Newspaper archives + text mining = rich sources of historical geo-spatial data

To cite this article: A Yzaguirre et al 2016 IOP Conf. Ser.: Earth Environ. Sci. 34 012043

View the article online for updates and enhancements.
Newspaper archives + text mining = rich sources of historical geo-spatial data

A Yzaguirre, M Smit and R Warren
Dalhousie University, Halifax, Nova Scotia, Canada
E-mail: yzaguirre@dal.ca

Abstract. Newspaper archives are rich sources of cultural, social, and historical information. These archives, even when digitized, are typically unstructured and organized by date rather than by subject or location, and require substantial manual effort to analyze. The effort of journalists to be accurate and precise means that there is often rich geo-spatial data embedded in the text, alongside text describing events that editors considered to be of sufficient importance to the region or the world to merit column inches. A regional newspaper can add over 100,000 articles to its database each year, and extracting information from this data for even a single country would pose a substantial Big Data challenge. In this paper, we describe a pilot study on the construction of a database of historical flood events (location(s), date, cause, magnitude) to be used in flood assessment projects, for example to calibrate models, estimate frequency, establish high water marks, or plan for future events in contexts ranging from urban planning to climate change adaptation. We then present a vision for extracting and using the rich geospatial data available in unstructured text archives, and suggest future avenues of research.

1. Introduction

Data is being created and utilized as you read this passage. It is finding new life in an app on a smart phone, assisting instructors in a classroom where participation is recorded through audience response systems, navigating newcomers to the best coffee in town, and flourishing online in various forms. We benefit from and contribute to the ever-growing mountains of digital information whose value and relevance varies from unknown to low to arguable to unquestionable. As countries move towards open data and crowdsourcing fuels new and major sources of geographic information (e.g. OpenStreetMap), it is certain that new, sustainable governance models will be required in the development of the “next-generation Digital Earth” [1] to assist with the curating of data and establishment of a system for its long-term preservation. It is also necessary to be able to discover, explore, and extract useful information from these datasets, in any format or space they may currently reside in, as either an expert or general user.

In recent years there has been great interest and activity in applying computational methods to extract knowledge from various document collections. This body of work interweaves several domains including Natural Language Processing (NLP), Geographic Information Systems (GIS), Information Retrieval (IR), and Geographic Information Retrieval (GIR). Automatic text analysis can aid what were once intensive and time-consuming manual methods through NLP tools that streamline the initial assembly of a database built from information in free or proprietary text documents. Text analysis is a means for converting such unstructured text documents into structured, geo-spatial objects that are readily placed into a geographic context that facilitates exploration and discovery.
Archives of regional newspapers can contain hundreds of years of rich cultural, social, and historical information that offer both business value to the newspaper, and cultural and social value to members of the region. For readers, there is unique, personal value in news articles about a region that are written and published by people from that region. This locality is what allows regional newspapers to compete with national and international news sources. However, the value is not easily extracted from these archives. These archives are typically unstructured and organized by date rather than by theme/subject or place. Metadata is often minimal and does not indicate which articles are important, which articles are outdated or duplicated, and which will be of long-term interest, particularly when the interest is driven by regional or personal interest.

Text mining seeks to extract information in unstructured text and convert it into structured data that allows the information to take on an easily searchable form. The structuring of data often includes links to other databases, which in the case of GIR and our study means links to gazetteers (worldwide geographical databases, or “geographical dictionaries”). However, the lack of human “geo-annotated” versions of what are often vast geo-spatial collections often complicates evaluation of GIR tasks and their performance. Evaluation datasets for localized data sources are often challenging to establish and largely unavailable due to time, cost, and feasibility constraints. A study of these challenges is included in [2], who focused on the GeoCLEF collection (169,477 documents derived from the Los Angeles Times and the Glasgow Herald) as a demonstration of these issues.

This paper describes tools and methodologies to extract valuable data from digital newspaper archives. We describe a pilot study already implemented, seeking specific geospatial data from an unstructured archive (Section 2); describe work in progress to present archival data to the general public using mechanisms informed by and driven by place (Section 3); and describe potential research directions to continue progress toward daylighting unstructured textual archives (Section 4).

