Knowledge Discovery in an Earthquake Text Database: Correlation between Significant Earthquakes and the Time of Day

TO BE PRESENTED AT THE 9TH INTERNATIONAL CONFERENCE ON SCIENTIFIC AND STATISTICAL DATABASE MANAGEMENT

Jeffrey A. Goldman, D. Stott Parker, and Wesley W. Chu
Computer Science Department
University of California
3436 Boelter Hall
Los Angeles, California 90095-1596
goldman@ucla.edu*

Abstract

In this paper, we take a real world application from a text database and present a case history. The techniques ultimately led to a discovery contradicting an accepted paradigm in seismology. Using simple, tailored, keyword extraction, we examined a text collection of earthquake data. A discovery was made when an unusual pattern emerged from the text. We then tested a more comprehensive numerical database, treating the the text discovery as a hypothesis. It was verified using a standard χ^2 *statistic. The hypothesis was significant earthquakes in the longitude regions that include California, occur more often in the morning hours than any other time of day.*

1. Introduction

We are just beginning to see the emergence of text-based knowledge discovery systems [6, 12, 14] while the Information Retrieval (IR) community has been tackling the same routing and querying issues for more than 30 years [7, 23, 22]. The best retrieval and routing techniques do a respectable job at ranking relevant documents to a given query [7]. There are a variety of techniques that perform at a similar level. The majority involve some type of threshold function based on a truncated keyword frequency analysis (vector space methods and probabilistic IR). In other words, a body of text is analyzed by its constituent words. Various techniques are used to build the core words. The ultimate goal of routing is finding documents that are similar, based on document feedback from the user. For retrieval, the goal is finding the right index terms in a collection so querying will return the appropriate documents.

Knowledge discovery and data mining invert this task. We now want to extract the similarities of the documents in our collection automatically [16, 17]. In this paper, we look at using the relative frequency of words in a small text database of earthquake information. Using these statistics, we extract all the knowledge that we can from a keyword breakdown approach. We then attach significance to the discovery by rigorously testing it in a much larger database. A χ^2 test confirms that our hypothesis is not the result of random observations.

2. Text verses Records

Text collections lack the imposed fields of a traditional database, so an extra step is needed in order to transform this information into a vector-like entry. In the IR community, this task is accomplished by creating a dictionary of words that are present in the entire collection of texts. Usually, the words are truncated, a process called "stemming", to allow words with the same root but different suffixes, to be counted together. However, it is unclear that this process degrades or enhances document matching [3]. In addition, the preprocessing also removes "stop words" such as: *and, the,* and *or*. Again, it is not clear that removing stop words helps or harms the matching process [19]. At the end of processing, a document is now a vector of normalized frequencies of words in the global dictionary. This normalization insures against longer documents from matching with everything. It also prevents shorter documents from matching with nothing. The goal is to count the documents equally in

*http://www.cs.ucla.edu/~ goldman

the collection. In addition, words are also weighted as to importance by various formulas. We now have a collection of vectors that we can measure and compare.

We take some comfort in the discovery task for databases that already have imposed fields. In a sense, our discoveries are limited by the database design [16]. However, usually the designer had some coherent idea in mind when the database was built. Thus, for record entry databases, the miner can take full advantage. In general, text collections are unstructured as a group but very structured individually. The discovery task requires building a "database" before we can search for patterns. However, we are now forced to make some design decision without the understanding of the collection.

As previously mentioned, the IR community is generally agreed that this task is accomplished by breaking up each document into its constituent words [4, 20, 26, 27, 30]. Very few models incorporate other information in retrieval. [2] allows for a query to utilize the table of contents in a meaningful way but not all collections of text have such a table. A parallel approach for retrieval tries to summarize text with phrases [24] to make matching more manageable. However, this research focus is about summarizing, not building similarities among several documents.

3. Experimental Tests

In this section we describe the setup and results of a discovery task for collections of text. Our goal is the eventual discovery of a set of usefully related terms. The text collection describes earthquake data. The data for this experiment was obtained from the U. S. Geological Survey. Each document is a short (~25K) text report of seismic activity during a given week for the planet. Specifically, the text breaks down information into the "San Francisco Bay Area," "Northern and Central California," "Long Valley Caldera," "United States," and "Planet Earth." The document also includes a table of dates, times, locations, and other numerical information about the individual quakes. There is an additional table of times, latitude/longitude, magnitude, and location name (i.e. the text name of the location). Each of these documents spans activity for seven days and does not include every measurable quake. Usually a magnitude of two or more is required for the California areas and four or more for the planet. We analyzed a collection of text spanning the entire 1995 calendar year. There are 49 texts (some holidays were taken and there were no reports for that particular week) of earthquake data. The total amount of data was approximately 1.34 megabytes.

One will note that this collection is basically uniform. There is little variation in document structure from one week to the next except for the locations and magnitudes of different earthquakes and the corresponding descriptive text.

Word	Frequency
GEY	1471
MOR	1136
ISLANDS	1010
ML	856
M2	848

Table 1. Most Frequent Words.

