*Advanced Engineering Forum* *Online: 2012-09-26*
*ISSN: 2234-991X, Vols. 6-7, pp 576-582*
*doi:10.4028/www.scientific.net/AEF.6-7.576*
*© 2012 The Author(s). Published by Trans Tech Publications Ltd, Switzerland.*

# Session Segmentation Method Based on Naïve Bayes Model

Li Ping[1,a], Cui Ming-Liang[1,b], Hou Zhen-Shan[1,c], Wei Liu-Liu[1,d], Ying Wen-Hao[1,e], and Zuo Wan-Li[1,f,*1]

[1]College of Computer Science and Technology, Jilin University, Changchun 130012, China

[a]liping2110@mails.jlu.edu.cn, [b]cuiml2110@mails.jlu.edu.cn, [c]houzs2110@mails.jlu.edu.cn, [d]weill2110@mails.jlu.edu.cn, [e]yingwh2110@mails.jlu.edu.cn, [f]zuowl@jlu.edu.cn

[1] Corresponding author

**Keywords:** Naïve Bayes Model; query log; session segmentation.

**Abstract.** Session segmentation can not only contribute a lot to the further and deeper analysis of user's search behavior but also act as the foundation of other retrieval process researches based on users' complicated search behaviors. This paper proposes a session boundary discrimination model utilizing time interval and query likelihood on the basis of Naive Bayes Model. Compared with previous study, the model proposed in this paper shows a prominent improvement through experiment in three aspects, which is: recall ratio, precision ratio and value F. Owing to its advantage in session boundary discrimination, the application of the model can serve as a tool in fields like personalized information retrieval, query suggestion, search activity analysis and other fields which is related to search results improvement.

## Introduction

To make the search engines provide users with personalized service, we need to comprehend users' searching habits, to find out user's interest. Using user's query words is an important way to understand its interest, but they are complicated. If only query words classification was fully understood, it would do a lot to interest mining. Classification for queries requires a criterion; so far the relatively recognized criterion has been session segmentation. Previous researches have different definitions about the meaning of session. We much more agree with Zhang Lei and accept that session is a group of sequences of activities related to each other not only through an evolving information need at a deeper, conceptual level but also through close proximity in time. We intend to group these activities together and refer them as a session. Previous documents [1,2,3,4,5,6,7,8] all utilize attribute time interval (TI) to achieve session segmentation. Other attributes like query likelihood (QL) and anchor likelihood (AL) are also used to divide session. Generally speaking, session segmentation methods based on attributes is broadly divided into two categories: time interval method and multiple attribute method.

**Time interval session segmentation method**. Time interval is recognized as a significant attribute in session segmentation [1,2,3,5,6,7]. Daqing He described and discussed the research based on two web logs: Excite and Reuter with a view to divide sessions, they utilized only one attribute time interval (TI), and tried different values for TI to find out how the quantity of queries in one session change when TI differs. And the conclusion they have derived is that when TI is between 10 and 15 minutes the query amount in one session tend to be stable. It's not unreasonable to conclude that 10 and 15 minutes is the critical interval with regard to session segmentation.

**Multiple attributes session segmentation method**. Documents [3,4,5,6] were written by same group of authors and gave the same definition for session. These papers adopt the same method to deal with query log and define that all the interactive information between the user and his search engine is a session. At last, they identify different topics which are contextual irrelevant, this work is called Topic Shift Identification. The emphasis of their work is how to divide sessions with two attributes in query log. The method they used is to discrete time interval into 7 ranks: divide 0 to 30 minutes into 6 parts averagely; and the bracket equal to or more than 30n minutes constitutes the seventh rank. The relationship between current query (Qc) and the previous query (Qp) is referred

This article is an open access article under the terms and conditions of the Creative Commons Attribution (CC BY) license
(https://creativecommons.org/licenses/by/4.0)

to as search pattern (SP). With regard to SP, it is also divided into 7 ranks averagely based on their likelihood in semantic relationship. 7 kinds of time intervals and 7 kinds of search patterns make up 49 combinations; any successive queries belong to one of the combinations. To make topic identification on such a data set, they have artificial neural network [4], Multiple Linear Regression [5] and Monte Carlo simulation [6]. Their common feature is that they have a better performance in continuation identification than shift identification. The three methods respectively in document [4,5,6] have serious defects in topic shift identification and as they have always stressed. There are two reasons accounting for that: (1) Time interval correctly utilized, 30minutes act as the maximum time interval without sufficient evidence; (2) Search pattern used are not proper, for granularity is too large.

