# Collocation Extraction Using Web Statistics

## Hsin-Hsi Chen, Yi-Cheng Yu, and Chih-Long Lin

Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
E-mail: hh_chen@csie.ntu.edu.tw

## Abstract

This paper mines collocations from two different web usage corpora, NTU proxy log and TTS search log. The precisions for NTU and TTS test data are 61.76% and 57.50%, respectively, by human judgment for 2% sampling of extracted collocations. For automatic evaluation, we submit extracted collocation to Google search engine, and the resulting page counts are used to compute the mutual information of the collocation. Experimental results show that total 43.27% and 42.65% of collocations mined from NTU and TTS corpora passed the examination of MIs.

## 1. Introduction

Web, which contains heterogeneous live data, is a natural resource for human language technologies. Several types of logs are kept on the Web. The famous examples are proxy logs in proxy severs, search logs in search engines, browsing logs in content providers, and so on. The logs can be employed to save Internet bandwidth, to decrease communication bottlenecks, to speed up the processing speed, to track users' behaviors, to improve service quality, *etc* (Cui, Wen, Nie and Ma, 2002; Hansen and Shriver, 2001;.Huang, Oyang and Chien, 2001; Srivastava, Cooley, Deshpande, and Tan, 2000; Zuckerman, Albrecht, and Nicholson, 1999) Logs express the information needs of users. They reflect common behaviors of a specific group or an individual user. They may be long-term or short-term behaviors. Several approaches have been proposed to mine users' behaviors from logs.

Search logs of search engine portal are different from proxy logs of proxy server in that the former is fan-in and the latter is fan-out. In other words, requests are merged into the search portal from different sites. On the other hand, requests are sent to different sites through proxy server. The logs on proxy are records of users clicking through search results, through browsing, or through new URLs. Users may open several windows at the same time to express different information needs. In other words, the records of different information needs may be mixed together. Thus, to tell out the sessions of users' information needs is the first step in log mining, and mining from proxy logs is more challenging than that from search logs.

This paper utilizes web log, and page counts of web search to extracting live word collocations which are embedded implicitly in search queries. Two log corpora – say, proxy log of National Taiwan University Computer and Information Networking Center (abbreviated as NTU corpus) and search log of TTS Group (abbreviated as TTS corpus), are adopted in collocation extraction. Page counts of Google search are used for verification.

NTU proxy log corpus of size 160GB was collected from February 22, 2002 to May 18, 2002. It contained various kinds of files, including web pages, images, audio, CGI, ASP, Java, *etc*. TTS search log corpus, which records queries on news and journal databases from January 20, 2002 to May 23, 2002, is comparatively much smaller (4.19MB).

This paper is organized as follows. Section 2 presents the collocation mining algorithm. Section3 shows the experiments using both NTU proxy log corpus and TTS search log corpus. Section 4 proposes an automatic collocation verification method using page counts of search engines. Section 5 concludes the remarks.

## 2. Collocation Mining with Web Log

A URL may be clicked through search result of a query, clicked through browsing, or clicked with a new URL that may be irrelevant or relevant to a query. Because the behaviors may be mixed together by multiple browsing windows, we cannot disambiguate the relationship trivially from proxy log. We postulate that a session is a series of queries by a single user made within a small range of time, and propose a threshold to define sessions. The following sections show an algorithm to mine collocations with web log. We discuss proxy log at first, and then extend the algorithm to deal with search log.

### 2.1 Step 1: Preprocessing

Query terms in log denote the objects that users request. There exists an URL field in each entry of proxy log. It records either search or browsing. To tell the function of each entry is the first step. We examine some famous search engines like *Google*, *Yahoo*, *Openfind*, and so on, and find that the URL of a search query has some fixed formats. An example of searching "ASP" with *Google* and *Yahoo* is shown as follows

  http://www.google.com/search?q=ASP&ie=UTF-8&
      oe=UTF-8&hl=zh-TW&lr=
  http://tw.search.yahoo.com/search/kimo?p=ASP&a=b

We formulate rules to match each entry of the proxy log and select the possible search queries. In this approach, some queries that do not meet the rules may be missed.

