

Query Heartbeat: A Strange Property of Keyword Queries on the Web

Karthik B.R.* Aditya Ramana Rachakonda Srinath Srinivasa

International Institute of Information Technology, Bangalore 560100, India
{karthik.b.r, aditya.ramana, sri}@iiitb.ac.in

Abstract

This paper illustrates a strange property that was encountered when analysing keyword query data. The objective of this work was to observe the temporal properties of keyword queries. However, we found that generic keyword queries that have large enough volumes tend to follow the same “attractor” distribution. This phenomenon was discovered in a 3-month dataset. However, a visually similar phenomenon was also apparent when the keywords were searched on Google.

1 Introduction

The temporal nature of keyword searches over internet search engines carry useful information about their semantics. As a result, analysing the temporal signatures of queries has been attracting increasing attention in recent times [1, 5, 7, 9].

The shape of the temporal signature of a keyword can be an indicator to a class of queries that it belongs to, giving us insights into when, where and how much of the keyword to expect at any time.

For instance, on Google Trends¹, the queries `vldb` and `icde` display predictable spikes corresponding to submission deadlines, paper deadlines and the conference, in their overall shape. Similarly, while the keyword `tsunami` previously had very little volumes, December 26 2004, shows a huge spike for this keyword. After this event, the keywords `tsunami` and `earthquake` both show similar patterns. Figure 1 shows another pair of queries: `watersport` and `aircondition`, which display similar temporal shapes.

However, while observing these query patterns, we discovered another deeper pattern. Keyword queries tend to have the *same* overall temporal shape, whenever their volumes are high enough, no matter what

they are querying for. Figure 2(a) shows daily keyword volumes for two query terms: `yahoo` and `access`. While the temporal signatures look vastly dissimilar on an absolute scale, they are remarkably similar when normalised. Figure 2(b) shows the same two queries, where query volumes are normalised by removing the mean and dividing by the variance. This shape seems to be recurrent across different query terms and is more likely to be seen as query volumes increase.

We call this temporal shape the “attractor pattern” or the “heartbeat” of web keyword queries. To substantiate our findings, we also found an inverse correlation between query volumes and the DTW (dynamic time warp) dissimilarity between a given query pattern and the attractor pattern.

We describe the process we used that led to this discovery in the next section.

2 Temporal Signature Generation

The primary objective of our work was to cluster keywords based on the *shape* of their temporal signatures.

In order to capture the shape of a temporal graph of a keyword query, we need to perform the following:

1. Remove the scale factor from the shape
2. Encode the shape in a form that makes them pairwise comparable

We will first explain query shape encoding before taking up the first point concerning normalisation.

Query logs were first aggregated to a day-wise resolution, by adding up the number of queries in a day to represent its volume.

In order to encode the temporal shape of queries to make them pairwise comparable, we considered the quantum of change in query volumes from one day to the next. The quantum change was then projected onto a set of seven *primitives* labelled A through G.

The primitive A represents a sharp increase in query volumes, while the primitive G represents a sharp decrease in query volumes from one day to the next. Primitives B, C, E and F represent lesser amounts of

*Currently with Web18 Software Services, Mumbai, India.



