"GeoPlot": Spatial Data Mining on Video Libraries*

Jia-Yu Pan, Christos Faloutsos
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA

{jypan, christos}@cs.cmu.edu

ABSTRACT

Are "tornado" touchdowns related to "earthquakes"? How about to "floods", or to "hurricanes"? In Informedia [14], using a gazetteer on news video clips, we map news onto points on the globe and find correlations between sets of points. In this paper we show how to find answers to such questions, and how to look for patterns on the geo-spatial relationships of news events. The proposed tool is "GeoPlot", which is fast to compute and gives a lot of useful information which traditional text retrieval can not find.

We describe our experiments on 2-year worth of video data ($\approx$ 20 Gbytes). There we found that GeoPlot can find unexpected correlations that text retrieval would never find, such as those between "earthquake" and "volcano", and "tourism" and "wine".

In addition, GeoPlot provides a good visualization of a data set's characteristics. Characteristics at all scales are shown in one plot and a wealth of information is given, for example, *geo-spatial clusters*, *characteristic scales*, and *intrinsic (fractal) dimensions* of the events' locations.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data mining—*spatial data mining on video digital libraries*; H.3.1 [**Content Analysis and Indexing**]: Abstracting methods; H.3.7 [**Digital Libraries**]: Collection

General Terms

Algorithms

Keywords

Spatial data mining, video data mining, fractal dimension, intrinsic dimension, pair correlation, correlation integral

*This material is based upon work supported by the National Science Foundation under Cooperative Agreement No. IRI-9817496, IIS-9988876, IIS-0083148, IIS-0113089, IIS-0209107 and by the Defense Advanced Research Projects Agency under Contract No. N66001-00-1-8936.

CIKM'02, November 4–9, 2002, McLean, Virginia, USA.

1. INTRODUCTION

Events are not related only through their subjects, but also through the locations and time they occur. Geospatial data mining exploits the geographic information associated with spatial objects and finds interesting patterns, trends, and relations among them.

The geographic association is even more prominent among news stories. Incidences happen at one location or locations nearby usually have common or related subjects, or causal relations, since the characteristics of an area affect the way of living and, as a result, the incidences occurred on it. On the other hand, related events will occur together at many places.

Keyword-based methods have long been studied to find the relationship among news events. Keywords are assigned to news events, either manually by human experts after understanding the subject of a news event, or automatically by computer programs which simply select words from the transcripts reporting the event, assuming that words in the transcripts reveal the subject of the event. However, two events occurred at nearby locations and have effects on each other can not be found to have relationship, if they do not have shared keywords. One way to incorporate the concept of "closeness" is to assign extra locational keywords (such as adjacent cities or the country in which it happens) to "connect" events that are close to each other. However, this method does not scale when more and more locations needed to be considered, and these extra keywords also introduce noise into the processing.

How do we find patterns of global geographic phenomena? Does one event often come with another? Or does it repel another event? To find global patterns like these, we can not consider one place at a time , instead, all locations have to be examined at the same time. Keyword-based methods, which link locations to locations by keyword expansion, are not suitable for this task. One problem is when relating two faraway locations, many words will have to be included, which may end up introducing too much noise into the subsequent inference. In addition, it is difficult to present the notion of distance (degree of closeness) on terms.

In this paper, we propose a tool, *GeoPlot*, for mining global geospatial pattern. In particular, rather than measuring the closeness of two events by counting shared locational terms, it examines the geographic information more directly, in the sense that the actual physical distance between two places is computed and used as an indicator of

"closeness". We find that GeoPlot is effective on spotting global cross-event geospatial patterns. It detects the patterns that keyword-based methods can give, and also gives novel patterns missed by the keyword-based methods.

The paper is organized as follows. In section 2, we present some related works on geospatial data mining. In section 3, we introduce *GeoPlot*, and show how it works and how to interpret it. In section 4, our proposed method of using GeoPlot on geospatial data mining is explained. Section 5 gives experimental results on real world data gathered from a video digital library. Several discussions are given in section 6. Section 7 concludes the paper.

2. RELATED WORK

Spatial data mining focuses on finding interesting patterns, rules and trends among spatial objects. Often, spatial objects are stored in spatial databases as tuples with spatial attributes. Spatial attributes could be topological, such as adjacency or inclusion information, or geometric, such as position (longitude/latitude) or the boundary polygon. For a general survey, see [7].

Several approaches have been studied on spatial data mining [7], such as generalization as rule searching [9], clustering, and association rule. Attribute-oriented induction [8] learns rules by generalizing attribute values using spatial concept hierarchy (e.g., California is a generalization of Los Angeles and San Francisco). The quality of the rules produced depends on that of the concept hierarchy used.

Clustering techniques are used in spatial data mining to cluster objects based on their spatial attributes. Similarity between objects is often defined based on the physical distances among spatial objects, or their terrain types (hill or riverside). Interesting information is then inferred from the clusters formed. One drawback of the clustering techniques is that they tend to focus on local characteristics and are computationally expensive.

