基于日志分析的中文输入法用户行为研究

许丹青，刘奕群，岑荣伟，马少平，茹立云，杨磊

智能技术与系统国家重点实验室，清华信息科学与技术国家实验室（筹），清华大学计算机系，北京，100084

邮箱：xudanqing06@gmail.com

摘 要：与拼音文字不同，用户在进行中文输入时需要借助输入法软件完成从拼音串到汉字串的转换过程，输入法因此成为中文用户进行人机交互的基础性工具，而输入法的相关技术研发也一直是学术界与产业界的关注热点。在中文输入法技术的研究中，用户的行为特点对输入法软件的词库建立、算法设计、交互方式设计与性能评价等多方面都有着至关重要的作用，但由于数据获取与分析的困难，这方面的相关研究尚不多见。本文利用某中文输入法在用户许可下收集的超过 4.1 亿条用户输入行为记录，进行了中文输入法用户行为的分析研究，针对不同类别应用程序的输入词频差异，不同用户在同类应用程序中的不同候选词条的选择等行为特点进行了挖掘分析，研究结果会对深入了解中文输入法用户行为，进而改进输入法软件性能具有一定的指导意义。

关键词：中文输入法，用户行为，日志分析

Research on User Input Behavior, Based on Log Analysis of a Chinese Input Method Editor

Danqing Xu, Yiqun Liu, Rongwei Cen, Shaoping Ma, Liyun Ru, Lei Yang

State Key Laboratory of Intelligent Technology and Systems,Tsinghua National Laboratory for Information Science and Technology,Department of Computer Science and Technology, Tsinghua University, Beijing,100084, China

Email: xudanqing06@gmail.com

Abstract: Different from alphabetic languages, input software is required to transform PinYin strings into characters for Chinese language. Input software therefore plays an important role in HCI process for Chinese users. In the research field of Chinese input method, it is important to look into users' behavior information to improve the performance of dictionary construction, algorithm and interaction designing as well as performance evaluation. However, there lacks such works due to the difficulties in collecting corresponding behavior data. With the help of a widely-used Chinese input software company, we collected user input logs under users' agreement which contain 410 million input strings. With analysis into these input logs, we focused on the following behavior features: input string length distribution, character/word/phrase selection for different kinds of application software and the adoption of abbreviations. Conclusions help us to better understand users' input behavior and show possible ways to improve input software designation.

Keywords: Chinese input software, User behavior, Log Analysis.

1 引言

随着互联网在中国的蓬勃发展，中国网民的比例越来越大，据统计，直至 2010 年初，中国网民已达到了 3.84 亿，位于全球首位[1]。中文输入法作为中国网民在互联网上必不可少的输入工具，其作用也变得越来越重要。与英文输入不同，中文的很多候选词条所对应的拼音输入相同，如果对于一个拼音输入而言，大多数用户都选择了某个词条作为最终的结果，那么我们就将这个词条排在越靠前的位置。这也是当前的很多输入法所采用的排序策略的主要思路。由于数据收集和分析的困难，前人对输入法的研究主要集中用自然语言处理的方法，将现有的一些分词算法和排序算法进行一些优化和改进[2][3]，而对目前输

入法的用户行为研究并不多见。本文中，我们将主要对经过用户许可收集的某中文输入法一天的数据进行分析，希望可以从中得到输入法用户行为的一些普适特征，为当前的输入法日后的改进有一定的帮助。由于日志数据规模较大，所以结论更具一般性，更能反映大多数用户的行为特征。

本文后续的内容组织如下：第二部分是对前人相关工作的一些概述，第三部分简单描述了我们采用的输入法格式，第四部分会从三个不同方面对输入法日志的用户行为进行分析，第五部分从分析的结果中尝试得出一些对当前输入法改进有用的启示和结论。第六部分是参考文献说明。

