PaperMiner - A real-time Spatio-Temporal Visualization for Newspaper Articles

Sangeetha Kutty
Queensland University of Technology, Science and Engineering Faculty,
Brisbane, Queensland,
4000, Australia
s.kutty@qut.edu.au

Richi Nayak
Queensland University of Technology, Science and Engineering Faculty,
Brisbane, Queensland,
4000, Australia
r.nayak@qut.edu.au

Paul Turnbull
University of Queensland, School of Historical and Philosophical Studies,
School of Humanities, University of Tasmania
Brisbane, Queensland,
4072, Australia
paul.turnbull@utas.edu.au

Ron Chernich
Queensland University of Technology, Science and Engineering Faculty,
Brisbane, Queensland,
4000, Australia

Gavin Kennedy
Smart Services Co-operative Research Centre,
Queensland Cyber Infrastructure Foundation,
Brisbane, Queensland,
4072, Australia
gavin.kennedy@qcif.edu.au

Kerry Raymond
Queensland University of Technology, Science and Engineering Faculty,
Brisbane, Queensland,
4000, Australia

July 5, 2018

Abstract

In 2005, the National Library of Australia (NLA) began a pilot project to selectively digitise back issues of major Australian newspapers to provide free public access to over 60 million digitised newspaper articles, dating from the first years of Australian colonisation to the early 1960s. Trove, a faceted search engine maintained by NLA, provides access to this very large collection. Unfortunately, Trove lacked any means to filter by location which raised the tantalising possibility of using advanced computational techniques to identify long-term patterns and trends in newspaper reportage of people, events, concepts and many other historical entities.

PaperMiner, which utilises text mining techniques for extracting metadata information was developed that enabled the inclusion of geolocations of the places cited in the newspaper articles, supported the searching of articles by location and visualising the results of searches using both location and time using a map of Australia. Using PaperMiner, researchers could see when and where the anti-Chinese leagues movement started in Australia and how it spread, to better focus their subsequent research.

PaperMiner can be used as a digital humanities tool to assist in research by replacing the tedium of a shallow scan through thousands of Trove search results with a more efficient method that draws the researchers' attention to more significant times and places where their time can be better spent in deeper analysis. In this article, we describe the techniques utilised in creating PaperMiner, and discuss its usability testing with a group of leading researchers in Australian history.

Keywords: Spatio-Temporal Visualization; Text mining; Historical Studies; Trove; News mapping, Named Entity Recognition, Australian newspapers; and Crowdsourcing

1 Introduction

This article reports on the development of PaperMiner[1], a prototype web-based service enabling the discovery and visual analysis of connections in time and space between people, places, concepts and many other historic entities within the 60 million articles comprising the National Library of Australia's Australian Newspaper Service. PaperMiner provides these capabilities through employing a combination of text mining, entity extraction, geotagging and crowdsourcing. The current prototype of the service was created using the Agile development methodology, open-source software programs and a standards-based approach to using geospatial data – all of which proved well suited to our having limited project funding yet high end-user expectations – challenges all too common in the new and still highly experimental field of digital humanities.

[1] http://203.101.226.97/

We first sketch the background to the development of PaperMiner, then briefly review current spatio-temporal visualisation techniques and the conceptual and technical challenges they address. Thirdly, we describe the techniques and technologies employed in creating the current working prototype of PaperMiner, while in the fourth part of the article we discuss how the various technologies and methodologies were used in its development. Section five describes the usability testing carried out with the end users its outcomes. We conclude by reflecting on future research and development of PaperMiner.

Since the beginnings of digital librarianship in the mid-1990s, the digitisation of historical texts and imagery by major research libraries, museums and other cultural institutions in the developed world has proceeded at a pace surpassing the expectations of even the most optimistic digital humanists. The National Library of Australia (NLA), for example, embarked on large-scale collection digitisation and the development of innovative federated discovery services aimed at advancing public knowledge of history and heritage in 1997. Since then the National Library of Australia has won international acclaim for its work in aggregating its own and other digital collections and providing discovery services in partnership with other Australia state libraries, museums and art galleries, so that now anyone curious about particular aspects Australia's history, heritage and culture can have free access to digital surrogates of relevant historical documents, images and other cultural artefacts via Trove, a faceted search engine maintained by NLA.

Perhaps the most remarkable and successful service accessible via Trove[2] is the Australian Newspaper Service. The service has its origins in a 2003 pilot project by the National Library of Australia and the Australian Academy of the Humanities, for which Australian Research Council funding was sought to selectively digitise newspapers of prime interest to scholars of Australian history, society and culture. Since then, the National Library of Australia has invested a considerable sum in digitising Australia's surviving newspaper record from the first years of colonisation to when issues of particular titles are still subject to copyright, and making this digitised content freely available in a range of downloadable formats, Currently, the Australian Newspapers Service provides access to over 60 million digitised newspaper articles. While originally envisaged as meeting the needs of researchers in the humanities and social sciences, the Australian Newspapers Service has proved immensely popular with ordinary Australians. Currently it attracts 100,000 engaged users on any day.

The creation by OCR of raw ASCII text to render the myriad articles comprising the Australian Newspapers Service searchable raises the exciting possibility of employing advances in data mining techniques, entity recognition and geo-tagging to identify and analyse relations between historical people, places, concepts and many other entities in past time and space. As much became clear to historians and information scientists based at several Queensland universities, after learning that the NLA was willing to provide us with a copy of the raw text enabling faceted searching of the Australian Newspapers Service for students' data mining projects. On

[2] https://trove.nla.gov.au/newspaper/?q=.

learning of this development, several historians were drawn to wonder whether it might be possible to apply data mining, entity recognition and visualisation techniques to the Australian Newspapers Service raw text to map historical phenomena, such as the outbreak and relative intensity of violence occurring with the spread of white settlement and dispossession of Aboriginal people in central and northern regions of Australia during the last third of the nineteenth century. Since the early 1990s, the extent to which violence was a systemic element within Australian settler colonialism has been the subject of lively public debate. Computationally-based analysis of the Australian Newspapers Service text appeared to offer the possibility of gaining not only a more accurate picture of where, when and to what extent violence accompanied the spread of white settlement, but also what factors might account for some frontier regions witnessing more clashes between Aboriginal men defending their ancestral country and armed settlers and mounted police. However, in the course of discussing the potential of data mining, entity recognition and visualisation to generate new insights into Australian frontier history, it became clear that much new knowledge of a myriad other aspects of Australia's history since the beginnings of European colonisation could be gained should it prove feasible to build a web-based analytical service capitalising on NLA's achievement of digitising the nation's vast newspaper record. And such a service would not simply benefit historians and other researchers across the spectra of the humanities and social sciences, but also enable university and school students, and the general public, to visually explore connections, patterns and trends as they developed over time in various Australian locations.

And thus, it was that research and development of a PaperMiner prototype began, with $40,000 seed funding from the Faculty of Arts at the University of Queensland and generous in-kind support from the Australian Government and industry funded Smart Services Co-operative Research Centre.

