Study Language Models with Specific User Goals*

Rongwei Cen, Yiqun Liu, Min Zhang, Liyun Ru, Shaoping Ma
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
crw@mails.tsinghua.edu.cn

ABSTRACT
Under different language contexts, people choose different terms or phrases to express their feelings and opinions. When a user is writing a paper or chatting with a friend, he/she applies a specific language model corresponding to the underlying goal. This paper presents a log-based study of analyzing the language models with specific goals. We exhibit the statistical information of terms and software programs, propose some methods to estimate the divergence of language models with specific user goals and measure the discrimination of these models. Experimental results show that the language models with different user goals have large divergence and different discrimination. These study conclusions can be applied to understand user needs and improve Human-Computer Interaction (HCI).

Categories and Subject Descriptors
H.1.2 [**Models and Principles**]: User/Machine Systems – *Human Information processing.*

General Terms
Measurement, Experimentation, Human Factors, Languages.

Keywords
Log Analysis, User Goal, Language Model.

1. INTRODUCTION
Language is one of the most important interaction mediums and tools in daily life. With different goals, users choose different words or phrases to express their feelings and opinions. If local and personalized language with specific goals can be studied and modeled, HCI will be more effective and receptive to users' needs.

Logs containing the language interactions between users and machines have been applied to studied user language model. In previous work, researchers used search query logs to analyze user's information needs [1]. The typical application of user query logs is search recommendation and query correction.

In this paper, we present a log-based study that estimates the characteristics of user language model with specific goals. We conducted this study using a log-based methodology since input method logs contain evidence of real user languages and provide coverage of many types of user goals.

2. USER LANGUAGE DATA
Input method is a system software program that people use to enter characters not found on their keyboards, such as Chinese and Japanese. The primary source of data for this study was the anonymous logs of input method given permission by Chinese users to log their text input. For users who give permission, all input text and corresponding software programs are logged. We gathered about 7.6 million entries of 42.6K users, and it contained 7106 different programs.

To help ensure experimental integrity, we normalized user input text phrases, and divided the Chinese phrases into terms using segmentation. After the preprocessing work, logs were formatted into a set of triples, user id, term and software program.

Before proceeding to analyze the user language models, we verify term frequencies and software program frequencies follow a powerlaw distribution: a few terms and programs are very frequent while a large number occur only a few times (Figure 1). To ensure the studied programs contain enough terms, we selected top frequency programs and corresponding terms and users.

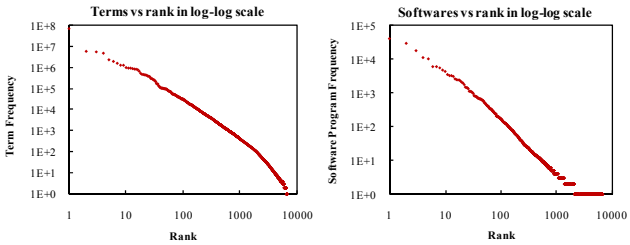

Figure 1. The powerlow distribution of term/software program frequencies vs ranks in input method logs.

3. METHODS
Under different language contexts, users have different goals and construct specific models. Intuitively, people choose different language terms or phases to express their goals. The models under different contexts present different characteristics. The models (vertexes) and the divergences (edges) between them form a graph. Figure 2 shows a divergence distance between these language models as an example.

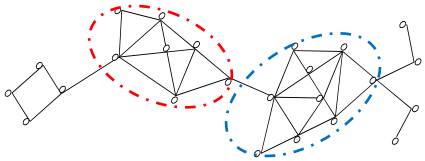

Figure 2. The graph of the models and the divergences.

Here, we introduce two methods to study the models with specific goals: divergence between different models and discrimination of

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

* Supported by Natural Science Foundation (60736044 , 60903107) and Research Fund for the Doctoral Program of Higher Education of China (20090002120005)

each model. These two methods capture intrinsic characteristics of language models and relationships between them.

3.1 Divergence between Language Models

In general, the Kullback-Leibler (KL) divergence is calculated between two term frequency distributions. The definition of the KL divergence is:

$$KL(p_q \| p_\tau) = \sum_w -p_q(w) \log \frac{p_q(w)}{p_\tau(w)}$$

where p_q and p_τ are the relative frequencies of term w in the distributions. While the KL divergence measures model similarity in an asymmetric fashion and the probability $p_\tau(w)$ may be zero, we adopt the Jensen-Shannon (JS) divergence [1] in a symmetric fashion. The JS divergence simply calculates the average KL divergence between each distribution and the mean of the two:

$$JS(p_q \| p_\tau) = \frac{1}{2} KL(p_q \| p_m) + \frac{1}{2} KL(p_\tau \| p_m)$$

where $p_m(w) = 0.5 p_q(w) + 0.5 p_\tau(w)$. This quantity has the advantage that the vocabularies of distributions don't need to be matched and smoothed. Here, we adopted JS divergence to quantify the models.

