Measuring the Meaning in Time Series Clustering of Text Search Queries

Bing Liu *
Dept. of Computer Science
Virginia Polytechnic Institute
Blacksburg, VA 24061
bingl@vt.edu

Rosie Jones
Yahoo! Research
3333 Empire Ave
Burbank, CA 91504
jonesr@yahoo-inc.com

Kristina Klinkner *
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
klinkner@stat.cmu.edu

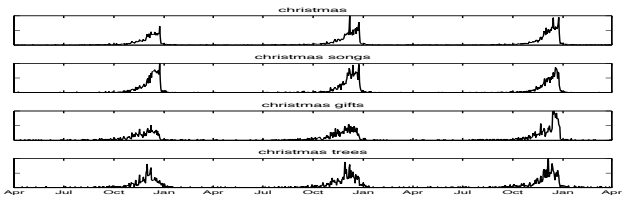

Figure 1: Daily web search query volumes of "Christmas" and seasonally related queries.

ABSTRACT

We use a combination of proven methods from time series analysis and machine learning to explore the relationship between temporal and semantic similarity in web query logs; we discover that the combination of correlation and cycles is a good, but not perfect, sign of semantic relationship.

Categories and Subject Descriptors:H.3.3 Information Storage and Retrieval Information Search and Retrieval [Clustering, Query Formulation]: H.3.5 Information Storage and RetrievalOnline Information Services [Web-based services]

General Terms: Experimentation, Measurements

Keywords: time series, query clustering, semantic similarity, Web search, query log analysis

1. INTRODUCTION

The time series of a query on a search engine, recording how many searches there were for that query per unit time, reflects changing patterns of collective behavior for the users. Queries which are semantically related may have similar patterns of popularity with users over time, so temporal characteristics might be used as features for query suggestion and query expansion, complementing content- and link- based approaches. Some of this information is also carried by static features, like within-session query co-occurrence but that lacks temporal context, such as an increase in searches on Christmas related terms during the holiday season, or a sudden peak related to a particular news event.

We know that temporal similarity is not *always* an accurate indicator of semantic similarity. For example, two unrelated events may take place on the same day, but queries about those events could have no semantic relationship. These are reflected by a temporal pattern with a major peak at a particular point in time, but no cyclic regularity. Since these false positives may occur when events happen to be correlated in time, but only, literally, coincidentally, we could make the basic assumption that events which co-occur regularly, on a weekly, monthly, yearly, etc. basis are more likely to be semantically related, and test that assumption.

Jones et al. [2] use within-session co-ocurence of terms to do query expansion. Recently a few groups have applied methods from time series analysis to query data: Vlachos et al [3] use Euclidean distance on leading Fourier coefficients for similarity matching. Chien and Immorlica [1] use cross-correlation to perform nearest neighbors search. While Vlachos et al use structure-based similarity, periodicity specifically, and Chien and Immorlica use shape-based similarity, we believe neither, on its own, is adequate to capture meaningful semantic relations. We use standard methods from time series analysis, cross-correlation and Fourier analysis; we check first cross-correlation (measuring shape-based similarity) over time and then filter by matching periodicity of queries (matching structure-based features).

2. EXPERIMENT

Our data is from search engine query logs from years 2002 to 2005, with around 1 million queries per day. We use queries appearing in a daily list of the top 10,000 most popular queries at least once in the time period. The *query time series* is represented as a sequence of daily counts of the number of users issuing that query. The data set contains 45,200 distinct queries with volumes on 1,096 days.

We perform two sets of clustering experiments. In the baseline experiment, we use only cross-correlation (Euclidean distance) as a similarity measure. In the second experiment, we initially check periodicity and partition the queries into 4 different periodic groups, then cluster based on cross-correlation to produce periodicity-constrained clusters. Since we restrict the search for periodicity to 4 frequencies: week, month, half a year, and one year, this partitions the original query time series into 5 subsets. Note that we treat *any other* periodicity which is not one of these 4 as *aperiodic*. Then complete link clustering is run within each subset; the summary of the results is shown in Table 1.

We manually labeled the clusters in several random samples, with the number of semantically related queries, the type of their semantic relation, and the broad topical category those queries belong to. We define the *semantic coher-*

*Work conducted while author was at Yahoo! Research.

CIKM'06, November 5–11, 2006, Arlington, Virginia, USA.
ACM 1-59593-433-2/06/0011.

leading period	fraction of all queries	proportion non-singleton
aperiodic	66%	0.69
a year	9%	0.85
half a year	2%	0.71
a month	1%	0.33
a week	22%	0.41

Table 1: Distribution of queries across different leading cycles, along with the proportion within the periodicity class which fall into non-singleton clusters. 66% of queries are aperiodic (recall that this just means without one of the 4 frequencies we check), and we are able to group 0.69 of those into clusters. Annual and semi-annual queries cluster well, while queries with monthly and weekly periodicity tend to produce singleton clusters. Weekly cycles are dominant.

ence coefficient of a group as $\mathcal{S}$, the proportion of semantically related queries in the group. If a group contains multiple semantically related sub-groups, the largest sub-group is used to calculate the coherence coefficient.