Also, even though there was significant numeric information in the table, the purpose of our experiment was to explore discovery among text information. Thus, only the text of the quake location was used.

Despite the recent objections, most of the IR community still uses stemming and the removal of stop words. We decided to do so also. We recorded terms as the same, regardless of case or different suffix: "CALIFORNIA" would be counted the same as "California" and "ISLAND" would count the same as "islands." Since the documents were approximately the same length, they were all counted equally. Stop words were also eliminated. In addition, words that were irrelevant but appeared specifically in this collection were ignored to yield the best possible word list. Thus, in this experiment we examine the core words that appear within some threshold in our collection. This is essentially how index words are chosen for routing and retrieval.

4. Results & Refinement of Word Extraction

For the first set of experiments, there were many irrelevant words and word fragments. Either they appeared in all documents because of their structure or there were common word abbreviations for this collection. It is important to note that even the most sophisticated document matching techniques based on the frequency information would not eliminate these words unless it was specifically tailored to do so (with term weighting, for example).

With the exception of "ISLANDS," which we will explore later, the top 5 words listed in Table 1 are irrelevant. For example, "GEY" and "MOR" are region codes to denote information about the velocity model used to locate the event. Since each quake is listed in the table, these codes appear for most of the entries. The only knowledge one could hope to discover from this frequency is which method is used the most. "ML" is used when for some reason, the magnitude of the quake is not listed explicitly in the table. "M2" is short for a magnitude 2.0 quake. These are abundant since they are often the only quakes that happen in a particular region. The next 12 words (not listed) are the abbreviations for the months of the year. Examining quakes for the entire year, one would expect to see the month abbreviations, es-

Word	Frequency
ISLAND, ISLANDS, island, or islands	1213
CALIFORNIA, California, or CALIF	845
COAST, or coast	740
CHILE, or Chile	587
San	417

Table 2. Most Frequent Location Words.

pecially when every quake in the table has a date[1]. Thus, simply using the most popular words is insufficient for discovery.

There were many other irrelevant words such as: "occurred," "miles," "area," and "earthquake(s)." Table 2 shows the frequency of the more useful words. We found "useful" to be words describing location. Word stems were used so that various forms of the same word were counted together. However, even when we further tailor our keyword frequency method, we still have problems. For example, the word "CALIFORNIA" always indicates an earthquake but "*California*" is used in the section heading in the texts to talk about Southern *California*, Northern and Central *California*, the Office of Emergency Services of the State of *California*, and others. Thus, this number is too high to represent earthquake activity in California as opposed to other regions. The same holds for the word fragment "San." Whether it is San Francisco, San Juan, San Simeon, or San Andreas, the word is mostly used to describe a heading. Therefore, further refining is needed.

In Table 3 we can finally draw some conclusions. From the data, our "discovered knowledge" is:

- The regions listed in Table 3 are places where the most earthquakes have occurred for 1995.
- Islands appear to be a particularly good hotbed for activity.
- Coastal targets are also frequent earthquake sites.
- Large bodies of water are prime targets although some regions (Chile, California, Argentina, and Alaska) were hit more than all of the water regions combined.

[1]This information was discovered after taking a close look at a few (sometimes only 1) of the individual documents by hand in order to further refine the keyword methods for our task.

Word	Frequency
ISLAND, ISLANDS, island, or islands	1213
COAST, or coast	740
CHILE, or Chile	587
CALIFORNIA, or CALIF	500
ARGENTINA, or Argentina	350
ALASKA, Alaska, or Alaskan	342
SEA, Sea, sea OCEAN, Ocean, ocean(s), oceanic, water(s), Lake(s), or lake	249
KURIL, Kuril, or Kurils	224
JAPAN, or Japan	224
NEVADA, or Nevada	156
HONSHU, or Honshu	145
FIJI, or Fiji	132
INDONESIA, or Indonesia	121
MEXICO, or Mexico	119
GREECE, or Greece	101

Table 3. Refined Location Words.

Word	Frequency
morning	165
evening	74
afternoon	62
night	21

Table 4. Earthquake Times.

Unfortunately, not all of these conclusions are solid. The text often uses the word "Coast" as a reference point for phrases such as: 50 miles off the coast of, near the coast of, etc. Certainly all of the proper places listed are known to be regions of active fault lines. Thus, this data merely reconfirms the obvious (for a seismologist). Moreover, the fact that most of the activity was in the water is owed to several fault lines located there and the majority of the planet being covered by water. However, the four regions that had more occurrences than all of these water areas are on the most active fault lines for 1995. In 1995, Argentina, Alaska, and Chile were particularly active. The data about "Island" is almost always used to describe a proper location. Statements describe either the island as a location of the quake or it was felt by inhabitants of the island. We can safely conclude that the earthquakes in 1995 were particularly partial to the islands of the world. Seismologists, however, would hardly consider this knowledge a discovery.