This paper proposes to adopt Bayes Model into session segmentation and is organized as follow: Segregation method based on Naive Bayes Model is presented (NBM-SBDM) in Section 2. Section 3 shows the experiment schemes, details and the results of the method. At last, we will conclude the method in the paper and propose directions for further research in Section 4.

## Session Boundary Discrimination Model (NBM-SBDM)

The center part of session segmentation is to discriminate whether each query is a session boundary, if it is, called *yes* class; otherwise, it is labeled *no* class. The session is correctly divided only when each query in it is exactly right discriminated. To realize it, we adopt Naïve Bayes Model.

**Data Set**. The query log from search engine is a successive text file, each line compresses *Access Time, User ID, Query Word, Hit URL,* and *URL ID*. Before formalizing the query log, we first give an explicit description for units in it.

Basic Data Type: *Time*, *Number* (0-9), *Word* (natural language word), *Punctuation*, *Alphabet*, *Symbol* (URL Symbol)

Unit of Query Log: *Access Time*:: =T (Time); *User ID*:: = (N(Number))*; *Query Word*::=(W (Word) +P (Punctuation))*; *Hit URL* ::=( A (Alphabet) +S (Symbol))*; *URLID* ::= (N (Number))*

Description of Query Log: *Query Record*::= Access Time× User ID ×Query Word ×HitURL ×URLID; *Query Log* ::=Record*

Definition 1. *Session*: It is a subset of successive sequence of retrieval behavior taking place in the course of retrieval process with one particular goal, its vector representation is $Session=(R_1,R_2,...,R_n)$, in which $R_1$ represent a query record.

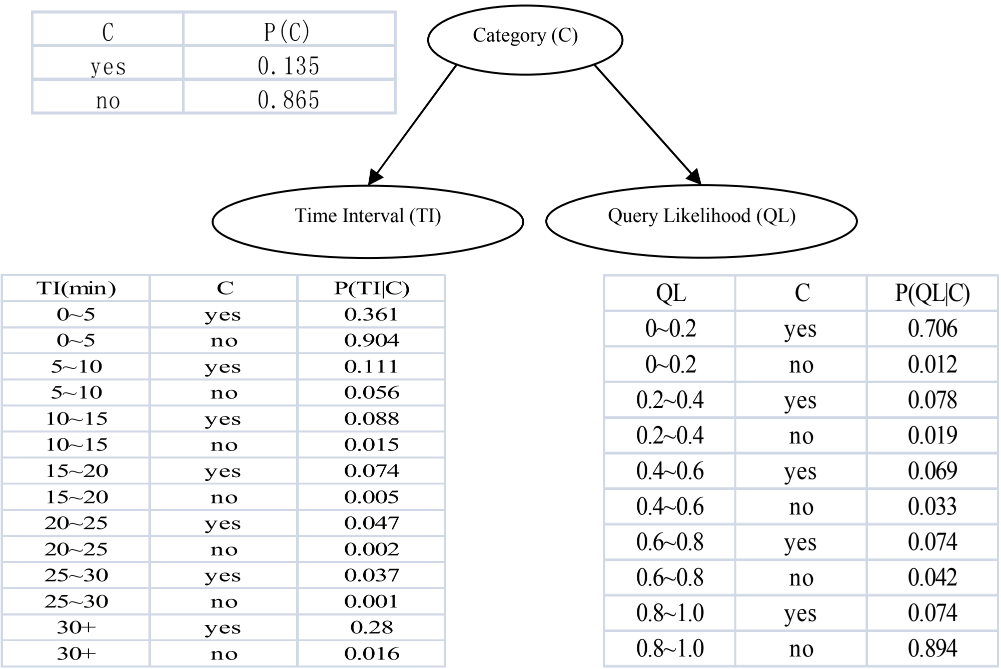
Definition 2. *Query Time Interval* (TI): TI is the time span between current query and the previous query, denoted by:

$$TI = T_{Qc} - T_{Qp} \tag{1}$$

Definition 3. *Query Likelihood* (QL): QL is an attribute to quantize the semantic likelihood between current query and previous query, denoted by:

$$QL = \frac{\sum_{t \in Qc \cap t \in Qp} \# t_{Qc} \times \# t_{Qp}}{\sqrt{|Q_c| \times |Q_p|}} \tag{2}$$