Association rule [1] has also been applied to spatial data mining. Spatial association rules are rules such as "*near(x, coast)* $\wedge$ *southeast(x, USA)* $\Rightarrow$ *hurricane(x), (70%)*", which says that if an object x is close to the coast and it is in southeast United States, then about 70% of the cases, x has hurricanes. Spatial association rules could reveal global relations among objects, not just the local ones.

Real-world data sets tend to have skewed distributions [15] and are self-similar [10]. Recent studies have found that the intrinsic (fractal) dimension is a good representation of real-world data [5] [6], where characteristics at both local and global scales are considered at the same time. It also spots non-linear correlations among objects. With these properties, the idea of intrinsic (fractal) dimension could be a good tool on spotting global spatial patterns inside real-world data sets. GeoPlot is based on this idea and aims to discover correlations among spatial objects, at both local and global scales. Related work includes the so-called K function [4] in spatial statistics [3], as well as the tri-plots [12].

3. BACKGROUND

In this section, we describe the concept of GeoPlot and how to interpret a GeoPlot to understand the characteristics of the data sets from which the GeoPlot is constructed.

A GeoPlot is defined on two given sets of points and is a plot which, given a distance r, indicates the number of point pairs that are not apart from each other by more than r. More specifically:

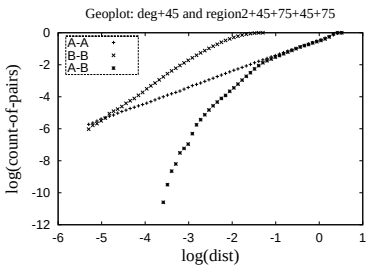


Figure 1: An example GeoPlot Curves labelled `A-A` and `B-B` are self-plots of the two data sets A and B, and `A-B` curve is the cross-plot. Set A has points along a 45^o big circle on the surface of the globe (Figure 2(c)). Set B has points inside a rectangle region between $[45^o, 75^o]$ longitude and $[45^o, 75^o]$ latitude (Figure 3(d)).

***Definition** 1.* Given two data sets A (with N_A points) and B (with N_B points), we define the **cross-plot** between the two datasets as the plot of

$$Cross_{A,B}(r) = \log\left(\frac{N_{A,B}(r)}{N_A \cdot N_B}\right) \text{ versus } \log(r),$$

where $N_{A,B}(r)$ is the number of point pairs (each consists of one point from A and B) within distance r.

***Definition** 2.* The **self-plot** of a given data set A (with N_A points) is the plot of

$$Self_A(r) = \log\left(\frac{N_{A,A}(r)}{\frac{N_A \cdot (N_A - 1)}{2}}\right) \text{ versus } \log(r),$$

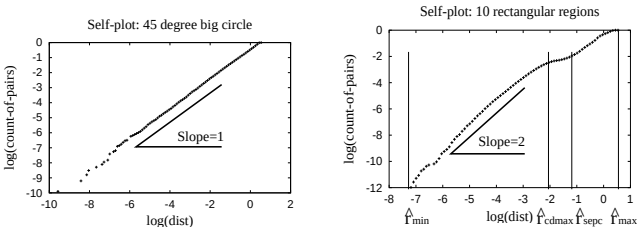
where $N_{A,A}(r)$ is the number of point pairs of set A within distance r. Note that the self-plot is indeed the "correlation integral" [10] of the data set A.

***Definition** 3.* The **GeoPlot** of two data sets A and B, is the graph which contains the cross-plot, $Cross_{A,B}(r)$, and the self-plots for both data sets, $Self_A(r)$ and $Self_B(r)$. Figure 1 shows an example of the GeoPlot .

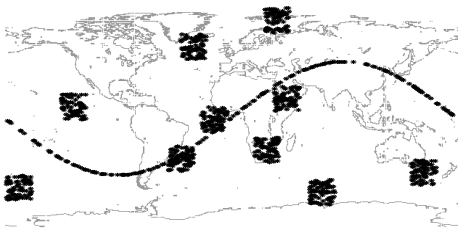
3.1 Characteristics of self-plots

We could observe interesting characteristics of a data set from its self-plot. Figure 2(a) shows the self-plot of a 45^o big circle on the globe (Figure 2(c)). Recall that self-plot is indeed the correlation integral whose *slope gives the intrinsic (fractal) dimension of the corresponding data set.* In this case, the slope is 1, since the circle is in fact an 1-dimensional object. Figure 2(b) is the self-plot of a set of clusters (Figure 2(d)). We can see *a flat portion (plateau) of the curve in the self-plot which indicates the existence of clusters* in the data set. In this case, the slope of the non-flat portion is 2, which corresponds to the intrinsic dimension of the clusters, which are 2-dimensional regions.

