Predicting Document Creation Times in News Citation Networks

Andreas Spitz\textsuperscript{1}, Jannik Strötgen\textsuperscript{2}, and Michael Gertz\textsuperscript{1}

April 23, 2018 — TempWeb 2018, Lyon

\textsuperscript{1} Database Systems Research Group
Heidelberg University, Germany

\textsuperscript{2} Bosch Center for Artificial Intelligence
Germany
David Cameron resigns after UK votes to leave European Union

PM announces resignation following victory for leave supporters after divisive referendum campaign

David Cameron has resigned, bringing an abrupt end to his six-year premiership, after the British public took the momentous decision to reject his entreaties and turn their back on the European Union.

How do you choose a new prime minister?

The prime minister is notionally picked by the Crown, but in practice the monarch has – for centuries – been obliged to pick the party leader who can command the support of most MPs in the House of Commons.

In the distant past this was established by taking informal soundings, but the Labour party from its inception elected its leader by a ballot of its MPs.

Just a year after he clinched a surprise majority in the general election, a visibly emotional Cameron, standing outside Number 10 on Friday morning alongside his wife, Samantha, said: “The will of the British people is an instruction that must be delivered.”

The prime minister campaigned hard in the divisive referendum on Britain’s relationship with the EU, appearing at hundreds of public events up and down the country to argue that Brexit would be an act of “economic self-harm”.

Hm, when did this happen again?
News Citation Networks
If Rosenstein is fired, this may be the timeline used to rationalize it

There are three possible explanations for the ongoing tension between congressional investigators and the Department of Justice.

The first is that the investigators, led by Rep. Devin Nunes (R-Calif.), are using Justice’s reluctance to share classified documents as a means of undercutting the investigation into Russian interference in the 2016 election and, more specifically, to cast Deputy Attorney General Rod J. Rosenstein as a bad actor to facilitate his firing. Rosenstein, as you may be aware, both appointed special counsel Robert S. Mueller III and has sole authority over Mueller and his investigation; ousting Rosenstein could severely hamper Mueller’s probe.
News Citation Network Overview

News articles from RSS feeds:
- Politics and business feeds
- 34 English news outlets
  (USA, UK, AUS, CAN, GER, CHN, QAT)
- 2 years (Nov 2015 - Oct 2017)
- 244.6 thousand articles
- 367.2 thousand edges

Used data:
- Hyperlinks in the article body
- Publication dates
- Temporal expressions
## News Outlet Statistics (sample)

<table>
<thead>
<tr>
<th>short</th>
<th>news outlet</th>
<th>days</th>
<th>⟨articles⟩</th>
<th>⟨temp exp⟩</th>
<th>other\textsubscript{in}</th>
<th>other\textsubscript{out}</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT</td>
<td>The Atlantic</td>
<td>334</td>
<td>7.2</td>
<td>10.5</td>
<td>16.7</td>
<td>50.6</td>
</tr>
<tr>
<td>BBC</td>
<td>British Bc. Corp.</td>
<td>730</td>
<td>8.1</td>
<td>6.5</td>
<td>19.1</td>
<td>8.0</td>
</tr>
<tr>
<td>DW</td>
<td>Deutsche Welle</td>
<td>334</td>
<td>1.2</td>
<td>6.1</td>
<td>48.1</td>
<td>5.9</td>
</tr>
<tr>
<td>FOX</td>
<td>Fox News</td>
<td>548</td>
<td>2.7</td>
<td>9.8</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>NPR</td>
<td>National Public Radio</td>
<td>334</td>
<td>0.4</td>
<td>8.4</td>
<td>63.6</td>
<td>58.5</td>
</tr>
<tr>
<td>NY</td>
<td>The New Yorker</td>
<td>548</td>
<td>3.0</td>
<td>13.2</td>
<td>33.5</td>
<td>30.6</td>
</tr>
<tr>
<td>NYT</td>
<td>New York Times</td>
<td>669</td>
<td>23.8</td>
<td>10.7</td>
<td>26.8</td>
<td>4.7</td>
</tr>
<tr>
<td>SMH</td>
<td>Sydney Morn. Herald</td>
<td>548</td>
<td>2.3</td>
<td>7.0</td>
<td>3.0</td>
<td>51.9</td>
</tr>
<tr>
<td>WP</td>
<td>Washington Post</td>
<td>548</td>
<td>62.7</td>
<td>9.4</td>
<td>13.7</td>
<td>5.1</td>
</tr>
</tbody>
</table>
Evolution of Network Metrics

