Term Weighting Approaches for Mining Significant Locations from Personal Location Logs

Zhengwei Qiu, Cathal Gurrin, Aiden R. Doherty, Alan F. Smeaton
CLARITY: Centre for Sensor Web Technologies, Dublin City University, Glasnevin,
Dublin 9, Ireland
{zqui, cgurrin, adoherty, asmeaton}@computing.dcu.ie

Abstract

In this paper, we describe experiments into the application of term weighting techniques from text retrieval to support the automatic identification of significant locations from a large location log, which we consider to be important for supporting many location-based social network applications. We identify the fact that the distribution of locations follows a similar shaped distribution to that of terms in a language and in so doing motivate our use of term weighting techniques. Using this information we then show that these proven techniques can be used to automatically identify social visits and "pass through" locations, as well as standard home and work locations. We also suggest that it is possible to classify whether an extended segment of personal location data may be a tourist trip, business trip or a typical working (at home) period of time.

Keywords : Location; Power-law distribution; GPS; important locations; text retrieval.

I. INTRODUCTION

These days, location-based services are an important element for supporting information seeking activities. Online applications, social networking applications, search engines, even our mobile phones can utilise location to personalise information delivery. There are numerous sources of location information, including GPS, mobile phone cell triangulation, and more recently WIFI location analysis which uses the ubiquity of WIFI networks to help identify the location of an online device.

With the ever increasing quantity of location-based services on offer, being able to identify important locations for any given individual will help to focus these services better. A mobile network service provider, for example could implement social applications on subscribers' mobile devices by maintaining the location of a subscriber for an extended period of time and analysing this information to identify important locations such as home, work, social locations and holiday travelling locations. There are any numbers of possible end-user services, such as lifelogging, security services, advertising, and so on. In our consideration, being able to identify significant places automatically for any given user is a key aspect of supporting any location-based social network. While this is possible using GPS and mathematical techniques, it will become prohibitively time-consuming if it is required for an entire mobile phone operator subscriber base. So, clearly an alternative is required; a fast and effective alternative that is robust to changes in a person's lifestyle, such as moving home, changing job or even altering a daily routine to begin or end work at obscure hours. In this paper we carry out an analysis of a three year location log for one individual. This location information is GPS data of the individual's movement, which we decompose into three-tier location names using a gazetteer, which we feel will be more representative of mobile cell triangulation, but also make the data more like natural text technologies. We report on the nature of the location data, comparing it to the nature of natural spoken text and illustrate how we can employ proven approaches to term weighting to identify important locations in the life of an individual, for any number of purposes, especially social networking.

II. BACKGROUND

In describing this research, we need to introduce the concepts of power law distributions of naturally occurring concepts, term weighting techniques for natural language text, as well as introducing and reviewing pre-existing work in this area.

A. Uses of Location data to support Information Seeking

Employing location data to support information seeking has received increasing attention of late. Simply employing location as a means to support information access has been extensively examined for the area of digital photo retrieval, where it was shown to be one of the key access mechanisms for digital photo retrieval [1]. In recent years we have noted the increase in the quantity of location tagged photos on flickr.com which is a very popular access methodology. Concerning the actual analysis of location data to mine important concepts, Wolf *et. al.* [2] report on a technique for analysing GPS trips to identify trip purpose automatically to maintain travel diaries. Liao *et. al.* [3] identify significant places using a trained modeling technique on a small dataset of four people. Ashbrook & Starner [4] describe a clustering technique to identify meaningful locations and evaluate it positively on multiple users. Finally Kang *et. al*. [5] describe a technique for automatically identifying significant places using clustering and evaluate it positively for short (2 day) logs from a small number of people. Where our research differs from all the above is that we report an experiment at mining significant locations, using a fast processing, threshold free approach, that is evaluated on a multi-year location archive.