2 Literature Review

The first question to address was obviously what techniques for the retrieval and assessment of information from large historical documentary corpora would best enable the exploration of past phenomena of potentially considerable social and cultural complexity. As is well known, traditional information retrieval systems and search engines usually match user queries against defined document collections and return results in the form of a linear list [3]. So, when searching large documentary corpora, queries can produce long and commonly disorganized listings of hundreds or thousands of items that are arranged by the search system on the basis of the relevance of the query entered by the user and the nature of the document collection. The resulting lists can then be sorted on the basis of dates or relevance. However, the identification of relevant information is consequently a time-consuming iterative process, by which users determine relevance based on assessing the information in the retrieved results. Users of popular web-based search engines often spend large amounts of time reformulating search terms to find information meeting their needs [18]. Search goals are often only partially fulfilled and the user is required not only to pose a series of queries, but also to navigate manually through what may be a very long list of search results to gain some insights about

their topic. There are several open access search engines such as CORE (https://core.ac.uk/search), BASE(https://www.base-search.net/) in addition to the popular search engines such as Google, Yahoo and Bing for searching a collection. But most of these systems return results in the form of linear result lists. Moreover, traditional information systems are known to make the process of information retrieval discontinuous and thus cause difficulties for users in utilising the search results [10].

Efficiently discovering relevant information in the Australian Newspapers Service corpus is less of a problem, thanks to NLA's development of a search facility based on classifying documents by characteristic features, or facets, which can be searched for and ordered in multiple ways. The most common mode of searching the Australian Newspapers Service corpus entails searching for particular entities using subject terms chosen by users, to which filters can be applied to restrict results within parameters such as date ranges and the content within nominated newspaper titles. Even so, the experience of Australian Newspapers Service users wanting to comprehensively investigate particular historical phenomena provides further confirmation, if it were needed, of the high frequency with which information overload and mismatch can occur when using conventional modes of information retrieval to find specific content in large documentary corpora [9].

As Peuquet points out, *to exist is to have being within space and time* [15]. Establishing the connections between entities in space and time is a necessary prerequisite for understanding material and cultural phenomena. There have been several efforts to develop ways of visualising complex entanglements of specifically located entities by extracting implicit knowledge and also loosely related information, thereby providing a holistic view of the essential attributes and qualities of the phenomenon in question [29]. An early work in this spatio-temporal arena is Space-Time Cube (STC), which has been used to assess four representational models for spatio-temporality: (1) space-time cubes, (2) sequential snapshots, (3) base state with amendments and (4) space-time composites. Kraak initially conducted a series of interactive operations on the space-time cube, and revised the cubes from the perspective of time, geographic and revisualisation [11]. More recent developments include Moving GeoPQL [7], which allows users to formulate spatio-temporal queries and generate visualisations of the results. Another initiative by Thibaud et.al [23] has focused on visualising complex data using a graph model and a spatial database to support the extraction of entities that are distant in time and space.

For an effective visualisation, data representations should provide comprehensive knowledge to end users. A data representation model needs to be grounded in intelligent information synthesis and knowledge extraction following cognitive principles. This has been a guiding principle in the development of PaperMiner. Another drawback of the existing spatio-temporal visualisation techniques is that models and methods originating from different disciplines can be complementary in various respects, but they are often hard to combine. To improve meaningful interpretations of visualisations and to enhance further interactive analytical operations, previous work included techniques from cartography, geographic information science, visualisation, and visual analytics [11]. They have provided users with easy to

understand simple operations, such as zoom in and out, pan in and out and how select an item; but they do not provide insights into these visualisations as illustrated by Zhong et. al. [29]. Hence, we sought to address these issues and develop the PaperMiner system so that it could facilitate providing insights to end users by relating the loosely coupled information.

A text mining tool Texcavator [25] produces wordclouds, timelines and visualisation
which focuses on presenting the role of reference cultures in debates about social issues and collective identities. However, our tool is openly accessible and utilises the Trove API to dynamically fetch information.

3 Background of the technologies used in PaperMiner

Several approaches were employed in creating the prototype PaperMiner system. We followed the Agile methodology due to the short duration and the nature of the project, which required the delivery of a stable version of PaperMiner meeting the analytical needs of the principal users of the system within a 6-month time-frame. Several approaches to extracting useful information from the newspaper articles with potential for visualization were evaluated, with the project team deciding to employ a particular tool for entity tagging that best suits user needs. We also adopted a crowdsourcing approach to enhancing and refining the geo-location of entities and sought to address issues relating to the interoperability of PaperMiner with Trove, the NLA's faceted search engine.

3.1 Agile methodology

The popular Agile software project development methodology entails iterative development over well-defined project phases, starting with understanding user requirements, then implementing and testing them with the group of individuals who specified the system's requirements [14]. This approach overcomes the shortcomings of popular sequential development methodologies such as Waterfall [22] [19]. The choice of the Agile methodology over more traditional development paths was also due to its emphasis on requiring active user involvement in defining the requirements necessary for a system such as PaperMiner to meet the prime need of users – in this instance the ability to discover and visually analyse significant connections, patterns and trends within Australian history. The Agile methodology places great emphasis on communication between users of a software tool or service and its developers from the outset. PaperMiner's development path thus entailed regular meetings of users and developers which first sought to identifying the core functionality of the system through describing how the users - a group of university-based historians with considerable experience in conventional modes of research, and some expertise in computationally-based analysis of historical data - imagined using the system in the context of their current and likely future research. The resulting scenarios formed the basis of implementation tasks and an initial attempt was made to prioritise the system's attributes on the basis of desirability, technical feasibility and likely costs. The resulting specification of the system's functional requirements was then presented to the users, who were asked to prioritise the identified requirements in terms of low, medium and high priority. Once this had been done the development of the system got underway with a series of fortnightly iterations including testing and demonstrating

what had been built to the users, who then either approved what had been achieved or requested that modifications be made. Thus, the stable prototype of PaperMiner emerged through incremental development with small, frequent system releases resulting from the users' engagement with the development team.

It is also worth emphasizing that the attraction of the Agile methodology was in no small measure due to our having a limited budget with which to build a working system satisfying core user requirements in a time frame of approximately six months.

3.2 Named Entity Recognition (NER)

Named Entity Recognition (NER) is a subtask of information extraction that identifies sequences of words in a sentence that represent entities, such as geographical locations, persons, organisations and cultural phenomena, into categories [27]. Some NER systems use additional features such as Part of Speech (POS) tags and external sources such as WordNet, Wikipedia and Gazetteers.

There are several tools that are available to conduct NER. The more popular include the Stanford Named Entity Recognizer [8], the Illinois Named Entity Tagger [17] and Senna from the NEC laboratory [5]. The Stanford tool, CRFClassifier, provides a general implementation (in arbitrary order) of linear chain Conditional Random Field (CRF) sequence models. Its features extend to enabling users to build sequence models [13]. The Illinois Entity Tagger [17] utilises external knowledge resources and non-local features that are the cluster of words generated from the given text to identify named entities in the text. This tagger performs well with data providing contextual information to the text commonly found on Webpages. Senna, on the other hand, is a simple and fast tool that not only outputs labelled POS tags and chunks (CHK) but also outputs semantic role labels (SRL) [6]. In this study, we chose Senna to annotate the dataset due to its accuracy and efficiency for the task at hand. We found that Senna yielded relatively higher precision and recall (above 90%) in the Trove dataset with faster processing speed compared to other tools such as CRFClassifier and the Illinois Entity Tagger.

