GLOBAL KEYWORD TRACKING IN ARCHAEOLOGY

A Thesis

by

YUE YAN

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Richard Furuta
Committee Members,	Yoonsuck Choe
	Luis F. M. Vieira de Castro
Head of Department,	Dilma Da Silva

May 2016

Major Subject: Computer Science

ABSTRACT

With the digitization of information, discoveries of events that previously took much human effort can now be found automatically. As example, we investigate several scandals in the art and antiques area that occurred between 1985 and 2005. In these events, the auction house Sotheby's was suspected to accept or even help the trading of smuggled paintings or antiques and the famous Getty Museum was exposed as purchasing antiques linked to treasure hunters. Discovering these secrets required the hard work of journalists, detectives, TV producers, and so on. The investigators were involved in illegal trades and various dangerous situations during their process of investigation. In comparison, today, with the access to digital version of large datasets, we are able to discover similar events using computationally-based techniques without the high risk and the cost of human labour needed before.

This thesis introduces our tool for extracting keywords, terms and peoples' names from news articles, books, and marking them on an interactive map. We use the *New York Times* as the main resource, extract location terms in each news articles using Gazetteer, extract keywords and people's names in each articles and reduce ambiguity using WordNet. Combining them, we are able to form location-keyword-time pairs for each articles, and together they form a database. Then we build an interactive map based on the database. The map is able to show the relationships between location and keywords. The linkages between two or more people or locations is able to show on the map. The demonstration was able to perform similar detection process as those journalists did in the late 90s.

The paper also introduces additional findings during the examination of the original datasets. As a news media outlet based in New York, we see evidence that the

New York Times turns out to focus much more on New York City and the United States compared with other countries. With the extraction of locations inside the articles, we were able to see the distribution of articles mentioning different countries differs a lot when comparing the different continents. Our visualization also shows how locations names were changed throughout time, and how the terms people use describing a certain object changes.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION: THE IMPORTANCE OF RESEARCH

In the early stages of archaeology around mid 19th century, the archaeologists were either wealthy people themselves, such as Heinrich Schliemann [9], or mostly supported by upper class, mainly motivated on adding antiques and treasures to their collections. However, in recent times, treasure hunting, especially in burial sites, is now condemned by most countries. In other countries, the usage of metal detection is still regulated by the government. In these cases, treasure hunters are considered as looters, together with grave robbers and art plunders. On one hand, looting events are usually consequences of war, natural disasters or regional in-stabling. During war, lots of arts and antiques have been shipped from their origins by the conquerers; for example the Nazis during the WWII, the British Empire in the beginning of 20th centuries, Napoleon Bonaparte and so many others. On the other hand, looting never stops during peace time in nearly every country. There are art thieves smuggling arts out of Italy, pottery shipping out from Greece, and statues smuggled from India every now and then. In lots of these cases, there are dealers involved and customers encouraging the thieves to keep doing their work. Among the customers, there are famous auction houses, national museums and a full chain of people involved, including museum curators, police officers, and governors. In the past, there have been people exposing such trading, such as Watson in *Sotheby's: the Inside Story [34], Chasing Aphrodite [8]*, and the many authors who have reported on the famous Getty Museum case in 2005. However, these cases take lots of people's efforts and have drawbacks like unreliable resources and risky or even dangerous methods. Considering these, we developed a system that could perform like the old methods or even better, showing the relationships between people and entities

and tracking movements of them, to help journalists, archaeologists or researchers in different areas to better conduct their investigators.

1.1 Previous Work

Peter Watson's book *Sotheby's: the Inside Story*, described his research finding out how Sotheby's was largely involved in illegal smuggling objects out of Italy, Greece and India. The book described true stories that exposed how Sotheby's was involved in these events. In one case, the authors created a bait that could catch the eyes of traders and Sotheby's dealers to smuggle artifacts out of Italy. Though the authors were breaking the laws by doing so, they still followed their plan in order to get to know the whole chain of trading. During these events, they included several innocent people and relatives in the event, raising unwanted risks and making the mission highly possible to be exposed. Another case described in his book involved the participation of a former Sotheby's administrator who had a vast amount of evidence that was retrieved directly from Sotheby's; these evidence required lots of time and special knowledge to examine. The book did have a very great influence on the art market and lots of people involved in these cases were put on trial. However, the case didn't have too much impact on those dealers' careers. Some of them seem to have retired after the scandal, but returned back to the art market after people gradually forgot about the story. During the process, Sotheby's business was never affected, and it always claimed that the smuggling actions were driven by their own profits, which had nothing to do with the company. The reason partly lies on that the documents are stolen from Sotheby's, and partly because the former administrator himself had some possibility of having forged the evidence papers.

A decade later after Peter Watson's original book, the famous Getty Museum was brought to the world's attention by its curator Marion True's case. True was indicted

by the Italian's government for illegally trafficking artifacts from Italy in 2005. The trail forced her to resign from her position at the Getty Museum, and forced the Museum to return antiques to Italy in 2006. Though with the evidence established, and people she dealt with, like Giacomo Medici, already sentenced to prison, she still escaped from further sanctions for the crime. One year later, Italy and Greece dropped their charges against her, after the museum returned several objects to Italy and Greece, claiming that the statute of limitations had expired. During the trial, she described herself as having to "carry the burden" for the Getty's Board of Directors. Ten years later, in 2015, Marion True spoke out to the public, still claiming that she didn't know where those artifacts come from, and once she found out the source, her group would have returned them.

In the book *Chasing Aphrodite*, the authors described the relation between the Getty Museum, especially Marion True, and looted artifacts from Greece and Italy. After the publication book, the authors launched a website, chasingaphrodite.com. The website included some stories that were not able to be published in the books, and updates frequently about things that happen recently. In some of their most recent stories, their resources came from anonymous sources, federal court records, various newspapers, websites and videos online. The stories told are still breathtaking, while the methods used to discover them and the evidence are much different. The digitization of information makes it possible for researchers like Peter Watson nowadays to do their research mostly online. The Internet provides a more up-to-date and larger dataset that a person could ever reach decades ago. On their recent posts regarding to how the Islamic State is getting their money from looted antiques in Syria, the author gets the resources from online news posts, discussions on Twitter, satellite screenshots, and online videos about hearings on Capitol Hill.

The changing of how journalists discover their stories shows us how the develop-

ment of technology will change the research methods in different areas. The Sotheby's and Getty Museum's cases shows how traditionally antiques could be transferred from looter to museums and auction houses and how they laundry the artifacts to make them look legal. Needless to say these kind of stories are important to reveal the dark side of art market. Nevertheless, these kind of investigations take lots of time and effort. The Sotheby's story was revealed by dangerous undercover actions with the coincidence of the evidence from the disgruntled former Sotheby's administrator. The authors of the book Chasing Aphrodite started running a website. They've continue doing what they did when writing the book. The emerging articles on the *Chasing Aphrodite* website also show that with proper usage of technologies on the Internet, people could make similar discoveries as they did before.

Considering such kind of situations, we hope to create a tool that could gradually help better follow similar steps to those carried out manually while saving lots of effort and making the process much easier. Thus would therefore help authors like Peter Watson or archaeologists in better performing these investigation.

1.2 Problem Statement

In the past, the investigation of antiques traveling, people's actions and money transfer relied solely on people. When Peter Watson started to investigate the transportation of paintings from Italy to London, his first approach was buying a painting himself and setting up a trap to see how the network works. After he started the investigation on Sotheby's, a former administrator in the Antiques Department at Sotheby's in London reached out to him, holding evidence that he retrieved from his office. The evidence brought by the former administrator contained original documents from inside Sotheby's ranging over a decade, involving locations from London to Jaipur, and in a very large amount. At that time, the author and the admin-

istrator had to examine them, find the linkage between evidence. Some of those documents required special knowledge to fully understand them.

The knowledge of part of the story drives the motivation, making the investigation have suspicion. Then he or she must know where to look for potential links and real evidence, either by knowing a potential informer, getting supporting files from them, or probably by devoting themselves in that field. The system must also have a dataset that can perform like the suitcases held by the administrator in Peter Watson's book, containing enough trustworthy evidence that could allow the formation of a deduction of the relationships among people. For our study the *New York Times* as our main resource. Ever since its founding in September 18, 1851, *New York Times* has continuously published a newspaper each day. It covers nearly every aspects of peoples' lives, including stories coming from all over the world, and describes both big and small events happened in the world on a daily basis.