In addition to the location terms, there were a few frequent words that were not stop words, document specific abbreviations, or irrelevant terms. They were "morning" and "creeping." "Creeping," turned out to be a reference to a particular segment of the San Andreas, i.e. "on the creeping segment of the San Andreas." Thus, it was used as a reference to whenever activity occurred on this segment. "Morning," however, was another matter. In the text, the times of the earthquakes were given locally. Thus, when it referred to early morning or afternoon, this was the time for that region. We looked at the entire lexicon word list by hand, (the vocabulary for these documents) for words associated with time. The days of the week posed a problem because the text would say: "for the seven-day period ending Wednesday..." Furthermore, one would not expect to see earthquakes on one particular day over another[2]. The word "day" was also a problem for the same reason. In addition, "day" appeared in phrases such as: the same *day*, the next *day*, a *day* later. Instead, the word "afternoon," described day activity. Other words were studied and removed for similar reasons. The words that we were reasonably sure described a particular time of day were: morning, afternoon, evening, and night (Table 4).

[2]Even accounting for this phrase, there was no pattern.

5. Discussion of Earthquake Times

We can draw two conclusions from the data. Since the majority of the text discusses regions of California, it is likely that if there is a pattern, it may only be relevant for the California region. One possibility is that earthquakes occur more often in the morning than any other time of day[3]. The other possibility is that the writers may simply have referred to "morning" for a greater period of the day than the other words. Morning may mean after midnight until 11:59am. If this is true, than morning is about as frequent as the rest of the terms and there is no meaningful pattern.

Considering the data as a suggestion of the hypothesis: "Earthquakes occur more frequently in the morning than any other time of day," we researched the claim. One text supported the theory and stated earthquakes occur most frequently between the hours of 4:00am and 8:00am in California [9] but did not back up the claim statistically. Instead, it was presented as an observation for earthquakes of magnitude 6 or larger since the 1900's. There have been several studies linking planetary motion and earthquake triggering [10, 13, 21, 25]. Most have focused on tidal forces and the interaction of the moon and earth. In a few cases, there have been minor correlations of tidal forces and the triggering of earthquakes [5, 15, 28, 29]. However, the general consensus of geologists is apparently that tidal forces play at best a small role [8, 10, 21]. Only [10] looked at the Sun's position for California. This work showed for Northern California, some non-uniform behavior was apparent for earthquakes larger than magnitude 6.0 over the day, when it was split into two equal parts centered around 6am/6pm and noon/midnight local time. One paper actually examined the local time of day for a specific region in San Diego county [1]. However, the purpose of the work was to allow seismologists to distinguish between small chemical explosions and actual seismological events. This work demonstrated how explosions of magnitude 3.6 or smaller can skew real events. Specifically, the small "earthquakes" had a significant peak time at noon and 5pm local Pacific time which coincided with blasting times for construction. Moreover, the Caltech seismology group stands by their statement that "big earthquakes occur in the morning is a common myth[4]."

In spite of suggestions of earthquake myths or man made blasting as the cause of all earthquake patterns associated with the time of day, the most promising support of our hypothesis came from Charles Richter himself. In [18], Richter states:

> Predictions based on positions of the sun and moon have to be regarded a trifle more seriously,

[3]For this data, each document stated that when it describes a time of an earthquake, it is local time. The tables on the other hand were all in Greenwich Mean Time (GMT).

[4]http://www.scec.gps.caltech.edu/eqmyth.html

since there is evidence that tidal forces may occasionally act as triggers for earthquakes otherwise on the point of taking place; in this way the dates and hours of occurrence may show a slight statistical correlation with the tides.

We could not find any current work on correlating tidal forces with earthquakes. Moreover, tides are on a 28 day cycle which would not necessarily correspond to the actual local time of day over a one year period. We decided to test our claim over a much larger dataset to determine if we had indeed discovered a pattern or if the unusual word imbalance for "morning" was owed to some other factor.

6. Testing the Hypothesis on a Larger Data Set

6.1. The Database

The database for the tests in this section is from the Northern California Earthquake Data Center (NCEDC). On the web[5], any user can set a variety of parameters and retrieve all of the earthquake information that exists in the database which was compiled by members of the Council of the National Seismic System (CNSS) in the CNSS Worldwide Earthquake Catalog. The database goes as far back as 1910 and contains earthquake magnitude, depth, time, location, and depth for the entire planet. Some of the older records are incomplete but all of them have the time and location. Like all scientific data, the entries are occasionally flawed and improve in quality over time (in this case because increasing numbers of higher quality seismic sensors were installed).

6.2. The Results

We first examined every earthquake between $-127.5°$ longitude and $-112.5°$ longitude from 1910 to 1996 inclusive. This corresponded to every recorded earthquake of magnitude larger than 0.0 for the slice of the earth including California, a significant portion of Canada, Alaska, and Mexico (hereafter, we will refer to this slice as the California region). We then created a frequency plot of the number of earthquakes by the hour of the day seen in Figure 1. The times in the database were recorded in Greenwich Mean Time (GMT) which is 7 or 8 hours ahead of the local Pacific time, depending on whether it is Daylight Savings Time or not. Figure 1 seems to suggest that the majority of earthquakes in the California region occur around noon and 3pm local time. The morning hours do not appear to have nearly as many earthquakes.

