Finding the news lead in the data haystack: Automated local data journalism using crime data

Måns Magnusson
Linköping University
mans.magnusson@liu.se

Jens Finnäs
Journalism++ Stockholm
jens.finnas@gmail.com

Leonard Wallentin
Journalism++ Stockholm
mejl@leowallentin.se

ABSTRACT

The amount of public data is increasing in most countries. At the same time fine-graned data, such as municipal or county levels data, are often underutilized by local newsrooms. There are many reasons to this, the financial situation of local newsrooms, the large scale of public datasets and the large amount of random noise in data. We use a Bayesian modeling approach to identify probabilistic news leads, i.e. news leads that are statistical signals in large scale noisy datasets. The news leads studied are peaks in the number of reported offences in Sweden, on a monthly basis and at the municipal level. The results are preliminary, but promising. During June and July of 2016, 25 articles were published in local newspapers where this approach was used. Although the results are promising, further work is needed to extend the results to other datasets and to identify other types of probabilistic news leads.

Keywords

Bayesian modeling, data monitoring, local journalism

1. INTRODUCTION

Access to open and public data is increasing in most countries.[4, 12] Data on most areas of society are disseminated as statistical or open data in different format easily accessed by journalists. In Sweden only, millions of datapoints are disseminated each month from different government agencies. These public data cover most areas in society, such as the number of reported offences, causes of death, unemployment and school results, to name just a few examples. It is not uncommon for public datasets of these types to be released every month. In many cases the datasets are available at a very fine grained level - by municipality, different age groups or, as in our case, by criminal offences.

These fine grained local data are a highly underutilized source for local journalism. There are multiple reasons for this. First of all, the amount of data becomes a threshold in itself. Analyzing all relevant data quickly becomes insurmountable using traditional, manual approaches. A second problem is that most public datasets are full of statistical noise of small (if any) journalistic or public interest. This makes filtering or identifying the story in the data haystack a major problem. Data monitoring or other similar continuous news application has also had limited success historically, due to short shelf time of news and the lack of an organizational culture of maintaining projects over time in news organizations.[10] All these problems, together with the tough financial situation of many local newsrooms, hinders this type of "data watchdogging" by local journalists.[8, 1] This is unfortunate, since the reduced economic and editorial resources of local newsrooms makes it, at the same time, more difficult for local journalists to hold local politicians and government to account.[7]

One approach to aid the local newsrooms is automating analysis and monitoring large public dataset at a detailed local level. Identified news leads or potential stories can then be delivered to local journalists for further investigation. We have created an embryo of such a news production system called *Marple: The Story Sniffer*, to examine the possibility of using Bayesian methods to find potential news leads that can be of real interest for local journalists. The project was awarded prototype funding from Google's Digital News Initiative in the spring of 2016 and is currently under development. This article will share and discuss some of the methods we have used so far as well as problems encountered.

We draw methodological inspiration from the area of public health surveillance. In this field of research a large amount of statistical methods has been developed to identify communicable disease outbreaks in large public health datasets.[9] Applying ideas and methods from the field of public health surveillance, we study the possibility of an automated large scale journalistic surveillance, with reported offences in Sweden as a first experiment.

2. RELATED WORK

Journalistic monitoring using computational methods has been put forward by many journalistic innovation theorists as a way of improving "watchdogging journalism".[7, 1] In [10] a continuous monitoring system of the Norwegian parliament is tried out. The approach of monitoring is appreciated by experienced political journalists, but at the same time one of the conclusions is that computing and systemizing raw data is not sufficient. The journalistic process also needs to filter out the interesting stories and put them in a

C+J '16 September 30– October 1, 2016, Palo Alto, CA, USA
ACM ISBN 999-9999-99-99)/99/99.
DOI: 99.999/999_9

journalistic context.[10]

The work of monitoring data in [10] is closely related to the first step of automation of journalism described in [6], *Collecting data*. Our approach extends the monitoring of data primarily to the two next steps, *Identifying interesting events* and *Prioritizing insights*. We do not just want to collect the public data, we also want to automate the identification of interesting local events in public data.

