Foundations and Trends® in
Information Retrieval
8:1

LifeLogging
Personal Big Data

Cathal Gurrin, Alan F. Smeaton,
and Aiden R. Doherty

now
the essence of knowledge

LifeLogging: Personal Big Data

Cathal Gurrin
Insight Centre for Data Analytics
Dublin City University
cgurrin@computing.dcu.ie

Alan F. Smeaton
Insight Centre for Data Analytics
Dublin City University
alan.smeaton@dcu.ie

Aiden R. Doherty
Nuffield Department of Population Health
University of Oxford
aiden.doherty@dph.ox.ac.uk

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Information Retrieval

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

C. Gurrin, A. F. Smeaton, and A. R. Doherty. *LifeLogging: Personal Big Data.* Foundations and Trends® in Information Retrieval, vol. 8, no. 1, pp. 1–107, 2014.

This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-60198-802-7

LifeLogging:
Personal Big Data

Foundations and Trends® in Information Retrieval

Volume 8, Issue 1, 2014

Editorial Board

Editorial Scope

Topics

Foundations and Trends® in Information Retrieval publishes survey and tutorial articles in the following topics:

- Applications of IR
- Architectures for IR
- Collaborative filtering and recommender systems
- Cross-lingual and multilingual IR
- Distributed IR and federated search
- Evaluation issues and test collections for IR
- Formal models and language models for IR
- IR on mobile platforms
- Indexing and retrieval of structured documents
- Information categorization and clustering
- Information extraction
- Information filtering and routing
- Metasearch, rank aggregation, and data fusion
- Natural language processing for IR
- Performance issues for IR systems, including algorithms, data structures, optimization techniques, and scalability
- Question answering
- Summarization of single documents, multiple documents, and corpora
- Text mining
- Topic detection and tracking
- Usability, interactivity, and visualization issues in IR
- User modelling and user studies for IR
- Web search

Information for Librarians

Foundations and Trends® in Information Retrieval, 2014, Volume 8, 5 issues. ISSN paper version 1554-0669. ISSN online version 1554-0677. Also available as a combined paper and online subscription.

Foundations and Trends® in Information Retrieval
Vol. 8, No. 1 (2014) 1–125

DOI: 10.1561/1500000033

now
the essence of knowledge

LifeLogging: Personal Big Data

Cathal Gurrin
Insight Centre for Data Analytics
Dublin City University
cgurrin@computing.dcu.ie

Alan F. Smeaton
Insight Centre for Data Analytics
Dublin City University
alan.smeaton@dcu.ie

Aiden R. Doherty
Nuffield Department of Population Health
University of Oxford
aiden.doherty@dph.ox.ac.uk

Contents

1 Introduction **3**

1.1 Terminology, definitions and memory 4
1.2 Motivation 8
1.3 Who lifelogs and why ? 11
1.4 Topics in lifelogging 15
1.5 Review outline 18

2 Background **19**

2.1 History 19
2.2 Capture, storage and retrieval advances 27
2.3 Lifelogging disciplines 36

3 Sourcing and Storing Lifelog Data **39**

3.1 Sources of lifelog data 39
3.2 Lifelogging: personal big data — little big data 45
3.3 Storage models for lifelog data 47

4 Organising Lifelog Data **51**

4.1 Identifying events 54
4.2 Annotating events and other atomic units of retrieval . . . 59
4.3 Search and retrieval within lifelogs 68
4.4 User experience and user interfaces 76

4.5 Evaluation: methodologies and challenges 80

5 Lifelogging Applications **85**
5.1 Personal lifelogging applications 86
5.2 Population-based lifelogging applications 90
5.3 Potential applications of lifelogging in information retrieval 92

6 Conclusions and Issues **97**
6.1 Issues with lifelogging 97
6.2 Future directions 103
6.3 Conclusion 105

Acknowledgments **107**

References **109**

Abstract

We have recently observed a convergence of technologies to foster the emergence of lifelogging as a mainstream activity. Computer storage has become significantly cheaper, and advancements in sensing technology allows for the efficient sensing of personal activities, locations and the environment. This is best seen in the growing popularity of the quantified self movement, in which life activities are tracked using wearable sensors in the hope of better understanding human performance in a variety of tasks. This review aims to provide a comprehensive summary of lifelogging, to cover its research history, current technologies, and applications. Thus far, most of the lifelogging research has focused predominantly on visual lifelogging in order to capture life details of life activities, hence we maintain this focus in this review. However, we also reflect on the challenges lifelogging poses to an information retrieval scientist. This review is a suitable reference for those seeking an information retrieval scientist's perspective on lifelogging and the quantified self.

