Gazpacho and summer rash: lexical relationships from temporal patterns of web search queries

Enrique Alfonseca **Massimiliano Ciaramita** **Keith Hall**
Google
Zürich, Switzerland
ealfonseca@google.com, massi@google.com, kbhall@google.com

Abstract

In this paper we investigate temporal patterns of web search queries. We carry out several evaluations to analyze the properties of temporal profiles of queries, revealing promising semantic and pragmatic relationships between words. We focus on two applications: query suggestion and query categorization. The former shows a potential for time-series similarity measures to identify specific semantic relatedness between words, which results in state-of-the-art performance in query suggestion while providing complementary information to more traditional distributional similarity measures. The query categorization evaluation suggests that the temporal profile alone is not a strong indicator of broad topical categories.

1 Introduction

The temporal patterns of word occurrences in human communication carry an implicit measure of their relationship to real-world events and behavioral patterns. For example, when there is an event affecting a given entity (such as a natural disaster in a country), the entity name will turn up more frequently in human conversation, newswire articles and web documents; and people will search for it more often. Two entities that are closely related in the real world, such as the name of a country and a prominent region inside the country are likely to share common events and therefore be closely associated in human communication. Finally, two instances of the same class are also likely to share common usage patterns. For example, names of airlines or retail stores are more likely to be used by day rather than by night (Chien, 2005).

In this paper we explore the linguistic relationship between phrases that are judged to be similar based on their frequency time series correlation in search query logs. For every *phrase*[1] available in WordNet 3.0[2] (Miller, 1995), we have obtained its temporal signature from query logs, and calculated all their pairwise correlations. Next, we study the relationship in the top-ranked pairs with respect to their distribution in WordNet and a human-annotated labelling.

We also discuss possible applications of this data to solve open problems and present the results of two experiments: one where time series correlations turned out to be highly discriminative; and another where they were not particularly informative but shed some light on the nature of temporal semantics and topical categorization:

- *Query suggestion*, i.e. given a query, generate a ranked list of alternative queries in which the user may be interested.
- *Query categorization*, i.e. given a predefined set of categories, find the top categories to which the query can be assigned.

Finally, we illustrate with an example another application of time series in solving information extraction problems.

Although query logs are typically proprietary data, there are ongoing initiatives, like the Lemur toolbar[3], which make this kind of information available for research purposes. Other work (Bansal and Koudas, 2007b; Bansal and Koudas, 2007a) shows that temporal information can also be extracted from public data, such as blogs. More traditional types of text, such as news, are also typically associated with temporal labels; e.g., dates and timestamps.

This paper is structured in the following way:

[1] We use the term *phrase* to refer to any single word or multi-word expression that belongs to a synset in WordNet. Examples of phrases are *person*, *causal entity* or *william shakespeare*. We focused on the nouns hierarchy only.

[2] http://wordnet.princeton.edu

[3] http://www.lemurproject.org/querylogtoolbar/

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 1046–1055,
Singapore, 6-7 August 2009.

Section 2 summarizes the related work. Section 3 describes the correlation analysis between all pairs of phrases from WordNet. Next, Section 4 describes the application to query suggestion, and Section 5 the application to labelling queries in topical categories. Section 7 summarizes the conclusions and outlines ideas for future research.

2 Related work

The study of query time series explores a particular instance of the so-called *wisdom of the crowds* effect. Within this area, we can distinguish two kinds of phenomena. Knowledge and resources assembled by people explicitly, either individually, such as the case of blogs, or in a collaborative way, as in forums or wikis. These resources are valuable for human-consumption and can also be exploited in order to learn computational resources (Medelyan et al., 2008; Weld et al., 2008; Zesch et al., 2008b; Zesch et al., 2008a). On the other hand, it is possible to acquire useful resources and knowledge from aggregating behavioral patterns of large groups of people, even in the absence of a conscious effort. There is extensive ongoing research on the use of web search usage patterns to develop knowledge resources. Some examples are clustering co-click patterns to learn semantically related queries (Beeferman and Berger, 2000), combining co-click patterns with hitting times (Mei et al., 2008), analyzing query revisions made by users when querying search engines (Jones et al., 2006), replacing query words with other words that have the highest pointwise mutual information (Terra and Clarke, 2004), or using the temporal distribution of words in documents to improve ranking of search results (Jones and Diaz, 2007).

