The Extraction of Events from Differently Heterogenous Social Networking Data for Life-Logging purposes: A System for the Automatic Extraction of Social Network Event Groups (SAESNEG)

Ben Blamey

May 1, 2014

Summary

Life-logging can (arguably) be defined as any practice of recording ones daily life in digital form, for the purpose of reference, review or reminiscence. It is strongly associated with the study of data collected from wearable cameras (a relatively niche activity), as well as evangelists who develop their own systems for storing emails, phone-calls, and a miscellanea of digitized documents. The quantified self movement, where the body is equipped with wearable sensors, to record exercise performance and physiological data can also be viewed as a form of life-logging, and is rapidly growing in popularity. However, this thesis explores opportunities for life-logging in the context of online social networks (OSNs) based on data available from the social media footprint of the average user, who is likely not a user of wearable cameras or other specialist equipment. Unlike existing approaches which require the use of new software or devices for data collection, our system theoretically makes a life-logging experience available to the wider public, by using the information already stored in their social network accounts their social media footprint.

In the first critical literature review chapter, a sound motivation for the work is synthesised from a number of existing ideas in the literature: beginning with a discussion of the user experience of OSNs; how they are considered to be 'focused on the now' with reminiscence-related activities neglected. The personal social media footprint tends to be distributed over multiple social networks, this fragmentation obstructs a comprehensive reminiscence experience, and can lead to loss of data - with death creating particular problems. A range of relevant work is critically reviewed; including efforts to develop user interfaces appropriate for the potentially high data volume created over a lifetime.

This thesis proposes SAESNEG (System for the Automated Extraction of Social Network Event Groups) a pipeline for the aggregation of a personal social media footprint, and its automatic organization into a set of events the so-called event identification task, facilitating a user experience focused around reminiscence, so that the user is able to successfully navigate the large, heterogeneous data in their social media footprint.

Moving to technical discussion, previous systems have tended to focus on the organization of a single kind of data, such as photos or tweets exclusively. The system proposed in this thesis is intended to handle a variety of kinds of social network datums found in a typical social media footprint, with a variety of image, text and other metadata, grouping this source data into a sequence of events, constituting a life story for the user. The author believes it is the first such unified system capable of doing this.

The concept of an event, its various guises and representations within different fields of data-mining is discussed: its historic influence on systems for organizing personal photo collections, and those for mining textual data sources from newswire, and later, micro-blogging services. We combine algorithms and techniques from these areas, as well as developing a number of novel techniques, into a single, extensible pipeline capable of handling a variety of data. Specifically, the characteristics of the research problem that distinguish it from similar systems include:

- A focus on the personal social media footprint, where data tends to be non-public. Many previous studies have addressed the event identification task in the context of large, public datasets; originating from Flickr and Twitter, for example. The abundance and redundancy characterizing these datasets means that high performance can be achieved with machine learning algorithms, based on features extracted by comparatively simple techniques focusing on public events, whilst excluding background noise. Conversely, the personal social media footprint for a single individual can be comparatively sparse, containing many private events with an order-of-magnitude less data available. This greatly reduces the data available for the training of machine-learning models, demanding more elaborate techniques. These new techniques extract richer information in order to maximize event-identification performance on the smaller dataset.

- Handling differently heterogeneous data, i.e. an input dataset where documents are structured differently, and contain different kinds of data to each other, rather than a uniformly heterogeneous dataset, a collection of one kind of document, as has been the focus of many previous studies, creating a subtly distinct research problem.

The implementation of the system combines a range of techniques from these different fields. The first step (Phase A), extracts information from the datums, focusing on natural language processing to make use of any available text content. New techniques are developed including a novel approach to handling temporal expressions, a novel CRF-based NER tagger incorporating semantic features from OpenStreetMap, and a grammatical parser for specific events such as birthdays. These contribute to the relevant NLP areas in their own right, and are evaluated independently of the pipeline. Image-content features are extracted using state-of-the-art techniques, and rules are defined to take advantage of structured metadata. These Phase A algorithms yield rich annotations to drive the event-identification phase (Phase B).

Phase B performs event identification, grouping the datums according to the underlying real-world event. A key focus of the thesis is the development of a number of similarity metrics, computed on

pairs of datums. These algorithms (the FESTIBUSK strategies) are a mixture of new and existing algorithms. In a development on previous work, low-level algorithms, which focus on single event facets (who/where/when) are combined with more high-level reasoning, drawing on information from multiple facets, to maximize event-identification performance.

For the final event identification step, a number of algorithms are evaluated datums are grouped into events based on the second tier of features extracted by the FESTIBUSK algorithms in phase B. The differently heterogeneous source data creates sparsity in the second tier feature set, creating a challenge for final clustering. Ground truth data for training and testing the algorithms is obtained from users in an interactive web-interface, which also includes the necessary ethics forms and instructional screencast.

The contributions of the research include: the pipeline itself and the specific research task identified, the various algorithms and techniques used in the components of the pipeline, some of which have been published independently. The thesis concludes with a critical reflection of the use of the event model, and how it underpins the pipeline and the techniques used, finishing off with discussion of various avenues for future work, and planned extensions to the system.