This idea of monitoring open data is not new.[3] One of the promises of computational journalism has been to free the journalist from the mundane task of discovering and obtaining facts. Instead the hope has been that journalists can focus on verification and explanation.[4] As mentioned in [1], the best algorithms will generate leads, hunches and anomalies to investigate further. To develop methods to find these news leads a combination of knowledge is needed, both the technical (and statistical) skills of data crunching and the journalistic expertise of refining the story and investigating.[1] But as has been pointed out by [12] there has been little focus on which algorithms (or statistical models) to use. Examples of algorithmic approaches to investigative reporting can be found in [1] and [12].

This limited research on explicit methods to identify news leads in large datasets has drawn our attention to research fields where large scale data monitoring has been used for a long time. In public health, monitoring and surveillance of public health data has continuously been used to identify anomalies and deviations that can indicate a possible disease outbreak.[2] The problem of identifying disease outbreaks is similar to that of identifying news stories. Most variations in data is just random noise, but sometimes there are deviations or anomalies that need further attention. In public health monitoring this has been handled by developing specialized statistical methods for continous public health surveillance.[9] The fact that multiple statistical methods have been developed to analyze reported counts of disease cases (incidence), together with previous experience of analyzing crime data in [1] has motivated our choice of starting out analyzing public data on the number of reported offences in Sweden. Similar to a public health surveillance system, the resulting alarms (potential news leads) are sent to local journalists for further investigation - in the same way as potential disease outbreaks are handed over to epidemiologists for further investigation.[11]

3. FINDING STORIES IN PUBLIC DATA

To approach the problem of how to continuously monitor public datasets for potential news leads we first define two main types of potential stories or news leads: deterministic and probabilistic news leads.

Deterministic news leads can be computed very easily and by the very definition it can be a story of interest. An example of a deterministic news lead can be a new maximum value in a time series of reported offences stretching over a ten year period. It is easy to identify computationally and it is often of journalistic interest.

Probabilistic news leads, on the other hand, are news stories that are relative to the time series at hand. To identify a probabilistic news lead we need to take previous variation and data into account and try to find the journalistic signal in the random noise. Examples of these approaches are the algorithms described in [12] on how to find neighborhoods with an unusual amount of crime last week. We cannot be sure that the last week's data points are just due to random noise or that there exist seasonal effects, but a sufficiently large or small value will *probably* be a news lead of interest. The larger the anomaly, and hence the smaller probability of just being random noise, the higher the probability of being an interesting potential story. To find those news leads we need to model the data using probabilistic models to take the random noise into account when trying to find relevant news leads.

Deterministic news leads are relatively easy to identify using simple computational approaches. The probabilistic news leads, on the other hand, make it necessary for us to combine both journalistic knowledge, computer science and statistics, especially in large scale datasets where thousands or maybe millions of time series are being analyzed. In this study we use a Bayesian approach to find news leads (see [5] for an introduction to Bayesian analysis). We use a Bayesian approach for pedagogical reasons and for transparency. Bayesian philosophy facilitates a more commonsense reasoning about unknowns and uncertainties in the form of probabilities. We can estimate the probability of a given datapoint being an anomaly or random noise, thus making it easier to reason about probabilistic news leads. The Bayesian approach, in contrast to more algorithmic machine learning approaches, is always based on a proper probability model. This makes the Bayesian approach transparent and easy to communicate. Given the probabilistic model and the inferencial approach, it is easy to reproduce the analysis on which the news leads are based.

4. EXPERIMENT

To explore the possibility of identified probabilistic news leads resulting in news articles, we created a system called *Marple: the story sniffer*. The system monitors the reported number of criminal offences in Sweden at a monthly basis for all the 290 municipalities and city districts in Stockholm, Gothenburg and Malmö (15 in total) for 25 different offence categories. That makes a total of almost 8,000 time series - and potential stories - every month. This data type, reported counts, is similar to reported disease cases in public health application, making public health surveillance models a first approach to study. There are also seasonal patterns in the data that further complicate the identification of potential news stories.

