Beyond Word Embeddings: Temporal Representations of Words using *Google Trends*

Md Enamul Haque
Spencer Center for Vision Research
Stanford University
Palo Alto, CA 94303 USA
Email: enamulh@stanford.edu

Aniruddha Maiti
Dept. of Computing and Information Science
Temple University
Philadelphia, PA 19140 USA
Email: aniruddha@temple.edu

Mehmet Engin Tozal
School of Computing and Informatics
University of Louisiana at Lafayette
Lafayette, LA 70504 USA
Email: metozal@louisiana.edu

***Abstract*—The main essence of word representation in a vector space is to preserve the similarity between the words. Traditional measures of word similarity retain the contextual or semantic affinity among the words. In this study, we propose an alternative word embedding scheme which considers the temporal relationships among the words. We employ the Google Trends search queries along with the respective time series information to represent words in a vector space. Our experiments show that the proposed representation is capable of incorporating temporal context that is otherwise unavailable in conventional word representations.**

I. Introduction

Conventional word representation in vector space is built on the idea of word similarity in contextual or semantic sense. Two words are close to each other in the vector space if they are contextually or semantically similar. The contextual or semantic information is captured from text documents in order to build the vector representation. Two obvious questions in this regard are a) what other type of similarities can be used to represent a word and b) what could be the alternative source of such information other than large volume of text corpus. It is true that the traditional text documents are by far the most informative source that is available today when it comes to word-related studies but it is probably fair to state that texts or sentences alone can not capture all the information related to a word. One such limitation of texts is that they are static in nature and the time related information can largely be missing or difficult to capture using only textual information. It is true that natural languages have words related to temporal sense, for example *'day', 'week', 'night', 'concurrently', 'simultaneously', 'frequently', 'repetitively', 'date', 'time', 'hour'* etc. But these words are not treated in a special way in conjunction with other words to impose temporal context in words. One motivating example of using time information in text-analysis is studies using microblog dataset where in addition to textual documents, time-stamps, indicative of the time of document generation, are available. A large number of studies have successfully used such metadata information for various tasks. The present study is performed in this larger context, where time-related information is accepted and well established to be useful in text-analysis. The rapid developments in the field of internet speed and connectivity have opened new avenues of data sources to capture temporal information. One such source of information is Google Trends [1]. Google Trends data can be thought of as a source to measure current level of interest shown by the population on a given word. The time series index related to the volume of queries reported in Google Trends has become important source of information to understand the present trends in the interest related to a particular search query [1]. There have been various studies across disciplines which make use of this passively generated digital data for predictive or now-casting [2] modelling. An important property that is often overlooked in this context is interesting temporal trends of different words/queries. We found that majority of the words demonstrate strong and often non-obvious seasonality in Google Trends. A natural consequence of this property exhibited by majority of the words is that we can use these temporal patterns to define the similarity among words from temporal perspective and represent them in a vector space.

Given the usefulness of Google Trends data to capture word similarity, we propose Temporal Representation of Words using Google Trends (`TeRGoT`). This proposed representation can solely be used as a supplement for conventional vector representation or it can be used as a complementary source of information along with traditional vector representations, for instance "BERT" [3], of words to enhance the quality of such traditional representations. Our experiments on Google Trends data corresponding to the top (3000) words of Wikipedia show some interesting temporal similarity that are captured using the proposed representation. We have demonstrated that temporal representations can often provide insights that are otherwise not easily captured in traditional vector representation of words derived using semantic context.

II. Google Trends : A real time indicator of interests

Google Trends data provides an index or measure of the relative volume of queries entered by users in Google search engine platform in a given time from a given geographical location. According to the information regarding the computation of this index available in Google Trends, the index does not represent the raw volume of the search queries. Given a single search term, the index is calculated from a sample

[1] https://www.google.com/trends

from all queries from a certain geographic region. First, the total query volume in the sample related to the search term is divided by the total number of queries in the sample. This relative ratio is then normalized in a scale between 0 and 100. Not every possible search query is included. Search terms with low volume are not reported (or appear to have an index 0). Additionally, Google Trends excludes "repeated searches from the same user over a short period of time" so that a single user can not influence the index by performing a frequent search of a given query [2]. There is a fair amount of online information available in the Google Trend web-page detailing the methodology of computing the index.

III. Motivation: Interesting Temporal Trends

We observed different types of trends corresponding to keywords in Google Trends. There are some natural seasonality patterns such as daily, weekly, monthly, and yearly. The majority of the words demonstrate at least four of these types of seasonality patterns. Some words might have more than one of such seasonal trends, moreover they might posses some other unique trends or a sudden burst in query volume influenced by a related event. The nature of the trend depends on the word itself and accordingly two similar trends indicate that there is a similarity between the concerned words in temporal sense. For the purpose of clarity, some different types of trends exhibited in Google Trends are discussed below.

A. Daily Trends

The daily trend is by far the most obvious trend present in almost every word. Due to people's sleep patterns, we can expect the nightly volume of the search query related to a search term will be low. But the normalization with respect to the total volume of search, as is done in Google Trends, is expected to eliminate this effect. It is interesting to note that in spite of the normalization, daily trend exists for majority of the search terms. Figure 1 shows the example of such daily

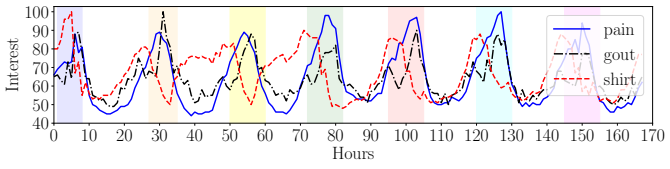


Fig. 1. Daily Trends

trends. Three words out of which first two, *pain* and *gout* are related words and the third one, *shirt*, apparently unrelated to the first two words shows daily trends.