3.3 Crowdsourcing

The explosive development of the World-Wide-Web since the mid-1990s has seen crowdsourcing employed in research projects in both the humanities and sciences. Perelman School of Medicine at the University of Pennsylvania engaged thousands of people to provide data and conduct analysis that has led to a significant improvement in the quality of medical research with reduced costs [16]. This instance of crowdsourcing is especially significant as biomedical research requires high standards of accuracy in data collection and processing. There have also been recent research ventures employing crowdsourcing to develop more effective means of detecting and responding to urban emergencies such as fires and extreme weather events [28]. In this PaperMiner project, crowdsourcing was adopted in an attempt to

mitigate the effects of the constraints on time, budget and resources on developing the system. We utilised a total of 10 volunteers who were interested in the development of PaperMiner, some of whom were experienced users of the NLA's Trove search service. These users were used as our crowd to run the Java-based application that uses the google maps API to lookup the geocode of the location identified in the newspapers. Due to the public license restriction of the google maps API, without the crowdsource option to run the google maps API, it would have taken the developer a few months to lookup the geocodes. This was a smaller number of volunteers than was desirable, but using a larger number of people would have exceeded the capacity of the development team to validate the information provided.

3.4 Search Service

Trove is a free search service provided by the National Library of Australia that allows users to investigate an extensive collection of documents, images and other media. At the time of developing the PaperMiner prototype, users of Trove could search more than 361 million articles using simple keywords, or more advanced query constructs[3]. Trove categorises this wealth of information into ten "zones" ranging from journal articles, books, newspapers and pictures to archived websites. Our key zone of interest was, of course, newspapers. Trove enables the searching of the Australian Newspaper Service, which comprises newspapers from the turn of the nineteenth century to the 1950s, when copyright restrictions begin to apply (although some copyright holders have licensed the non-commercial reproduction of titles beyond the 1950s).

To enable text searching of Australian Newspapers Service content, images of the original newspaper pages have been created, and individual articles identified and transcribed by Optical Character Recognition (OCR). The ability of OCR software to accurately convert newspaper articles to searchable text is of course determined by the legibility of the original newsprint. For much of the nineteenth century, type was extensively reused and thus subject to wear in ways limiting the extent to which OCR software can be trained to read newspapers using a particular typeface. The National Library of Australia has sought to improve the accuracy of the Australian Newspapers Service search text by encouraging readers to provides corrections which are automatically incorporated into the search text and immediately made available to all (a list of recent corrections available at http://trove.nla.gov.au/newspaper/recentCorrections).

The designers of Trove have also provided an HTML based Application Programming Interface (API), which allows the service to be incorporated into third party systems. Trove, like any request/response service with an API, is susceptible to degradation when heavily used, as often happens on weekends when many Australians browse the Australian Newspapers Service in the course of researching their family history or a great many other aspects of Australia's history. Users of the API are thus required to obtain a "key" which must be passed to Trove with each request. This allows the Trove server to track users and prevent individual API users making excessive requests. Currently both commercial and non-commercial keys

[3] http://trove.nla.gov.au/

are available, with the latter designed for personal use and free of cost. PaperMiner requires users to obtain and use a non-commercial key before they can use the system.

4 The proposed PaperMiner system

The primary objective of PaperMiner is to allow users to visually analyse the content of Australian newspapers published since 1803 that the National Library of Australia has digitised in association with Australian state and municipal libraries and made searchable by the Trove search engine. As previously mentioned, the use of the Agile methodology allowed iterative development of the system through the phases of defining requirements, then designing, implementing and testing them. Core user requirements were that PaperMiner allow users to:

- Geo-spatially locate places where articles within the Australian Newspapers Service were published.
- Geo-locate places referred to in articles.
- Represent the results based on the date when articles were published.
- Relate results based on similarity of terms that are not keyword terms.
- Map the progress of conversations within the newspaper articles based on the search results.

PaperMiner creates visual clusters based on the time and space information for a given query. The information about time (when) and location (where) along with query (what) enable the users to understand the cause and effect relationships, and thus the nature and structure of the processes. Figure 1 shows how visual clusters are obtained by plotting the time and space information along with the query. The size of the circle in this figure indicates the number of references for a given place in all of the articles text that were retrieved for a given query. This helps to identify the relative importance of a given location in a particular context. The interpretation of these views helps to generate information about events which in turn helps to form relationships and characteristics thereby deriving geo-temporal relationships and patterns as new knowledge. Figure 2 shows the proposed PaperMiner system and its workflow that will be discussed in the following subsections.

