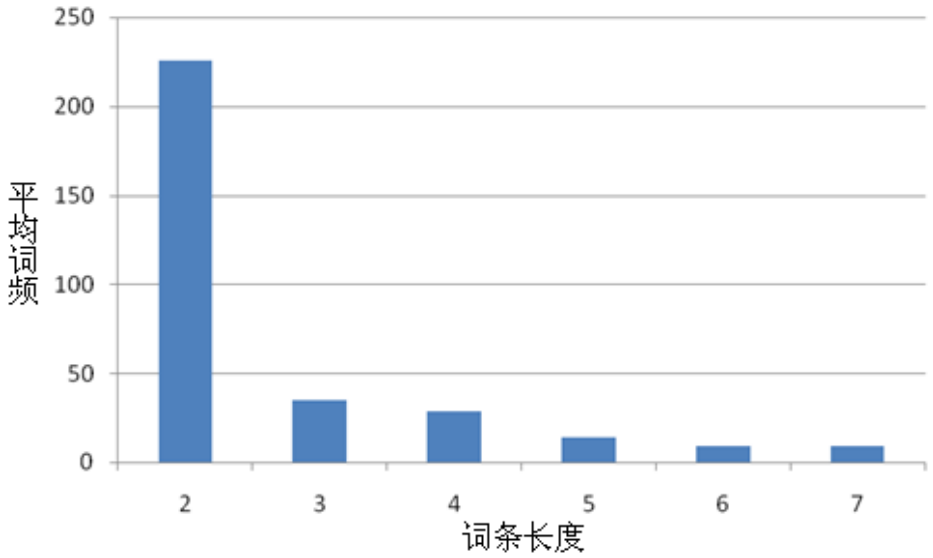

图3.19 不同词长查询日志纯汉字词条平均词频分布

3.4.1.3 纯非汉字词分布情况

经过计算整理，查询日志有纯非汉字词条 2635666 个(3.22%)，总词频 308942288 (17.49%)。排名前 20 的词如表 3.12。

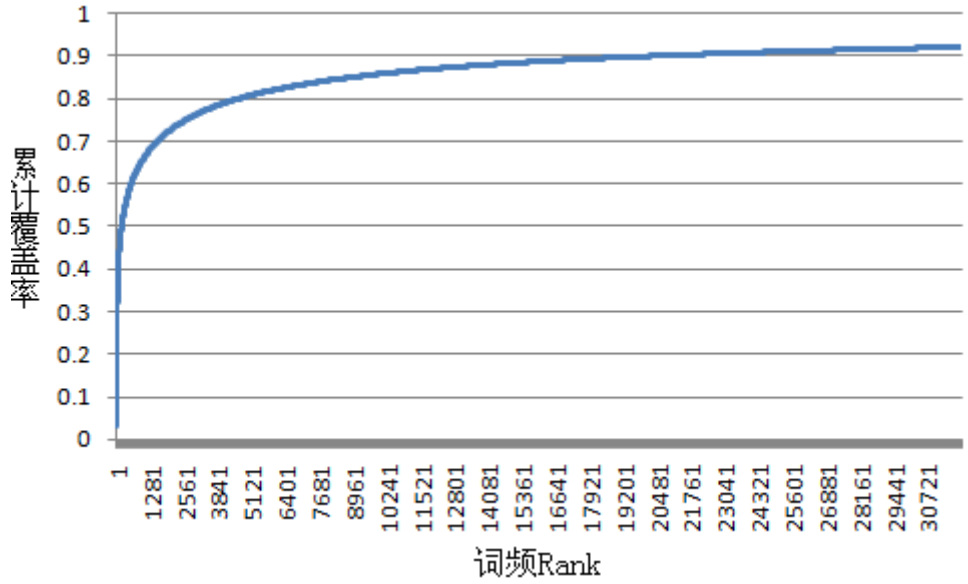

图3.20 查询日志纯非汉字词条累计覆盖率曲线

表3.12 查询日志中排名前20的纯非汉字词条

词	词频	累计词频	累计覆盖率	Rank
m p 3	10309694	10309694	0.033370938	1
b a i d u	8596338	18906032	0.061195999	2
h t t p	8370998	27277030	0.088291668	3
n b a	5772458	33049488	0.106976252	4
x i a o n e i	3592199	36641687	0.118603663	5
d n f	3552056	40193743	0.130101137	6
51	3293443	43487186	0.14076152	7
c o m	3144447	46631633	0.150939625	8
163	3078896	49710529	0.160905551	9
x i x i	2829046	52539575	0.17006275	10
s i t e	2727204	55266779	0.178890301	11
q q	2711787	57978566	0.18766795	12
y y	2516175	60494741	0.195812433	13
3 g p	2412581	62907322	0.203621597	14
d j	2271002	65178324	0.210972491	15
w w w	2226115	67404439	0.218178092	16
4399	2113998	69518437	0.225020788	17
y o u k u	2112167	71630604	0.231857557	18
v a g a a	1927935	73558539	0.238097994	19
h a o 1 2 3	1898048	75456587	0.244241692	20

累计覆盖率曲线如图 3.20。

前 32000 个词 Log(词频)关于 Log(Rank)曲线如图 3.21，线性相关系数 -0.9991，符合 Zipf 定律：

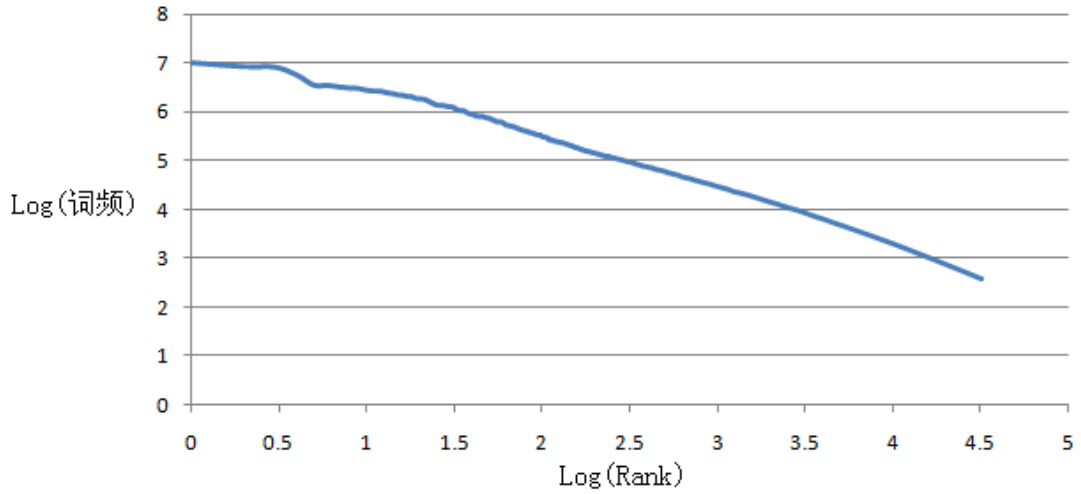

图3.21 查询日志纯非汉字词条Log(词频)关于Log(Rank)的曲线

不同词长分布情况如表 3.13。

表3.13 查询日志中纯非汉字词条在不同词长上的分布

词长	总个数	总词频	平均词频
2	10669	40699000	3814.697
3	73083	79006278	1081.049
4	371010	64376205	173.5161
5	680066	52930138	77.83088
6	812721	46815836	57.60382
7	688117	25114831	36.49791
全部	2635666	308942288	117.216

词条数随词长变化图如图 3.22，可见 6 字词之前递增，6 字词最多。

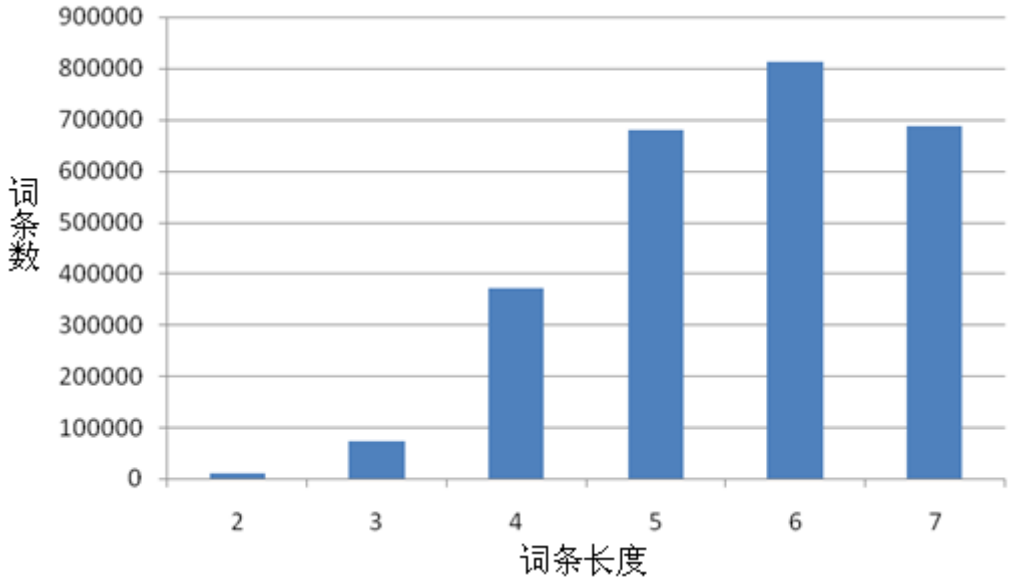

图3.22 不同词长查询日志纯非汉字词条词条数分布

总词频随词长变化图如图 3.23， 2 字词较低，3 字词最高之后递减。这可能因为英文缩写等因素，双字词的区分度较差，3 字词能更好地表达查询目的。

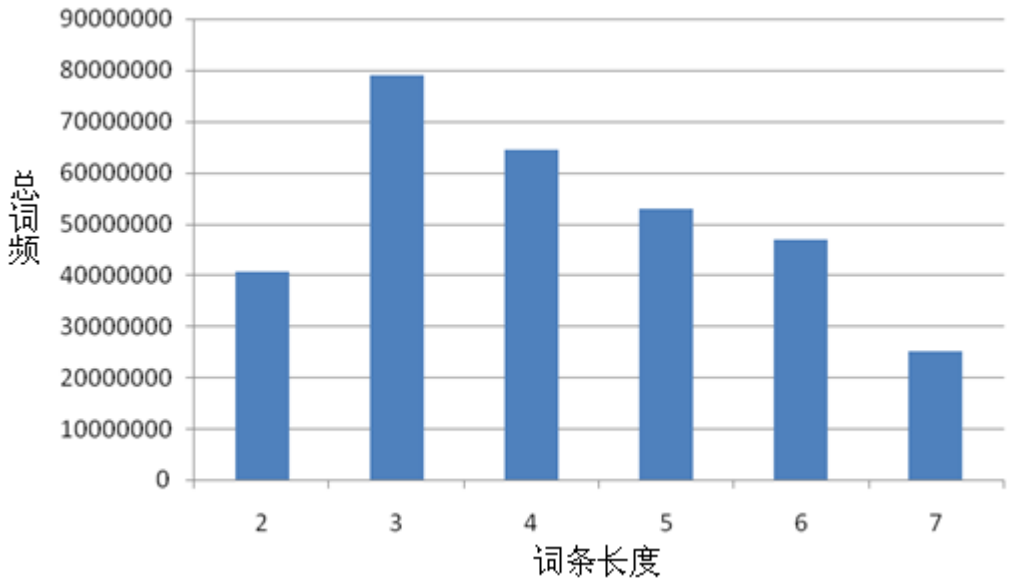

图3.23 不同词长查询日志纯非汉字词条词频分布

平均词频随词长变化图如图 3.24，2 字词远高于多字词。

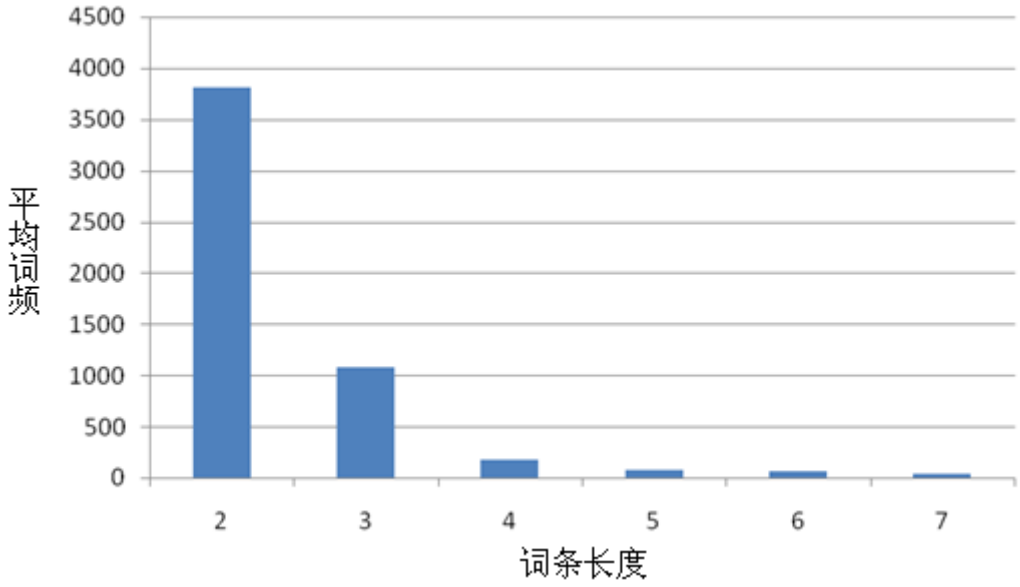

图3.24 不同词长查询日志纯非汉字词条平均词频分布

3.4.1.4 混杂词分布情况

经过计算整理，查询日志有词条 12022051 个(14.67%)，总词频 184910402 (10.47%)。排名前 20 的词如表 3.14。

表3.14 查询日志中排名前20的混杂词条

词	词频	累计词频	累计覆盖率	Rank
q q 空间	1845848	1845848	1.00%	1
q q 头像	1457836	3303684	1.79%	2
q q 个性签名	1443931	4747615	2.57%	3
q q 网名	1405206	6152821	3.33%	4
迅雷 5	1188490	7341311	3.97%	5
q q 空间克隆	931773	8273084	4.47%	6
3 g p 电影下载	837652	9110736	4.93%	7
d n f 官网	702950	9813686	5.31%	8
d n f 外挂	653123	10466809	5.66%	9
a 片	627655	11094464	6.00%	10
q q 表情	608989	11703453	6.33%	11
q 吧	583325	12286778	6.64%	12
迅雷 5 下载	583211	12869989	6.96%	13
q q 下载	527978	13397967	7.25%	14
3 g p 转换器	526247	13924214	7.53%	15
4 3 9 9 小游戏	515634	14439848	7.81%	16
y y 下载	501030	14940878	8.08%	17
n b a 直播	477798	15418676	8.34%	18
3 6 0 安全卫士	474792	15893468	8.60%	19
q q 空间代码	473123	16366591	8.85%	20