Figure 1: Google Trends for the queries watersport and aircondition

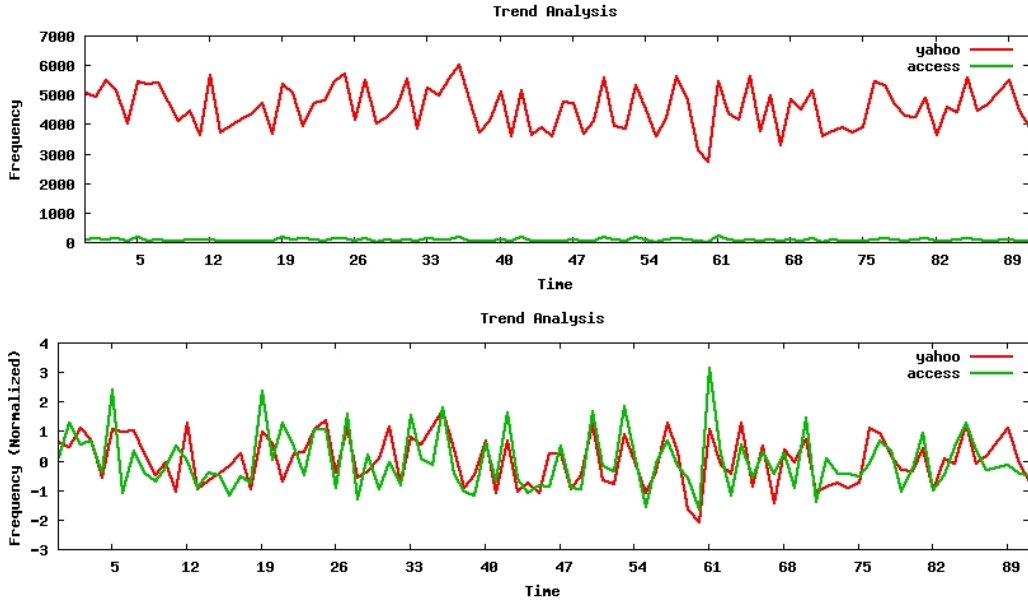


Figure 2: Absolute and normalised volumes for the terms yahoo and access on the AOL dataset

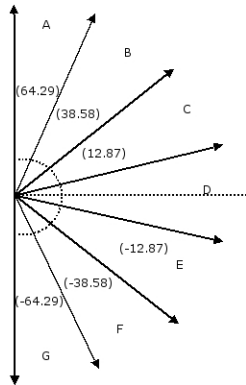


Figure 3: Primitives for representing quantum of change

increase and decrease in volumes respectively. Primitive D represents an insignificant or no change in query volumes from one day to the next.

The primitives are assigned by mapping the slope of the query graph in one time interval onto a radial region as shown in Figure 3. The mapping process is described below.

The slope of the line depicting change in query volumes can be determined as:

$$\Delta q = \frac{q_{t+1} - q_t}{\Delta t} \quad (1)$$

where q_t is the query volume at time t . The denominator Δt defines the resolution. For a low value of Δt the Δq value or the slope would be high and all changes would concentrate near A and G and similarly for a high value of Δt all changes would be low and the slope would be in D region. In order to minimise such biases in the calibration, we need to set Δt to a value such that it *maximises the entropy* of points falling across the different radial regions.

3. The data is collected over a period of three months from 1st March, 2006 to 31st May, 2006.

The dataset contains some sensitive user information which has been removed before performing the analysis. We used only the query aggregates and eliminated the user id field completely. As at the time of submitting this paper, the dataset is no longer available at its original source.

For each keyword in the dataset, a temporal signature, namely a time series, was prepared with a resolution of one day. The volumes depicted in this time series were normalised according to equation 4 and were reduced to primitives. For this dataset, we found that $\Delta t = \frac{1}{1.6}$ gave the maximum entropy in the distribution across primitives.

When the keywords were clustered, we consistently encountered the formation of one large cluster. Based on this, we suspected that the temporal shapes of keywords were similar.

When we compared the time series from different keywords we observed that they were identical. This was in spite of the lack of any similarity between the keywords semantically. The only common factor seemed to be the volume of the queries. The higher the volume of a pair of keywords the more their similarity was.

This similarity in shape was also visible on Google Trends (now on Google Insights³). Figure 5 shows Google Trends for the queries: **book**, **california**, **women**, **hair** and **access** in order from top. Figure 6, shows a more recent query on Google Insights for the same keywords. In Figure 7, we queried for each of the five words separately on Google Insights and overlapped them using an image processing tool.

Even though they are strikingly different from the trends from Figure 1, all of them have very similar shapes. We called this strange temporal shape, as the “heartbeat” of web queries.

The “heartbeat” on the AOL dataset is shown in Figure 8 where the respective search terms are: **google**, **yahoo**, **ebay**, **sale**, **lyric**. The respective query volumes are 552 332, 417 368, 215 302, 174 773, 200 284⁴. Figure 9 shows these queries after their volumes were normalised as shown in equation 4.

