By their apps you shall understand them: mining large-scale patterns of mobile phone usage

Trinh-Minh-Tri Do
Idiap Research Institute
Switzerland
tri.do@idiap.ch

Daniel Gatica-Perez
Idiap Research Institute
EPF Lausanne
Switzerland
gatica@idiap.ch

ABSTRACT

Mobile phones are becoming more and more widely used nowadays, and people do not use the phone only for communication: there is a wide variety of phone applications allowing users to select those that fit their needs. Aggregated over time, application usage patterns exhibit not only what people are consistently interested in but also the way in which they use their phones, and can help improving phone design and personalized services. This work aims at mining automatically usage patterns from apps data recorded continuously with smartphones. A new probabilistic framework for mining usage patterns is proposed. Our methodology involves the design of a bag-of-apps model that robustly represents level of phone usage over specific times of the day, and the use of a probabilistic topic model that jointly discovers patterns of usage over multiple applications and describes users as mixtures of such patterns. Our framework is evaluated using 230 000+ hours of real-life app phone log data, demonstrates that relevant patterns of usage can be extracted, and is objectively validated on a user retrieval task with competitive performance.

1. INTRODUCTION

Mobile smartphones are the epitome of ubiquitous multimedia devices. Phones are equipped to shoot photos and video, listen to music, browse the web, estimate our location, and communicate via voice, text, and multimedia. Smartphones are also increasingly seen as large-scale, unintrusive sensors of human activity recording data, related both to the physical and social pace of people's lives and to how we interact with our devices. This brings an enormous potential to the use of phones as part of large-scale behavioral studies, using actual sensor and application data as an extension to traditional ways of collecting behavioral information (typically through questionnaires and other forms of self-reports) [3, 17, 16, 15].

Among the many data types currently available on smartphones, a fundamental one is represented by phone applications. Whether pre-installed or available for download from highly popular sites (like the successful iPhone App store or Nokia Ovi store), phone applications tell much about what we do and like as users, and how we relate to our devices in the context of daily life. People have different ways of using their mobile phone depending on their needs, interests, and situational contexts. For business purposes, a user may use mainly voice calls during working time at the office, but SMS or email in a public space. A teen user may send and receive a lot of SMS during the whole day but rarely call. The potential of automatically understanding patterns of phone app usage from populations of users is significant, ranging from collecting unbiased data about the popularity of specific applications, to characterizing users' preferences conditioned on a number of key contextual cues (time, location, social situation), and to improve phone design and personalized mobile services [12, 19, 22, 9]. Analyzing phone usage data is critical to the mobile industry [18], including phone operators, manufacturers, advertising companies, and service providers, but it is also a challenging domain, especially for the design of computational frameworks that use the data efficiently, extract patterns of usage robustly, produce insights beyond the ones obtained with standard tools, and are capable of making predictions based on available data. In this emerging area, statistical machine learning methods have become the predominant modeling tool.

We present a novel probabilistic framework to automatically mine patterns of mobile phone usage at large scale. Based on the availability of large-scale log data of phone apps, our objectives are to discover, in a principled unsupervised way, the daily patterns of joint usage of phone apps at specific times of the day for a population of users, and to estimate the probability that each user has of conforming the discovered usage patterns. In other words, we address two fundamental questions: Which are the main emergent patterns of usage? Can we summarize user behavior by a mixture of usage patterns?. Our work has three contributions. In the first place, we propose a method based on (i) a novel data representation (bag-of-apps), which incorporates the level of usage for a set of commonly used apps anchored by the time of the day in which they are used, and (ii) a probabilistic topic model that infers the underlying structure of usage patterns and users. Our method is able to discover, without any supervision, meaningful patterns of usage behavior, such as using the phone mainly for communication during the day, or jointly using the camera and photo gallery in the afternoon, and estimates who are the most likely users

MUM '10, December 1-3, 2010, Limassol, Cyprus.