[5] http://quake.geo.berkeley.edu/cnss/catalog-search.html

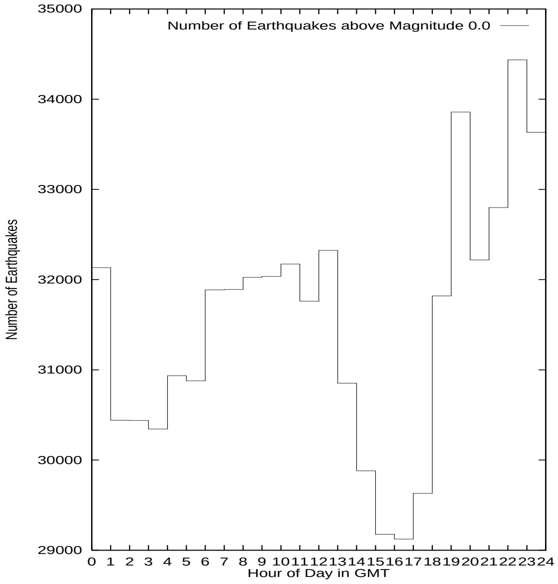

Figure 1. Frequency of Earthquakes in the California Region from 1910-1996.

After a closer examination of some of the original text data, we refined our hypothesis. It appeared that only earthquakes of notable size were mentioned in the text. This made sense because there are so many small earthquakes that are not even felt in California. Moreover, it is possible that the data from the CNSS often records events that are really a result of explosions. As it turned out, the data in Figure 1 was gathered just prior to a refinement of the database allowing the user to separate blasting from natural events. We decided to look at earthquakes for the same region and time period, however, only if their magnitude was at least 4.0 and only if it was not categorized as a blasting event. Discussions with seismologists confirmed that, for Californian seismic data, this should eliminate most, if not all man-made explosions or other non-seismic events. We further felt that earthquakes of this magnitude were probably significant enough on a weekly basis to be mentioned in the text data. When the frequency was plotted, the results were dramatically different. In Figure 2, we see distinct peaks between 14:00 and 16:00 GMT which corresponds to between 6:00am and 8:00am local Pacific Time. We had our first independent corroboration of the hypothesis. In these particular plots, we grouped together the magnitude ≥ 4.0 earthquakes into two hour blocks by even numbered hours ($0-2$, $2-4$, etc.) and odd numbered hours ($1-3$, $3-5$, etc.) to take into account the variation over the solar year, i.e. the position of the sun in the sky over the course of the year, as well as to enhance the visualization of the patterns.

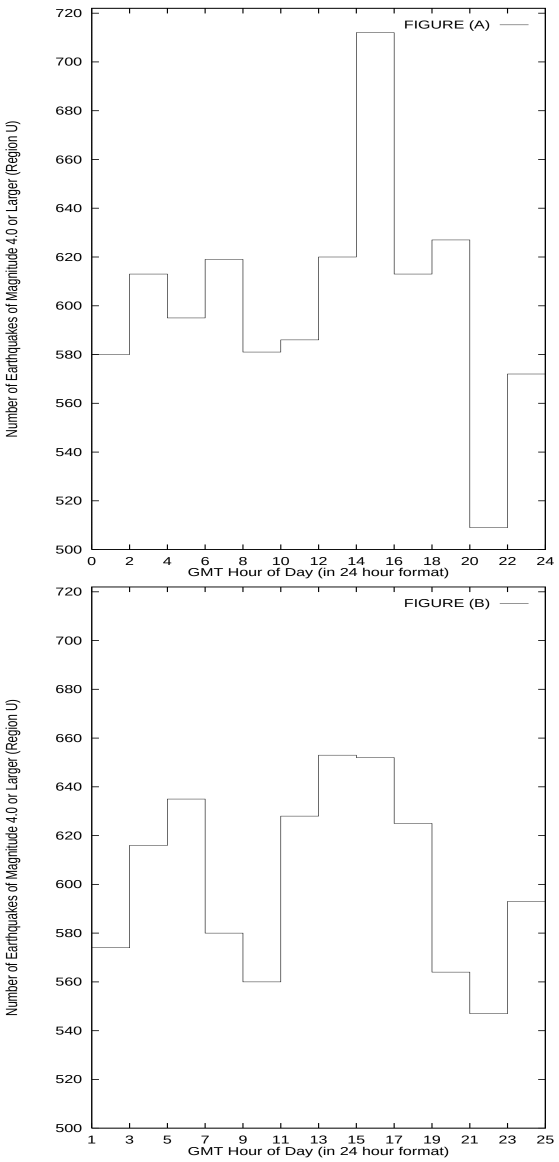

Figure 2. Frequency of Earthquakes in the California Region of magnitude $\geq$4 from 1910-1996. Figure (A) groups the frequencies into two hour blocks on the even numbered hours. Figure (B) does the same for odd numbered hours.

6.3. Statistical Confirmation

Seismologists have long held that significant earthquakes have as likely a chance occurring in the middle of the night as they do in the middle of the day. For our region of California however, our frequency plots would suggest otherwise. To be more rigorous, we performed a χ^2 test to see if our data is significantly different than randomly occurring earthquakes throughout the day.