News articles never mention unrelated things in one article. Consequently, when two names are mentioned in one articles, we can assume that these two people are largely related. If these two people were at the same place at a certain time, it would be very safe to say that they have met with each other or to say that they know each other. If a person was mentioned in one article, along with some other entities such as type of antiques, or museum names, then the person had something to do with those entities. In situations like Getty Museum's scandal, before 2005, the year when the Italy brought Marion True on court, there were already lots of articles in the *New York Times* suggesting that the objects in Getty Museum were stolen property, and that Marion True was involved in it, since she stated that their artifacts resources were clean; stories like this could even date back to the 1990s. In 1996, the *New York Times* launched its website, and since then, they gradually have digitized earlier editions, making access to older articles simpler and thus making the

building of a tool based on their articles possible. All news articles since 1851 now are accessible through the website in different ways. For articles after 1978, readers are able to access full text, and for articles prior to that, they provide API access to the abstract and top paragraphs. In addition, the *New York Times* API provides access to keywords for the articles and location names about the article. The location names could be where the article was written and sometimes the locations mentioned in the articles. These help make it easier to collect information from articles.

Given the *New York Times* dataset, users could be enabled to perform similar approaches as that Peter Watson follow manually doing his investigation on the Sotheby's case. The user should be able to tracking antiques transportation and moving, and see the person involved or other related entities at different stages. Also, simple search like where the person had been can also be very useful to do pairwise comparison. Apart from that, the relationship among people could be easily revealed; for example how they were connected with each other, and if there was anybody who also showed up at this place with this person. The tool should also be able to show the crossover of two peoples' tracks and those common entities or names that they both are involved with. To demonstrate such kind of information that is highly related to location, a map can be very useful for display. Thus, we decided to focus on an interactive website to track the information and show users the results based on their queries.

Trading looted antiques requires the transportation of objects. For example, some paintings are not allowed to be auctioned in Italy, while they still have a very large market. The undercover story by Peter Watson showed that the auction houses will help the transportation from Italy to London. Before the transportation, the painting was located in Naples, Italy; after the transportation, it will be shown in an auction happening in London. The tool should be able to show such information

about the painting's location again a goal supported by a map. With the item shown at different places at different times, someone related to the item must have transported the item.

Google Map is now the most widely used map in the world. It provides an API for the developer to access location information like latitude and longitude, and a friendly interface as well. Location pins can be easily appended to the map, along with an information box that allows interactive actions. Apart from that, Google map also allows developers to embed their own control tools to interact with the map, which provides more flexibility over the interactive map. Moreover, it is very easy to display new data with the use of Google Map, which allows user add another data layer or feed the data directly to the maps and display. Considering these facts, we devoted to use Google Map as our platform.

All in all, we're aiming at building an interactive web application that can show the relationship between entities and people mentioned in *New York Times* articles based on their locations. The information demonstrated on the map would first be generated from *New York Times* articles, and analyzed to find the closest match of potential entities or relationships. For queries for two or more entities, the map should be able to show those articles that both or all of these entities are related to.

1.2.1 Contributions

The contribution for the research is to build a system that better tracks traces of antiques or famous names using techniques and resources on the Internet. The system should be able to answer questions like the following:

1. How many museums could Giacomo Medici possible have connections with?

2. Was Marion True in contact with Giacomo Medici between 1980 and 2000?

3. How did the popularity of Marion True's topic change with respect to time?

4. Which museum have Apulian vases in the United States?

5. What kinds of African art are in the Getty Museum?

6. What kinds of antiques did Marion True deal with?

To answer these questions, there are several functions that the system should be able to perform. First of all, the system should be able to demonstrate the information on articles from *New York Times*. The articles are written in natural language and need to be formatted and processed before applying a visualization tool. The system should be able to let the user track the transportation of an item or the traveling of people. When looking at a specific location, it is expected to show who has been here, who has connection with this place, or what kinds of items have connection with this place. When searching for a specific topic or specific people, the system should be able to show the locations that the topic or people are related to, and other entities and names that have links to the people or this topic. The system should also be able to show how the topic changes with respect to time - when was the topic brought up to surface and when did it fade. Besides that, the system should be capable of narrowing the search result by changing the time and zooming in the map. When looking at two terms, such as a person's name and an entity, the system should be able to show the locations and topics that both of these terms have connections with, and other information such as timescale should also be changed along with the terms. Moreover, there are many topics and articles related to one single search; the system should also find an optimized method to sorting the results.

The following parts of the paper is organized as follow: Section 2 describes related work in detail. Section 3 describes the assumptions and implementation methods,

including the data analyzing part and the visualization part. Section 4 describes evaluation of the methods. Section 5 is the conclusion of the paper and Section 6 discusses future work.

2. LITERATURE REVIEW

This chapter discusses the previous related work of the thesis. In the first section, we discuss the motivational works for the research about how journalists' investigations were run in the past. Next, we discuss previous work on article crawling, including using APIs and building crawlers, followed by various related work on entity extraction and geolocation extraction. Finally, we discuss some previous work on map-based demonstrations.

2.1 Loot Tracking in Archaeology

This research is motivated by the investigations done by Peter Watson in the 90's about illegal transporting and trading related to Sotheby's. In this section, we discuss how journalists do their work.

2.1.1 Sotheby's: the Inside Story

Sotheby's the Inside Story describes how Peter Watson discovered and revealed the scandal of Sotheby's relationship with tomb raiders and antiques or art smugglers, and the final outcome of his investigation that started in late 1980s. This book became one of the motivations for our research.

The book first describes the trial of James Hodges, a former administrator in the antiques department at Sotheby's, from the very beginning to the end of the trail, and several operations Watson conducted after the trial. The book, together with a television show, was describes true stories, about how auction houses and museums were involved in smuggling, tomb raiding and looting events. In 1985, the author Peter Watson was informed that Sotheby's was selling lots of smuggled antiques by the curator in the Greek and Roman Antiquities department of the British Museum.

After he heard of this, Watson called Felicity Nicholson, the head of the antiquities department at Sotheby's for comments. The call was put through by an administrator who was at his lunch hour, reading Peter Watson's previous book, *The Caravaggio Conspiracy*. This man, James Hodges, knew that what his boss told Peter was not the entire truth. Four years later, the administrator realized that he was dragged deeply in smuggling events. The accounts used for traveling were under his name, and he felt he would be the scapegoat once the case was exposed. Worried about the situation, he decided to find evidence to keep himself safe, and hence started collecting evidence from Sotheby's. In 1989, Hodges turned over to Peter Watson three suitcases full of documents acquired from Sotheby's before Hodges was put on trial. Watson found over 500 pages useful materials, including evidences that clearly showed that Sotheby's had links with notorious tomb raiders and dealers. However, his presentation to the court as witness wasn't powerful enough for the public to realized that. Peter Watson and his colleagues then traveled to Italy, India and Switzerland to further investigate details. Sotheby's continued to claim that the antiques in their auction catalogues were legally imported from Switzerland, but due to a flaw in Switzerland's law they didn't need to prove the origins of the antiques before arrival at Zurich or Geneva. This allowed cases such as that notorious antiquities dealer Giacomo Medici, who worked with Italian local grave diggers to obtain vases or statues on ground, then shuttered them in pieces, and sent those fragments to Switzerland warehouses. There the fragments were glued back together and were sent to London or New York for auctions "legally". In Italy, Watson witnessed the actions of tomb raiders and even had interactions with them. Once after he was on the site, he received a threatening call from Giacomo Medici himself. In India, he were able to talk to one of the local antique dealers about the transportation and smuggling arrangements when he had business with

the Sotheby's. In Zurich, he found the warehouse of Medici, filled with hundreds of antiques, in pieces, half restored or fully recovered. After the investigation, to make their story more trustworthy, he made a bait for the Sotheby's and experienced the steps himself. He brought a painting in Italy, and claimed he wanted the painting sold in London to reach a better price, Sotheby's contacted him saying that they needed several contacts information addressed outside of Italy; one in London, and one in Australia. Several months later, the painting was found in the catalogue of one auction in Sotheby's London.

The result of the investigation included two TV programs and one book. The first TV program was aired in 1995, while the second one was aired in 1997. The book was published in 1997 in both the United States and England. The followup of the book included the arrest of Medici, who was out of jail in 1997, and museums and Sotheby's stopped dealing with unprovenanced antiques, Sotheby's at once stopped their auctions at London, and the publication of the book made many people in the field comment that Sotheby's was not the only one dealing with smuggling events, making the author start to investigating Christie's at a later time.

This book described in detail about how journalists, archaeologists and other researchers link evidences together and look for traces about secret smuggling trades. The existence of evidence brought by James Hodges was one of the most important steps in Peter Watson's investigation. Prior to that, he only had slight suspicion due to the appearance of antiques in Sotheby's catalogue but had no direct evidence. With Hodges' documentation, he was able to find the names, locations, and made further connections with people doing the smuggling.