All time series are analyzed when the data is publicated and if anomalies in the series are identified they are flagged as "alarms" or potential news stories. These alarms are then reviewed by an editor (one of the authors of this paper) and screened to be interesting enough. Finally the news leads are sent to local newsrooms through Nyhetsbyrån Siren, a national news provider that more or less all local newsrooms in Sweden subscribe to.

4.1 Method

To identfy news leads we use a relatively simple conjugate Bayesian model. We assumed the data followed a Poisson model with parameter λ following a vague gamma prior. To calculate the predictive distribution of the reported offences we used the conjugacy property and for computations we used the `surveillance` R package.[9] One of the main benefits of the conjugacy property of the model is that it is possible to do calculations analytically, considerably reducing the computational cost of inference.

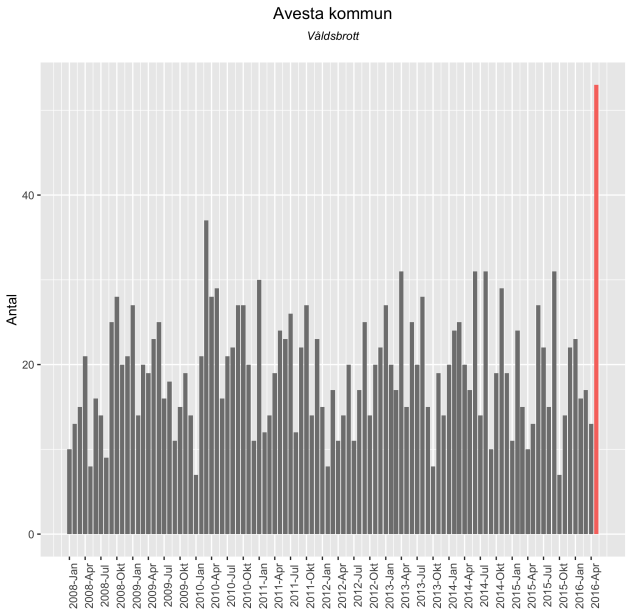

Figure 1: Example of a news lead in Avesta municipality, frequencies of reported violent crimes. The red bar highlights the alarm identified.

Figure 2: Example of a story that Tidningen Nordost ran based on a news lead in Gothenburg from Marple. The title translates "Significant increase in drug use and vandalized vechicles".

To handle seasonality, three different parameter settings have been used to estimate three different versions of the model. The first model uses data from the same month five years prior to the current data point. The second model uses data from the same months, three months before and three months after the data point of interest. The last model uses only the last three months leading up to the month of interest. Using all models we estimated the λ parameter and a prediction interval with an upper threshold value with α set to 0.001. These parameters were chosen ad hoc to find "obvious" alarms and to reduce the number of false alarms. If the datapoint was outside the prediction intervals of all three models, the data point was flagged as an alarm. This approach is ad hoc and should be refined, but it works as a first example of a more simplistic probabilistic approach to identify potential news leads.

The analyses were conducted for all approximately 8,000 time series of the data disseminated in June and July 2016. The resulting newsleads are presented to the local journalist with a simple auto-generated heading and a descriptive text, a chart and a csv file with the underlying data. Figure 1 shows an example of what a news lead can look like. Here we found an unusual amount of violent crime in the municipality of Avesta.

5. RESULTS

These preliminary results consist of two months of monitoring reported number of offences, data disseminated in the beginning of June and July 2016. Each month all approximately 8,000 time series were analyzed. The result is presented in Table 1.

The alarms that were disqualified and not selected to be sent to local newsrooms were excluded in some cases due to model defects resulting in strange alarms for municipalities

Table 1: Preliminary Results

Month	Alarms	Selected	Publications
June	74	32	14
July	55	22	11

with high frequencies, in some cases due to more journalistic considerations, such as small peaks in less interesting offence categories (such as the category "other types of thefts"). Most of the published stories were short news items, but in some cases the local journalists took the leads one step further, making interviews and adding cases to put the data in context. An example of a resulting news item can be found in Figure 2.

