

gmail identica wakoopa happy factor google latitude del.icio.us facebook curetogether twitter last.fm plazes google calendar organic indoor location rememberthemilk slife plazes nike+ skydeck wesabe bedpost skyhook fitbit livestrong labpixies calorie counter trixie tracker

Watching through the Web

Using Web-based life tracking and context sensing services for long-term modeling and prediction of information needs

While some of our information needs occur just once, many recur repeatedly - whenever we are in some situation or perform a particular activity. In particular, needs surrounding our personal information, such as our documents, e-mails, contacts, and personal notes tend to recur often, such as whenever we are in certain places, with certain people, at certain times of the day, and whenever we work on certain tasks or projects.

In this line of research, we examine the feasibility of using a mixture of life-tracking Web sites and in-browser instrumentation to sense activity and build models that can be used to predict the user’s information needs.

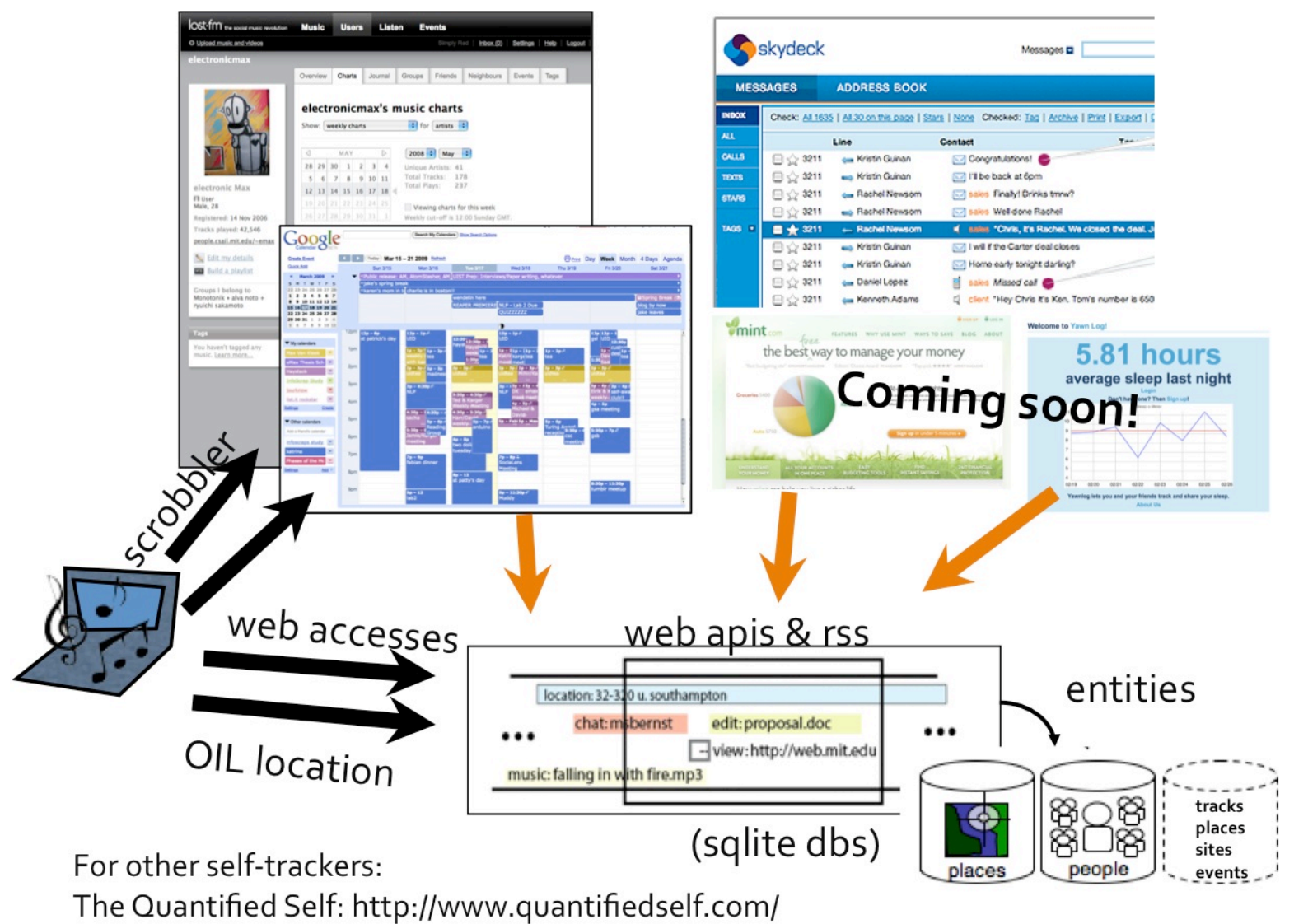
User activity tracking using the Web

A large number of web sites offer tools to facilitate the tracking of many low-level statistics of our lives - from the the things we do, to how we feel, to the people we meet, the things we eat, and places we go. Some of these sites track these activities automatically through sensors or through integration with software on the user’s personal devices. Others instead facilitate the manual logging of activity through mobile apps and compelling GUIs.

We have compiled a partial list of such services available in July 2009 and the kinds of contexts or activities they sense. The last few activities listed are not yet tracked by any service known to us and thus we have implemented our own instrumentation for these activities in our architecture for Personal Lifetime User Modeling (PLUM):

type of activity tracked	sensing technique	providers
e-mail access	software / web page instrumentation	Gmail Instrumentation (PLUM) for Greasemonkey
web page accesses	browser instrumentation	WatchMe (PLUM) for Firefox
music listening activity	media player instrumentation	Audioscrobbler (by Last.fm) for iTunes, Windows Media Player
physical location (outdoor)	WiFi for laptops, AGPS (w/ Cell tower) for mobile phones	Plazes, Google Latitude, Skyhook, W3C Geolocation