2.1.2 chasingaphrodite.com

The website is created and managed by the author of Chasing Aphrodite, Jason Felch, aimed at hunting for looted antiquities in the world's museums, as stated on the front page. After the book, he started publishing news on similar topics on the website. He keeps track of the museums' returning their collections to their origins, and reports his new discovers in this area using various documents he has access to, including videos of hears at Capitol hill, news reports around the world, Google maps, street view, satellite pictures and photos sent from all over the world. Some of the most recent articles revealed the relationship between ISIS and antiques, about how they were support in the conflicts by antiques, and customers behind them.

The change of techniques for discovering the links among people and antiquities have changed slightly with the usage of computer and Internet. The website gave us an idea about how the modern way to discovering traces should look like, and provided trustworthy results about antiquities smuggling that could be used in our research.

2.2 Article Crawling

The *New York Times* launched their website in 1996, and made all their articles since 1851 available online, with free access to full-contents for articles later than 1978. For articles earlier than 1978, the abstract and main paragraph are available through an API. These articles cover news happening in the world every day, but are mainly focused in the United States, with some local news from the New York area.

2.2.1 API

The traditional media industry is very cautious about new things, especially given the complicated legal licensing of linked data. A case study of the BBC (British Broadcasting Corporation) [24] shows that despite their willingness to share their metadata and information with their users, and even allow them to download content, they are always reluctant to let people modify the content, and do not provide an useful API. This might be due to the lack of competition and profits for the industry company. There are already several researches using the *New York Times* dataset through its API[23]. There are also some people who have studied into the dataset itself and discovered interesting findings. Zitnik discovered bias about different countries, how different countries were mentioned in different areas, and the dynamics of countries' names usage [35]. They used the article search API to query specific country name and analyzing the query result to see how many articles mentioned a specific country. However, their research didn't consider the United States. Diakopoulos [6] and Pierson [25] studied the comments on the NYT articles using the API, both using the Times Community API. Diakopoulos studied compared the quality of editor picked comments with comments before the selection using the average relevance score. Pierson discovered that women's comments tend to receive more recommendations and focused on different areas compared with men. Pierson's finding was shared with the developer team of the NYT and helped them to make their forum more equal. There are also several studies using the NYT dataset as a benchmark for different datasets, for example, Huang proposed a method to help classify tweets using categories in the *New York Times* [11].

2.2.2 Crawler

In order to retrieve the content of articles from different websites, we need to crawl those websites thoroughly. Though the New York Times provides an article API for developers to access, the query results only contain keywords, headline, first paragraph, URL and some other metadata about the article. Further steps need to be taken after retrieving the URL with the API. For other sites that don't provide an API with a sitemap, we are able to visit all the webpages of the website. The HTML for each page can be easily retrieved, but for our study we only need the content of main articles and some metadata. News articles websites like the *New York Times* last for a very long period, and change their webpage pattern every now and then. Iraklis Varlamis proposed an automatic crawler that combined the information gathered through the process of crawling that could change with changes in the webpages' patterns [31]. Varlamis's crawler analyzed the XPath of web pages and categorized the tags, extracting article contents from them. Other methods including using RSS feed to extract news article pages links and extract article content from webpages [26], which is also efficient for crawling and further processing such as summarization.

2.3 Entity Detection

Entity Detection is one of the main challenges in the research. With the entities having different forms in the articles, we need to find some methods to efficient extract them from articles, and reduce ambiguity at the same time.

2.3.1 Entity Extraction

With open access to data resources on websites, entity extraction became a very popular topic in natural language processing for many years. The intuitive method

for entity extraction is comparison with an existing database or dictionary. Krieger used DBpedia as dictionary to get relationships among semantics in articles [14]. However, this would require the access to a complete word dictionary, which is not available in many situations, especially when people or companies' names or location names need extraction. To extract other entities in documents, people started with the analysis of words and gramma in different languages, and used the features in gramma to better extract the entities [2], the intuitive method would be discovering capitalized keywords in articles, combined with conjunction or segmentation. Rau discussed various kinds of company names and designed an algorithm about them [27]. Later on, researchers tended to improve the accuracy using machine learning technologies to better help with the accuracy, some used Hidden Markov Chains to increase the accuracy of entity extraction [3]. Considering the situation of using abbreviation or references of surnames of people, researchers have applied approximate extraction [15, 33, 4], calculating the distance between each words to perform approximate search for keywords. This leads to proposing new filter algorithms using pruning methods for faster calculation and better accuracy [15].

2.3.2 WordNet, BabelNet

With the building of WordNet as proposed by Fellbaum in Princeton [21, 20] people could be able to reduce ambiguity between words or terms using the WordNet sense set. In the same period, people came up with different databases for similar use, such as DBpedia. However, these database are still very limited compared with the large amount of terms and people's names referenced every day in the world. Thus, people came up with the idea of combining databases like WordNet with large online libraries like WikiPedia. Makris's group proposed a method to annotate text using entities from Wikipedia and using WordNet to determine the dominate sense words

to reduce ambiguity [17]. Cornolti compared among various annotating methods using Wikipedia [5]. Another successful case for this is BabelNet, which combines Wikipedia with WordNet and provides wiki links to each of the search results. For each search query, it will list all possible senses of the query and the Wikipedia link of its dominate word.

2.4 Geolocation Extraction

Finding accurate geolocation names inside news can be hard, because the relevant keywords are hard to identify as geolocations with information retrieval systems. For example, Washington could be a name, a state, a city, a university or a road. In natural languages, locations are even more vague, such as South France. Some people use geolocation positions and minimum bounding boxes to improve the accuracy of semantic matching [16]. There are also some research that focuses on summarizing geolocation of an article using *Geographical Gazeeter* to find location tags [1]. Wikipedia is a very useful tool when it comes to Geotagging [28], it helps in disambiguating cities having several names in history.

Some combined semantic topic of articles on Wikipedia and related geolocation to built a location based article recommendation [29]. When evaluated with the *New York Times* corpus for fifteen famous locations, the algorithm using Wikipedia outperformance BOW(Bag of Words), LDA(Latent Dirichlet Allocation), ESA(Explicit Semantic Analysis) and PESA(Probabilistic Explicit Semantic Analysis).

2.5 Demonstration

There are many research projects that demonstrate their results on maps. Mauder used isomap to see the relationships among trades and people [18]. Some projects are implemented using Google Maps [32, 7, 12], which provides accurate location and automatic mapping given latitude and longitude using the Google Map API. Overlay

maps or features can be easily applied on the map, and the map itself can be easily embedded with data or other data [30].

A new visualization tool called D3 is also applied by many people to demonstrate information [13], and many tools could be added with JavaScript to improve better performance or provide stylish views [10]. D3 could also be used to build maps, and combined with Google Map to provide better interaction experiences.

While displaying a large amount of data on a map, the crowding situation happens every now and then. We collect all articles from the New York Times, so the amount of data is large, since some places have lots of keywords. In this case, the pins on map could squeeze together. One possible solution for this is to summarize them, and show the summarized result, instead of all the pins [22].

3. IMPLEMENTATION

The overall flowchart for our the system is shown in Figure 3.1. It contains four main parts: data retrieval; analysis of articles for locations and entities; building the database; and generating the demonstration. Both will be discussed in turn next. The data processing parts are illustrated in the system flowchart.

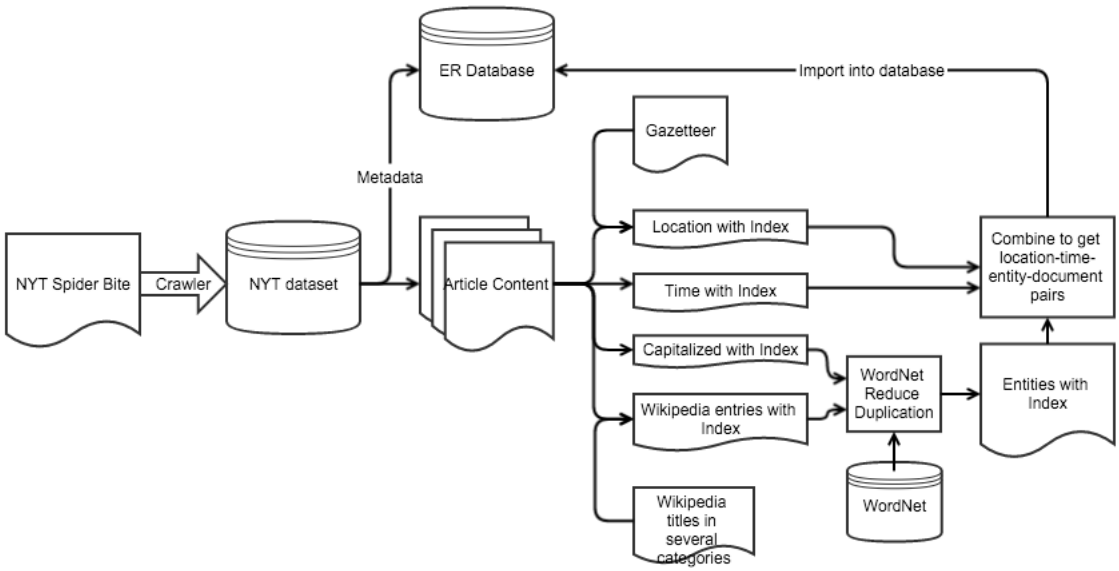


Figure 3.1: Data Process Flowchart

3.1 Data Retrieval

3.1.1 New York Times

The articles in *New York Times* can date back to 1851. Since the NYT does not provide free access to full contents for articles before 1978, we use a crawler to retrieve the article contents and this ensures we could get the information for all arti-

