Event-Related Query Classification with Deep Neural Networks

TempWeb 2020 @ WWW Taipei, Taiwan

April 21st, 2020

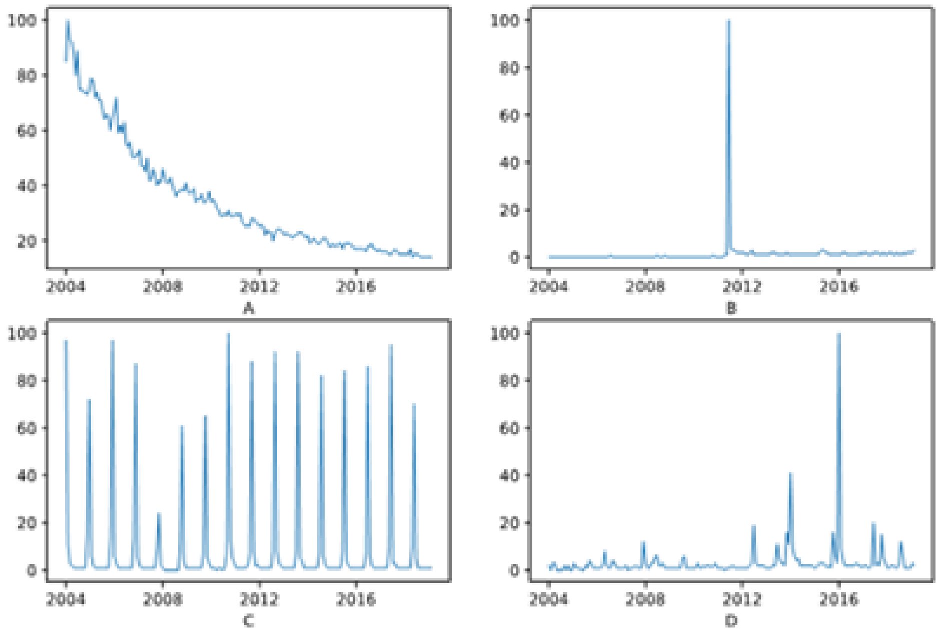
Authors: **Sahaj Gandhi**, Behrooz Mansouri, Ricardo Campos, Adam Jatowt

Ci2.ipt Smart Cities Research Center
Smart Cities Research Center
LIAAD INESC TEC
LIAAD INESC TEC
Kyoto University
Rochester Institute of Technology

Introduction

Introduction

- Pattern on time-series of queries using:
  - Document publish time
  - Query issue time
- Pattern on time-series:
  - Non-event
  - One Time Spike
  - Periodic
  - Aperiodic



Different classes of event-related queries. A. Folk Music (Non-Event Query) B. Death of Steve Jobs (One-time-only) C. Golden Globe (Periodic) D. Ronaldo’s injury (Aperiodic).

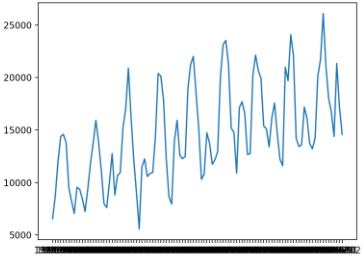
Research Goals

- Detecting different type of event-related queries
- Reducing the model dependency on language
  - Removing the need to use text and language parsing tools
- Using the following deep neural networks on time-series data to classify each of the event-related queries into their type:
  - Convolutional Neural Networks (CNNs)
  - Recurrent Neural Networks (RNNs) like Long Short Term Memory (LSTM) Network
  - A combination of the above two models

Previous work

Event-related query classification with:

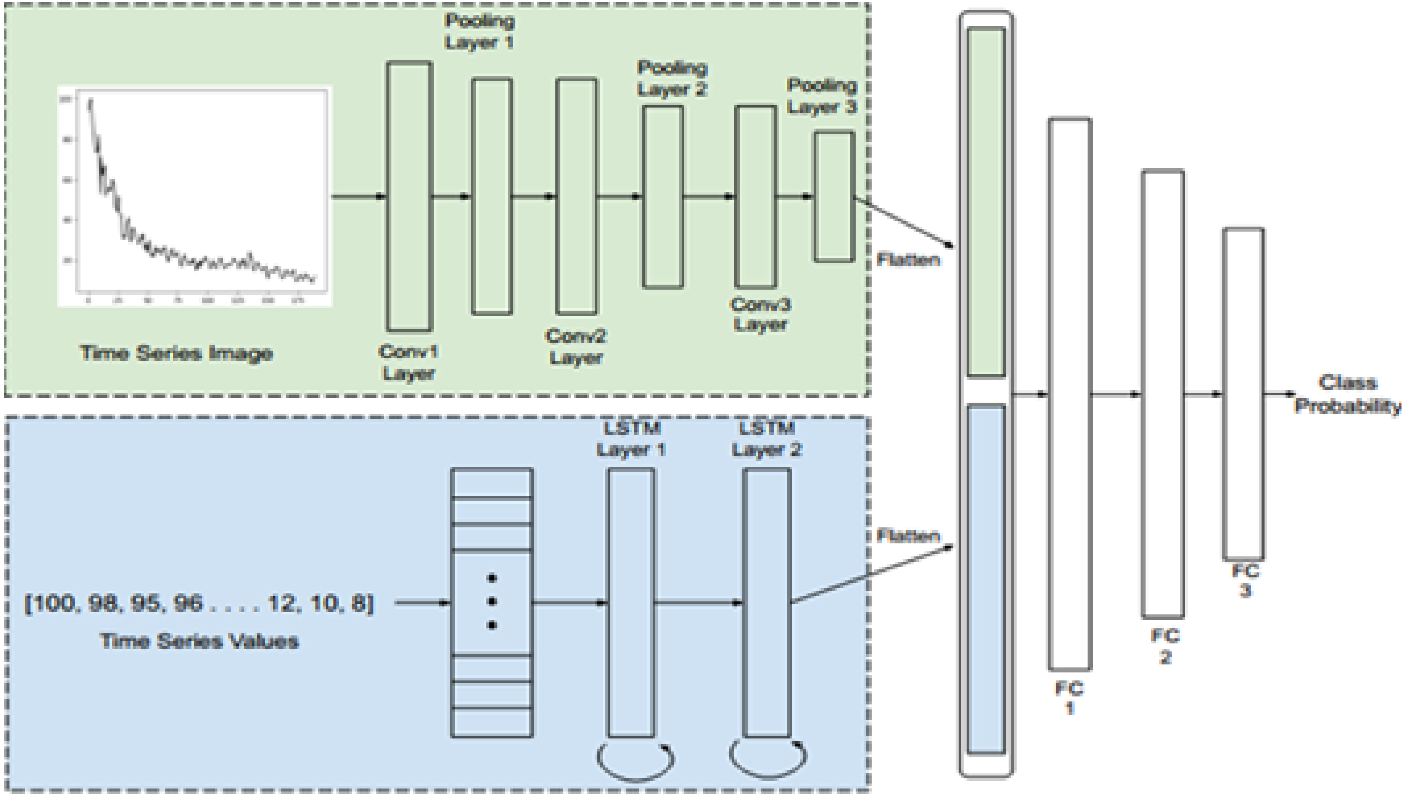
- Text + Time series data
    - Seasonal queries [1]
    - Temporal ambiguous queries [2]
- Time series data
    - Seasonal queries [3]



Our work differs from the previous ones in using only time-series data and using deep neural network to classify query type.

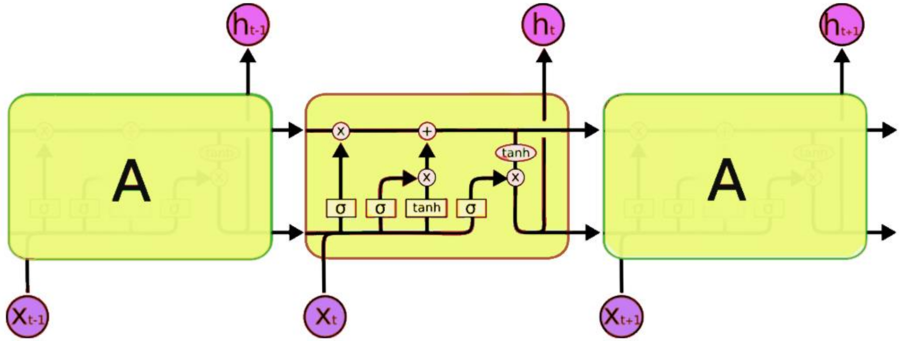
[1]Mansouri, Behrooz, et al. "Detecting seasonal queries using time series and content features." *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*. 2017.
[2]Mansouri, Behrooz, et al. "Learning temporal ambiguity in web search queries." *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 2017.
[3]Shokouhi, Milad. "Detecting seasonal queries by time-series analysis." *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. 2011.