Within this second category, an important area is dedicated to the study of time-related features of search queries. News aggregators use real-time frequencies of user queries to detect spikes and identify news shortly after the spikes occur (Murata, 2008). Web users' query patterns have also proved useful for building a real-time surveillance system that accurately estimates region-by-region influenza activity with a lag of one day (Ginsberg et al., 2009). Search engines specifically developed for real-time searches, like Twitter search, will most likely provide new use cases and scenarios for quickly detecting trends in user search query patterns.

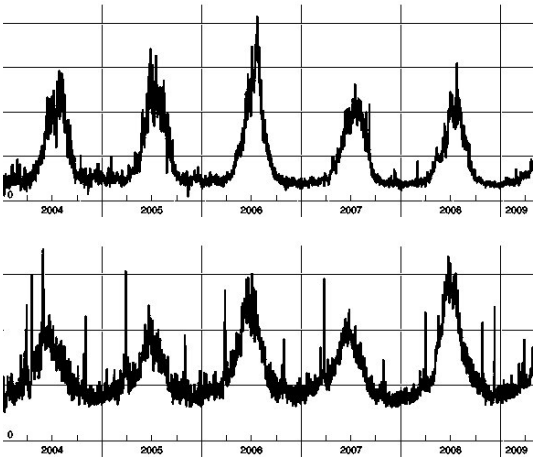

Figure 1: Time series obtained for the queries [gazpacho] and [summertime] (normalized scales).

Our study builds upon the work of Chien (2005), who observed that queries with highly-correlated temporal usage patterns are typically semantically related, and described a procedure for calculating the correlations efficiently. We have extended the analysis described in this work, by performing a more extensive evaluation of the kinds of semantic relationships that we can find among temporally-similar queries. We also propose, to our knowledge for the first time, areas of applications in solving well-established problems which shed some light on the nature of time-based semantic similarity. This work is also related to the analysis of temporal properties of information streams in data mining (Kleinberg, 2006) and information retrieval from time series databases (Agrawal et al., 1993).

3 Time-based similarities between phrases

Similarly to the method described in Chien (2005), we take a time interval, divide it into equally spaced subintervals, and represent each phrase of interest as the sequence of frequencies with which the phrase was observed in the subintervals. In our experiments, we have used as source data the set of fully anonymized query logs from the Google search engine between January 1st, 2004 and March 1st, 2009.[4].

These data have been aggregated on a daily basis so that we have the daily frequency of the

[4]Part of this data is publicly available from `http://www.google.com/trends`

queries of interest for over five years. The frequencies are then normalized with the total number of queries that happened on that day. The normalization is necessary to avoid daily and seasonal variations as there are typically more queries on weekdays than on weekends and fewer queries during holiday seasons than in the rest of the year. It also helps reducing the effect deriving from the fact that the population with Internet access is still monotonically growing, so we can expect that the number of queries will become higher and higher over time.

Given two phrases and their associated time series, the similarity metric used is the correlation coefficient between the two series (Chien, 2005). For illustration, Figure 1 shows the time series obtained for two sample queries, *gazpacho* and *summertime*, whose time series yield a correlation of 0.92. Similar high correlations can be observed with other queries related to phenomena that occur mainly in summer in the countries from which most queries come, like *summer rash*.

3.1 WordNet-based evaluation

In this section, we describe a study carried out with the purpose of discovering the traditional lexico-semantic relationships which hold between the queries that are most strongly related according to their temporal profiles.

For this evaluation, we have taken the nominal phrases appearing in WordNet 3.0. Given that users, when writing queries, typically do not pay attention to punctuation and case, we have normalized all phrases by lowercasing them and removing all punctuation. Next, we collected the time series for each phrase by computing the normalized daily frequency of each of them as exact queries in the query logs. The computation of the pairwise correlations was performed in parallel using the MapReduce infrastructure running over 2048 cores with 500 MB of RAM each. The total execution (including data shuffling and networking time) took approximately three hours.

Next, we represented the data as a complete graph where phrases are nodes and the edge between each pair of nodes is weighted by their time series correlation. Using a simple graph-cut we obtained clusters of related terms. A minimum weight threshold equal to 0.9 was applied;[5] thus, two phrases belong to the same cluster if there is a path between them only via edges with weight over 0.9.

The previous procedure produced a set of 604 clusters, with highly different sizes. The first observation is that 70% of the phrases in WordNet do not have a correlation over 0.9 with any other phrase, so they are placed alone in singleton clusters. There are several reasons for this. The clusters obtained are very specific: only phrases that have a very strong temporal association have temporal correlations exceeding the threshold. This is combined with the fact that we are using a very restricted vocabulary, namely the terms included in WordNet, which is many orders of magnitude smaller than the vocabulary of all possible queries from the users. Few phrase pairs in WordNet have a temporal association and popularity strong enough to be clustered together. Finally, many of the phrases in WordNet are rare, including scientific names of animals and plants, genuses or families, which are not commonly used. Therefore, the clusters extracted here correspond to very salient sets of phrases. If, instead of WordNet, we choose a vocabulary from known user queries (cf. Section 4), there would be many fewer singleton clusters, as the options of similar phrases to choose from would be much larger.