cles throughout all the time. We built a Python program that uses the GET request provided by the NYT to retrieve the metadata of articles, including the URL, first paragraph, abstract, title, author, location where the article is written, and the time when the article is published. The request URL is in the following format:
`http://api.nytimes.com/svc/search/v2/articlesearch.json?&sort=oldest&document_type=article&fl=web_url%2Clead_paragraph%2Cabstract%2Cheadline%2Ckeywords%2Cpub_date%2Cbyline%2C_id&begin_date=19000101&api-key=sample-key`. The request retrieves metadata of articles since 01/01/1990. The result with the earliest publication date was returned first with its url, lead paragraph, abstract, headline, keywords, publication and id. By adding `&page=1`, we were able to retrieve the second page of the results. However, the NYT limits the number of requests sent each day for each authorized key. For the article search API, which was used to retrieve article metadata, each API key is only allowed to send 10,000 calls per day. Compared to over 15 millions articles on the website, this number is too small to retrieve sufficient articles for our next steps. Further, with so many requests needed to be sent to retrieve metadata, the program runs relatively slowly when compared with a crawler. Considering this, and that articles before 1978 can't be accessed with API, we retrieve the dataset using two different methods, one for articles before 1978, and another for articles after that. For articles prior to 1978, we used multiple accounts running simultaneously using the API to retrieve the metadata, and stored the results in a .json file.

For articles published after 1978, we built a crawler to retrieve the full content of the articles. For the first step, we need to have the URL address for each articles on the website. Most websites have a sitemap with its creation that lists the URLs of all webpages contained in the website; nytimes.com is not an exception. Its sitemap is modified as a spider bite, with article URLs categorized based on the time of

publication and stored in different years. For each year, the articles are divided by time order into several parts for each month. Thus, we build a crawler to access all links in the spider bite to access all articles on the *New York Times*. Considering that some articles have multiple pages, we use the found links of articles to generate instead a link to their printed version pages, which have less advertisements, or decorations, and put all pages of the article together. After this, we were able to get a full list of URLs for articles published by the NYT after 1978.

With the URL lists, we are able to retrieve the full contents of articles on NYT.com. The retrieval program was initially built using Python, and it was able to retrieve a full list of all the articles based on the url lists. However, when it come to retrieving the full context of the articles, the Python program took too much time. So we took a further step, and transferred the Python program to a C++ application that ran faster on accessing full context page. We also ran several executions of the application for different years simultaneously, so that the articles were able to be fully retrieved within a couple days.

3.1.2 Chasing Aphrodite Website

The website has very useful articles. However, the number of articles is not too large. Similarly as for the *New York Times* website, we used the sitemap to retrieve all the articles.

The program used in the previous section can be directly applied to this website. Since the dataset is so small compared to nytimes.com, the program didn't take too long. Another tricky part about the website is that all the articles are in blog style, and don't have a page separator. On the other hand, all the articles only have one page, which make it even more easier to retrieve the full contents from this website.

3.1.3 Extract File Content

After the previous steps, the articles on nytimes.com and chasingaphrodite.com were stored locally in HTML files. Those files contained the full content and metadata for the articles. For each article, the file content part only takes around 10% of space of the file, the rest of the html files are supporting codes, metadata, or advertisements. Considering the file size and benefits of further processing, we choose to analyze the file to retrieve the metadata information into database, and extract main content of the articles. We extracted article title, url, publish time to the database, assigned each file a unique id based on their publish time, and save the main article content as text file using their id. In that way, the total size of the data shrink from 63GB to 5GB.

3.1.4 Sotheby's: Inside Story

The Sotheby's book is the motivation of our research, and we hoped we could use our tool to follow the same process so that we could achieve similar result. Thus, we wanted to include the book into our dataset. Using OCR techniques, we transferred the book into digital copies, to better prepare for the following process.

3.2 Entity Extraction

3.2.1 Name Extraction

To extract the entities from the dataset, we need to considering what kind of entities we hope to get from the files. The first kind of entity to be extracted is the names mentioned in the files. People's name follow certain rules. The rule is usually 'FirstName, LastName', 'LastName, FirstName', 'FirstName MiddleNames LastName', in these cases, they're a consecutive string of capitalized words. In other cases, especially in some European names, there could be lower case letters in the

middle of names, such as Leonardo da Vinci. Considering all various kind of cases, we find the similarity pattern is that the both contain multiple capitalized words, with other words in the middle of them. Thus, we use regular expression to represent the names: $re = [A-Z][a-z] + (?= \ s[A-Z])(?: \ s[A-Z][a-z]+)*$

We built a program that extracted the index of people's names in each article, and stored the name with their location in each articles into name.json file.

3.2.2 Wikipedia

Another kind of entities would be the entities contained in the file, they could be either area or subject names such as "Nautical Archaeology", terms that refer to a certain time period such as "Renaissance", or certain kind of antiquities, such as "Mummy". These kind of words sometimes cannot be picked by the previous step for the following reasons:

- They are not capitalized words, such as "art theft".
- They are single words, such as "Mummy" or "Renaissance" mentioned before.
- They contain numbers or special characters, for example "20 Exchange Place".

Considering these, we decided to use Wikipedia to help us better select the entities from the articles. We used Wikipedia's API to retrieve all terms mentioned in Wikipedia under the art, history or archaeology categories since we were mainly focused on helping research archaeology areas. This approach can support research in other areas. The crawler can be easily extended to other areas while providing similar results. There were around 14,000 entries from Wikipedia. We built a program that extracted the index of entities in the Wikipedia dictionary we just created, and stored the entity name and its position in the article into a .json file.

3.2.3 BabelNet, WordNet

Another situation we needed to consider are different ways to describe the same things. Take President George Bush as an example. People could refer him using President Bush, George W. Bush, or simply George Bush. Considering that his father George H.W. Bush was also once the president, people sometimes refer him as Bush the younger. To detect all of the various different descriptions as Bush, we applied WordNet and BabelNet on the Keywords database to eliminate the ambiguity of the words. In BableNet, two words describing the same entity will be linked to entity objects with same ID. For each entry in the Keywords database, we searched for its ID in the WordNet and BableNet, and then changed the name of the current entry to the item's name, which was usually the most popular words describing the entry with the minimum ambiguity. When there were multiple items about an entry, we picked the one with the lowest id, since the sense id in WordNet is assigned based on its popularity, with the most common sense numbered as 1. As in the previous example, both President George Bush, and President Bush will be treated as George W. Bush. After this process, we eliminated the duplications in the database.

3.3 Location Extraction

3.3.1 Usage of Gazetteer

To extract the location names in the articles, we first used GeoName gazetteer, which contained the information about all cities in the world with population larger than 1,000 people.The information included name, latitude, longitude, population, state, country, and all other forms of the name.

3.3.2 Reduce Ambiguity

There are several cities in the world that belong to different countries or different states, but share the same name. Considering this situation, we first preprocessed the location gazetteer to generate unique location names, for duplicated location names, we keep the location with larger population. For example, for London in England and London in Canada, we kept London in England in the gazetteer and deleted all other Londons. For faster access, the location names were stored in a two level hash table in .json format.

3.4 Database Build

We analyzed the positions of locations, names, entities, and some time stamps in each article, and combined them. For each location and time, we treated any name or entities appearing right after the location and time as meant? that the person had been at this location or the entity was mentioned at this location at that time. Once a new location is found, we treat everything mentioned after this location's position as happening at this later position. The same rule applied to time. In that case, we kept a currentLocation variable and a currentTime variable, with the increase of index, once a new location or new time is found, we update currentLocation or currentTime. After the process, we are able to get several [time, location, entity/name] pairs, and this information is stored in the database.