累计覆盖率曲线如图 3.25。

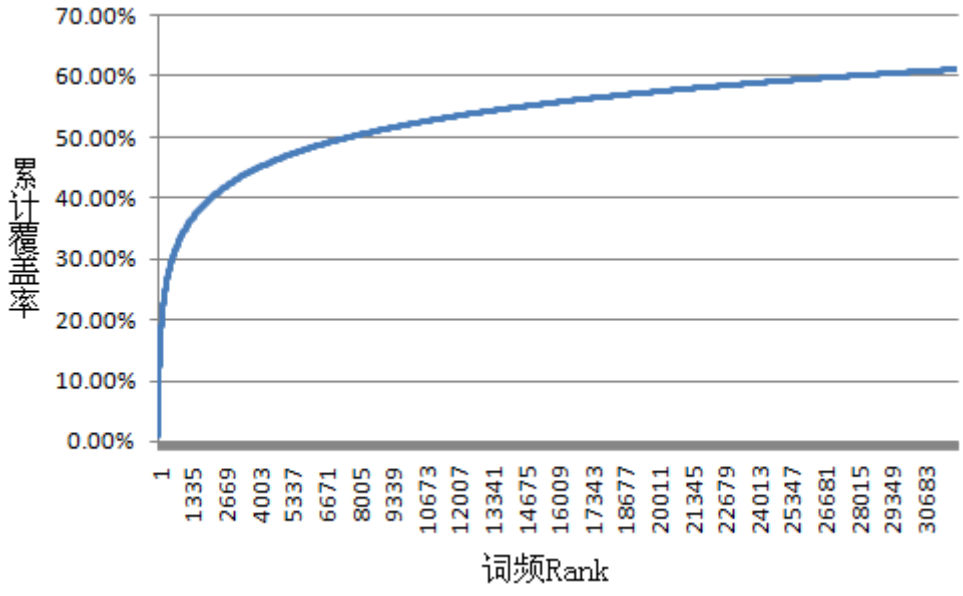

图3.25 查询日志混杂词条累计覆盖率曲线

前 32000 个词 Log(词频)关于 Log(Rank)曲线如图 3.26，线性相关系数 -0.9998：

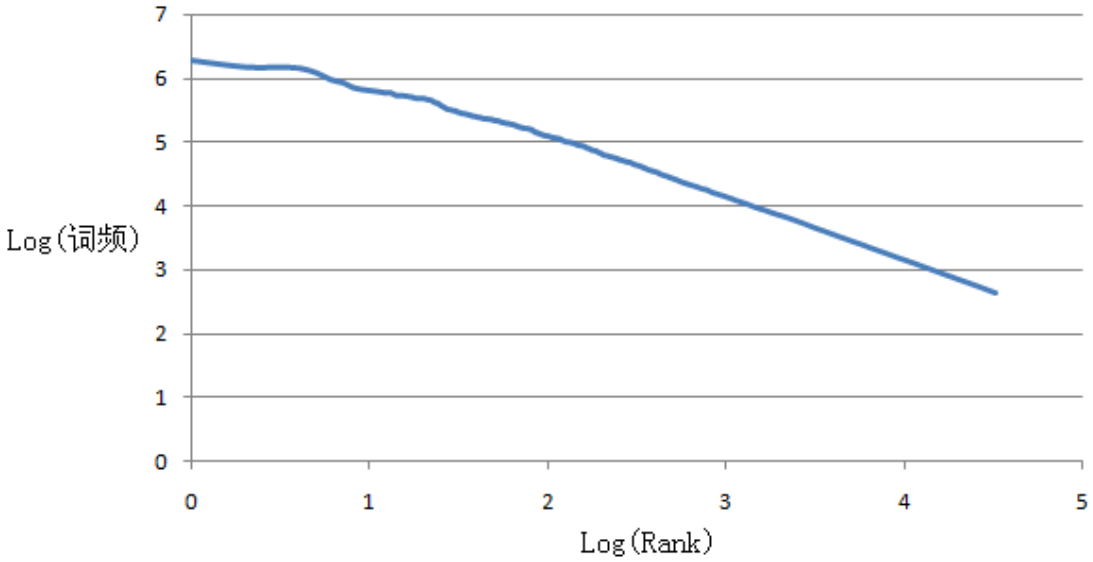

图3.26 查询日志混杂词条Log(词频)关于Log(Rank)的曲线

不同词长分布情况如表 3.15：

表3.15 查询日志中混杂词条在不同词长上的分布

词长	总个数	总词频	平均词频
2	55412	4093539	73.87459
3	410920	11017285	26.81126
4	1132792	33967075	29.98527
5	2394505	38055994	15.89305
6	3766070	47613256	12.64269
7	4262352	50163253	11.76891
全部	12022051	184910402	15.38094

词条数随词长变化图如图 3.27，可见递增，7 字词最多。

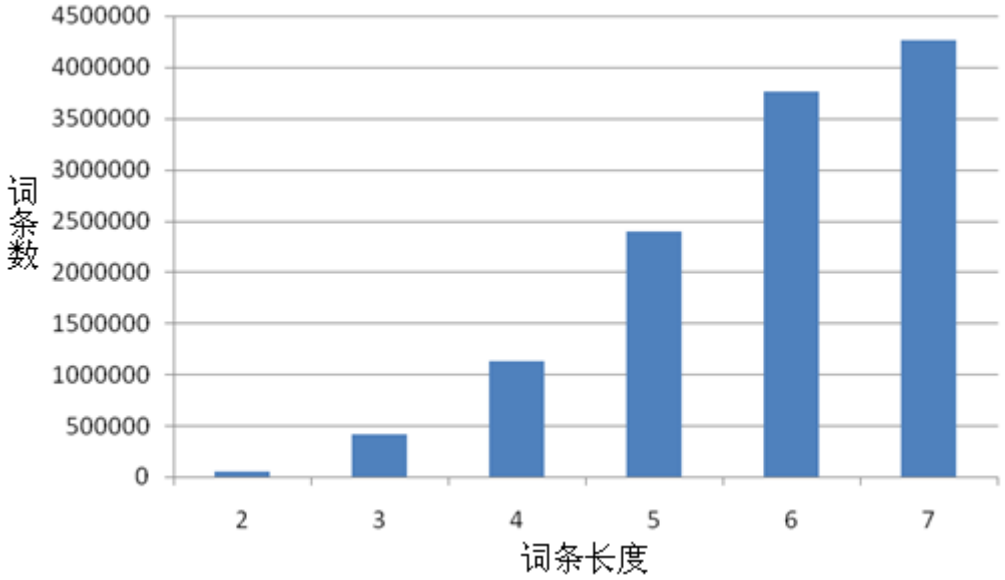

图3.27 不同词长查询日志混杂词条词条数分布

总词频随词长变化图如图 3.28， 随词长递增。对混杂词条来说，可能长词条更能表达查询目的。

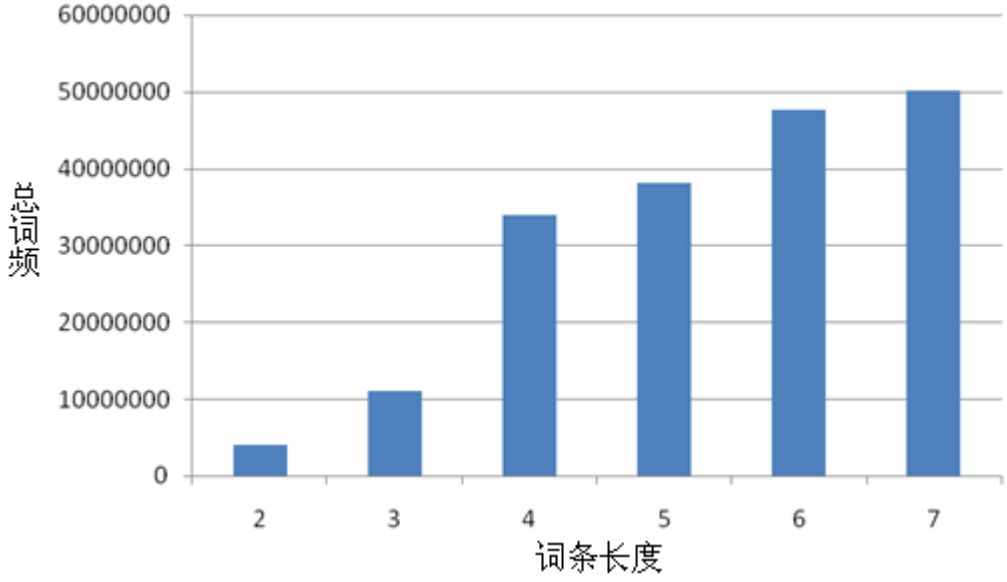

图3.28 不同词长查询日志混杂词条词频分布

平均词频随词长变化图如图 3.29，2 字词较高。

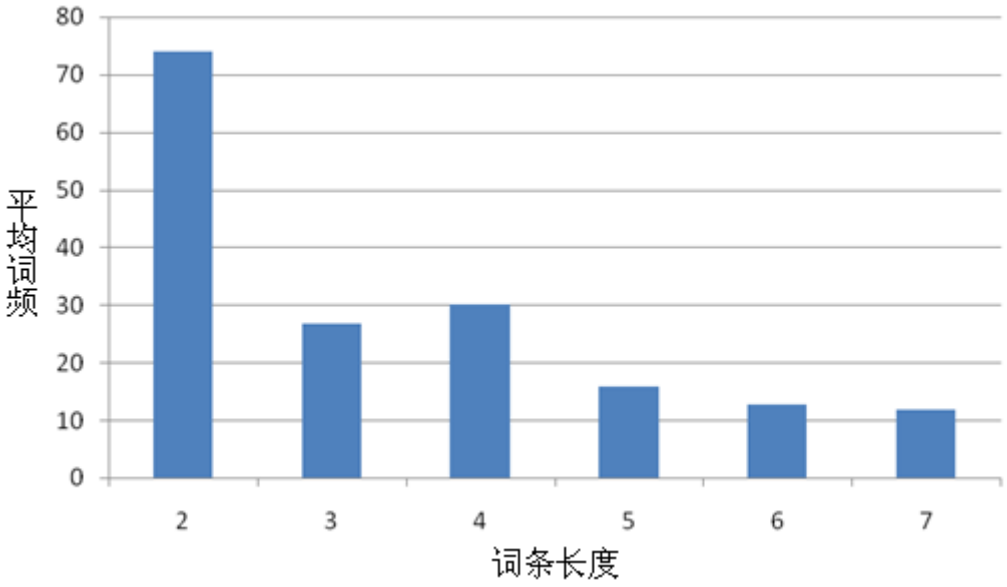

图3.29　不同词长查询日志混杂词条平均词频分布

3.4.2　三千常用词分布情况

三千常用词在查询日志中的分布情况如表 3.16：

表3.16　三千常用词在查询日志中的分布情况

词长	三千常用词	交集词条数	词频	平均词频	非交集词条数
1	1009	0	0	0	1009
2	2571	2565	30397746	11850.97	6
3	204	201	959653	4774.393	3
4	26	26	17513	673.5769	0
5	3	3	3529	1176.333	0
6	0	0	0	0	0
7	2	2	361	180.5	0
全部	3815	2797	31378802	11218.74	1018