B. Weekly Trends

To demonstrate weekly trend we will use two related keywords *salary* and *paycheck*. The Google Trend data related to these two keywords shown in Figure 2. We found similar weekly trend for the word *suicide*, a keyword used in [4] to model suicide occurrences.

[2]Source: Google Trends (https://www.google.com/trends)

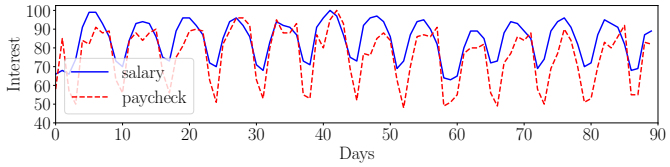


Fig. 2. Weekly trends

C. Monthly Trends

Similar to the daily and weekly trends, monthly trends are also prevalent. An example is presented in Figure 3 using two words having obvious monthly trends as they corresponds to two particular date of a month.

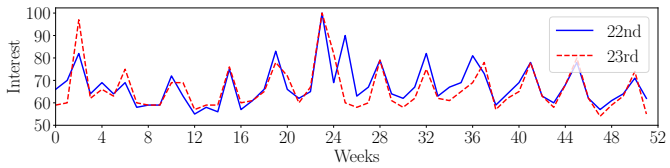


Fig. 3. Monthly trends

D. Yearly Trends

In our experience, yearly patterns are by far the more informative than other types of seasonal trends. We consider two examples for yearly patterns. The first example is comprised of words related to an event, Halloween, shown in Figure 4. The trends corresponding to these keywords experienced a burst of query during the time when the event occurs in each year.

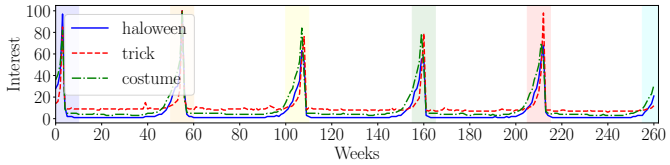


Fig. 4. Yearly trends related "Haloween" event

In the second example, some keywords used in a popular online website which reports the popularity of programming languages : PyPL [5] is used. PyPL prepares the search term by concatenating the word "tutorial" after the programming language name in order to compare different programming languages. For example, to compare java with python, PyPL uses queries "Java Tutorial" and "Python Tutorial" to measure their relative popularity. This strategy helps to weed out the effect of spurious unrelated search queries containing the same spelling as that of a programming language. As the purpose of this study is not to compare the programming languages, we can use the original keywords without concatenating the word *programming* at the end. The single word search terms are found to have much higher index in Google Trends than multiple words search terms. We inspected the trends of two popular programming languages Java and Python. We found that these words has a yearly trend in which the interest dropped to an extent during the month December and January. This yearly phenomenon can be attributed to the fact that a long holiday

season starts at the end of December and continues until the first half of January. During these days as people more likely not to work and thus, as shown in Figure 5, relative volume of queries related to these programming languages are low.

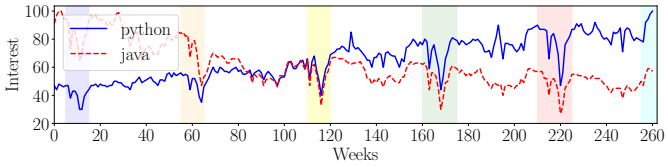


Fig. 5. Yearly trend between "python" and "java" programming language

IV. Capturing Temporal Context of Words from Google Trends

To capture the temporal context of words and to represent the words based on these temporal contexts, trends corresponding to six different time spans are collected from Google Trends. Google Trends provide trends related to keywords in seven default spans : past one day, past one week, past one month, past three months, past one year, past five years and entire available time beginning from 2003. According to our observations, trends corresponding to the beginning years of Google Trends are not so reliable as these were the beginning years of this web based tool. For this reason, only first six default time spans are considered. We took top 3000 words from Wikipedia according to their frequency of occurrence and then collected their trends corresponding to these six time series indicating the trend index of corresponding time spans. Let us denote the time series corresponding to word w as T^w_{1d}, T^w_{7d}, T^w_{1m}, T^w_{3m}, T^w_{1y}, T^w_{5y}. It may be noted that there are overlapping information in these six time series. It might seem that 5 year time-span is capable to capture all the information. But, as the resolution of the data is limited by the default Google Trends settings, it is difficult to capture all the seasonal patterns from only one series. For example, weekly trends are best captured from 1m or 3m data. Daily trends are best captured from the weekly trends etc. For this reason all these six trend data corresponding to a particular word is used. These six time series are then concatenated to form the combined time series corresponding to a word.