From the phrases that belong to clusters, 25% of the WordNet phrases do not have strong daily temporal profiles. The typical pattern for these terms is an almost flat time series, usually with small drops at summertime and Christmas (when seasonal leisure-related queries dominate). Therefore, these phrases were collected in just one cluster containing them all. Typical examples of the elements of this set are names of famous scientists and mathematicians (Gauss, Isaac Newton, Albert Einstein, Thomas Alva Edison, Hipprocrates, Gregor Mendel, ...), common terms (fertilization, famine, macroeconomics, genus, nationalism, ...), numbers and common first names, among other things. It is possible that using sub-day intervals might help to discriminate within this cluster.

The items in this big cluster contrast with periodical events, which display recurring patterns (e.g., queries related to elections or tax-returns), and names of famous people and other entities which appeared in the news in the past few years. All of these are associated with irregular, spiky time series. These constitute the final 5% of the

[5] This threshold is the same used by Chien (2005), and was confirmed after a manual inspection of a sample of the data

Type	Pairs	Examples
Synonyms	283	(angel cake, angel food cake), (thames, river thames), (armistice day, Nov 11)
Hyponym/hyperonyms	86	(howard hughes, aviator), (muhammad, prophet), (olga korbut, gymnast)
Siblings in hyponym taxonomy	611	(hiroshima, nagasaki), (junior school, primary school), (aids, welt)
Meronym/holonyms	53	(tutsi, rwanda), (july 4, july), (pyongyang, north korea)
Siblings in meronymy taxonomy	7	(everglades, everglades national park), (mississipi, orleans)
Other paths	471	(maundy thursday, maundy money), (tap water, water tap), (gren party, liberal)
Not structurally related	1009	(poppy, veterans day), (olympic games, gimnast), (belmont park, horse racing)

Table 1: Relationships between pairs of WordNet phrases belonging to the same cluster.

phrases belonging to small, highly focused, clusters.

Table 1 shows the relationships that hold between all pairs of phrases belonging to any of the smaller clusters. Out of 2520 pairs, 283 belong to the same synset, 697 are related via hyponymy links, 60 via meronymy links, and 471 by alternating hyponymy and meronymy links in the path. When the phrases were polysemous, the shortest path between any of their meaning was used. About 40% of the relations do not have a clear structural interpretation in WordNet.

The majority of pairs are related via more or less complex paths in the WordNet graph. Interestingly, even the structurally unrelated terms are characterized by transparent relations in terms of world knowledge, as it is the case between *poppy* and *veteran day*. Note as well that sometimes a WordNet term is used with a meaning not present in WordNet or in a different language, which may explain why *aids* has a very high correlation with *welt* (AIDS and welt are both hyponyms of *health problem*, but the correlation may be explained better by the AIDS World Day, *Welt Aids Tag* in German), and it also has a very high correlation with *sida*, defined in WordNet as a genus of tropical herbs, but which is in fact the translation of AIDS into Spanish. These observations motivated an additional manual labelling of the extracted pairs.

3.2 Hand labelled evaluation

As can be seen in Table 2, most of the terms that constitute a cluster are related to each other, although the kinds of semantic relationships that hold between them can vary significantly. Examples of the following kinds can be observed:

- **True synonyms**, as in the case of *november* and *nov*, or *architeuthis* and *giant squid*.
- **Variations of people names**, especially if a person's first name or surname is typically used to refer to that person, as in the case of *john lennon* and *lennon*, or *janis joplin* and *joplin*. Sometimes the variations include personal titles, as it is the case of *president carter* and *president nixon*, which are highly correlated with *jimmy carter* and *richard nixon*.
- **Geographically-related terms**, referring to locations which are located close to each other, as in the clusters {*korea, north korean, south korea, pyongyang, north korea*} and {*strasbourg, grenoble, toulouse, poitiers, lyon, lille, nantes, reims*}.
- **Synonyms of location names**, like *bahrain* and *bahrein*.
- **Derived words**, like *north korea* and *north korean*, or *lebanese* and *lebanon*.
- **Generic word optionalizations**, which happen when one word in a multi-word phrase is very correlated to the phrase, as in the case of *spanish inquisition* and *inquisition*, or *red bone marrow* and *red marrow*, where the most common interpretation for the shortened version of the phrase is the same as for the long version.
- **Word reordering**, where the two related phrases have the same words in a different order, as in the case of *maple sugar* and *sugar maple*, or *oil palm* and *palm oil*.
- **Morphological variants**: WordNet does not contain many morphological variants in the main dataset, but there are a few, like *station of the cross* and *stations of the cross*.
- **Acronyms**, like *federal emergency management agency* and *fema*.
- **Hyperonym-hyponym**, like *fern* and *plant*.
- **Sibling terms** in a taxonomy, as in the cluster {*lutheran, methodist, presbyterian, united methodist church, lutheran church,methodist church, presbyterian church,baptist, baptist church*}, which contains mostly names of Christian denominations.
- **Co-occurring events in time**, as is the case of *hitch* and *pacifier*, both titles of movies which were launched at almost the same

hydrant,fire hydrant
inauguration day,inauguration,swearing,investiture,inaugural address,inaugural,benediction,oath
indulgence,self indulgence
insulation,heating
interstate highway,interstate, intestine,small intestine
iq,iq test
irish people,irish,irish potato,irish gaelic,gaelic,irish soda bread,irish stew,st patrick,saint patrick,leprechaun, march 17,irish whiskey,shillelagh
ironsides,old ironsides
james,joyce,james joyce
janis joplin,joplin
jesus christ,pilate,pontius pilate,passion of christ,passion,aramaic
jewish new year,rosh hashana,rosh hashanah,shofar
john lennon,lennon
julep,mint julep,kentucky derby,kentucky
keynote,keynote address
kickoff,time off
korea,north korean,south korea,pyongyang,north korea
l ron hubbard,scientology
leap,leap year,leap day,february 29
left brain,right brain
leftover,leftovers,turkey stew
linseed oil,linseed
listeria,listeriosis,maple leaf
lobster tail,lobster,tails
lohan,lindsay
loire,rhone,rhone alpes
looking,looking for
lutheran,methodist,presbyterian,united methodist church,lutheran church,methodist church,presbyterian church, baptist,baptist church
mahatma gandhi,mahatma
malignant hyperthermia,hyperthermia
maple sugar,sugar maple
martin luther,martin luther king,luther,martin,martin luther king day
matzo,matzah,matzoh,passover,seder,matzo meal,pesach,haggadah,gefilte fish
mestizo,half blood,half and half
meteorology,weather bureau
moslem,muslim,prophet,mohammed,mohammad,muhammad,mahomet
movie star,star,revenge,film star,menace,george lucas
mt st helens,mount saint helens,mount st helens
myeloma,multiple myeloma
ness,loch ness,loch ness monster,loch,nessie
new guinea,papua new guinea,papua
november,nov
pacifier,hitch
papa,pope,vatican,vatican city,karol wojtyla,john paul ii,holy see,pius xii,papacy,paul vi,john xxiii,the holy see, vatican ii,pontiff,gulp,pater,nostradamus,ii,pontifex
parietal lobe,glioma,malignant tumor
particle accelerator,atom smasher,hadron,large,tallulah bankhead,bankhead,tanner
pledge,allegiance
president carter,jimmy carter
president nixon,richard nixon,richard m nixon
sept 11,september 11,sep 11,twin towers,wtc,ground zero,world trade center
slum,millionaire,pinto
strasbourg,grenoble,toulouse,poitiers,lyon,lille,nantes,reims
valentine,valentine day,february 14,romantic
aeon,flux
alien,predator
anne hathaway,hathaway
architeuthis,giant squid
basal temperature,basal body temperature
execution,saddam hussein,hussein,saddam,hanging,husain
flood,flooding
george herbert walker bush,george walker bush
intifada,palestine
may 1,may day,maypole

Table 2: Sample of clusters obtained from the temporal correlations.

Type	Clusters
True synonyms	19
Variations of people names	42
People names with and without titles	4
First name and surname from the same person	4
Geographically-related terms	18
Synonyms of location names	4
Derived words	4
Word optionalizations	87
Word reordering	7
Morphological variants	1
Acronyms	1
Cross-language synonyms	3
Hyperonym/hyponym	10
Sibling terms	10
Co-ocurring events in time	8
Topically related	38
Unrelated	72

Table 3: Results of the manual annotation of 2-item clusters.

time. A particular example of this is when the two terms are part of a named entity, as in the case of *quantum* and *solace*, which have a similar correlation because they appear together in a movie title.