B. Distribution of Naturally Occurring Phenomena

Many naturally occurring phenomena can be seen to adhere to a certain type of distribution often referred to as a power law. Power laws are used in mathematics when one wishes to relate one quantity to the power of another. A power-law implies that small occurrences of a concept are extremely common whereas large occurrences are extremely rare. For example, if once considers the height of buildings in the world, there are very few tall buildings but there are very many small buildings. Power-laws are common to both manmade and naturally occurring phenomena [6], such as internet growth models, and in the distributions of word frequencies in language [7].

A power-law distribution when plotted on a graph is seen as a straight line with a distinctive slope on a log-log plot, or an extreme L shaped graph (hugging the axes) on a linear graph. See Figures 1-3 for examples. This is the characteristic signature of data that follows a power-law distribution.

Examining the word frequencies in a language, we have processed 371,610 unique words from 156,358 english language news articles as a representation of natural languge text. In Figure 1, we plot a power-law distribution of this textual data. We are interested to identify the shape of the distribution for a comparative analysis to the distribution of locations in a travelog.

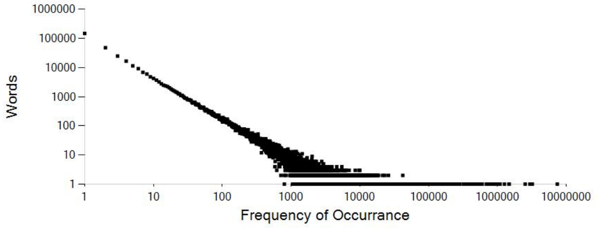

Figure 1. Plot of a Power-law distribution of words from english language text on a log-log scale, showing the frequency of occurrance of every unique word in the text corpus.

If location data follows a similar slope of distribution, then this suggests that the distributions (and consequently the characteristics) of the two types of data are similar. This suggests that we could employ term weighting text retrieval techniques successfully on location data from a travelog, because if the distribution of locations in a person's location log follows a power law, then it is likely that applying text search 'term' (in our case location) weighting techniques can help us to automatically identify important locations in a person's life, just as term weighting helps us to locate important terms in a document collection for ranking purposes (e.g. as successfully applied by the search engine provider Google).

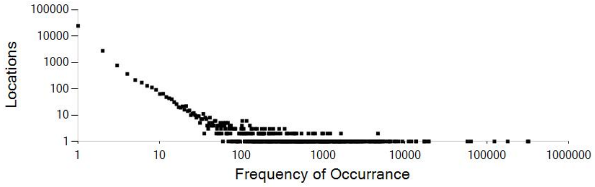

Figure 2. Plot of a Power-law distribution of location data on a log-log scale, showing the number of places visited where the user lingered for lengths of time from 1 to 1,000,000 minutes.

As can be seen from Figure 2, the power-law distribution of the location data looks very similar to the term frequency plot in Figure 1.

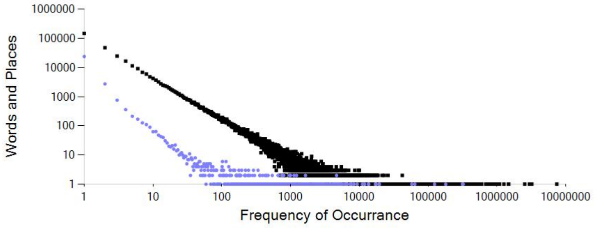

Figure 3. Plot of a Power-law distribution of locations (left) and words (right).

Examining both distributitions on a single graph (as in Figure 3) the similarities between the two plots are clearly visible. The primary difference is in the slope of the distributions, with the term frequecies having a much higher frequecny of occurrance and hence a much steeper slope. Therefore it is clear that locations follow the naturally occurring phenomena of a power law, and our conjecture is that the term weighting techniques from information retrieval research that are successfully applied to text data could also be applied to location data.

*C. Term Weighting using TF*IDF*

In text retrieval, one of the initial challenges faced by researchers in the field was how to identify the most important terms in a piece of text, and in a language as a whole. Finding a solution to this problem would allow for retrieval of ranked lists of documents and not just sets of documents based on Boolean logic. The solution lies in the work of Luhn [8] which states that "*the frequency of word occurrence in an article furnished a useful measurement of word significance*". This means that that we can use term frequency information from both documents and the language as a whole to identify and weight highly the important terms in a language.