2 相关工作概述

由于 web 数据规模庞大且数据格式等比较杂乱，因此，现有的输入法必须经过分词，过滤，排序等操作才能得到能够供用户使用的词条。同时由于与用户体验联系紧密，这部分的工作主要集中在产业界，目前研究界使用较多的是用户查询词日志，前人多次使用查询日志对用户需求进行分析[4]，最典型的使用就是查询推荐和查询纠错。而目前的产业界的输入法分析改进主要集中在全局的词库生成算法和词条排序算法，通过词库生成算法生成格式规整的词库，然后再通过词频等一系列的特征进行权重计算，对候选词条进行最终的排序。

前不久，中国科学院研究生院的张玮[5]等人提出了一种结合分类模型的中文输入法，他们针对现有输入法很少用的候选词本身的特性，将输入法词库进行了类别标注，同时根据用户当前输入串的上下文判定当前的输入语境类别，提高属于当前类别词库的词条在整个词库中对应权重，这样符合上下文语境的词条就会排在相对靠前的位置。这种对输入法词库进行类别标注并在输入法系统中集成分类引擎的方法，可以在一定程度上提高用户的输入效率和体验度的，然而，其缺乏大规模数据集的训练和测试。

此外，输入法的用户交互行为也很重要。之前，我们就某中文输入法某一天的日志进行了简单的分析[6]，主要从用户使用的应用程序的相对熵值和用户半径这两个特征简单分析了各个应用程序类别的用户行为，研究结果显示有着相同用户需求的应用程序之间有着相似的属性。基于之前的工作，本文我们将从其他多个方面对输入法的用户行为进行详细分析。

3 输入法日志格式说明

该实验中所采用的日志是来自于某中文输入法“用户体验提升计划”记录的用户行为日志，该部分日志得到了行为信息收集对象的同意，完全为匿名记录。我们选取了 20091101

表 1： 输入法日志格式

ID	系统分配给用户的标识号
Time	系统时间
String Inputed	用户输入的拼音串
Chinese Chosen	用户最终选定的中文候选词条
String Needed	候选词串所对应的标准拼音串
Application	用户正在使用的应用程序

当天的日志数据，其涵盖了 4.1 亿条用户输入条目，417 万个独立用户和 8232 个应用程序。其中每条用户输入条目包含的信息如表 1。

4 基于输入法日志的用户行为分析

4.1 词频与应用程序使用频度的分析

该实验中，我们对日志中的所有词条以及所有应用程序按照其使用频度进行了排序，并分析了使用频度和其对应的排序之间的关系，如图 1。

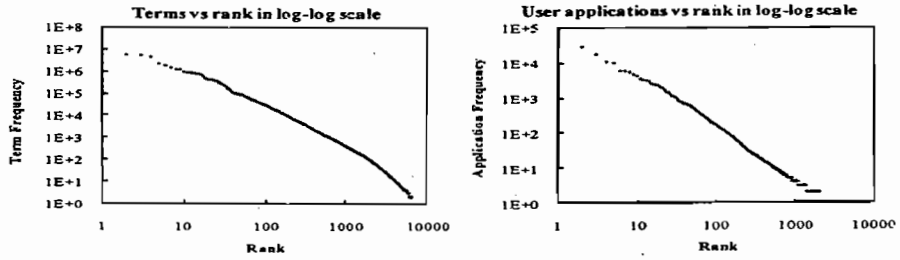

图 1：词条和应用程序的使用频度与对应排名的关系

图 1 显示，应用程序和词条的使用频度是符合幂律分布的，即极少数的应用程序和词条的使用频度占了所有应用程序频度的绝大多数。统计发现，排名前 10 的应用程序的频度占了输入法总数的 77.4%，而排名第 1 的应用程序—— 腾讯 QQ 聊天程序的使用频度占了总使用频度的 44.2%。这里，为了更加方便的对排名前 10 的应用程序进行深入的研究和分析，我们在表 2 列出了使用频度排名前 10 的应用程序的名称。