The database consists of three main tables: Articles, which stores the information about each article, including ID, URL, written time, title, and its source, in which ID is the primary key; Locations, which is the location dictionary, containing name as the primary key, latitude, longitude, and whether it is a city or a country; Keywords, that stores the entities or names with its location, year, count of appearances and the ID of article mentioned it. For the same entity or name mentioned in different

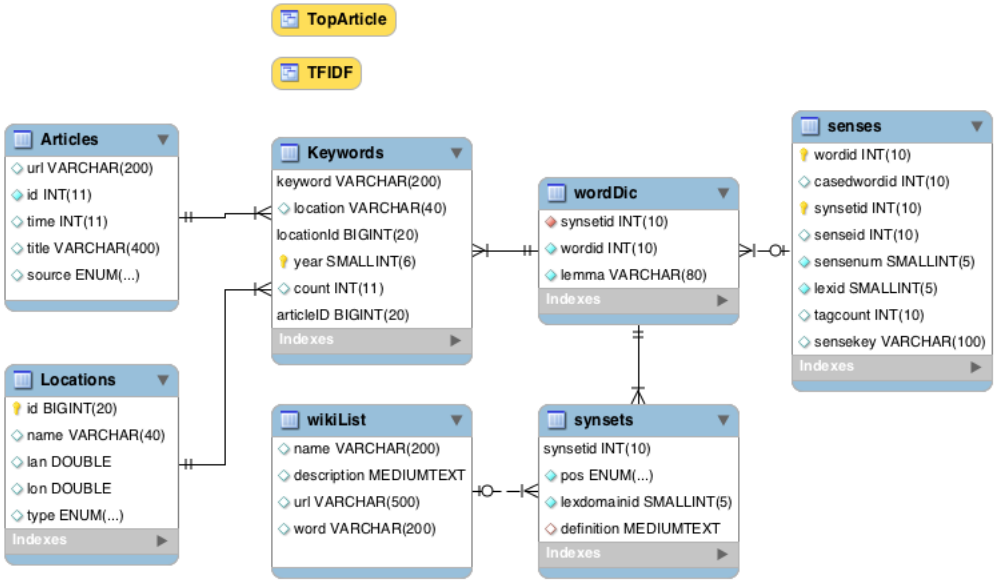
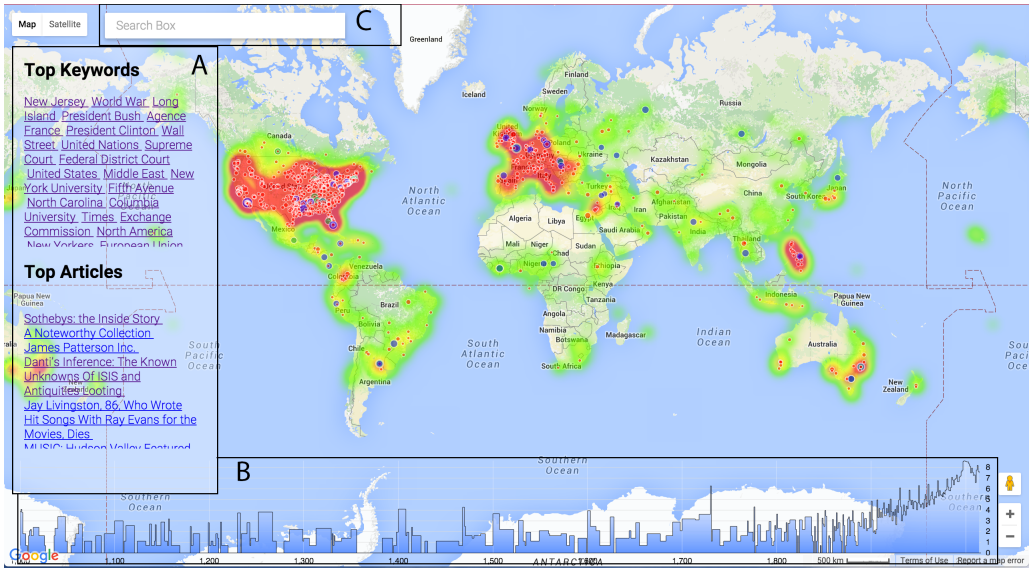


Figure 3.2: ER Diagram Of The Database

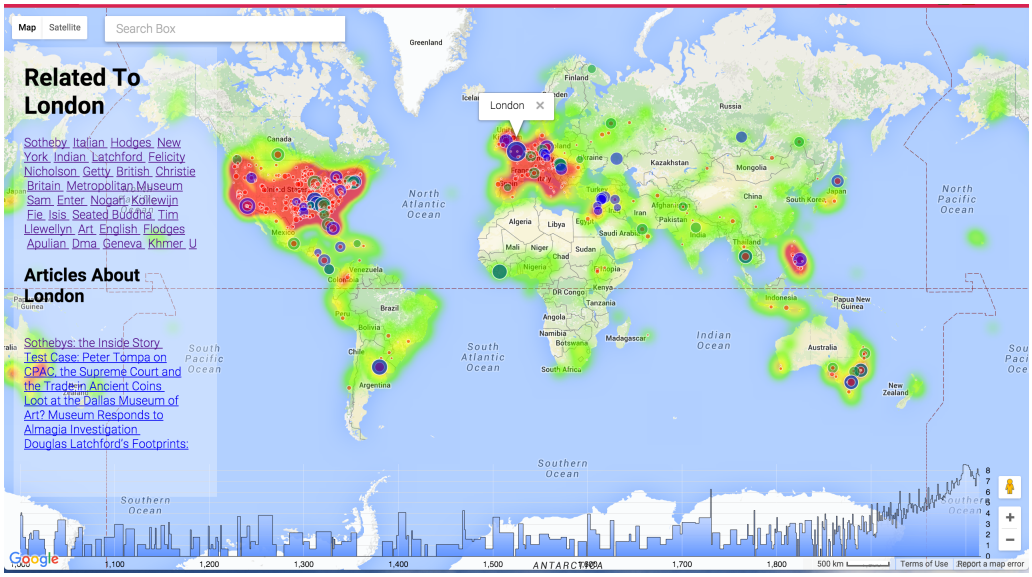
articles, location, or time, they will be stored in the database as different entries. The relationship among the tables are shown in Figure 3.2.

3.5 Demonstration Website

The website is based on a map and provides several interaction methods with it. We first created the website with underlying Google Map. This allows users to zoom in and out, drag smoothly, and provide useful API to get the status of map and set control functions for the map. To show the transportation and relations based on the locations, we display the entities and names on their locations as markers, and show the keywords, locations and articles related to it, and its appearances throughout the time. To demonstrate the most mentioned keywords or terms about a specific keyword, we use the left sidebar to display top keywords and top articles, as shown in Section A of Figure 3.3a. The year slider in Section B at the bottom shows the change of numbers related to the current search. The search box at the top in



(a) Demonstration After Initialization



(b) System Status After Search For London

Figure 3.3: Demonstrations

Section C can be used to search for specific keywords or entities. Considering there are millions of articles, this makes the number of entries in the Keywords table very large. For performance reasons, we took a 1/100 sample from the articles randomly.

We found that the overall distribution over time and space was relatively the same while the processing and interaction processes were much faster. Figure 3.3a shows the interface of the website after initialization.

3.5.1 Markers

The webpage first loads all the keywords from the database, creating a marker for each of the entries in the Keywords table, the marker containing information such as location name and how many keywords are mentioned at this location. With the latitude and longitude of the location, the marker points to a specific location on the map, while the color shows the source of the location, and the size of marker shows relatively how popular the location is.

Upon clicking the markers or keywords in the side column, the map sends a request to the temporary result table, and updates the temporary table based on the query. Then it refreshes all the markers, keyword column, article column and year slider using the updated table, as shown in Figure 3.3b. The parsing algorithm applied to search box also makes it available to show the relationships between two entities.

3.5.2 Top Keywords, Top Articles

The side column is used to show top keywords and top articles related to the current search query. In the initial state, it shows the overall most popular keywords and most popular resource. For each search query, there could be too many keywords or related articles to display in the available space. So it became necessary to show users the most useful keywords or articles. Consequently, we use TF-IDF (Term Frequency - Inverse Document Frequency) to evaluate the importance of certain keywords or articles. The TF-IDF value is the product of TF value and IDF value. Term Frequency describe the number of occurrence of term t in document d

(Equation 3.1), or in the case when the frequency numbers are too large, use the log value of the frequency (Equation 3.2).

$$tf(t,d) = f_{t,d} \tag{3.1}$$

$$tf(t,d) = 1 + log(f_{t,d}) \tag{3.2}$$

The Inverse Document Frequency describes the importance of the term in the dataset Equation 3.3. N is the total number of documents in the dataset. $|\{d \in D : t \in d\}|$ describes the number of documents that contain the term t.

$$idf(t,D) = log(\frac{N}{|\{d \in D : t \in d\}|}) \tag{3.3}$$

To make the initializing process faster, we calculated the TF-IDF values for all keywords and stored the results in a new table, so that we don't need to perform the calculation every time the page is opened. For other search queries, once a search is performed, the results are stored in a temporary table in the database, and each follow up search based on the query is performed using this table. The order of keywords is also determined by the TF-IDF score of each entry.