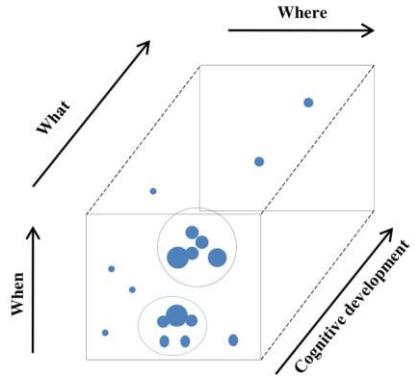


Figure 1: Visualisation dimensions for PaperMiner

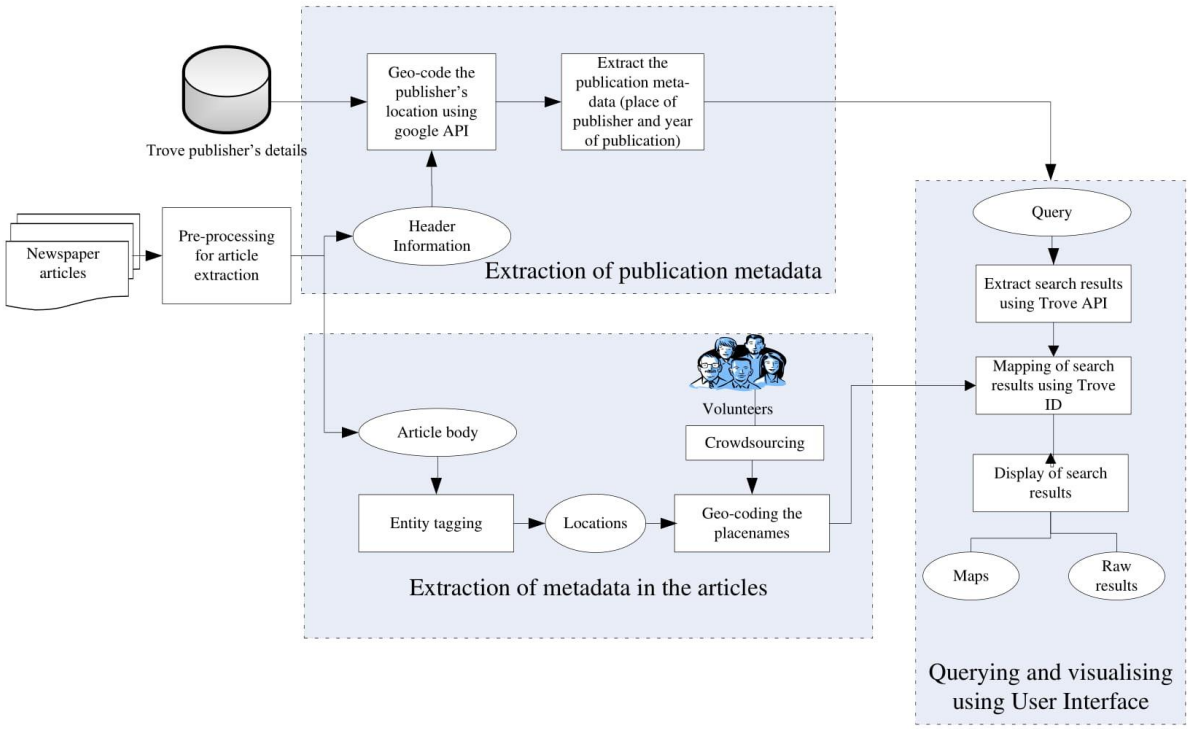


Figure 2: The proposed PaperMiner System: Workflow

4.1 Pre-processing for article extraction

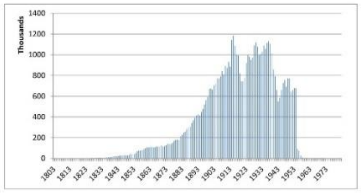
The OCR text used to search the Australian Newspapers Service was provided to us as a single large text file of 260GB. The text reflects the classification of content in the Australian Newspapers Service as one of six types: Literature, Detailed Lists Results, Guides, Advertising, Articles, Family Notices and Others. Figure 3(a) shows the number of documents that belong to each of the six types. Figure 3(b) and 3(c) shows the number of publication titles based on place of publication and year of publication respectively. The historians for whom PaperMiner was designed were mainly interested in articles due to the information stored within being highly relevant to their research aims. These newspaper articles occupied more than 75% of the collection with a coverage of events over most of Australia.

Zone Name	Number of articles
Literature	12,289
Detailed lists, results, guides	7,516,250
Advertising	11,058,829
Article	58,549,810
Family Notices	703,846
Others	3

(a)

Places of publication	No. Of publication titles
Sydney	35
Adelaide	19
Hobart	14
Perth	14
Brisbane	11
Melbourne	11
Darwin	8
Kiama	8
Queanbeyan	7

(b)



(c)

Figure 3: Statistics of the dataset

Each article in the collection was identified by the category to which it had been assigned by the designers of Trove in associated metadata. A new document was then created with the name of its identifier (id) from metadata. The next step was to attach the content for the given document. Pre-processing of the content was undertaken to remove line-breakers and unnecessary characters so that content was in the form of sentences for ease of processing by the parser as shown in Figure 4. This process resulted in the identification of 441,147,822 unique terms with 631,466,061 as the total number of terms in a data collection of 58,549,810 articles.

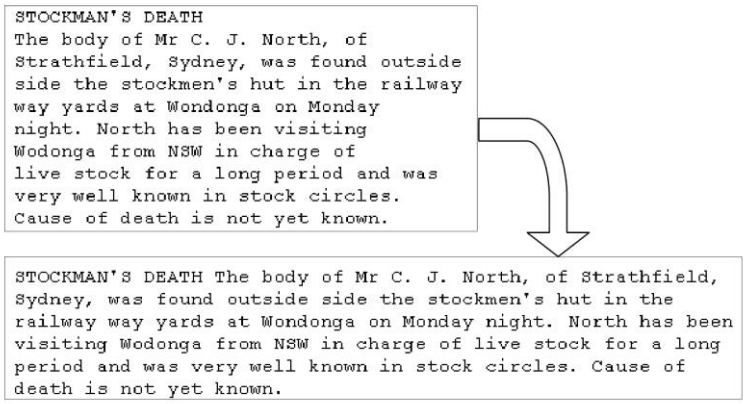


Figure 4: Transformation of Original Content Format

4.2 Extraction of the publication metadata

Metadata was created by identifying place of publication and the year of publication in each article. This information assisted in identifying where the article was published and its relevance to the given query. The place of publication and the year of publication information were included in the header of each of these articles. There were found to be 384 publication titles in the entire collection with 143 unique places of publication for these titles. Figure 3(c) shows the years of publication in this collection containing articles from
1803 to 1982 spanning 180 years. It can be seen that the collection provided to us contain more number of articles from 1915 to 1938. Due to the copyright issues, only a handful of articles were included in the collection beyond 1960s. From the user's perspective, the year of publication helped provide a temporal overview of the search results in terms of time span and important time periods relevant to a query. For instance, a query employing the term "celestials" – a derogatory term for Chinese men commonly used by Australians of British descent – returned a large number of articles, the majority of which could immediately be seen to have been published between 1880 to 1930 with peaks in the years 1892, 1894, 1896 and 1902. A screenshot of the output for the query "`celestials'" is shown in Figure 6. Figure 3(b) list the top-10 places of publication for these titles. What the graph reflects are the periods of intense agitation occurring in the decade before the foundation of the Commonwealth of Australia in 1901 aimed at preventing Chinese migration. What the visualisation discloses will come as no surprise to historians, but it indicates how the PaperMiner system can help in the

discovery of trends and patterns in Australian history that merit closer scrutiny of the relevant newspaper articles.

4.3 Extraction of metadata in the articles

This step involved extracting metadata within articles, notably place names mentioned in articles, such as country names, city names, suburbs or landmarks. This metadata helps in the discovery of where events referred to the article have occurred. Many articles provide place names in brief text appearing either at the start or end of the article. However, we were not only interested in these 'datelines', which generally provide information about where an article was published, but not the location where an event occurred. We also aimed to register places discussed in an article relevant to the events or other phenomena described therein. Named Entity Recognition (NER) was utilised to extract place names within articles using the NER tool, SENNA [6]. SENNA builds on the idea of deep learning of extracting useful features from unlabelled text. This unsupervised learning phase occurs using auto-encoders and neural networks language models. It allows the mapping of words into another space of representation that has lower dimensionality. SENNA maps every word available in its 130-thousand-word dictionary to a vector of 50 floating numbers. These vectors are then merged into a sentence. The same architecture is then trained on different tasks using annotated text to generate different classifiers. The big advantage of taking this approach is the lesser amount of engineering that it requires to solve multiple problems. Also, SENNA has been noted to perform much better in comparison to other state-of-the-art entity tagger both in terms of accuracy and speed [2].