In our data, there were 7,641 earthquakes of magnitude 4.0 or larger occurring in the region $-127.5°$ to $-112.5°$ longitude from 1910 to 1996 inclusive. Testing against a uniformly distributed hypothesis, we used 24 possible time slots (0-23). The χ^2 statistic (of [11]) for n observations and k degrees of freedom is:

$$V = \sum_{s=1}^{k} \frac{(Y_s - np_s)^2}{np_s} \tag{1}$$

where each Y_s is the number of earthquakes from (and including) GMT hour s to GMT hour $s{+}1$ and p_s is the expected value for a given hour of the day. For the data in Figure 2, the value for V is 57.07. The value of the χ^2 distribution for 23 degrees of freedom for $p = 99\%$, is 41.64. Thus, we would expect V to be larger than 41.64 less than 1% of the time if earthquakes occurred randomly during the day. Instead, the statistics suggest very strongly that the time of day and the number of earthquakes are strongly correlated.

6.4. Reinforcing Earthquake Myths

To quote the Caltech earthquake myth web page directly[6]:

> MYTH #1
>
> "Big earthquakes always happen in the early morning."
>
> This myth may be so common because we want it to be true. Several recent damaging earthquakes have been in the early morning, so many people believe that all big earthquakes happen then. In fact, earthquakes occur at all times of day. The 1933 Long Beach earthquake was at 5:54 p.m. and the 1940 Imperial Valley event was at 9:36 p.m. Even recently, the 1990 Upland earthquake was at 3:43 p.m. and the 1989 Loma Prieta event was at 5:02 p.m. It is easy to notice the earthquakes that fit the pattern and forget the ones that don't.

We decided to look at "big" earthquakes to see if there was a correlation there as well. We took "big" to mean earthquakes of magnitude 5.5 or larger. If we were to look at

[6] http://www.scec.gps.caltech.edu/eqmyth.html

only earthquakes of magnitude 6.0 or larger, there were less than 130 earthquakes in the California region from 1910-1996. There was some concern that the results would not be relevant due to an insufficient number of samples. The frequency plots for earthquakes, magnitude 5.5 or larger are shown in Figure 3. There were 352 earthquakes and the χ^2 value, $V = 39.23$. Once again, we are presented with strong evidence that earthquakes in the California region do not occur randomly during the day but in fact occur in the morning.

6.5. A Closer Look at Other Time Zones

The text data had more information about California than any other region. However, the results for California were so compelling, we examined all the other regions of the world as well. Breaking up the earth into 24 equal "pie wedges," corresponding to the GMT time zones (See Figure 4[7]), we calculated the total number of earthquakes of magnitude 4 or larger by hour as before. Some of the regions had even higher values for V than California. Table 5 shows the region, the value of V and the approximate $1-p$ value[8]. This table shows how unlikely the outcome was if attributed to randomness. The regions for which the $1-p-$value entries are not labeled, have what we would expect if the earthquakes occurred at random times. However, the majority of the regions do not appear to be random. In fact, when all of the data is taken together (see Figure 5), the value for V suggests anything but an equally distributed number of earthquakes throughout the day.

Expert seismologists, when shown some of our results, raised concerns about the sensitivity of seismic detection over the world, particularly for earlier recorded earthquakes. Specifically, because detection can be better at night (when there is less ambient noise), artificial patterns can emerge in some regions. However this basic concern is diminished by our focusing on earthquakes of high magnitude. Another common concern deals with the analysis of aftershocks. We did a careful inventory of aftershocks, and found they do not affect the statistics here significantly. Generally we have studied such concerns very carefully and do not believe that they affect the validity of our results. Nevertheless, interpretation of our results and categorical dismissal of such concerns are matters for further work.

7. Conclusion

In this experience, the keyword frequency information alone, was too weak for data mining tasks. However, it provided what turned out to be an outstanding, validated hypothesis. The efforts to obtain the keyword frequency infor-

[7] This map is presented here courtesy of the Royal Observatory of Hong Kong.

[8] Table 6 lists the χ^2 values for easy comparison.

