Extraction of Spatio-Temporal Data for Social Networks

Judith Gelernter, Dong Cao, Kathleen M. Carley

School of Computer Science, Carnegie-Mellon University,
5000 Forbes Ave., Pittsburgh, PA 15213, U.S.A.
{gelern, kathleen.carley}@cs.cmu.edu, dongcaocmu@gmail.com

Abstract. It is often possible to better understand group change over time through examining social network data in a spatial and temporal context. Providing that context from a text analysis perspective requires identifying locations and associating them with people. This paper presents our GeoRef algorithm to automatically do this person-to-place mapping. It involves the identification of location, and uses syntactic proximity of words in the text to link location to person's name. We describe an application using the algorithm based upon a small set of data from the Sudan Tribune divided into three periods in 2006 for the Darfur crisis. Contributions of this paper are (1) techniques to mine for location from text (2) techniques to mine for social network edges (associations between location and person), (3) use of the mined data to make spatio-temporal maps, and (4) use of the mined data to perform social network analysis.

Keywords: Social network analysis, text mining, geographic data mining, media monitoring, Geo-IR, GIR, topic georeferencing, spatio-temporal tagging, geo-parse

1 Introduction

Texts can help us answer the question "who is where?" But automatically identifying person–location pairs in texts requires understanding the text. This paper describes how to construct an artificially intelligent algorithm that "understands" which words are places with the help of an authoritative place list called a gazetteer. The algorithm associates entity-place pairs with one another as the basis of a two-mode, people-by-location network.

Social network analysis uses advanced mathematical techniques and statistical analysis to examine the relationships among group members. These members, or "entities" are represented by nodes, and the relationships among the nodes are represented by links (also called edges), that in a diagram are shown as lines between nodes. Nodes may be people, organizations, locations, events, resources, topics, etc.

Identifying locations. Location is particularly valuable in analyzing certain kinds of networks. In epidemiology, geographic context proves more important than personal contact in understanding the spread of disease [1]. In disaster response, location of events and how they change over time can allow relief efforts to be coordinated efficiently [2]. In crime investigation as well as prevention, location is used to spot patterns and learn where to enforce preventive measures [3].

Identifying locations in a text is a complex problem. Named Entity Recognition typically includes identifying names of locations as well as people and organizations [4]. Named Entity Recognition can reach almost 80% accuracy [5], but it is clear that automatically identifying location accurately is harder. For example, a group participant in GeoCLEF 2005, a conference devoted to geographic information retrieval (of which geospatial data mining is a part), had only 41% of documents relevant to a query identified [6].

Time in networks. Newspaper articles often begin with month and day of writing, so mining the date is straightforward. Including time information in a network allows us to examine how a community evolved, and how different network indices may be viewed in series [7]. Often this is shown visually with a series of diagrams or maps.

We are interested in the network measure known as centrality. According to the glossary of the social network analysis software, ORA, centrality is the nearness of an entity to all other entities in a network [8]. The calculation of closeness is the inverse of the sum of the shortest distances between each entity and every other entity in the network. It has been shown that centrality measures may be robust in light of missing data [9]. Visualizations of data in time series allows us not only to see changes from one time period to another, but also they may allow us to make inferences about data that is missing. For example, if we have evidence of a foreign presence in Kassala in February, March and May, one may infer that foreign presence resident in Kassala in April as well.

In §2 below we describe related projects and in §3 we discuss the types of difficulties involved in mining for location words.[1] We describe the data in §4, and the GeoRef algorithm in §5. We provide several ways to evaluate our work in §6, both in demonstrating the accuracy of the algorithm's location-identification capabilities, and in dividing the data into spatio-temporal maps and giving network statistics for each of three time periods to show who was involved where. Discussion about algorithm optimization and generalizability follows in §7. We summarize in §8 what we believe are our main contributions.

2 Related Work

Much research has focused on the extraction of social networks from texts [10],[11]. As named entity identification has improved, so too has the extraction of social

[1] The footnote numeral is set flush left and the text follows with the usual word spacing.

network data. Here we take the extraction of social network data as a given. Our interest is in extracting locations and in linking those locations to the social network.

Finding locations. Locations represent a particular challenge within named entity identification [12]. In addition to the standard approaches and heuristics, gazetteers are used. Gazetteers differ in scope, coverage, accuracy, and specificity of entries. Choice of gazetteer by necessity will influence match results. The gazetteer can supply additional background knowledge that is helpful in data analysis. Some researchers use existing gazetteers such as the National Geospatial Intelligence Agency gazetteer[2] or GeoNames,[3] while others generate them automatically [13] or derive them from Wikipedia [14]. Researchers have extended the problem of finding location names in text to identifying regions that signify places, such as "downtown" or "by the docks" [15], [16]. Such are not generally found in gazetteers.

Location-mining software has gone commercial. MetaCarta, [4] will locate places named in a document or text stream. Yahoo! Placemaker[5] has a web service to do the same. These applications might use gazetteers that are not inclusive enough to find small towns or vague regions named in the text. But using very large gazetteers can slow processing.