Our Model Overview



LSTM Network-based Architecture

- This model in the past has been used to classify temporal information
- They help understand the underlying structure of the frequency data due to its ability to retain important information over time
- This type of models usually suffer from the problem of vanishing gradients

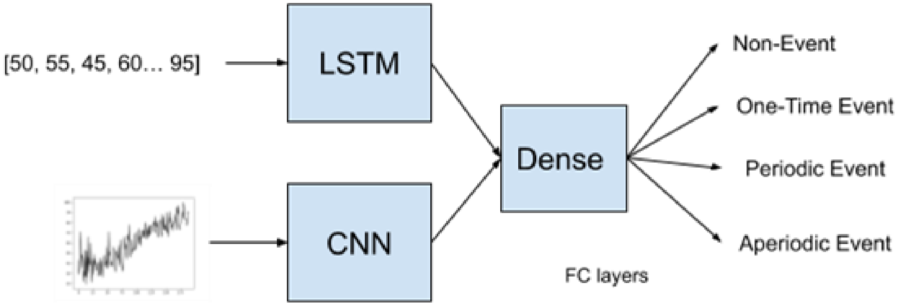


CNN-based Architecture

- This type of an architecture works well to understand the spatial structure of query data
- Works on data being represented visually instead of just numbers that represent frequencies
- Works similar to how human identify patterns in frequency data

LSTM + CNN Architecture

- LSTM networks help analyze the temporal structure of data
- CNN-based networks analyze the spatial structure
- Using both together would help use the pros of both these networks, while avoiding their cons



DATASET

- 600 queries (150 per category).
- Includes queries from English, Spanish, French, Persian, Chinese, and Russian.
- Google trend data (2004 to 2019) which shows the query frequency on scale of [0, 100].
- Used 3 annotators with agreement level of 0.79.

DATASET

Event-related Class	Event-related Queries
One-time-only	北京奥运会(Chinese- Beijing Olympics), The Hateful Eight movie, Ibrahim tatlises vurdun (Turkish- Ibrahim tatlises shot).
Periodic	Winter Olympics, نوروز(Persian- Nowruz), День отца (Russian, Father’s day).
Aperiodic	Éclipse de lune (French-Lunar eclipse), 津波(Japanese-Tsunami), Lampard injury.
Non-event	Puppe für Kinder (German- Doll for children), Church songs, Dynamic programming.

Event-related Classes and Query Instances

Experimental Result

Model	Precision	Recall	F1-Score
SVM	0.72 ▼	0.69 ▼	0.70 ▼
Naïve Bayes	0.65 ▼	0.58 ▼	0.61 ▼
Decision Tree	0.62 ▼	0.66 ▼	0.64 ▼
LSTM	0.82 ▼	0.82 ▼	0.82 ▼
CNN	0.85 ▼	0.80 ▼	0.82 ▼
LSTM+CNN	**0.88**	**0.86**	**0.87**

Event-related Query Classification results with neural network and non-neural network models on test data. Boldface indicates the best results.