This is achieved using various approaches to term weighting, with the most well known being the TF*IDF ranking technique [9]. TF*IDF ranking associates term importance weights with terms in documents by employing two term frequency components, TF and IDF. TF refers to Term Frequency, and is basically a measure of how important a term is to a document, by simply counting its frequency of occurrence in a document. IDF is a global document collection score that identifies how important the term is to the document collection as a whole. The more a term occurs across all documents, the less discriminating it is as a query term and the lower its IDF value is. The less a term occurs consequently means that the term is more descriminating and hence more desirable as an aspect of document ranking and will have a higher IDF value. IDF is basically the inverse of score called DF (Document Frequency) which is a count of the number of documents that a term occurs in.

TF*IDF weighting allows for the calculation of a term importance weight for the occurrence of a unique term in a document, and is calculated using the following formula (1) where: wij represents the weight assigned to a term Tj in a document D_i. , tf_{ij} = frequency of term T_j in document D_i, N = number of documents in collection and dfj = number of documents where term T_j occurs at least once:

$$w_{ij} = tf_{ij} \cdot \log\left(\frac{N}{df_j}\right) \tag{1}$$

It is our conjecture that, since the distributions of locations (a naturally occurring phenomena) follows a similar distribution law to the distribution of words in natural language (see Figure 3) - then it should be possible to utilize text IR term weighting techniques, such as TF*IDF to automatically identify important locations in a person's location log. We do however note that where location log analysis differs from text IR is that in text IR, the least useful words are the words that occur most often, for example 'and', 'the'. However, in location log analysis the most frequently occurring terms are likely to be the most important places in a person's life i.e. home and work locations. In addition for this research, the concept of a document needs to be defined in terms of location logs. Hence, we assume that a document is a month of location log, giving a total of 39 individual documents. Identifying an individual trip unit as a document does not make sense as a trip will typically not contain an extended time period anchored in one location.

III. Data Employed

Since we are analysing location logs to mine important information, it is necessary to have access to large location logs. For this research, we are fortunate to have access to a three year location log (39 months), captured using GPS location at ten second intervals, from mid November 2005 to January 2009. The owner of the location log was dedicated enough to turn on the GPS recorder device at any time he was moving from one location to another. This recording was achieved over 99.5% of all available days, and as such is the most complete location log available to us.Within this travelog, a trip to work, on holiday, or to the shop was recorded, but walking from one building to another in a place of work was not, primarily due to the start-up time of turning on the GPS device.

From this data, we calculated the location of the individual at every second over the 39 month period. When there were any breaks in the GPS coverage, the last known point was employed. These locations were converted from raw GPS points into a three part hierarchical location textual description, e.g. GPS co-ordinate 30.299982, -97.591782 is converted to the following place name: Walter E. Long Lake, Texas, United States. This is achieved by querying a gazetteer of over 7 million entries for the nearest entry to any given GPS point. The rationale for taking this approach (i.e. not using the raw GPS values and a clustering technique) is because firstly, this better replicates phone network cell location analysis and secondly, this technique is faster and inherently more scalable to support its employment for a large subscriber base.

IV. Analysing the Location Log

As stated the location log extended over 39 months and was created by one individual, which contained location data from 43 countries. Any movement was logged by the individual, including walking, driving and any airline flights taken during this time. The lifestyle of the individual is such that a reasonable amount of international travel was undertaken during this period, averaging about twelve international trips per year. Since airline location is included, the number of countries visited seems artificially high, since flying over a country would result in it being given a location log. In Table 1 we show the countries actually visited by the location logger, while ignoring countries that were stayed-in for less than one hour, and locations that were over sea and therefore not associated with any one country (286 hours). The data is further subdivided into countries that the user visited (normal text) and countries that the user simply passed-through while in an airplane (italic text).