3.5.3 Time Scaling

The time scale shows how popular the current query is throughout the time. Once a query is performed, the time scale is updated to show the change of numbers based on the current change. The time scale bar can also be used to narrow the query result. If the user drags the time scale around or zooms in and out, the markers on the map will change with it.

3.5.4 Search Box

The search box should be able to perform a search query. For a basic search query, the search box sends a request to the database, checks if the query is a location, and creates a temporary result table using all entries in the Keyword table related to current query. For queries that are trying to get a result based on multiple keywords, the search box first analyze the input, generates a SQL search query based on the input, and creates a temporary result table.

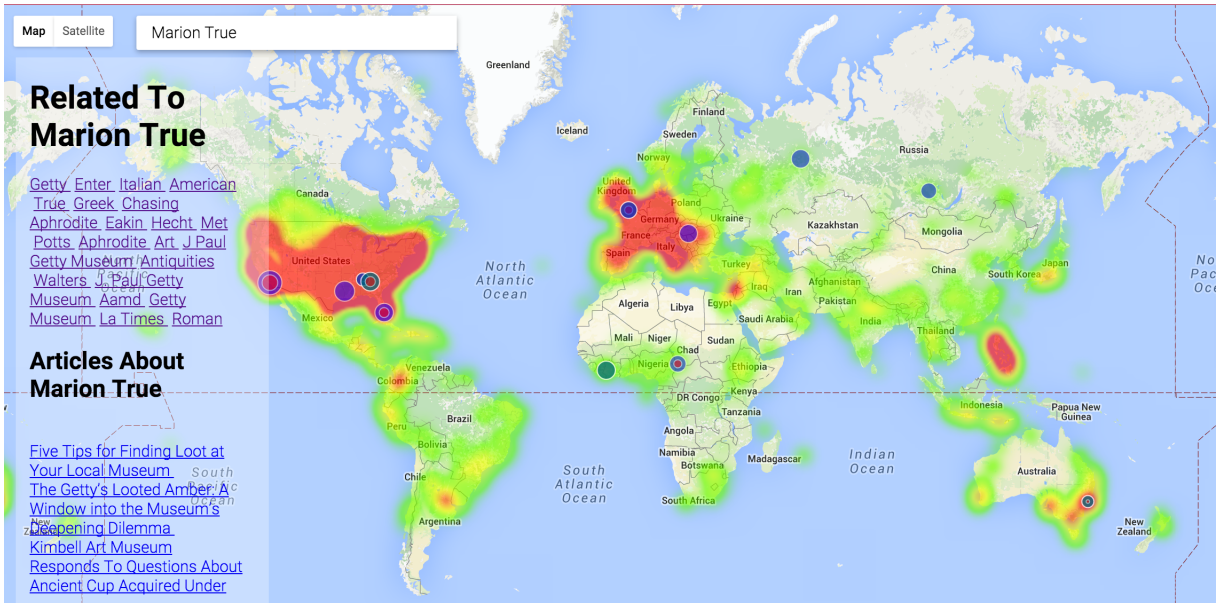


Figure 3.4: System Status After Search Of Marion True

Figure 3.4 shows the view after search for Marion True. We can see that she is mostly related to Los Angeles, London, Italy, Australia and some parts of Unites. This matches with what we know of her path, places and people she deal with.

3.5.5 Heat Map

The map also has an underlaying heat map that shows the overall result which will not change based on the search query. This allows us do a comparison with the current search result. In the most mentioned areas and locations, the heat map tends to be more red and in those green areas, the number of entities are fewer.

4. EVALUATION

4.1 Entity Extraction Accuracy

To calculate the accuracy of Entity Extraction, we take several documents as sampler, manually identify the entities and locations inside the documents, and calculate how many keywords and locations also appeared in the database. Equation 4.1 was used to evaluate how successful the system is in extracting existing terms. Equation 4.2 was used to calculate the redundancy of the database.

$$ExtractionAccuracy = \frac{\#\ of\ pairs\ appeared\ in\ the\ database}{Total\ \#\ of\ pairs} \tag{4.1}$$

$$Redundacy = \frac{Total\ \#\ of\ pairs\ in\ database}{Total\ \#\ of\ pairs} \tag{4.2}$$

From the result we can clearly see the differences among different resources. First

Table 4.1: Accuracy Of Extraction

	Sotheby's	chasingaphrodite.com	New York Times
Extraction Accuracy	0.1875	0.5674	0.6132
Redundancy	1.96	1.32	1.24

of all, the content of the book included lots of discussions, conversations, and details of trading and transportation. This evidence is necessary in the investigation process, but they generated a very large redundancy in the processing of digital files. This is especially seen during conversations when two or more people were having a discussion, the book mentions their names every time they speak. However, it is

their words that have more information other than their own names. Similar things happen in the book many times, mainly because the book was written in a journal form, including every detail of the investigation process, while makes it not so formal compared with news articles or even website posts. What's more, since the *Sotheby's: the Inside Story's* original dataset was generated using OCR technology, there was a very large inaccuracy generated during the process.

Since we aimed at automating the research process done by Peter Watson, we compared our database with his book, to see whether our tools were able to locate the evidence he discovered and processed. We conducted several searches on our database, and excluded the book's part of the database, so that we could see whether that information could be found in other resources. However, the result was not very good. For the Sotheby's company, it can be clearly seen from the map that their traces can be found all over the world. Also, for famous names or those big names in the books like Giacomo Medici or Marion True, their names are scattered around the world, and appear more often in those areas mentioned in the books. For example, Giacomo Medici is related to Italy, London, and India, and Marion True is highly related to Los Angeles. We also can see top keywords about Marion True including Getty Museum, Aphrodite, other museum names and Giacomo Medici. However, smaller names like Felicity Nicholson, a department head of Sotheby's London, can not even be found in New York Times, let alone the alias of Peter Watson or the colleagues he works with. In fact, their names were barely mentioned in *New York Times* website.

4.2 Analysis Over Dataset

Since the tool analyzed all articles in *New York Times*, it is able to show some interesting facts about the *New York Times* dataset itself. In this section, we discuss

what we found about the dataset.

4.2.1 Area Coverage

Whether in the demonstration of complete dataset or a sampler of the dataset, it's clear that the coverage of markers around the world are scattered differently, as we can see from Figure 3.3a. Zitnik also discussed about this in their research about the *New York Times* [35]. It is understandable that the number of markers in the United States is much larger than the rest of the word, since the *New York Times* company itself is located in New York. Also, the heating in Europe shows that people in the United States tend to focus more on European countries than Africa or Asia, which is also true considering the deep connections between Europe and North American. Table 4.2 shows the number of articles mentioned different areas in the world, and Table 4.3 and Table 4.4 shows show the top ten cities and countries mentioned in the New York Times dataset respectively.

Table 4.2: Article Location Distribution

Africa	Asia	Europe	North America	South America	Middle East
19,301	16,453	100,572	1,193,364	5,791	2,520

We can see from the Table 4.2 the lack of coverage about Africa and Asia, this could be caused by two reasons: first, the error generated during the analyzing process; second, the lack of coverage existing in the original dataset. To see which one has more effect we need to see the coverage of the original *New York Times* dataset. To check the coverage of the total coverage of NYT articles, we used the

Table 4.3: Most Mentioned Cities

City	Count
New York City	326544
Los Angeles	47771
San Francisco	30924
New Orleans	12949
San Diego	11834
Staten Island	10881
Las Vegas	10795
Kansas City	7029
New Haven	6571
Notre Dame	6064

Table 4.4: Most Mentioned Countries

Country	Count
United States	1091012
North Korea	19499
Hong Kong	18619
Mexico	17063
South Korea	15733
South Africa	13325
Saudi Arabia	10294
Colombia	9196
New Zealand	6730
Dominican Republic	4741

Table 4.5: Number Of Articles Distribution In Original NYT Dataset

Region	Africa	Asia	Europe	United States
# of Articles	271,750	172,634	691,517	3,302,694

NYT API to look for articles mentioning or talking about different regions. The results for different regions are shown in Table 4.5. From the table we can clearly see the distribution of articles in different regions over the world. This matches with the distribution discovered in the analyzed result and the distribution of the map. Based on the result, we can see that the bias existed in the original dataset and was not generated during the processing procedure.