由于查询日志没有记录单字，所以全部单字都没有出现。

没有出现的二字词有：赛球、惯于、哪话、成万、少陪、枝对。

没有出现的三字词有：留声片、人家儿、那么着。

交集部分的词条数随词长的分布图如图 3.30，基本和三千常用词除单字外的情况相同。

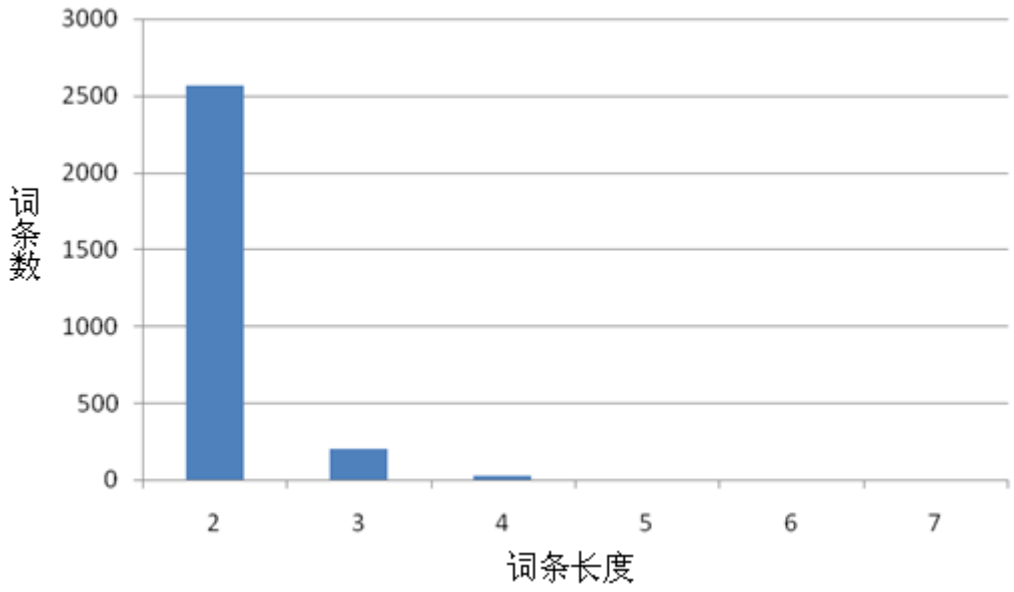

图3.30　不同词长三千常用词在查询日志中的词条数分布

交集部分词频随词长变化图如图 3.31，2 字词远远高于其他，这与多字词太少有关：

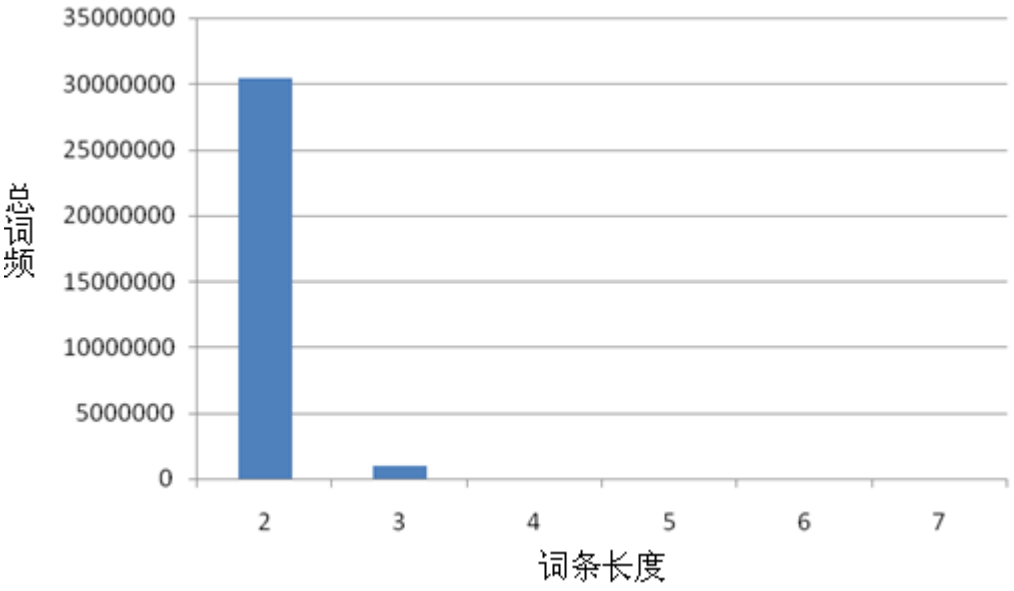

图3.31　不同词长三千常用词在查询日志中的词频分布

交集部分平均词频随词长变化图如图 3.32，2 字词较高：

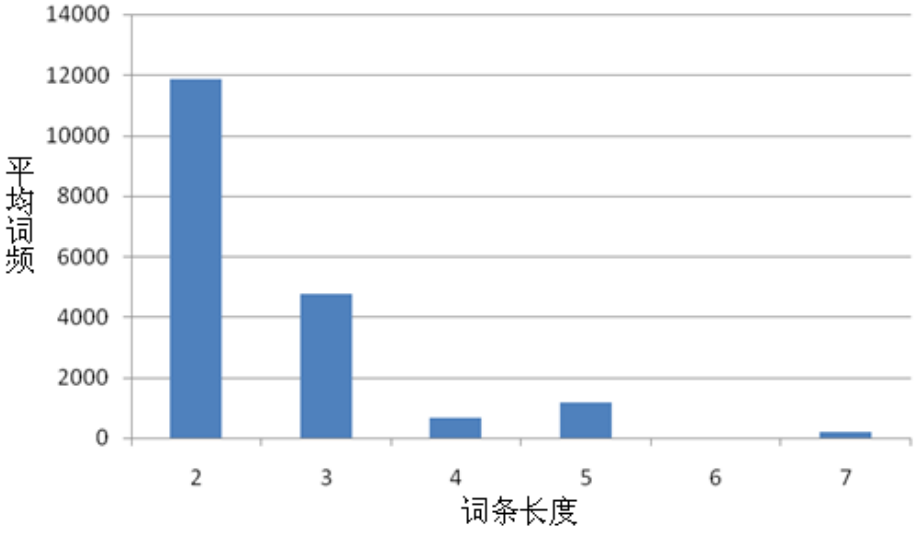

图3.32　不同词长三千常用词在查询日志中的平均词频分布

查询日志中词频前20的三千常用词如表3.17,可见多是有具体意义的实词：

表3.17　查询日志中词频前20的三千常用词

词	词频	Rank
电影	3339722	1
小说	2886274	2
地图	1431557	3
土豆	1104685	4
音乐	695445	5
歌曲	658019	6
鸭蛋	545180	7
翻译	538737	8
游戏	474880	9
五月	382888	10
军事	335856	11
苹果	275286	12
笑话	263551	13
同志	255167	14
萝卜	235271	15
汽车	215356	16
电话	195704	17
英语	189495	18
火车	183029	19
新闻	181657	20

词频在 10 以下的三千常用词如表 3.18。

表3.18　查询日志中词频小于10的三千常用词

词	词频	词	词频
暂且	10	教给	6
灶火	10	水管子	6
凑巧	10	划子	5
为着	10	受累	5
动身	10	筐子	5
原先	10	眼前的	5
总归	10	免不了	5
板擦	10	不好过	5
好好地	10	老早	4
归于	9	便帽	4
免得	9	撒开	4
来着	9	通讯处	4
医务所	9	农业社	4
秆子	8	有空来玩	4
水瓢	8	这会	3
退掉	8	除开	3
跟前	8	这么些	3
电车站	8	树林子	3
向来	7	沓子	2
爽直	7	那会	2
后年	7	那么些	2
不可不	7	手腕子	2
对不住	7	快快地	2
匙子	6	灵便	1
想出	6	这么着	1
纸烟	6	大前天	1
外头	6	铁甲车	1

这些词中，除了时代发展等原因使用较少的一些词如农业社、铁甲车等，还包括一些虚词、口语词汇等实际意义较少或指代不明确的词，比如这么着、大前天等。这也是受查询词用途的影响。

3.5 小结

用户词库虽然包含大量词条，但是使用非常集中，不到 1%的词条即可覆盖绝大多数词频。因为输入法用于文字交流较多，和语言常用词有较高的相似性。三千常用词可以覆盖用户词库大部分词频。

查询日志的词条由于使用用途的特点，和平常语言中使用词条情况差别较大，表示查询目的特征的实词频度较高，平常语言中常用的虚词等频度较低。

第4章 基于输入法用户词库和查询日志的 wiki 中文常用词条

4.1 实验概述

本实验的目的在于结合不同语料数据，找到一种衡量词条常用程度的方法，进而得到现在网络环境中的常用词表。Wiki（维基）是用户产生的百科知识库，用户编辑后的词条也要经过审核和公众监督。相对输入法用户词库和查询日志，Wiki 词条更为严谨，一些异体字、“火星文”、色情词汇的频度受到遏制。因此以 Wiki 词条为基础，结合输入法用户词库和查询日志等其他语料，可以尝试得到中文常用词条。

4.2 数据介绍

4.2.1 输入法用户词库和查询日志

本实验使用的输入法用户词库数据和查询日志数据和第三章使用的相同。用户词库数据是 2010 年 3 月 15 日的词库数据，记录了词条长度不超过 7 的词条，共 111659347 个词条，总词频 327029776076，平均词频 2928.817。查询日志数据是 2009 年全年数据，记录了 2-7 字词，对英文、数字等非汉字字符进行了全角化处理。查询日志有词条 81970629 个，总词频 1766113757。因为本实验关注中文词条，因此只选择了查询日志中的纯汉字词，共 67312912 个，总词频 1272261067。

4.2.2 Wiki 链接词数据

本实验使用的是中文维基百科页面所有中文链接词的统计词表。链接词即是 Wiki 内容页面中以超链接形式存在的词条，可以点击进入该词条的 Wiki 页面。但是并非所有链接词都有对应的 Wiki 页面，因此链接词数量比 Wiki 词条要多。Wiki 中文链接词共有 1314388 个，总词频 314920483。

4.2.3 Sogout 网页串频数据

Sogout 互联网语料库收录了互联网各种类型的 1.3 亿个网页数据，有 2000 亿汉字，没有分词，所以无法得到总词条数、总词频等信息。本实验使用的是 Wiki 词条在 Sogout 上的串频数据，共出现词条 947807 个，总词频 245374207469。

4.3 Wiki 中文链接词条在不同数据集下的分布情况

4.3.1 Wiki 中文链接词条在输入法用户词库的分布

Wiki 中文词条在用户词库中出现了 530956 个，总词频 189059763501，覆盖率 57.81%。词频前 20 的词条如表 4.1。

表4.1 Wiki中文链接词条在用户词库中词频前20的词条

词条	用户词库词频	wiki 词频
啊	3990088189	6940
就	2940184432	215167
在	2564723683	1511811
好	2474029080	127232
的	2414096285	5416612
我	2382157145	225577
有	2361758812	1227785
没	2225026282	161022
去	2075258235	112076
要	1974892907	422094
都	1964884870	299195
那	1934999163	116946
什么	1773690404	29848
恩	1762991445	78978
说	1737688923	227967
还	1635249977	133837
个	1621634553	736854
看	1533318146	104526
现在	1489592537	57578
哦	1345621758	836

不同词条数累计覆盖用户词库的覆盖率情况如表 4.2。

表4.2 在用户词库中Wiki中文链接词条不同词条数的覆盖率

词条数	累计覆盖率
10000	54.89%
20000	56.34%
30000	56.89%
40000	57.19%
50000	57.36%
100000	57.68%

可见前 100000 个词条已经覆盖了绝大多数词频。累计覆盖率曲线如图 4.1，经过计算，Wiki 中文链接词在用户词库的分布符合 Zipf 定律。

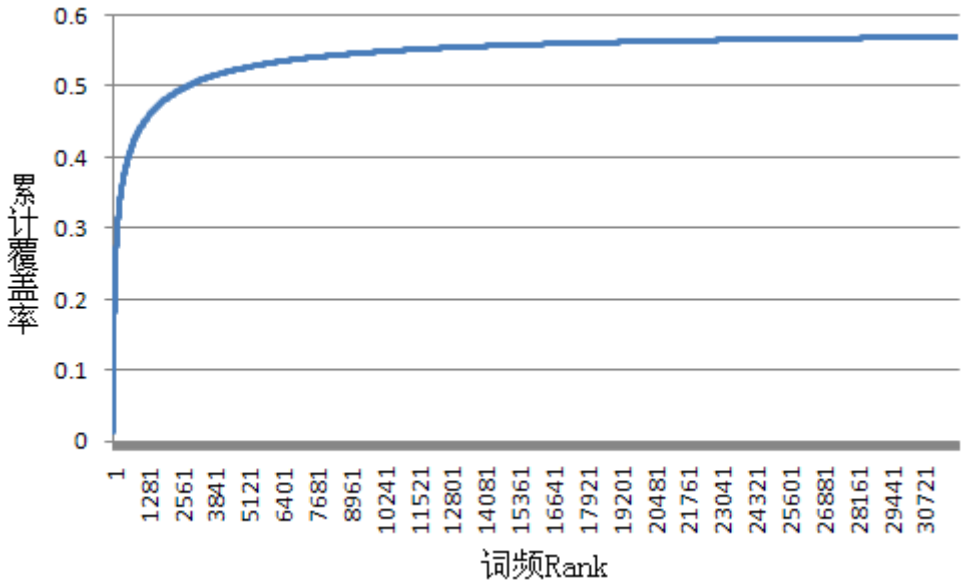

图4.1 在用户词库中Wiki中文链接词条的累计覆盖率曲线

4.3.2 Wiki 中文链接词条在查询日志的分布

Wiki 中文链接词在查询日志中出现了 494348 个，总词频 412799061，覆盖率 32.45%。词频前 20 的词条如表 4.3。

不同词条数覆盖查询日志词频的情况如表 4.4。

前 100000 个词已经覆盖了绝大多数词频。累计覆盖率曲线如图 4.2，经过计算，Wiki 中文链接词在查询日志的分布符合 Zipf 定律。

表4.3 Wiki中文链接词条在查询日志中词频前20的词条

词条	querylog 词频	wiki 词频
百度	6678435	3438
视频	4292585	7903
下载	3688498	5058
电影	3339722	121198
人体艺术	3048060	28
小说	2886274	40861
迅雷	2718826	402
开心网	2667120	50
海阔天空	2539768	118
搜索	2342214	7261
盘龙	2341789	555
优酷	2290672	76
校内网	2171512	72
酷狗	2171319	40
火影忍者	2033307	1832
小沈阳	1720109	33
淘宝网	1665133	74
土豆网	1515161	100
五月天	1452890	604
地图	1431557	15982