physical location (indoor)	Wifi for laptops	Organic Indoor Location (PRUNE)
mobile phone communications (calls placed, messages sent and received)	interface with mobile service providers	Skydeck, (soon: Google Voice)
running and walking activity	hardware	fitbit, Nike+, runkeeper (Other heart-rate monitors like Polar are catching up to Web2.0)
sleep	hardware, manual logging	fitbit, yawnlog
nutrition (calories consumed)	manual logging	livestrong.com, labpixies calorie counter, tweet what you eat, trixie tracker
credit card use	interface with financial institutions	wesabe.com (open API), mint.com (closed API)
social bookmarking	browser toolbar (and integration)	del.icio.us

type of activity tracked	sensing technique	providers
to-do items	manually input via various software clients, SMS	rememberthemilk
microblogging	manually input via various software clients, SMS	twitter.com, identica
private events (calendars)	manually input via various software clients, web interfaces	Google Calendar, iCal
public events	contributed by other web users	Any RSS feed on the Web
travel plans	manually input	dopplr
sexual activity	manually input	bedposted
recovery from sickness	manually input	curetogether
instant messages/chat	IM client or gateway instrumentation	PLUM, (??)
e-mail	IMAP server or e-mail client instrumentation	PLUM, (??)
document editing / interaction	application instrumentation	slife, vibe logger, PLUM (??)
face to face social interaction	automatic sensing	??



Data integration challenges

While many of the above services expose data via public Web APIs and RSS/ATOM feeds, using these data together is challenging in practice. Two immediate sets of challenges are **data integration** and **incompleteness/data sparsity**.

Towards **data integration**, data sources typically use different structures and forms to express information; in order to comparing data obtained from different sources, the system must be able to map or convert, the disparate representations to make them alike. This is the schema alignment problem. Second, since these separate data sources typically represent entities people, places and things in different ways without the use of universal identifiers, the system must be able to identify when two entities are the same (co-referring entity resolution).