One way to improve the ability of a computer to recognize location is to use a specialized gazetteer. [17] devised a Location Aware Topic Model to discover topics and the location related to that topic. They extend the gazetteer by adding words with implied locations not found in a standard gazetteer. For example, they add leaders and connect leaders to their country (Barack Obama to the United States), events to countries (Olympics 2008 to Beijing, China), and groups to region (Sunni to the Middle East). Such a gazetteer, however, is difficult to prepare with any degree of thoroughness.

Techniques for spatiotemporal knowledge discovery have been described by [18]. Geospatial data mining begins with toponym resolution, or attaching a location to a place named in a text [19]. The difficulty is that not all location words are associated with actual locations, in what is called non-geo/geo ambiguity (is "Mobile" a phone or a town in Alabama?). The other problem is geo/geo ambiguity, introduced when there are several places with the same name [20].

Linking people to location. Extracting relations between entities is substantially harder than entity recognition, and state-of-the-art systems perform less well on this task. Most relation extraction work assumes that entities have been identified correctly. Main methods for extracting relations between entities are to discover verb relations [21], construct concept graphs based on rules [22], or find syntactic proximity based on inference. The limitation of the syntactic proximity techniques is that they tend to miss links that are implied in the text. For example, while they

---

[2] National Geospatial Intelligence Agency gazetteer for download at http://earth-info.nga.mil/gns/html/The footnote numeral is set flush left and the text follows with the usual word spacing.

[3] The footnote numeral is set flush left and the text follows with the usual word spacing.

[4] http://www.metacarta.com

[5] http://developer.yahoo.com/geo/placemaker/guide

typically identify linkages in the same sentence, they less often find linkages that are expressed later in the paragraph.

3 Mining a text for location words: Implications for the GeoRef algorithm

Geoparsing is the identification of place names in a text, it is the backbone of the novel GeoRef algorithm. Geo-coding assigns latitude and longitude to a location [23].

We give examples of potential errors in mining locations from our data domain of the Sudan Tribune, and then describes what was done in the GeoRef algorithm to resolve those errors. We find two types of errors: place names that are not recognized as places, and non-place names that are taken to be place names incorrectly. We show how each is treated in our GeoRef algorithm. Then we conclude with only a mention of the related problem of deciding which of multiple versions of the same place name in a gazetteer is the one referred to in the text.

3.1 Place names not recognized as places

Location words are recognized as places based on matches with the gazetteer. The types of errors of places not recognized – large places, small places, places with multiple spellings, and imprecise regions – result in type I error (omission of the correct response), and may all be improved by adjusting the gazetteer. This section illustrates in italics each difficulty in the context of our data domain.

Large places. These are regions which correspond to a geographical area larger than a country, which do not appear in standard gazetteers. We need to list countries that comprise the region in order to determine the geographic coordinates.

Examples:

> …[T]he regime hopes to have a fig-leaf international presence with which to cultivate support among the *Arab and Islamic worlds*, and from stalwart economic partners Russia and China. wk35_4j

Algorithm solution:

1) Add Arab world and Islamic world to gazetteer
2) Resolve multi-country regions into those countries that comprise them for the gazetteer
3) Enter geographic coordinates of the centroid of the multi-country region in the gazetteer

Small places. Towns or neighborhoods known locally may not appear in a world gazetteer. In the examples below, the reporter supposed even Sudan Tribune readers would not know the location of "Deleige" and "Tawilla", so the towns are followed in the same sentence with geographical descriptors in parentheses.

Examples:

> The humanitarian organization Tearfund reported the death of a member of its relief team in *Deleige* (Wadi Saleh), West Darfur. wk 30_6l
>
> …Minawi and his soldiers deny responsibility for the violence in *Tawilla* (west of el-Fasher in North Darfur) and other towns in the region … wk 30_6l

Algorithm solution:

1) Add small towns to gazetteer
2) Enter geographic coordinates of each small town centroid to the gazetteer

Multiple spelling. In names transliterated from other languages, multiple spellings may be a problem. "Kordofan" appears as "Kurdufan" in GeoNames, for example. Also, punctuation might be lax. The U.S. without punctuation matches the pronoun "us," and so might not be found in a text.

Examples:

> …a rally at Zeribah in *North Kordofan*, central Sudan." wk 30_6l
>
> "It is another disaster for *US* policy." wk 30_6l

Algorithm solution:

1) Add multiple spellings of the same place to gazetteer
2) Recognize upper case "US" as the United States. We cannot use lack of punctuation only, or else each predicate pronoun "us" will be resolved geographically

Imprecise regions. Many imprecise regions such as "north of the city" or "down by the docks" do not necessarily correspond to a precise geographic area. However, rather than lose a geographic reference, we bound these regions artifically. For example, the Upper Nile region in Sudan comprises 3 states: the Upper Nile (state 20), Jonglei (22), and Unity (19). We resolve "southeast Upper Nile region" as the state of Jonglei, although it might be more precise to use the centroid of the eastern portion of Jonglei.