C. Gurrin, A. F. Smeaton, and A. R. Doherty. *LifeLogging: Personal Big Data.* Foundations and Trends® in Information Retrieval, vol. 8, no. 1, pp. 1–125, 2014.
DOI: 10.1561/1500000033.

1

Introduction

Lifelogging represents a phenomenon whereby people can digitally record their own daily lives in varying amounts of detail, for a variety of purposes. In a sense it represents a comprehensive "black box" of a human's life activities and may offer the potential to mine or infer knowledge about how we live our lives. As with all new technologies there are early adopters, the extreme lifeloggers, who attempt to record as much of life into their "black box" as they can. While many may not want to have such a fine-grained and detailed black box of their lives, these early adopters, and the technologies that they develop, will have more universal appeal in some form, either as a scaled-down version for certain applications or as a full lifelogging activity in the years to come.

Lifelogging may offer benefits to content-based information retrieval, contextual retrieval, browsing, search, linking, summarisation and user interaction. However, there are challenges in managing, analysing, indexing and providing content-based access to streams of multimodal information derived from lifelog sensors which can be noisy, error-prone and with gaps in continuity due to sensor calibration or failure. The opportunities that lifelogging offers are based on the fact that

a lifelog, as a black box of our lives, offers rich contextual information, which has been an Achilles heel of information discovery. If we know a detailed *context* of the user (for example, who the user is, where she is and has been recently, what she is doing now and has done, who she is with, etc...) then we could leverage this context to develop more useful tools for information access; see the recent FNTIR review of Contextual Information Retrieval, Melucci (2012). This valuable contextual information provided by lifelogging to the field of information retrieval has received little research attention to date.

Before we outline the content of this review we will introduce and define what we mean by lifelogging, discuss who lifelogs and why they do so, and then introduce some of the applications and core topics in the area.

1.1 Terminology, definitions and memory

There is no universal or agreed definition of lifelogging and there are many activities which are referred to as lifelogging, each producing some form of a lifelog data archive. Some of the more popular of these activities include quantified-self analytics[1], lifeblogs, lifeglogs, personal (or human) digital memories, lifetime stores, the human black box, and so on.

In choosing an appropriate definition, we refer to the description of lifelogging by Dodge and Kitchin (2007), where lifelogging is referred to as"*a form of pervasive computing, consisting of a unified digital record of the totality of an individual's experiences, captured multi-modally through digital sensors and stored permanently as a personal multimedia archive*". The unified digital record uses multi-modally captured data which has been gathered, stored, and processed into semantically meaningful and retrievable information and has been made accessible through an interface, which can potentially support a wide variety of use-cases, as we will describe later.

A key aspect of this definition is that the lifelog should strive to record a totality of an individual's experiences. Currently, it is not

[1] http://quantifiedself.com

possible to actually record the totality of an individual's experiences, due to limitations in sensor hardware. However, we take on-board the spirit of this definition and for the remainder of this review, we assume that lifelogging attempts to capture a detailed trace of an individuals actions. Therefore, much of the lifelogging discussion in this review is concerned with multimodal sensing, including wearable cameras which have driven many first generation lifelogging efforts.

Because lifelogging is an emergent area[2], it is full of terminology that is not well considered and defined. Therefore, for the purposes of this discussion, we regard the lifelogging process as having the following three core elements:

- *Lifelogging* is the process of passively gathering, processing, and reflecting on life experience data collected by a variety of sensors, and is carried out by an individual, the lifelogger. The life experience data is mostly based on wearable sensors which directly sense activities of the person, though sometimes data from environmental sensors or other informational sensors can be incorporated into the process;

- A *Lifelog* is the actual data gathered. It could reside on a personal hard drive, in the cloud or in some portable storage device. The lifelog could be as simple as a collection of photos, or could become as large and complex as a lifetime of wearable sensory output (for example, GPS location logs or accelerometer activity traces);

- A *Surrogate Memory* is akin to a digital library, it is the data from the lifelog and the associated software to organise and manage lifelog data. This is the key challenge for information retrieval, to develop a new generation of retrieval technologies that operates over such enormous new data archives. Given the term surrogate memory, we must point out that this does not imply any form of cognitive processes taking place, rather it is simply the digital li-

[2]Although lifelogging has been around for several decades in various forms, it has only recently become popular.

brary for lifelog data, which heretofore has been typically focused on maintaining a list of events or episodes from life;

It is important to consider that lifelogging is typically carried out *ambiently* or passively without the lifelogger having to initiate anything. There have been a number of dedicated individuals who are willing to actively try to log the totality of their lives, but these are still in the very significant minority. For example, Richard Buckminster Fuller manually logged every 15 minutes of activity from 1920 until 1983, into a scrapbook called the Dymaxion Chronofile, as described in Fuller et al. (2008). More recently Gordon Bell's MyLifeBits project, Bell and Gemmell (2007) combined active and passive logging by using wearable cameras and capturing real-world information accesses. Another example of active logging is Nick Feltron's Reporter app, which allows an individual to manually log whatever life activity they wish in as much detail as they desire. Reporter will periodically remind the user to 'report' on the current activities.