Where, #t is a variable to define how many times t term appears, $Q_c$ represents current query and $Q_p$ represents previous query.



| C | P(C) |
|---|---|
| yes | 0.135 |
| no | 0.865 |

| TI(min) | C | P(TI\|C) |
|---|---|---|
| 0~5 | yes | 0.361 |
| 0~5 | no | 0.904 |
| 5~10 | yes | 0.111 |
| 5~10 | no | 0.056 |
| 10~15 | yes | 0.088 |
| 10~15 | no | 0.015 |
| 15~20 | yes | 0.074 |
| 15~20 | no | 0.005 |
| 20~25 | yes | 0.047 |
| 20~25 | no | 0.002 |
| 25~30 | yes | 0.037 |
| 25~30 | no | 0.001 |
| 30+ | yes | 0.28 |
| 30+ | no | 0.016 |

| QL | C | P(QL\|C) |
|---|---|---|
| 0~0.2 | yes | 0.706 |
| 0~0.2 | no | 0.012 |
| 0.2~0.4 | yes | 0.078 |
| 0.2~0.4 | no | 0.019 |
| 0.4~0.6 | yes | 0.069 |
| 0.4~0.6 | no | 0.033 |
| 0.6~0.8 | yes | 0.074 |
| 0.6~0.8 | no | 0.042 |
| 0.8~1.0 | yes | 0.074 |
| 0.8~1.0 | no | 0.894 |

Fig. 1 NBM-SBDM

**Session Boundary Discrimination Model**. Document [8] finds out that the performance of session segmentation when only time interval and query likelihood are taken into account is no worse than the result when time interval, query likelihood and anchor likelihood three attributes are all considered. Thus, to make the experiment process concise, we might as well select time interval and query likelihood to do session segmentation job and assume attribute TI and QL are conditionally independent over class variable C. With this assumption, the session boundary discrimination model based on Naïve Bayes Model is as shown in Fig.1. The prior probability is acquired by manually labeling part of Sogou query log. The statistical data is presented in Table 1.

Table 1 Manual Statistical Data

| attributes | quantity |
|---|---|
| query | 3102 |
| Session | 1203 |
| Session boundary | 1203 |
| None session boundary | 1899 |
| User | 911 |

Hereinto, prior probability P (C=c), P (TI|C), P (QL|C) is computed by the formulas:

$$P(C=c)=\frac{N(c)}{the\ quantity\ of\ queries} \tag{3}$$

$$N(c)=\begin{cases} the\ quantity\ of\ session\ boundary & c=yes \\ the\ quantity\ of\ non\ session\ boundary & c=no \end{cases} \tag{4}$$

$$P(TI=(ti_1,ti_2)\,|\,C=c)=\frac{the\ quantity\ of\ Qc\big(ti\in(ti_1,ti_2)\cap Qc\ is\ c\ kind\ session\big)}{N(c)} \tag{5}$$

$$P(QL=(ql_1,ql_2)\,|\,C=c)=\frac{the\ quantity\ of\ Qc\big(ql\in(ql_1,ql_2)\cap Qc\ is\ c\ kind\ session\big)}{N(c)} \tag{6}$$

Here ($ti_1$,$ti_2$) is a bracket for TI in Fig.1, and ($ql_1$,$ql_2$) is a bracket for QL. Variable c is used to label whether Qc is a session.