TABLE 1. COUNTRIES VISITED AND HOURS SPENT THERE.

Country	**Hours**	**Country**	**Hours**	**Country**	**Hours**
Ireland	19,824	Norway	4,073	China	1,316
South Korea	750	UK	625	US	268
Singapore	253	Hong Kong	133	Finland	131
France	101	Japan	82	Denmark	56

Germany	54	*Sweden*	*41*	*Russia*	*33*
Holland	29	*Poland*	*15*	*India*	*5*
Belgium	*4*	*Ukraine*	*3*	*Malaysia*	*3*
Mongolia	*2*	*Estonia*	*2*	*Afghanistan*	*2*
Canada	1	*Belarus*	*1*	*Turkmenistan*	*1*
Latvia	1	*Pakistan*	*1*		

The number of GPS points logged, countries visited, actual named places visited and the average duration spent at each location is shown in Table 2, year by year, and in total. The reason for the average time spent at each location dropping significantly was because the individual moved from a location with no car (2005) to a location where a car was necessary (2006).

TABLE 2. SUMMARISING THE LOCATION LOG

Year	**# Points**	**# Countries**	**# Places**	**Avg Time (mins)**
All	1,021,607	43	27,508	61
2009	20,878	6	1,096	33
2008	361,312	36	14,527	36
2007	318,638	33	11,492	45
2006	305,869	11	7,091	74
2005	14,910	8	388	174

V. ANALYSING LOCATION DISTRIBUTION

Since a person's movement is a natural occurrence, one would expect that a small number of places would be very frequently visited (such as work or home) and a large number of places would simply be passed through (see Table 1). As shown in Figure 4, which is a log-log (axes) plot of the places and the minutes spent in these places over the entire 39 month archive, this is indeed the case, because the expected power law characteristic is clearly evident.

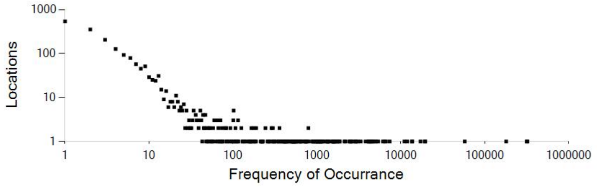

Figure 4. Log-log plot of the distribution of location data within the home country of the individual.

However, if one removes any non-home country data from the distribution, it still results in a similar distribution (though with a different slope), as shown in Figure 5.

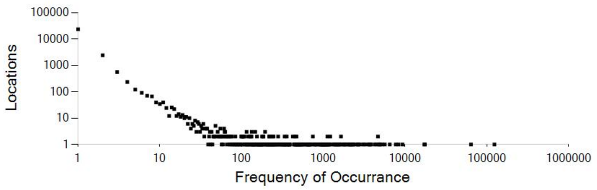

Figure 5. Log-log plot of the distribution of location data out of the home country of the individual.

A further type of location that we can examine is holiday destinations. Examining only holiday locations visited in the 39 month period, we note that any one country selected follows a power-law distribution, see Figure 6. All holiday destinations in our log adhere to this style of distribution in our location log.

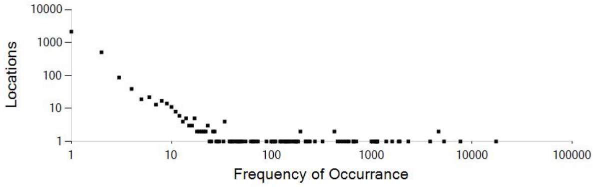

Figure 6. Log-log plot of the distribution of location data in a country visited a number of times for holidays.

Taking any country individually from the travelog, we note that all locations in our log follow a power-law distribution fairly accurately, except work trips away from home, where the distribution still appears roughly like a power-law, but the correlation to a trendline would be far less and the line is significantly closer to the origin. This was found to be the case for all work related travel locations. See Figure 7 for one such example. We assume this is due to the fact that on a work-related trip, the user will not be visiting many locations when compared to being on a holiday.