In Figure 2(b), we label 4 meaningful characteristic scales, namely, $\hat{r}_{min}$, $\hat{r}_{max}$, $\hat{r}_{cdmax}$, and $\hat{r}_{sepc}$.



(a) (Self-plot) 45^o big circle (b) (Self-plot) 10 regions



(c) 45^o big circle and 10 rectangular regions on the globe

Figure 2: Self-plots **(a),(b) self-plots of synthetic data sets shown in (c).**

- $\hat{r}_{min}(\hat{r}_{max})$ denotes the minimum(maximum) distance between two points of a given data set. In other words, $\hat{r}_{min}$ is the smallest distance where the count of pairs is not zero, and $\hat{r}_{max}$ is the distance that for distances bigger than $\hat{r}_{max}$, the counts remain the same.
- $\hat{r}_{cdmax}$ denotes the maximum diameter of the clusters.
- $\hat{r}_{sepc}$ denotes the inter-cluster distance.

Observation 1. **Characteristic scale** A plateau in the self-plot indicates the existence of clusters in the corresponding data set.

Observation 2. **Local/Global distribution behavior** The behaviors of a distribution which appear at scale less than (left to) the characteristic scale (plateau) can be considered as local behaviors. Those appear at scale greater than (right to) the characteristic scale are the global behaviors.

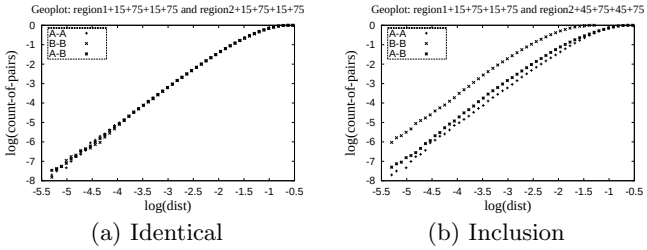
Observation 3. **Intrinsic dimension** If the self-plot is linear, its slope reflects the intrinsic (fractal) dimension of the corresponding data set.

3.2 Interpreting GeoPlot

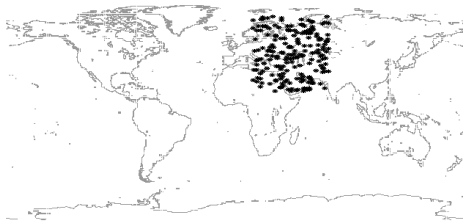
The relation between two data sets of locations on the globe (e.g. identical or disjoint) affects the look of their GeoPlot. A particular type of relation between two data sets causes a particular appearance of GeoPlot. In Observation 4, we list 5 possible relations and refer each case to its corresponding GeoPlot.

Observation 4. **GeoPlot Rules** We catalog the relations between 2 data sets (A and B) into 5 cases and show an example GeoPlot for each case. These are the rules for finding hidden relations between the distributions of 2 data sets from their GeoPlot.

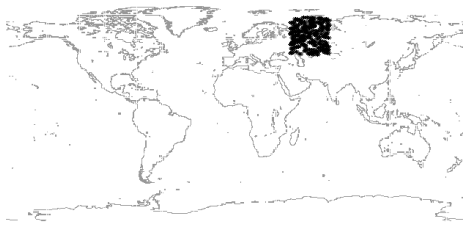
1. **Identical distributed**: The two data sets are from distributions statistically identical (Figure 3(a)). In other words, they have similar spatial distributions.



(a) Identical (b) Inclusion



(c) $[15^o, 75^o]$ longitude, $[15^o, 75^o]$ latitude



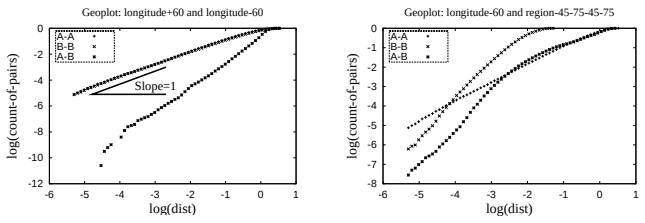
(d) $[45^o, 75^o]$ longitude, $[45^o, 75^o]$ latitude

Figure 3: GeoPlot : Identical and inclusion **(a) GeoPlot of 2 statistically identically distributed data sets. Both of them have points in the region as shown in (c) ((c) is one of them). (b) GeoPlot of (c) and (d). The coverage of the data set in (c) includes that in (d). Self-plot: A-A, B-B curves. Cross-plot: A-B curve.**

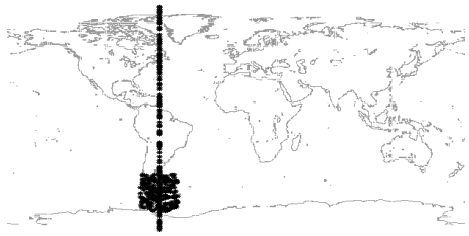
2. **Inclusion**: The distribution of one data set is included in that of the other data set (Figure 3(b)).
3. **Same dimension but not identical**: The two data sets have the same intrinsic (fractal) dimension but are not identical (Figure 4(a)).
4. **Dominating at different scales**: Distributions of two data sets may dominate each other at different scales. Figure 4(b) is an example, where data set A is the set of points scattered along meridian -60^o (Figure 4(c)), and data set B is the set of points in a rectangular region (Figure 4(c)).B dominates A at small (local) scale but is dominated by A at large scale. This is because distribution of B is more compact (localized), while A has more pairs separated at larger distances than B does, which makes A's characteristics remain significant at large scale.