To derive the temporal representation, the combined vectors associated with top 3000 words in Wikipedia are then used to form the data matrix $D \in \mathbb{R}^{3000\times 547}$. Robust Principal Component Analysis is then performed on the data set D to transform the 3000 points with dimensionality 547 in a low-rank and a sparse matrix.. We used the entire 547 dimension. However, we can apply dimensionality reduction method on top of our method as well. The current `TeRGoT` method uses raw components from RPCA. We postulate that the words represented in this manner captures the temporal information associated with the Google Trends.

A. Robust Principal Components

Robust principal components is a variant of Principal component analysis [6], [7]. The main idea behind robust PCA is that the errors in the dataset can occur in large quantity but sparsely (that is with only a few entries). So a balance between the sparseness of the error and the quantity of the error is considered to calculate the principal components optimally. Robust PCA is ideal for the purpose of the current effort as in Google Trends, although words show strong seasonality, sometimes an increased interest might change the well established seasonality pattern. Robust PCA will be capable to amend such irregularities.

Robust PCA solves the following optimization problem:

$$\begin{aligned} \min \quad & ||\mathbf{L}||_* + \lambda||\mathbf{S}||_1 \\ \text{s.t.} \quad & \mathbf{D} = \mathbf{L} + \mathbf{S} \end{aligned} \tag{1}$$

where λ is a positive regularization parameter. RPCA can recover underlying low-rank structure in the data even with the presence of outliers, noise, or large error.

To solve the RPCA problem in Equation 1, a Lagrange multiplier Y is introduced in [8] as follows:

$$\mathbb{L}(\mathbf{L}, \mathbf{S}, \mathbf{Y}, \mu) = ||\mathbf{L}|| + \lambda||\mathbf{S}||_1 + \frac{\mu}{2}||\mathbf{D} - \mathbf{L} - \mathbf{S}||_\mathbf{F}^2 \tag{2}$$

where μ is a positive scalar and $||\cdot||_F^2$ denotes Frobenius norm.

Our data matrix D of Google Trends has size $m \times n$ where each row of D represents temporal interest of a word. In our dataset m and n are $3,000$ and 547, respectively. Our primary goal of using RPCA to model the Google Trends data is to identify the temporal trends among words. We use sparse component of the transformed data D to find such trends.

V. Experiments

A. Dataset Description

At first we selected top 3000 words from Wikipedia-corpus [3] to use them as query words for google trends (Figure 6). While it might be desirable to increase the vocabulary size to a large number, due to unavailability of API, we have restricted our study on top $3,000$ Wikipedia words and used the corresponding data from Google Trends.

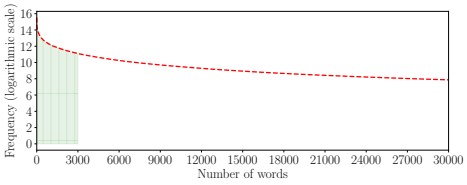


Fig. 6. Word distribution in Wikipedia corpus.

Multiple user options are available in Google Trends. Given a word, the trend can be *search term* based where any search query containing the word contributes to the trend of the word. On the other hand, the search can be *topic based* as well. Topic based search can be performed on several named entities. In this study, we focused on search term based trends. Furthermore, there are multiple categorized search available for different domains, for instance "Arts and Entertainment", "Auto and Vehicles", "Books and Literature",

[3] https://github.com/IlyaSemenov/wikipedia-word-frequency

"Business and Industrial" etc. These categorization is useful for disambiguation where the word having the same spelling might indicate different objects. In our study we included all available categories for simplicity. Google Trends also provides options to include search trends in only one type of platform (for example "image search", "news search" or "Youtube search"). In this study, we selected the default "web search" to collect the data. Regarding the time span of the trend time series, Google Trends comes with default options of past one day, past one week, past one month, past three months, past one year, past five years and entire available time beginning from 2003. In this study, the first six default time spans are considered as trends corresponding to the beginning years of Google Trends seems noisy. While, in principle, past 5 year data include all five other time spans, due to the limited resolution of time series provided by Google Trends we used all of them together to make the data more informative.

In Google trends, the user can select a certain geographic region to observe the trend restricted to that particular region. In this study, data associated with entire United States is considered.

B. Temporal vs Contextual Representation

To compare the proposed temporal representation, `TeRGoT`, we selected two state of the art methods `BERT` [3], and `GloVe`[9]. Both of these two pretrained models are trained on Wikipedia data which implies these methods rely on the context of the words to capture the similarity. It might be noted that our proposed method does not need to be trained on large corpus such as Wikipedia. Simple selection of words are sufficient to download the corresponding trends and form subsequent representation. The vector representation of words in `BERT` is of length 1024 whereas that in `GloVE` is 50. For `TeRGoT`, we used RPCA[4] to compute the low rank and sparse matrices corresponding to the original Google Trends data. The representation of each of the selected $3,000$ words is of length 547. To identify the similar words from these three different representations, we applied "Agglomerative Clustering" [10] to identify 50, 70, 100, 150, 200, 250 and 300 clusters.

C. Results

The results following the clustering associated with the three representations `BERT`, `GloVe` and `TeRGoT`, suggests that the proposed representation using `TeRGoT` can capture temporal association of words more distinctly than the other two.

