Discovery and Raking of Significant Trails

Dikaios Papadogkonas, George Roussos and Mark Levene

Birkbeck College, University of London
Malet Street
London WC2E 7HX, UK
++44 20 7631 6324

{dikaios, gr, mark}@dcs.bbk.c.uk

ABSTRACT

Trail-based representation has been identified as a useful approach in exploiting context histories. Indeed, trails provide within a single framework an effective way to record, structure and represent interactions between users and physical or digital resources. Yet, this approach faces several challenges before it becomes practical at large scale since trail-based navigation requires considerable storage and computational resources when carried out naively. In this position paper, we propose an architecture for a query engine for trails and associated discovery and ranking techniques that can address such performance considerations and thus support processing of large numbers of recorded trails. In particular, we propose a stochastic model for the representation of trails and trail aggregates, and suitable data structures for efficient storage, filtering and retrieval which result in significantly improved performance. Furthermore, we propose algorithms and associated metrics that we use to rank trails and identify significant ones. Such significant trails can be subsequently used to guide search and navigation. We conclude by illustrating the use of these techniques in the context of three case studies.

Categories and Subject Descriptors

E.2 [**Data Storage representations**]: Composite structures, Contiguous representations, Hash-table representations, Linked representations, Object representation. H.3.3 [**Information Search and Retrieval**]: Information filtering, Query formulation, Relevance feedback, Retrieval models, Search process, Selection process.

General Terms

Algorithms, Performance, Experimentation.

Keywords

Trails, stochastic models, probabilistic grammars, suffix tree, spatiotemporal data mining.

1. INTRODUCTION

In this position paper we report on our investigations in developing efficient and effective tools for the discovery and ranking of significant trails in ubiquitous computing environments. These techniques are general in that they can be employed at any level of abstraction and incorporate whatever types of user or service interactions are deemed appropriate. Our long term aim is to explore such techniques for the development of navigational assistance tools in the situation brought about by ubiquitous computing.

In this position paper we consider a particular context for the application of these techniques, whereby material and digital objects coexist in a single environment which possesses both spatial and information characteristics. This situation presents new challenges to users or visitors in finding their way to particular objects or information that they are seeking. Indeed, by blurring the boundaries between the physical and the digital, ubiquitous computing constructs mixed spaces that are the source as well as the repository of massive amounts of information created by their use, a fact that severely limits the capability of humans to navigate effectively this situation. As already identified in the previous ECHISE workshop and elsewhere, trails [3] can play a critical role for the development of effective navigation tools in this case and they are at the core of the techniques discussed in this poster.

In this position paper we present our work on formalizing trail representation and discuss techniques that allow their aggregation and filtering. In particular, we propose a number of alternative metrics for the discovery and ranking of significant trails and illustrate their use in the context of three data sets. We present this work as a first step towards the development of search and navigation tools suitable for ubiquitous computing.

2. TRAIL RECONSTRUCTION

In this position paper we consider trails to be sequences of interactions between users and locations, material or digital artifacts and other users (though these ideas can be extended directly to other types of interactions including service and robot interactions). Trails can be observed and recorded by most ubiquitous computing environments albeit at different degrees of detail. The most common type of trail would be one that results from the use of a location sensing technology whereby users are timed and traced while traversing a particular area. At the other end of complexity, rich records of interaction can be recorded using a variety of sensing capabilities and would include exchanges of information with digital artifacts and other users, setting or following hyperlinks between physical objects, detailed descriptions of such interactions for example proximity and orientation towards an object, or creating and browsing annotations of specific places.

Irrespective of their type, in our approach all *interactions* are captured and represented as a node within a network representation in the form of a probabilistic grammar (cf. Figure

1). Sequences of interactions recorded for a specific user within a particular session are called *trails* and are represented as directed paths across the network graph. Each node can have several items of metadata associated with it for example duration of visit, related semantic information (e.g. name, location co-ordinates and so forth), the identifiers of the users that have visited and various computed statistics for example frequency and average time spent. Clearly, a considerable a mount of pre-processing is required for the captured log files to be translated into suitable format so that (non-spurious) interactions can be identified, trails constructed and filtered so that they can be added to the graph representation capturing all the recorded information. We call this graph the *interaction network*. This representation is suitable for this task as it preserves all the information recorded by each individual trail separately.

