Seasonal Web Search Query Selection for Influenza-Like Illness (ILI) Estimation

Niels Dalum Hansen
University of Copenhagen / IBM Denmark
nhansen@di.ku.dk

Kåre Mølbak
Statens Serum Institut, Denmark
krm@ssi.dk

Ingemar J. Cox
University of Copenhagen, Denmark
ingemar.cox@di.ku.dk

Christina Lioma
University of Copenhagen, Denmark
c.lioma@di.ku.dk

ABSTRACT

Influenza-like illness (ILI) estimation from web search data is an important web analytics task. The basic idea is to use the frequencies of queries in web search logs that are correlated with past ILI activity as features when estimating current ILI activity. It has been noted that since influenza is seasonal, this approach can lead to spurious correlations with features/queries that also exhibit seasonality, but have no relationship with ILI. Spurious correlations can, in turn, degrade performance. To address this issue, we propose modeling the seasonal variation in ILI activity and selecting queries that are correlated with the residual of the seasonal model and the observed ILI signal. Experimental results show that re-ranking queries obtained by Google Correlate based on their correlation with the residual strongly favours ILI-related queries.

ACM Reference format:
Niels Dalum Hansen, Kåre Mølbak, Ingemar J. Cox, and Christina Lioma. 2017. Seasonal Web Search Query Selection for Influenza-Like Illness (ILI) Estimation. In *Proceedings of SIGIR'17, August 7-11, 2017, Shinjuku, Tokyo, Japan, ,* 4 pages.
DOI: http://dx.doi.org/10.1145/3077136.3080760

1 INTRODUCTION AND BACKGROUND

The frequency of queries in web search logs has been found useful in estimating the incidence of *influenza-like illnesses* (ILIs) [3, 6, 8, 9, 11]. Current methods use two core discriminative features for ILI estimation: (i) past ILI activity, and (ii) the frequency of queries in web search logs that correlate strongly with past ILI activity. There are two problems with this approach.

The *first problem* is that not all queries whose frequency is strongly correlated to ILI activity are necessarily informative with respect to ILIs, and hence discriminative as a feature for ILI estimation. At the most general level, this is an example of the fundamental issue of correlation not being causation. In the case of estimating ILI, this is exacerbated by the seasonal nature of influenza. In fact, it has been previously observed that previous methods can identify queries that have a very similar seasonality but are clearly not related to ILI. For example, the query *"high school basketball"* has been found to have a high correlation with ILI activity [8] even though it is obviously unrelated to ILI. The seasonality of high school basketball accounts for this correlation. Queries unrelated to the ILI activity will not be useful in the case of irregular ILI activity, e.g. an off season influenza outbreak. Additionally, changes in for example the high school basketball schedule would result in changes in the ILI estimates.

The *second problem* is that by using two types of features that are strongly correlated to each other (past ILI activity, and queries whose frequencies are strongly correlated to past ILI activity), we may compromise diversity in the representations one would expect from the features. Better estimations may be produced by using features that *complement* each other, regardless of their between–feature correlation.

Motivated by the above issues, we propose an alternative approach to selecting queries. Our approach consists of two steps. (1) We model the seasonal variation of ILI activity, and (2) we select queries whose search frequency fits aspects of this seasonality. Specifically, we present two variations of our algorithm: select queries that correlate with our seasonal model of ILI, and select queries that correlate with the residual between the seasonal model and observed ILI rates.

Our results are two fold. (i) Experimental evaluation of our seasonal query selection models for ILI estimation against strong recent baselines (of no seasonality) show that we can achieve performance that is overall more effective (reduced estimation error), and requires fewer queries as estimation features. With respect to error reduction we see that selecting queries fitting regular seasonal ILI variation is a better strategy than selecting queries fitting ILI outbreaks. (ii) Selecting queries that fit seasonal irregularities result in much more semantically relevant queries. These queries are surprisingly not the ones that result in the best predictions.