More examples of the “heartbeat” observed on AOL dataset are shown in Figure 9 where the respective search terms are : **google**, **book**, **car**, **picture** with respective volumes being 552 332, 93 649. 147 993, 291 505 and in Figure 11 where the respective search terms are : **google**, **house**, **school**, **mexico** with respective volumes being 552 332, 118 010, 245 838, 31 687.

Figures 12 and 13 shows these queries when their volumes were normalised as shown in equation 4.

³<http://google.com/insights/search>

⁴The volumes are the number of queries observed on that keyword in the dataset.

The important factor here was the volume; while the mean daily volume for the query **watersport** was around 5, the term **book** was queried at an average of around 1100 times in a day.

In order to verify the correlation with volume, we plotted the DTW distance between all queries q in our dataset with the query having the highest volume (which, incidentally was the keyword **google** in the AOL dataset).

Figure 14 shows how the DTW distance between a keyword q against the keyword **google** changes as the difference in the respective query volumes increase. It is apparent that as query volumes increase, the DTW distance to the “heartbeat” exemplified by **google** decrease. When the correlation between the log of the query volume and the corresponding DTW distance with **google** was computed, we obtained a value of -0.7865, indicating a high negative correlation between volumes and the DTW dissimilarity with the heartbeat. This correlation is also evident in figure 15, which shows the variation of DTW distance of each of the approximately 20 million queries against the log of query volume.

The similarity between the volume of all the queries put together and the volume of **google** is very high, as shown in figure 16. Even though the volume of all the queries put together and the volume of **google** are vastly different, the “heartbeat” emerges when the volumes of all the queries are summed up.

Another common factor across these queries showing the “heartbeat” is that, they all represent *generic* terms that don’t have periodic spikes and are not affected by external events. One of the high volume queries in AOL was the term **myspace**, whose time series was significantly different from the heartbeat. A major source of distortion was the occurrence of sudden spikes in the query volumes. We estimated the dates when some of the major spikes occurred and searched the Web for any events corresponding to **myspace** around that time. Sure enough, we found the following events corresponding to some of the major spikes:

1. On March 31st 2006, the market share of visits to Myspace Video was reported as having increased by 1242%⁵.
2. On May 2nd 2006, Myspace denied access to a user who had created a profile in the name of Barack Obama, a Chicago Senator⁶.
3. On May 22nd 2006, two teenagers were charged with illegal computer access into Myspace and at-

⁵<http://weblogs.hitwise.com/leeann-prescott/2006/04/myspace-video-a-youtube-killer.html>

⁶<http://www.thatpoliticalblog.com/serendipity/archives/827-Barack-Obama-steals-a-fans-MySpace-page.html>

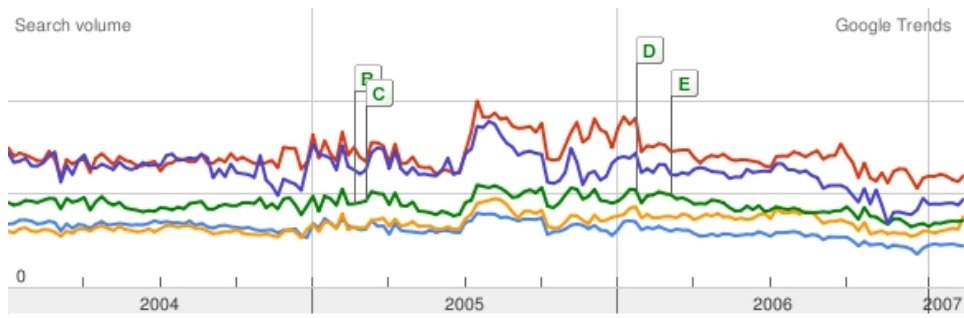


Figure 5: Google Trends for the queries: `book`, `california`, `women`, `hair` and `access` in order from top retrieved on August 10, 2007.

