Retrieval of Similar Travel Routes Using GPS Tracklog Place Names

Aiden R. Doherty, Cathal Gurrin, Gareth J. F. Jones, Alan F. Smeaton
Centre for Digital Video Processing & Adaptive Information Cluster
Dublin City University
Glasnevin, Dublin 9, Ireland
adoherty@computing.dcu.ie

ABSTRACT
GPS tracklogs provide a valuable record of routes travelled. In this paper we describe initial experiments exploring the use of text information retrieval techniques for the location of similar trips from within a GPS tracklog. We performed the experiment on a dataset of 528 individual trips gathered over a seven month time period from a single user. The results of our preliminary study suggest that traditional text-based information retrieval techniques can indeed be used to locate similar and related tracklogs.

Categories and Subject Descriptors
H.3.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval – *Retrieval models*

General Terms
Algorithms, Experimentation.

Keywords
GPS, trip matching.

1. INTRODUCTION
In recent years, the GPS (Global Positioning System) has become an accepted aspect of modern life. Small handheld GPS devices are becoming commonplace and many automobiles now contain GPS navigation screens. We are becoming accustomed to the ready availability of GPS information and this will only increase with the advent of GPS enabled mobile phones and the emergence of location tracking using cell-based telephony triangulation. Since employing GPS data allows the accurate location of an object to within a few metres on the planet, this data could be put to any number of uses, from retrospectively geo-stamping digital photos to simply keeping a record of the movements of an individual.

In this paper we explore the problem of matching trips from GPS tracklog data. Given the likely increase in the number of GPS enabled devices in the near future, we believe this will become an important issue. There could be many reasons for wanting to do this: finding most traveled routes by a number of people, matching routes for tourists in a new city, recommendation of routes based on similarities of previous travels, building a digital memory of human activities, and many more. In this investigation we focus on the employment of traditional textual information retrieval techniques to index and match trips automatically segmented from a user's tracklog. Text information retrieval models enable the significance of indexing attributes, both within individual documents and across a collection of documents, to be captured using term weighting, and used effectively in the identification of potentially relevant documents. We believe that the characteristics of GPS tracklogs mean that these term weighting and document matching techniques will transfer well to this quite different dataset. Also text retrieval methods have been shown to scale well to very large document collections.

We have developed an initial dataset for these experiments by taking GPS data collected by one individual over a seven month period from November 2005 to June 2006. The individual carried a GPS device with them at all times and the device appended to the tracklog as they moved location. These trips were gathered from four different countries over this period. A total of 148,500 location points were contained within the dataset, representing 528 individual trips.

The experimental process involved the data being automatically segmented into separate trips, the GPS co-ordinates of these trips were converted to place names using a gazetteer lookup, and then indexed using the BM25 text retrieval model. We conducted a preliminary retrieval experiment on a small subset of the trips to determine how effective our text-based retrieval system was at finding similar trips. The BM25 text retrieval model is known to scale up well to deal with very large amounts of data, and has been proven to be very effective in retrieving documents from hundreds of millions of web pages.

The following sections describe some existing related work and give more details of our investigations. Finally we discuss our initial results and speculate on future work.

2. RELATED WORK
There has been some previous work on finding similar trips, however the focus of that work has been on the starting and ending points of a trip [1][2][3]. In this paper we consider the route, not just the start and end locations.

Other papers in literature are mainly interested in deriving the purpose of each trip [1][3][4]. In this paper our motivation is

SIGIR GIR'06, August 10, 2006, Seattle, USA.

solely focused on effectively matching trips following similar routes, rather than just matching start/end points.

Ashbrook and Starner [5] use K-means clustering to group many similar GPS points to have just one co-ordinate representing that cluster. In our work we convert all GPS co-ordinates into place names and our motivation for doing this is that it allows us the possibility to investigate the use of textual information retrieval techniques for trip matching.

Morris et al. [6] do matching on the whole trip in their paper. They automatically build up a model of the various possible routes in that area, and use Markov chains to match trips. At each node they compute the probability of the next node that a person is likely to go to. However their test set was confined to a very limited geographic space (a nature park). We are not convinced that this approach would scale up very large tracklog archives collected from multiple users over extended periods. We believe that a traditional text retrieval model, such as BM25, is well proven in terms of dealing with large amounts of data and providing fast and accurate retrieval facilities.

