Query Log Analysis

Wensi Xi, Abdur Chowdhury, Kush Sidhu and Greg Pass

American Online, Inc.
xwensi@vt.edu , {Cabdur, Ksidhu35, GregPass1,}@aol.com

Abstract. The rapidly growing World Wide Web provides an enormous amount of information for Internet users all across the world; however, not all of this information is equally important. Determining the most preferred topics on the Internet is of great interest to commercial information providers. This paper provides a method to compare and analyze user query logs, and answers the questions: do users' queries change over time, to what extent do they change, and finally, is it possible to predict future users' queries by examining past users' queries.

1. Introduction

The topics on the net that Internet users prefer are of great interest to commercial web information providers (e.g. search engine companies). Users' information needs and preferences can be determined by examining user query logs. This paper makes use of some statistical methods to measure the similarity of various query logs and answers questions such as: how different is the user's information need over time, and can future users' information needs be determined by examining past query logs?

Before going into our research method, we will introduce a few terminologies.

1.1 Log Files

This research is based on the set of log files collected from American Online, Inc.'s AOL Search service (http://www.aol.com), from May, 2001 to April, 2002. The log file is a ranked listing of unique queries ordered by the total number of searches for that query for each month.

Table 1. A Sample of log file

Query	Number of Searches
yahoo	2053960
google	1064029
ebay	764470
hotmail	745435
test	718311
….	…….

Some queries appear more often than others; these are the queries of most interest to the information provider. Taking the number of times a unique query occurs divided by the total number of all queries in a given month produces the percentage that this query contributes to total usage for that month.

Below is a cumulate curve for the query logs under study from March 2002.

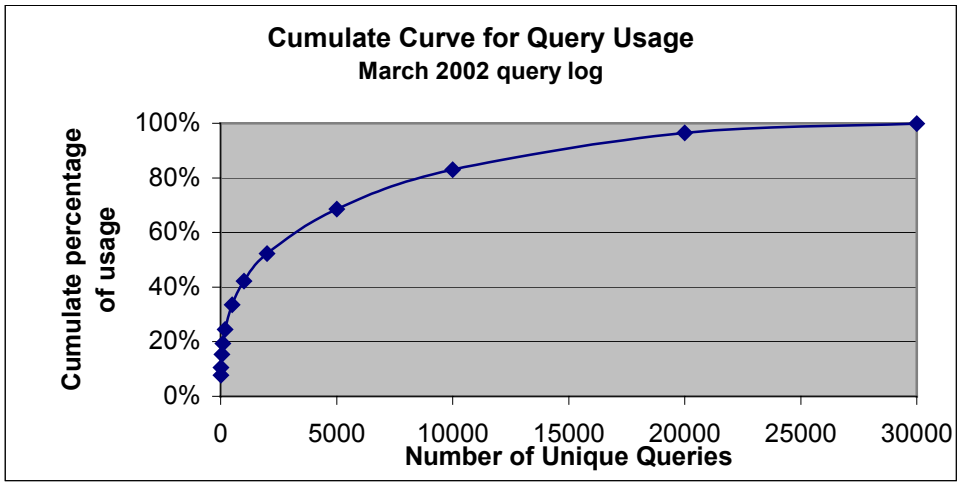

Fig. 1. Cumulate Curve for Query Usage

The curve is fairly skewed, indicating that a small number of queries (approx. 10,000) contribute a great amount to the total usage of the service (over 80%). Theoretically, if this small percentage of queries can be predicted and the search results improved, the whole system can be improved for 80% of the service provided.

1.2 Overlap Rate

The easiest way to measure the similarity between query logs for different months is to measure the *overlap rate*. *Overlap Rate* is introduced in [1]. Suppose we have a series of log files *L1, L2, ..., Ln*, the *Overlap Rate* is defined as:

$$R = \frac{L_1 \cap L_2 . \quad . \quad . \cap L_n}{L_1 \cup L_2 . \quad . \quad . \cup L_n}$$

It can be rendered as the total number of common queries vs. the total number of unique queries.