Our main results are: (i) We demonstrate that Google Correlate retrieves many non-relevant queries that are highly correlated with a times series of historic ILI incidence, and that the ILI-related queries are not highly ranked; (ii) re-ranking these queries based on their correlation with a residual signal, i.e. the difference between a seasonal model and historic data, strongly favours ILI-related queries; (iii) the performance of a linear estimator is improved based on the re-ranked queries. To our knowledge, the seasonal examination of ILI activity for query selection in automatic ILI estimation is a novel contribution. Seasonal variation has, however, been studied for other medical informatics tasks, such as vaccination uptake estimation [4, 5].

2 PROBLEM STATEMENT

The goal is to estimate ILI activity at time t, denoted y_t, using observed historical ILI activity (reported e.g. by the Centers for Disease Control and Prevention (CDC) in US) and query frequencies in web search logs. This is most commonly done by (i) submitting to Google

SIGIR'17, August 7-11, 2017, Shinjuku, Tokyo, Japan
 978-1-4503-5022-8/17/08...$15.00
DOI: http://dx.doi.org/10.1145/3077136.3080760

Correlate[1] a file of historical ILI activity, and receiving as output web search queries (and their frequencies) that are most strongly correlated to the input ILI data. Then, y_t can be estimated with a linear model that uses only web search frequencies [3] as follows:

$$y_t = \alpha_0 + \sum_{i=1}^{n} \alpha_i Q_{t,i} + \epsilon, \tag{1}$$

where n is the number of queries, $Q_{t,i}$ is the frequency for query i in the web search log at time t, the αs are coefficients, and ϵ is the estimation error.

Including historical ILI activity data can improve the estimations of Eq. 1, for instance with an autoregressive model [11], as follows:

$$y_t = \beta_0 + \beta_1 t + \sum_{j=1}^{m} \beta_{j+1} \cdot y_{t-j} + \sum_{i=1}^{n} \beta_{i+m+1} \cdot Q_{t,i} + \epsilon, \tag{2}$$

where m is the number of autoregressive terms, and the βs are coefficients to be learned. With $m = 52$ and $n = 100$, Eq. 2 corresponds to the model presented by Yang *et al.* [11].

Most ILI estimation methods (exceptions include [7]) that use web search frequencies use all queries found to be correlated to ILI activity, i.e. in Eq. 1 and Eq. 2 n corresponds to *all* strongly correlated queries, and query selection is typically left for the model regularisation, for example using lasso regularisation. In the next section we present a novel way of selecting which among these correlated queries to include in the estimation of y_t according to how well they fit the seasonal variation of ILI activity.

3 SEASONAL QUERY SELECTION

We reason that among the queries whose frequency is correlated with past ILI activity, some queries may fit the ILI seasonal variation better than others. This is supported by the literature [8]. We further reason that this fit of queries to seasonal ILI variation may not be sufficiently captured by simply measuring the correlation between the frequency of those queries and ILI activity. Based on this, we (i) present two models to represent seasonal variation of ILI activity, and (ii) select queries based on these seasonal models.

3.1 Step 1: Model seasonal ILI variation

We model seasonal variation in two ways. The first model is the Serfling model [10], chosen because of its simplicity and expressiveness. The Serfling model (Eq. 3) uses pairs of sine and cosine terms to model seasonality, and a term linear in time to model general upward or downward trends. We use this model with weekly data and one yearly cycle (details on data are given in Section 4), resulting in the following ILI estimation model:

$$y_t = \beta_0 + \beta_1 t + \beta_2 \sin\left(\frac{2\pi t}{52}\right) + \beta_3 \cos\left(\frac{2\pi t}{52}\right) + \epsilon, \tag{3}$$

where the βs denote model coefficients and ϵ the error, i.e. residual.

For the second model we use a yearly average (YA). Here the expected value of y_t is calculated as the average value of N seasons of ILI activity data,

$$\hat{y}_t = \frac{1}{N} \sum_{i=0}^{N-1} y_{(t \bmod S) + i \cdot S}, \tag{4}$$

where S is the season length in weeks, in our case 52.