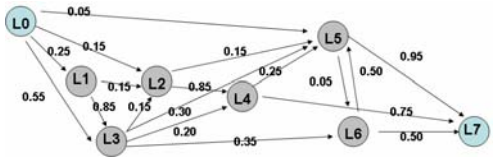

Figure 1. The interaction network records interactions of visitors and landmarks as a probabilistic grammar.

A critical ingredient for making this complete trail information usable in practice is the development of an efficient way to store and query this repository. We have extended the suffix tree data structure [5] to develop a data primitive suitable for this task which we call the *interaction tree*. This extended tree structure and associated algorithms can represent and efficiently query the interaction network (cf. Figure 2). More importantly, the interaction tree provides an effective mechanism to identify so-called *significant trails* that is trails that best match specific measures of prominence.

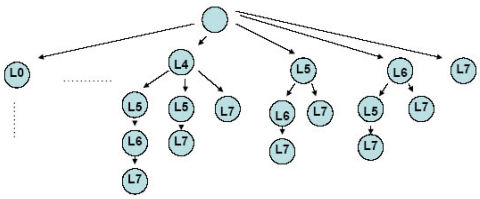

Figure 2. The interaction tree representing the interaction network of Figure 1.

3. QUERY ENGINE ELEMENTS

The trail reconstruction ideas have been implemented in a query engine (cf. Figure 3). The engine consists of a sessioning tool for reading data collected from the ubiquitous computing environment in consideration. Raw interaction data are often unsuitable for use as they include redundant, incomplete or inaccurate information or in other cases can only record spurious interactions that are not of any relevance at the level of abstraction where the engine works at. In other cases, several lower level events at a lower level of abstraction can be aggregated into higher levels events that are suitable for recording or specific thresholds (e.g. in terms of length of interactions) might be considered to separate spurious from relevant events. It is also important and indeed a challenge which often depends on the particular environment to separate event sequences into sessions before processing them. Such sessioning involves the identification of start and end points a problem that do far does not have a general solutions that covers all possible situations.

After sessioning, trails are incorporated into the probabilistic tree data structure which represents all recorded trails. The details of this data structure are beyond the scope of this position paper and are recorded elsewhere. Finally, a framework for the definition of suitable metrics used to identify significant trails and rank them according to significance is included as well as associated algorithm that can parse the interaction tree and return these significant trails and their importance scores. Different alternatives of possible metrics are briefly discussed in the following section.

Finally, we have implemented a user interface (cf. Figure 4) which allows users to query directly the recorded trails repository setting different metrics and ranking criteria. As we are still exploring the relative metric of different metrics it has been deemed appropriate to allow greater flexibility for such interactive exploration. In the long term, we expect that applications will directly query the engine through appropriate interfaces.

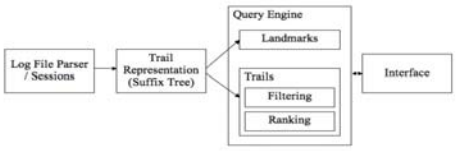

Figure 3. System architecture of the query engine.

4. RANKING TRAILS

The interaction tree also maintains links to metadata associated with individual nodes for two reasons: first, it makes possible to retrieve individual trails in addition to aggregates. Second, at each node it stores several weights measuring particular statistical properties, so as for example to allow for queries that relate to popularity or temporal characteristics of user activity and their orientation towards landmarks. These three characteristics of interaction between visitors and landmarks can be used in many cases to identify significant trails.

A significant trail is a sequence of interactions with a starting and ending point, such that it satisfies one of more the following criteria:

- It is one of the top-n trails in respect of trail popularity.
- It is one of the top-n trails in respect of average time (or some other time-related statistical measure) spent interacting with the landmarks in the trail.
- It is one of the top-n trails in respect of relevance of the landmarks to a chosen subject matter (for example steam locomotion technology exhibits within a science museum).
- It is one of the top-n trails in respect of one of the above criteria for a chosen demographic group.
- In addition, the above criteria may be combined and weighted for example, one of the top-n trails in respect of trail popularity that last at least a specific time period.

Significant trails are identified within a particular class of trails-for example all trails that start at an entrance landmark to the particular environment navigated and end at the exit landmark.

Other classes can be defined by selecting for example, all trails between any two specific locations; all trails that start at the entrance and last more than or less than a fixed period of time; all trails that pass through a specific landmark (e.g. a coffee shop); all trails of a fixed length; all trails that start and end at the same place (i.e. all cyclic trails); all trails that occur within a specific time, e.g. in the morning, in the afternoon, on one or every Tuesday, or during the Christmas period to list only a few. Significant trails can be inferred from the interaction network via a variety of tree-transversal methods.