In Table I a few selected clusters are shown. Cluster numbers starting with T, G, and B denote `TeRGoT`, `GloVe` , and `BERT` clusters respectively. The cluster names are followed by the cluster number in the corresponding agglomerative clustering. For the purposes of qualitative evaluation, at first, a keyword is selected and the corresponding clusters are identified for all three methods. Along with the cluster number, a few selected words of the clusters are also shown. For instance, the first row of Table I shows clusters that contains the word "miami", a name associated with a city. The cluster numbers of the `BERT`, `GloVe` and `TeRGoT` are $B18$, $G12$ and $T1$ respectively. From the table, it can be observed that `TeRGoT` gives preference to words that are associated with an event or game while for Glove the words are strictly geographical. For `BERT`, the words in the cluster $B18$, which contains the word "miami", does not seem to strongly suggest any geographical or temporal association [5]. The interest during a particular event or game is responsible for the similarity patterns in trends of words observed in `TeRGoT` cluster $T1$. Similarly, the word "Kentucky" is associated with the word "horse" in `TeRGoT` cluster $T5$. These two words are most likely to be associated with the trend related to "kentucky derby", a horse race held annually in Louisville, Kentucky, United States, on the first Saturday in May. This type of association is missing in the clusters $G49$ and $B29$ found in `GloVe` and `BERT` containing the word "horse". Similar event based words can be found in $T6$ which is associated with Christmas related annual event. Similar temporal association of words are also clearly inferable from other examples shown in Table I, for instance $T27$ in the last row is associated with Super bowl which is played in February.

Similar as in Table I, we also present some clusters in Table II for qualitative evaluation. For instance, the first row of Table II shows clusters that contains the word "olympics", a name associated with Olympic games. The cluster numbers of the `BERT`, `GloVe`, and `TeRGoT` are $B80, G33$ and $T34$, respectively. The result shows that both `TeRGoT` and `GloVe` provides "Olympic" and sports related words in the same clusters. However, `BERT` does not show similar temporal association of "Olympic" with other sports and awards such as "championship" and "gold". Similarly, the word "goal" is associated with the words "world", "cup", and "matches" in `TeRGoT` cluster $T42$. These words are most likely to be associated with the trend related to "World Cup" football, an international football/soccer event held every four years. This type of affiliation is lacking in the clusters $G6$ and $B76$ count in `GloVe` and `BERT` containing the word "goal". SImilar temporal and contextual association of words are also clearly noticeable from other examples shown in Table II, for instance "voting" is related to presidential election in United States.

We publish all clustering results [6] for 50, 70, 100, 150, 200, 250 and 300 clusters suggested by GAP score analysis discussed in Figure 13. Due to space constraints, only a few interesting clusters corresponding are shown here. Please refer to the tiny URL for individual cluster member information in details.

D. Seasonal Trends Associated with Words in the Identified Clusters

To understand and validate the intuition behind the use of temporal trends of words for vector representations we selected words from a cluster and observed their original trend. The idea was that different types of similar trends would contribute in the formation of different clusters.

[4]https://github.com/ShunChi100/RobustPCA

[5]only partial cluster member representation are shown

[6]http://tiny.cc/tergot

TABLE I
QUALITATIVE ANALYSIS ON 50 CLUSTERS

Cluster + Key	TeRGoT	GloVe	BERT
T1, G12, B18 + miami	baltimore, dallas, defensive, detroit, divisions, intended, losses, miami, network, offensive, pittsburgh, praised, schedule, seattle, starred, stream, team's, touchdown, uk, vs, wings	..., jersey, kansas, kentucky louisiana, maryland, massachusetts, miami, michigan, minnesota, mississippi, missouri, montreal, ...	alex, anderson, anna, anne, berlin, carl, ceo, chris, christ, dallas, diego, dna, dr, dublin, edward, frank, german, gordon, harry, hong, iowa, jane, jr, ...
T5, G49, B29 + horses	administrative, celebrated, champions, considerable, criticized, draft, graduating, horses, jazz, kentucky, may, round, subsequently, war	...,birth, boy, boys, child, children, couple, couples, daughters, dog, family, father's, female, friends, girl, girls, horse, horses, ...	abbey, access, account, accused, adaptation, ..., wounded, wrestling, young, younger, youth
T6, G5, B14 + christmas	christmas, december, present, santa	addition, anniversary, appearances, arranged, attend, ..., various, visit, visited, visiting, youth	allen, american, americans, arkansas, birmingham, christian, christmas, clark, edinburgh, elizabeth, england, frederick, george, georgia, germany, greece, greek, hamilton, ...
T8, G46, B43 + nfl	championship, january, nfl	..., legend, nba, ncaa, nfl, performance, performances, play, played, player, players, ...	abc, bbc, dvd, ltd, nfl
T10, G12, B44 + alabama	alabama, arkansas, athletics, club's, coaching, competitions, filmed, guitarist, iowa, kent, liverpool, maryland, mississippi, oklahoma, school, tennessee	G12	alabama, america, arizona, atlanta, carolina, florida, harris, indiana, indonesia, korea, louisiana, malaysia, minnesota, russia, russian, texas
T27, G46, B29 + bowl	bowl, february, super	G46	B29