2. Pilot study: locating flood events in newspaper archives
The pilot study of this paper is borne out of a collaboration with Nova Scotia Environment, a government department of the Province of Nova Scotia, Canada, and a major regional newspaper, The Chronicle Herald. The novel technical tools, algorithms, and approaches are described in detail in a previous technical paper [3]; here we focus on the process and the lessons we can learn to apply to future projects. This project was driven by interest in extracting information about flood events in the province of Nova Scotia from newspaper articles and into a database that can be explored through a web map application; the result is available at http://nsfloodhistory.management.dal.ca. Flood databases are typically created manually from archives, media, news articles and insurance reports, and take years to build. Nova Scotia’s had not been updated since 1987. The standard for other provinces in this region is a basic textual interface to retrieve descriptions of flood events from a database; for example, the neighbouring province of New Brunswick offers http://www.elgegl.gnb.ca/0001/en/Home/Main and the federal Public Safety Canada department offers http://cdd.publicsafety.gc.ca/).

In this section, we provide a system overview (2.1), our methodology of using off-the-shelf NLP and machine learning (ML) tools to apply classic text-mining techniques on a regional, digital newspaper archive of 2 million news articles to find flood events over the last 20 years (2.2), our approach to evaluation using Amazon’s Mechanical Turk to employ human assessors to perform the same tasks and compare the results (2.3), and our visualization interface (2.4).

2.1. System overview
Our system architecture is shown in figure 1. We rely on the GeoNames.org gazetteer (http://www.geonames.org/) to weight the relevance of the documents for place names in Nova Scotia. This allows us to filter the corpus for non-ambiguous locations. Specifically, documents whose list of location matches contain more place names outside of Nova Scotia
than locations only in Nova Scotia are filtered out to resolve referent ambiguity, which is
the well-known problem in GIR of distinguishing a document’s location between two places
with the same name, e.g. Sydney, Nova Scotia versus Sydney, Australia. See Vasardani [4]
for a review of literature on geographic information retrieval based on place names.

2.2. Methodology
Our initial dataset is an unstructured corpus of 2 million articles provided by The Chronicle
Herald; after a manual data cleaning/filtering step to remove non-news articles from this
corpus, we have 1,210,476 digital articles in XML format (6.54 GB) dating from January
1992 to February 2015. Though XML documents have some structure, in our case most
of the metadata is not relevant to our task; we do use the XML structure to isolate the
free text of the news article, the article’s headline, the publication date, and a unique ID
of the article to generate appropriate back references for later tasks.

Our goal is extraction of a thematic and spatial event: we seek to identify articles that
describe flooding events in Nova Scotia, and produce a structured corpus that includes
a date for the event, a geo-tagged location, a snippet from the full story text, a list of
other places mentioned in the article, and a full citation of the article (author, headline,
publish date).

In our text analysis, we use the Natural Language Toolkit (NLTK) [5], a suite of
program modules to support natural language processing in Python, to assist with pre-
processing of the text data such as part-of-speech tagging of words (e.g. nouns, verbs,
adjectives), and tokenization of text into words or sentences. Our measure of “relevance
of information” in the text relies on the calculation of varied TF-IDF weights which is done
using scikit-learn [6], an open source machine learning library for Python. TF-IDF stands
for “Term Frequency - Inverse Document Frequency” and is a measure that reflects the
importance of a term in a document collection: the more times the term appears in the
document, the higher its TF-IDF score, but this score is also offset by the frequency of

Figure 1. Overview of the system architecture from extraction to user interactivity
the term in the document collection. It is common to refer to the document collection as a corpus, and so we use these two terms interchangeably in this paper; similarly, spatial locations will be referred to as toponyms, place names or named entities.

2.2.1. Filtering and geo-locating We perform a series of rule-based filters to facilitate our search for floods in The Chronicle Herald’s digital newspaper archives. The first of these includes a keyword search in the document’s text for an appearance of the word ‘FLOOD’, which trivially provides us with a collection of flood-related news. This reduces our total of news articles from 1.2 million to 13,659 articles over a period of two decades. A majority of these articles are not relevant to our search task, which brings our focus to isolating more true positives in the collection.

In a second filtering step we use a domain-specific vocabulary to identify words and phrases that we expect true positives to include. For our pilot study this vocabulary develops from a historical flood database of events prior to 1987 that include details on causes, description, magnitude, region, river, and damage information for a collection of 183 flood events in Nova Scotia. We use the frequency of these vocabulary words in a news article to yield a score for the article, which we normalize against its total words to give a percentage: we find that a relevance threshold of ≥ 3% determines which articles to keep in our corpus. It is common practice in text mining to set an arbitrary threshold that is tuned according to manual review of your information extraction goals. In our third filtering step we exclude newspaper articles that do not display a variety of identified vocabulary words in their free text; in our study this worked to filter out stories about other types of violent or damaging events. At this point our corpus contains 1,223 news articles, about 9% of the original data set. The final filtering step involves the GIR-related task of identifying those news articles that took place in our region of interest, Nova Scotia.