Towards **data incompleteness**, some data sources may omit information that other sources provide, or may provide information at a different degree of precision or granularity. For example, locations may provide precise room-level granularity, a street address, or latitude/longitude.

Our approach has been to create a framework that reduces the complexity of these challenges as follows:

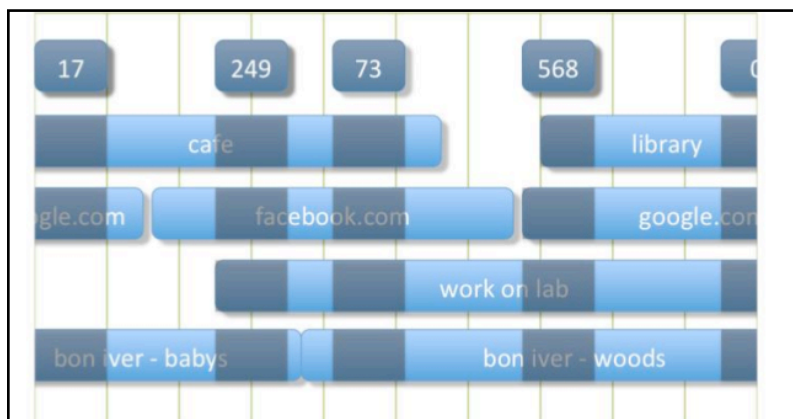
1. Use an simple, consistent internal world model (RDF) consisting of *entities* and *events*
2. Support the community-creation of import filters (to solve the alignment problem)
3. Use inverse-functional property heuristics to merge like entities, and empower end-users to repair problems
4. Incorporate notion of probability of correctness, granularity and data freshness in internal representations

Modeling from activity data

Based on integrated data, user activity and situation data can be brought into recommender models for information items.

Our initial simple approach watches how activities and situations/contexts overlap with information accesses over time. From this, estimate how likely each information item is useful given the user’s active context. This likelihood is combined with a textual context-based relevance term extracted from the user’s active document(s) and information item contents to yield a posterior likelihood of relevance. We are planning on improving this algorithm soon.

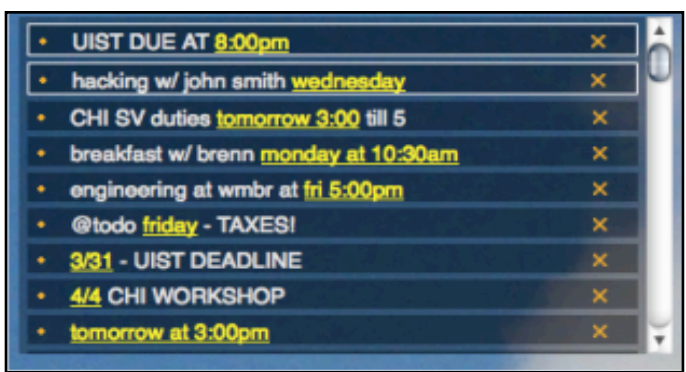
Step 1. Overlapping activities in the log are used to estimate joint probabilities of activities and notes/information items



Step 2. For a new situation, the posterior probability of needing each note /information item is estimated given the current known context

$$\begin{aligned} & \text{Relevance}(note_i | \text{note contents}, \text{user context}) \\ &= P(note_i | Tr(note_i, now), C_1, C_2, \dots, C_{|C|}) \\ &\propto Tr(note_i, now) P(note_i) P(C_1, C_2, \dots, C_{|C|} | note_i) \\ &\propto Tr(note_i, now) P(note_i) \prod_c P(C_c | note_i) \end{aligned}$$

Step 3. Notes with high posterior likelihoods of relevance are ranked and made visible



For more info: <http://plum.csail.mit.edu>

Max Van Kleek
mc schraefel
David Karger

emax@csail.mit.edu
mc+sigir@ecs.soton.ac.uk
karger@csail.mit.edu