TABLE II
QUALITATIVE ANALYSIS ON 100 CLUSTERS

Cluster + Key	TeRGoT	GloVe	BERT
T34, G33, B80 + olympics	bronze, count, entrance, field, gold, individual, iv, medal, meters, metres, olympic, olympics, participated, team, women's	bronze, champion, champions, championship, championships, crown, cup, doubles, finals, gold, grand, match, matches, medal, medals, miss, olympic, olympics, prix, ...	olympic, olympics
T42, G6, B76 + goal	brazil, cup, goal, matches, mexico, soccer, sweden, world	..., finished, finishing, first, fourth, gain, goal, goals, lead, leading, leads, led, losing, lost, major, minute, minutes, open, opening, penalty, ...	abandoned, afterwards, appointed, back, captain, ..., thereafter, ultimately, victory, win, wins,
T35, G44, B89 + voting	..., markets, measure, measures, mental, neither, officially, policies, proposal, registered, series, standing, supporters, thousand, thousands, voted, voting	..., nominated, nomination, parliament, parliamentary, presidential, republican, seat, seats, senate, senator, speaker, vote, voted, votes, voting,	..., representatives, republican, revolution, revolutionary, seat, seats, secretary, socialist, vote, voted, votes, voting

One such example can be observed in Figure 7, where, the trend corresponding to words in cluster $T6$ (refer to Table I) are presented. We can clearly observe a yearly seasonality of the interests in the search queries associated with the words which is quite explainable as these words are associated with Christmas, a yearly event observed in the month of December.

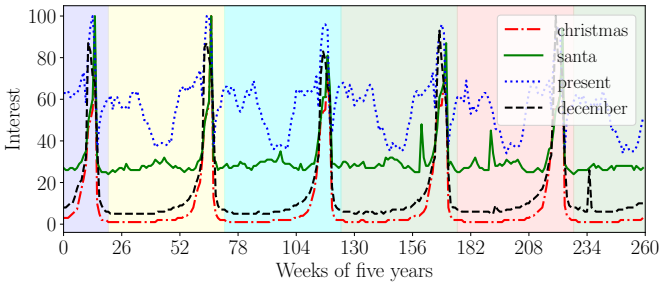


Fig. 7. Unique yearly trend that makes the words in the cluster $T6$ distinguishable

In Figure 7, the Y-axis represents search interest relative to the highest point on the chart. A value of 100 is the peak popularity for the term. A value of 50 means that the term is half as popular. A score of 0 means there was not enough data for this term.

Similarly, words having different types of seasonality have been grouped together to form a cluster in case of TeRGoT. Different such examples are presented in Figure 8, 9, 10, and 11. Figure 8 shows a yearly trend of most searched keywords related to hurricanes' projected track that hit Florida and the damage done by the disaster afterwards. At the same time, people also searched for different channels such as "The weather channel" to get relevant news of the disaster. Therefore all the keywords "track", "florida", "hit", "damage", and "channel" are grouped in the same cluster.

Daily trends can be observed in Figure 9 that shows relevant keywords searched during a three month period relevant to Olympic games.

Figure 10, on the other hand, presents a weekly trend

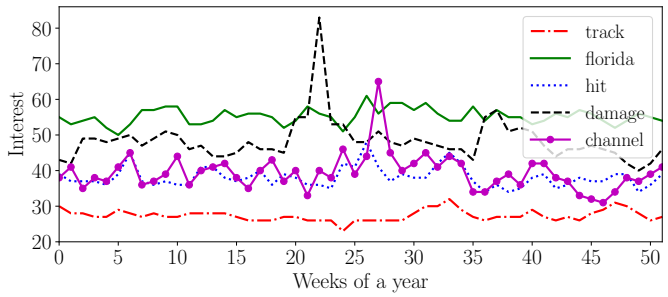


Fig. 8. One year shows monthly trend for "Hurricane" event in Florida

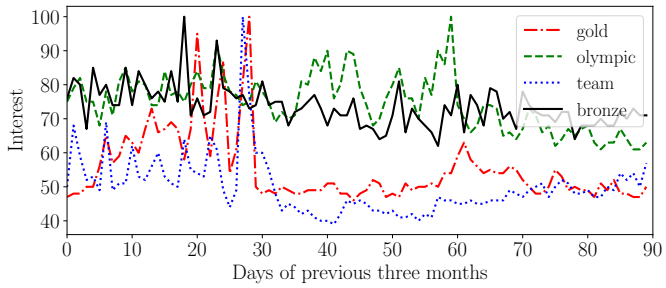


Fig. 9. Three month provides daily interest trend of the past three months. this particular scenario describes search words related to olympics.

which is very distinguishable. This particular example shows "churches" and "sunday" are relevant with respect to weekly religious event. Therefore, these kinds of words are put in the same group using the method TeRGoT.

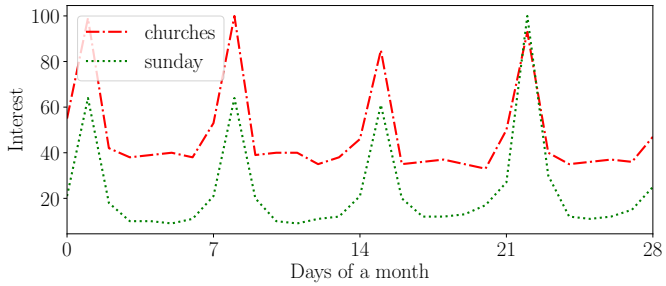


Fig. 10. One month provides weekly interest trend

Finally, Figure 11 shows that people were interested to search for the disaster in the Carribean island and how the property were damaged. They were also interested to search in the restoration process. As a result these words are clustered in $T248$ under 300 clusters. There were other words such as "coast", "residents", "destroyed", and "villages" in the same cluster.