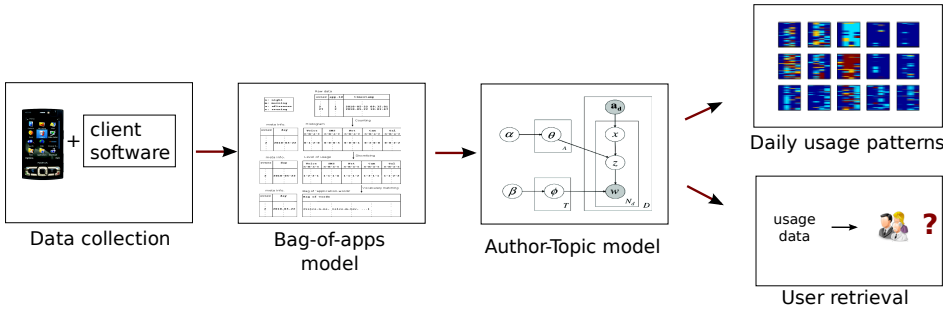

Figure 1: Overview of our method.

to interact with the phone in that particular manner. In the second place, we conduct our analysis on a large-scale data set collected with Nokia N95 smartphones, and involving 111 people and 230 000+ hours of real-life data. In the third place, we objectively validate our framework in the context of a user retrieval task, and show that it is feasible to correctly retrieve users from phone app data, which is a step forward towards predictive tasks. Overall, our framework is effective and extensible to multiple applications and contextual anchors.

The structure of the paper is as follows. Section 2 reviews related works on large-scale phone app data analysis. We present an overview of our methodology in Section 3. Section 4 described our data collection framework. Sections 5 and 6 present in detail our approach for mining phone app usage patterns. Section 7 presents the experimental results and the discussion. Finally, we draw conclusions in Section 8.

2. RELATED WORK

Our work is inscribed within an emerging body of work that is investigating the possibilities of analyzing human behavior at large-scale using mobile phones as sensors of activity. In all the works described in this section, mobile phones are equipped with software application that record logs of application usage, and in several cases many other phone sensors.

Eagle and Pentland [6, 5] pioneered the Reality Mining concept, conducting an extensive analysis of mobile phone data recorded with Nokia 6600 phones from 100 MIT students and staff over an academic year, extending the Context phone application developed by Raento et al. [20]. Specifically regarding phone application usage, in [6], the authors presented a summarized description of the most popular phone apps used in their data, finding that despite the availability of sophisticated features, the most common use of the phone was still communication, with voice clearly dominating over SMS and email. In [5], the authors proposed to use phone calls, in addition to Bluetooth, to define pairwise links between people and in this way infer friendship networks, as an alternative to questionnaire-based, self-reported data. In these studies, however, the question of what recurrent *joint* patterns of phone app usage can be mined from the data was not addressed.

Large-scale phone-based data collection and analysis efforts have also been conducted in industry, with special interest in user modeling and personalized services [12, 18]. As one example, France Telecom's Orange organized the KDD Cup 2009 using large marketing mobile phone databases [9]. Based on statistical usage of phone users (updated monthly), the task was to predict the propensity of customers to churn (switch provider), buy new products or services (appetency), or buy upgrades or add-ons proposed to them to make the sale more profitable (up-selling). The problem was addressed as a supervised learning task, and most participants in the evaluation proposed models based on black-box classifiers, from input features to desired labeling. These models can reasonably perform the tasks but can not explain the reasons behind user decisions. In a notable case, Nokia has conducted several studies over the past few years [19], and summaries of their main findings have sometimes appeared as part of press releases. In a study conducted in 2007 in three Western European countries and involving 540 "early-adopter" users equipped with S60 devices during three months, it was reported an increase in the overall time that people spent using their phone compared to previous years, and also that the increase in the use of certain apps, including web browser, music player, and email, was rather significant. In other related work, Verkasalo et al. [10, 22] recently presented an analysis of users and non-users of smartphone applications, investigating the relation between actual phone usage and a number of variables under the Technology Acceptance model (TAM) proposed by Davis [4], including perceived usefulness and enjoyment, social norms, behavioral control, and potential technological barriers. In particular, they studied three apps: internet, games, and maps, and used a data set provided by over 570 Finnish users who participated in a two-month study. Our work differs from this work in several ways. First, we are interested in what joint patterns of usage emerge from people's regular use of the phone, while Verkasalo studied the use of single apps. In the second place, our modeling tools are also different, moving from correlation analysis to a probabilistic approach based on topic models. Furthermore, our work integrates the time of day as contextual anchor in modeling, which introduces the assumption that usage is temporally grounded. Finally, the specific apps investigated in our work are different, and were used over a significantly larger period of time (up to eight months of time).