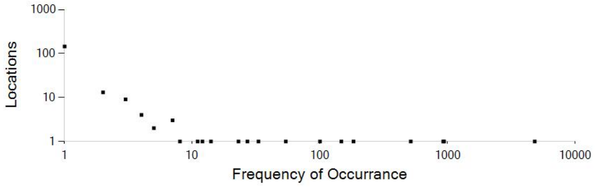

Figure 7. Log-log plot of the distribution of location data in a country visited only for work related activities.

An overvation of these distributions on a per-county level suggests that by analysing the slope of a trendline over the distribution and the correlation of the data to this trendline, it is possible to accurately identify the purpose of a trip to any given location, which is a task for further research, and not presented in this paper.

We now examine the possibility of utilising common term weighting techniques to automatically identify significant locations from the travelogs.

VI. Mining Important Concepts

Utilising proven term weighting techniques to identify significant locations should be both faster and less reliant on rules than other clustering based techniques (see Section 2). As mentioned earlier there are a number of components of a popular term weighting scheme such as TF*IDF. By employing the components of TF*IDF term weighting (TF, DF, IDF) we now define four weighting techniques for identifiying the importatant locations from personal location logs. They are:

- **TF** - the *Term Frequency*, which if taken globally can identify the most commonly occurring locations in a month's location log, which we would expect would refer to the individuals place of home and work.
- **DF** - the *Document Frequency*, which if assuming a monthly subdivision of the data into documents, would identify the locations that are repeatedly visited across various months, e.g. we visit regularly, such as family home locations, as well as typical home/work places.
- **TF*IDF** - the complete TF*IDF term weighting technique, which should identify the most important locations that may not be visited every month, but at which we spend some time in those months that we do visit. We consider that these would be social locations, such as relatives or favourite places to visit.
- **TF*DF** - In addition we define TF*DF, which in text retrieval would highly weight common terms within a month (such as the home and work locations), in a manner similar to TF, but increasing the weighting of locations that are visited every month.

Given these measurements, we were interested in locating a number of important location types:

- **Home/Work locations** - the locations we most frequently visit, which would be important for many locations applications, such as our user had purchased a new home, so we expected to locate both homes.
- **Social locations** - the locations that are most similar to the important terms in a language, and the locations that we attend not every day. We would expect to be able to identify important social locations from the archive automatically, the locations that the individual returns to again and again, such as family home, relations home and socialising locations.
- **Extended visit locations** - such as places where the individual has spent some time, but not reoccurring too frequently. For example, holiday locations or work travel locations.
- **Pass-through locations** - such as the places we pass through very often, but rarely stop at. These are exemplified by short linger/stay durations which occur frequently. An example of these locations are the places that we pass through on our way to work every day.

VII. Experimentation Results

In the following tables we identify the accuracy of identifying the four location types using each of the four algorithms previously described. To achieve these figures we calculate the precision at cut-off levels (1, 3, 5 & 10) for each month and then calculate the overall average precision. In Table 3 we show the best performing algorithms to identify (in bold) the four location types.

TABLE 3. AVERAGE PRECISION FOR 1,3,5 AND 10 LOCATIONS.

	Home/Work		**Social Visit**		**Long Visit**		**Passing Through**	
TF*DF **P@** *1/3* *P@ 5/10*	**1.0**	**0.83**	0.0	0.0	0.0	0.16	0.0	0.01
	0.59	**0.51**	0.06	0.07	0.23	0.24	0.22	0.32
TF*IDF	0.26	0.28	**0.45**	**0.49**	0.16	0.15	**0.58**	**0.58**
	0.28	0.29	**0.38**	**0.35**	0.15	0.17	**0.58**	**0.59**
TF	0.97	0.8	0.0	0.11	0.03	0.15	0	0.14
	0.57	0.51	0.16	0.17	0.19	0.19	0.26	0.32
DF	0.58	0.55	0.0	0.01	**0.37**	**0.37**	0.11	0.14
	0.50	0.46	0.04	0.07	**0.28**	**0.25**	0.28	0.36