Then we will extract the search terms in the queries, which denote the objects that users request.. The Chinese search terms may be in terms of either Big5 code

or UTF-8 code.   The two examples shown as follows are in Big5 and UTF-8 codes, respectively.   Search terms are underlined, and codes are marked.

http://www.google.com/search?q=%A8%CA%B3%A3 %A6a%B9%CF& hl=zh-TW & lr=

http://www.google.com/search?q=%E7%A1%AC%E9 %AB%94+%E8%80%81%E4%BC%AF&source id=opera&num=0& ie=utf-8 & oe=utf-8

We decode the search terms according to the codes adopted.

## 2.2  Step 2: Determination of a Session

A session is meant to capture a single user's attempt to fill a single information need (Silverstein *et al.*, 1998). In proxy log, an IP just tells us an individual user instead of its concrete information need.   Figure 1 sketches a threshold approach to determine a session, where $t_{i+1}-t_i >$ threshold.   We will group the query terms in the same session.
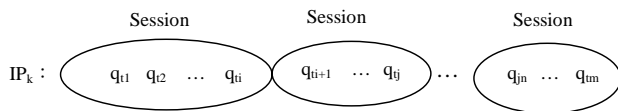


**Figure 1. Determination of a Session**

Now the problem is the setting of the threshold.   We sample the data from NTU proxy log corpus.   Table 1 shows the statistics.

**Table 1. The Statistics of Sampling**

| Total days | 20 |
|---|---|
| Total number of IPs | 30,151 |
| Total number of Queries | 190,453 |

Figure 2 shows the threshold setting and the number of sessions.   The number of sessions increases very quickly, when threshold increases.   It reaches to a stable state after threshold is set to 500 seconds.   Thus, we choose 500 seconds as our threshold.   Table 2 shows the statistics of NTU proxy log corpus if 500 seconds are considered as a boundary of an information need.
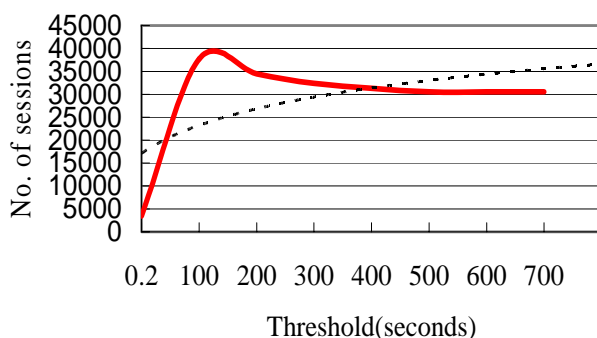


**Figure 2. Threshold Setting**

**Table 1. Statistics of Query Log**

| Number of sessions | 51,125 |
|---|---|
| Total queries | 190,453 |
| Average queries per session | 3.73 |
| Number of unique queries | 81,707 |
| Average unique queries per session | 1.6 |

## 2.3  Step 3: Collocation Extraction

Query terms within a session are assumed to be correlated.   To get reliable statistics, we only consider those sessions that consist of more than one query.   Besides, frequency of queries is also important.   We only keep queries of more than 5 occurrences

After filtering, an association matrix $M=(f_{ij})$ with *number-of-distinct-query-terms* rows and *number-of-sessions* columns is defined.   Symbol $f_{ij}$ denotes the frequency of a query term $t_i$ in a session $x_j$. Multiplication of M and the transpose of M is a term-term association matrix T.   The value $c_{u,v}$ in T denotes a correlation between terms $t_u$ and $t_v$.   A scalar association matrix $S=(s_{u,v})$ is further computed from T by using Cosine of correlation vectors $c_u$ and $c_v$.   The terms v of larger scalar correlation values $s_{u,v}$ with term u are called *collocates* of u.

## 2.4   Extension to Search Log

The above method is also applicable to search log except that the determination of a session is much easier. An entry in TTS search log corpus includes the database that users consult, IP, login time, action time, and search terms.   Login time denotes when a user logins with a specific IP, and action time specifies when the specific queries are submitted.   User's IP, login time and action time are useful cues to find a session.   The following shows the session partition strategy.

(1)   If IPs of two continuous search queries are different, these two queries are partitioned into two different sessions.

(2)   If two continuous search queries come from the same IP, we consider the conditions as follows further.

(a)   If their login time is the same and the difference of the two action times is larger than $threshold_1$ then these two queries are postulated to belong to two different sessions.