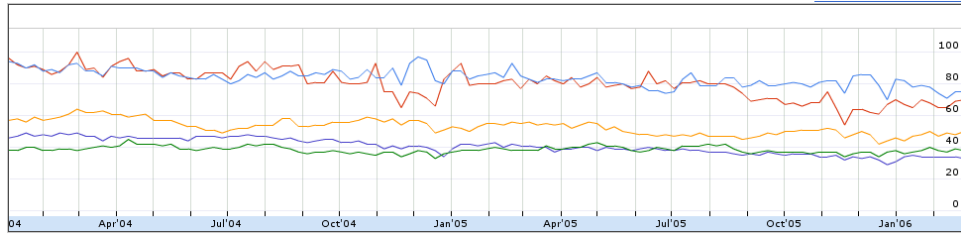


Figure 6: Google Insights for the queries: `book`, `california`, `women`, `hair` and `access` in order from top retrieved on August 7, 2008.

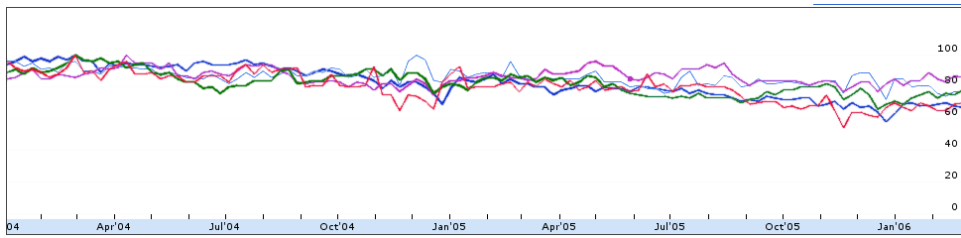


Figure 7: Manual overlap for the queries: `book`, `california`, `women`, `hair` and `access` in order from top retrieved from Google Insights on August 7, 2008.

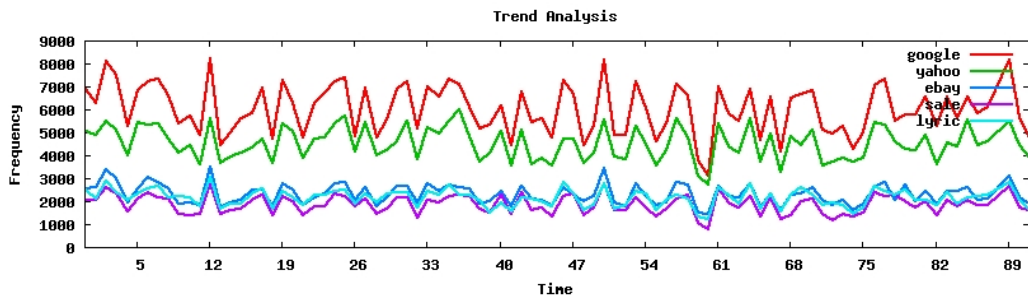


Figure 8: Attractor Distribution on AOL dataset

tempted extortion worth \$150,000⁷.

Hence, the heartbeat seems to be the characteristic of *generic* search terms having large enough volume and which are unlikely to be affected by periodicity or

⁷http://www.theregister.co.uk/2006/05/25/myspace_hack_charges/

external events.

One interesting characteristic, of the heartbeat observed, was its *rise and fall* pattern across all queries as shown in table 2. The conditional probability that a *rise* in volume would be followed with a *fall* in volume was more than double that of the conditional probability that a *rise* would be followed by a *rise*. A similar