- **Average Degree**
- **Clustering Coefficient**
- **Undirected Diameter**
- **Average Path Length**

The graphs show the evolution of these network metrics from January 2016 to July 2017, categorized into different networks: aggregated, politics, and business.
Exploring Citation Chains

Bob Corker tirade encapsulates five reasons why Trump has failed at governing

It's What Bob Corker Does Next That Counts

It's a Shame the White House Has Become an Adult Day Care Center

Why Did Jeff Sessions Really Meet With Sergey Kislyak?

Your cheat sheet for all the investigations into Russia and Trump

Sessions met with Russian envoy twice last year

Six big takeaways from the extraordinary congressional hearing on Russian hacking

The CIA concluded that Russia worked to elect Trump.

Donald Trump just took the first step toward working with Vladimir Putin

The many problems with Donald Trump’s call for Russia to spy on Hillary Clinton

Transcript: Donald Trump on NATO, Turkey’s Coup Attempt and the World

Article Publication Time Prediction
Task Definition: Publication Time Prediction
Available News Citation Network Data

Predict article publication times from:

- Citation network topology
- Publication dates of adjacent articles
- Temporal expressions in adjacent articles

Not the metadata of the article itself
Not the article content
Available News Citation Network Data

Predict article publication times from:

- Citation network topology
- Publication dates of adjacent articles
- Temporal expressions in adjacent articles
- **Not** the metadata of the article itself
- **Not** the article content
Feature Extraction
Network Topology Features

Node degree-based features:

- Incoming degree
- Outgoing degree
- Undirected degree
Network Topology Features

Node degree-based features:
- Incoming degree
- Outgoing degree
- Undirected degree

Centrality-based features:
- Betweenness centrality
- Incoming closeness centrality
- Outgoing closeness centrality
- Page Rank centrality
Network Topology Features

Node degree-based features:
- Incoming degree
- Outgoing degree
- Undirected degree

Density-based features:
- Undirected local clustering coefficient

Centrality-based features:
- Betweenness centrality
- Incoming closeness centrality
- Outgoing closeness centrality
- Page Rank centrality
Temporal Network Features
Temporal Expression Features

Correlation of temporal expressions:

- **good** with publication dates of referencing articles (incoming edges)
- **bad** with publication dates of referenced articles (outgoing edges)
Temporal Expression Features

Correlation of temporal expressions:

- **good** with publication dates of referencing articles (incoming edges)
- **bad** with publication dates of referenced articles (outgoing edges)
Missing Features and Imputation

Missing features

- 30.8% of feature values are missing
- 89.6% of articles are missing at least one feature
Missing Features and Imputation

Missing features

- 30.8% of feature values are missing
- 89.6% of articles are missing at least one feature

Imputation of missing values

- Column mean of the feature
Used regression methods:

- **BASE**: Baseline (average publication date of adjacent articles)
- **LR**: Linear regression
- **BAY**: Bayesian ridge regression (Laplace model)
- **RF**: Random forest
- **GB**: Gradient boosting (Laplace distribution, decision trees)
- **SVM**: Support vector machine (radial kernel)
- **NN**: Neural network (feedforward, one hidden layer)
# Evaluation Results: Mean Absolute Error (days)