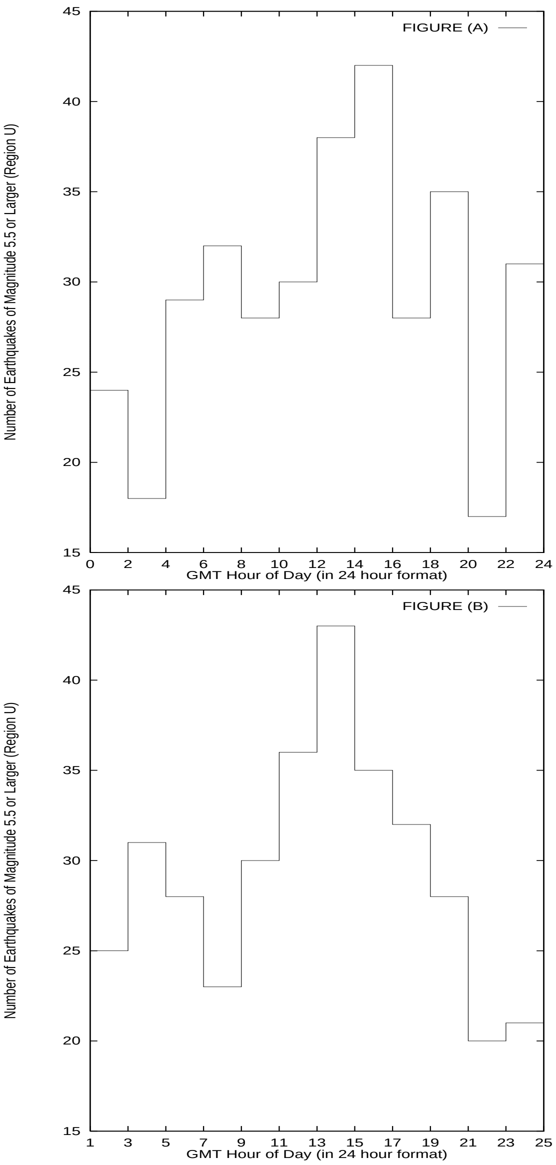

Figure 3. Frequency of Earthquakes in the California Region of magnitude $\geq$5.5 from 1910-1996. Figure (A) groups the frequencies into two hour blocks on the even numbered hours. Figure (B) does the same for odd numbered hours.

22 WORLD MAP OF TIME ZONES

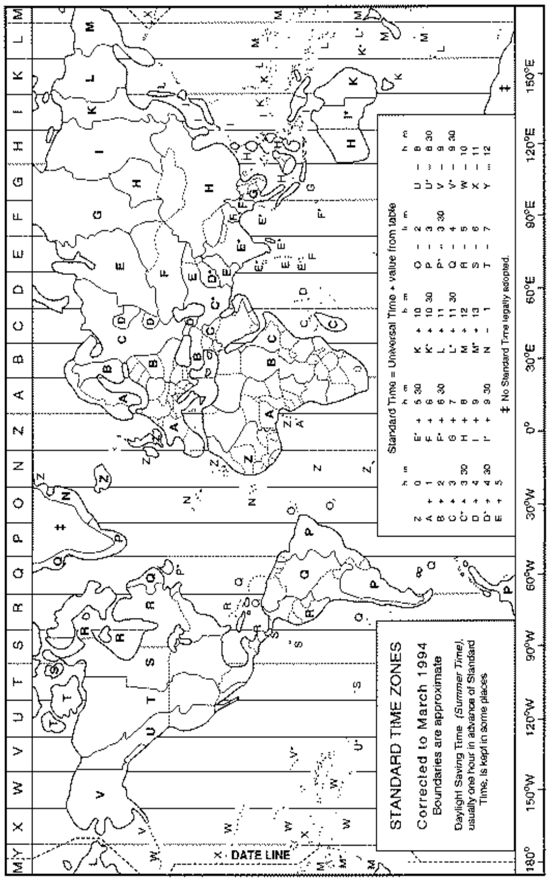

Figure 4. Timezone Map of the Earth.