4. PROPOSED METHOD

We would like to find relations among events such as "earthquake", "hurricane", "storm", and "flood". We look at the locations where these events occur, and check how one event is associated with another.



(a) A:Meridian+60^o, B:-60^o (b) 2 data sets in (c)



(c)-60^o meridian, and the region between [-45^o,-75^o] longitude and [-45^o,-75^o] latitude

Figure 4: GeoPlot : intrinsic dimension and local/global scale **(a) 2 data sets ($+60^o$, $+60^o$ meridians) have the same intrinsic dimension 1 (= self-plot slope), but are not statistically identical (do not overlap with the cross-plot). (b) 2 data sets (shown in (c)) dominate each other at different scales. Self-plot: A-A, B-B curves. Cross-plot: A-B curve.**

The locations where a specific event occurs are collected from transcripts about that event. A transcript is considered to be about an event if the words describing that event appear in the transcript. In the context of traditional information retrieval, the set of transcripts about an event could be viewed as the result of query using the word of the event. Specifically, the process contains 3 steps:

1. Transcripts from a large video digital library [14] are examined, and if the words describing an event are mentioned in a transcript, the locations mentioned in the same transcript are associated with that event. For example, if words "Iran" and "earthquake" are both mentioned in a transcript, then Iran is associated with the event "earthquake". In our experiments, we consider only transcripts from news segments, and drop those of commercials. Several domain-specific heuristics are also used to associate locations to events (e.g., drop locations which are too common, like "Atlanta", where the headquarter of the news agency is, and "United States").

2. The collected locations are then mapped to 3-dimensional points on the surface of a sphere (globe) of radius 0.5 (according to their longitudes and latitudes).

3. GeoPlots of event pairs are then constructed based on the geodesic distances among the locations associated with the events. These GeoPlots are then examined (using the rules listed in Observation 4) to find how an event is related to another.

5. EXPERIMENTS

Event	Locations	Event	Locations
beach	84	earthquake	38
flood	131	hurricane	56
island	128	storm	215
tornado	51	volcano	19
tourism	28	wine	33

Table 1: Experimental events and the number of their associated locations

Topic 1	# Doc.	Topic 2	# Doc.	# Shared
flood	133	storm	236	29
flood	133	hurricane	49	10
earthquake	46	volcano	23	3
beach	87	hurricane	49	3
storm	236	unemployment	87	2
tourism	24	wine	21	0
tornado	42	volcano	23	0
beach	87	earthquake	46	0

Table 2: Transcript sets of the selected events

5.1 Data sets

The locations on which the events occurred are collected from a news video digital library. A 2-year long collection (from January 1, 2000 to December 31, 2001) of 33,632 transcripts are examined. We dropped segments of commercials and kept 8,499 transcripts. Table 1 listed some of the events we collected.

Table 2 shows the number of transcripts that the 2 events in an event pair share with each other. Note that the event pairs we used in our study share only a small amount of transcripts. This reduces the chances of falsely declaring two events to be similar when in fact they are not.

5.2 Textual similarity

We compare the GeoPlot to the traditional text retrieval technique to see if GeoPlot can give us interesting information which the traditional text retrieval misses.

For a pair of events, we compare their GeoPlot with the cosine similarity value calculated from the 2 sets of transcripts associated with them. GeoPlot gives information about how similar the two events (sets of locations) are distributed. On the other hand, cosine similarity has been used to determine similarity between two documents, and can be slightly modified to report similarity between two sets of transcripts.

Table 3 lists the event pairs we compared and their cosine similarity scores. In our experiment, we used the *dtb term weighting scheme* [11] to compute the similarity between two sets of documents. The dtb scheme is reported as one of the well-performed term weighting scheme that gives high mean average precision at TREC-7 [13]. The original dtb scheme is designed for comparing two documents. We slightly modified the dtb scheme to compare two sets of transcripts[1]. In the following, we call the cosine similarity score between the transcripts of event "A" and those of event "B" **the dtb**

[1]We change the *d-factor* to be the term frequency on the set of transcripts, i.e., the sum of the term frequency on transcripts in the set. The *b-factor* is also changed, where *the length of document* is replaced by *average document length in the set*. The *t-factor*, the idf factor, is not changed.