Mr	NNP	B-NP	O	-	I-A1	O	I-A1	O
C.	NNP	I-NP	B-PER	-	I-A1	O	I-A1	O
J.	NNP	I-NP	I-PER	-	I-A1	O	I-A1	O
North	NNP	E-NP	E-PER	-	I-A1	O	I-A1	O
,	,	O	O	-	I-A1	O	I-A1	O
of	IN	S-PP	O	-	I-A1	O	I-A1	O
Strathfield	NNP	S-NP	S-LOC	-	I-A1	O	I-A1	O
,	,	O	O	-	I-A1	O	I-A1	O
Sydney	NNP	S-NP	S-LOC	-	I-A1	O	I-A1	O
,	,	O	O	-	E-A1	O	E-A1	O
was	VBD	B-VP	O	-	O	O	O	O
found	VBN	E-VP	O	found	S-V	O	O	O
outside	JJ	B-NP	O	-	B-AM-LOC	O	O	O
side	NN	E-NP	O	-	E-AM-LOC	O	O	O
the	DT	B-NP	O	-	B-A1	O	O	O
stockmen	NN	E-NP	O	-	I-A1	O	O	O
's	POS	B-NP	O	-	I-A1	O	O	O
hut	NN	E-NP	O	-	I-A1	O	O	O
in	IN	S-PP	O	-	I-A1	O	O	O
the	DT	B-NP	O	-	I-A1	O	O	O
railway	NN	I-NP	O	-	I-A1	O	O	O
way	NN	I-NP	O	-	I-A1	O	O	O
yards	NNS	E-NP	O	-	I-A1	O	O	O
at	IN	S-PP	O	-	I-A1	O	O	O
Wondonga	NNP	S-NP	S-LOC	-	I-A1	O	O	O
on	IN	S-PP	O	-	I-A1	O	O	O
Monday	NNP	B-NP	O	-	I-A1	O	O	O
night	NN	E-NP	O	-	I-A1	O	O	O
.	.	O	O	-	I-A1	O	O	O

Figure 5: Output of NER from Senna

For NER, Senna uses the CONLL tagset which basically contains four types of phrases:

person names (PER), organizations (ORG), locations (LOC) and miscellaneous names (MISC). The NER column uses tags in the IOBES format. This consists of the tags B, E, I, S or O where S is used to represent a chunk containing a single token. Chunks of length greater than or equal to two always start with the B tag and end with the E tag. I is a token inside a chunk, O is a token outside a chunk. A chunk is a Named Entity. As a result, in the NER column, four types of categories and entities tagged with the label LOC which includes S-LOC, B-LOC, I-LOC and E-LOC are stored. For example, in Figure 5, entities Strathfield, Sydney and Wondonga are extracted and stored as geo-locations.

4.3.1 High performance computing (HPC)

Though Senna can process each article in this collection in less than 5 seconds, the number of articles to process was about 60M. The sheer number of documents poses difficulty. We used the high performance computational systems, the Silicon Graphics International Corp (SGI) Altix XE computational cluster that includes 128 compute nodes with 64-bit Intel Xeon processors, was used. The HPC resources provided the ability to process multi-threaded applications simultaneously. The collection of articles was packaged into smaller tar files with each tar file containing 10K documents. The purpose of tarring helped to reduce the processing time as the time in reading the tarred file was less than the individual file. At each execution 10 documents were processed and there were 120 processes were running in a given time. This helped the entire collection to be processed in less than 2 weeks' time which would have otherwise taken months or years without this facility.

4.3.2 Geocoding

Geocoding can be defined as a process of converting location addresses (e.g. 1600 Amphitheatre Parkway, Mountain View, CA) into geographic coordinates (e.g. latitude 37.423021 and longitude -122.083739) to be located on the world map. The Geocoding API is generally designed for geocoding static (known in advance) addresses for placement of application content on a map. The API based implementation is suitable for our application where we can identify the place names in the articles using entity tagging.

Geocoding is a time and resource intensive task due to our very large number of documents therefore instead of identifying locations for every article, the locations of articles were indexed and then the geocodes were looked up. However, there is a drawback of this technique which is the loss of context information. For instance, there could be two different places with the same name – i.e. the Brisbane suburb of Oxley in Queensland, Australia and Oxley in Victoria, Australia. The Australian gazetteer of place names was used to resolve these types of conflicts by including names of popular locations over the unpopular.

Due to budgetary constraints, we could only utilise the public license of google maps API. This constrained us to 2,500 query requests per day. Often our requests exceeded the 24-hour limit imposed on the public license, and the Geocoding API stopped working for a few days. Crowdsourcing was used to overcome this problem. A group of people who were interested in this project served as crowdsourcing volunteers. Each volunteer was allocated 2,500 locations to look-up per day. We created a Java-based application that used Geocoding API to lookup

the locations identified in the newspaper articles. These volunteers ran the Java-based application on their personal computers that would determine the geocode for a given location, and collate all the geocodes for the given batch. The second task was to manually check locations were correctly coded. A total of 10 volunteers completed the given tasks in about a week. Locations and corresponding geocodes were collated from all the volunteers, then indexed along with their corresponding geocodes in database tables for easy look-up. Using the indexed locations, the articles were then parsed to map their locations with their corresponding location indices. The geocoded locations for the place names and publishers for each of the articles were then stored in a database using their Trove identifiers which were then matched to the Trove identifiers of the records retrieved for a given query.
Not all locations, however, were correctly tagged. The geocode of Formosa, for example was found to refer to a small town in Argentina when the place mentioned in the article of interest was the former name of Taiwan, and many locations were incorrectly attributed to smaller towns in the United States.

4.4 User Interface

The User Interface (UI) was built using Javascript and HTML and hosted on Apache 2.2 with Tomcat server 7.0.34 using Catalina for JavaServer Pages(JSP) and MySQL5.5 as the database. The UI was developed adopting the focal eight "golden principles" of Human Computer Interaction (HCI) design [21], with a particular focus on striving for consistency, minimal user input, reduction of errors, and ease of traversal.
A fixed layout was designed with all the web pages using the same layout of navigation icons and menu items. Adhering to the UI design principle of reducing errors, users were directed to enabled menu options, for example query and view functions, only after accessing a user login / register page, and then common items such as Help, Partners and Contact. This is also a good security feature that helps prevent malicious attacks on the system by unknown users or malware. Additionally, the adoption of the minimal user input design principle allowed users to use click and select rather than typing more text which helps to avoid several typo errors. The ease of traversal principle was adopted in the view of the design (explained in detail below) enabling users to switch between the results of their queries appearing on a google map, with temporal information presented on an interactive "drag-able" timeline, or as lists of "raw results". Users were also given the ability to sift through results without needing to wait until all results have been fetched, which has the advantage of allowing users to adapt queries when initial search results appear unsatisfactory.

Once the users have logged into PaperMiner, they are presented with the following menu options: Query, View, and User Details, which allow users to control query execution, display selection and user management.

The query menu item allows users to create and execute queries, examine the progress of query, and to save and re-execute queries. Within the query option there are three submenus namely New, Current and Saved.

New - This menu item allows users to perform searches of the Australian Newspapers Service on user provided text (query), which can be a single word, multiple words, or phrases enclosed in quotes. The search is directed to the Australian Newspapers Service via the Trove search engine, with a small number of records returned to provide a quick initial response, followed by the gradual return of further results. The search continues in the background and returns the results of the first one hundred results.

Current - This menu option is the default menu option that allows users to examine the state of their current or previous query. Users have the ability to pause and resume their query at will. The status of the query is also provided, i.e. the number of records so far returned, the total number of records that a given query has found, and the time the query has spent retrieving records from Trove. Since Trove restricts users of an API Key to a maximum of 100 records per request, PaperMiner gives users the option of saving a given query so that it can be viewed later. Also, there is an option, "Revise query," that allows users to refine their search.