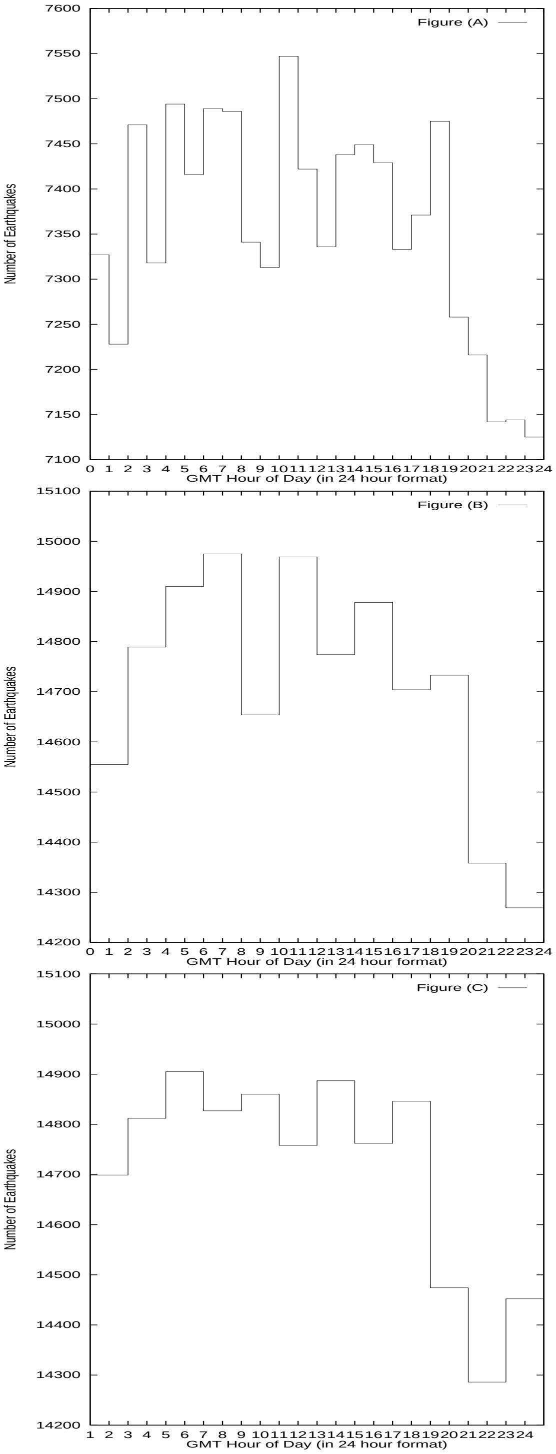

Figure 5. Frequency of Earthquakes on the Planet of magnitude ≥ 4 from 1910-1996. Figure (A) groups the frequencies by hour. Figure (B) groups the frequencies into two hour blocks on the even numbered hours. Figure (C) does the same for odd numbered hours.

Min.	Max.	Time	V value	Approximate
Longitude		Zone		$(1-p)$-value
-172.5	-157.5	GMT-11	26.24	–
-157.5	-142.5	GMT-10	22.68	–
-142.5	-127.5	GMT-9	27.15	–
-127.5	-112.5	GMT-8	57.07	$> 99\%$
-112.5	-97.5	GMT-7	41.81	$> 99\%$
-97.5	-82.5	GMT-6	71.95	$> 99\%$
-82.5	-67.5	GMT-5	91.27	$> 99\%$
-67.5	-52.5	GMT-4	31.02	–
-52.5	-37.5	GMT-3	43.69	$> 99\%$
-37.5	-22.5	GMT-2	51.80	$> 99\%$
-22.5	-7.5	GMT-1	34.95	90%
-7.5	7.5	GMT	15.93	–
7.5	22.5	GMT+1	20.21	–
22.5	37.5	GMT+2	32.26	90%
37.5	52.5	GMT+3	32.76	90%
52.5	67.5	GMT+4	38.22	97.5%
67.5	82.5	GMT+5	28.17	–
82.5	97.5	GMT+6	26.97	–
97.5	112.5	GMT+7	30.76	–
112.5	127.5	GMT+8	60.21	$> 99\%$
127.5	142.5	GMT+9	24.07	–
142.5	157.5	GMT+10	27.57	–
157.5	172.5	GMT+11	58.18	$> 99\%$
172.5 -180	180 -172.5	GMT+12 (GMT-12)	48.48	$> 99\%$
Entire Planet			46.40	$> 99\%$

Table 5. Regions and their Statistical Significance.

mation were minimal. With a pattern as strong as the one we have seen, only a small amount of text was needed to capture it. We then took care in the word frequency counting process to ensure we removed the meaningless terms and did not overquantify others. This type of information can be provided by a domain expert.

Once a meaningful collection of words was observed, delving more deeply required human intervention. Although the CNSS system was invaluable in verifying our claims, considerable effort was required for our data extraction. What was really needed is the ability to quickly retrieve frequency information only and then be able to test against a randomly distributed assumption as we have shown in this paper. This seems especially necessary in any type of time series type database. Without this feature, individual records had to be gathered for each region, parsing was needed to extract the frequency by hour, and finally the frequencies were used to calculate the χ^2 statistic.

There have been several instances when other seismologists claimed to find a pattern for particular regions with frequency of events and the time of day. Most of the time, there have been major flaws such as: insufficient sampling, dependent data (aftershocks), unknowing inclusion of man-made events, and dependency on the measuring equipment. Further refinements and tests of our work is the subject of another paper. However, in discussions with the experts in the field, the California regional observations are more than likely real. Moreover, the region of the Kuril Islands and New Guinea (GMT+10) may be a new finding.

We can only speculate as to why strong earthquakes occur in the morning hours for the California region; for that matter, why there is an apparent non-random frequency of earthquakes in particular times around the globe. Seismologists have never attached any weight to the occurrence of earthquakes at any particular time of the day. This work suggests that the correlation between earthquakes and the time of day requires further analysis. It also suggests that the practical approach we have taken can work well for knowledge discovery in text databases.

Appendix

$(1-r)$	$\chi^2_r(23)$ 23 degrees of freedom
0.010	10.20
0.025	11.69
0.050	13.09
0.100	14.85
0.900	32.01
0.950	35.17
0.975	38.08
0.990	41.64
0.995	~44.2
0.999	~49.9
0.9999	~57.2

Table 6. Chi-Square Values; The observed value, "V," will be larger than the value in Column 2, 100 − Column 1 percent of the time if the data occurs randomly over the time periods. For example, with 23 degrees of freedom, "V" should be larger than 41.64 only 1% of the time. Moreover, "V" should be smaller than 13.09 only 5% of the time.

Acknowledgments

Jeffrey Goldman would like to thank his wife Melissa for suggesting we look at larger earthquakes.