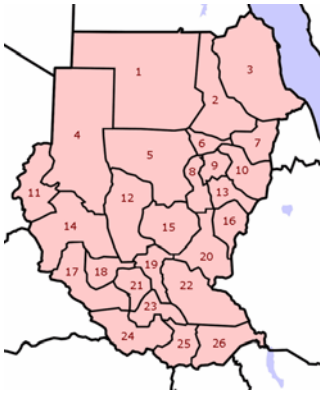

Fig. 1. States of Sudan. Image from Wikipedia http://en.wikipedia.org/wiki/States_of_Sudan, Retrieved November 2010

Examples:

…. between militias in the *southeast Upper Nile region* and the areas around Sudan's main oil fields which are in the south." wk 30_6l

Many fled from *south and western Sudan* during a famine in the 1980s. wk 35_7l

Algorithm solution:

1) Add to gazetteer the names of imprecise regions
2) Associate geographic coordinates with those imprecise regions

3.2 Non-place names mistaken for places

The words in italics in the examples below all are listed in the gazetteer as place names. In context, however, they do not refer to places.

Common words. Ambiguities are created when country names are transliterated. "Nor" and "Both," according to the GeoNames gazetteer, happen to be populated places in Sudan's Upper Nile. We surmount this problem by filtering the gazetteer for common words.

Examples:

With Chad's government neither willing *nor* able to protect rural populations, a massive increase in violence and civilian destruction seems *both* imminent and inevitable. wk 40_dd

Algorithm solution:

1) Create a separate list of geo-words that are also common words
2) For any geo-word found also on the list of common words (such as mobile, or nor), only attach geographic coordinates if that word is immediately preceded in the same sentence by another place name, or immediately followed in the same sentence by another place name.

Named Entities. Titles and organizations might include geographical names that may be a source of confusion in data mining.

Examples:

Those countries that have the required military assets must be ready to deploy them." ("Darfur Descending," The *Washington* Post, January 25, 2006) wk 28_av

'France is taking steps to stop the genocide as fast as possible,' he said on Radio *Monte Carlo* … " wk35_4j

Algorithm solution:

1) List commonly-appearing phrases that contain geo-words, such as New York Times.
2) Do not attach geographic coordinates to geo-words found within those phrases

Metonymy. The literary conceit known as metonomy borrows the name of one thing to stand for another with which it is associated. Metonomy creates confusion especially in news articles because a capital city often indicates that country's government.

Examples:

… the inability of the donors conference to compel *Khartoum* to accept a robust UN force … wk 28_av

"… to allow a UN mission into Darfur to replace an African Union force that has been unable to stem the violence *Washington* called genocide. wk 28_av

Algorithm solution:

1) Do not attach geographic coordinates to capital cities if followed by the words "regime" or "government". Example: In the phrase "Khartoum regime," Khartoum would not be considered a place.

2) Admittedly, this solution is inadequate in many cases of metonomy. Researchers are encouraged to work on this problem.

3.3 Which is the correct match in the gazetteer?

Confusion arises when there are two or more places with the same name. Leidner [12] relates rules that have been used by different researchers to resolve the problem of two places of the same name in the gazetteer that match a place named in the data. Examples of such disambiguation rules are: among same-named places in the gazetteer, choose the place that is higher in the geographical hierarchy (country above city), that is more populous, that is within the geographic domain of the data, or that is closer in distance to other non-ambiguous places named in the data.

4 Data and resources for data processing

News articles. Our team decided that mining articles after the separation of Southern Sudan and as the Darfur conflict escalated in 2006 would provide insight into the events of the time. Several thousand short pieces in the form of news reports, commentary, and an occasional published letter make up the full data set that was downloaded mostly from the Sudan Tribune.[6] We selected 11 files randomly from among these for creating, refining and testing the GeoRef algorithm. File size differs because article length differs depending upon who wrote the article, the significance of week's events, and so on. One article might contain 500 words, while another's might have closer to 5000. Locations in the 11 articles were annotated manually by coders guided by instructions in the Appendix. We divided the texts into three sets according to the time periods they represented: January 2006 (two files), March—April 2006 (4 files), and May—July 2006 (5 files). These divisions were made to balance the number of people's names found in the files, because the first two files were particularly rich in names.

Thesaurus. We manually created an external thesaurus for people's names to tag the names of people in the data. Very few of these people are political officials or foreign dignitaries, so our preliminary experiments with a thesaurus less fitted to the domain proved useless.

Gazetteers. GeoNames is attractive for our domain context because it includes alternate spellings, and many of our place names are transliterated from Arabic. We could not use the entire gazetteer since it would slow processing greatly. Instead, we limit the gazetteer to the GeoNames features of continent, first- and second-order administrative divisions, seat of a first-order administrative division, independent

[6] Recall that in 2005, a separate government formed in southern Sudan in opposition to the dominating regime, and in early 2006, Sudan rejected United Nations peacekeeping efforts. Many died during the conflict between north and south in 2006. Sudan accepted African Union peacekeeping help in November 2006, and then accepted United Nations peacekeepers in early 2007.

political entities and dependent political entities, territories, zones and buffer zones. Only for Sudan did we use entities in all feature classes. This serves the purposes of speeding processing without sacrificing resolution of location names in the data. Our gazetteer excerpt has 25,926 entries (8% for the world, and 92% for Sudan states, villages and topographical features).

5 The GeoRef algorithm

The three main tasks of the novel algorithm are to find location words in the data, identify locations with a precise area for which we can give latitude and longitude coordinates and a spatial hierarchy (for a city, for instance, it outputs also state and country). Also, it associates locations and other entities in text, such as topics or people. Figure 2 presents an overview, which is followed by an outline of the steps the algorithm follows. Plotting people-location pairs is a form of georeferencing, hence the algorithm name, GeoRef.