表 2：排名前 10 对应的应用程序

排名	应用程序名称	应用程序说明
1	QQ	腾讯 QQ 即时通信软件（即时通信类）
2	DNF	地下城与勇士（游戏类）
3	iexplore	IE 浏览器（浏览器类）
4	winword	Word 文档程序（办公编辑类）
5	wow	魔兽世界（游戏类）
6	360se	360 浏览器（浏览器类）
7	my	梦幻西游网络游戏（游戏类）
8	wxcel	Office 表格程序（办公编辑类）
9	allim	阿里旺旺程序（应用类）
10	client	一些网络游戏客户端（游戏类）

从表 2 可以看出，使用频度排名前 10 的应用程序主要分为四类：即时通信类，游戏类，浏览器类和办公编辑类。本文中，我们将针对这四个主要类别和这 10 个应用程序的词频等特征展开分析。

4.2 各个主要应用程序间一元，二元词频相似度分析

在这些主要应用程序的基础上，我们还分别对它们的一元与二元高频词进行统计，根据两个应用程序之间的一元（或是二元）词频的前 200 个词中所包含的相同词的个数，从而计算两个应用程序之间的相似度。假设应用程序 a 一元（或是二元）词频的前 200 个词的集合为 Va，应用程序 b 的一元（或是二元）词频的前 200 个词的集合为 Vb，我们定义 a 与 b 的相似度计算公式如公式 1。

$$Sim_{a,b} = \frac{|Va \cap Vb|}{200} \qquad \text{公式 1}$$

另外，除了对应用程序之间进行了词频相似度的计算外，我们还对两个类别之间的应

用程序进行了平均相似度的计算。假设有 k 个程序属于某一类别 A(i=1…k)，n 个程序属于某一类别 B(j=1…n)，类别 A 和类别 B 之间的一元（二元）词频的平均相似度计算公式如公式 2。

$$Sim_{A,B} = \frac{\sum_i \sum_j sim_{i,j}}{k * n} \qquad \text{公式 2}$$

4.2.1 各种主要应用程序之间词频相似度对比

我们对输入法中的主要应用程序（见表 2）的一元词频的 top200 进行了提取，并分别根据公式 1 和公式 2 进行了一元相似度和二元相似度的计算，表 3 显示了各个程序之间的相似度结果，其中下图的上三角表示的是一元相似度的计算结果，而下三角表示的是二元相似度的对应的计算结果。

表 3：主要应用程序的词频相似度

Application	qq	fetion	dnf	iexplore	winword	wow	360se	excel	client	maxthon
qq	—	0.79	0.36	0.56	0.21	0.45	0.59	0.12	0.42	0.56
fetion	0.76	—	0.31	0.48	0.18	0.43	0.51	0.14	0.38	0.52
dnf	0.23	0.18	—	0.25	0.08	0.56	0.25	0.1	0.62	0.25
iexplore	0.53	0.49	0.15	—	0.35	0.33	0.87	0.29	0.3	0.81
winword	0.08	0.18	0	0.19	—	0.1	0.3	0.38	0.1	0.32
wow	0.37	0.31	0.38	0.23	0.03	—	0.34	0.09	0.64	0.33
360se	0.53	0.48	0.15	0.86	0.18	0.24	—	0.21	0.3	0.79
excel	0.005	0.14	0.05	0.05	0.16	0.04	0.04	—	0.12	0.2
client	0.36	0.29	0.48	0.22	0.02	0.57	0.21	0.01	—	0.29
maxthon	0.46	0.45	0.11	0.74	0.22	0.22	0.76	0.05	0.18	—