References

[1] D. C. Agnew. The use of time-of-day seismicity maps for earthquake/explosion discrimination by local networks, with an application to the seismicity of San Diego county. *Bulletin of the Seismological Society of America*, 80(3):747–750, June 1990.

[2] R. Baeza-Yates and G. Navarro. Integrating contents and structure in text retrieval. In *The Third Text Retrieval Conference (TREC-3)*, 1994.

[3] K. W. Church. One term or two? *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 310–318, 1995.

[4] W. S. Cooper and A. Chen. Experiments in the probabilistic retrieval of full text documents. In *The Third Text Retrieval Conference (TREC-3)*, 1994.

[5] D. D. Dionysiou, E. N. Sarris, and P. G. Tsikouras. Newtonian and post-Newtonian tidal theory: Variable G and earthquakes. *Earth, Moon, and Planets*, 60(2):127–140, 1993.

[6] R. Feldman and I. Dagan. Knowledge discovery in textual databases (KDT). *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, 1995.

[7] D. Harman. Overview of the fourth text retrieval conference (TREC-4). *The Fourth Text Retrieval Conference (TREC-4)*, 1995.

[8] S. Hartzell and T. Heaton. The fortnightly tide and the tidal triggering of earthquakes. *Bulletin of the Seismological Society of America*, 79(4):1282–1286, 1989.

[9] R. L. Iacopi. *Earthquake Country: How, Why, and Where earthquakes strike in California*. Fisher Publications, 1996.

[10] S. Kilston and L. Knopoff. Lunar-solar periodicities of large earthquakes in southern California. *Nature*, 304:21–25, July 1983.

[11] D. E. Knuth. *The Art of Computer Programming, 2: SeminumericalAlgorithms*. Addison-Wesley, Reading, Massachusetts, second edition, 1981.

[12] K. Lagus, T. Honkela, S. Kaski, and T. Kohonen. Self-organizing maps of document collections: A new approach to interactive exploration. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996.

[13] R. M. C. Lopes, S. R. C. Malin, A. Mazzarella, and A. Palumbo. Lunar and solar triggering of earthquakes. *Physics of the Earth and Planetary Interiors*, 59:127–129, 1990.

[14] K. J. Lynch and H. Chen. Knowledge discovery from historical data: An algorithmic approach. Technical report, MIS Department of Mathematics, University of Arizona, 1994.

[15] A. Palumbo. Lunar and solar tidal components in the occurrence of earthquakes in Italy. *Geophysics Journal of the Royal Astronomical Society*, 84:93–99, 1986.

[16] G. Piatetsky-Shapiro and W. J. Frawley. *Knowledge Discovery in Databases*. AAAI Press, 1991.

[17] G. Piatetsky-Shapiro, C. Matheus, P. Smyth, and R. Uthurusamy. KDD-93: Progress and challenges in knowledge discovery in databases. *Artificial Intelligence Magazine*, 15, No. 3:77–82, Fall 1994.

[18] C. F. Richter. *Elementary Seismology*. W. H. Freeman and Company, San Francisco, 1958.

[19] E. Riloff. Little words can make a big difference for text classification. *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 130–136, 1995.

[20] S. E. Robertson, S. Walker, M. M. Beaulieu, M. Gatford, and A. Payne. Okapi at TREC-4. In *The Fourth Text Retrieval Conference (TREC-4)*, 1995.

[21] P. A. Rydelek, I. S. Sacks, and R. Scarpa. On tidal triggering of earthquakes at Campi Flegrei, Italy. *Geophys. Journal Int.*, 109(1):125–137, 1992.

[22] G. Salton. *The SMART retrieval system: Experiments in Automatic Document Processing*. Prentice Hall, 1971.

[23] G. Salton and M. Lesk. Computer evaluation of indexing and text processing. *Journal of the Association for Computing Machinery*, 15:1:8–36, 1968.

[24] G. Salton, A. Singhal, C. Buckley, and M. Mitra. Automatic text decomposition using text segments and text themes. *Proceedings of the Dual-use Technologies and Applications Conference*, 1995.

[25] M. Shimshoni. Evidence for higher seismic activity during the night. *Geophysical Journal of the Royal Astronomical Society*, 24:97–99, 1971.

[26] A. Singhal, C. Buckley, M. Mitra, and G. Salton. Document length normalization. Final report, TR-95-1529, Cornell University, 1995.

[27] A. Singhal, C. Buckley, M. Mitra, and G. Salton. Pivoted document length normalization. Final report, TR-95-1560, Cornell University, 1995.

[28] G. P. Tamrazyan. Time origination of earthquakes in the Kuril-adjoint-to-Kamchatka region and the local time (lunar, solar). *Journal of Physics of the Earth*, 14(2):41–48, 1966.

[29] H. Tsuroka, M. Ohtake, and H. Sato. Statistical test of the tidal triggering of earthquakes: contribution of the ocean tide loading effect. *Geophysical Journal Int.*, 122:183–194, 1995.

[30] Z. W. Wang, S. K. M. Wong, and Y. Y. Yao. An analysis of vector space models based on computational geometry. In *SIGIR 92*, pages 152–160, 1992.