In Fig.1, given C, variable TI and QL are conditional independent; C=yes represents that current query is a session boundary (the beginning of a session), C=no means current query is not a session boundary. And session boundary model is expressed in this way:

$$Border\left(ti,ql\right)=\arg\max_{c} P\left(C=c\mid TI=ti,QL=ql\right) \tag{7}$$

Abstractly, SBDM is a conditional probability model P (C|TI, QL); class variable C is dependent on two attributes TI and QL. The problem is that to present probability table based on the model is too complicated, we therefore reformulate the model to make it more tractable:

$$P\left(C\mid TI,QL\right)=\frac{P\left(C\right)P\left(TI,QL\mid C\right)}{P\left(TI,QL\right)} \tag{8}$$

In practice we are only interested in the numerator of that fraction, since the denominator does not depend on *C* and the values of the features *TI and QL* are given, so that the denominator is effectively constant. The numerator is equivalent to the joint probability model P(C, TI, QL) which can be rewritten as follows:

$$P\left(C,TI,QL\right)=P(C)P(TI\mid C)P(QL\mid C,TI) \tag{9}$$

Now the "naive" conditional independence assumptions come into play; it means that the joint model can be expressed as:

$$P\left(QL\mid C,TI\right)=P\left(QL\mid C\right) \tag{10}$$

Under the above independence assumptions, the conditional distribution over the class variable can be expressed like this:

$$P\left(C,TI,QL\right)=P\left(C\right)P\left(TI\mid C\right)P\left(QL\mid C\right) \tag{11}$$

The discussion so far has derived the simplified model Border (ti, ql):

$$Border\left(ti,ql\right)=\arg\max_{c} P\left(C=c\right)P\left(TI=ti\mid C=c\right)P\left(QL=ql\mid C=c\right) \tag{12}$$

**NBM-SBDM Algorithms**: (1) *Data Preparation.* Queries in original data are sorted by time. There are lots of users who are engaged in information retrieval even at the same second. Successive queries of one user need to be extracted. In order to discriminate session of different users' queries, we tidy the data by ID sorting to make the queries successive for one user. (2) *Session Boundary Discrimination*. Compute attribute TI and QL and discriminate session boundary with the NBM_SBDM proposed in the paper.

*SBDM algorithm*:

```
Input:    Query Log
Output: Session sequence
list<search>User;        //initialize query list
while (i!=File1.eof())
    (id (i), time (i), query (i)) <-(File1);      //read query log and store the data
    i++;
Endwhile
User. Sort (comp_id); //sort query log by user ID
ICTCLAS_Init();       // initialize word segmentation module
while (i!=User. begin())
    TI (i)=Time Interval(time(i))-Time Interval(time(i-1));      // time interval
    Segmentation (query (i-1)); Segmentation (query (i));         //word segmentation
    QL (i)=Similarity(query(i-1),query(i));     // the likelihood of query words
    i--;
```

```
Endwhile
ICTCLAS_Exit();          //exit the module
Discretize(TI); Discretize(QL);     //discrete time interval and query likelihood
while (j!=File2.eof())
    P<-(File2);   //read prior probability table
while (i! =User. end ())
    if(ComputeC_Yes_TQ(i,P)>ComputeC_No_TQ(i,P)) // posterior probability
        session boundary<-query (i);
    i++;
Endwhile
Return
```

## Experiment

**Evaluation criteria of the session segmentation**. Document [8] uses both evaluation criteria, that is, the query boundary evaluation criterion and the session for the evaluation criterion. In order to facilitate analysis and discussion, we use common evaluation framework [9,10], namely, the use of standard criterion to check boundaries. In order to carry out evaluation better, we may assume that the two situations: (1) *Situation A:* Session boundaries have not been identified by the discrimination model; (2) *Situation B:* Non-session boundaries have been regarded as session boundaries by the discrimination model. These performance criterions are used to demonstrate the performance of SBDM for discriminating session boundaries, and their formulations are as follows:

$$P_{yes}=\frac{N_{boundary\&correct}}{N_{boundary}} \tag{13}$$

$$R_{yes}=\frac{N_{boundary\&correct}}{N_{trueboundary}} \tag{14}$$

$$T_{yes}=\frac{2\times P\times R}{P+R} \tag{15}$$

Where, $N_{boundary}$ means number of queries labeled as session boundaries by SBDM; $N_{trueboundary}$ represents number of queries labeled session boundaries by manual annotation; $N_{boundary\ \&\ correct}$ means number of queries labeled session boundaries by SBDM and manual annotation. $P_{no}$, $R_{no}$ and $T_{no}$ have similar definitions.

**Experimental results**: We discriminate the session boundaries and non-session boundaries in training set manually. In Table 2, we will display the comparison of the result between Decision Tree and SBDM Algorithm.