E. Analysis on the Clustering Performance

In this section we perform some analysis with the clustering results to understand the nature of the clustering achieved by using RPCA on three types of representations.TeRGoT focuses on temporal trends and it is expected that the number of words associated with an event is limited. Therefore, it is expected that most of the clusters in TeRGoT will be formed

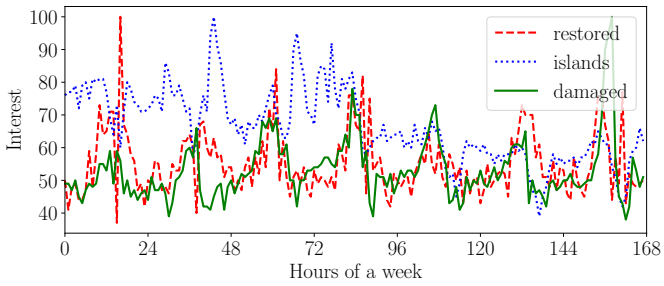


Fig. 11. Seven day provides hourly interest trend for the past weeks

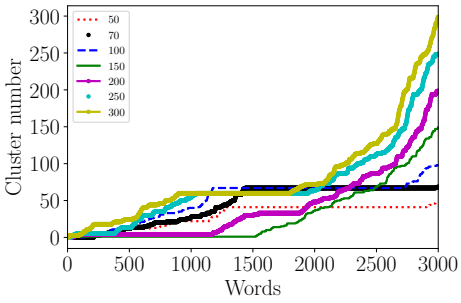


Fig. 12. Group plot

by a few keywords associated with an event. All other common words having no distinct temporal seasonality will form a large cluster. In the Figure 12, this phenomenon is shown for all seven clustering experiments with 3000 words. In each clustering experiment, we can observe a large cluster represented by a horizontal stretch which indicate the cluster number is same for all these words. The presence of a large cluster in each clustering experiment confirms that most of the words have a common trend which makes them indistinguishable from one another. On the other hand, the presence of several small clusters indicates unique trends associated with a few keywords that make them distinct from the common large group.

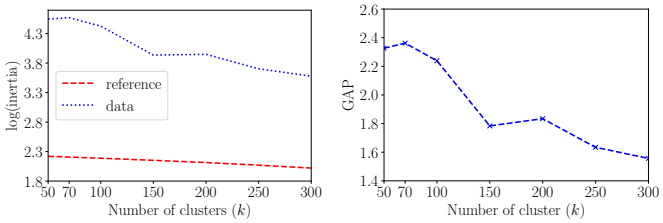
For comparison purposes, we calculated Silhouette score [11] of clusters identified from three different representations : TeRGoT, BERT and GloVe. Silhouette score measures both the intra-cluster affinity and the inter-cluster separation and ranges between -1 and $+1$. The result is presented in Table III which shows that TeRGoT demonstrates better cluster till 200 clusters. The scores for BERT are all negative which demonstrates poor clustering of words. GloVe clusters are very suitable with respect to inter and intra cluster distance and affinity, respectively. However, TeRGoT offers temporal representation of words which is unavailable in BERT and GloVe.

In addition to the Silhouette scores, we present Gap statistic which is a standard method for determining the number of clusters in a dataset [12]. The Gap statistic standardizes the graph of within-cluster dispersion by comparing to its expectation under an appropriate null reference distribution (Figure 13(a)). The number of clusters k needs to be chosen so that the rate of Gap score changes slow down. The authors in [12] suggests the 1-standard-error method to find optimal

TABLE III
SILHOUETTE SCORES FOR TERGOT, BERT, AND GLOVE

Method	50 Clusters	70 Clusters	100 Clusters	150 Clusters	200 Clusters	250 Clusters	300 Clusters
BERT	-0.0198	-0.0208	-0.0111	-0.0207	-0.0206	-0.0210	-0.0184
GloVe	0.0480	0.0551	0.0513	0.0591	0.0586	0.0613	0.0681
TeRGoT	0.0327	0.0149	0.0124	0.0192	0.0009	-0.0229	-0.0348

number of clusters k: $Gap(k) \geq Gap(k+1) - s_{k+1}$, where s denotes standard deviation. We can see from Figure 13(b) that the rate of change slows down after cluster count 100, which is considered as an ideal number of clusters for TeRGoT with respect to the current dataset.



(a) inertia plot (b) GAP score plot

Fig. 13. Inertia and GAP score

VI. DISCUSSION

Temporal trend based similarity measure of words have some interesting aspects. Similarity in trends between two keywords can happen in two ways. These words might be searched individually. In this scenario, the similar resulting trends indicate a purely temporal similarity. On the other hand, two words might comprise a frequently searched phrase such as "4th July". In this second case, similar trends in "4th" and "July" represent a semantic relation as well for these two words comprise the phrase "4th July". Thus, representation derived from search-query trends can capture both temporal and contextual similarities, although it might be argued that the effect of *context* is limited by the short length of the search queries. This limitation can be overcome by using temporal representation in addition to conventional context based representation such as word2vec [13].

Unlike conventional context based representations such as BERT, GloVe, or word2vec, the proposed time series based representation has spatial aspect as well given that the user option is present in Google Trends to restrict the trends in a certain geographic region. However, unavailability of sufficient number of search queries might restrict such an analysis.