Event 1	Event 2	dtb score
flood	storm	0.509
flood	hurricane	0.433
earthquake	volcano	0.298
beach	hurricane	0.350
storm	unemployment	0.366
tourism	wine	0.187
tornado	volcano	0.166
beach	earthquake	0.243

Table 3: Cosine similarity scores

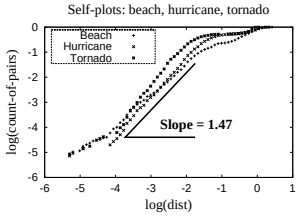


Figure 5: Self-Plots The self-plots of events "beach", "hurricane", and "tornado". The intrinsic dimension of their distributions at local scale is about 1.47.

score of "A" and "B".

Next, we present experimental results on (a) how useful are the self-plot and GeoPlot on revealing underlying relations, (b) when the analysis of the self-plot and GeoPlot agrees with the cosine similarity value, and (c) when they disagree.

5.3 Case studies

We are interested in questions such as "*do earthquakes always happen at where volcanos are?*", or "*do hurricanes and storms always cause floods?*" To find patterns giving additional information on these questions, we conduct case studies on selected events using the GeoPlots (self-plots and cross-plots). We investigate characteristics of an event from the behavior of its self-plot. The relation between two events is examined using their GeoPlot. The relations found from examining the GeoPlots are compared with the cosine similarity scores to see how the findings from GeoPlot compare to those from the cosine similarity score.

5.3.1 Self-plots analysis

Self-plots of several events have clear plateaus. These include events such as "beach", "hurricane", and "tornado". Figure 5 shows the self-plots of these three events. Detailed statistics of these self-plots are shown in Table 4.

Table 4 shows the statistics about the plateaus in the self-plots of events "beach", "hurricane" and "tornado". To make the result more understandable, we converted the distance scales into miles[2]. The plateau in a self-plot indicates the existence of clusters, and $\hat{r}_{cdmax}$ (the scale where the plateau starts) on a self-plot is the (maximal) diameter of the clusters. We know that the United States (not including Alaska and Hawaii) spans about 3,000 miles from east to west, and about 2,000 miles from north to south. From Table 4, we found that "beach" and "tornado" have clusters with diameter between 2,000 and 3,000. This suggests that the clusters of "beach" and "tornado" are probably caused by locations in the United States. We checked this result with the map of "beach" (Figure 6(c)), and found that the cluster is indeed on the United States. The map of "tornado" (not shown) has the same result.

[2]The conversion is done as the following: for $\hat{r}_{cdmax}$=-1.105, the distance in miles is $exp(-1.105) * \frac{7926}{1} = 2625.2$, where 1 is the diameter of the Earth we used in our implementation and 7926 is the true diameter of the Earth in miles.

Event	$\hat{r}_{cdmax}$	plateau width	slope of plateau
beach	-1.105	0.191	0.131
hurricane	-0.723	0.385	0.039
tornado	-1.200	0.477	0.015
Event	$\hat{r}_{cdmax}$(miles)	plateau width (miles)	
beach	2625.2	552.5	
hurricane	3846.4	804.9	
tornado	2387.3	1459.2	

Table 4: Analysis of the plateaus in the self-plots Measurements at the top are in the unit of scale, and those at the bottom are in miles. Recall that $\hat{r}_{cdmax}$ is the diameter of the cluster, which is also the scale at which the plateau starts. The width of the plateau is ($\hat{r}_{sepc}$-$\hat{r}_{cdmax}$).

As for the event "hurricane" which has clusters of diameter greater than 3,000, we checked the map of its distribution (Figure 6(d)) and found that its cluster spans beyond the United States reaches the Caribbean Islands, which explains why this cluster has diameter greater than the size of the United States (3,000 miles).

We summarize the results of the self-plot analysis in the following:

Observation 5. **Characteristic scale** In our study, the geographic distribution of an event tends to have one single cluster of the size of the United States. In our implementation, this characteristic scale ($\hat{r}_{cdmax}$) is about -1.1 (2,638 miles). This might due to the fact that our transcripts are from a US-based news agency which focuses more on news occurred in the United States.

Observation 6. **Intrinsic dimension** The intrinsic dimension of the geographic distribution of an event is the slope of its self-plot. For example, the intrinsic dimension of the three events "beach", "hurricane" and "tornado" at local scale is about 1.47 (Figure 5). In comparison, the intrinsic dimension of the points scattered uniformly in a region is 2, and that of the points scattered along a line is 1. Therefore, a dimension of 1.47 indicates a distribution not uniformed on the surface of the globe.

Observation 7. **Similarity determination** Cosine similarity score provides a mean to rank the degree of similarity. However, sometimes it is hard to choose the cutting threshold of "similar" and "dissimilar". On the other hand, visualizing the degree of similarity by displaying self-plots help tell the similar from the dissimilar quickly.