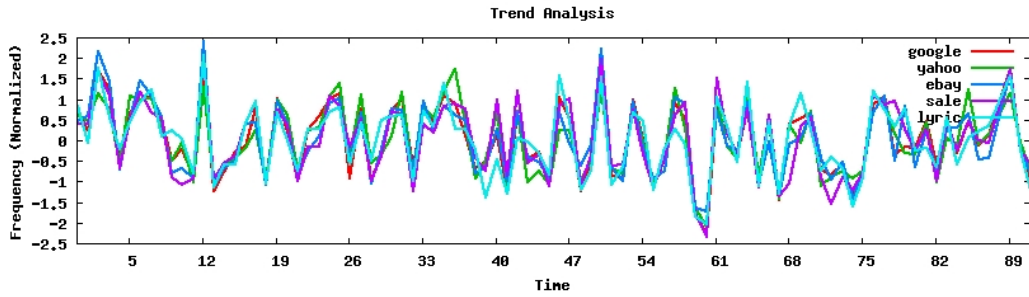


Figure 9: Normalised Attractor Distribution on AOL dataset

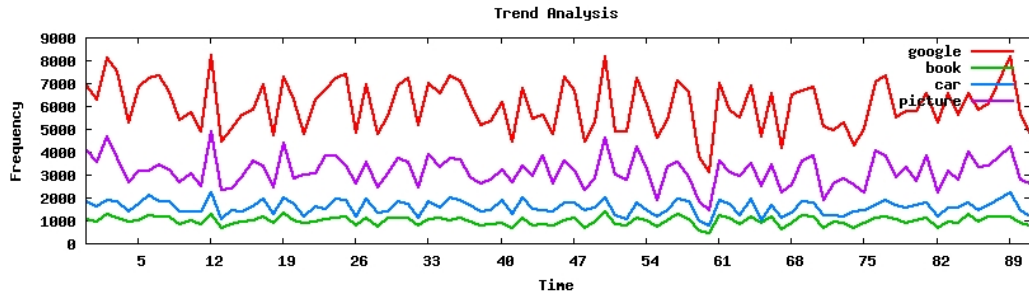


Figure 10: Attractor Distribution on AOL dataset

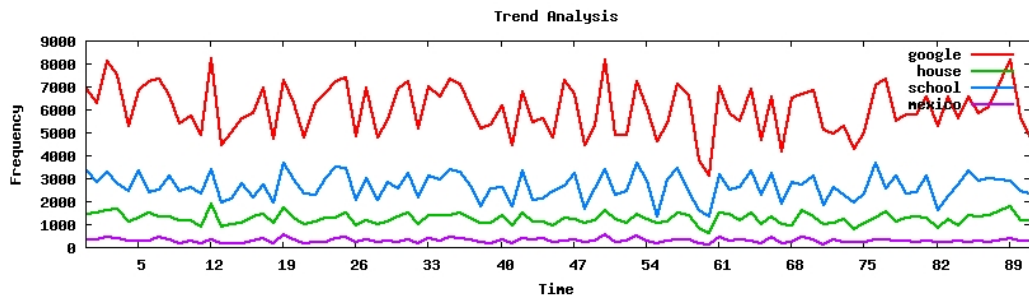


Figure 11: Attractor Distribution on AOL dataset

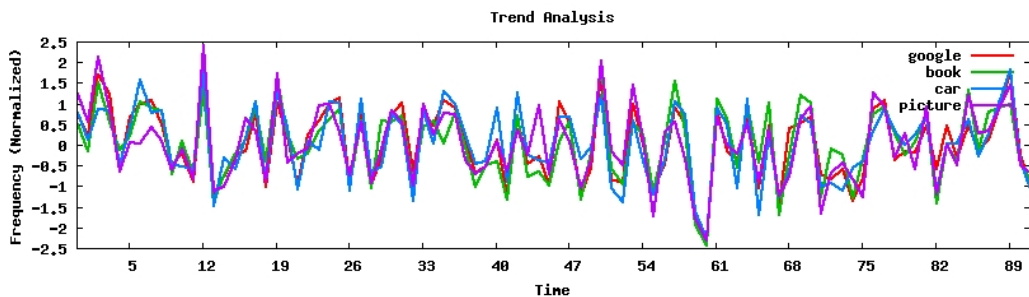


Figure 12: Normalised Attractor Distribution on AOL dataset

pattern was observed after a *fall* where it was more likely there would be a *rise* in the volume immediately after the *fall*. The heartbeat tended to oscillate more between peaks and troughs in volume rather than continually exhibiting an increasing or a decreasing trend

in volume.

It would be very interesting to discern other such characteristics of the heartbeat and perhaps also obtain a generative model using the hints given by the *rise and fall* pattern. Typically datasets depicting