1.3 Correlation Coefficient

In order to measure the stability of the ranks of queries in logs, we adopt the *Correlation Coefficient* method []. The *Correlation Coefficient* summarizes the direction and magnitude of association between two variables. It is a widely used and cited statistical method. There are other methods to measure the similarity of two ranked lists (e.g. *Spearman* Rank Correlation and *Kendoll's Tau* Correlation). The reasons we chose *Correlation Coefficient* are:

- The query logs do not contain exactly the same set of queries; the *Spearman* and *Kendoll's Tau* rank correlations both require the set being compared to contain the same elements.
- The *Overlap Rate* of two query logs does not affect the *Coefficient Correlation*. However, it may affect *Spearman* and *Kendoll's Tau* score [].

2. Analysis Methods

The *Overlap rate* and *Correlation Coefficient* were used to analyze the query logs from May 2001 to April 2002. These two measures were taken using the top 1,000 queries from each month. Additionally, an *overlap rate* curve was generated for the whole year for the top 10, 100, 1,000, 10,000, 20,000 and 30,000 queries from each month.

3. Results

Below are the monthly, quarterly, and semi-yearly *Overlap* and *Correlation* scores across the full year for the top 1,000 queries:

Table 2. Correlation and Overlap results for top 1000 queries

Month-to-month	**Correlation Coefficient**	**Overlap Rate**
May 01 – Jun 01	0.90	0.87
Jun. 01 – Jul. 01	0.94	0.92
Jul. 01 – Aug. 01	0.93	0.91
Aug. 01 – Sep. 01	0.89	0.85
Sep. 01 – Oct. 01	0.91	0.88
Oct. 01 – Nov. 01	0.91	0.88
Nov. 01 – Dec. 01	0.91	0.88
Dec. 01 – Jan. 02	0.85	0.84
Jan. 02 – Feb. 02	0.91	0.90
Feb. 02 - Mar. 02	0.93	0.90
Mar. 02 –Apr. 02	0.95	0.92
Quarterly	**Correlation Coefficient**	**Overlap Rate**
May.01 –Aug. 01	0.83	0.82
Sep. 01 – Dec.01	0.80	0.78
Jan. 02 – Apr. 02	0.87	0.86
Semi yearly	**Correlation Coefficient**	**Overlap Rate**
May. 01- Oct. 01	0.80	0.77
Nov. 01- Apr. 01	0.79	0.80

Table 3. Average and Standard Deviation for results of top 1000 queries

	Average for correlation coefficinet	**Standard deviation for correlation coefficient**	**Average for cverlap rate**	**Standard deviation for cverlap rate**
Month-to-month	0.905	0.025	0.881	0.026
Quarterly	0.833	0.035	0.820	0.040
Semi yearly	0.795	0.007	0.785	0.021

The top 1,000 queries are on average 88.1% the same month-to-month, 82.0% the same quarterly, and 78.5% the same on a semi-yearly basis. The fairly low standard deviations

for Overlap and Correlation scores found from table 3 indicate that the queries and their ranks in the logs are fairly stable. However, we also notice that as the time span increases, the two measurements decrease: the query logs will be more different as time goes by.

The question now becomes, are those same queries remained same from time to time? Below is the query Overlap curve over one full year.