We use NLTK’s part-of-speech tagger and chunking modules to identify a list of possible toponyms referenced in the article. Sometimes locations tagged by NLTK are people or organizations; we can disambiguate locations from those that are person’s names by excluding the place if the entity is preceded by a title (e.g. Mme, Col, Ms, Mr, Dr, Rev, etc.). Though in general this step comprises a ‘generate-and-test loop’ where we annotate the text with ‘locations’, generate patterns from these annotations, extract information, and then use this information as a new instance to annotate text, generate more patterns, and so on (when to stop this loop process is non-trivial and requires manual input). The goal of this cycle is to identify a list of locations for a given newspaper article that we next use to positively filter those articles that likely occurred in Nova Scotia.

The Geonames gazetteer helps to tag places in the province of Nova Scotia. Using each document’s list of extracted toponyms, we score an article based on the number of places it contains that are place names in Nova Scotia. In our study at least two non-ambiguous location matches are set as threshold to keep the article in the corpus. We remove articles whose lists contain more geographic locations that are also place names outside of Nova Scotia than those place names that are only in Nova Scotia. Such a heuristic approach is effective if one is primarily interested in identifying geographic names in a document and not in the broader task of named entity recognition and text categorization.

Our final task is to extract a single location to associate with the “flood-relevant” newspaper article. In the spirit of simplicity, we search and select from the article the toponym that occurs nearest to all occurrences of the word “FLOOD” in the free text. We evaluate the performance of this choice against human assessors, though we do comment here that the challenge of identifying a region or point of geographic interest in GIR tasks is quite complicated and well-studied in other areas, e.g. see [7, 8, 9, 10, 11, 12, 13, 14, 15], and references therein.

2.2.2. Summarization When working with closed domain data it can be the case that their proprietary, protected, or private nature means that reproduction of the full source document text is not allowed or legal. This is the case for our study where the news
articles themselves are valuable intellectual property and so cannot be displayed to users of the database in their complete, original form. This gives us an opportunity to explore a summarization method that identifies the most relevant sentences of the article to include in the final database. We refer to these summaries as *snippets*.

A well-chosen snippet helps users of the database quickly understand the content and substance of the article, which improves usability. Our approach for snippet extraction begins by manually selecting 18 relevant articles from the first year of events in the corpus (1992), where relevance is a boolean query for an article describing a flood event in Nova Scotia, or not. This subset of articles form an *exemplar sub-corpus* that we use to build a collection of (manually extracted) relevant snippets that help us to automatically extract snippets for each document in the remainder of the corpus. Relevant snippets that we manually select include sentences about the extent of flood damages, the cause of a flood, the regional locations affected, and informative quotes from local residents impacted by the event. We refer to this curated collection of flood story snippets as our *seed snippets*.

Our approach for automatic extraction of snippets is to identify similarity using the TF-IDF measure, where each sentence in a news article is now assumed to be a document itself. First we pre-process the data using NLTK’s tokenize tool to segment the news article’s free text into its individual sentences. Sentences that include “flood”, “damage”, or “cause” are ranked as potential candidates for the story snippet; the lemmatized forms of flood, damage, and cause in our keyword search help to capture relevance in more phrases.

Using the methods in scikit-learn’s feature extraction module we convert the text to lower-case, remove punctuation/accents, filter stop words, and compute the TF-IDF scores for each sentence; specifically, we rely on the TfidfVectorizer and fit_transform methods to convert our collection of “documents” (in this case, news article sentences) to a matrix of TF-IDF scores. The NumPy [16] Python package is used to manipulate this final matrix, where we offset each document’s TF-IDF scores with the seed snippet’s scores, returning the highest scoring sentence (i.e. the sentence most similar to the seed snippet) as the relevant snippet for the news article.

To cultivate variety in the final database, we also allow the collection of seed snippets to grow dynamically throughout the extraction process, where extracted snippets also become seed snippets for other news articles in the corpus. The result of this variety is a collection of both relevant and interesting story snippets in the final database.

2.3. Evaluation
Evaluation uses Amazon’s Mechanical Turk service (MTurk) to recruit human assessors, provide them with tasks, and capture their responses for comparison with our own results and manual assessment (see [3] for details) at a large scale.