- **Topically-related terms**, as the cluster {*jesus christ, pilate, pontius pilate, passion of christ, passion, aramaic*}, or the cluster containing popes and the Vatican. A similar example, *execution* is highly correlated to *saddam hussein*, because his execution attracted more interest worldwide during this time period than any other execution. Interestingly, topical correlation emerges at very specific granularity.

For the manual analysis of the results, we randomly selected 332 clusters containing only two items (so that 664 phrases were considered in total). Each of these pairs has been classified in one of the previous categories. The results of this analysis are shown in Table 3.

4 Application to query suggestion

Query suggestion is a feature of search engines that helps users reformulate queries in order to better describe their information need with the purpose of reducing the time needed to find the desired information (Beeferman and Berger, 2000; Kraft and Zien, 2004; Sahami and Heilman, 2006; Cucerzan and White, 2007; Yih and Meek, 2008). In this section, we explore the application of a similarity metric based on time series correlations for finding related queries to suggest to the users.

As a test set, we have used the query suggestion dataset from (Alfonseca et al., 2009). It contains a set of 57 queries and an average of 22 candidate query suggestions for each of them. Each suggestion was rated by two human raters using the 5-point Likert scale defined in (Sahami and Heilman, 2006), from irrelevant to highly relevant. The task involves providing a ranking of the suggestions that most closely resembles the human scores. The evaluation is based on standard IR metrics: precision at 1, 3 and 5, and mean average precision. In order to compute the precision- and recall-based metrics, we infer a binary distinction from the ratings: related or not related. The inter-annotator agreement for this dataset given the binary classification as computed by Cohen's Kappa is 0.6171.

Method	P@1	P@3	P@5	mAP
Random	0.37	0.37	0.37	0.43
Web Kernel	0.51	0.47	0.42	0.51
Dist. simil.	0.72	0.63	0.60	0.64
Time series	0.74	0.63	0.53	0.67
Combination	0.79	0.68	0.60	0.69

Table 4: Results for the query suggestion task.

We used three baselines: the average values that would be produced by a random scorer of the candidate suggestions, Sahami and Heilman (2006)'s system (based on calculating similarities between the retrieved snippets), and a recent competitive ranker based on calculating standard distributional similarities (Alfonseca et al., 2009) between the original query and the suggestion. Please refer to the referenced work for details.

In order to produce the ranked lists of candidate suggestions for each query, due to the lack of training data, we have opted for the unsupervised procedure described in the previous section:

1. Collect the daily time series of each of the queries and the candidate suggestions.
2. Calculate the correlation between the original query and each of the candidate suggestions provided for it, and use it as the candidate's score.
3. For each query, rank its candidate suggestions in decreasing order of correlation.

Finally, taking into account that the source of similarity is very different to the one used for distributional similarity, we tested the hypothesis that

a combination of the two techniques would be beneficial to capture different features of the queries and suggestions. We have trained a linear mixture model combining both scores (time series and distributional similarities), using 10-fold cross validation.

The results are displayed in Table 4. For evaluating the results, whenever a system produced a tie between several suggestions, we generated 100 random orderings of the elements in the tie, and report the average scores.

Using distributional similarities and the temporal series turned out to be indistinguishable for the precision scores at 0.95 confidence, and both are significantly better than the similarity metric based on the web kernel. The combination produced an improvement across all metrics, although not statistically significant at p=0.05.

This is quite a positive finding as the time series method relies on stored information requiring only simple and highly optimized lookups.

5 Application to query categorization

The results from the manual evaluation in Section 3.2 support the conclusion that time series from query logs provide powerful signals for clustering at a fine-grained level, in some cases uncovering synonyms (may 1st, may day) and even causal relations (insulation, heating). A natural question is if temporal information is correlated with other types of categorizations. In this section we carry out a preliminary exploration of the relation between query time series and query categorization. To this extent we adapt the data from the KDD 2005 CUP (Li et al., 2005), which provides a set of queries classified into 67 broad topical categories. Since the data is rather sparse (678 queries) we applied Fourier analysis to "smooth" the time series.