4.2.2 Location Name Change

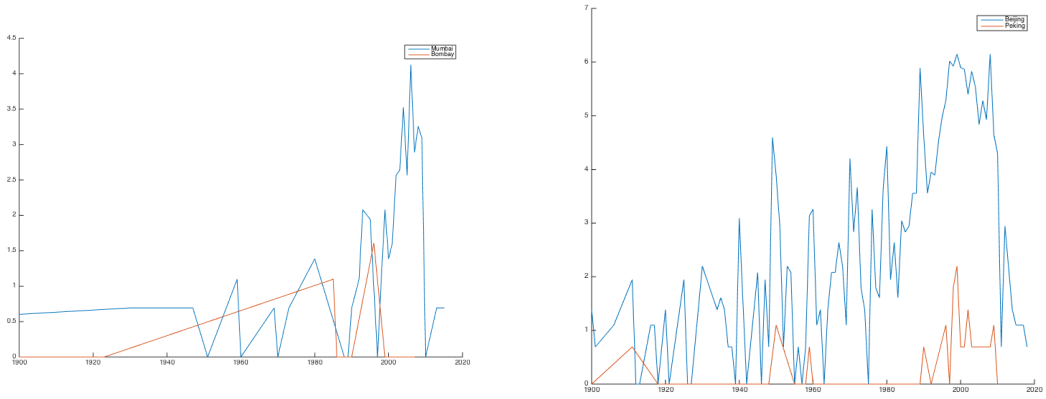
Several different cities have changed their names during some point in history, because of political or other reasons. Examples are the cities in Australia that changed German or German-sounding names during Word War I, Bearbrass was

changed to Melbourne; Beijing once had names of Peking, Khanbaliq, Yanjing due to the change of spelling methods or regime change; some cities' names will change to its native language after the independence of its country, such as Bombay to Mumbai; some cities changed their Latin names to modern languages, for example Lyon was once called Lugdunum. Table 4.6 shows the top 10 cities that have the largest number of alias. With the change of a city's name, media will start to refer

Table 4.6: Top 10 Cities That Changed Names

City	# of Names	Former Names
New York City	5	New York, Big Apple, Neuva York, Nova York
New Orleans	5	Big Easy, The Big Easy, Orleans Parish, Nueva Orleans
Sao Paulo	4	Sao Paolo, San Paolo, San Paulo
San Francisco	3	San Fransisco, Sao Francisco
St. Louis	3	Saint Louis, San Louis, San Luis
Ciudad Juarez	3	Paso del Norte, El Paso del Norte
Ciudad Guayana	3	Guayana City, San Tome
San Ildefonso	3	La Granja, La Granja de San Ildefonso
San Lorenzo de El Escorial	3	El Escorial, San Lorenzo del Escorial
Stare Mesto	3	Old Town, Ciudad Vieja
De Haan	3	Le Coq, Den Haan

to that city using its new name. This kind of process can be seen from the change of the number of articles mentioning different aliases of a city's name. In Figure 4.1a and Figure 4.1b, we show the base 10 logarithm of number of entities happened at Mumbai and Beijing. Mumbai changed from its old name, Bombay, in 1996. From



(a) Mumbai

(b) Beijing

Figure 4.1: Name Changes

Figure 4.1a we can clearly see the rise of Mumbai right before 2000, and the number of entities about Bombay drop to zero before 2000; this follows the change of its name in 1996.

Beijing is another case. In Figure 4.1b, we can see the number of entities talking about Beijing and Peking, a former spelling method of Beijing. The Chinese government introduced the Pinyin system in 1958, and in 1979, Chinese government began to use Beijing in all of its foreign language publication. However, some media didn't change until 1990s, and BBC still referred the city as Peking in its 1989 report [19]. We can clearly see the peak of both Peking and Beijing in 1989, and the rise of Beijing right before 1960 and 1979.

4.2.3 Other Name Changes

There're several ways to refer to certain entity, and there're different methods to call a person. For example, President George Bush can be referred as George W. Bush, President Bush, or George Bush Jr. WordNet could identify different

references to certain word, and link them to its most used name. In Table 4.7, we showed the changes of words that were mentioned using the largest number of nick names. Figure 4.2 shows the number of articles that mentioned George Bush. We

Table 4.7: Various Name Use

Sense ID	Sense Name	#	Nick Names or Other References
108158374	Drug Enforcement Administration	3	Dea, Drug Enforcement Administration, Drug Enforcement Agency
108190414	European Union	3	European Community, European Economic Community,European Union
109147066	John F. Kennedy International Airport	3	Kennedy, Kennedy International,Kennedy International Airport
110895055	George H. W. Bush	3	George Bush, George Herbert Walker Bush, President Bush
111380017	Vincent van Gogh	3	Van Gogh, Vincent Van Gogh, Gogh
100521115	Mardi Gras	2	Fat Tuesday, Mardi Gras
100949739	Substance abuse	2	Substance Abuse, Drug Abuse
101314011	World War I	2	Great War, First World War
101314587	World War II	2	World War II, Second World War
102704730	New York Stock Exchange	2	American Stock Exchange, Amex

can see that the first peak started from 1988, and slowly goes down from 1992. This matches with the period when George H. W. Bush served as the president of United States. A second high peak started from 2000, and reaches its highest in 2004, and slowly decreases until 2009, when George W. Bush left office as president. The two highest peaks matche with his two campaign periods, and the third one matches the

invasion of Iraq led by him in 2003.

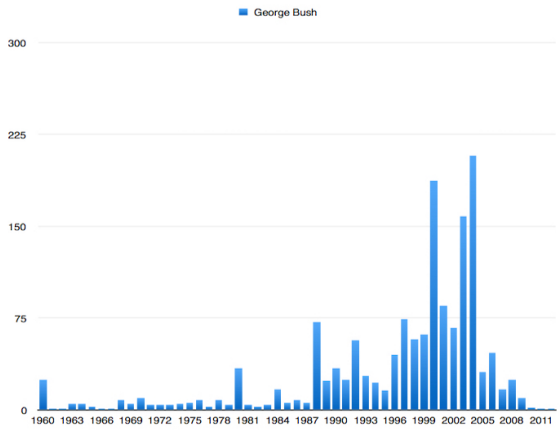


Figure 4.2: Mentions Of George Bush

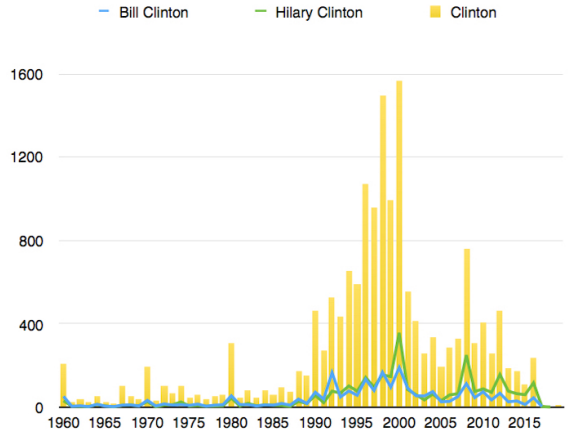


Figure 4.3: Mentions Of Clinton

Figure 4.3 also shows the change of Clinton. From the chart we can see the rise of Clintons in early 90s, when Bill Clinton campaigned and became the president. We especially see a peak before 2000, because of the scandal in 1998. Also in 1998, the number of articles mentioning Hillary Clinton outnumbered those about Bill Clinton because of the scandal. After that, the Clintons silent for a while until Hillary Clinton became a Senator and turn the Secretary of State in 2009. During this period she always outnumbered her husband since his retirement from the president.

From the analysis on George Bush and the Clintons, we can see that time scale can be one parameter to be used on disambiguation. For the same phrase in different times, it could represent different entity and as a consequence, we could use the time to filter entity in future work.

5. CONCLUSION

The paper introduced a tool that help researchers track entities and people in the world, and see the relationships between them. The tool used *New York Times* as its resource to get the location and entities name in each articles, and extract the relationships from articles, applied WordNet and Wikipedia to reduce ambiguity and link entities to Wikipedia, used Google Maps and D3 to generate interactive map demonstrating keywords, their relationships, and articles related to them.

The paper also discussed the distribution of the data resource, the *New York Times*, and showed the clear bias among different locations, countries and regions. We found that the NYT dataset itself has very large bias. It is mainly focused on the United States and the New York region and focuses more on Europe than Asia or Africa. We also analyzed different resources and see the differences among them in terms of word use and reference of names. News articles like *New York Times* clearly have less conversations and are more formatted, while books containing more original data and interesting stories and thus have more duplications in names due to conversations. We also compare the occurrence of several words in our database with some of the big events happened that in the world, and found out that they followed a similar pattern, rising and decreasing at similar moments. This shows that our system is able to successfully discover the entities in *New York Times*.