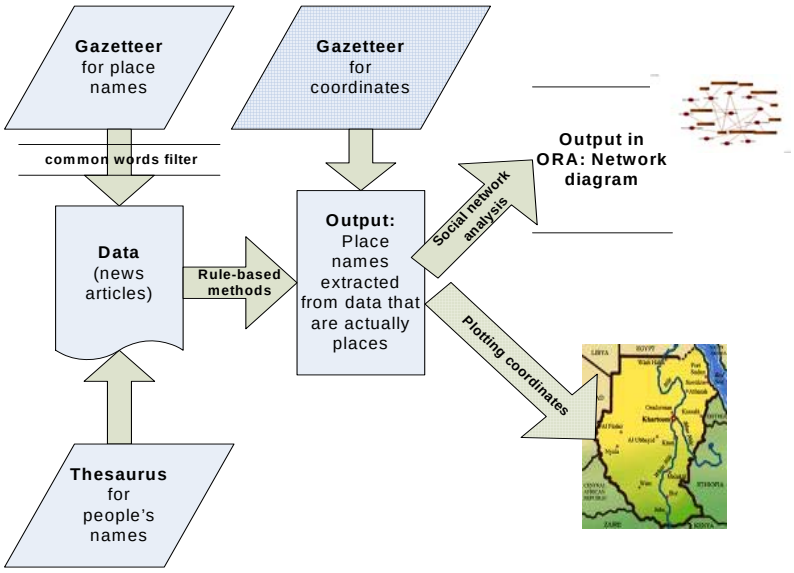

Fig. 2 The algorithm runs a thesaurus over the data to find people, and then a gazetteer to find potential geo-words to link to those people. Note that plotting topic-location pairs is a form of geo-referencing. The Sudan map is re-printed with permission Mill Hill Missionaries.

Steps followed by the algorithm

1. *Mine for people's names*. The identification of the names in the given texts was done manually. These names were assembled into a list. Then this list was used to identify the people's names automatically in order to prepare the text for finding associated locations.

2. *Mine for location.* The process of mining for location was automated. We used the very large gazetteer GeoNames, with small city and town entries for Sudan only, and we filtered the gazetteer for common names. Additional heuristics to disqualify place names in instances of metonymy and when the place name is embedded in other named entities are invoked.

3. *Associate each person's name with a location in text*. To find a location to match with a person, the program mines the geo-word closest to the person's name that is in the same clause. If the sentence contains two geo-words, it mines the geo-word that is before rather than after the person's name. If there is no geo-word in same sentence as the name, it looks in the same paragraph, first in the sentence immediately before, then immediately after, then anywhere in the paragraph. When no locations are found, it looks to the article title and first paragraphs for a geo-word.

4. *Determine link strength.* We assign relative strength to the name-location connection based on how certain we are that the connection is valid. This we infer from the distance in the document between linked words. If the geo-word is found anywhere in the same paragraph as the name, the link to the name is considered strong; if the program must find a geo-word in the title or in the article's first paragraph, the link is considered weak. In cases where no geo-word is found in any of the places suggested, we consider the connection to be invalid because no pair can be made with any degree of assurance. We describe this in pseudocode below.

   ```
   If geo-word occurs in same paragraph = 2 (strong link)
   Else if geo-word occurs in title or
     first paragraph of article  = 1 (weak link)
   Else  = 0
   ```

5. *Enrich location output*. The geo-word mined from the text is enriched with the upper levels in the spatial hierarchy. It is also associated with the geographic coordinates of that region's centroid (the geometrical center of the region) so that the location can be plotted on a map as a single point. This is done automatically.

6 Testing the GeoRef algorithm

We describe separately a test for GeoRef location-identification, and a test for GeoRef associating locations with people.

6.1 Location Identification

An annotator was given 11 texts downloaded from the Sudan Tribune and asked to list every location found, following only a few guidelines (reported in the Appendix). Even a paid participant can tolerate only so much of a dull task without succumbing to fatigue and error. We wished to retain validity by asking a single annotator to go through all the documents. Our sample size, therefore, was limited by human attention constraints.

Procedure. We illustrate below an excerpt from the text, and examples of what the manual coder and what the algorithm selected as location words from that text (Table 1). The manual coding was used as a benchmark to judge the degree of algorithm accuracy both at the corpus (collection) and at the document (text) level.

Table 1: Comparison of one manual coder's decisions about what locations this paragraph contains to the GeoRef output

Manual coding	GeoRef
Southern Sudan	Southern Sudan
England	White Nile
Block Ba	England
Junglei	White Nile
Southern Sudan	Junglei
Junglei	Southern Sudan
Bor	White Nile
	Junglei
	Bor

White Nile Petroleum and Dinka Bor Extinction [article title]
Tuesday 2 May 2006 23:30. By Deng Ajak Jongkuch [file: week17]
May 1, 2006 On April 25 of 2005, the government of Southern Sudan signed a 10 year oil exploration licensing contract with a White Nile Limited owned by a former Middlesex and England cricket star Phil Edmonds. The agreement gave White Nile Limited a right to explore oil in Block Ba which covered an area of 67,000 km of State of Junglei. The area is believed to have about 6 billion of barrels oil in reserved. According to agreement, a government of Southern Sudan owned oil company Nile Petroleum Corporation will own 155 shares and 40% in stakes while White Nile Limited will retain 60% of stakes. According to CPA of wealth sharing modelity <sic>, 2% of oil revenues will go to where oil exploration will take place. The government of Junglei is responsible for the 2%, not the Bor County.