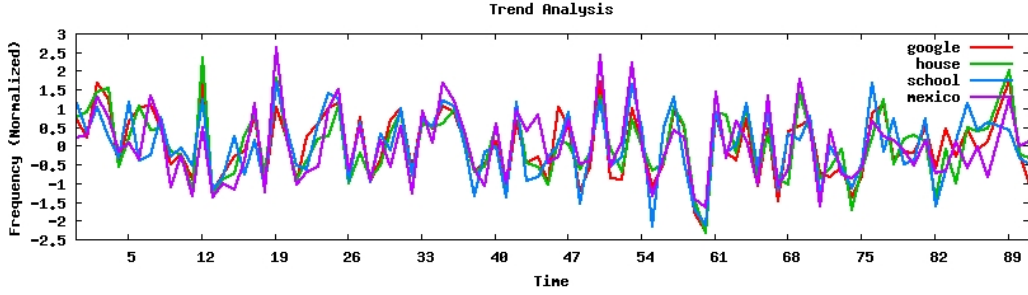


Figure 13: Normalised Attractor Distribution on AOL dataset

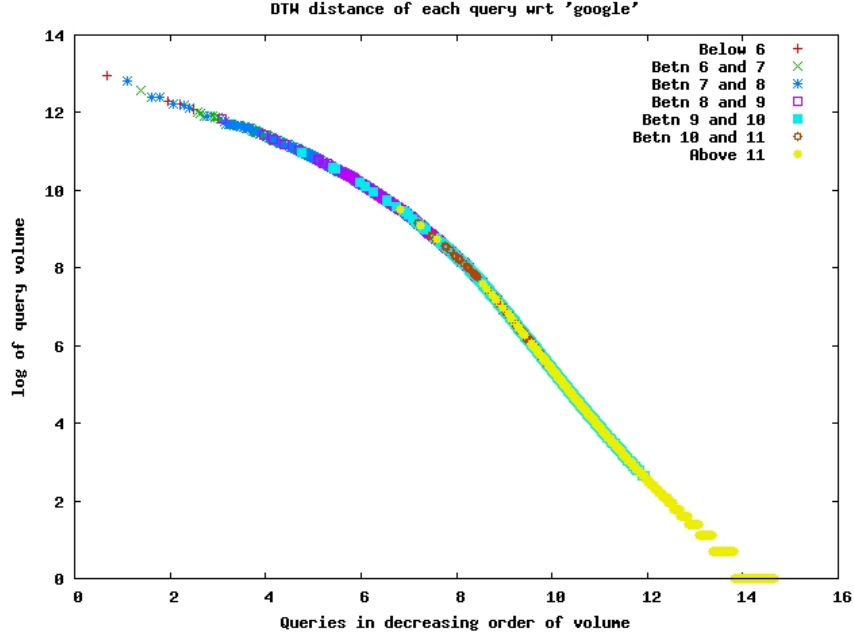


Figure 14: Plot showing change in DTW distance to google against query volume

Rise	$P(fall rise)$	0.68
	$P(rise rise)$	0.27
	$P(constant rise)$	0.05
Fall	$P(rise fall)$	0.60
	$P(fall fall)$	0.28
	$P(constant fall)$	0.12

Table 2: Transition probabilities

large number of autonomous processes, display long range dependencies (LRDs) or self similarity in their structure. Unfortunately, the dataset we had was only for 3 months, which was too little information to verify long-range dependencies. It would be desirable to verify the heartbeat on a larger dataset.

4 Related Literature

Vlachos et al. [10] propose a method for finding temporal similarity between queries based on Euclidean distance between query demand patterns over time.

They also propose a method to detect bursts in query demand. But, euclidean distance is brittle when comparing temporal sequences [4].

Chen and Immorlica [2] use correlation coefficient to find similar queries with respect to a given query. They also propose a mapping function which helps in finding related queries in real time. Liu et al. [5] use cross correlation as a distance measure. They partition the queries into periodic groups and then cluster them. Due to this, queries with no periodic behaviour will not belong to the same cluster. C. Silverstein et al. [8] try to find correlations between keywords in queries. They use chi-squared test and correlation coefficient to find the related queries and strength of the correlation. As an example, the keywords `www` and `com` are highly correlated. Keywords that occur in a phrase are considered similar and their correlations are calculated.

X. Shi et al. [7] define similar queries to be queries which can be reformulated by adding, deleting or changing some words of the original query string.