通过表 3 的词频相似度结果的对比，我们可以很明显的看出：有着相同的用户需求的应用程序的词频相似度要明显高于其他不同需求的程序。如程序“qq”与“fetion”的之间的相似度明显高于其和“dnf”，“winword”等程序。在本节中，我们主要就高频词条的相似度作为衡量应用程序的用户行为的一个特征。表 3 的结果显示，相同类别的应用程序之间的高频相似度很高，即相同类别的应用程序有着相似的用户行为。接下来，为了进一步深入探索用户需求与应用程序的关系，接下来我们对主要应用程序类别之间的一元平均相似度和二元平均相似度分类别进行观察比较。

4.2.2 各类主要应用程序之间词频的相似度对比

表 3 显示了相同类别的应用程序的词频相似度很高，因此，我们将各个主要应用程序根据其用户需求的不同进行了类别分类，并且对类别之间和类别之内的平均相似度按照公式 2 进行计算。类别内部的平均相似度是所有这个类别内部的应用程序相似度的平均值，与此类似，类别之间的平均相似度是分属于两个类别的应用程序相似度的平均值。图 2 显示了类别之间的一元词频的平均相似度的结果。

图 2 显示，不同类别内部的平均相似度也是不同的。其中，浏览器类的高频词条的相似度最高，达到了 0.85。浏览器类的词频分布相对集中，需求比较明确，而办公和游戏类用户群体差异较大，需求比较分散，这也在一定程度上解释了浏览器内部的平均相似度最高，而办公和游戏类相似度较低的结果。另外，我们还发现，不同类别之间的相似度也是不同的，聊天程序和浏览器的相似度达到了 0.5 左右，而和办公编辑类的只有 0.1。这表明了聊天类和浏览器类之间有着一定的相似度。为了进一步的对进行类别之间的比较，我

们也按照类似的方法对二元词频相似度进行了统计，统计结果见图 3。

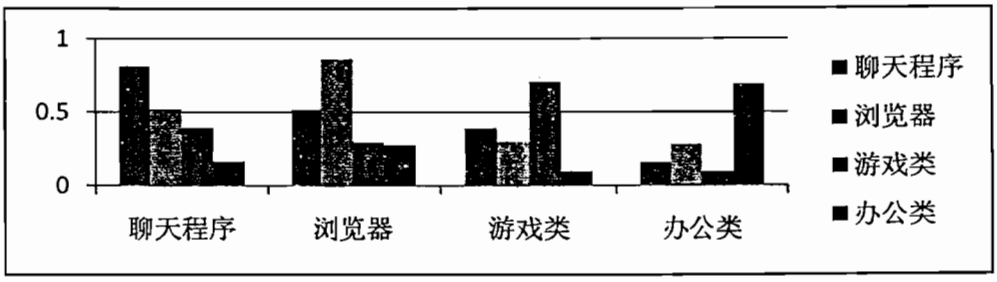

图 2：不同类别之间的一元词频相似度

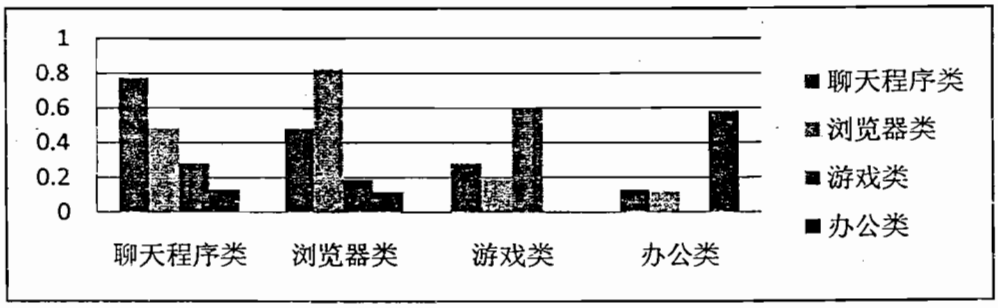

图 3：不同类别之间的二元词频相似度

将二元词频的相似度和一元词频相似度进行对比，各类程序之间的相似度趋势几乎没有发生变化。浏览器类内部的相似度依然最高，而办公类和游戏类之间的平均相似度几乎为 0，即办公类的高频词和游戏类的高频词几乎没有重合，这与办公类和游戏类就用户需求相差较大的原因有一定的关系。通过上述的一元词频和二元词频相似度的分析，我们可以假设，如果对现有输入法词库和排序策略按照应用程序的类别属性进行更新和改进，应该可以在一定程度上提高用户的满意度。