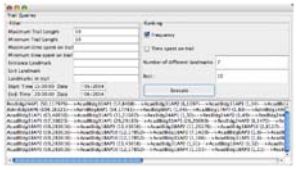

Figure 4. The query engine interface and textual output.

We have developed mechanisms to compute efficiently significant trails based on a fine grain heuristic defined on the interaction network, extending similar techniques developed for the web [1]. Nevertheless, choosing an appropriate metric that captures well the characteristics of a particular system is a challenging issue which we are currently investigating.

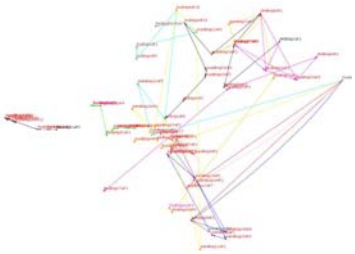

Figure 5. Top 10 trails in terms of frequency and of length 10 with at least 7 distinct landmarks visited.

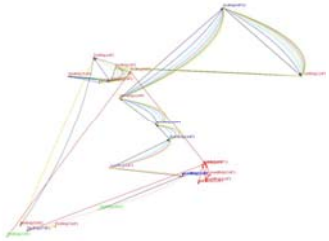

Figure 6. Top 5 trails in terms of time spent and of length 10 with at least 7 distinct landmarks visited.

5. CASE STUDIES

We use the query engine to discover and rank significant trails within three data sets that are available to us. The first case study relates to the wireless access network traces collected by the Crawdad project at the Centre for Mobile Computing at Dartmouth [4] (cf. Figures 5 and 6). This is a simple data set that only records connectivity patterns of users of the campus-wide wireless network based on the syslog records of the access points. The second case employs log files created by imote devices in several experiments run by the Computer Laboratory at the University of Cambridge [2]. Similar to the Dartmouth data set, these log files also record interactions using the wireless interface of the imotes though in this case the devices are either carried by individual users or embedded in artifacts (cf. Figure 7). The third experiment relates to the study of user orientation patterns in the London Zoo conducted for the re-design of their signage system. In this case, landmarks were selected beforehand to represent places of interest to the evaluation team within the Zoo for example, specific exhibits and other recreation areas (cf. Figure 8).

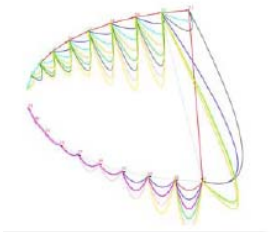

Figure 7. Top 10 significant trails by time and frequency combined, of size 10, with 7 different landmarks.

6. CONCLUSION

In this position paper we introduce a framework for the discovery and ranking of significant trails within ubiquitous computing systems. These techniques provide an efficient and effective means for the development of navigational assistance tools which in the longer-term we expect to help users avoid “being lost in ubiquitous computing space”. We believe that the navigation problem in ubiquitous computing environments is a significant one and we expect that the techniques developed here would be the first step towards the construction of search and navigation engines and associated tools for ubiquitous computing.

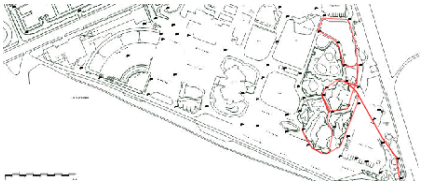

Figure 8. Best trail from entrance gate to exit.

7. REFERENCES

[1] J. Borges and M. Levene, A fine grained heuristic to capture web navigation patterns. *SIGKDD Explorations*, Vol. 2, pp. 40-50, 2000.

[2] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass and J. Scott, Pocket Switched Networks: Real-world mobility and its consequences for opportunistic forwarding, Technical Report UCAMCL-TR-617, University of Cambridge, Computer Laboratory, February, 2005.

[3] S. Clarke and C. Driver. "Context-Aware Trails". *IEEE Computer*, Vol. 37, No. 8. pp. 97-99, August 2004.

[4] D. Kotz and K. Essien, Analysis of a campus wide wireless network, Wireless Networks, vol.11, pp. 115-133, 2005.

[5] P. Weiner, Linear Pattern Matching Algorithms, *Proc. 14th IEEE Annual Symp. on Switching and Automata Theory*, pp. 1-11, 1973