Table 2 Result Comparison

| Evaluation Criterion | Type | Decision Tree | SBDM |
|---|---|---|---|
| P | yes | 0.974 | 0.952 |
|  | no | 0.937 | 0.952 |
| R | yes | 0.841 | 0.919 |
|  | no | 0.991 | 0.971 |
| F | yes | 0.903 | 0.935 |
|  | no | 0.963 | 0.961 |

The experiment results show that P, R, F are all greater than 90 percent for both session discrimination and non-session discrimination. Allowing for that session and non-session are inverse, the performance of the proposed model in precision ratio, recall ratio and value F is great.

The value F in SBDM Algorithm is greater than that in Decision Tree for session discrimination as Table 2 shows. Compared with the results achieved by Decision Tree [8] which adopts the same evaluation criterion as ours, the method in paper improved value F for class *yes* and kept a high value F for *no* class simultaneously. By horizontal comparison, the method achieves an improvement in session boundary discrimination. And session discrimination is much more significant than non-session discrimination when it comes to session segmentation. Because session identification errors will lead to queries with different intentions separated into one session, it will bring about many noises to session analysis.

## Summary

This paper outlined current status of session boundary discrimination, and gives a description for the significance of session segmentation to users' retrieval behaviors in personalized search engine field and proposes a new approach based on Naïve Bayes Model to discriminate session boundaries. This idea is initiative in the field. Experimental result shows that this model performs well in session boundary discrimination work. In fact, this target is session segmentation job has always been striving for.

Despite the good job done by the model, there is still some advancement that can be achieved. As the model proposed in this paper is quite dependent on prior probability like P(C), P(TI|C), P(QL|C) and the prior probability rely a heavily on the sample manually labeled from training set, there are more or less discrimination errors in testing test set. It is reasonable to consider that more volume of training set will make for a more stable prior probability. However, oversize sample can be intractable for manually labeling. With respect to this, we will shift our attention after this paper and do studies about how sample size of training set can make a difference in the performance of session discrimination.

## Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant No.60973040; the National University Student Innovation Program.

## References

[1] Craig Silverstein, Monika Henzinger, Hannes Marais, et al. Analysis of a very large Web search engine query log [J]. In SIGIR Forum, fall 1998, 33(1): 6- 12.

[2] Daqing He, Ayse Goker. Detecting session boundaries from Web user logs[C].Proceedings of the 22nd annual colloquium on information, 2000.pp:57-66.

[3] H. Cenk Ozmutlu, Fatih cavdur. Application of automatic topic identification on excites web search engine data logs [J].Information Processing and Management: an International Journal, 2005, 41(5):1243-1262.

[4] Seda Ozmutlu, Fatih Cavdur. Neural network applications for automatic new topic identification [J]. Online Information Review, 2005, 29(1): 34-53.

[5] Seda Ozmutlu, H. Cenk Ozmutlu, Amanda Spink. Automatic New Topic Identification in Search Engine Transaction Logs using Multiple Linear Regression[C].Proceedings of the 41st Hawaii International Conference on System Sciences. 2008. pp: 140-140.

[6] Seda Ozmutlu, Huseyin C. Ozmutlu, Buket Buyuk.Using Monte-Carlo Simulation for Automatic New Topic Identification of Search Engine Transaction Logs[C].Proceedings of the 2007 Winter Simulation Conference.2007.pp: 2306-2314.

[7] Huijia Yu, Yiqun Liu, Min Zhang, Liyun Ru, Shaoping Ma. Research in Search Engine User Behavior Based on Log Analysis [J]. Journal of Chinese Information Processing, Vol.2007, 21(1):109-114.

[8] ZHANG Lei, LI Yanan, WANG Bin, LI Peng, JIANG Zaifan. Session Segmentation Based on Query Logs of Web Search [J].Journal of Chinese Information Processing, 2009, 23(2):54-61.

[9] Xiangji Huang, Fuchun Peng, Aijun An, DaleSchuurmans. Dynamic Web Log Session Identificationwith Statistical Language Models [J]. Journal of the American Society for Information Science and Technology, 55 (14): 129021303.

[10] FANG Qi, LIU Yiqun, ZHANG Min, RUN Liyun, MA Shaoping Swarm Intelligence Based Topic Identification for Sessions in Web Access Log [J].Journal of Chinese Information Processing. Vol.2011, 25(1):35-40.