Another important difference between the proposed method and conventional context based representations is that the proposed method does not require training in the sense that, unlike GloVe, BERT and other similar models, our approach does not require training using large text corpus. This makes the proposed method *language independent* in the sense that only the keyword and its historical time series from Google trends is sufficient to construct the representation, whereas in the conventional models, two words from two different languages might need two separate corpus corresponding to the respective language in order to construct the representation.

VII. RELATED WORK

Healthcare research and disease tracking: Google Trends data has been widely used in healthcare research and disease tracking purposes [14], [15], [16], [17]. It is also used as a novel data source for women's health research in [18]. For tracking different diseases such as tuberculosis, influenza, flu, and insomnia [19], [20], [21], [22], [23], the trend data is used with various statistical and survey models. The data is used for dynamic forecasting of "Zika" epidemics using linear regression and auto-regressive integrated moving average (ARIMA) model [24]. Authors in [25] discussed how disease outbreaks can be predicted before a week using Google Trend "flu" data. Recently popular non-cigarette tobacco use is tracked in [26]. Bariatric surgery, a surgery for reducing obesity, interest around the world is monitored and analyzed in [27], [28] using the same data.

Market analysis: There are several studies that investigate the use of Google Trends data to forecast near-term values of economic indicators such as automobile sales, unemployment, claims, and consumer confidence [1], [2]. Financial markets portfolio risk estimation is performed with the help of market index search interest on Google Trend [29]. Similarly, authors in [30] found patterns that may be interpreted as "early warning signs" of stock market moves. On the other hand, private consumption [31], unemployment rate, and new car sales in south-western Europe [32] is predicted based on the trend data. Tourism demand [33] and tourist inflow [34] are also measured using different seasonality trend analysis using the data.

Crypto-currency trends analysis: Digital currencies such as BitCoin search trends and Wikipedia relationship is studied in [35]. The authors show that the Google Trend interest data has a significant impact on the search query and price. Crypto-currency price behaviour is analyzed based on both linear and non-linear dependency test on Bitcoin, Ethereum, Ripple and Litecoin data in [36], [37]

Information technology and software trend: Software engineering and multiple technological topic related trends are analyzed to find cause-effect analysis of one topic or between multiple topics in [38], [5]. The authors in [39] uses social media and search volume data from Google Trend to find financial market influence. The properties of Google Trends is used as a measure of issue salience on four issues, fuel prices, the economy, immigration and terrorism in [40].

VIII. CONCLUSIONS

This paper proposes a word representation based on the temporal context of the interest shown by internet users in a popular search engine. The word representation derived from the time series in Google Trends shows that the proximity

of the words in the temporal sense is captured effectively using the proposed representation. A comparison with BERT and GloVe reveals that the representation derived from the textual context is not always sufficient to capture information associated with time. The proposed representation can be used in conjunction with any available context-based representation to enhance the quality of the representation and to improve subsequent tasks performed using these representations. A study is performed with the proposed temporal representation to show the nature and usefulness of such representations. This novel representation throws new insights on the nature of the internet search queries associated with different words. The availability of search query related data may inspire similar works in other domains.

REFERENCES

[1] H. Choi and H. Varian, "Predicting the present with google trends," *Economic Record*, vol. 88, pp. 2–9, 2012.

[2] Y. Carrière-Swallow and F. Labbé, "Nowcasting with google trends in an emerging market," *Journal of Forecasting*, vol. 32, no. 4, pp. 289–298, 2013.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[4] L. Kristoufek, H. S. Moat, and T. Preis, "Estimating suicide occurrence statistics using google trends," *EPJ data science*, vol. 5, no. 1, p. 32, 2016.

[5] P. Carbonnelle, "Pypl popularity of programming language," *URL: http://pypl. github. io/PYPL. html*, 2018.

[6] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM (JACM)*, vol. 58, no. 3, p. 11, 2011.

[7] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Advances in neural information processing systems*, 2009, pp. 2080–2088.

[8] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *arXiv preprint arXiv:1009.5055*, 2010.

[9] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[10] D. Beeferman and A. Berger, "Agglomerative clustering of a search engine query log," in *KDD*, vol. 2000, 2000, pp. 407–416.

[11] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[12] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 2, pp. 411–423, 2001.

[13] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[14] S. V. Nuti, B. Wayda, I. Ranasinghe, S. Wang, R. P. Dreyer, S. I. Chen, and K. Murugiah, "The use of google trends in health care research: a systematic review," *PloS one*, vol. 9, no. 10, p. e109583, 2014.

[15] A. Valdivia and S. Monge-Corella, "Diseases tracked by using google trends, spain," 2010.

[16] M. Moccia, R. Palladino, A. Falco, F. Saccà, R. Lanzillo, and V. B. Morra, "Google trends: new evidence for seasonality of multiple sclerosis," *J Neurol Neurosurg Psychiatry*, vol. 87, no. 9, pp. 1028–1029, 2016.

[17] C. Pelat, C. Turbelin, A. Bar-Hen, A. Flahault, and A.-J. Valleron, "More diseases tracked by using google trends," *Emerging infectious diseases*, vol. 15, no. 8, p. 1327, 2009.

[18] S. D. Fazeli, R. C. Carlos, K. S. Hall, V. K. Dalton *et al.*, "Novel data sources for women's health research: mapping breast screening online information seeking through google trends." *Academic radiology*, vol. 21, no. 9, pp. 1172–1176, 2014.