Scoring. We measure results at two levels—the corpus level and the document level.

At the corpus level, we identify the percent of location words identified manually that were also identified by GeoRef. For each document, we count the number of time GeoRef identified a location correctly, missed a location (type I error), and added a location not found in the manual benchmark (type II error).

We score a location correct if the place is an exact match with the manual coding or if the place is a subset of that found by the manual coder. So if, for instance, the manual location is Junglei, and algorithm found Bor which is in Junglei, the GeoRef algorithm found Junglei too, so the output is considered correct. We also score a GeoRef location to be a match with that of the manual coder if it is higher in the hierarchy. So for example, if the coder found Southern Sudan and the algorithm found Sudan, we score the algorithm to be correct.

Results: Corpus level and Document level accuracy. At the corpus level, the GeoRef algorithm yielded 62% accuracy, with a standard deviation (spread around the mean) of 18%. We get this number by adding together true positives, true negatives and false positives according to the formula.

$$\text{accuracy} = \frac{\text{number of true positives} + \text{number of true negatives}}{\text{numbers of true positives} + \text{false positives} + \text{false negatives} + \text{true negatives}}$$

This 62% accuracy represents a significant improvement over the 41% accuracy reported by one of the participating groups in the 2005 GeoCLEF geographic information retrieval conference [6]. Document level results are arranged by file number in Table 2.

Table 2 shows the number of locations mined from text correctly (True Positive or TP), the number of locations found that are not actually locations (False Positive or FP), and the number of locations that should have been found that were not (False Negative or FN). Then we calculate precision and recall statistics as well as the combination of precision and recall called the F-measure. These are classic information retrieval evaluation tests for whether all locations are found correctly (precision), and for whether all locations found should be found (recall). The F-measure combines the two. [7]

Limitations of experiment 1. This experiment based on manual coding as benchmark is limited by the sample size of the data. Sample size is limited by human constraints in that asking a person to coding too much introduces error from fatigue. As it is, in our small data set, our coder missed numerous locations due to lapses in attention, especially in articles that are longer. The result is that many correct locations are marked wrong in the algorithm output. The accuracy of the algorithm, therefore, is somewhat higher than the statistics suggest.

Discussion of experiment 1. What accounts for the errors? GeoRef found places that are not actually places (type II error), whether because they are within the names of organizations or companies, or because they are capitals that stand for countries as a form of metonymy. GeoRef omitted the names of villages too small to appear in

[7] *Introduction to Information Retrieval* by Manning, Raghavan and Schütz, 2008

the gazetteer (type I error). Type II error is far more common here that Type I error, both because metonymy and other errors of place name ambiguity occur regularly and are not well-managed by the algorithm, and because the data coder missed places that are legitimate.

Table 2 Document level accuracy on the GeoRef algorithm.

File name	Correct (TP)	Incorrect (type II) (FP)	Missing (type I) (FN)	Precision (%)	Recall (%)	Accuracy (%)	F-measure (%)
wk_1	24	3	1	89	96	86	92
wk_2	105	92	8	53	93	51	68
wk_9	19	0	3	100	86	86	93
wk_10	10	4	2	71	83	63	77
wk_11	9	3	2	75	82	64	78
wk_14	10	3	3	77	77	63	77
wk_17	31	33	16	48	66	39	56
wk_22cm	11	11	4	50	73	42	59
wk_22 3q	16	12	13	57	55	39	56
wk 22xs	31	7	1	82	97	79	89
wk 26	23	8	2	74	92	70	82
Mean	26.27	16	5	**71**	**82**	**62**	**75**
Standard deviation	**27.34**	**26.75**	**5.11**	/	/	**18**	/

6.2 People-location link identification

We measure the people–location links using the same 11 files as in the location experiment.

Procedure. The people thesaurus was used to code for people and the GeoNames gazetteer excerpt was used to code for locations. The 11 files were separated into groups representing the three time periods, and each group was run independently.

Scoring. We measure the number of nodes and links first. The data was input into the social network analysis software ORA [8], [24] to find the number of links shown in Table 3.

Result: Links.

Table 3: Distributions of codes in each of the three periods, with link statistics supplied by the ORA social network analysis software.

Distribution of Codes

	Time period 1	Time period 2	Time period 3
Number of people	12	12	7
Number of locations	11	10	6
Number of links	48	24	20
Total	**71**	**46**	**23**

Result: Maps.

The GeoRef algorithm attaches geographic coordinates to locations. To create a visualization, we associate the coordinates of the person with his paired location and then plot each using Google Earth. We label each red site indicator with the number of occurrences of that entry in the data set. The color of label and the color of the tear drops are otherwise arbitrary. The maps for the three time periods (Fig. 3, 4, 5) give an idea of who were the actors in different phases, where they or their influence was, and how the situation changed as the year progressed.