In addition, we compared vocabularies to see whether the same terms are being used in both language models. To calculate the relative overlap between the vocabularies of two models (denoted V_q and V_τ), we used the Overlap coefficient [2] and the vocabularies are made by the top N words in each model:

$$Overlap = \frac{|V_q \cap V_\tau|}{N}$$

3.2 Discrimination of Language Models

We adopted two metrics to measure term distribution of each language model, entropy and radius. The entropy of software program is defined as:

$$Entropy = \sum_w -p(w) \log p(w)$$

The radius measure whether language models of each user are concentrated on similar distribution for the same program. We defined the radius as the average divergence between each user model and the centroid of the program, defined below:

$$Radius = \frac{1}{|U|} \sum_{u \in U} \sum_w p(w) \log \frac{p(w)}{p_{centroid}(w)}$$

where $p_{centroid}(w) = \frac{1}{|U|} \sum_{u \in U} p_u(w)$, U is the set of users who install the corresponding software program.

4. EXPERIMENTAL RESULTS

We adopted the metrics to measure the divergence between different program models and discrimination of each one.

Table 1. The JS divergence between some program models

	wps	winword	iexplore	maxthon	fetion	qq	wow	game
wps	-	0.09	0.21	0.21	0.33	0.33	0.40	0.42
winword	0.09	-	0.19	0.20	0.34	0.35	0.42	0.44
iexplore	0.21	0.19	-	0.04	0.16	0.13	0.22	0.24
maxthon	0.21	0.20	0.04	-	0.16	0.14	0.22	0.24
fetion	0.33	0.34	0.16	0.16	-	0.04	0.15	0.15
qq	0.33	0.35	0.13	0.14	0.04	-	0.11	0.11
wow	0.40	0.42	0.22	0.22	0.15	0.11	-	0.08
game	0.42	0.44	0.24	0.24	0.15	0.11	0.08	-

The JS divergence was adopted to measure the divergence between programs, and the result shows that some programs have similar language models and these models have the same user goals. Table 1 details the divergence between some programs. From the table, we know that the language models on *fetion* and *qq* software programs have smaller divergence due to the similar application context (Instant Messenger program, IM).

Table 2. The vocabulary overlaps of the top 200 words of some IM and explore software programs.

	qq	fetion	msn	IExplore	Maxthon	360se
qq	-	0.76	0.66	0.53	0.46	0.53
fetion	0.76	-	0.59	0.49	0.45	0.48
msn	0.66	0.59	-	0.50	0.47	0.46
IExplore	0.53	0.49	0.48	-	0.74	0.86
Maxthon	0.46	0.45	0.51	0.74	-	0.76
360se	0.53	0.48	0.50	0.86	0.76	-

Table 2 shows the vocabulary overlaps of some IM and explore programs. The vocabularies are normalized using segmentation and stopword, and N is set as 200. From Table 2, we can see that the overlaps are high when the models have the same user goals.

Table 3 show the characteristics of a sub set of models and it shows the similar user goals have the similar characteristics. The Game Client and IM have less discrimination, since the most of users are chatting with each other. The two metrics have a similar performance, and the correlation is 0.889.

Table 3. The entropy and radius of some applications

Application	Entropy	Radius	User Goal
wow	10.3	3	Game Client
qq	10.8	3.5	IM
game	10.1	3.8	Game client
war3	10	4	Game Client
msnmsgr	10.8	4.2	IM
fetion	10.9	4.8	IM
wps	12.7	6.4	Text editor
winword	13.3	7.1	Text editor
theworld	12.8	7.3	Browser
maxthon	13	7.4	Browser
kwmusic	11.9	7.5	Music
notepad	12.7	7.6	Text editor
ppstream	11.2	7.7	Video
kugoo	12.1	7.8	Music

5. CONCLUSIONS

In this paper, we study user language models with different user goals through user input log analysis. We measured the models from two levels, divergence between different models and discrimination of each model. The results shows the language models with the same goals are similar, and the models with chatting goals have small entropy/radius. The conclusions of this study can be applied to understand user need and improve Human-Computer Interaction (HCI).

6. REFERENCES

[1] Downey, D., Dumais, S., Liebling, D., and Horvitz, E. 2008. Understanding the relationship between searchers' queries and information goals. CIKM '08. 449-458.

[2] C. D. Manning and H. Schutze. Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, Massachusetts, 1999.

[3] Jianhua Lin. Divergence measures based on the shannon entropy. IEEE Trans. Infor. Theory, 37:145-151, 1991.