5.3.2 GeoPlot analysis

We are also interested in finding relations among events. Self-plot analysis in the previous section reveals interesting characteristics of an individual event, but for finding relations among events, we have to use the GeoPlot. We use the

GeoPlot of a pair of events to determine whether the two events have similar spatial distributions.

In the following, we will show:

1. GeoPlot successfully detects event correlations which are missed by the textual similarity function.
2. If the textual similarity function finds correlation between events, so will GeoPlot.
3. Even then, GeoPlot gives more information about how two events are correlated at different scales.

***Result** 1.* GeoPlot can find new correlations which the textual similarity function misses.

Justification 1.1 Figure 6 shows cases where GeoPlots and the textual similarity function disagree. Here, curves in the GeoPlots are overlapped. By the rules in Observation 4, this suggests relations exist between "earthquake" and "volcano", and between "beach" and "hurricane", while the dtb scores are low (suggest dissimilarity) as 0.298 and 0.350 (Table 3), respectively. However, we know that there are indeed spatial relations between earthquakes and volcanos (both are closed to geological faults, e.g., the Philippines and Italy), and hurricanes and beaches (hurricanes are formed above oceans and get high news coverage when they hit the land). In these cases, GeoPlot reveals hidden relations that the textual similarity function misses.

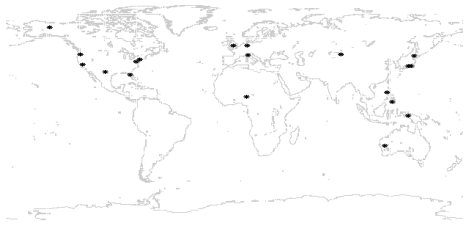
Justification 1.2 Figure 7 shows another interesting pair of events which GeoPlot finds similar: "tourism" and "wine". They are missed by the textual similarity function with dtb score 0.187. A plausible explanation is that tourists prefer mild climate (Florida, California) where grapevines naturally grow. This also demonstrates the usefulness of the GeoPlot on spotting novel, hidden relations among events.

***Result** 2.* GeoPlot seems more general than the textual similarity function and can discover relations (similarly or dissimilarly distributed) between events which are also predicted by the textual similarity function.

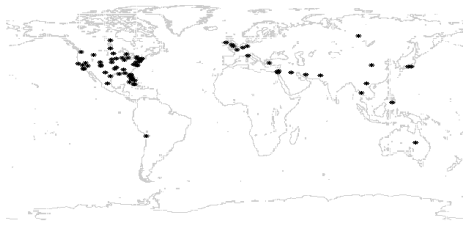
Justification 2 Figure 8(c) shows the GeoPlot of the event pair "flood" and "storm". By the rules in Observation 4, we found that the spatial distributions of "flood" and "storm" are statistically similar, since the self-plots and cross-plot in the GeoPlot overlap. The dtb score of "flood" and "storm" is 0.509 (Table 3), which is relatively higher than those of other pairs. Therefore, "flood" and "storm" are considered distributed similarly by both methods, i.e., they agree on this case. This finding suggests that *the place where a flood occurs is likely to have had a storm.*

One possible explanation for Result 2 is as follows: for textual similarity to report "similar", the two sets of transcripts about the two events must have features such as high shared-term frequency. A location mentioned in the set of transcripts about one event is then very likely to also appear in the set of transcripts about the other event. This sharing of associated locations will cause the overlapping of self-plots and cross-plot in the GeoPlot of the two events, which by our rules in Observation 4, is a sign of spatial distribution similarity. The explanation for that GeoPlot generally agrees with textual similarity function on *dissimilar* event pairs is similar to the explanation given above for the "similar" case.

(a) Earthquake

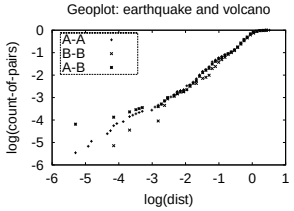


(b) Volcano

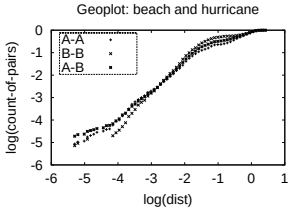


(c) Beach

(d) Hurricane



(e) A:Earthquake, B:Volcano



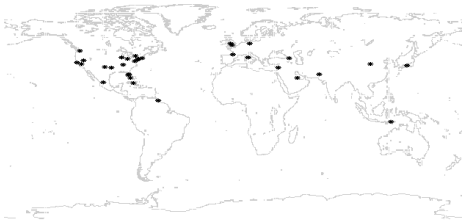
(f) A:Beach, B:Hurricane

Figure 6: GeoPlot disagrees with cosine similarity **Distributions of (a) "earthquake", (b) "volcano", (c) "beach", (d) "hurricane". GeoPlot of (e) "earthquake" and "volcano" (dtb score: 0.298), and (f) "beach" and "hurricane" (dtb score: 0.350).**

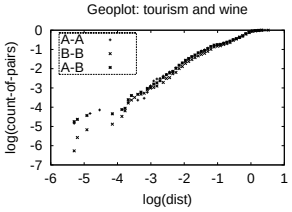
***Result** 3.* GeoPlot gives more information about how two data sets are correlated (how does the degree of correlation change as the scale changes), not just a single-valued indicator of degree of correlation as the textual similarity function gives us.