Fig 3 (above): Time Period 1: Jan 2006 (files wk1, wk2) ©2010 Google, Map ©2010 Tele Atlas

Fig. 4 (above) Time period 2: March-April 2006 (files for wk 9,10,11,14) ©2010 Google, Map ©2010 Tele Atlas

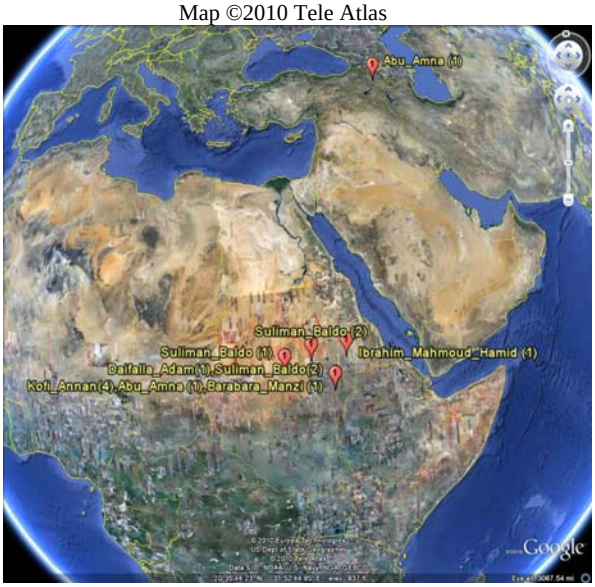

Fig 5 (above): Time period 3: May-July 2006 (wk 17, 22_1c, 22_3q, 22xs, 26) ©2010 Google, Map ©2010 Tele Atlas

Interpretation of maps. Data analysis is based on examination of the Fig. 3, 4 and 5 maps for the three time periods. At the beginning of 2006, we see that influence was concentrated in the east and south of Sudan with dominant actors being Jan Pronk, UN-appointed special envoy to Sudan and Kofi Annan who was Secretary General to the United Nations. The influence of the United Nations diminishes somewhat in the spring, according to the documents mined, with prominent actors being Sudanese. Abdallah Moussa Abdalla, secretary-general of the Beja Congress Party in Port Sudan, and the Sudanese who were arrested in the Gadaref state, al-Almin al-Hajj the president and Hassan al-Masri, the treasurer showed the center of the action moving to Sudan's north. Then by the late spring, Kofi Annan of the United Nations and Suliman Baldo who is the deputy director of the International Center for Transitional Justice make foreign presences again prominent, with the center of concern being Sudan's east.

This interpretation is some reflection of actual happenstance since it derives from news articles. However, a more accurate picture would result from a much larger data set. Also, there are errors generated by incorrect association of persons with locations. Other errors are caused by differences in precision of the data mined such that some people are located by city while others are located to the level of Sudan only (and hence appear in numbers in the country's center).

Evaluation of maps. The utility of the maps will depend upon the map user's purposes. Historians, political scientists, and anthropologists, for example, might use our annotated maps in series to follow where people were acting over time. We could annotate maps based on historical documents as easily as current newspapers, as is done here.

Our annotations show a person's village, city, state or country, depending on what hierarchical level of location is named in the text. We could improve the utility of the maps in future by drawing a bounding box around the location being mapped to indicate the region indicated. This is not a present function of the software, however.

Result: Network performance.

We are interested in calculating which people and which locations are most central to the group of people in the network constructed for each of the three time periods. The people and locations mined and linked by GeoRef are input into the social network analysis software ORA. The set of people–location pairs for each time period is formed into a network and then reduced in order to calculate the statistics. Finally, we run the ORA "All Measures Report" on each network for each of the three time periods to produce the tables below.

The actors in these networks extracted from news data are not necessarily linked to one another, yet the centrality position vis à vis the others makes it the measure we discuss here. Degree centrality measures the number of direct links an entity has. Because links flow out from people and in to the locations, we are interested in the "Centrality-Out Degree" measure for people and the "Centrality-In Degree" measure for locations. The software calculates centrality numerically, giving people and places numerical scores based on the number of linkages they have. These raw values

are then normalized on a scale between 0 and 1. We take only the top 5 results for each measure.

Time period 1

***Table 4A**: Period 1, Centrality-In Degree for Locations*

Rank	Location	Scaled Value
1	SOUTHERN SUDAN	1.000
2	SUDAN	0.786
3	WILAYAT AL KHARTUM	0.500
4	HAMESH KHOR	0.143
5	AFRICA	0.071

***Table 4B.** Period 1, Centrality-Out Degree for People*

Rank	Agent	Scaled Value
1	KOFI_ANNAN	1.000
2	PRONK	0.438
3	LAM_AKOL	0.250
4	AMNA_DIRAR	0.125
5	ALI_EL-SAFI	0.125

Time period 2

***Table 5A**. Period 2, Centrality-In Degree for Locations*

Rank	Location	Scaled Value
1	PORT SUDAN	0.455
2	KASSALA	0.182
3	AL QADARIF	0.091
4	THE EAST	0.091
5	SUDAN	0.091

Table 5B. *Period 2, Centrality-Out Degree for People*

Rank	Agent	Scaled Value
1	ABDALLAH_MOUSSA_ABDALLAH	0.500
2	AL-AMIN_AL-HAJJ	0.100
3	HASSAN_AL-MASRI	0.100
4	SULIEMAN_DERAR	0.100
5	SIMA_SAMAR	0.100