Clearly TF*DF is the best way to locate work/home locations, though it is only a little better than TF which is easier and faster to calculate. DF shows promise for being able to locate places of long visits, though the precision values are not as high as expected, and this needs more work. Finally TF*IDF does find significant social and visiting (holiday) locations, as expected, though once again not as successfully as hoped. Finally, a P@Rel (precision at total number of possible relevant items) evaluation gives a score of 1.0 for identifying the three home/work locations when using TF and TF*DF ranking techniques. Simple location frequecny analysis (TF) is not effective at finding any location other than Home/Work, and in this case, we have shown TF*DF to be more effective.

One remaining issue with these results is that we have not yet separately identified the place of work and the place of home. To this end, we employ our only rule-based assumption into the process. *Home is dominant after 6pm and before 6am, while work is dominant after 6am and before 6pm*. Assuming this , we employ TF*DF ranking to calculate home and work locations and found P@1 for home to be 1.0 and P@1 for Work to be 1.0, which is as expected. Since the user moved home during the logging period, we note that p@2 is actually 1.0 also, which illustrates robustness of the process and the proposed techniques.

VIII. CONCLUSION

In this work we examined the location distributions of a large location log gathered by one individual in an effort to automatically identify significant locations. We noted that the location data follows a similar distribution to natural language text and as such we were able to successfully employ popular term weighting techniques from text retrieval to identify significant locations automatically. We found most success in identifying significant home and work locations, though social locations can also be identified using TF*IDF location weighting. This technique, is robust and efficent in that it builds on proven, effective and efficient term weighting techniques from information retrieval, which makes it inherently scalable to large datasets.

For future work, we will analyse data from more users, examine these issues of scale and efficiency, improve the term weighting algorithms and examine what other location types are important for a social location network scenario. Finally, we note that the experiments and results reported in this work are dependent on the individual wearer, though we feel that most people will produce similar distributions of location data, and consequently that these techniques will be effective for most users.

ACKNOWLEDGEMENTS

This material is based upon works supported by the Science Foundation Ireland under Grant No. 07/CE/I1147.

REFERENCES

[1] Naaman, M., Harada, S., Wang, Q., Garcia-Molina, H. and Paepcke, A. Context Data in Geo-Referenced Digital Photo Collections. In proceedings ACM MM 2004, October 2004.

[2] J Wolf, R. Guensler, and W Bachman. Elimination of the travel diary: Experiment to derive trip purpose from GPS travel data. Transportation Research Record, 1768, 2001.

[3] Liao, L., Fox, D., and Kautz, H. 2007. Extracting Places and Activities from GPS Traces Using Hierarchical Conditional Random Fields. Int. J. Rob. Res. 26, 1 (Jan. 2007).

[4] Ashbrook, D. and Starner, T. 2003. Using GPS to learn significant locations and predict movement across multiple users. Personal Ubiquitous Comput. 7, 5 (Oct. 2003), 275-286.

[5] Kang, J. H., Welbourne, W., Stewart, B., and Borriello, G. 2004. Extracting places from traces of locations. In Proceedings of the 2nd ACM international Workshop on Wireless Mobile Applications and Services on WLAN Hotspots (Philadelphia, PA, USA, October 01 - 01, 2004). WMASH '04.

[6] Mitzenmacher M (2001) A Brief History of Generative Models for Power Law and Lognormal Distributions. In Proceedings of the 39th Annual Allerton Conference on Communication, Control, and Computing, pp. 182-191.

[7] Adamic L and Humberman B (2001) The Web's Hidden Order. In: Communications of the ACM, Vol. 44, No. 9.

[8] Luhn, H.P. The automatic creation of literature abstracts. IBM Journal of Research and Development (159165), 1958

[9] Robertson, S.E., and Sparck Jones, K. (1997). Simple, Proven Approaches to Text Retrieval. University of Cambridge Technical Report.