2.3.1. Event To evaluate the accuracy of the filtering of news articles, we start with the 13,659 articles after the first filtering step, which simply looked for the ‘FLOOD’ keyword. We randomly select 864 news articles, half articles that appear in our final database, and half articles that do not, and we ask Mturk workers to score the article from strongly positive (a definite match for a flood event) to strongly negative (irrelevant as a flood event). There are no false negatives when we compare their results with the flood articles identified by our approach. If we aggregate their responses into categories of “FLOOD” and “NO FLOOD”, we identify a false positive error rate of about 4.9%, or in other words, a total accuracy of over 95% for identifying flood events in the newspaper articles.

2.3.2. Geo-location We use the news articles that human assessors identify as describing flood events to evaluate geo-location annotation. We ask a human to tag the article with a GeoNames toponym provided to them in a multiple choice list, and each toponym is tagged by two people. They both agree on a geo-tag for a news article 60% of the time; when they agree, our system agrees with them 64% of the time. When the human assessors disagree,
Figure 2. Design of the web application. The color of a map marker represents temporal season. An info window appears when a marker is clicked, displaying the story snippet, season, a full citation, other places, and an error report link. The web application and the raw data are available online: http://nsfloodhistory.management.dal.ca.

Our accuracy falls to 34%. As another benchmark for the potential accuracy of our geotagging system, we randomly selected 20 news articles and hand-coded their location; interestingly, both our system and the hired human assessors had a 60% accuracy against this data set. This suggests that location extraction is difficult to assess even for human readers, who need geographic knowledge of the region being discussed to be accurate [7].

2.4. Visualization

The final database of news articles contains 572 records of flood events in Nova Scotia from 1992 to 2015. Since it is common for events to take place at the same location, we opt to display each record at a unique point by algorithmically spreading out each point in circles centered at the original location’s longitude and latitude coordinates, with the oldest dates nearest the center and increasing with time as spiral out clockwise. Clicking on a marker opens an info window for that record with the story snippet, a list of other locations mentioned in the article, the season that the story occurred in, a full citation for the record, and a link to the errata page to report the record for re-evaluation as illustrated in figure 2. The List View button directs the user to a full list of all records, which can be sorted at the top by Year, Location, Citation, Snippet, or Other Places.

The database is searchable using any of the radio buttons in the left sidebar: address, year, keyword (e.g. causes of flood events), and seasons. For instance, a user can decide to view all articles that occurred within 10 km of Cole Harbour over the last 10 years during the Summer, of which there are 11 news articles. We currently host the database on Google’s Fusion Tables and use HTML, CSS, and Javascript for the interactive web map (no server-side code necessary); some other systems include the open-source Leaflet.js and Neatline, or CartoDB.
3. Vision for mining relevant geospatial data from archives

Our process of identifying the needs of the stakeholders in our pilot study produced the following set of objectives for connecting users to archival data in a way that maximizes utility: relevance, identifying only the data sought is important when expectations are set by Google’s web search example; summarization, crafting a summary that reflects the story’s value; geo-spatial significance, presenting accurately data tied to place, including the density of data; interactivity, providing navigation tools that allow users to navigate through the data collection in an intuitive way that can be tuned to their own interest; and visualization, presenting data in an intuitive and accessible format that exploits temporal and spatial features.

General users do not seek to experience archival data through the exploration of fonds or collections. The continued growth of data, the variety of records, and the additional forms of media employed suggest that manual examination of archives will become increasingly difficult and intractable even for expert users. Automated methods, tested and evaluated, combined with effective visualization, are capable of providing positive interactions with valuable archival content for both types of users. Automated text analytics can operate at Big Data scale to produce meaningful summarizations and visualizations.

For example, we are undertaking a project to connect newspaper articles to individuals’ experience of place. Using a mobile application, users will be shown archival articles or photos relevant to their location. This requires the geolocation of the unstructured data using a method similar to the one described in section 2. It also requires understanding which articles are important or relevant; location is a key indicator of relevance, but we are also exploring original placement in the paper, the proximity to locations mentioned in the article, the occurrence of follow-up articles or similar articles in the corpus, and the people mentioned in the article. Understanding how the display of relevant articles and information impacts an individual’s sense of place is also a focus of interest for our work; we envision a mobile application capable of displaying information of relevance to the current location of the users, identifying news stories that are important to both a user and the location, and exploring how users experience “place” through photos.

Of course this vision is not limited to newspaper articles presented through mobile apps. Other sources of data could include digitized archives, digital cultural heritage collections, and digital museums, presented through, for example, augmented reality exhibits where real-world environments and context-based digital information are dynamically blended. The incorporation of visual data can be accomplished using modern algorithms capable of generating semantic descriptions of images (e.g. [17]) and employing our approach, or through the development of new algorithms specifically for this task.