While such dedicated lifelogging is currently atypical, most of us often explicitly record aspects of our lives such as taking photos at a social event. In such cases there is a conscious decision to take the picture and we pose and smile for it. Lifelogging is different, in that by default it is always-on unless it is explicitly switched off and it operates in a passive manner. Therefore the process of lifelogging generates large volumes of data, much of it repetitive.Thus the contents of the lifelog are not just the deliberately posed photographs at the birthday party, but the lifelog also includes records of everything the individual has done, all day (and sometimes all night), including the mundane and habitual.

Compare this to the recently popular field of quantified self analytics. Quantified self is considered to be a movement to incorporate technology into data acquisition on aspects of a person's daily life in terms of inputs (e.g. food consumed, quality of surrounding air), states (e.g. mood, arousal, blood oxygen levels), and performance (mental and physical). While there is a level of ambiguity in terms of the cross-over between quantified self and lifelogging, this review assumes that the key difference between lifelogging and quantified self analytics is that

quantified self is a domain-focused effort at logging experiences (e.g. exercise levels, healthcare indicators) with a understanding of the key goals of the effort, whereas lifelogging is a more indiscriminate logging of the totality of life experience where the end use-cases and insights will not all be understood or known at the outset of lifelogging.

Considering how to organise these vast lifelog data archives, we believe that lifelog data should be structured in a manner somewhat similar to how the brain stores memories. While a debate on human memory models is beyond the scope of this review, we select the Cohen and Conway (2008) model of human memory due to the fact that many other memory scientists who have ventured into the application of lifelogging; for example Doherty et al. (2012); Pauly-Takacs et al. (2011); Silva et al. (2013), all refer to this model. Cohen and Conway's model suggests that the memory of specific events and experiences should be called our episodic memory. It is autobiographical and personal, and can be used to recall dates, times, places, people, emotions and other contextual facts. Our semantic memory is different and is our record of knowledge, facts about the real world, meanings and concepts that we have acquired over time. While our episodic memory is personal, our semantic memory is shared with others and is independent of our own personal experiences or emotions since its contents can stand alone and are abstract. It is suggested that our semantic memory is generally derived from our episodic memory in the process that is learning new facts or knowledge from our own personal experiences, as described in Cohen and Conway (2008) For lifelogging, much of the focus thus far has been on supporting and generating surrogates of episodic memory.

Based on such a model, one would consider a typical day being segmented into a series of events of various durations. Figure 1.1 shows a timeline of a day with events represented by an image and various metadata sources. Dressing and self-grooming, preparing food, eating, travel on a bus, watching TV, listening to music, working on a computer, taking part in a meeting, listening to a presentation, doing gardening, going to a gym, and so on, are all examples of everyday events. Some of these events are regular and repetitive. For example, many of us eat the same or similar breakfasts each day at approximately the

same time and in the same place. Going to a movie or attending a party is probably a rarer occurrence, perhaps weekly or monthly. While debate exists on the formation of human memories, the view presented in this review is that lifelogging creates a lifelog which is similar to the Cohen and Conway (2008) model of episodic memory. A lifelog captures the "facts" around the episodes in our lives but not their emotional interpretation.

A lifelog does not typically capture or store semantic memory, so when we want to know the capital city of Azerbaijan (Baku) or the winners of the 2000 FA Cup (Chelsea), we don't ask a lifelog, we go to Wikipedia or we search the web. As of now, we do not refer to a lifelog for such semantic facts. Therein lies one of the real challenges in lifelogging: how to search a lifelog for relevant information given that the IR techniques we have developed over the last several decades are developed to search semantic rather than episodic memory. We shall return to this point later.

Other use-cases of lifelogging are broad and varied, such as the ability to detect and mine insights from our daily lives, in a Quantified Self type of analysis. We will return to a detailed discussion of the use-cases later. Whichever use-cases we employ, in order to maximise the potential of lifelogging (as with any technology), we should map this new technology into our lives and develop the technology in support of, rather than to try to change, our lives around the technology. Thus at the outset we should ask ourselves what are the characteristics and structures which form the organisation of our lives where we can use lifelogging to build upon.

1.2 Motivation

Lifelogging is becoming more accessible to everyone due to data capture becoming more feasible and the availability of inexpensive data storage technologies. Gordon Bell from Microsoft was one of the first to fully embrace digitising his life as part of the MyLifeBits project (Gemmell et al. (2002, 2006)) at Microsoft Research and this helped raise the profile of lifelogging. Lifelogging alone can generate large volumes of