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

Figure 1: Fit of the Serfling model (Eq. 3) and the Yearly Average model (Eq. 4) to historical ILI data (described in Section 4).

Fig. 1 shows the fit of the two models, i.e. the Serfling model (Eq. 3) and the YA model (Eq. 4) to historical ILI activity data. We see that the Serfling model fits the general seasonality, but not more complex patterns representing higher transmission. It does not, for example, model differences between the start and end of the ILI season. This is better captured by the YA model.

3.2 Step 2: Query selection

Having modelled seasonal ILI variation, the second step is to approximate how well queries fit the seasonal variation of ILI activities modelled by Eq. 3-4. We do this in two ways:

Seasonal correlation. We compute the Pearson correlation between the query frequencies and the ILI seasonal model, i.e. Eq. 3 or 4. We then select queries that are most strongly correlated to the ILI activity model.

Residual correlation. We compute the Pearson correlation between the query frequencies and the residual between the ILI seasonal model and the historical ILI activity. We then select queries that are most strongly correlated to the residual, i.e. *unexpected variations in ILI activity* (possible outbreaks).

The four query selection methods are denoted (i) Seasonal (Serfling), (ii) Seasonal (YA), (iii) Residual (Serfling), and (iv) Residual (YA).

4 EVALUATION

Experimental Setup. We experimentally evaluate our seasonality-based query selection methods using two types of data: weekly ILI activity data and Google search frequency data. The ILI activity data is from the US CDC for the period 2004-6-6 to 2015-7-11[2] (inclusive). The CDC reports this in the form of weighted ILI activity across different US regions. ILI activity for a region corresponds to the number of ILI related visits to health care providers compared to non-influenza weeks, e.g. an ILI activity of 2 corresponds to twice as many visits as in non-epidemic weeks.

We retrieve query search frequencies from Google Correlate with the location set to the US. Specifically, we use the 100[3] queries that

[2] https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html
[3] This is a maximum number set by Google Correlate.

have the highest correlation with the ILI activity from 2004-6-6 to 2009-3-29 according to Google Correlate. Google normalizes the search frequencies for each query to unit variance and zero mean, i.e. we do not know the actual search frequencies. We use the interval 2004-6-6 to 2009-3-29 because it represents a non-epidemic period (it excludes the 2009 pandemic of H1N1 influenza virus that caused highly irregular ILI activity). The 100 queries are shown in Tab. 1. Only 21 of the 100 queries are related to ILI (in bold).

We compare our query selection methods (Section 3) to the following three baselines: (Tab. 2 baseline i) uses the top-c queries to estimate ILI activity, where the top-c are chosen to minimise the RMSE error, i.e. if we use $c+1$ queries the RMSE increases; (Tab. 2 baseline ii) using ***all*** 100 queries to estimate ILI activity; (Tab. 2 baseline iii) using no queries, only past ILI activity, i.e. an autoregressive model. For (i) and (ii), the query ranking is determined by Google Correlate. For (iii), two autoregressive models are fitted: one using 3 autoregressive terms [8, 11] and one with 52 autoregressive terms [11]. This setup is similar to the setup of Yang *et al.* [11]. We implement baseline (iii) using Eq. 2 where m is set to 3 and 52 terms, respectively, and $n=0$. Similarly to [8, 11], we evaluate estimation performance by reporting the root mean squared error (RMSE) and Pearson correlation between the estimations and the observed historical ILI activity.

For all runs, we use data from 2004-6-1 to 2009-3-29 for training, and data from 2009-4-1 to 2015-7-11 for testing. The training data is used to fit Eq. 3-4, and to calculate the correlation scores as described in Section 3.2. Estimations are made in a leave-one-out fashion where data prior to the data point being estimated is used to fit the estimation model. Each model is retrained for every time step using the 104 most recent data points (exactly as in Yang *et al.* [11]). We determine the number of queries n in Eq. 1-3 by iteratively adding the next highest ranked query, where query rank is given by either Google Correlate (for the baselines), or by the four variants of our algorithm, specifically (i) correlate seasonal (Serfling), (ii) correlate seasonal (YA), (iii) correlate residual (Serfling), and (iv) correlate residual (YA). The models are fitted using lasso regression, where the hyper-parameter is found using three fold cross-validation on the training set.