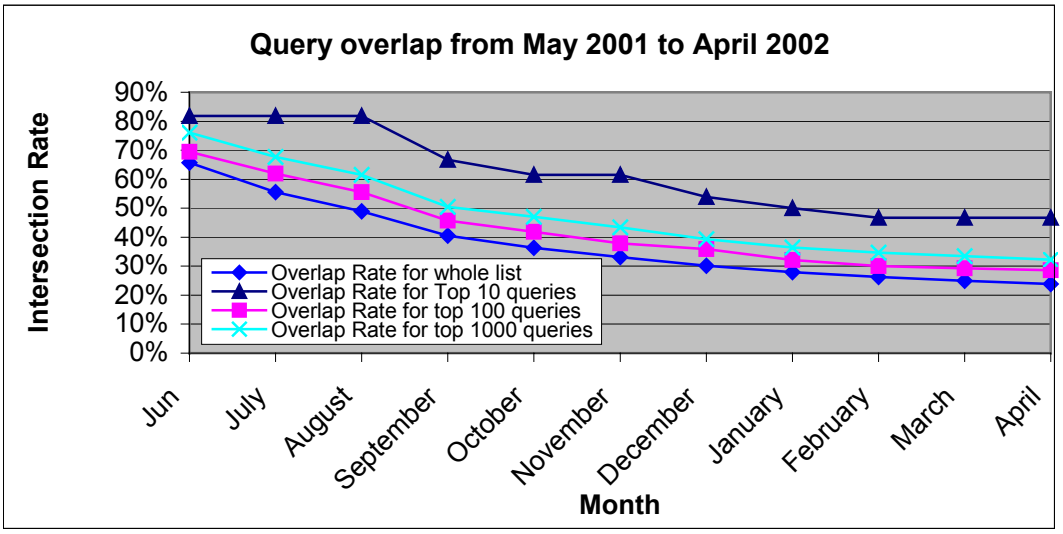

Fig. 2 Query Intersection Curve for the whole year

Figure 2, shows:

1. Queries that survived the whole year only account for 20-30% of all unique queries.
2. The slope of the overlap curve tends flatten, indicating that the longer a query survives in a log file, the more likely it is to remain in the logs in the future.

In order to explain how important those overlapped queries are, the occurrences of the queries was measured instead of the number of unique queries. Two ratios are calculated:

1. Total overlapped query occurrences. Total occurrence number (a constant) for a whole year.
2. Total overlapped query occurrences vs. total occurrence number at different level of top queries

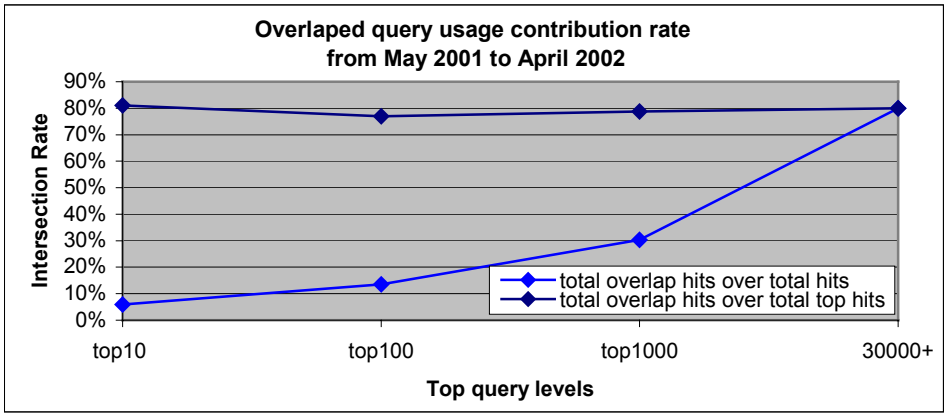

Fig. 3 Overlapped query usage contribution rate

The black line is quite stable, meaning the overlapped queries always count for 80% of total usage in each level of top queries.

The total overlapped queries accounts for 80% of total queries on AOL Search, however, they only account for 23% of total unique queries over a full year. This tells us that finding the small number of overlapped queries will help in improving a major part of the AOL Search service.

4 Conclusions and Future Work

Based on the findings from this work, we can answer the questions at the beginning of this paper:

1. The users' information needs are quite stable from time to time. More specifically, they only change 20% over time.
2. Future user queries can be predicted by examining users' past query logs (at an 80% confidence level).
3. Tracking user logs is a routing problem, and it is possible to produce the tools to help in identifying routing query lists.

As tracking user logs is a routing problem, the next step is to produce tools or hierarchies to help in identifying and qualifying the benefits of routing query lists.

References Query log similarity is examined in order to determine:

1. If tracking user queries success is a routing problem that is measurable.
2. If its possible to produce the tools to help in identifying and qualifying the benefits of routing query lists.

1. C. Badue, R. Baeza-Yates, B. Ribeiro-Neto and N. Ziviani. Distributed Query Processing Using Partitioned Inverted Files. *In Proceedings of SPIRE 2001*, IEEE CS Press, Laguna San Rafael, Chile, pp. 10-20, November 2001.
2. A. Chowdhury, D. Grossman, O. Frieder, C. McCabe, "Analyses of Multiple-Evidence Combinations for Retrieval Strategies", In *Proceedings of the 24th International Conference on Information Retrieval,* ACM-SIGIR, September 2001.