表4.4 在查询日志中Wiki中文链接词条不同词条数的覆盖率

词条数	累计覆盖率
10000	26.32%
20000	28.59%
30000	29.68%
40000	30.33%
50000	30.77%
100000	31.76%

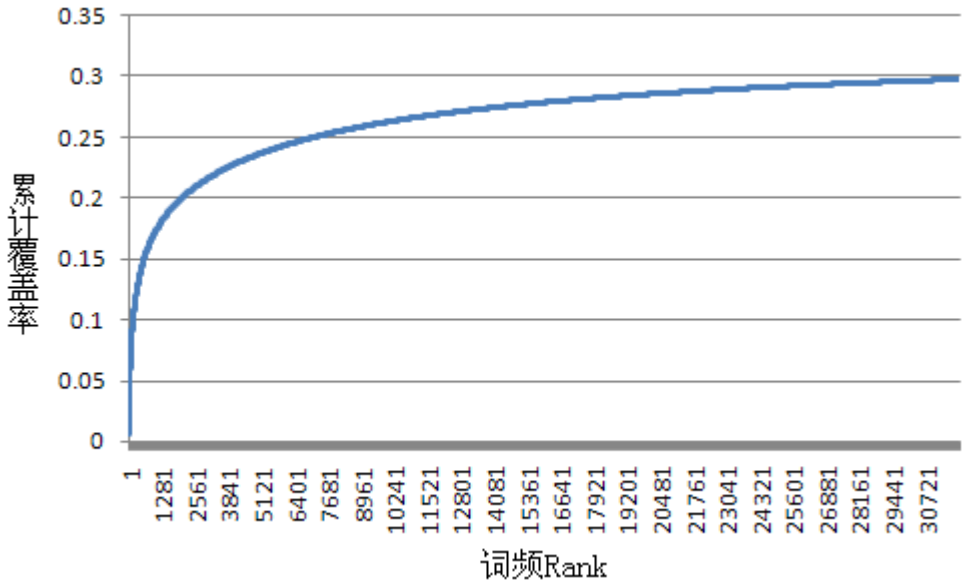

图4.2 在查询日志中Wiki中文链接词条的累计覆盖率曲线

4.3.3 Wiki 中文链接词条在 Sogout 串频数据的分布

Wiki 中文链接词在 Sogout 中出现了 947807 个，总词频 245374207469。词频前 20 的词条如表 4.5。

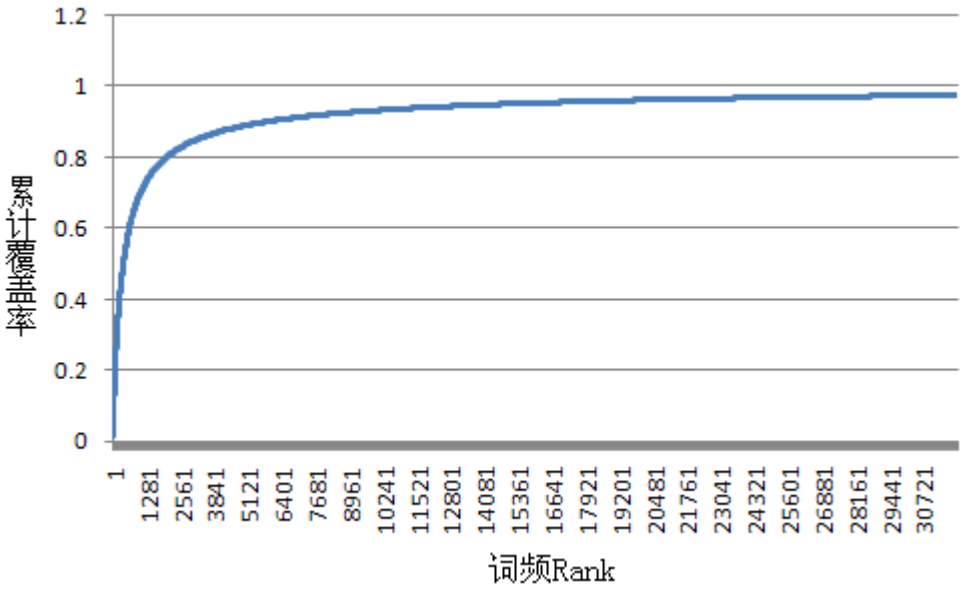

图4.3 在Sogout数据中Wiki中文链接词条的累计覆盖率曲线

表4.5 Wiki中文链接词条在Sogout数据中词频前20的词条

词条	Sogout 词频
的	3401650474
网	1313271905
一	1303128036
人	1284674315
中	1198054990
大	1022941757
新	979347248
有	975193257
我	960649164
不	935123534
国	909503868
在	879083071
上	826572077
是	822630228
发	791464908
用	678057514
时	669167484
电	659621330
文	655706509
个	652011076

由于 Sogout 数据没有整体词频信息，因此只能计算和 Wiki 共有部分的覆盖率，不同词条数覆盖共有部分词频情况如表 4.6。

表4.6 在Sogout串频数据中Wiki中文链接词条不同词条数的覆盖率

词条数	累计覆盖率
10000	93.53%
20000	96.45%
30000	97.66%
40000	98.32%
50000	98.74%
100000	99.58%

和前两个数据集相同，前 100000 个词已经覆盖了绝大多数词频。累计覆盖率曲线如图 4.3，经过计算，Wiki 中文链接词在 Sogout 数据中的串频符合 Zipf 定律。

4.3.4 小结

Wiki 中文链接词在三个数据中的分布都符合 Zipf 定律，而且前 10 万个词都覆盖了绝大多数词频，因此考察常用词时可以以这些词为基础。

从共有词条数量和总覆盖率情况来看，Sogout 数据和 Wiki 链接词相关性最好，用户词库其次，查询日志较差。

4.4 基于不同数据集的 Wiki 常用词条

实验的基本思路是利用用户词库、查询日志和 Sogout 数据的常用词条，综合 Wiki 数据的信息，得到 Wiki 常用词条。通过之前的实验可以知道，在用户词库、查询日志和 Sogout 数据中出现的 Wiki 词条，前 100000 个即可覆盖绝大部分词频。因此对 3 个数据集中出现的 Wiki 词条，只取前 100000 个高频词条，过滤掉大部分低频词。将三组高频词合并，首先得到一个高频词条集合。合并的结果，高频词条集合包含 154651 个词条。这些词条包含 4 个维度的概率信息：用户词库概率、查询日志概率、Sogout 串频概率、Wiki 链接词概率。用不同方法计算综合概率或平均概率，可以得到一个综合各个数据集的常用词概率的度量。

使用算术平均值计算平均概率，得到的前 50 个常用词如表 4.7。

表4.7 使用算术平均值计算综合概率的前50个常用词

词条	用户词库概率	查询日志概率	Sogout 概率	Wiki 概率	算术平均值
的	0.007382	0	0.013863	0.0172	0.009611
在	0.007842	0	0.003583	0.004801	0.004056
有	0.007222	0	0.003974	0.003899	0.003774
人	0.003492	0	0.005236	0.005699	0.003607
年	0.000783	0	0.002412	0.011186	0.003595
一	0.001424	0	0.005311	0.005762	0.003124
啊	0.012201	0	0.000229	2.20E-05	0.003113
我	0.007284	0	0.003915	0.000716	0.002979
中	0.000659	0	0.004883	0.006124	0.002916
就	0.008991	0	0.001014	0.000683	0.002672
国	6.71E-05	0	0.003707	0.006668	0.00261
个	0.004959	0	0.002657	0.00234	0.002489
大	0.000919	0	0.004169	0.004755	0.002461
好	0.007565	0	0.001694	0.000404	0.002416

续表4.7 使用算术平均值计算综合概率的前50个常用词

词条	用户词库概率	查询日志概率	Sogout 概率	Wiki 概率	算术平均值
是	0.001328	0	0.003353	0.004907	0.002397
要	0.006039	0	0.001685	0.00134	0.002266
日	0.000829	0	0.002335	0.005529	0.002173
上	0.002286	0	0.003369	0.002387	0.002011
没	0.006804	0	0.000708	0.000511	0.002006
都	0.006008	0	0.000853	0.00095	0.001953
和	0.003191	0	0.001646	0.00281	0.001912
会	0.002379	0	0.002359	0.002868	0.001901
说	0.005314	0	0.001377	0.000724	0.001854
月	0.000917	0	0.001598	0.004841	0.001839
去	0.006346	0	0.000584	0.000356	0.001821
为	0.000666	0	0.002117	0.004406	0.001797
看	0.004689	0	0.002111	0.000332	0.001783
到	0.003106	0	0.00204	0.001976	0.00178
来	0.003645	0	0.001806	0.001591	0.001761
不	0.000426	0	0.003811	0.002682	0.00173
那	0.005917	0	0.000565	0.000371	0.001713
他	0.003662	0	0.00136	0.001785	0.001701
用	0.001772	0	0.002763	0.002152	0.001672
下	0.002493	0	0.002555	0.001426	0.001619
新	0.000345	0	0.003991	0.002064	0.0016
网	0.000217	0	0.005352	0.000751	0.00158
还	0.005	0	0.000758	0.000425	0.001546
发	0.001243	0	0.003226	0.00161	0.00152
什么	0.005424	2.35E-05	0.000436	9.48E-05	0.001494
学	0.000395	0	0.001951	0.003595	0.001485
行	0.000977	0	0.002432	0.002325	0.001434
恩	0.005391	0	8.53E-05	0.000251	0.001432
点	0.00266	0	0.002388	0.000639	0.001422
家	0.00106	0	0.002508	0.002005	0.001393
时	0.00034	0	0.002727	0.00246	0.001382
多	0.002269	0	0.001715	0.001464	0.001362
百度	9.53E-05	0.005249265	9.22E-05	1.09E-05	0.001362
地	0.000306	0	0.002442	0.002629	0.001344
文	7.67E-05	0	0.002672	0.002486	0.001309
小	0.000956	0	0.00246	0.001799	0.001304

观察可以发现，用算术平均度量的前 50 个常用词中，查询日志的影响较小，大部分概率都为 0。而且常用词以单字词为主，这与用户词库、Sogout 和 Wiki 的常用词是一致的。

使用调和平均计算平均概率，前 50 个常用词如表 4.8。

表4.8 使用调和平均值计算综合概率的前50个常用词

词条	用户词库概率	查询日志概率	Sogout 概率	Wiki 概率	调和平均值
中国	0.000432	0.000632465	0.001451	0.001519	0.000763
公司	0.001347	0.000421986	0.001233	0.000582	0.000709
游戏	0.001076	0.001150368	0.000759	0.00025	0.000562
北京	0.000575	0.001990898	0.00079	0.000236	0.000517
电影	0.000367	0.008090271	0.000432	0.000385	0.000515
时间	0.001678	0.00028751	0.000912	0.000323	0.000484
上海	0.000461	0.001276267	0.000531	2.26E-04	0.000432
网站	0.00039	0.0003463	0.000812	0.000227	0.000361
图片	0.000369	0.003130568	0.000803	0.00014	0.00035
音乐	0.000162	0.001684673	0.000486	0.000299	0.000328
大学	2.93E-04	0.000244044	0.000302	0.00068	0.000325
电视	0.000311	0.000158951	0.000315	0.000543	0.000275
资料	0.00032	0.000161075	0.00047	0.000303	0.000271
空间	0.000538	0.00043045	0.000404	0.000112	0.000256
日本	0.000133	0.000609357	0.000198	0.000726	0.000256
工作	0.001489	0.0001072	0.000392	0.000207	0.00023
新闻	0.000116	0.000440053	0.000896	0.000176	0.000226
美国	0.000128	0.000251713	0.000217	0.000682	0.000224
问题	0.00157	9.85E-05	0.000343	0.00022	0.000219
有限公司	0.000141	0.000784986	0.000432	0.000128	0.000216
留言	0.000155	0.000187867	0.000339	0.000253	0.000214
小说	0.000114	0.006991821	0.000295	0.00013	0.0002
广州	0.000211	0.000472317	0.000261	0.000104	0.000197
电脑	0.00088	0.00023995	0.000254	8.78E-05	0.000194
管理	0.000416	7.58E-05	0.000692	0.000253	0.00019
设计	0.000292	0.000102287	0.000337	0.00019	0.000187
移动	0.000183	0.000295567	0.000105	0.000314	0.000186
香港	9.24E-05	0.000288878	0.000162	0.000915	0.000186
英语	0.000224	0.00045904	0.000198	0.000101	0.000185
软件	0.00035	0.000408002	0.000523	6.65E-05	0.00018
学校	0.000686	0.000115264	0.000112	0.000282	0.000177