Time period 3

Table 6A: *Period 3, Centrality-In Degree for Locations*

Rank	Location	Scaled Value
1	THE EAST	1.000
2	ENGLAND	0.091
3	KASSALA	0.091
4	WILAYAT AL KHARTUM	0.091
5	SUDAN	0.091

Table 6B*: Period 3, Centrality-Out Degree for People*

Rank	Agent	Scaled Value
1	KOFI_ANNAN	0.833
2	SULIMAN_BALDO	0.667
3	ABU_AMNA	0.500
4	PHIL_EDMONDS	0.167
5	IBRAHIM_MAHMOUD_HAMID	0.167

Interpretation of network performance results. We see from the tables above that the central location in the first period is southern Sudan and the major actors are those associated with the United Nations, Kofi Annan and Jan Pronk. In the second period, the central actors are associated with Sudanese politics and the most central regions are northern Sudan. In the third period, the focus is more in eastern Sudan, in Kassala and also Khartoum.

The data is the same as the data as was used for the geographical maps, so it is not surprising that the top actors and locations resemble that in the three maps. This mathematical presentation validates the mapped visualizations.

7 Algorithm optimization

We have in coding GeoRef made two decisions that optimize processing time at the expense of gazetteer coverage and algorithm generalizability. We discuss both choices here.

GeoNames contains over 10 million geographical names, with the main download file being 878 MB.[8] We suggest that those using GeoNames as an external referent use only a gazetteer excerpt unless other optimization methods such as parallel processing will be used.

The few rules in the algorithm for metonymy and for place names embedded in organization names are generalizable to other news domains. These rules will be of limited use in other text domains. Metonymy is unusual outside of the political domain, so should not be a drawback. Other domains, however, such as business or history might have a fair number of place names embedded in titles or organization names. We recommend these as areas of further research.

In this paper, we have mined the date of the newspaper articles that appears in each article's beginning. This date mining will not generalize beyond news media. Extending temporal data mining to other sources can use natural language processing methods. For example, we could translate an event into a date (Thanksgiving), or to add time to an event ("a conference that occurred last week"). The algorithm would get the date of the event either by data mining or from a temporal thesaurus.

Another way we optimized in coding GeoRef is to allow a frankly superficial level of understanding of the people-location link. Deeper exploration of the people-location link remains for future work. Is the person travelling to the (linked) place? Does he start in the place? Is he in the place temporarily? Was he in the place at one time but no longer?

8 Contributions in summary

This paper describes how to mine news articles for the names of people and locations using a novel GeoRef algorithm in preparation for social network analysis. We have offered heuristics for mining location, and for associating a person's name with a location. We use a spatio-temporal network approach and an application with data from news articles. We map the data and also give network statistics to compare change over time in whom and where are the network actors.

We present a network with node labels enriched beyond what is possible through data mining alone. The gazetteer supplies upper levels of the spatial hierarchy in additional to geospatial coordinates, so that for example, given city, the algorithm supplies state/province and country. Even so, non-optimal levels of accuracy imply that the spatiotemporal methods must be re-examined. Methods for evaluation might also be improved.

---

[8] The size quoted is as of November 2010

Acknowledgments

Thanks are due to Michael Bigrigg for his insights into the network analysis. This work was supported in part by the Air Force Office of Sponsored Research (MURI: Computational Modeling of Cultural Dimensions in Adversary Organizations, FA9550-05-1-0388), the Army Research Institute W91WAW07C0063, and the Army Research Office ERDC-TEC W911NF0710317. Additional support was provided by the Center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Air Force Office of Sponsored Research, Army Research Institute, the Army Research Office, or the U.S. government.

References

1. Chen, Y-D, Tseng, C., King, C-C, Wu, T-S, Chen H.: Incorporating geographical contacts into social network analysis for contact tracing in epidemiology: A study on Taiwan SARS data. In: D. Zeng et al. (eds). Intelligence and Security Informatics: Biosurveillance, LNCS vol 4506, pp. 23-36 New Brunswick, NJ (2007)
2. Ashish, N. Eguchi, R., Hegde, R., Cuyck, C., Kalashnikov, D., Mehrotra, S., Smyth, P., Venkatasubramanian, N.: Situational awareness technologies for disaster response. In: H. Chen, E. Reid, J. Sinai, A. Silke, B. Ganor (eds.) Terrorism Informatics: Knowledge Management and Data Mining for Homeland Security. Springer, New York (2008)
3. Skillicorn, D.: Knowledge discovery for counterterrorism and law enforcement. CRC Press, Boca Raton, FL (2009)
4. Giuliano, C., Lavelli, A., Romano, L.: Relation extraction and the influence of automatic named-entity recognition. ACM Transactions on Speech and Language Processing 5(1), pp.2:1-2:26 (2007)
5. Hassell, J., Aleman-Meza, B., Arpinar, I.B.: Ontology-Driven Automatic Entity Disambiguation in Unstructured Text In: Cruz et al. (eds.): ISWC 2006, LNCS, vol. 4273, pp. 44–57 (2006)
6. Gey, F., Larson, R., Sanderson, M. Joho, H., Clough, P., Petrasi, V.: GeoCLEF: The CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview. In: C. Peters et al. (eds.): CLEF 2005, LNCS, vol. 4022, pp. 908–919 (2006)
7. Danowski, J.A., Cepela, N.T.: Automatic mapping of social networks of actors from text corpora: Time series analysis. In N. Memon, R. Alhajj (eds.) International Conference on Advances in Social Network Analysis and Mining, 2009. ASONAM '09. 20-22 July, 2009, Athens, Greece, pp. 137–142 (2009)
8. ORA [Organizational Risk Analysis software], v. 2.0.8. Center for Computational Analysis of Social and Organizational Systems (CASOS), Institute for Software Research, Carnegie Mellon University. Copyright Kathleen Carley, 2001-2010. http://www.casos.cs.cmu.edu/projects/ora/
9. Borgatti, S.P., Carley, K.M., Krackhardt, D.: On the robustness of centrality measures under conditions of imperfect data. Social Networks 28(2), pp. 124–136 (2006).