Saved - PaperMiner allows each user to save up to 20 queries for convenient re-execution. Selecting the option "Query Saved" displays past queries, allowing users to select one for re-execution by clicking the query description next to a check box. This takes users to the New Query view, in which the users can run or modify the query string. There is also an option enabling saved queries which are no longer required to be deleted, with deletion restricted to one query at a time to prevent accidental loss.

PaperMiner's View menu enables users to exploit its real analytical power. As useful as comprehensive lists of references to phenomena of interest mentioned in newspaper articles may be, reading each to determine its relevance to a user is time consuming. The View menu provides users with the ability to geospatially and temporally visualize their search results, as well as providing different ways of viewing and sorting them.

Map View - Marker pins are displayed on a geo-spatial (currently google) map denoting either the places where articles referring to phenomena of interest were published, or geo-tagged place names appearing in the article. A pair of radio buttons allow users to swap between viewing pins representing places of publication and places mentioned in articles. Selecting the "Place names" option displays colour-coded markers for locations data found in search results.

Timeline - A timeline indicates how the results are distributed by year of publication, with the density of results for a year, and references to particular locations indicated by the colour of the timeline bar and map marker pins. A legend to the right of the map explains the colour coding. Further assisting users to assess their findings are two sliders linked to the timeline which allow the date range of the map pins to be set to a specific number of years. Users can also easily move from viewing the results of their query to a "Raw Results "view, showing only the records retrieved from the Australian Newspapers Service for the date range set in the map view in ascending date order. The listing of raw results can then be sorted using radio buttons at the bottom of the display; and by clicking on an item in the list, a "snippet" of the full article text around the occurrence of given search term(s) are provided (the length of the

snippet is determined by the Trove search engine). Figure 6 is a screenshot of the visual output for the query “celestials” with place names contained in relevant articles displayed.

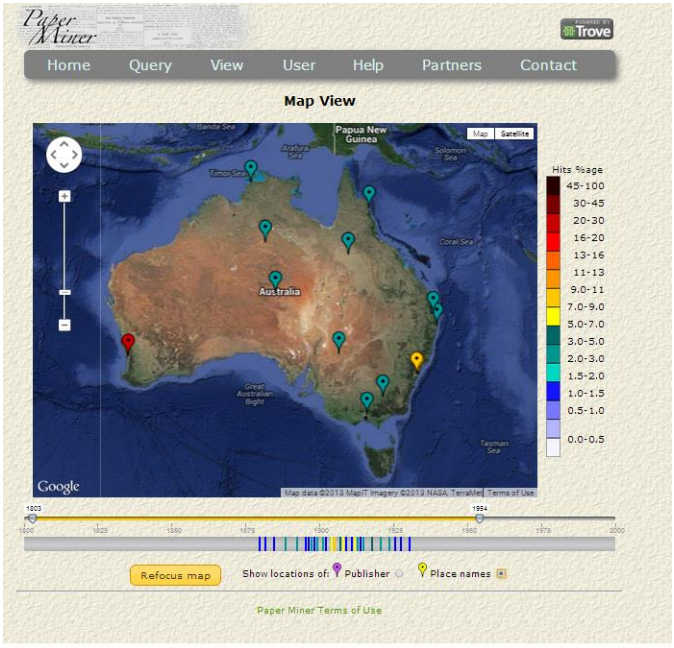


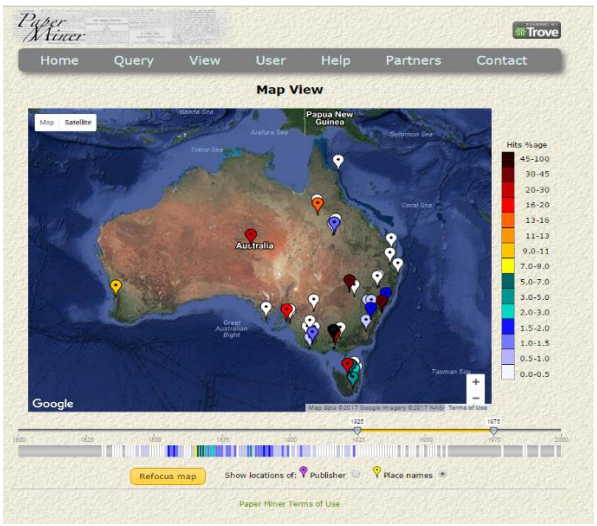
Figure 6: Map display of the places mentioned in articles returned on using the query term Celestials

Edit Australian Newspapers Service Text - PaperMiner also provides users with the option of loading the full text of articles in the Australian Newspapers Service via a radio button. Additionally, there is a “Fix in Trove" option which directs users to the relevant article in the Australian Newspapers Service via the Trove search engine, from where they can correct errors in the OCR text of the article. After saving their changes to the Australian Newspapers Service search text, users can press a “Refresh Record” button to update the view of the text with PaperMiner with the changes they have made.

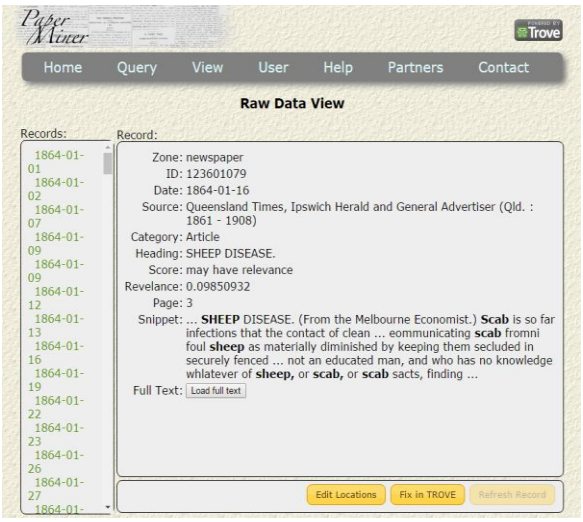
Edit Locations – In the raw results view PaperMiner users can assist each other by improving the quality of searches through adding geo-codes to places mentioned in articles or revising or removing existing geo-codes. When users are presented with a view of an article within the Australian Newspapers Service (snippet, plus full text) they are also provided with a list of place names associated with the article, if any. Should a location be incorrect - say Newcastle, England, when Newcastle, Australia is referred to in an article - the location can be deleted by clicking a check-box beside the name of the place and pressing “Remove Location”. This marks the location as “struck-through” so that it will not be inadvertently restored by future users. Locations can also be added to records. To reduce the risk of adding incorrect location codes, users are prompted to search the PaperMiner database, or the Google Geocoding facility to add new place names.