Justification 3 Figure 8(d) shows the GeoPlot of "flood" and "hurricane". The dtb score of "flood" and "hurricane"

(a) Tourism

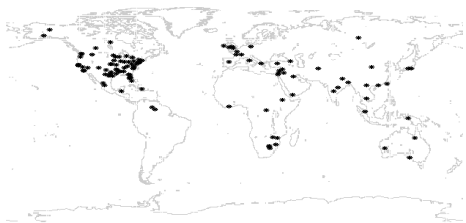


(b) Wine

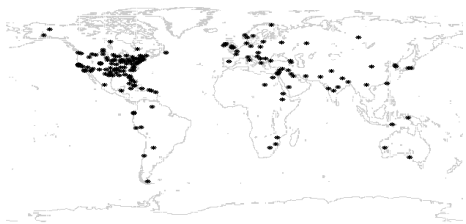


(c) A:Tourism, B:Wine

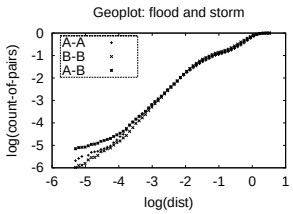
Figure 7: Interesting relation **Distributions of (a) "tourism", and (b) "wine". (c) GeoPlot of "tourism" and "wine" (dtb score: 0.187).**



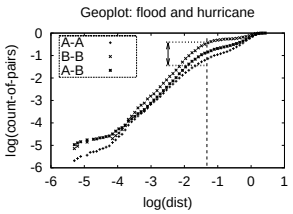
(a) Flood



(b) Storm



(c) A:Flood, B:Storm



(d) A:Flood, B:Hurricane

Figure 8: GeoPlot agrees with cosine similarity **Both methods predict the events similar. Distributions of (a) "flood", (b) "storm". GeoPlot of (c) "flood" and "storm" (dtb score: 0.509), and (d): "flood" and "hurricane" (dtb score: 0.433). Note the discrepancy among the curves in (d) at large scale (at ≈ -1.3, i.e. 2,160 miles).**

is 0.433 (Table 3), which is relatively high and indicates that they are related. Their GeoPlot gives the same implication that the two events have similar geographic distributions and they are related. Furthermore, GeoPlot also points out that this correlation only happens at local scale, specifically, $\leq 2,160$ miles, which corresponds to log(dist) $\leq$ -1.3. This is shown at the portion left to the plateaus in the GeoPlot, where the cross-plot (A-B curve) is overlapped with one of the self-plot (A-A curve). By the rules in Observation 4, this overlapping indicates the distribution of event A ("flood") includes that of event B ("hurricane") at local scale. From the self-plot analysis in the previous section, which shown that the portion left to the plateaus corresponds to the cluster located at the United States and the Caribbean Islands on the map, we concluded that this GeoPlot provides a knowledge that *floods and hurricanes in the United States and the Caribbean Islands are correlated*. We note that this is a more reasonable result than just saying "flood" and "hurricane" distributed similarly, since the correlation between "flood" and "hurricane" is only true at west Atlantic Ocean. In fact, the word "hurricane" is not even used for those happened outside west Atlantic, they are called "typhoon" in west Pacific, or "monsoon" in the Indian Ocean. Hence, analysis of GeoPlot gives us more information about the local/global characteristics of the distributions of the two events.

6. DISCUSSION

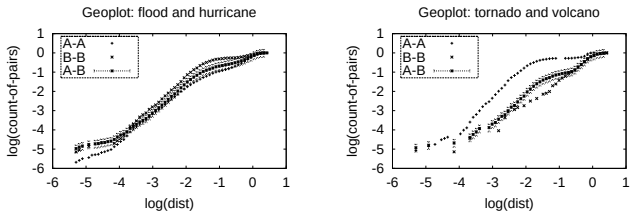
GeoPlot is an effective tool for finding hidden relations between two events. To determine the relation between the geographic distributions of the two events, we examine the closeness among the self-plot curves and the cross-plot curve in their GeoPlot.

Figure 9 shows the GeoPlots with the 20% error bar along the cross-plots (A-B curves). The 20% error bar could be used as a visual aid to determine the closeness of the curves. We consider curves are close-by when they are inside the range covered by the error bar. In the figure, the error bar shows that all curves in (a) are close-by at small scale, but not at large scale. Also, the A-A (self-plot) curve is closed to the A-B (cross-plot) curve at all scales. This result is the same as our analysis in Justification 3.