(b)   If login times of these two queries are different and the action time of the first query minus the login time of the second query larger than $threshold_2$, then they are partitioned to two sessions.

(c)   Otherwise, the two queries are assumed to be in the same session.

To determine the two thresholds, i.e., $threshold_1$ and $threshold_2$, we examine the relationship between login time and number of sessions (Figure 3), and the relationship between action time and number of sessions (Figure 4).   The thresholds, 600 and 720 seconds, are selected, respectively.   Table 2 summarizes the statistics of TTS search log corpus after this setting.
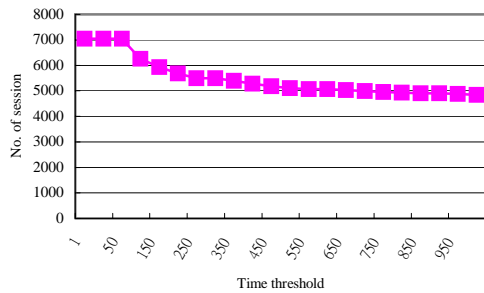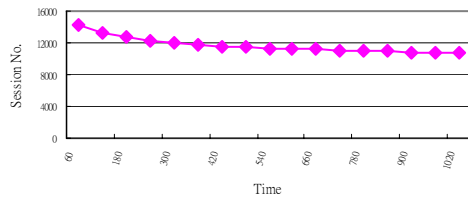
**Figure 3. Login Time and Number of Sessions**



**Figure 4. Action and Number of Sessions**

**Table 2. Statistics of Search Log**

| | |
|---|---|
| Number of sessions | 6,018 |
| Total queries | 34,426 |
| Average queries per session | 5.72 |
| Number of unique queries | 20,048 |
| Average unique queries per session | 3.53 |

## 3. Experiments

Some interesting collocations shown below are extracted from NTU log corpus. Pair {w1, w2} denotes a collocation. Here, each Chinese word is in both Han character and Pinyin.

(1) English-Chinese bilingual terms, e.g., {cocaine, "古柯鹼" ("gu ke jian")}, {ebook, "電子書" ("dian zi shu")}, {E.+coli, "大腸桿菌" ("da chang gan jun")}, {lottery, "樂透彩" ("le tou cai")}, *etc*.

(2) English-Chinese companies/products, e.g., {HP, "惠普" ("hui pu")}, {norton, "賽門鐵克" ("sai men tie ke")}, {pioneer, "先鋒" ("xian feng")}, *etc*.

(3) Abbreviation, e.g., {"行政院青輔會" ("xing zheng yuan qing fu hui", national youth commission), "青輔會" ("qing fu hui", NYC)}, *etc*.

(4) Relationships, e.g., {"大世紀" ("da shi ji", Theater), "二輪電影" ("er lun dian ying", second-run film)}, {"考題" ("kao ti", examination questions), "補習班" ("bu xi ban", cram schools)}, {"百日咳" ("bai ri ke", pertussis), "流行病學" ("liu xing bing xue", epidemiology)}, *etc*.

Because requests are sent to Internet through proxy server, the above collocations cover various kinds of subjects, and may be in bilingual form. In contrast, requests submitted to TTS are operated on special databases, so that collocations extracted from TTS corpus

are different from those from NTU corpus. The following show that some terms appearing in both corpora have different collocations.

(5) NTU: {MP3, "軟體下載" ("ruan ti xia zai", software download)},
TTS: {MP3, "成大MP3" (National Cheng Kung University MP3 event)}

(6) NTU: {"馬來西亞" ("ma lai xi ya", Malaysia), "大鵬旅行社" ("da peng lu xing she", travel agency)}, …
TTS: {"馬來西亞" ("ma lai xi ya", Malaysia), "馬哈迪" ("ma ha di", Mahathir Mohamad)}, …

The first example shows that the collocation from NTU corpus denotes where to download MP3, however, the collocation from TTS corpus denotes an event concerning MP3. The second example shows that the collocates of "Malaysia" mined from NTU corpus are "travel agency", because users' foci are "traveling in Malaysia". In contrast, the collocations extracted from TTS corpus are Malaysia Premier, *etc*. They focus on political and economical affairs. These examples demonstrate that collocations have strong relationship with source of training materials.