4.3 各个主要应用程序之间的 KL 距离分析

上一节中，我们从高频词频的相似度对不同应用程序类别的用户行为习惯进行了分析，下面我们将选定相对熵和用户半径这两个指标作为另一种衡量应用程序之间的差异的特征[7]。首先，相对熵的定义如下：

$$Entropy = \sum_{w} -p(w)\log p(w) \qquad \text{公式 3}$$

其中 p(w)代表某个词 w 的概率分布，Entropy 代表的是整个应用程序的相对熵。如果某个应用程序的相对熵很小，表示其对应的用户行为相对较为集中。

用户半径衡量的是应用程序中的每个用户到该应用程序总体词频分布的中心点的距离，每个单独的用户的半径作为一个独立的点，没有统计意义，从而我们将应用程序的用户半径定义为此应用程序中所有用户半径的平均值，其是衡量用户行为另一个重要的指标。用户半径越小，代表此应用程序的用户群的行为特征有着较高的相似性。应用程序对应的用户半径定义如下：

$$Radius = \frac{1}{|U|}\sum_{u \in U}\sum_{w} p(w)log\frac{p(w)}{P_{centroid}(w)}$$ 公式 4

其中 $P_{centroid}(w) = \frac{1}{|U|}\sum_{u \in U} P_u(w)$，U 表示所有安装此应用程序的用户集合，|U|代表安装此应用程序的用户总数。接下来，我们用上述提到的两个指标对各个主要应用程序之间的用户行为特征进行分析，各个主要应用程序的熵值与用户半径见表 4。

这一节中，我们主要采用应用程序的相对熵和用户半径作为衡量应用程序的用户行为的主要特征，表 4 显示了属于同一类的应用程序在熵值和用户半径有着相似的属性。游戏

表 4：主要应用程序的熵值与半径

应用软件	熵值	用户半径	软件种类
qq	10.8	3.5	即时通信类
wow	10.3	3	游戏软件类
game	10.1	3.8	游戏软件类
war3	10	4	游戏软件类
msnmsgr	10.8	4.2	即时通信类
fetion	10.9	4.8	即时通信类
wps	12.7	6.4	编辑器类
winword	13.3	7.1	编辑器类
theworld	12.8	7.3	浏览器类
maxthon	13	7.4	浏览器类
kwmusic	11.9	7.5	音乐软件类
ppstream	11.2	7.7	视频软件类
kugoo	12.1	7.8	音乐软件类
iexplore	13.2	7.9	浏览器类
notepad	12.7	7.6	编辑器类

软件类和即时通信类的相对熵值和用户半径都比较小，这表示这两类的应用程序之间的差别相对较小，可能是因为此类应用程序的用户之间的联系相对紧凑一些，而相对而言，浏览器类，休闲娱乐类等的用户群分布较为分散。之前的高频词相似度是从不同应用程序的词频分布进行衡量的，而这一节主要是就用户分布情况进行比较。将两者的结果进行综合，可以就应用程序类别的词频使用情况和用户群有了较全面的评估。另外，本文中，我们还将继续采用一些其他特征，对中文输入法的用户行为进行更全面的分析。

4.4 其他用户行为分析

4.4.1 基于应用程序类别的首词条命中率和翻页分析

首词条命中率，即排在第一位的候选词条被命中的比例，翻页率则是用户在输入法提供的候选结果中进行翻页的比例。首词条命中率和翻页率一直被认为衡量输入法效率很主要的两个指标，因此，我们将对各个类别对应的首词条命中类别和翻页率进行了统计，统计结果如图 7。