5.1 The KDD CUP data

The KDD Cup 2005 [6] introduced a query categorization task and dataset consisting of 800,000 unlabeled queries for unsupervised training, and an evaluation set of 911 queries, 111 for development and 800 for the final evaluation. The systems submitted for this task can be quite complex and made full use of the large unlabeled set. Our goal here is not to provide a comparative evaluation, but only

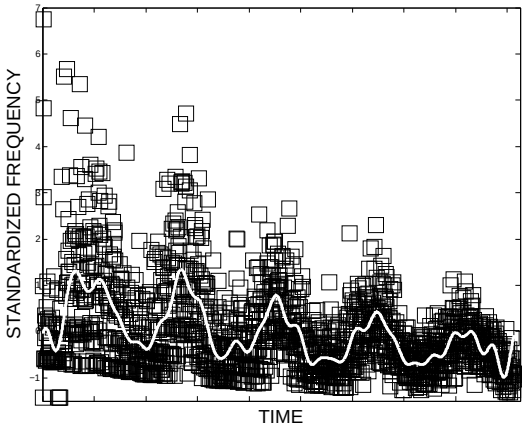

Figure 2: RDFT reconstruction for the query "brush cutters" using the first 25 Fourier coefficients. The squares represent the original time series datapoints, while the continuous line represents the reconstructed signal.

to use the labelled data[7] in a simplified manner to better understand the semantic properties of query time series. Each query in the dataset is assessed by three editors who can assign multiple topic labels from a set of 67 categories belonging to seven broad topics: Computers, Entertainment, Information, Living, Online Community, Shopping and Sports. We merged the KDD Cup development and test set, out of the 911 queries we were able to retrieve significant temporal information for 678 queries. We joined the sets of labels from each assessor for each query. On average, each query is assigned five labels.

5.2 DFT analysis

Assessing the similarity of data represented as time series has been addressed mostly my means of Fourier analysis; e.g., Agrawal et al. (1993) introduce a method for efficiently retrieving time series from databases based on Discrete Fourier Transform (DFT). Several other methods have been proposed, e.g., Discrete Wavelet Transform (DWT), however DFT provide a competitive benchmark approach (Wu et al., 2000).

We use DFT to generate the Fourier coefficients of the time series and Reverse DFT (RDFT) to reconstruct the original signal using only a subset of the coefficients. This analysis effectively compresses the time series producing a smoother approximate representation. DFT can be computed efficiently via Fast Fourier Transform (FFT), with

[6] http://www.sigkdd.org/kdd2005/kddcup.html

[7] The KDD Cup dataset is probably the only public query log providing topical categorization information.

Method	Accuracy	± std-err
Random	0.107	0.03
MostFrequent	0.490	0.07
DFT-c10	0.425	0.06
DFT-c50	0.456	0.05
DFT-c100	0.502	0.05
DFT-c200	0.456	0.04
DFT-c400	0.506	0.05
DFT-c600	0.481	0.06
DFT-c800	0.478	0.04
DFT-c1000	0.466	0.05

Table 5: Results of the KDD dataset exploration.

complexity $O(n \log n)$ where n is the length of the sequence. The approximate representation is useful not only to address sparsity but can also be used to efficiently estimate the similarity of two time series using only a small subset of coefficients as in (Agrawal et al., 1993). As an example, Figure 2 shows the original time series for the query "brush cutters" and its reconstructed signal using only the first 25 Fourier coefficients. The reconstructed signal captures the essence of the periodicity of the query and highlights the yearly peaks registered for the query in spring and summer.

5.3 Experiment and discussion

To explore the correlation between the structured temporal representation of queries provided by the time series and topical categorization we run the following experiment. Each KDD Cup query was reconstructed via RDFT using a variable number of coefficients. The set of 679 queries was partitioned in 10 sets and a 10-fold evaluation was performed. For each fold we trained a classifier on the remaining 9 folds. We used an average multi-class perceptron (Freund and Schapire, 1999) adapted to multi-label learning (Crammer and Singer, 2003). Each model was trained on a fixed number of 10 iterations. The accuracy of each model was evaluated as the fraction of test items for which the selected highest scoring class was in the gold standard set provided by the editors. As a lower bound we estimated the accuracy of randomly choosing a label for each test instance, and as a baseline we used the most frequent label. The latter is a powerful predictor: baselines based on class frequency outperform most of the systems that participated in the KDD Cup (Lin and Wu, 2009).

Table 5 reports the average accuracy over the 10 runs with relative standard errors. Each DFT-based model is characterized by the number of coefficients used for the reconstruction. Two main patterns are noticeable. First, none of the differences between the frequency-based baseline and the DFT models is significant, this seems to indicate that temporal structure alone is not a good discriminator of topic, at least of broad categories. In retrospect, this is somewhat predictable. The temporal dimension is a basic semantic component of lexical meaning and world knowledge which is not necessarily associated with any broad, and to some extent subjective, categorization. An inspection of the patterns found in each category shows in fact that similar patterns often emerge in different categories; e.g., "Halloween costume" and "cheesecake recipe" have a similar yearly periodical pattern with spikes in early winter, while monotonically decaying patterns are shared across all categories; e.g., between computer hardware and kids toys.