## 4. Collocation Verification Using Page Counts

Mutual information (MI) measures how a word $w_1$ is correlated to another word $w_2$. We further employ MI to verify the collocation $\{w_1, w_2\}$ extracted from log corpora. Page counts of $w_1$, $w_2$ and $(w_1, w_2)$ by using Google simulate their frequencies.

Total 6,446 and 5,301 collocations are derived from NTU and TTS corpora, respectively. Neglecting the cases with zero page counts, 5,320 and 3,605 collocations remain, respectively. Figures 5 and 6 show MI distribution for NTU and TTS test data, respectively. Figure 5 is much like a normal distribution. The smallest MI is -8 and the largest is 21. The largest collocation group is of MI 8.0-8.9. It happens similarly in Figure 6. The smallest and the largest MIs are -23 and 23. The largest group locates within 0-.0.9.
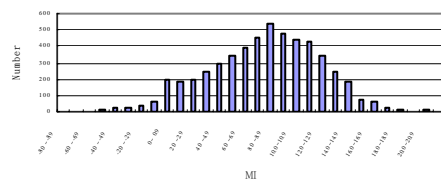


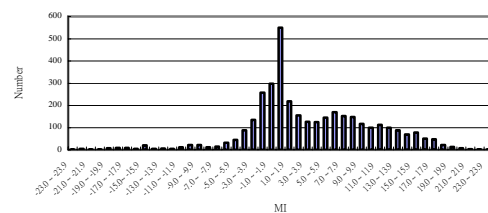**Figure 5. MI Distribution of NTU Test Data**



**Figure 6. MI Distribution of TTS Test Data**

To determine the thresholds, we select 2% of pairs of MIs and ask 9 persons to assess the results. The pairs that more than 5 assessors say yes are considered as correct. The precisions are 61.76% and 57.50% by human judgment for NTU and TTS test data.

Then we compute the average MIs of these correct pairs and regard it as a threshold. Finally, 8.38 and 1.76 are selected for NTU and TTS test data. Total 43.27% and 42.65% of the collocations extracted from NTU and TTS corpora passed the examination of MIs. Page count is a rough statistics. Some collocation is relevant, but its MI is low using page count. The famous example is "佛教" (Buddhism) and "天主教" (Catholicism). They may not appear in the same web page quite often.

## 5.　Concluding Remarks

In this paper, web serves as both training and evaluation corpora for human language technologies. Two different kinds of log corpora are employed to mine collocations. The experimental results show that different collocations are extracted for the same query term. That reflects characteristics of live log corpora. The precisions of 61.76% and 57.50% for NTU and TTS test data are achieved respectively under the manual evaluation.

Because Chinese has segmentation problem and the sentences in web pages are not segmented, a web page in which a search term is matched successfully does not always contain it. For example, a web page of theme "澳門聯網" ("ao men lian wang", MacaoLink) is reported for a query "門聯" ("men lian", gatepost couplet). In this way, the page counts do not exactly reflect the frequency of a term. That also affects the automatic evaluation using page counts. How to deal with this problem is indispensable when web statistics is employed in Chinese language processing.

## References

Cui, Hang, Wen, Ji-Rong, Nie, Jian-Yun and Ma, Wei-Ying (2002) "Probabilistic Query Expansion Using Query Logs," *Proceedings of the Eleventh International Conference on World Wide Web*, 2002, pp. 325-332.

Hansen, Mark and Shriver, Elizabeth (2001) "Using navigation Data to Improve IR functions in the Context of Web Search," *ACM CIKM*, 2001, pp. 135-142.

Huang, Chien-Kang, Oyang, Yen-Jen, and Chien, Lee-Feng (2001) "A Contextual Term Suggestion Mechanism for Interactive Web Search," *Proceedings of the First Web Intelligence Conference*, pp. 272-281.

Silverstein, C. *et al*. (1998) "Analysis of a Very Large Alta Vista Query Log," Technical Report 1998-014, Digital Systems Research Center.

Srivastava, Jaideep, Cooley, Robert Deshpande, Mukund and Tan, Pang-Ning (2000) "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," *SIGKDD Explorations*, Vol. 1, No. 2, pp. 12-23.

Zuckerman, I., Albrecht, D., and Nicholson, A. (1999) "Predicting User's Requests on the WWW," *Proceedings of the Seventh International Conference on User Modeling*, pp. 275-284.