We also know that how individuals experience place is important when considering development projects that impact areas of long-term local, historic, or cultural importance. Archival data has the potential to inform the approval process (e.g. social impact assessments) of such projects, by comparing the text and images from archives to text and images from social media and other contemporary sources.

As geo-spatially-minded disciplines become more intertwined and overlapping in their research endeavours, we agree with Southall et al. [18] that there continues to be a need for cultural gazetteers that describe places in addition to placing them as markers on a map. We see the integration of a discrete global grid into the system architecture of interactive thematic search engines, such as that demonstrated in the Frankenplace http://frankenplace.com/ system by [19], and the enrichment of gazetteers through volunteered geographic information [20] as being two promising steps towards the realization of this achievable vision of the next generation Digital Earth.

4. Conclusion

We have described a pilot study for extracting geospatial data relevant to a focused interest from digital newspaper archives, summarizing those articles, and presenting them to end users on a digital map. This pilot study has informed our vision for employing analytics,
archival data, visualization, and interactive data discovery to provide rich user experiences with data of substantial cultural and social value.

We also believe our pilot study can be improved upon. For example, the classic text-mining approach of TF-IDF weights and treating text as a “bag-of-words” performs well in the task of event detection. We are interested in the use of more complex methods in our workflow to manipulate and study the complete corpus as a Big Data problem in the context of geographic information retrieval. A next step towards this goal is the use of topic modeling (see [21] for an interesting, technical review of this field), which was used by [22] to discover strike events in the New York Times newspaper archives.

References

[1] Goodchild M F, Guo H, Annoni A, Bian L, de Bie K, Campbell F, Craglia M, Ehlers M, van Genderen J, Jackson D et al. 2012 Proc. of the National Academy of Sciences 109 11088–11094
[2] Stokes N, Li Y, Moffat A and Rong J 2008 Int. J. of Geographical Information Science 22 247–264
[3] Yzaguirre A, Warren R and Smit M 2015 IEEE International Conference on Big Data pp 2027–2035
[4] Vasardani M, Winter S and Richter K F 2013 International Journal of Geographical Information Science 27 2509–2532
[5] Bird S, Klein E and Loper E 2009 Natural Language Processing with Python (O’Reilly Media)
[6] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M and Duchesnay E 2011 Journal of Machine Learning Research 12 2825–2830
[7] Clough P D, Joho H and Purves R 2006 Advances in Information Retrieval (Springer) pp 548–552
[8] Hill L L 2000 Research and advanced technology for digital libraries (Springer) pp 280–290
[9] Leidner J L 2007 Toponym resolution in text Ph.D. thesis University of Edinburgh
[10] Montello D R, Goodchild M F, Gottsegen J and Fohl P 2003 Spatial Cognition & Computation 3 185–204
[11] Palacio D, Sallaberry C and Gaio M 2012 Advances in Spatial Data Handling and GIS (Springer) pp 65–78
[12] Pasley R C, Clough P D and Sanderson M 2007 Proceedings of the 4th ACM Workshop on Geographical Information Retrieval GIR ’07 (ACM) pp 72–82
[13] Purves R S, Clough P, Jones C B, Arampatzis A, Bucher B, Finch D, Fu G, Joho H, Syed A K, Vaid S and Yang B 2007 Int. J. of Geographical Information Science 21 717–745
[14] Schockaert S and De Cock M 2007 Proc. of the ACM SIGIR Conference on Research and Development in Information Retrieval SIGIR ’07 (ACM) pp 167–174
[15] Wang W and Stewart K 2015 Computers, Environment and Urban Systems 50 30 – 40
[16] Van Der Walt S, Colbert S C and Varoquaux G 2011 Computing in Science & Engineering 13 22–30
[17] Karpathy A and Fei-Fei L 2015 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 3128–3137
[18] Southall H, Mostern R and Berman M L 2011 International Journal of Humanities and Arts Computing 5 127–145
[19] Adams B, McKenzie G and Gahegan M 2015 Proceedings of the 24th International Conference on World Wide Web WWW pp 12–22
[20] Gao S, Li L, Li W, Janowicz K and Zhang Y 2014 Computers, Environment and Urban Systems
[21] Blei D M and Lafferty J D 2009 Text mining: classification, clustering, and applications 10 34
[22] Hendrickx I, Düring M, Zervanou K and Van Den Bosch A 2013 Proc. of the 3rd workshop on annotation of corpora for research in the humanities pp 25–36

Acknowledgements

We gratefully acknowledge the support of Nova Scotia Environment and The Chronicle Herald.