An alternative to the approach explored in this paper is to match the geographical regions described by a set of points by employing spatial ranking methods which compare their degree of overlap [7]. While this would give a measure of similarity between a GPS tracklog query and a set of potentially related tracklogs, it would not capture certain aspects covered by our approach. For example, by only looking at the locations of points without taking account of their frequency, the potential significance of locations within the tracklog would not be captured. Also no measure of the similarity of segments of the tracklog enclosed with shapes describing geographical regions would be made; thus potentially very interesting similar paths described within quite dissimilar regions would be missed completely in the comparison.

3. EXPERIMENTAL PROCEDURE

GPS tracklogs record a stream of locations and are not ordinarily segmented into separate trips when extracted from a GPS device. Tracklogs may thus cover a number of separate activities spread out over a period of time. In order to separate a tracklog into separate events we employ an initial trip segmentation process prior to evaluating trip similarity measures.

The segmented trip logs are then represented as text documents prior to indexing the trips using BM25 and our subsequent evaluation.

3.1 Trip Segmentation

To segment our GPS data into individual trips, we use a simple method discussed by Gemmell et al. [8]. They *"…simply look for gaps in space or time above a given threshold (e.g. time gap > 90 min, or location gap > 1 mile) in order to divide the data…"*. To explain this approach, please consider Table 1 on the top right of this page, which shows GPS log points (latitude, longitude) and the timestamp of these points from a trip.

As suggested by Gemmell et al. [8], we look at the time interval between successive GPS points. If we consider only the first row and the second row, we can calculate that there is just a ten second interval between them. Therefore we conclude that due to the small time interval between them, it is highly probable that they both belong to the same trip.

Time Recorded	Latitude	Longitude
2005-12-03 16:32:20	69.64965	18.95511
2005-12-03 16:32:30	69.64965	19.95511
2005-12-03 16:32:40	69.649651	18.955107
2005-12-04 12:57:00	69.88103	18.955107
2005-12-04 12:57:10	69.68805	18.99932
2005-12-04 12:57:20	69.688	18.99923

Table 1 Trips to be segmented in GPS tracklog

However if we consider the third row and the fourth row, we can quickly ascertain that the time interval between them of over twenty hours, is much more than the recommended threshold of ninety minutes. Therefore we deduce that the fourth row is the start of a new trip. In this way we segment all the rows in our table of GPS tracklog data into trips.

Once all trips from the GPS tracklog have been segmented, we convert these trips into a textually indexable format. To achieve this, all GPS co-ordinates are converted into corresponding place names e.g. GPS co-ordinate 53.385787,-6.258795 is converted to the following place name: Glasnevin_Dublin_IRELAND. This is achieved by querying a gazetteer of over 7 million entries for the nearest entry to any given GPS point. All 148,500 GPS points from our tracklog were converted to place names.

```
Glasnevin_Dublin_IRELAND Glasnevin_Dublin_IRELAND
Glasnevin_Dublin_IRELAND Glasnevin_Dublin_IRELAND
Santry_Dublin_IRELAND Santry_Dublin_IRELAND Santry
Santry_Dublin_IRELAND Santry_Dublin_IRELAND Santry
Santry_Dublin_IRELAND Santry_Dublin_IRELAND Santry
Santry_Dublin_IRELAND Santry_Dublin_IRELAND Santry
TowerHouse_Dublin_IRELAND TowerHouse_Dublin_IRELAN
TowerHouse_Dublin_IRELAND TowerHouse_Dublin_IRELAN
TowerHouse_Dublin_IRELAND TowerHouse_Dublin_IRELAN
KildonanHouse_Dublin_IRELAND KildonanHouse_Dublin_
KildonanHouse_Dublin_IRELAND KildonanHouse_Dublin_
Cappoge_Dublin_IRELAND Cappoge_Dublin_IRELAND Capp
```

Figure 1 Sample trip document

After segmentation and naming steps we then have a document for each individual trip which contains all the place names of the GPS co-ordinates for that trip. As our GPS device records new co-ordinates every ten seconds, these documents can become quite large. Obviously many of the place names are repeated in the documents, but this is addressed by the similarity matching algorithm as an indication of location importance.

3.2 Trip Similarity Calculation

For retrieval from the trip archive we employ the BM25 indexing model [9]. For our initial experiments the BM25 parameters were selected based on prior experience with text archives. We feel this is justified as an initial examination of the frequency of occurrence of the locations within the 148,500 GPS points indicates that the locations suggest a Zipfian distribution [10], in which very few locations occur very frequently, and very many locations occur rarely. We plan to use the outcome of our preliminary experiment to suggest better tuned parameters for future work. As would be expected, no stopword removal or stemming was required. The individual locations within the trip documents were used as indexing terms.