The second interesting finding is the trend of the DFT system results, higher at low-intermediate values, providing some initial promising evidence that DFT analysis generates useful compressed representations which could be indexed and applied efficiently. Notice that the sequences reconstructed using 1,000 coefficients reproduce almost identically the original signals.

6 Applications in information extraction

Time series from query logs are particularly relevant for phrases that refer to entities which are involved in recent events. Therefore, we expect them to be useful for solving other applications that require handling entities, such as named entity recognition and classification, relation extraction or disambiguation.

To illustrate this point, we mention an example of relation extraction between actors and movies: movies usually have spikes when they are released, and then the frequency again drops sharply. At the same times, when a movie is released, the search engine users have a renewed interest in their actors. Figure 3 displays the time series for the five most recent movies by Jim Carrey (as of march 2009), and the time series for Jim Carrey. As can be seen, the spikes are at exactly the same points in time. If we add up the series (a) through (e) into a single series and calculate the correlation with (f), it turns out to be very high (0.88).

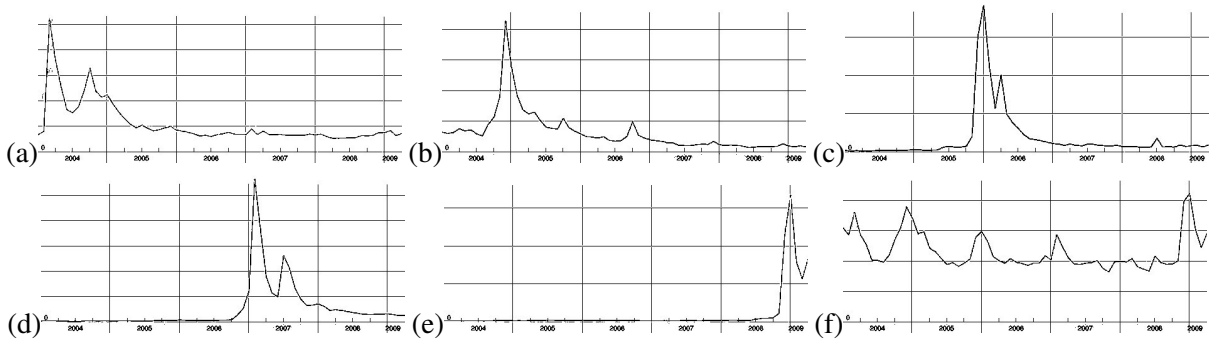

Figure 3: Time series obtained for the five most recent movies with Jim Carrey, and (f) time serie for the query [jim carrey] (normalized scales).

System	Precision	Recall	F-measure
Random	0.24	0.14	0.17
Time series	0.53	0.66	0.57

Table 6: Results for the query suggestion task.

To validate the hypothesis that this data should be useful for identifying related entities, we have performed a small experiment in the following way: by choosing five popular actors[8] and the cinema movies in which they appear since the year 2004, obtained from IMDB[9]. Using the time series, for each actor we choose the combination of movies such that, by adding up the time series of those movies, we maximise the correlation with the actor's time series. It has been implemented with a greedy beam search, with a beam size of 100. The results are shown in Table 6. The random baseline randomly associates the movies from the dataset with the five actors.

We do not believe this to be a perfect feature as, for example, actors may have a peak in the time series related to their personal lives, not necessarily to movies. However, the high correlations that can be obtained when the pairing between actors and movies is correct, and the improvement with respect a random baseline, indicates this is a feature which can probably be integrated with other relation extraction systems when handling relationships between entities that have big temporal dependencies.

7 Conclusions and future work

This paper explores the relationships between queries whose associated time series obtained from query logs are highly correlated. The use of time series in semantic similarity has been discussed by Chien (2005), but only a very preliminary evaluation was described, and, to our knowledge, they had never been applied and evaluated in solving existing problems. Our results indicate that, for a substantial percentage of phrases in a thesaurus, it is possible to find other highly-related phrases; and we have categorized the kind of semantic relationships that hold between them.

We have found that in a query suggestion task, somewhat surprisingly, results are comparable with other state-of-the-art techniques based on distributional similarities. Furthermore, information obtained from time series seems to be complementary with them, as a simple combination of similarity metrics produces an important increase in performance..