6. DISCUSSION AND FUTURE WORK

The results of our experiment clearly show potentialwhen it comes to finding news stories by monitoring large public datasets. The models used are, however, still far from perfect. Two main areas of improvement are needed. First we need to reduce the number of alarms identified by the models that are disqualified in the editorial process. Second, we need to see how the editorially selected stories can be used more by the local journalists.

The first problem is primarily due to the all too simple and ad hoc modeling approach. The models did, in general, a good job in small municipalities where the number of reported offences are low. But for larger municipalities, such as Stockholm city, the simple Poisson model was too rigid, resulting in several "false" alarms.

The second problem, increasing the usage by the local journalists, is harder to solve. Our experience from this first small experiment has given us the insight that local jour-

nalists are not used to work with this type of probabilistic news leads. They have problems as to using it and we believe that each news lead need to be presented with a more obvious journalistic angle with more context, text and visualization of the identified alarms.

Another problem we have encountered during this project is the risk that anomalies in the data might be due to errors or changing methods in the data collection rather than events in the real world. The quality of an automated story finder will never be better than the quality of the data, making full automation and no editorial step difficult in the near future.

For now we also only identify crime peaks, but there are of course many other ways to define a news lead in a dataset like this: a declining or rising trend, a development that differs from similar or neighboring municipalities and so on.

We are now trying to improve the probabilistic modeling. First of all we need a model that can solve the initial problems of the Poisson model - identifying too many false anomalies in larger municipalities. But we also want to extend and create probabilistic models for other types of news leads such as sudden differences between neighboring municipalities, new trends and better handling of seasonal effects by the usage of state-space models. The introduction of probabilistic models also makes it interesting to use short time predictions and deliver these as news stories, such as what type of criminal offences are expected to increase the coming month.

It remains to be seen how easily these models will scale to other public datasets such as unemployment, immigration, house prices and causes of death. How much statistical tailoring does every new topic require? A more elaborate version of the platform will also have to include more elaborate editorial logic. If, for example, every municipality in a county reports a peak in unemployment the story should probably be framed with a county angle, rather than a municipality one.

7. REFERENCES

[1] M. Broussard. Artificial intelligence for investigative reporting: Using an expert system to enhance journalists' ability to discover original public affairs stories. *Digital Journalism*, 3(6):814–831, 2015.

[2] B. C. Choi. The past, present, and future of public health surveillance. *Scientifica*, 2012, 2012.

[3] S. Cohen, J. T. Hamilton, and F. Turner. Computational journalism. *Communications of the ACM*, 54(10):66–71, 2011.

[4] T. Flew, C. Spurgeon, A. Daniel, and A. Swift. The promise of computational journalism. *Journalism Practice*, 6(2):157–171, 2012.

[5] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*, volume 3. Chapman & Hall/CRC Boca Raton, FL, USA, 2014.

[6] A. Graefe. Guide to automated journalism. *Tow Center for Digital Journalism. Janeiro*, 2016.

[7] J. V. Pavlik. Innovation and the future of journalism. *Digital journalism*, 1(2):181–193, 2013.

[8] C. Royal. The journalist as programmer: A case study of the new york times interactive news technology department. In *International Symposium on Online Journalism*, 2010.

[9] M. Salmon, D. Schumacher, and M. Höhle. Monitoring count time series in r: Aberration detection in public health surveillance. *Journal of Statistical Software*, 70(1):1–35, 2016.

[10] E. Stavelin. Computational journalism. *When Journalism Meets Programming. PhD, The University of Bergen*, 2014.

[11] S. M. Teutsch and R. E. Churchill. *Principles and practice of public health surveillance*. Oxford University Press, USA, 2000.

[12] M. L. Young and A. Hermida. From mr. and mrs. outlier to central tendencies: Computational journalism and crime reporting at the los angeles times. *Digital Journalism*, 3(3):381–397, 2015.