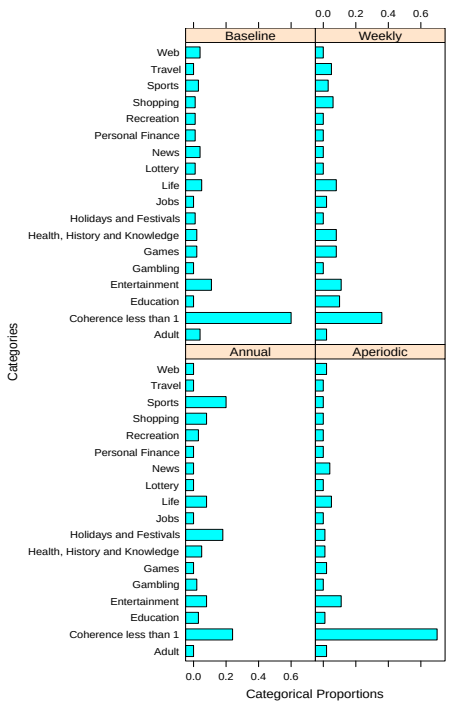

Figure 2: Categorical proportions for sample clusters with $S = 1$, each bar represents a proportion of the clusters for a given semantic category, grouped by different periodicities. Note the much larger fraction of $\mathcal{S} < 1$ for the aperiodic and baseline clusters.

We randomly sampled 100 clusters from the 4,751 clusters of the baseline experimental results. There were 63 clusters with $\mathcal{S} \geq 0.5$, including 40 "perfect" clusters with $\mathcal{S} = 1.0$. To repeat the experiment with periodicity constraints, we randomly sampled 50 clusters, for each of the 5 periodic subsets, from our experimental results, with the exception of the monthly subset. There are only 31 clusters in that subset, all of which are used as samples. We pool all 5 sets and the results for semantic coherence are that for $\mathcal{S} = 0.5$, 79% of the clusters are coherent and "perfect" clusters account for 53.5%, compared to 40% for the baseline case.

Similarity Source	Precision
Timeseries nearest neighbor	61% (43/70)
Session and Timeseries	94% (29/31)

Table 2: Temporal and within-session information together find highly relevant query pairs. However, 33% of semantically-similar pairs of queries identified by temporal clustering had negligible within-session similarity. Those not found using session information are shown in Table 3.

xfiles → thexfiles.com
playstation 2 codes → cheatcodecentral.com
bay area fireworks → fireworks in los angeles
lunar eclipse may 2003 → lunar eclipse may 15
valentinecards → valentine greetings
captaincode.com → playhousedisney.com
al jezzera → al jazera
mjnews → mjnews.us
death of john ritter → johnritter.com
al roker → merlinsantana.com
golf courses → camping
beheading video berg → nick berg video beheading
maine election results→ colorado election results
janet jackson superbowl breast pictures →
janet jackson flash super bowl

Table 3: Pairs of queries which were found to be highly similar using temporal cross-correlation, which had no within-session similarity in our data sample.

The proportion of topics for different periodicities and for the baseline case (looking at clusters with $S = 1$) are in Figure 2. Entertainment is a large fraction in all cases. Compared with the baseline, the proportions of holiday-related and sports-related clusters increase for the yearly queries; for the weekly queries, the education-related clusters stand out.

We also found that when both within-session information and temporal information suggest that a pair of queries are similar, they are correct 94% of the time. However, within-session similarity can be found for only 29 of the 43 query-pairs correctly identified as similar by temporal correlation, a recall of 67%. In Table 3 we see the query pairs correctly identified as similar by cross-correlation, which were not identified by session information.

3. CONCLUSIONS AND FUTURE WORK

Correlation among query log time series can help identify semantically coherent clusters, but also yields many "false positives". Constraining all the queries in a cluster to have the same leading period helped, but still 27% of clusters were not coherent. Combining time-series and session similarity may lead to the best results for identifying semantically related queries.

4. REFERENCES

[1] S. Chien and N. Immorlica. Semantic similarity between search engine queries using temporal correlation. In *the 14th International World Wide Web Conference*, pages 2–11, Chiba, Japan, 2005.

[2] R. Jones, B. Rey, O. Madani, and W. Greiner. Generating query substitutions. In *WWW '06: Proceedings of the 15th international World Wide Web conference*, 2006.

[3] M. Vlachos, P. S. Yu, and V. Castelli. On periodicity detection and structural periodic similarity. In *the 5th SIAM International Conference on Data Mining*, Newport Beach, CA, 2005.