Results. As noted earlier, Google Correlate identifies the top-100 queries, but only 21 of these are ILI-related. Our four algorithms re-rank the 100 queries. Fig. 2 plots the number of ILI-related queries as a function of the number of rank-ordered queries. The solid curve is based on the original ranking provided by Google Correlate. We observe that both the (i) Seasonal (Serfling) and (ii) Seasonal (YA), re-rank the queries such that, in general, the ILI-related queries are ranked worse. The Residual (Serfling) generally performs similarly to or worse than Google Correlate in favouring ILI-related queries. In contrast, Residual (YA) re-ranks the queries such that almost all ILI-related queries are favoured. Of the top-21 queries, 19 are ILI-related. All 21 ILI-related queries are within the top-23. The only two non-related queries in the top-23 are ranked at 19 and 21. Clearly re-ranking queries based on Residual (YA) strongly favours ILI-related queries much more than Google Correlate or our other three variants.

For each ranking of the queries, we select the top-n queries that either minimise the RMSE or maximise the Pearson correlation. This is done for the Linear model of Eq. 1 and for the autoregressive model of Eq. 2, Tab. 2 shows the results. For the Linear model (column 1), we observe that Residual (YA) performs best w.r.t. RMSE and Pearson correlation, though the latter is not significant. Note that in both

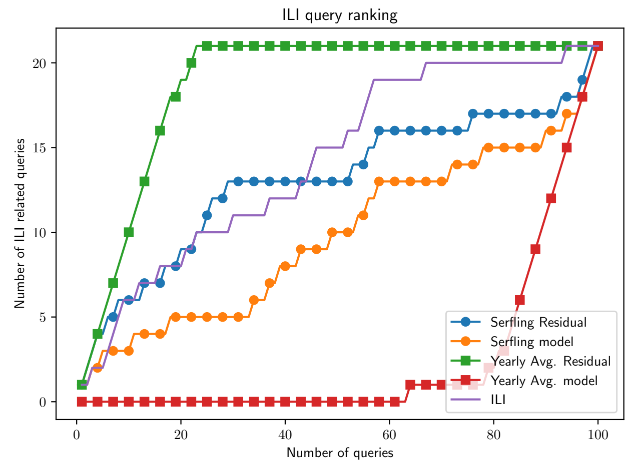

Figure 2: Portion of ILI related queries in the top n queries for each of the five ranking methods.

cases, (i) the number of queries needed by Residual (YA) is significantly less than for the other three variants and (ii) the two baselines performed worse.

For the autoregressive models, we observe that the Seasonal (Serfling) model performs best w.r.t. RMSE and Pearson correlation. This is achieved with relatively few queries (5, 9, or 11). However we note that of the top-5, -9 or -11 queries only 3, 3 or 4, resp. are ILI-related. In general, autoregressive models perform well when the signal has a strong autocorrelation. However, should the signal strongly deviate from seasonal patterns, then it is unlikely that the ILI estimates would be accurate.

5 CONCLUSION

The incidence of influenza-like illness (ILI) exhibits strong seasonal variations. These seasonal variations cause Google Correlate to identify a large number of non-relevant queries (80%). Many of the relevant queries are not highly ranked. Estimating the incidence of ILI with non-relevant queries is likely to become problematic when ILI deviates significantly from its seasonal variation.