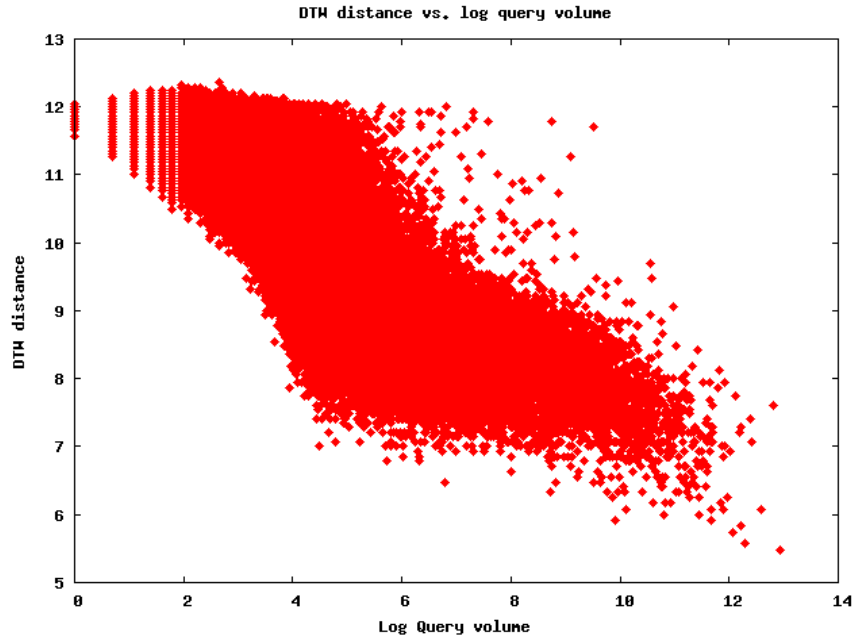


Figure 15: Plot showing the DTW distance to google against log of query volume

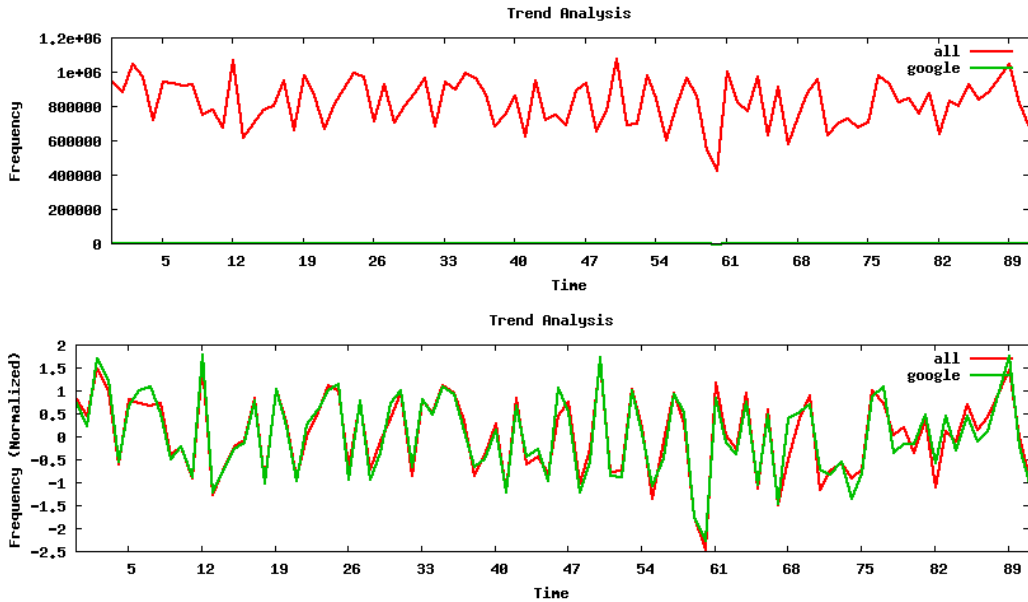


Figure 16: Absolute and normalised volumes for the total volume of all queries and the query google on the AOL dataset

They use a modified version of the traditional approach of mining association rules to find statistical associations between a given input query and other queries. They use levenshtein distance to calculate the similarities between queries. For example, they consider "adobe" and "adobe photoshop as similar queries. But queries like tsunami and earthquake which have a similar temporal behaviour will not be considered similar.