<table>
<thead>
<tr>
<th></th>
<th>BASE</th>
<th>LR</th>
<th>BAY</th>
<th>NN</th>
<th>RF</th>
<th>GB</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>66.72</td>
<td>60.46</td>
<td>59.61</td>
<td>26.88</td>
<td>24.98</td>
<td><strong>22.66</strong></td>
<td>26.19</td>
</tr>
<tr>
<td>in</td>
<td>88.88</td>
<td>66.48</td>
<td>87.55</td>
<td>34.03</td>
<td>32.25</td>
<td><strong>27.49</strong></td>
<td>32.29</td>
</tr>
<tr>
<td>out</td>
<td>87.32</td>
<td>59.54</td>
<td>40.24</td>
<td>32.52</td>
<td>30.10</td>
<td><strong>26.68</strong></td>
<td>30.77</td>
</tr>
<tr>
<td>in+out</td>
<td>18.68</td>
<td>55.45</td>
<td>54.95</td>
<td>12.62</td>
<td><strong>11.23</strong></td>
<td>12.76</td>
<td>14.31</td>
</tr>
</tbody>
</table>
Distribution of Absolute Errors

The graph shows the distribution of absolute errors for different regression methods across various datasets. The x-axis represents the regression methods, which include BASE, LR, BAY, NN, RF, GB, and SVM. The y-axis indicates the absolute error in days, with values ranging from 0 to 250 days.

The graphs are divided into four sections:
- **all**: Shows errors for all datasets combined.
- **in**: Focuses on errors within a specific dataset.
- **out**: Focuses on errors outside a specific dataset.
- **in+out**: Combines both in and out errors.

Each box plot represents the distribution of absolute errors for each regression method in the respective dataset category. The error bars indicate the interquartile range (IQR) and the median is marked by a horizontal line within the box.
Recall by Varying Absolute Error

![Graph showing recall by varying absolute error for different methods and data sets.]

- **Method:** BASE, LR, BAY, NN, RF, GB, SVM
- **Data Sets:** all, in, out, in+out
- **X-axis:** Absolute error (days)
- **Y-axis:** Recall (percentage of predictions < absolute error)
Feature Importance: Random Forest

Feature importance: random forest

<table>
<thead>
<tr>
<th>Feature type</th>
<th>Related Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>network topology</td>
<td></td>
</tr>
<tr>
<td>temporal expression</td>
<td></td>
</tr>
<tr>
<td>temporal network</td>
<td></td>
</tr>
</tbody>
</table>
Feature Importance: Gradient Boosting

```
Feature importance: gradient boosting

max (T_{out})
min (T_{in})
deg_{out}
µ (T_{out})
min (Dist)
deg_{in}
σ (T_{out})
σ (T_{in})
span (T_{in})
max (T_{out})
µ (X_{in})
min (T_{out})
µ (Dist)
C_{out}
max (X_{in})
max (Dist)
span (X_{in})
σ (X_{in})
span (T_{out})
min (X_{in})
C_{out}
σ (Dist)
C_{cl.out}
σ (cl.out)
C_{cl.in}
cc
```

Feature type:  
- network topology
- temporal expression
- temporal network

relative importance

10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1} 10^{0} 10^{1} 10^{2} 10^{3} 10^{4} 10^{5}
Summary & Resources
Summary

News citation networks:

- Focus on anchored links inside the article body
- Constructed like a citation network between articles

Publication date prediction:

- Can be framed as a regression problem
- Average prediction error of 3 weeks
- Temporal network features are most discriminative
Data and implementation are available online:

- [data] News citation network (including URLs)
- [data] Temporal annotations
- [code] Publication date prediction

https://dbs.ifi.uni-heidelberg.de/resources/data/
Resources

Data and implementation are available online:

- [data] News citation network (including URLs)
- [data] Temporal annotations
- [code] Publication date prediction

https://dbs.ifi.uni-heidelberg.de/resources/data/

Thank You!
Questions?

Interested in more network-based news analysis? Click here:

*Exploring Entity-centric Networks in Entangled News Streams*
Track: Journalism, Misinformation and Fact Checking III
Wednesday, 15:40 - 17:00, Salle Rhône 2

shameless advertisement