We proposed a new approach to ILI estimation using web search queries. The novelty of our approach consists of re-ranking queries derived from Google Correlate. We first developed two models of the seasonal variation in ILI. The first is an analytical Serfling model. The second is an empirical model based on yearly averages. Four methods of re-ranking queries were then examined. The first two re-rank the queries based on their correlation with the two seasonal models. The second two re-rank queries based on their correlation with the residual between the seasonal models and historical ILI activity.

Experimental results showed that re-ranking queries based on Residual (YA) strongly favoured ILI-related queries, but re-ranking queries based on the two seasonal models, Seasonal (Serfling) and Seasonal (YA) led to rankings that were worse than those of Google Correlate.

When ILI estimates were based on both queries and autoregression the best performance was obtained when queries were re-ranked based on Seasonal (Serfling). Future work is needed to determine why, but we reason that (i) autoregessive models perform better florida spring, **influenza symptoms**, **symptoms of influenza**, new york yankees spring training, yankees spring training, **flu incubation**, **flu incubation period**, **flu duration**, florida spring training, spring training locations, **influenza incubation**, florida spring break, spring break dates, **flu fever**, sox spring training, new york mets spring training, **bronchitis**, red sox spring, spring training in florida, snow goose, spring break calendar, spring training in arizona, red sox spring training, mlb spring training, **flu report**, baseball spring training, mariners spring training, wrestling tournaments, spring training, golf dome, **flu recovery**, city spring, wrestling singlets, spring training sites, boys basketball, **type a influenza**, yankee spring training, spring training tickets, las vegas march, indoor track, harlem globe, spring break panama city, girls basketball, panama city spring break, cardinals spring training, ny mets spring training, ny yankees spring training, **flu symptoms**, minnesota twins spring training, concerts in march, spring training map, **tessalon**, boston red sox spring training, **flu contagious**, **symptoms of the flu**, events in march, seattle mariners spring training, singlets, **influenza contagious**, **influenza incubation period**, spring break schedule, spring vacation, **treating the flu**, college spring break, basketball boys, college spring break dates, boys basketball team, **respiratory flu**, atlanta braves spring training, **acute bronchitis**, march madness dates, spring break florida, braves spring training, college basketball standings, in march, braves spring training schedule, high school boys basketball, spring break ideas, spring break miami, banff film, addy awards, grapefruit league, spring clothing, spring collection, banff film festival, st. louis cardinals spring training, april weather, spring break family, red sox spring training schedule, miami spring break, nj wrestling, spring break getaways, spring break date, high school boys, march concerts, high school basketball, indoor track results, **tussionex**, globetrotters, orioles spring training

Table 1: The 100 queries retrieved from Google Correlate. We treat queries in bold as ILI related.

RMSE *(the lower, the better)*

		Linear (Eq. 1)	#q.	Autoregressive 52 (Eq. 2)	#q.	Autoregressive 3 (Eq. 2)	#q.
Seasonal (Serfling)		0.398	97	**0.280**	5	**0.280**	11
Residual (Serfling)		0.407	98	0.309	98	0.312	98
Seasonal (YA)		0.394	96	0.311	100	0.298	68
Residual (YA)		0.390	79	0.309	49	0.310	47
Not Seasonal	*baseline i*	0.413	33	0.309	68	0.312	46
Not Seasonal all q.	*baseline ii*	0.416		0.310		0.314	
ILI History	*baseline iii*	n/a		0.348		0.333	

Correlation *(the higher, the better)*

		Linear (Eq. 1)	#q.	Autoregressive 52 (Eq. 2)	#q.	Autoregressive 3 (Eq. 2)	#q.
Seasonal (Serfling)		0.948	96	0.973	5	**0.974**	9
Residual (Serfling)		0.946	98	0.968	100	0.967	98
Seasonal (YA)		0.948	99	0.969	99	0.971	68
Residual (YA)		0.949	77	0.969	59	0.966	47
Not Seasonal	*baseline i*	0.942	86	0.968	68	0.967	46
Not Seasonal all q.	*baseline ii*	0.941		0.967		0.967	
ILI History	*baseline iii*	n/a		0.959		0.962	

Table 2: Root mean squared error (RMSE) and Pearson Correlation of our seasonal ILI estimation methods and the three baselines. Bold marks the best score. #q denotes the number of queries used in the estimation.

when the signal has strong autocorrelation, i.e. is strongly seasonal, and (ii) this strong seasonality was present in our dataset, i.e. there was little deviation from the seasonal models. If, however, strong deviations did arise, we expect that models based on autoregression and queries re-ranked based on correlation with seasonal models will perform much worse.