续表4.8 使用调和平均值计算综合概率的前50个常用词

词条	用户词库概率	查询日志概率	Sogout 概率	Wiki 概率	调和平均值
使用	0.000275	6.73E-05	0.00042	0.000545	0.000176
中文	9.48E-05	0.000310019	0.000181	0.000239	0.00017
学生	0.000249	9.92E-05	0.000207	0.000163	0.00016
生活	0.000386	7.44E-05	0.000411	0.00014	0.000156
系统	0.000484	5.89E-05	0.000316	0.000268	0.000154
电话	0.001478	0.000474081	0.000407	4.74E-05	0.000152
银行	0.000252	0.000106369	0.000248	0.000112	0.000152
什么	0.009382	7.24E-05	0.000436	9.48E-05	0.000149
国家	0.000165	7.03E-05	0.000201	0.000597	0.000148
企业	0.000206	0.000108072	0.00053	8.77E-05	0.000146
我们	0.004359	5.56E-05	8.14E-04	0.000125	0.000146
教育	0.000125	8.15E-05	0.000341	0.00023	0.000145
汽车	0.000107	0.000521687	0.000445	6.80E-05	0.000142
南京	0.000142	0.000962205	0.000126	7.96E-05	0.00014
内容	0.000225	5.09E-05	0.000631	0.000336	0.00014
主题	7.51E-05	0.000236488	3.07E-04	1.26E-04	0.000139
广告	0.000201	0.000117924	0.000997	6.92E-05	0.000138
介绍	3.12E-04	0.000155419	0.000201	6.85E-05	0.000137
天津	0.000118	0.000457381	0.000158	8.03E-05	0.000136

在计算调和平均值之前，对各个数据的概率值进行了 Good-Turing 平滑处理。由于调和平均值高的条件之一是各个维度上都不能太低，因此查询日志没有收录的单字词的调和平均值普遍较低。算术平均值排名第一的“的”字在调和平均值中只能排到 5 万多。通过观察可以发现调和平均值排序的常用词中有实际意义或带有较强查询目的性的词较多，这与查询日志的特点比较一致。

用不同数据集的词条数规模进行加权，计算调和平均值，排名前 50 的常用词如表 4.9。

在计算之前同样进行了 Good-Turing 平滑处理。观察结果，与调和平均值的结果类似，排名较高的多是有实际意义的双字词。单字的排名较低，也是在 5 万左右才会出现单字。

表4.9 使用加权调和平均值计算综合概率的前50个常用词

词条	用户词库概率	查询日志概率	Sogout 概率	Wiki 概率	加权调和平均值
游戏	0.001076	0.001150368	0.000759	0.00025	0.001074
北京	0.000575	0.001990898	0.00079	0.000236	0.000774
公司	0.001347	0.000421986	0.001233	0.000582	0.000736
手机	0.000883	0.001565307	0.000697	1.73E-05	0.000734
电话	0.001478	0.000474081	0.000407	4.74E-05	0.000729
视频	0.000518	0.010398524	0.001298	2.51E-05	0.00066
下载	0.000569	0.00893516	0.000801	1.61E-05	0.000632
上海	0.000461	0.001276267	0.000531	0.000226	0.0006
时间	0.001678	0.00028751	0.000912	0.000323	0.00059
电影	0.000367	0.008090271	0.000432	0.000385	0.000572
图片	3.69E-04	0.003130568	0.000803	0.00014	0.000543
价格	0.00054	0.000887734	0.000354	2.05E-05	0.000519
中国	0.000432	0.000632465	0.001451	0.001519	0.000495
美女	0.000521	0.002569187	0.000431	1.05E-05	0.000492
空间	0.000538	0.00043045	0.000404	0.000112	0.000479
名字	0.000551	0.000375543	4.03E-05	0.000112	0.000434
照片	0.000651	0.000313808	0.00013	5.36E-05	0.000433
电脑	0.00088	0.00023995	0.000254	8.78E-05	0.000424
女人	0.000474	3.80E-04	0.000259	2.15E-05	0.000379
网站	0.00039	0.0003463	0.000812	0.000227	0.000371
软件	0.00035	0.000408002	0.000523	6.65E-05	0.000359
妈妈	0.00045	0.00035106	8.42E-05	1.27E-05	0.000326
开心	0.000419	0.000532997	4.44E-05	7.43E-06	0.000306
地址	0.000293	0.000379364	0.00034	2.34E-05	0.000294
朋友	0.001225	1.33E-04	0.000227	5.06E-05	0.000288
英语	0.000224	0.00045904	0.000198	0.000101	0.000274
大学	0.000293	0.000244044	0.000302	0.00068	0.000273
意思	1.23E-03	0.000131185	3.30E-05	4.86E-05	0.000273
论坛	0.000203	0.000790245	0.000755	2.62E-05	0.000264
广州	0.000211	0.000472317	0.000261	1.04E-04	0.000264
免费	0.000207	0.00064924	0.00045	2.82E-05	0.000263
男人	0.000491	1.74E-04	0.000163	1.81E-05	0.000261
妹妹	3.15E-04	0.000280354	2.53E-05	1.77E-05	0.000256
工作	0.001489	1.07E-04	0.000392	0.000207	0.000253
深圳	0.000192	6.45E-04	0.000307	4.48E-05	0.000253
音乐	0.000162	1.68E-03	0.000486	0.000299	0.000247
老师	0.000489	0.000147873	5.96E-05	3.08E-05	0.000243

续表4.9 使用加权调和平均值计算综合概率的前50个常用词

词条	用户词库概率	查询日志概率	Sogout 概率	Wiki 概率	加权调和平均值
密码	0.000457	0.000162657	0.000224	1.48E-05	0.00024
学校	0.000686	1.15E-04	0.000112	2.82E-04	0.000237
在线	0.000224	4.64E-04	0.000732	1.07E-05	0.000236
问题	0.00157	9.85E-05	0.000343	2.20E-04	0.000236
资料	0.00032	1.61E-04	4.70E-04	0.000303	0.000234
电视	0.000311	1.59E-04	0.000315	0.000543	0.00023
任务	0.001031	0.000105587	4.94E-05	6.43E-05	0.000229
聊天	0.000439	0.000181751	7.19E-05	7.78E-06	0.000224
百度	0.000165	1.62E-02	9.22E-05	1.09E-05	0.000224
介绍	3.12E-04	0.000155419	2.01E-04	6.85E-05	0.000222
地图	0.000145	0.003467859	0.000229	5.07E-05	0.000222
移动	1.83E-04	0.000295567	0.000105	3.14E-04	0.000213
幸福	0.0003	0.00017442	6.10E-05	1.61E-05	0.000211

通过之前字、词分析等的若干结论，查询日志往往会有不同于语言学意义上分布的特点，带有较强的目的性和实用性。因此我们考察的用户词库、查询日志和 Sogout 数据出现的 Wiki 词条分布上的相关性。将 3 个数据集两两求交集，并计算交集部分词条在不同数据集上排名（Rank）的 Spearman 相关系数，结果整理如表 4.10。

表4.10 三个数据集交集部分Rank的相关系数

	用户词库	查询日志	sogout
用户词库	-	0.6479	0.7829
查询日志	0.6479	-	0.6551
sogout	0.7829	0.6551	-

可见查询日志和其他数据的相关性相对较差，Sogout 数据的相关性最好。接着考察同一词条在不同数据集上 Rank 差的情况。用户词库与查询日志的交集中，Rank 差超过 Max(用户词库 Wiki 词条数, 查询日志 Wiki 词条数)的 20%的词条共有 154597 个，占交集词条数 421347 的 36.69%。用户词库与 Sogout 数据的交集中，Rank 差超过 Max(用户词库 Wiki 词条数, Sogout 中 Wiki 词条数)的 20%的词条共有 84282 个，占交集词条数 527406 的 15.98%。查询日志与 Sogout 数据的交集中，Rank 差超过 Max(查询日志 Wiki 词条数, Sogout 中 Wiki 词条数)的 20%的词条共有 123220 个，占交集词条数 486162 的 25.35%。从交集数量和 Rank 差的角度来看，查询日志和其他数据的相关性也相对较差。

在计算调和平均值时，尝试不使用所有维度的信息，而只采用 3 个维度的信息，发现不考虑查询日志时，结果变化较大，其他情况结果变化较小。不使用查询日志概率信息的调和平均值，排名前 50 的常用词如表 4.11。

表4.11 不考虑查询日志，使用调和平均值计算综合概率的前50个常用词

词条	用户词库概率	查询日志概率	Sogout 概率	Wiki 概率	调和平均值
的	0.012769	1.76E-10	0.013863	0.0172	0.014382
人	0.00604	1.76E-10	0.005236	0.005699	0.005639
在	0.013566	1.76E-10	0.003583	0.004801	0.005346
有	0.012492	1.76E-10	0.003974	3.90E-03	0.005101
一	0.002463	1.76E-10	0.005311	5.76E-03	0.003907
个	0.008577	1.76E-10	0.002657	2.34E-03	0.00326
是	0.002298	1.76E-10	0.003353	4.91E-03	0.003201
上	0.003955	1.76E-10	0.003369	0.002387	0.003097
会	0.004115	1.76E-10	0.002359	0.002868	0.002954
大	0.00159	1.76E-10	0.004169	0.004755	0.00278
和	5.52E-03	1.76E-10	0.001646	0.00281	0.002622
用	0.003065	1.76E-10	0.002763	2.15E-03	0.002602
到	0.005372	1.76E-10	0.00204	0.001976	0.002537
年	0.001354	1.76E-10	0.002412	1.12E-02	0.002415
中	0.00114	1.76E-10	0.004883	0.006124	0.002409
日	0.001433	1.76E-10	2.33E-03	0.005529	0.002295
下	0.004312	1.76E-10	0.002555	1.43E-03	0.002265
来	0.006305	1.76E-10	0.001806	1.59E-03	0.002237
发	0.002149	1.76E-10	0.003226	1.61E-03	0.002148
行	0.00169	1.76E-10	0.002432	0.002325	0.002094
要	0.010446	1.76E-10	0.001685	1.34E-03	0.00209
家	0.001834	1.76E-10	2.51E-03	2.00E-03	0.002079
他	0.006334	1.76E-10	1.36E-03	1.78E-03	0.002064
月	0.001585	1.76E-10	0.001598	4.84E-03	0.00205
多	0.003925	1.76E-10	0.001715	1.46E-03	0.001972
为	0.001153	1.76E-10	0.002117	0.004406	0.001915
小	0.001653	1.76E-10	0.00246	0.001799	0.001914
这	3.65E-03	1.76E-10	1.54E-03	1.50E-03	0.001884
我	0.0126	1.76E-10	0.003915	7.16E-04	0.001733
出	0.001294	1.76E-10	0.001925	2.14E-03	0.001706
能	0.004078	1.76E-10	0.001507	1.11E-03	0.00166
对	0.002862	1.76E-10	0.001269	1.37E-03	0.001608

续表4.11 不考虑查询日志，使用调和平均值计算综合概率的前50个常用词

词条	用户词库概率	查询日志概率	Sogout 概率	Wiki 概率	调和平均值
开	1.99E-03	1.76E-10	2.07E-03	1.13E-03	0.001607
天	0.00121	1.76E-10	0.002271	0.001386	0.001509
加	0.001533	1.76E-10	0.001601	1.40E-03	0.001507
不	0.000736	1.76E-10	0.003811	0.002682	0.001505
点	0.004601	1.76E-10	2.39E-03	6.39E-04	0.001363
说	0.009191	1.76E-10	0.001377	7.24E-04	0.001354
过	0.002729	1.76E-10	0.001061	1.10E-03	0.001351
后	0.000935	1.76E-10	0.001361	2.31E-03	0.001341
学	0.000683	1.76E-10	0.001951	3.60E-03	0.00133
号	0.002736	1.76E-10	9.67E-04	0.001109	0.001304
都	0.010393	1.76E-10	0.000853	0.00095	0.001293
可	0.000918	1.76E-10	1.55E-03	1.61E-03	0.001274
分	0.000683	1.76E-10	2.26E-03	2.16E-03	0.001267
新	0.000597	1.76E-10	3.99E-03	2.06E-03	0.001245
里	1.51E-03	1.76E-10	9.42E-04	1.40E-03	0.001231
时	0.000587	1.76E-10	0.002727	2.46E-03	0.001212
就	1.56E-02	1.76E-10	0.001014	6.83E-04	0.001193
得	0.002226	1.76E-10	8.86E-04	1.04E-03	0.00118

结果和算术平均类似，单字的排名较高。考察加权调和平均值，排名前 50 的常用词如表 4.12。

表4.12 不考虑查询日志，使用加权调和平均值计算综合概率的前50个常用词

词条	用户词库概率	查询日志概率	Sogout 概率	Wiki 概率	加权调和平均值
在	0.013566	1.76E-10	0.003583	0.004801	0.012985
的	0.012769	1.76E-10	0.013863	0.0172	0.012816
有	0.012492	1.76E-10	0.003974	0.003899	0.011969
就	0.015552	1.76E-10	0.001014	6.83E-04	0.011315
我	0.0126	1.76E-10	0.003915	7.16E-04	0.010394
要	0.010446	1.76E-10	0.001685	1.34E-03	0.009302
好	0.013086	1.76E-10	0.001694	4.04E-04	0.009197
都	0.010393	1.76E-10	0.000853	0.00095	0.008589
没	0.011769	1.76E-10	0.000708	0.000511	0.008477
个	0.008577	1.76E-10	0.002657	0.00234	0.00817
说	9.19E-03	1.76E-10	0.001377	0.000724	0.007762
去	0.010977	1.76E-10	0.000584	3.56E-04	0.007328
那	0.010235	1.76E-10	0.000565	0.000371	0.00704