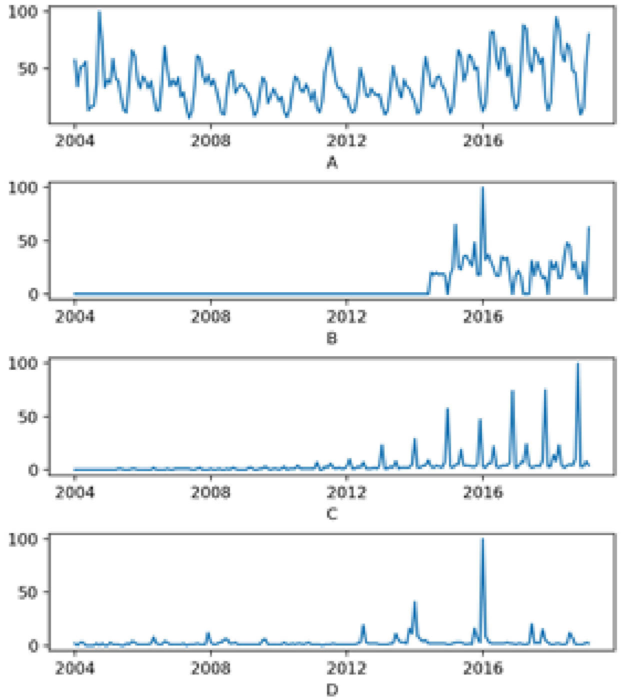
Experimental Result

		Non-Event	OneTimeOnly	Periodic	Aperiodic
Ground Truth	Non-Event	13	0	1	1
	OneTimeOnly	0	13	0	2
	Periodic	0	0	12	3
	Aperiodic	0	1	0	14

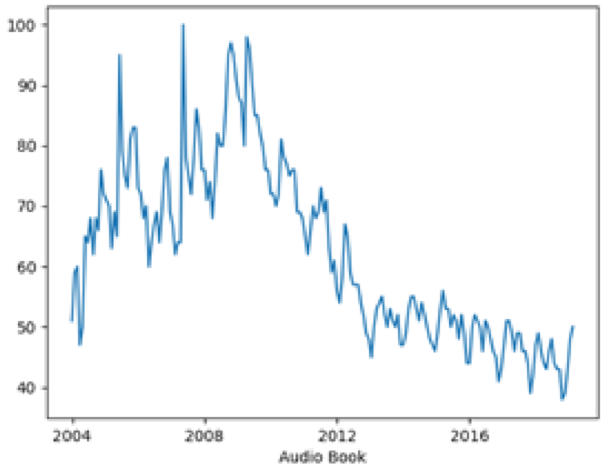
Confusion Matrix for LSTM+CNN Model.

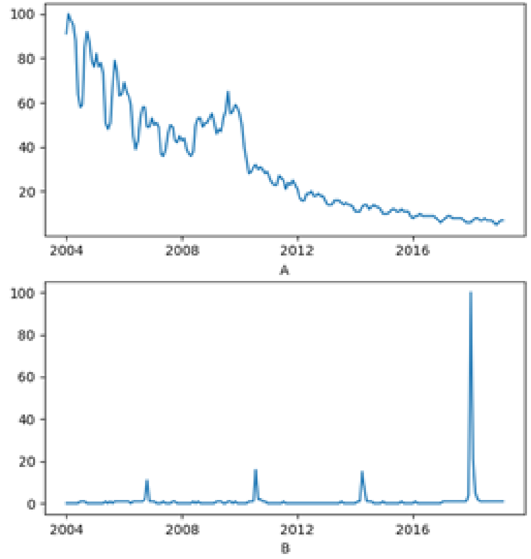
Result Analysis - Misclassification

A) Newton’s Second Law (Non-Event classified as Periodic)
B) BlackLivesMatter (One-time only classified as Aperiodic)
C) День отца (Father’s Day in Russian) (A seemingly periodic event being classified as aperiodic)
D) Ronaldo Injury (Aperiodic, but classified as one-time only event)



Result Analysis





Time-series related to query "Audio Book".

Time-series related to query "Encyclopedia" (a) and "Asian Game" (b).

- Presented a new model for classifying event-related queries by considering four classes: “periodic”, “aperiodic”, “one-time-only”, and “non-event”.

- Our model uses only the time-series data built upon query frequency, which allows it to work for any language.

- We used 600 different queries (150 per category) and studied the effectiveness of our model.

- We plan to study how we could extend the current model also to classify different categories of periodic and aperiodic queries.

- Make use of content features. Based on that, we plan to study how using the embedding models such as word2vec or Bert.

- The prediction of the event type instead of doing the detection.

Thanks for your attention.

This work was financed by the project PTDC/CCI-COM/31857/2017 (NORTE-01-0145-FEDER-03185) and also by the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, through national funds, and co-funded by the FEDER, where applicable.

For further questions please contact the authors.