10. Carley, K.M.: Coding Choices for Textual Analysis: A Comparison of Content Analysis and Map Analysis. In: Marsden P. (ed), Sociological Methodology, vol. 23, pp. 75−126. Blackwell, Oxford (1993)
11. Carley, K.M.: Network Text Analysis: The Network Position of Concepts. In: C. Roberts (ed.), Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts (chapter 4 pp. 79–100). Lawrence Erlbaum Associates, Hillsdale, NJ (1997)
12. Leidner, J.L.: Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names. Unpublished doctoral dissertation, University of Edinburgh, United Kingdom. Retrieved January 8, 2008 from http://hdl.handle.net/1842/1849 (2007)
13. Kozareva, Z.: Bootstrapping named entity recognition with automatically generated gazetteer lists. Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Morristown, NJ, pp.15–21 (2006)
14. Popescu, A., Grefenstette, G.: Spatiotemporal mapping of Wikipedia concepts. JCDL '10, June 21-25, 2010, Gold Coast, Queensland, Australia, pp. 129–138 (2010)
15. Purves, R., Clough, P, Joho, J.: Identifying imprecise regions for geographic information retrieval using the web. In: Proceedings of the GIS Research UK 13th Annual Conference (2005), pp. 313–318. Retrieved April 11, 2010 from http://www.dcs.gla.ac.uk/~hideo/pub/gisuk05/gisuk05.pdf
16. Twaroch, F.A., Jones, C.B. Abdelmoty, A.I.: Acquisition of vernacular place names from web sources. In: I. King, R. Baeza-Yates (eds), Weaving Services and People on the World Wide Web, pp. 195-214. Springer, Heidelberg (2009)
17. Wang, C., Wang, J., Xie, X., Ma, W-Y: Mining geographic knowledge using location aware topic model. GIR '07 November 9, 2006, Lisbon, Portugal, pp. 65–70 (2006)
18. Roddick, J.F. and Lees, B.G.: Spatio-temporal data mining paradigms and methodologies. In: H.J. Miller, J. Han (eds.) Geographic data mining and knowledge discovery, 2nd ed., (pp. 27–44) CRC Press, New York (2009).
19. Buttenfield, B., Gahegan, M., Miller, H. Yuan, M.: Geospatial data mining and knowledge discovery. Retrieved April 7, 2010 from http://www.ucgis.org/priorities/research/research_white/2000%20Papers/emerging/gkd.pdf (2001).
20. Amitay, E., Har'El, N, Sivan, R., Soffer, A.: Web-a-Where: Geotagging Web Content. SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, U.K., pp. 273–280 (2004)
21. Pazienza, M.T., Pennacchiotti, M., Zanzotto, F.M.: Discovering verb relations in corpora: Distributional versus non-distributional approaches. In: A. Ali, R. Dapoigny (eds.) IEA/AIE 2006, LNAI vol. 4031, pp. 1042–1052 (2006)
22. Xu, X., Mete, M., Yuruk, N.: Mining concept associations for knowledge discovery in large textual databases. SAC '05, March 13-17, 2005, Santa Fe, New Mexico, U.S.A., pp. 549–550 (2005)
23. Roongpiboonsopit, D., Karimi, H.A.: Quality assessment of online street and rooftop geocoding services. Cartography and Geographic Information Science 37(4), pp. 301–318 (2010)
24. Carley, K.M., Reminga, J., Storrick, J. Columbus, D.: ORA user's guide 2010. Carnegie Mellon University, School of Computer Science, Institute for Software Research, Technical Report, CMU-ISR-10-120. Retrieved October 14, 2010 from http://www.casos.cs.cmu.edu/publications/papers/CMU-ISR-10-120.pdfTopic

Appendix

Excerpt from directions given to data coder to determine what constitutes a place name.

What is a location? noun
City, State, Country
Named landscape features: River Nile
Non-specific region (in the east)

How to record locations named in the text? Please indicate what place is actually meant. Example: "The West = Europe and the U.S." ; "The east = Eastern Sudan"; "Eritrean capital = Asmara"

Locations within an organization name can be omitted because they are not counted as place names

Locations used as metonomy for a gov't can be omitted
Ex. Khartoum = Sudan
Ex. Washington = U.S.