续表4.12 不考虑查询日志，使用加权调和平均值计算综合概率的前50个常用词

词条	用户词库概率	查询日志概率	Sogout 概率	Wiki 概率	加权调和平均值
还	0.008649	1.76E-10	0.000758	4.25E-04	0.006586
看	0.00811	1.76E-10	0.002111	0.000332	0.006252
人	0.00604	1.76E-10	5.24E-03	0.005699	0.006028
来	0.006305	1.76E-10	0.001806	1.59E-03	0.005974
他	0.006334	1.76E-10	0.00136	1.78E-03	0.005972
和	0.00552	1.76E-10	0.001646	2.81E-03	0.005354
到	0.005372	1.76E-10	0.00204	0.001976	0.005196
现在	0.007879	1.84E-05	0.000241	1.83E-04	0.004483
点	0.004601	1.76E-10	2.39E-03	6.39E-04	0.00426
下	0.004312	1.76E-10	2.55E-03	1.43E-03	0.004189
会	0.004115	1.76E-10	0.002359	2.87E-03	0.004069
什么	0.009382	7.24E-05	0.000436	9.48E-05	0.004051
恩	0.009325	1.76E-10	8.53E-05	0.000251	0.003997
上	0.003955	1.76E-10	0.003369	0.002387	0.003919
能	4.08E-03	1.76E-10	1.51E-03	1.11E-03	0.0039
想	0.004783	1.76E-10	0.000719	2.83E-04	0.003878
多	0.003925	1.76E-10	0.001715	1.46E-03	0.003809
这个	0.005121	2.27E-06	0.000259	2.94E-04	0.003795
这	0.00365	1.76E-10	0.001537	1.50E-03	0.003549
打	3.94E-03	1.76E-10	7.43E-04	3.50E-04	0.003412
一个	0.003362	1.02E-05	0.000588	0.000937	0.003143
不是	0.004453	1.22E-05	0.000217	1.85E-04	0.003108
用	0.003065	1.76E-10	0.002763	0.002152	0.003047
我们	0.004359	5.56E-05	8.14E-04	1.25E-04	0.003043
做	0.003818	1.76E-10	0.000438	2.07E-04	0.003011
自己	0.003585	1.24E-05	0.000387	2.36E-04	0.002903
她	0.003327	1.76E-10	0.000453	2.76E-04	0.002813
对	0.002862	1.76E-10	0.001269	1.37E-03	0.002797
号	0.002736	1.76E-10	9.67E-04	0.001109	0.00265
过	0.002729	1.76E-10	0.001061	0.001097	0.002648
又	0.003004	1.76E-10	2.79E-04	3.47E-04	0.002564
知道	0.004865	5.78E-05	1.91E-04	7.63E-05	0.00251
一	0.002463	1.76E-10	5.31E-03	5.76E-03	0.002491
等	2.53E-03	1.76E-10	7.58E-04	1.12E-03	0.002449
把	0.002904	1.76E-10	0.000367	2.42E-04	0.002447
你	2.63E-03	1.76E-10	0.001728	2.47E-04	0.002357
是	0.002298	1.76E-10	3.35E-03	4.91E-03	0.002318

结果类似，不过在加权调和平均值的结果中双字词排名有所提升。

综上来看，查询日志是使用的数据集中最有特点的一个，因为其使用环境不同于口语或书面语等自然语言使用环境。词条分布上也是实际意义较强、查询目的性较强的常用词为主。但是正是由于其特点，往往能给予不同的信息。在基于不同数据集衡量 Wiki 词条常用程度这方面，可能根据使用目的不同分成不同情况效果较好。如果主要目的是得到偏向于语言学上的常用词，不使用查询日志会得到较好的结果，因此查询日志的词频分布与自然语言差别较大。如果主要目的是得到有较强实意的常用词，添加查询日志信息能得到较好的结果。

4.5 小结

本章主要讲述了以 Wiki 中文链接词条为基础，通过其他数据集配合衡量 Wiki 中文链接词条。实验中发现查血日志数据和其他数据集相比有较大差异和自身特点，其词频分布与自然语言有较大差异。因此利用这些数据集衡量 Wiki 词条的常用程度应从不同目的出发。如果主要考察自然语言意义上的常用词，不使用查询日志信息较好。如果主要关注有实际意义和较强目的性的常用词，使用查询日志能得到更多信息。

第5章 基于输入法输入数据的常见拼音错误模式抽取

5.1 实验背景概述

英文的自动查错、纠错研究已经有一定历史。很多编辑软件和网络应用都可以自动对输入错误的英文进行提示或修正。中文的查错、纠错如果能在输入工具中进行，对中文用户来说是很方便的。但是通过计算机输入中文的常见错误之前鲜有研究，这是在中文输入工具中进行错误提示和修正的基础。本实验主要研究的是用户在用拼音输入法输入中文过程中的常见错误。通过带有输入过程的用户输入数据中，抽取常见的错误模式。

5.2 数据介绍

5.2.1 小白狗输入法数据

小白狗输入法是在搜狗输入法 3.0 版本的基础上，为了收集用户输入数据而专门制作的特殊版本。为了保护用户隐私，该版本输入法在搜狗输入法论坛中提供下载，并说明了会记录用户输入行为。愿意下载并使用的用户，也会进行匿名处理。因为是小范围提供的性质，小白狗输入法的用户数相比搜狗输入法要少很多。小白狗输入法的记录数据样例如下：

```
[1209299276s+582250ms]pk: KeyUp:"Backspace"     :POWERPNT.EXE
[1209299276s+582942ms]pk: KeyDown:"E"       :POWERPNT.EXE
[1209299276s+583051ms]pk: KeyUp:"E"     :POWERPNT.EXE
[1209299276s+583410ms]pk: KeyDown:"N"       :POWERPNT.EXE
[1209299276s+583500ms]pk: KeyUp:"N"     :POWERPNT.EXE
[1209299276s+583505ms]pk: KeyDown:"G"       :POWERPNT.EXE
[1209299276s+583587ms]pk: KeyUp:"G"     :POWERPNT.EXE
[1209299276s+584023ms]pk: KeyDown:"Space":POWERPNT.EXE
[1209299276s+584023ms]sc:  完成       :POWERPNT.EXE
[1209299276s+584119ms]pk: KeyUp:"Space"     :POWERPNT.EXE
[1209299276s+588253ms]pk: KeyDown:"W"       :POWERPNT.EXE
[1209299276s+588368ms]pk: KeyUp:"W"    :POWERPNT.EXE
[1209299276s+588484ms]pk: KeyDown:"A"       :POWERPNT.EXE
[1209299276s+588654ms]pk: KeyUp:"A"     :POWERPNT.EXE
```

[1209299276s+588655ms]pk: KeyDown:"N" :POWERPNT.EXE
[1209299276s+588784ms]pk: KeyUp:"N" :POWERPNT.EXE

数据以行的方式记录，可以分为两类，一类是上屏数据，记录了输入法输出中文字符的时间、字符和输出的应用程序。如

[1209299276s+584023ms]sc: 完成 :POWERPNT.EXE

另一类是按键数据，记录了用户在操作输入法时的键盘按键情况，记录了时间、按下/抬起按键、按键、应用程序的信息。如

[1209299276s+582942ms]pk: KeyDown:"E" :POWERPNT.EXE

5.2.2 大白狗输入法数据

大白狗输入法和小白狗输入法类似，也是记录用户输入数据的搜狗输入法特殊版本，并面对更多用户。记录之前也经过用户同意。小白狗输入法的缺点是记录了几乎所有按键信息，包括了大量与输入法无关的数据。当用户量增多的时候，浪费也会增多。因此大白狗输入法做了一些改进，不记录用户输入拼音过程中的按键信息，只记录上屏信息以及上屏后的按键信息，不记录按键抬起的信息。按键信息方面，也只记录了删除、数字键等少量按键。大白狗输入法数据样例如下：

[2009-03-22 14:38:09 505ms]sc: 0 danyuange 单元格 danyuange :WINWORD.EXE
[2009-03-22 14:38:11 993ms]pk: KeyDown:"," :WINWORD.EXE
[2009-03-22 14:38:17 590ms]sc: 0 gusuanchu 估算出 gusuanchu :WINWORD.EXE
[2009-03-22 14:38:21 500ms]sc: 0 fafu 发福 fafu :WINWORD.EXE
[2009-03-22 14:38:22 407ms]pk: KeyDown:"Backspace" :WINWORD.EXE
[2009-03-22 14:38:22 589ms]pk: KeyDown:"Backspace" :WINWORD.EXE
[2009-03-22 14:38:25 360ms]sc: 0 fanfu 反复 fanfu :WINWORD.EXE
[2009-03-22 14:38:36 163ms]sc: 0 xuanzhuande 旋转的 xuanzhuande :WINWORD.EXE
[2009-03-22 14:38:45 765ms]sc: 0 zhu 住 zhu :WINWORD.EXE
[2009-03-22 14:38:46 585ms]pk: KeyDown:"Backspace" :WINWORD.EXE
[2009-03-22 14:38:54 344ms]sc: 0 zhuluoxuanjiang 主螺旋桨 zhuluoxuanjiang :WINWORD.EXE
[2009-03-22 14:38:56 12ms]sc: 0 he 和 he :WINWORD.EXE

大白狗数据同样以行为单位，分为上屏数据和按键数据两类。上屏数据记录了上屏时间、上屏词在输入法候选中的位置（0 代表默认词条）、用户输入拼音、上屏词条、输入法将用户拼音补全后的拼音（比如用户只打了“l”输出

“了”，输入法记录时会在上屏词条后记录补全后的拼音“le”）、应用程序。如

[2009-03-22 14:38:25 360ms]sc: 0 fanfu 反复 fanfu :WINWORD.EXE

按键数据则与小白狗输入法基本相同，记录了时间、按下的按键、应用程序。如

[2009-03-22 14:38:46 585ms]pk: KeyDown:"Backspace" :WINWORD.EXE

5.3 错误拼音模式抽取方法

因为用户输入数据中并没有标明哪些是错误拼音，需要用自动方法来判断。因为数据记录的是一系列用户按时间顺序的操作，我们可以一定程度上还原用户输入的过程。实验中采取的判断输入错误的总体思路是：如果用户输入了一段拼音序列或词条后，又在短时间内将拼音序列或词条进行少量修改，则可以认为后输入的词条拼音是前者的一种修正，可以将两者作为一对错误输入与正确输入的样例。我们将一对错误拼音序列与正确拼音序列称之为一个错误对。实际上判断一段用户输入是否有错误非常困难，理论上无法区分用户是在对原输入进行改正，还是输入了一个无关的新词。

由于小白狗数据和大白狗数据的差异性，我们希望从不同数据中得到不同信息。按照修改输入错误的时间，可以将修改分成两类：在词条上屏前对拼音序列进行修改，在词条上屏后对词条进行修改。小白狗数据量少且冗余信息多，但是记录了词条上屏前的输入序列。因此对小白狗数据，只抽取词条上屏前在拼音序列中的错误对。对于大白狗数据，只抽取词条上屏后的错误对。

抽取错误对是错误拼音模式抽取的第一步。在收集了大量错误对后，将错误对进行简化、合并可以得到一系列错误模式。

5.3.1 小白狗输入法数据错误对抽取方法

小白狗数据只抽取词条上屏前在拼音序列中的错误对。由于小白狗数据记录了大量和输入法无关的键盘操作，包括操作系统和应用程序的操作等，因此抽取小白狗数据采用的规则是：

1. 输入法处于激活状态。通过用户按键信息可以还原输入法是否处于关闭或待机状态。

2. 用户对输入法拼音序列窗口的拼音序列进行了修改，并将修改后的拼音序列上屏。通过按键信息还原可以判断是否在输入法拼音序列窗口中进行操作。这条规则说明前后两个拼音序列很有可能是错误对。
3. 用户输入、修改、上屏过程中没有关闭或清空输入法拼音序列窗口。这头规则的目的是尽量减少另打新词被当成错误对的情况。

如果满足以上条件，则将原始拼音序列和上屏时的拼音序列作为一个错误对。比如以下数据片段：

```
[1209437778s+12509475ms]pk: KeyDown:"B"   :WINWORD.EXE
[1209437778s+12509542ms]pk: KeyUp:"B" :WINWORD.EXE
[1209437778s+12509863ms]pk: KeyDown:"A"   :WINWORD.EXE
[1209437778s+12509921ms]pk: KeyUp:"A":WINWORD.EXE
[1209437778s+12510015ms]pk: KeyDown:"T"   :WINWORD.EXE
[1209437778s+12510091ms]pk: KeyUp:"T" :WINWORD.EXE
[1209437778s+12510246ms]pk: KeyDown:"I"    :WINWORD.EXE
[1209437778s+12510324ms]pk: KeyUp:"I"  :WINWORD.EXE
[1209437778s+12510643ms]pk: KeyDown:"Backspace"    :WINWORD.EXE
[1209437778s+12510702ms]pk: KeyUp:"Backspace" :WINWORD.EXE
[1209437778s+12510736ms]pk: KeyDown:"A"   :WINWORD.EXE
[1209437778s+12510816ms]pk: KeyDown:"I"    :WINWORD.EXE
[1209437778s+12510858ms]pk: KeyUp:"A":WINWORD.EXE
[1209437778s+12510899ms]pk: KeyUp:"I"  :WINWORD.EXE
[1209437778s+12511782ms]pk: KeyDown:"Space"    :WINWORD.EXE
[1209437778s+12511782ms]sc: 吧台   :WINWORD.EXE
```

用户在输入序列 bati 后将 ti 删除修改为 batai，并选择“吧台”输出，可以从中抽取 bati -> batai 作为一个错误对。

5.3.2 大白狗输入法数据错误对抽取方法

大白狗数据无法完全还原用户操作，因此在判断是否错误对上更加困难。虽然无法还原输入过程，但是可以利用用户操作序列跟踪光标位置。实验中采用了三种方法抽取错误对：

方法一：

着重于单字的拼音。随时记录最新的 10 个汉字的位置、字符、完整拼音、用户拼音。当用户移动光标，对记录中的汉字进行删除并重新输入时，对比重新输入的汉字拼音与原拼音，如果相似度大于阈值，则将原拼音与重新输入的