From an analysis on a query categorization task the initial evidence suggests that there is no strong correlation between broad topics and temporal profiles. This agrees with the intuition that time provides a fundamental semantic dimension possibly orthogonal to broad topical classification. This issue however deserves further investigation. Another issue which is worth a deeper investigation is the application of Fourier transform methods which offer tools for studying the periodic structure of the temporal sequences.

[8] Ben Stiller, Edward Norton, Jim Carrey, Leonardo Dicaprio, and Tom Hanks.

[9] www.imdb.com.

References

R. Agrawal, C. Faloutsos, and A.N. Swami. 1993. Efficient similarity search in sequence databases. In *Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms*, pages 69–84.

E. Alfonseca, K. Hall, and S. Hartmann. 2009. Large-scale computation of distributional similarities for queries. In *Proceedings of North American Chapter of the Association for Computational Linguistics - Human Language Technologies conference*.

N. Bansal and N. Koudas. 2007a. BlogScope: a system for online analysis of high volume text streams. In *Proceedings of the 33rd international conference on Very large data bases*, pages 1410–1413.

N. Bansal and N. Koudas. 2007b. BlogScope: Spatio-temporal analysis of the blogosphere. In *Proceedings of the 16th international conference on World Wide Web*, pages 1269–1270.

D. Beeferman and A. Berger. 2000. Agglomerative clustering of a search engine query log. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 407–416.

S. Chien. 2005. Semantic similarity between search engine queries using temporal correlation. In *Proceedings of the 14th international conference on World Wide Web*, pages 2–11.

K. Crammer and Y. Singer. 2003. Ultraconservative online algorithms for multiclass problems. *Journal of Machine Learning Research*, 3:951–991.

S. Cucerzan and R.W. White. 2007. Query suggestion based on user landing pages. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 875–876.

Y. Freund and R.E. Schapire. 1999. Large margin classification using the perceptron algorithm. *Machine Learning*, 37:277–296.

J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, and L. Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature*, 457, February.

R. Jones and F. Diaz. 2007. Temporal profiles of queries. *ACM Transactions on Information Systems*, 25(3):14.

R. Jones, B. Rey, O. Madani, and W. Greiner. 2006. Generating query substitutions. In *Proceedings of the 15th international conference on World Wide Web*, pages 387–396.

J. Kleinberg. 2006. Temporal dynamics of on-line information streams. In *Data Stream Management: Processing High-Speed Data*. Springer.

R. Kraft and J. Zien. 2004. Mining anchor text for query refinement. In *Proceedings of the 13th international conference on World Wide Web*, pages 666–674.

Y. Li, Z. Zheng, and H. Dai. 2005. KDD Cup-2005 report: Facing a grat challenge. *SIGKDD Explor. Newsl.*, 7(2):91–99.

D. Lin and X. Wu. 2009. Phrase clustering for discriminative learning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*.

O. Medelyan, C. Legg, D. Milne, and I.H. Witten. 2008. *Mining meaning from Wikipedia*. Dept. of Computer Science, University of Waikato.

Q. Mei, D. Zhou, and K. Church. 2008. Query suggestion using hitting time. In *Proceeding of the 17th ACM conference on Information and knowledge management*, pages 469–478.

G.A. Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

T. Murata. 2008. Detection of breaking news from online web search queries. *New Generation Computing*, 26(1):63–73.

M. Sahami and T.D. Heilman. 2006. A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th international conference on World Wide Web*, pages 377–386.

E. Terra and C.L.A. Clarke. 2004. Scoring missing terms in information retrieval tasks. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 50–58.

D.S. Weld, F. Wu, E. Adar, S. Amershi, J. Fogarty, R. Hoffmann, K. Patel, and M. Skinner. 2008. Intelligence in Wikipedia. In *Proceedings of the 23rd Conference on Artificial Intelligence*.

Y. Wu, D. Agrawal, and A. El Abbadi. 2000. A comparison of DFT and DWT based similarity search in time-series databases. In *Proceedings of the 9th International ACM Conference on Information and Knowledge Management*, pages 488–495.

W. Yih and C. Meek. 2008. Consistent Phrase Relevance Measures. *Workshop on Data Mining and Audience Intelligence for Advertising*, page 37.

T. Zesch, C. Muller, and I. Gurevych. 2008a. Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In *Proceedings of the Conference on Language Resources and Evaluation*.

T. Zesch, C. Muller, and I. Gurevych. 2008b. Using Wiktionary for computing semantic relatedness. In *Proceedings of the Conference on Artificial Intelligence*, pages 861–867.