Alan Walks Wales
Data and Challenges

Alan Dix

28/5/2014

Summary

I have extensive quantitative data (bio-sensing and location) and qualitative data (text, images and some audio). There are challenges in analysing individual kinds of data, including merging similar data streams, entity identification, time-series and textual data mining, dealing with provenance, ontologies for paths, and journeys. There are also challenges for author and third-party annotation, linking the data-sets and visualising the merged narrative or facets of it.

Data Available

Quantitive Data

GPX – location traces (Garmin and phone app)

EDA – skin conductance

ECG – heart activity – about 60 days with about 30 nights too

Accelerometers – on both wrist and chest sensors

Qualitative Data

blogs – about 2-3,000 words per day

images – between 250 and 650 per day, all timestamped

audio blogs – very variable typically 6-12 a day

Links to data and documentation at:
http://alandix.com/alanwalkswales/data/

Challenges

Here are some of the challenges I have thought of … I'm sure there are more

Location data

This seems simplest data, but still has surprising complexity.

GPX files come from two different sources. There are gaps where batteries ran out, I turned on the devices late, and sometimes extraneous data, where I forgot to turn off before getting on a bus. In addition some photos have GPS and timestamp, some photos are timestamped, but I can identify precisely where they are. What **tools help merge** these, and allow **human 'fixing' of gaps**, glitches, etc. How to record **provenance** or final data.

Interestingly GPX files can store routes, which are sequences locations only, for planned ways to go and tracks, which are sequences of timestamped locations, for where you have actually been. Of course hand edited tracks will included a mix of sensed locations with and hand-added locations from maps without precise times.

Representing a location on a route can even be problematic as a route can cross through the same point twice (e.g. Menai Bridge to Isle of Anglesey) and distance is ill defined for a fractal such as the Welsh Coast.

Bio data

This is rather specialist, but has been converted to CSV to make it easier to access. Are there patterns in the ECG that recur and maybe correlate with terrain, weather etc. Are there trends visible over the length of the trip, esp. in the night-time ECG.

Qualitative data

I am part way though hand marking key entities (places, names, etc.). For automatic entity resolution the blogs have good locality structure for place names. Where there blogs are hand-marked, these can be a gold standard to check algorithms. For sentiment analysis form the text, I can still recall pretty much what I felt like as I was walking … of course not the same as whet I felt like wen I was writing abut a particular day … so can verify algorithmic tagging.

Semantics of blogs – can narrative ontologies be used to mark-up this kind of account. The blogs sometimes refer to incidents during the walk, sometimes mention memories of previous times in the walk, or childhood; sometimes mention world events or lapse into philosophical reverie; sometimes make mention of the time of writing.

Audio data has not been transcribed, and I've ha trouble getting speech-to-text to work. The recordings typically had a lot of wind noise. Are there techniques to get reasonable recognition despite this? Can the non-speech part of the audio be analysed for sentiment analysis ?

Images: GPS or timestamp tells you where they were taken form, not the location of hat they represent (e.g. distant mountain). For landscapes , is it possible to combine the known location of the photographer (from GPS/timestamp) with satellite elevation date, to identify peeks automatically?

Linking

Timestamps can be used to connect the quantitative data … except care needed as of course each device has its own time, albeit all as close as possible to BST.

Place names can link the blogs to GPS traces, and GOS traces can be used to help disambiguate place names in blogs. Of course the blog can easily refer to names out of sequence "as I looked at the rock I remembered XXX" – indeed the challenge here is that the narrative effectively refers to two places (where it is and what I'm thinking of), and two times.

Although automatic methods can establish a putative connection, there will need to be hand 'tweaking', what sort of tool support would help, bot bespoke, but also generically – what data tool should I have had available that would have helped me do this linking … and could help now – what is the Excel of data linkage?

… and of course how does one represent this kind of linkage

Augmentation

As well as adding mark-up myself, if someone else uses the data (e.g. someone asked me for photo links as they were interested in the images of maps along the way), how can their tagging, mark-up, cleaned data, derived data, be added back into the over data set whilst maintaining provenance etc.

What tools would help both my own and third-party mark-up?

Visualisation

In the blogs I've included a small number of images, but each day has hundreds of images, plus audio blogs as well as the text. In addition there are cross-references in the text, etc. How do you represent this kind of thing to a reader?

In addition I'd like to be able to view a theme (assuming the text/images) are marked up, say 'signage', so that I see just the relevant text, but in a way that makes it easy to see this in context.

Similarly when some analysis algorithm points out interesting incident in the ECG data it would be good to be able to easily see the audio, images and blog close to the time … not each fo these has different 'densities', so 'close' means different things.