This work complements the use of information retrieval and machine learning methods in the wider area of medical and health informatics [1, 2].

Acknowledgments. Partially supported by Denmark's Innovation Fund (grant no. 110976).

REFERENCES

[1] Radu Dragusin, Paula Petcu, Christina Lioma, Birger Larsen, Henrik Jørgensen, Ingemar J. Cox, Lars K. Hansen, Peter Ingwersen, and Ole Winther. 2013. Specialised Tools are Needed when Searching the Web for Rare Disease Diagnoses. *Rare Diseases* 1, 2 (2013), e25001–1 – e25001–4. DOI: http://dx.doi.org/10.4161/rdis.25001

[2] Radu Dragusin, Paula Petcu, Christina Lioma, Birger Larsen, Henrik Jørgensen, and Ole Winther. 2011. Rare Disease Diagnosis as an Information Retrieval Task. In *Advances in Information Retrieval Theory - Third International Conference, ICTIR 2011, Bertinoro, Italy, September 12-14, 2011. Proceedings (Lecture Notes in Computer Science)*, Giambattista Amati and Fabio Crestani (Eds.), Vol. 6931. Springer, 356–359. DOI: http://dx.doi.org/10.1007/978-3-642-23318-0_38

[3] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457, 7232 (2009), 1012–1014.

[4] Niels Dalum Hansen, Christina Lioma, and Kåre Mølbak. 2016. Ensemble Learned Vaccination Uptake Prediction using Web Search Queries. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, Snehasis Mukhopadhyay, ChengXiang Zhai, Elisa Bertino, Fabio Crestani, Javed Mostafa, Jie Tang, Luo Si, Xiaofang Zhou, Yi Chang, Yunyao Li, and Parikshit Sondhi (Eds.). ACM, 1953–1956. DOI: http://dx.doi.org/10.1145/2983323.2983882

[5] Niels Dalum Hansen, Kåre Mølbak, Ingemar J. Cox, and Christina Lioma. 2017. Time-Series Adaptive Estimation of Vaccination Uptake Using Web Search Queries. In *Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017*, Rick Barrett, Rick Cummings, Eugene Agichtein, and Evgeniy Gabrilovich (Eds.). ACM, 773–774. DOI: http://dx.doi.org/10.1145/3041021.3054251

[6] Vasileios Lampos, Andrew C Miller, Steve Crossan, and Christian Stefansen. 2015. Advances in nowcasting influenza-like illness rates using search query logs. *Scientific reports* 5 (2015).

[7] Vasileios Lampos, Bin Zou, and Ingemar Johansson Cox. 2017. Enhancing Feature Selection Using Word Embeddings: The Case of Flu Surveillance. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017.* 695–704. DOI: http://dx.doi.org/10.1145/3038912.3052622

[8] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The parable of Google Flu: traps in big data analysis. *Science* 343, 14 March (2014).

[9] Mauricio Santillana, André T Nguyen, Mark Dredze, Michael J Paul, Elaine O Nsoesie, and John S Brownstein. 2015. Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS Comput Biol* 11, 10 (2015), e1004513.

[10] Robert E Serfling. 1963. Methods for current statistical analysis of excess pneumonia-influenza deaths. *Public health reports* 78, 6 (1963), 494.

[11] Shihao Yang, Mauricio Santillana, and SC Kou. 2015. Accurate estimation of influenza epidemics using Google search data via ARGO. *Proceedings of the National Academy of Sciences* 112, 47 (2015), 14473–14478.