上述的图表显示，即时通信类两者的质量评估总体是最好的，而游戏类的则相对较差。同时，首条命中率和翻页率随不同的应用程序类别也是不同的。究其原因，可能是因为不同的拼音串在不同的应用程序类别中，用户所需要的中文候选词条可能互不相同。这样可能导致在一些类别的应用程序中首次命中率下降，翻页率上升，从而导致用户体验度下降。

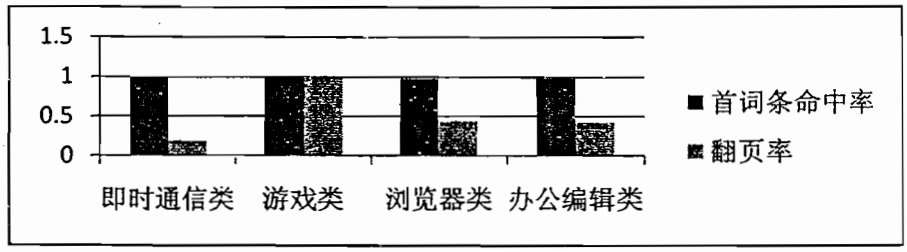

图 4：各种应用程序的首词条命中率和翻页率对比

5 结果与讨论

衡量中文输入法产品的优劣有很多指标，其中包含首条词条命中率，翻页次数，内存占用量，简拼词使用率等，这些都直接影响着用户的满意度。本文我们主要从两个方面对输入法日志进行了分析：首先，就一元，二元词频相似度分析而言，我们发现有着相似用户需求的应用程序（同一类别）之间的词频相似度要明显高于不属于同一类别的相似度。而且不同类别内部的平均相似度也是不相同的。其中，浏览器类别内部的相似度最高（达到 0.85），表明浏览器类的词条需求较为集中。其次，我们对应用程序的熵值与用户半径进行了定义，并用其各大主要应用程序进行类别分析。分析发现，用户需求相似的应用程序在熵值和用户半径有着相似的属性。其中游戏软件类和即时通信类的熵值和用户半径比较小，表明这两类所对应的用户群相对较为集中。同时，我们发现，当前的首词条命中率和翻页率也是随着应用程序类别发生变化的。即时通信类别的效果最好，而游戏类别相对较差。

词库和排序策略是整个输入法的核心。目前输入法在整个应用程序体系中均采用了一致的词库和排序算法，然而很多情况下能够达到全局最优的条件未必会是局部最优。目前即时通信类用户占了整个输入法用户的 50%左右，于是当前的词条排序大都是基于即时通信类用户最优的，而其他类用户和即时通信类用户需求之间有一定的差异，从而导致了当前的词条排序不能很好的满足其他类用户的输入需求。通过我们的分析发现，如果我们可以对于不同用户需求的应用程序，选用不同的词库和排序算法，则会使得不同应用需求的用户的满意度都得到一定的提升。

用户需求是输入法最切实的质量评估，然而也是最抽象的，有时可能一个应用程序会对应着不同的用户需求，所以为应用程序定位其对应的用户需求，从而进行针对性改进，是我们下一步研究的方向。

参 考 文 献

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

[2] Cen Z,Lee Kai-Fu L. A new statistical approach to Chinese pinyin input. The 38th Annual Meeting of the Association for Computational Linguistics,2000.

[3] Manning CD, Schutze H. Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, Massachusetts, 1999.

[4] Downey D, et al. Understanding the relationship between searchers' queries and information goals. CIKM, 2008, 449-458.

[5] 张玮，等. 一种结合分类模型的中文输入法. China Academic Journal Electronic Publishing House. 1994-2010.

[6] Rongwei C, et al. Study Language Models with Specific User Goals.WWW, ACM, 2010.

[7] Jianhua Lin. Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory, 1991, 37:145-151.