Ji-Rong Wen et al. [11] propose an approach to cluster queries using user logs . In this approach, queries are considered similar if users clicked on the same documents. Also, they consider queries and selected documents to have a stronger relationship than between these queries and other documents. They do not match temporal behaviour of queries, but rely on keyword similarity and target documents.

Qiakun Zhao et al. [13] use click-through data

from query logs to find similar queries. They use a marginalised kernel function to find similarity between any two temporal sequences. In the marginalised kernel function, they compare the keywords in queries to find the similarity between queries. They also use click-through data for different queries to compute the similarity measure.

Zhiyong et al. [12] propose two methods to find query recommendations from query logs. They consider the similarities between keywords present in queries. In the first method, they first calculate the similarity between all queries by computing their similarity in every session. In the second method, they calculate the similarity between the terms in queries and their inverse document frequency.

Eytan Adar et al. [1] present a study where they compare query logs, blog posts and news articles. They use [11] to cluster queries and find topics. They use events from one source and try to predict those in another. They use the cross correlation function to find the correlation between two time series.

In [9], re-finding behaviour among users is studied. A behaviour is considered re-finding if users click on the same result for different queries.

In figure 1, we see that queries `watersport` and `aircondition` show a similar temporal behaviour. Though they have no common keywords and also may not lead the user to select the same documents, we can still say that they are similar. In this work, we consider only the temporal behaviour of queries, by which we can find similarity between disparate queries.

5 Conclusions

In this work, we showed that disparate keyword queries on the Web across disparate sources exhibit strange central limit properties. This work also shows that the volume of the query determines the closeness of the query distribution with the attractor distribution. In the near future, we also intend to find the generative model for the attractor distribution.

References

- [1] E. Adar, D. D. Weld, B. N. Bershad, and S. D. Gribble. Why we search: Visualizing and predicting user behavior. In *Proceedings of WWW2007*, Banff, Alberta, Canada, May 2007.
- [2] S. Chien and N. Immorlica. Semantic similarity between search engine queries using temporal correlation. In *Proceedings of WWW2005*, Chiba, Japan, May 2005.
- [3] R. M. Gray. *Entropy and Information Theory*. Springer Verlag, New York, 2007.
- [4] E. J. Keogh and M. J. Pazzani. Scaling up dynamic time warping for datamining applications. In *Knowledge Discovery and Data Mining*, pages 285–289, 2000.
- [5] B. Liu, R. Jones, and K. Klinkner. Measuring the meaning in time series clustering of text search queries. In *Proceedings of 15th ACM Conference on Information and Knowledge Management (CIKM)*, Arlington, VA, USA, 2006.
- [6] C. S. Myers and L. R. Rabiner. A comparative study of several dynamic time-warping algorithms for connected word recognition. *The Bell System Technical Journal*, 60(7):1389–1409, 1981.
- [7] X. Shi and C. C. Yang. Mining related queries from search engine query logs. In *Proceedings of WWW2006*, Edinburgh, Scotland, 2006.
- [8] C. Silverstein, M. Henzinger, H. Marais, and M. Moricz. Analysis of a very large web search engine query log. Technical Report 1998-014, Digital System Research Center, October 1998.
- [9] J. Teevan, E. Adar, R. Jones, and M. Potts. History repeats itself: Repeat queries in yahoo’s logs. In *Proceedings of 29th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR ’06)*, Seattle, WA, USA, August 2006.
- [10] M. Vlachos, C. Meek, Z. Vagena, and D. Gunopulos. Identifying similarities, periodicities and bursts for online search queries. In *SIGMOD ’04: Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, pages 131–142, New York, NY, USA, 2005.
- [11] J.-R. Wen, J.-Y. Nie, and H.-J. Zhang. Query clustering using user logs. *ACM Transactions on Information Systems*, 20(1):59–81, 2002.
- [12] Z. Zhang and O. Nasraoui. Mining search engine query logs for query recommendation. In *Proceedings of WWW2006*, Edinburgh, Scotland, 2006.
- [13] Q. Zhao, S. C. Hoi, T.-Y. Liu, S. S. Bhowmick, M. R. Lyu, and W.-Y. Ma. Time-dependent semantic similarity measure of queries using historical click-through data. In *Proceedings of WWW2006*, Edinburgh, Scotland, May 2006.