[19] X. Zhou, J. Ye, and Y. Feng, "Tuberculosis surveillance by analyzing google trends," *IEEE transactions on biomedical engineering*, vol. 58, no. 8, pp. 2247–2254, 2011.

[20] S. Cho, C. H. Sohn, M. W. Jo, S.-Y. Shin, J. H. Lee, S. M. Ryoo, W. Y. Kim, and D.-W. Seo, "Correlation between national influenza surveillance data and google trends in south korea," *PloS one*, vol. 8, no. 12, p. e81422, 2013.

[21] M. Kang, H. Zhong, J. He, S. Rutherford, and F. Yang, "Using google trends for influenza surveillance in south china," *PloS one*, vol. 8, no. 1, p. e55205, 2013.

[22] M. T. Malik, A. Gumel, L. H. Thompson, T. Strome, and S. M. Mahmud, ""google flu trends" and emergency department triage data predicted the 2009 pandemic h1n1 waves in manitoba," *Canadian Journal of Public Health*, vol. 102, no. 4, pp. 294–297, 2011.

[23] J. S. Albrecht, E. M. Wickwire, A. Vadlamani, S. M. Scharf, and S. E. Tom, "Trends in insomnia diagnosis and treatment among medicare beneficiaries, 2006–2013," *The American Journal of Geriatric Psychiatry*, vol. 27, no. 3, pp. 301–309, 2019.

[24] Y. Teng, D. Bi, G. Xie, Y. Jin, Y. Huang, B. Lin, X. An, D. Feng, and Y. Tong, "Dynamic forecasting of zika epidemics using google trends," *PLoS One*, vol. 12, no. 1, p. e0165085, 2017.

[25] H. A. Carneiro and E. Mylonakis, "Google trends: a web-based tool for real-time surveillance of disease outbreaks," *Clinical infectious diseases*, vol. 49, no. 10, pp. 1557–1564, 2009.

[26] P. A. Cavazos-Rehg, M. J. Krauss, E. L. Spitznagel, A. Lowery, R. A. Grucza, F. J. Chaloupka, and L. J. Bierut, "Monitoring of non-cigarette tobacco use using google trends," *Tobacco control*, vol. 24, no. 3, pp. 249–255, 2015.

[27] F. Linkov, D. H. Bovbjerg, K. E. Freese, R. Ramanathan, G. M. Eid, and W. Gourash, "Bariatric surgery interest around the world: what google trends can teach us," *Surgery for Obesity and Related Diseases*, vol. 10, no. 3, pp. 533–538, 2014.

[28] D. A. Telem and A. D. Pryor, "Google trends: Is it a real tool to predict the future of bariatric surgery or merely a marketing landmine?" *Surgery for Obesity and Related Diseases*, vol. 10, no. 3, pp. 538–539, 2014.

[29] L. Kristoufek, "Can google trends search queries contribute to risk diversification?" *Scientific reports*, vol. 3, p. 2713, 2013.

[30] T. Preis, H. S. Moat, and H. E. Stanley, "Quantifying trading behavior in financial markets using google trends," *Scientific reports*, vol. 3, p. 1684, 2013.

[31] S. Vosen and T. Schmidt, "Forecasting private consumption: survey-based indicators vs. google trends," *Journal of Forecasting*, vol. 30, no. 6, pp. 565–578, 2011.

[32] N. Barreira, P. Godinho, and P. Melo, "Nowcasting unemployment rate and new car sales in south-western europe with google trends," *NETNOMICS: Economic Research and Electronic Networking*, vol. 14, no. 3, pp. 129–165, 2013.

[33] I. Önder and U. Gunter, "Forecasting tourism demand with google trends for a major european city destination," *Tourism Analysis*, vol. 21, no. 2-3, pp. 203–220, 2016.

[34] S. Park, J. Lee, and W. Song, "Short-term forecasting of japanese tourist inflow to south korea using google trends data," *Journal of Travel & Tourism Marketing*, vol. 34, no. 3, pp. 357–368, 2017.

[35] L. Kristoufek, "Bitcoin meets google trends and wikipedia: Quantifying the relationship between phenomena of the internet era," *Scientific reports*, vol. 3, p. 3415, 2013.

[36] A. Arratia and A. López-Barrantes, "Do google trends forecast bitcoins? stylized facts and statistical evidence," *Stylized Facts and Statistical Evidence (August 3, 2019). ITISE*, 2019.

[37] N. Smuts, "What drives cryptocurrency prices?: An investigation of google trends and telegram sentiment," *ACM SIGMETRICS Performance Evaluation Review*, vol. 46, no. 3, pp. 131–134, 2019.

[38] J. Rech, "Discovering trends in software engineering with google trend," *ACM SIGSOFT Software Engineering Notes*, vol. 32, no. 2, pp. 1–2, 2007.

[39] T. Dergiades, C. Milas, and T. Panagiotidis, "Tweets, google trends, and sovereign spreads in the giips," *Oxford Economic Papers*, vol. 67, no. 2, pp. 406–432, 2014.

[40] J. Mellon, "Internet search data and issue salience: The properties of google trends as a measure of issue salience," *Journal of Elections, Public Opinion & Parties*, vol. 24, no. 1, pp. 45–72, 2014.