In order to evaluate the effectiveness of trip matching using BM25, we selected ten trips at random and employed them as 'more like this' queries to the BM25 ranking engine. The top 10 matching trips were examined for each query.

Sample Trip	Precision @ 10	Precision @ 5
1	0.90	1.00
2	0.80	1.00
3	0.20	0.20
4	0.90	0.80
5	0.90	0.80
6	0.70	0.80
7	0.67	1.00
8	0.10	0.20
9	1.00	1.00
10	0.80	1.00
Average	0.70	0.78

Table 2 Precision values for trip similarity

Sample Trip	Precision @ 10	Precision @ 5
1	0.30	0.40
2	0.80	1.00
3	0.20	0.20
4	0.90	0.80
5	0.60	0.60
6	0.20	0.40
7	0.22	0.40
8	0.10	0.20
9	1.00	1.00
10	0.80	1.00
Average	0.51	0.60

Table 3 Trips that are similar and of the same length too

4. EXPERIMENTAL RESULTS

We visually investigated the results of all ten randomly chosen trips for different types of trips via a map interface. Each ranked result trip was observed on a map and judged as being similar or not to the query trip. This was a simple binary decision. In making a decision about each trip similarity, trips that did not follow the same route (road) for a significant part of the trips were judged as not similar. This would reduce the perceived effectiveness of trip matching and perhaps a less strict definition of trip similarity would be more applicable in many cases, such as that of a tourist trying to locate other routes people took through a certain area.

Figure 2 illustrates the plotting of a GPS log upon a map, such as was used for trip similarity evaluation. The top ten ranked results were chosen for examination, thereby providing the precision at 5 and 10 results presented in Table 2 and Table 3 above.

Figure 2 Sample trip

One aspect of trip matching that should not be ignored is the fact that trips are naturally going to be of differing lengths, yet still converge with another trip over part of the triplog. In table 1 we have not taken trip distance into consideration and matched trips where a reasonable part of the triplog converges.

In Table 3 we only consider trips that are a similar length to the query trip to be relevant. Naturally the average precision values are decrease in this instance. Sample trips three and eight have noticeably low precision values. The reason for this is that this reference trip was a highly infrequent trip undertaken by the user. Therefore it was not possible to retrieve ten relevant documents.

From the small sample of trips we selected, the top ranked result was generally a well matched trip. Of course it should be expected that for results of a query trip x, trip x would itself often be the top ranked result. In the set of trips that we selected, this was the case on seven occasions. On the other three occasions the query trip was also ranked among the top ten results.

We also found it interesting to look at the least similar matches for different images returned in the top ten results. Displayed below in Figure 3 is the least similar match from the top 10, for the trip shown in Figure 2:

Figure 3 Result trip is shorter than reference trip

The result trip in Figure 3 is obviously shorter than the reference trip in Figure 2. However it follows closely a part of the trip in Figure 2. This may have been retrieved because it is a better match than other longer trips which do not follow a substantial amount of the same path as the reference trip. In such a scenario, in this experiment, this trip would not be judged as relevant, however, as mentioned, in alternative scenarios (such as the tourist guide) this trip could be assumed relevant in that it could suggest new and alternative routes for a tourist to take. This recommendation of a new route could easily be coupled with a visual guide of the sights along another route, possibly extracted from a shared archive of location stamped digital photos such as that presented by O'Hare et al [11].

On the other hand there are instances where the resultant trips are actually longer than the queried trip. Please consider the reference trip (short line in bottom right hand corner of Figure 4) and one of its least impressive results in a visual sense (Figure 5) on the next page:

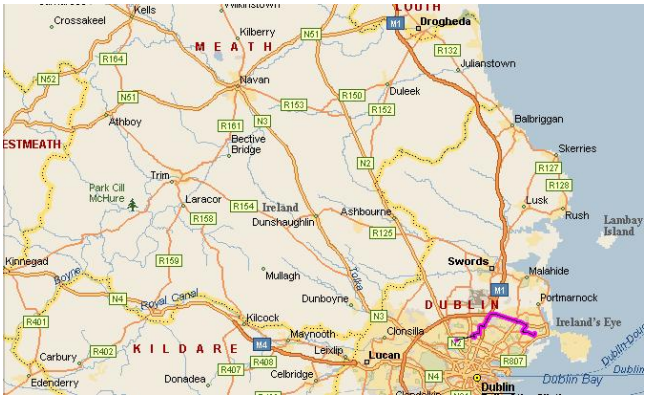

Figure 4 Reference trip

Figure 5 Result trip is longer than reference trip

The result trip here in Figure 5 is obviously much longer than the reference trip in Figure 4. In this case the trip may not have been segmented properly, as the start of the trip in Figure 5 is an exact match to the reference trip in Figure 4. However the remainder of the trip in Figure 5 appears that it should be a separate trip. The first section of this trip would have obtained a high score, and this may be a possible explanation for why it was retrieved in the top ten results.

These experimental results (especially Table 2 and 3) suggest that the BM25 model provides useful facilities to accurately match trips from GSP tracklogs. The inherent scalability of text retrieval techniques such as BM25 makes it ideal for matching trips on demand from very large trip archives. An interesting aspect of taking a 'bag of words' BM25 approach is that trips can be matched irrespective of direction. Subjects taking routes in opposite directions will often have been exposed to the same information during their travel. We plan to explore further the usefulness of this observation and to contrast it with a possible alternative retrieval method using direction dependant indexing units.

5. CONCLUSIONS AND FUTURE WORK

In this paper we present our initial work on trip matching based on GPS tracklogs using the BM25 text retrieval model. We believe the problem of trip matching will increase in importance in the future, and BM25 makes an ideal candidate for supporting matching, given that it is a proven model for dealing with very large datasets.

The findings presented in this paper are early results of our work. We still plan to tune the BM25 parameters from a testset of our collected data. The results of the small experiment in this paper are quite good, therefore it will be exciting to find out how much better they will be with tuned parameters.

We also have plans to compare this trip matching technique to more conventional trip matching tools, as well as other text retrieval techniques. Finally, we plan to expand the size of our dataset to incorporate tracklogs from other users and examine the effect this has on both retrieval performance, and also on our underlying assumptions of the applicability of such retrieval models as BM25 to the task in hand.

6. ACKNOWLEDGMENTS

This work is part-funded by the Irish Research Council for Science Engineering and Technology and is partially supported by Science Foundation Ireland under grant 03/IN.3/I361.

7. REFERENCES

[1] Wolf J, Guensler R, and Bachman W. *The Elimination of the Travel Diary: An Experiment to Derive Trip Purpose from GPS Travel Data.* Annual meeting of the Transportation Research Board, Washington, D.C., January 7-11, 2001.

[2] Jan O, Horowitz A.J, and Peng. *Using GPS Data to Understand Variations in Path Choice.* Transportation Research Record, 1725, pp 37-44, 2000.

[3] Schonfelder S, and Samaga U. *Where do you want to go today? – More observations on daily mobility.* Swiss Transport Research Conference, March 19-21, 2003.

[4] Zhou J, and Golledge R. *An Analysis of Variability of Travel Behavior within One-Week Period Based on GPS.* IGERT Conference, UC Davis, U.S., April 2000.

[5] Ashbrook D, and Starner T. *Using GPS to learn significant locations and predict movement across multiple users.* Personal and Ubiquitous Computing, 2003.

[6] Morris S, and Gimblett R, Barnard K. *Probabilistic Travel Modeling using GPS.* International Congress on Modelling and Simulation, Melbourne, 12-15 December, 2005.

[7] Larson, R.R, and Frontiera, P., *Spatial Ranking Methods for Geographic Information Retrieval (GIR) in Digital Libraries* European Digital Libraries Conference (ECDL 2004), Bath, pp45-56, 2004.

[8] Gemmell J, Aris A, and Lueder. *Telling Stories with MyLifeBits.* ICME 2005, Amsterdam, July 6-8, 2005.

[9] Robertson S.E, Walker S, Jones S, Hancock Beaulieu M.M, and Gatford M. *Okapi at TREC-3.* TREC-3, the 3rd Text Retrieval Conference, pp 109-127, NIST, 1995.

[10] Zipf, G. *Selective Studies and the Principle of Relative Frequency in Language.* Harvard University Press, Cambridge, MA, 1932.

[11] O'Hare N, Gurrin C, Lee H, Murphy N, Smeaton A.F, and Jones G. *Digital Photos: Where and When?* ACM Multimedia 2005 - 13th ACM International Conference on Multimedia 2005, Singapore, 6-12 November 20