拼音作为一个错误对。比如用户之前输入了“塌秧能”（拼音 tayangneng），之后删除 3 个字符重新输入“太阳能”（拼音 taiyangneng）。因为“太”和“塌”的位置匹配，对比它们的拼音“ta”和“tai”，如果相似度大于阈值，ta -> tai 就被抽取为一个错误对。

方法二：

和方法一较相似，不同之处在于，方法二着重于词条整个拼音而不是单字拼音。记录中记录最新的若干词条的位置、字符、完整拼音、用户拼音。当用户进行删除修改操作时，定位到记录中对应位置的原词条，与新词条整个拼音进行对比，如果相似度大于阈值，抽取旧、新词条的拼音作为一个错误对。

方法三：

和方法二有相似之处，但是模糊化了位置对应。当用户进行删除修改操作时，除了对比新词条拼音与原词条拼音，还将原词条相邻词条的拼音与新词条拼音进行对比，选择相似度最高的一对，如果相似度大于阈值，抽取为一个错误对。

实验中计算相似度的公式为：

$$Similarity(s1,s2) = 1–ED(s1,s2)/max(L1, L2) \quad (5\text{-}1)$$

其中 s1、s2 代表两个拼音串，ED 代表编辑距离，L1、L2 分别代表 s1、s2 的长度。实验中使用的阈值是 0.5。

5.3.3 从错误对抽取错误模式的方法

当抽取了一定数量的错误对后，对每个错误对，利用最小编辑距离算法计算它们从错误拼音转换成正确拼音的最短路径，将这个转换路径中涉及到的拼音作为一个错误模式。比如 bati -> batai 这个错误对，将会抽取出“~ -> a”这个错误模式，其中“~”代表空字符。如果较长的错误对存在若干不连续的错误模式，将它们分别抽取。最后将错误模式汇总，可以得到一些高频错误模式。

5.4 实验结果及分析

5.4.1 小白狗数据

从小白狗数据中，一共抽取了 519 个频度大于 10 的错误模式。这 519 个错

误模式总频度为 36917。表 5.1 列出了最常见的 25 种错误模式，这 25 种错误模式已经覆盖了一半以上的频度。前 25 个错误模式中，删除、增加 1、2 个字符占大多数，打错字符的错误有“I - U”“U - I” “I - O” “O - I”四种，这四种的共同特点是替换字符在键盘上相邻且都是元音字母，这种错误出现的原因很可能是输入时按到了相邻按键。

表5.1　前25个高频小白狗数据错误模式

original	correct	frequency
~	I	1894
~	G	1893
~	N	1844
G	~	1733
~	A	1337
~	U	1180
~	H	973
~	E	913
~	O	784
I	~	768
H	~	691
~	NG	651
A	~	553
N	~	543
U	~	517
~	AN	516
E	~	429
O	~	377
I	U	324
U	I	306
~	Z	257
~	S	246
I	O	243
~	Y	240
O	I	234
Total		19446(52.67%)

将错误模式分类成删除、插入、替换、其他四种，并汇总在表 5.2。可以发现删除类错误占据了一半以上的总频度，说明用户在使用输入法上屏前发现的输入错误中，少打某个按键是大多数。编辑距离为 1 的错误占据了大多数频度，

说明用户在上屏前发现的输入错误中，以小错误为主，很少出现大错误。

表5.2 小白狗数据中四种错误类型统计

total frequency	insertion	deletion
edit distance 1	6893	13539
edit distance 2	922	4569
edit distance 3	235	2171
edit distance 4	25	569
edit distance 5	0	103
total	8075	20951
	substitution	others
	6125	1766

5.4.2 大白狗数据

对于大白狗数据，将三种方法分别实验并观察对比之后，发现方法三的效果较好。分析原因为：

1. 方法一只对比字的拼音，但是输入法实际输入时是以词为单位，只对比字不符合需求。有些错误会导致前后词条的字数不同，比如“什么 shenme”错打成“神恶魔 shenem”，按字匹配会不准确。
2. 用户输入的随意性较强，经常出现多删少删的情况，这会导致完全匹配位置信息的方法一、方法二匹配位置不准，从而无法抽取错误对。
3. 方法三的好处是对比词的信息，又在一定程度上模糊位置信息，可以让因为用户误操作而无法在位置上完全对应的错误对被抽取。

下面以方法三的抽取结果进行抽取结果分析。

因为大白狗数据量较大，只保留了频度大于 100 的错误模式，低频的错误模式很可能是符合抽取规则但不一定是错误的用户修改。最终得到了 785 个错误模式，总频度 6472041。前 25 个高频错误模式如表 5.3，可以发现，前 25 个错误模式已经占据了约一半的总频度，其中全部是编辑距离为 1 的错误模式。说明用户常见错误都是小错误。

将大白狗数据错误模式汇总如表 5.4，发现和小白狗数据不同，删除、插入、替换三类错误数量相当，没有明显差别，其他类的错误只占很少一部分。

表5.3 大白狗数据中前25个高频错误模式

original	correct	frequency
i	~	242965
~	i	225097
n	~	201022
a	~	192670
~	n	180874
~	e	166864
~	a	165188
e	~	160294
~	g	159664
g	~	156333
u	~	150933
~	u	146135
~	o	123897
o	~	109623
~	h	103219
h	~	102472
d	~	89560
b	~	79410
w	~	66987
y	~	63244
s	~	60446
i	o	59168
l	~	59059
o	i	55804
a	i	54863
Total		3175791(49.07%)

表5.4 大白狗数据错误模式分类汇总

	total frequency
insertion	2050393
deletion	1964394
substitution	2295891
others	161363
total	6472041

5.4.3 实验结果分析

1. 输入错误与选择错误

在使用输入法输入汉字的过程中，可能出现两种错误，一种是拼音拼写错误，一种是拼音正确的情况下选择候选词错误。我们把前者称之为输入错误，后者称之为选择错误。在抽取大白狗数据错误对的同时，我们把拼音没有改变，汉字发生变化的前后修改对也抽取出来，这些修改对可以视为选择错误。实验中抽取了 2678837 个选择错误，比输入错误的数量 6472041 少了很多，说明用户使用输入法时发生的错误大部分还是在输入拼音时发生。原因可能是输入法组词功能越来越完善，很多情况下默认结果就是所需词条。而输入法候选词条一般是 5 个，发生错误的机会比用键盘输入拼音要少一些。

2. 语言错误与键盘错误

使用输入法输入汉字不同于平常语言交流或写字，要经过计算机这条中间渠道，因此因为计算机特点发生的错误也会存在。我们可以将错误发生原因分成两类，一类是拼音错误，即用户本身记错拼音；一类是键盘错误，即用户在知道正确拼音的情况下输入时打错。因为无法得知用户在输入时的状态，这两者错误很难通过外界观察完全区分。不过通过汇总的错误模式结果，可以发现一些规律。

表 5.5 和表 5.6 分别是两个数据上高频替换错误的信息。首先可以发现，常见错误主要发生在元音的组成字母中，这与其本身的使用频度较高有关。其次，高频替换错误大多数是相邻按键（如 i-o，m-n，u-i）或者距离较远需要两手分别输入的按键。这些高频错误的发生，键盘因素很可能起到了很大影响，比如输入时错按成相邻按键，或者左右手的输入顺序弄错。

表5.5 大白狗数据中前20个高频替换错误

original	correct	frequency
i	o	59168
o	i	55804
a	i	54863
a	e	52092
i	a	45389
e	a	44670
i	u	40551
u	i	37429

续表5.5 大白狗数据中前20个高频替换错误

original	correct	frequency
i	e	33338
e	i	24865
u	a	24537
n	o	22290
n	i	22081
a	u	22071
i	n	21778
d	l	20714
b	n	20287
o	n	19938
h	d	19874
m	n	19534

表5.6 小白狗数据中前20个高频替换错误

original	correct	frequency
I	U	324
U	I	306
I	O	243
O	I	234
A	E	212
I	N	198
N	I	197
E	A	179
N	O	129
E	I	90
I	E	88
E	U	87
B	N	85
O	N	84
H	G	79
Z	C	75
S	Z	74
G	H	72
M	N	67
O	U	63

在汉语中，n、g 是鼻辅音，可以和某些元音组成鼻元音。在某些方言中，

有 n、g 和没有 n、g 有时候并不区分，导致输入拼音时会发生错误。汉语中一共包含 39 个元音，其中有 16 个鼻元音、10 个单元音和 13 个复合元音。有些方言中对这些元音的使用有不同差别。表 5.7 和表 5.8 列出了两个数据中前 20 个高频插入删除错误。这些错误中，由 n、g 引起的错误占据了一定比例，这些错误很有可能是由方言等原因引起的语言错误。

表5.7　大白狗数据中前20个高频插入删除错误

original	correct	frequency
i	~	242965
~	i	225097
n	~	201022
a	~	192670
~	n	180874
~	e	166864
~	a	165188
e	~	160294
~	g	159664
g	~	156333
u	~	150933
~	u	146135
~	o	123897
o	~	109623
~	h	103219
h	~	102472
d	~	89560
b	~	79410
w	~	66987
y	~	63244

表5.8　小白狗数据中前20个高频插入删除错误

original	correct	frequency
~	I	1894
~	G	1893
~	N	1844
G	~	1733
~	A	1337
~	U	1180
~	H	973

续表5.8 小白狗数据中前20个高频插入删除错误

original	correct	frequency
~	E	913
~	O	784
I	~	768
H	~	691
~	NG	651
A	~	553
N	~	543
U	~	517
~	AN	516
E	~	429
O	~	377
~	Z	257
~	S	246

3. 错误长度

通过实验结果可以看出，编辑距离为 1 的错误模式占据了大多数词频，编辑距离大于 3 的错误模式占据很小比例，说明用户在输入拼音时大部分错误都是很小的错误，很少有长的连续错误发生。这是符合常识的。用户用输入法输入过程中，如果发生了编辑距离大于 2 的错误，对输入法组词会有较大影响，大部分用户都会及时发觉。

4. 小白狗数据与大白狗数据的结果差异

从两个数据的常见错误模式和错误长度来看，两者还是比较统一的。两者结果的最大差别，就是小白狗数据中删除错误占据很大比例，而大白狗数据中删除错误和插入、替换错误数量相当。造成这种结果的原因，可能是用户在输入词条上屏前，希望能利用输入法的自动组词功能少输入一些字符，加快输入速度。如果输入短拼音序列后发现组词效果不理想，会修改增加序列长度，就造成了删除类错误。这个现象从另一个角度验证了用户在使用输入法时总是倾向于输入短内容的结论。

5.5 小结

本实验通过两种输入法用户输入序列数据，希望能得到用户输入拼音时的常见错误模式。采用的方法是模拟还原用户的输入过程，通过用户对输入过的

内容的修改操作寻找错误-正确的拼音对，从而进一步整理成错误模式。将输入法输入错误分类成拼音错误和选择错误两类，将拼音错误分类成语言错误和键盘错误两类。通过分析实验结果，可以得出以下几个结论：

1. 输入错误的数量要高于选择错误。
2. 键盘错误中的相邻按键按错和左右手分别按键顺序错误可能是导致很多替换类错误的主要因素。
3. 方言中鼻辅音不分等语言错误可能是导致很多插入删除类错误的主要因素。
4. 大部分错误都是小错误，很少会出现较长的错误。
5. 用户总是倾向于输入较短内容得到想要的结果。

第6章 女书拼音输入法的设计与实现

6.1 背景概述

在对输入法有了较深的了解和调研之后，与清华大学中文系合作，进行了输入法设计与实现的实践，完成了基于 Win32 平台的女书拼音输入法。女书是流传在中国湖南省江永县潇水流域的女性专用文字，是世界上目前为止发现的唯一的女性专用文字，有很高的人文价值。女书是湖南省目前唯一申请世界非物质文化遗产项目，由于旅游业和文化研究等原因，近几年逐渐受到关注。

女书字为斜体，呈“多”字型，是汉字楷体的变体（图 6.1）。女书是记录当地方言的表音文字，一个女书文字代表的当地发音可能对应若干不同汉字，一个汉字在当地发音可能有多种形式从而对应多个女书文字，因此女书文字与汉字之间有多对多映射关系，如图 6.2。

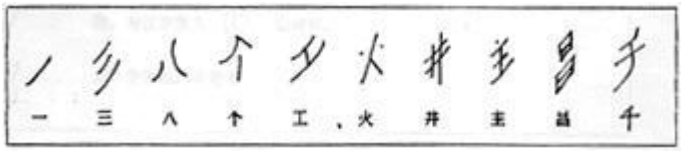

图6.1 女书文字样例（上排为女书，下排为所参考的汉字）

女书字	汉语拼音	女书字	汉语拼音
	yi		ren
	er		ba
	ru		shi
	ri		piao
	liang		pi
	qi		yu
	cha		pian
	cuo		biao

图6.2 女书文字与对应汉语拼音示例

清华大学中文系赵利明教授率领的女书团队经过多年的调研研究，整理了一系列女书相关研究成果[9][10]，揭开了女书的神秘面纱。截止到目前的研究，女书基本字约有 300 多个，不同于汉字的表意文字，这些表音的女书基本字足