PaperMiner aims to visually illuminate connections between past phenomena described in Australian newspapers as shown in Figure 7. Firstly, the list of articles returned on the query

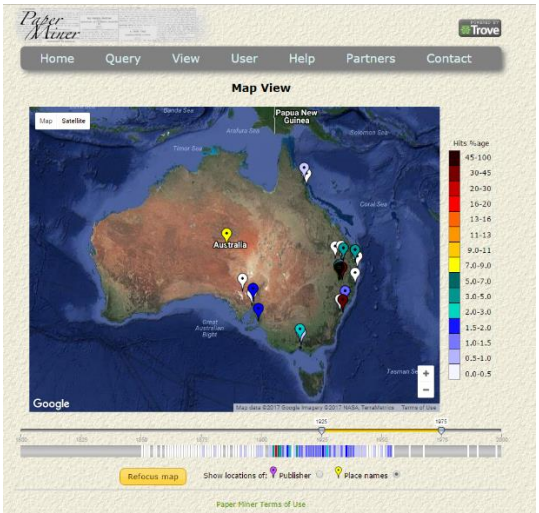
"sheep scab" provides new insight into the extent to which wool production in various parts of Australia was, at various times, affected by sheep becoming infected with this disease caused by parasitic mites. Secondly the term "Myall Creek" returns a wealth of articles concerning the massacre of innocent Aboriginal men, women and children at Myall Creek in New South Wales by local white settlers in 1838. Eleven of the twelve killers were arrested, brought to trial and hanged. The articles highlight how settler society responded to Myall Creek and the colonial administration applying the law equally to Aboriginal people and settlers, and how the massacre and capital punishment of its perpetrators figured in public and government responses to frontier violence until the turn of the twentieth century.



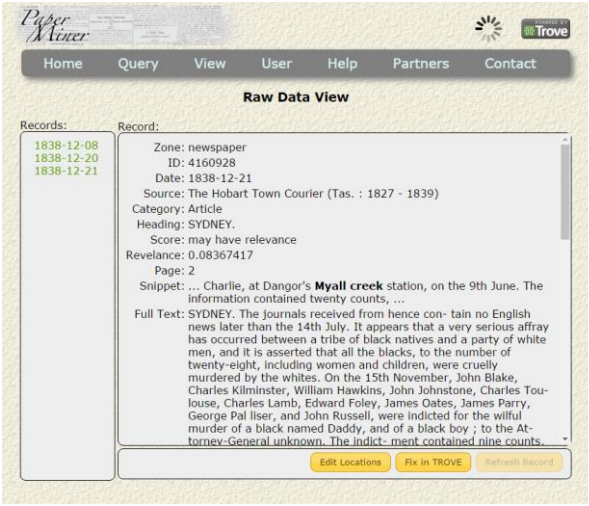
(a) Visualisation of "`Sheep Scab'" query on map



(b) Raw results of "`Sheep Scab'" query



(c) Visualisation of "`Myall Creek'" query on map



(d) Raw results of "`Myall Creek'" query

Figure 7: Visualisation Output for two queries – "Sheep scab" and "Myall Creek"

PaperMiner's UI not only displays the results of user queries in forms aiding historical analysis, but also enables users to collectively improve the quality of search results by providing them

with the means to add or edit the geo-location of places named in Australian newspaper articles, and the capability to correct errors in the OCR created text employed to render Australian Newspapers Service content searchable.

5 Usability testing

Usability testing was employed to evaluate how well the version of PaperMiner built within the resource and time constraints we experienced met the needs and interests of professional historians; but we were also interested to learn how amateur historians and novice users engaged with the system. Our pool of evaluators was small. It comprised several experienced historians, one amateur historian, and three "novice" users. However, previous research suggests that five users are enough to evaluate a system such that 80% of usability problems are identified [24, 26].

5.1 Usability Evaluation Criteria

Our testing was focused on understanding the realness (or relevance), validity, efficiency, consistency, accuracy, recall and emotional response for PaperMiner. These metrics can be defined as:

Realness (or relevance) - Identifying if a given usability problem is a real usability problem or not by comparing with the standard list of usability problems.

Validity - The ratio of the number of real usability problems over the total number of usability problems that were identified.

Efficiency - How much time, and how many steps, are required for people to complete basic tasks? (For example, complete a given task such as search using a given query, visualisation on map, display of raw results and save queries.)

Consistency - How the different web pages follow consistency in layout and navigation?

Accuracy - How many mistakes did the evaluators make? (And were they "fatal" or recoverable with the right information?)

Recall - How much does the evaluators remember afterwards or after periods of non-use?

Emotional response - How does the evaluator feel about the tasks completed? Is the person confident, stressed? Would the user recommend this system to a friend?

5.2 Tasks selected

Each of the users were presented with the following two tasks:

- To run a given set of queries and visualise the results on (a) Map (b) Raw results.

- To run a given set of queries and save the results and use it for future use.

Table 1: Sample Queries

Queries			
Celestials	Flooding	Cyclone	Sheep scab
Fossils	Snake bite	Drought	Myall creek
Manslaughter	Emancipists	Inquest	Locusts
Camels	Coroner	Influenza	Aborigines
Gold	Native police	Plaque	Natives

Table 2: Usability test results

Criteria						
	Historian 1	Historian 2	Student 1	Student 2	UI expert	Total
Realness	4	4	3	3	4	18
Validity	4	4	4	3	4	19
Efficiency	4	5	4	3	4	20
Consistency	5	5	5	5	5	25
Accuracy	5	5	4	3	4	21
Recall	5	5	4	3	4	21
Emotional response	5	5	4	3	5	22

The evaluators were provided with 20 sample queries as shown in Table 1. They were asked to evaluate the system based on these queries using a Likert scale of 0 to 5 with 0 representing very dissatisfied and 5 for very satisfied.

With every evaluation of a system such as PaperMiner there will be usability problems. However, a majority of users captured about 40% of the usability problems, out of which 35% proved to be valid issues as indicated in the usability tests described in Table 2. The efficiency of the system was high, with users completing most of the set tasks in the allocated time. Consistency, recall and emotional response rated very high. There was one issue in respect of recall, the login procedure, although the fact that options were disabled helped users to recall the procedure.

6 Conclusion

In this paper, we have described the aims and functionality of PaperMiner, an interactive web-based system that enables geo-spatial and temporal analysis of diverse historical phenomena in one of the world's largest online newspaper archive. PaperMiner provides this service by applying data mining techniques to some 260 gigabytes of OCR created text rendering the content of Australia's newspaper record over two centuries searchable. The system at its

current stage of development has proven to allow expert and novice users to explore events, trends and patterns in Australian history so as to generate new knowledge and insights beyond what would be achievable by using the faceted search engine hosted by the National Library of Australia to explore the content of the Australian Newspaper Service in an intuitive manner. Instead of a standalone tool, PaperMiner is offered as an interactive web- based tool that is accessible using browsers without the need to download any specific plug-in.
Our future work will focus on mining for people's names and identifying the association among them for a given query. We also plan to work on the evolution of news, automatic identification of duplicate articles in the collection that will enable the search engine to group similar articles to help the end users to save time by avoiding them in reading the same news again that were published elsewhere.

In this paper, we present an interactive web-based tool that is accessible using browsers without the need to download it as separate software. More specifically, we present a novel way to visualise information in real-time and spatio-temporal space for several million newspaper articles. It not only visualises the name and place of an article but also visualises the places that were referenced in the article. The reasoning behind using the places that were referenced in the article instead of the place of publication is that in many instances an incident or event could have occurred elsewhere but will be published in a different place. Hence, including only the place of publication may not be useful for those incidents. Crowdsourcing was utilised in a novel way to support with geotagging.