6. FUTURE WORK

There are several aspects about the tool that can be improved:

- The accuracy can be further increased. The algorithm for analyzing related keywords and locations is simple to use. However, it only consider the relative location in the articles between words, without considering the structure of the sentences or paragraphs. This could be more accurate in some aspects, but also introduces some errors since the situation are not always correct. Furthermore, the parsing of time in each article is only looking for numbers that make sense as a year marker. However, this can't be always true. There are also several ways to locate a time in articles, such as looking for a preposition before the number. These techniques can be applied to further improve the accuracy of the identification of the article. What's more, some Machine Learning techniques can also be applied to do topic recognition to further increase the accuracy of recognition part.

- There are words that are not capitalized but still form a specific term or entity that can't be recognized by the processing program. Considering this situation, we already introduced Wikipedia to help recognize phrases in the Art, History and Archaeology areas. Further steps include using more categories in Wikipedia to further mark entities inside articles without capitalized words.

- Not all capitalized words form an entity, some are capitalized only because they are the first word of a sentence. Some of the keywords are formed by an ordinary starting word and followed by a specific term. In Section 3, we already manually edited some of these words, such as In, On, As, About. However,

there are still some terms contain words capitalized in the beginning of articles.

- The major source was the *New York Times*, which is located in United States, and has to have some bias. Though NYT corporate does have a separate Chinese website, in which some articles are translated from English articles, some are originally written in Chinese. Most of the articles have an English version that can be accessed directly from the NYT English website, however, some of them only provide Chinese version, which makes it hard to analyze. A possible way to eliminate the bias of NYT in the Asia area is to include the Chinese version into consideration, which require a another whole system to process. Another possible way to reduce the bias of NYT is combining it with other data sources, such as *The Times* in London, the BBC in London, and the *Times of India* in Delhi.

REFERENCES

[1] Mirna Adriani and Monica Lestari Paramita. Identifying location in Indonesian documents for geographic information retrieval. In *Proceedings of the 4th ACM workshop on Geographical information retrieval*, pages 19–24. ACM, 2007.

[2] Daniel M Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. Nymble: A high-performance learning name-finder. In *Proceedings of the fifth conference on Applied natural language processing*, pages 194–201. Association for Computational Linguistics, 1997.

[3] Daniel M Bikel, Richard Schwartz, and Ralph M Weischedel. An algorithm that learns what's in a name. *Machine learning*, 34(1-3):211–231, 1999.

[4] Kaushik Chakrabarti, Surajit Chaudhuri, Venkatesh Ganti, and Dong Xin. An efficient filter for approximate membership checking. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 805–818. ACM, 2008.

[5] Marco Cornolti, Paolo Ferragina, and Massimiliano Ciaramita. A framework for benchmarking entity-annotation systems. In *Proceedings of the 22nd international conference on World Wide Web*, pages 249–260. International World Wide Web Conferences Steering Committee, 2013.

[6] Nicholas A Diakopoulos. The editor's eye: Curation and comment relevance on the New York Times. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1153–1157. ACM, 2015.

[7] Susan Dunnavant. Create interactive web illustrations with Google Maps. In *Proceedings of the 38th Annual ACM SIGUCCS Fall Conference: Navigation and Discovery*, SIGUCCS '10, pages 267–268, New York, NY, USA, 2010. ACM.

[8] Jason Felch and Ralph Frammolino. *Chasing Aphrodite: the hunt for looted antiquities at the world's richest museum*. Houghton Mifflin Harcourt, 2011.

[9] Michael L Galaty and Charles Watkinson. *Archaeology under dictatorship*. Kluwer Academic/Plenum Publishers, 2004.

[10] Jonathan Harper and Maneesh Agrawala. Deconstructing and restyling D3 visualizations. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST '14, pages 253–262, New York, NY, USA, 2014. ACM.

[11] Yin-Fu Huang and Chen-Ting Huang. Mining domain information from social contents based on news categories. In *Proceedings of the 19th International Database Engineering and Applications Symposium*, IDEAS '15, pages 186–191, New York, NY, USA, 2014. ACM.

[12] Colin J. Ihrig. Finding yourself using geolocation and the Google Maps API. *XRDS*, 19(1):72–74, September 2012.

[13] Abhijit Jain. Data visualization with the D3.JS Javascript library. *Journal of Computing Sciences in Colleges*, 30(2):139–141, December 2014.

[14] Katrin Krieger, Jens Schneider, Christian Nywelt, and Dietmar Rösner. Creating semantic fingerprints for web documents. In *Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics*, WIMS '15, pages 11:1–11:6, New York, NY, USA, 2015. ACM.

[15] Guoliang Li, Dong Deng, and Jianhua Feng. Faerie: efficient filtering algorithms for approximate dictionary-based entity extraction. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 529–540. ACM, 2011.

[16] Ivre Marjorie Machado, Rafael Odon de Alencar, Roberto de Oliveira Campos Junior, and Clodoveu Augusto Davis Jr. An ontological gazetter for geographic information retrieval. In *GeoInfo*, pages 21–32, 2010.

[17] Christos Makris, Yannis Plegas, and Evangelos Theodoridis. Improved text annotation with Wikipedia entities. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, pages 288–295. ACM, 2013.

[18] Markus Mauder, Eirini Ntoutsi, Peer Kröger, and Gisela Grupe. Data mining for isotopic mapping of bioarchaeological finds in a central European Alpine passage. In *Proceedings of the 27th International Conference on Scientific and Statistical Database Management*, SSDBM '15, pages 34:1–34:6, New York, NY, USA, 2015. ACM.

[19] Michael. From Peking to Beijing: A long and bumpy trip. `https://www.lostlaowai.com/blog/china-stuff/from-peking-to-beijing-a-long-and-bumpy-trip/`, 08 2011.

[20] George A Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[21] George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. Introduction to WordNet: An online lexical database. *International journal of lexicography*, 3(4):235–244, 1990.

[22] Till Nagel, Erik Duval, and Andrew Vande Moere. Interactive exploration of geospatial network visualization. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, pages 557–572, New York, NY, USA, 2012. ACM.

[23] New York Times. Article search API v2. `http://developer.nytimes.com/docs/read/article_search_api_v2`, 08 2015.

[24] Tassilo Pellegrini. Semantic metadata in the news production process: Achievements and challenges. In *Proceeding of the 16th International Academic MindTrek Conference*, MindTrek '12, pages 125–133, New York, NY, USA, 2012. ACM.

[25] Emma Pierson. Outnumbered but well-spoken: Female commenters in the New York Times. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '15, pages 1201–1213, New York, NY, USA, 2015. ACM.

[26] Arockia Anand Raj and T. Mala. Cloudpress 2.0: A next generation news retrieval system on the cloud with a built-in summarizer. In *Proceedings of the International Conference on Advances in Computing, Communications and Informatics*, ICACCI '12, pages 945–951, New York, NY, USA, 2012. ACM.

[27] Lisa F Rau. Extracting company names from text. In *Artificial Intelligence Applications, 1991. Proceedings., Seventh IEEE Conference on*, volume 1, pages 29–32. IEEE, 1991.

[28] Diana Santos, Nuno Cardoso, Paula Carvalho, Iustin Dornescu, Sven Hartrumpf, Johannes Leveling, and Yvonne Skalban. Getting geographical answers from Wikipedia: the GikiP pilot at CLEF. In *CLEF*, 2008.

[29] Jeong-Woo Son, A-Yeong Kim, and Seong-Bae Park. A location-based news article recommendation with explicit localized semantic analysis. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 293–302, New York, NY, USA, 2013. ACM.

[30] James D. Teresco. Using the Google Maps API with highway mapping data as a pedagogical tool: Demonstration. *Journal of Computing Sciences in Colleges*, 26(6):58–60, June 2011.

[31] Iraklis Varlamis, Nikos Tsirakis, Vasilis Poulopoulos, and Panagiotis Tsantilas. An automatic wrapper generation process for large scale crawling of news websites. In *Proceedings of the 18th Panhellenic Conference on Informatics*, PCI '14, pages 1:1–1:6, New York, NY, USA, 2014. ACM.

[32] Jan Vosecky, Di Jiang, and Wilfred Ng. Limosa: A system for geographic user interest analysis in Twitter. In *Proceedings of the 16th International Conference on Extending Database Technology*, EDBT '13, pages 709–712, New York, NY, USA, 2013. ACM.

[33] Wei Wang, Chuan Xiao, Xuemin Lin, and Chengqi Zhang. Efficient approximate entity extraction with edit distance constraints. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 759–770. ACM, 2009.

[34] Peter Watson. Sotheby's: The Inside Story. *Virginia Quarterly Review*, 74(3):104, 1998.

[35] Marinka Zitnik. Dynamics of news from the New York Times. *XRDS*, 21(1):64–66, October 2014.