以覆盖当地土话的发音。女书研究进一步开展需要对女书进行数字化建设，如建立字符集合、字体、电子图书馆和申请国际标准编码等。女书拼音输入法是女书数字化建设中的重要工作，是录入女书文字重要的辅助工具。女书拼音输入法的核心功能，就是根据用户输入的汉语拼音，生成女书的编码，并输出到应用程序。本输入法是基于 Win32 系统，使用了 Win32 平台的 IME 机制和接口。

6.2 Win32 平台的 IME 机制介绍

Win32 平台的 IME 机制是主要面对东亚用户的一种输入法机制。由于女书拼音输入法的功能特性和中文拼音输入法相近，因此也可以采用 IME 机制。IME 机制的基本原理，是应用程序、Win32 系统、输入法管理器（Input Method Manager，IMM）、输入法（Input Method Editor，IME）之间的信息传递。Win32 系统监控应用程序中的键盘事件，将键盘事件传递给 IMM，IMM 根据需要将键盘事件再传递给 IME 调用输入法。IME 转换后的文字编码，再通过 IMM 和 Win32 系统传递给应用程序，完成输入法的输出。IMM 是 Win32 系统管理输入法的程序。IME 实际上是提供给 IMM 调用的一个动态链接库（DLL）。完成 IME 机制输入法的核心，就是实现 IMM 需要调用的接口函数。Win32 系统规定的 IME 接口共有 19 个函数，列举如下：

1. ImeInquire
2. CandWndProc
3. CompWndProc
4. StatusWndProc
5. IMEConversionList
6. UIWndProc
7. ImeConfigure
8. ImeDestroy
9. ImeEscape
10. ImeSetActiveContext
11. ImeProcessKey
12. NotifyIME
13. ImeSelect

14. ImeSetCompositionString
15. ImeToAsciiEx
16. ImeRegisterWord
17. ImeUnregisterWord
18. ImeGetRegisterWordStyle
19. ImeEnumRegisterWord

这 19 个接口函数中，并不是每个函数都需要实现。实现一个基本的输入法，主要需要完成 UI 接口和编码转换两个功能。

输入法的 UI 显示主要涉及到三个窗口：Status Window 状态窗口、Composition Window 编码窗口、Candidates Window 候选窗口。如图 6.3，状态窗口用于显示输入法状态（模式、中/英文标点、全角/半角等），编码窗口用于显示当前已完成编码的字符和需要编码的拼音序列，候选窗口用于显示候选词条和对应选择按键。19 个接口函数中的 CandWndProc、CompWndProc、StatusWndProc、UIWndProc 主要对应了 UI 显示方面的功能。

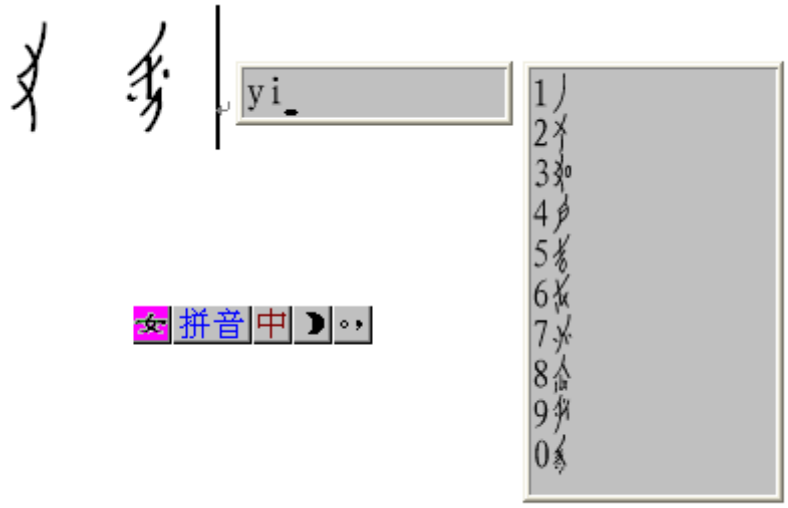

图6. 2　女书拼音输入法

19 个接口函数中编码转换的核心函数是 ImeProcessKey 和 ImeToAsciiEx。ImeProcessKey 函数负责处理键盘事件，是编码的发动者。ImeToAsciiEx 函数则负责将输入序列转换。

Win32 系统提供了一些系统函数和数据结构负责不同系统程序、函数间的通信、调用等，具体的说明请参考[11]等相关资料，在此不做详细论述。

6.3 女书拼音输入法的实现原理

为了正常显示女书文字，首先需要一个女书字体。女书字体由清华大学中文系赵丽明教授的女书研究团队提供，包含 457 个女书文字，包含了女书基本字和一些常见变体。

码表是输入法根据输入序列产生文字编码的根据。女书拼音输入法使用的码表同样是赵丽明教授的团队提供。码表共包含 1179 条拼音对应关系，囊括了所有字体中出现的女书文字。每个女书文字平均有 2.58 个对应关系，也就是平均上说，很多女书文字在汉语拼音意义上是“多音字”。女书文字对应的汉语拼音数量如表 6.1，只有 1 个拼音的女书文字最多，对应 2 个拼音的女书文字数量其次，他们占据了大部分女书文字，个别女书文字对应了很多汉语拼音，最多甚至到 15 个。

表6.1 女书文字对应汉语拼音数量汇总

单个女书字对应汉语拼音数	女书字数	单个女书字对应汉语拼音数	女书字数
1	171	7	7
2	121	8	9
3	60	9	4
4	34	10	1
5	33	13	1
6	15	15	1

女书拼音输入法包括三个文件：输入法文件、词库文件和拼音规则文件。词库文件用于记录用户输入过的词条，用户再次输入时可直接组词。拼音规则文件负责记录女书文字对应的拼音中合法的拼音规则，助于输入法组词。女书文字对应的所有拼音是汉字对应所有拼音的子集，有些拼音找不到对应的女书文字，因此拼音规则文件助于现有女书文字的拼音组词。图 6.4 展示了用户拼音不符合拼音规则文件记录时，会把当前拼音序列按最大匹配拆分成多个拼音分别寻找对应文字的情况。使用拼音规则文件的原因是当女书文字拼音规则的变化时便于扩展；另外编写其他可以参考汉语拼音规则的输入法时，可以方便的完成修改。

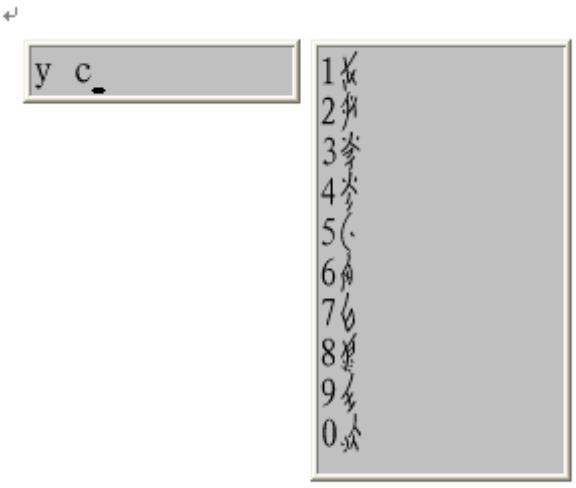

图6.4 不符合拼音规则文件记录时的输入序列处理样例

输入法启动开始初始化时，输入法会初始化各个 UI 窗口，并读取拼音规则文件和词库文件。IME 接口函数是被动地对系统和 IMM 发出的要求和信息进行处理的一系列函数，包括输入法初始化/启动/休眠/关闭请求、反馈输入法状态请求、各种键盘事件、编码开始/结束/输出的请求等等。当输入法用户输入拼音序列时，每个键盘事件都会传递给输入法管理器 IMM。输入法管理器调用 ImeProcessKey 函数判断键盘事件的处理方式：不做处理返回给系统、改变输入法状态、进行编码等等。当输入法判断需要进行编码处理时，会调用 ImeToAsciiEx 函数进行编码转换。对每个拼音序列，参考拼音规则文件的记录，用最大匹配原则对拼音序列进行分割。对分割后的每个拼音串，查找码表中的对应女书文字并在候选窗口中显示给用户，用户原则后即完成了一个拼音串的编码转换。以此类推将用户输入的整个拼音串转换完毕，并转换完毕后的文字编码发送给输入法管理器。输入法管理器将转换后的编码输出给目标应用程序显示。以上过程完成了女书拼音输入法基本的将拼音转换成女书文字编码的基本功能，也是女书拼音输入法实现的基本原理。

6.4 小结

本章主要介绍了 Win32 平台下的 IME 输入法机制，介绍了 19 个接口函数中若干重要函数的功能，并介绍了利用 IME 机制实现女书拼音输入法的原理。

由于女书还处在研究中，因此女书拼音输入法保留了很多可扩展性，方便对女书拼音规则的扩充进行处理。

女书拼音输入法的实现利用了女书源于汉字，和中文拼音输入法有一定相似性的特点。理论上与中文输入原理相似的其他语言文字也可以用类似女书拼音输入法的方式实现，甚至在女书拼音输入法的基础上修改扩展实现。这种实现拼音输入法的思路，希望对其他语言文字输入法的研发有所帮助。

第7章 结论

7.1 论文成果总结

本文可以分为两个主要部分。第一部分的主要目的是利用输入法用户词库和搜索引擎查询日志等数据，分析当前网络环境下不同渠道的语言特点，并与传统的中文常用语对比分析，并且希望结合不同环境下的数据得到当前网络环境的常用词表。

首先，本文对输入法和搜索引擎的用字情况进行分析。通过输入法、搜索引擎、媒体、传统中文四个渠道的数据进行单独、交叉的分析，总结了不同渠道数据的异同，以及各个频度段上的特点。

其次，本文对输入法和搜索引擎的用词情况进行分析。对不同数据的整体情况、不同词条长度的情况进行统计分析，并与传统的中文常用词进行对比。

之后，本文结合严谨性相对较好的 Wiki 数据，综合输入法用户词库、查询日志和 Sogout 数据三个大规模语料，尝试了若干种评价词条常用程度的方法，并对实验结果进行了分析。

第二部分主要涉及到输入法的具体功能和一些细节。首先，本文介绍了通过输入法用户输入序列数据抽取拼音输入中常见错误模式的方法。通过小白狗输入法和大白狗输入法两种数据，关注错误发生的不同时期，通过不同方法抽取出常见错误模式。并对实验结果和错误可能发生的原因进行分析，得到了若干结论。其次，结合之前在输入法上的了解和研究，实现了女书拼音输入法，介绍了 Win32 平台下输入法的运行机制和实现原理。女书输入法目前已经用于女书研究中。

7.2 课题研究展望

1. 通过用户词库、查询日志、Wiki 数据等的分析，结合更多语言学知识，得到当前网络中常用词表和新兴常用词表。
2. 改进拼音错误模式抽取的方法，改进输入法自动纠错功能。
3. 结合人体工学知识，细化键盘等因素对输入错误的影响。

4. 对女书拼音输入法进行改进，通过女书文字的词频分布特性等实现智能组词等功能。

参考文献

[1] 中国互联网络信息中心(CNNIC). 第 25 次中国互联网络发展状况统计报告. 2010

[2] 中国互联网络信息中心(CNNIC). 第 24 次中国互联网络发展状况统计报告. 2009

[3] 国家语言资源监测与研究中心. 中国语言生活状况报告 2007. 北京：商务印书馆, 2008

[4] 余慧佳, 刘奕群, 张敏, 茹立云, 马少平. 基于大规模日志分析的网络搜索引擎用户行为研究.中文信息学报, 2006, 21(1): 109~114

[5] Aminul Islam and Diana Inkpen. Real-Word Spelling Correction using GoogleWeb 1T 3-grams. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2009. 1241~1249

[6] Eric Brill and Robert C. Moore. An improved error model for noisy channel spelling correction. In: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2000. 286~293

[7] Yiqun Liu, Rongwei Cen, Min Zhang, Shaoping Ma, and Liyun Ru. Identifying Web Spam with User Behavior Analysis. In: Proceedings of the 4th international workshop on Adversarial information retrieval on the web. New York: ACM, 2008. 9~16

[8] 郑林曦编. 普通话三千常用词表. 北京：语文出版社, 1992

[9] 赵丽明, 等. 女书用字比较. 北京：知识产权出版社, 2006

[10] 赵丽明编著. 中国女书合集. 北京：中华书局, 2005

[11] David Iseminger 主编. Win32 开发人员参考库. 北京：机械工业出版社, 2001

[12] 王鹏，孙茂松. Win32 平台下女书拼音输入法的设计与实现. 见：第五届全国青年计算语言学研讨会论文集, 2010. 508~514

致　谢

衷心感谢我的导师孙茂松教授对本人的精心指导！孙老师积极参与并指导研究工作，乐于与学生交流沟通。他的科学精神和高尚品德将使我终生受益。

感谢清华大学搜狐联合实验室和清华大学自然语言处理组的老师和同学们！是他们的帮助，为本研究工作提供了优越的实验条件。与他们的交流合作，也是本研究工作得以顺利进行的保障。

本课题承蒙清华搜狐联合实验室项目和国家自然科学基金编号 60873174 资助，特此致谢。

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：　　日　期：

个人简历、在学期间发表的学术论文与研究成果

个人简历

1986 年 8 月 15 日出生于天津市。

2004 年 8 月考入清华大学计算机系计算机科学与技术专业，2008 年 7 月本科毕业并获得工学学士学位。

2008 年 9 月免试进入清华大学计算机系攻读工学硕士至今。

发表的学术论文

[1] 王鹏，孙茂松. Win32 平台下女书拼音输入法的设计与实现. 见：第五届全国青年计算语言学研讨会论文集, 2010. 508~514