This system provides researchers, historians and school students with unprecedented possibilities to access and analyse the big data collection of public media. It enables the fine-grained analysis of large-scale document collections and empowers scholars to group the concepts which help to identify and distinguish long-term patterns in digitised newspaper articles semi-automatically. This is a generic system that can be implemented to suit the requirements of any application domain and not just limited to newspaper articles.

Several technical solutions have been proposed to deal with this problem such as relevance ranking, personalisation and dual feedback to extract useful information among the returned search results [4]. However, the most promising developments have come from appreciating the capacity of users to make greater sense of the returned results when they are presented in visual forms that can be analytically manipulated so that connections and patterns of potential conceptual, chronological and geo-spatial significance can be seen [12].

For example, the name “celestials” was commonly used with derogatory racialist intent by Europeans to describe Chinese migrants to Australia of the later nineteenth and early twentieth centuries. A query using the term returns links to a great many articles within the Australian Newspapers Service containing references to Chinese men and women, which PaperMiner’s timeline shows to have been published between 1845 and 1935. The accompanying geographical map shows both the places at which newspapers containing articles referencing “celestials” were published, while also allowing users to see places mentioned in the returned articles. Comparing the timeline and map, users can see the term appears with greatest

frequency in the 1850s and 1860s when gold discoveries in New South Wales and Victoria led to over 40,000 Chinese men migrating to Australia. Their presence and the agitations and violence on the part of British, Irish and other European men who flocked to the gold fields it provoked, is well known. But what is interesting about PaperMiner's representation of the results of querying "celestials" is the presence of articles in Sydney and Melbourne based papers published during the economic depression of the 1890s pointing to growing anti-Chinese sentiment in these cities.

References

[1] Facts and Figures. http://www.nla.gov.au/facts-and-figures, July 2013.

[2] Al-Rfou, R., and Skiena, S. Speedread: A fast named entity recognition pipeline. CoRR abs/1301.2857 (2013).

[3] Baeza-Yates, R. A., and Ribeiro-Neto, B. Modern Information Retrieval.
Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[4] Carlson, C. N. Information overload, retrieval strategies and Internet user empowerment. 169-173.

[5] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. Natural language processing (almost) from scratch. J. Mach. Learn. Res. 12 (Nov. 2011), 2493-2537.

[6] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. P. Natural language processing (almost) from scratch. Journal of Machine Learning Research 12 (2011), 2493-2537.

[7] DUlizia, A., Ferri, F., and Grifoni, P. Moving geopql: a pictorial language towards spatio-temporal queries. GeoInformatica 16, 2 (2012), 357-389.

[8] Finkel, J. R., Grenager, T., and Manning, C. Incorporating non-local information into information extraction systems by gibbs sampling. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (Stroudsburg, PA, USA, 2005), ACL '05, Association for Computational Linguistics, pp. 363-370.

[9] Fischer, G. Context-aware systems: the 'right' information, at the 'right' time, in the 'right' place, in the 'right' way, to the 'right' person. In Proceedings of the International Working Conference on Advanced Visual Interfaces (New York, NY, USA, 2012), AVI '12, ACM, pp. 287-294.

[10] Jaakkola, T., and Siegelmann, H. Active information retrieval. In Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic (Cambridge, MA, USA, 2001), NIPS'01, MIT Press, pp. 777-784.

[11] Kraak, M.-J. Visualization viewpoints: beyond geovisualization. IEEE Computer Graphics and Applications 26, 4 (2006), 6-9.

[12] Kules, W., Wilson, M. L., Schraefel, M., and Shneiderman, B. From keyword search to exploration: How result visualization aids discovery on the web. Technical report, University of Southampton, 2008.

[13] Lafferty, J. D., McCallum, A., and Pereira, F. C. N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the Eighteenth International Conference on Machine Learning (San Francisco, CA, USA, 2001), ICML '01, Morgan Kaufmann Publishers Inc., pp. 282-289.
[14] Misra, S., Kumar, V., Kumar, U., Fantazy, K., and Akhter, M. Agile software development practices: evolution, principles, and criticisms. International Journal of Quality & Reliability Management 29, 9 (2012), 972-980.
[15] Peuquet, D. J. Representations of space and time. The Guilford Press, New York, 2002.
[16] Ranard, B., Ha, Y., Meisel, Z., Asch, D., Hill, S., Becker, L., Seymour,A., and Merchant, R. Crowdsourcingharnessing the masses to advance health and medicine, a systematic review. Journal of General Internal Medicine (2013), 1-17.
[17] Ratinov, L., and Roth, D. Design challenges and misconceptions in named entity recognition. In CoNLL (6 2009).
[18] Rezaei, B., and Muntz, A. Methods and apparatus for visualizing, managing, monetizing, and personalizing knowledge search results on a user interface, Sept. 2014. US Patent 8,843,434.
[19] Royce, W. W. Managing the development of large software systems: concepts and techniques. In Proceedings of the 9th international conference on Software Engineering (Los Alamitos, CA, USA, 1987), ICSE '87, IEEE Computer Society Press, pp. 328-338.
[20] Schneider, C., and von Briel, F. Crowdsourcing Large-Scale Ecological Monitoring: Identifying Design Principles to Motivate Contributors. Springer US, Boston, MA, 2013, pp. 509-518.
[21] Shneiderman, B., and Plaisant, C. Designing the User Interface: Strategies for Effective Human-Computer Interaction, 5th ed. ed. Addison Wesley Longman, Reading, Mass, 2009.
[22] Sommerville, I. Software engineering, 9th ed. ed. Pearson, Boston :, 2011.
[23] Thibaud, R., Del Mondo, G., Garlan, T., Mascret, A., and Carpentier, C. A spatio-temporal graph model for marine dune dynamics analysis and representation. Transactions in GIS 17, 5 (2013), 742-762.
[24] Tullis, T., and Albert, B. Chapter 5 - issue-based metrics. In Measuring the User Experience (Second Edition), T. Tullis and B. Albert, Eds., second edition ed.,Interactive Technologies. Morgan Kaufmann, Boston, 2013, pp. 99 - 120.
[25] van Eijnatten, J., Pieters, T., and Verheul, J. Using texcavator to map public discourse. 59.
[26] Virzi, R. Refining the test phase of usability evaluation: How many subjects is enough? Human Factors 34 (1992), 457-468.
[27] Wagner, W. Steven bird, ewan klein and edward loper: Natural language processing with python, analyzing text with the natural language toolkit. Lang. Resour. Eval. 44, 4 (Dec. 2010), 421-424.
[28] Xu, Z., Liu, Y., Yen, N., Mei, L., Luo, X., Wei, X., and Hu, C. Crowdsourcing based description of urban emergency events using social media big data. IEEE Transactions on Cloud Computing PP, 99 (2017), 1-1.
[29] Zhong, C., Wang, T., Zeng, W., and Mller Arisona, S. Spatiotemporal visualisation: A survey and outlook. In Digital Urban Modeling and Simulation, S. Arisona, G. Aschwanden, J. Halatsch, and P. Wonka, Eds., vol. 242 of Communications in Computer and Information Science. Springer Berlin Heidelberg, 2012, pp. 299-317.