As for the GeoPlot in Figure 9(b), the three curves are far away from one another at all scales, which indicates dissimilarity of the geographic distributions of the two events: "tornado" and "volcano".

7. CONCLUSIONS

We have introduced a new tool, GeoPlot, to find spatial patterns within a single group of video clips and across two groups of video clips, where video clips are grouped according to their common topic (or called events). The idea is



(a) A:Flood, B:Hurricane (b) A:Tornado, B:Volcano

Figure 9: GeoPlot with error bar **The 20% error bar of the cross-plot is shown with the GeoPlots. GeoPlot of (a) "flood" and "hurricane" (dtb score: 0.433), and (b) "tornado" and "volcano" (dtb score: 0.166).**

to extract place names from the transcripts of video clips, put these places on the globe, and find patterns across the spatial distributions of points. We showed that GeoPlots have significant advantages over the textual (cosine) similarity functions:

- **Spot new patterns** GeoPlots can detect geographic closeness of two groups of video clips, that textual similarity function might miss. (Result 1)
- **Capture known patterns** GeoPlot does not miss relations that the textual similarity function gives. (Result 2)
- **Full information** GeoPlots give whole functions, as opposed to just single numbers. They can reveal characteristics at all scales, clusters (plateaus), and intrinsic dimensions (slopes), which are not captured in the textual similarity score. (Result 3)

We showed how to use GeoPlots to discover

- **Clusters and intrinsic dimension** Using self-plot, we can determine whether a set of points (places) is clustered (plateaus in the self-plot), or self-similar (the linearity of the self-plot). The slope of the self-plot is also the intrinsic dimension of the geographic distribution of the places. (Figure 5)
- **Event correlation** Using GeoPlot and the characterization rules (Observation 4), we could determine whether two groups of video clips have similar (or dissimilar) geographic distributions (Figure 9). Moreover, the relation between two groups of video clips can be identified as global (Figure 8(c)) or local (Figure 8(d)) features, which can help discover (clarify) the source of the correlation.

In addition, the visualization of GeoPlot is user-friendly that a viewer can quickly tell whether two sets of points are similarly distributed (at local scale or global scale) or not.

Moreover, GeoPlots can be computed quickly in linear time $O(N)$ on the number of points N, using the so-called "box-counting plots" from the fractal theory [2] [12].

This is the first step towards a new tool for video data mining. Future work could include the time (date) dimension of the news events and explore the evolving patterns of the news events.

8. REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. *Proc. of ACM SIGMOD*, May 1993.

[2] A. Belussi and C. Faloutsos. Estimating the selectivity of spatial queries using the 'correlation' fractal dimension. *Proc. of the VLDB Conference*, 1995.

[3] N. A. C. Cressie. *Statistics for Spatial Data*. Wiley, 1991.

[4] P. M. Dixon. Ripley's k function. Technical report 01-18, Department of Statistics, Iowa State University, December 2001.

[5] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. *Proc. of SIGCOMM*, 1999.

[6] C. T. Jr., A. Traina, L. Wu, and C. Faloutsos. Fast feature selection using the fractal dimension. *XV Brazilian Symposium on Databases (SBBD)*, 2000.

[7] K. Koperski, J. Adhikary, and J. Han. Spatial data mining: Progress and challenges. *Proc. of SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 1996.

[8] W. Lu, J. Han, and B. C. Ooi. Discovery of general knowledge in large spatial databases. *Proc. of Fareast Workshop on Geographic Information Systems*, 1993.

[9] T. M. Mitchell. *Machine Learning*. WCB/McGraw-Hill, 1997.

[10] M. Schroeder. *Fractals, Chaos, Power Laws: Minutes From an Infinite Paradise*. W. H. Freeman and Company, 1991.

[11] A. Singal, C. Buckley, and M. Mitra. Pivoted document length normalization. *Proc. of the Nineteenth ACM SIGIR Conference*, August 1996.

[12] A. Traina, C. Traina, S. Papadimitriou, and C. Faloutsos. Tri-plots: Scalable tools for multidimensional data mining. *Proc. of ACM KDD*, August 2001.

[13] E. M. Voorhees and D. K. Harman. Overview of the seventh Text REtrieval Conference (TREC-7). *Proc. of the Seventh Text REtrieval Conference (TREC-7)*, 1999.

[14] H. Wactlar, M. Christel, Y. Gong, and A. Hauptmann. Lessons learned from the creation and deployment of a terabyte digital video library. *IEEE Computer*, 32(2):66–73, February 1999.

[15] G. K. Zipf. *Human Behavior and Principle of Least Effort: an Introduction to Human Ecology*. Addison Wesley, 1949.