Finally, Farrahi et al. [7] first proposed the use of topic models to location mobile data. Using a small number of

manually-defined location labels (e.g. being at home or at work), the authors showed that Latent Dirichlet Allocation (LDA) could learn the dominant daily routines, in terms of location transitions, for the population recorded in the Reality Mining data. They also presented routine extraction results with the Author-Topic Model (ATM). While promising, no objective evaluation of the results produced by their methods was presented. Furthermore, the importance of phone app usage as part of the description of human daily patterns was not explored. We address both of these open issues in our work: we demonstrate that mining daily phone app logs can extract meaningful patterns of usage, and we propose a novel way of evaluating the performance of topic models in the context of user retrieval.

3. OVERVIEW OF OUR METHOD

Our approach for analyzing usage data is described in Figure 1. Its principal components are:

- Data collection: At the low level, a software application is installed into smartphones, which are then distributed to volunteers for gathering data. Data from the set of users are uploaded dailly to a central server for analysis.
- Bag-of-apps model: The raw data is transformed into a novel high-level representation in order to employ statistical methods efficiently. A day in user's life is represented by a bag-of-application, that describes the usage level of each application at a given period of the considered day.
- Author-Topic model: We propose the use of a probabilistic model that can infer the underlying structure of data based on the bag representation. The model explicitly encodes daily usage patterns as a latent variable, and discovers these patterns along with the users who are most likely to display such patterns by fitting model parameters to data.
- Applications: A primary application of the learned model is to discover joint application usage patterns, which can be visualized graphically. We also use the learned model to perform a user retrieval task: finding relevant users from some querying (e.g. prototypical) usage data.

Each of these blocks are presented in detail in the next sections.

4. COLLECTING PHONE APP DATA FROM REAL-LIFE USAGE

We use a server-client architecture built around the Nokia N95 smartphone in order to collect data [14]. The software client was designed to detect and record *all* phone applications logs (including system apps, pre-installed user apps like the camera or the calendar, and any user-downloaded apps), storing the logs in the phone's memory. Each time an application is opened, the client software captures the event and stores it (together with the timestamp) in memory. The client was installed in the phone and runs in the background in a non-intrusive way, starting automatically at startup, and able to record data on a 24/7 basis as long as the phone is on. The event logs are then uploaded daily to a server, typically done at night, via a user-defined wifi connection, which results in a fully automated solution. On the server, the raw data is simply a table of three columns: owner id, application id, and timestamp. Each row corresponds to a captured event.

	Voice	SMS	Internet	Camera	Gallery
#events	121874	85733	36658	16620	12660

Table 1: Applications that was considered in this study. The total number of events for each application are also shown.

The real-life data in this study was collected in a large-scale fashion [14]. The data comes from 111 volunteer users of a European population. Our sample is mainly composed of educated, middle class individuals in the 20-40 years age bracket, and contains a mix of university students and professionals living in urban and suburban environments, linked by either professional or social links. All participants were previous users of mobile phones, although most of them did not own an advanced smartphone before the study. Clearly, no claims are made that this population reflects accurately all types of phone users, but it does constitute a population that is less homogeneous than that featured in other studies (e.g. the Reality Mining data). Users carried their smartphone as their actual (and only) mobile phone, and therefore used them in real conditions. Users had different participation times, between one and eight months. The data used for this work was collected between October 2009 and May 2010. Each user was given an ID and all data was anonymized [14], so no personal information is used for experiments.

While our mining framework is designed for general types of apps (e.g. those in iPhone App store or Nokia Ovi store), an initial analysis of the raw collected app data (reported in Section 7.1) led us to consider the 5 most used applications (Cf. Table 1) in this work:

- *Voice*: An event consists of an incoming or outgoing phone call considered as a binary variable (i.e., phone call duration is not considered). Missed phone calls are not considered as a Voice event.
- *SMS*: Sent or received SMS.
- *Internet*: web browsers and e-mail clients. We consider both native applications and user-installed application (Opera, GMail).
- *Camera*: The native camera application for taking picture and recording video.
- *Gallery*: A route into images and videos, to sound clips, etc. It is considered as the direct route into stills and videos that have been shot with the camera.

The above constitutes the input data on which our framework, described in the next two sections, is applied.

5. BAG-OF-APPLICATIONS MODEL

In order to employ statistical methods efficiently for the analysis, these raw data need to be preprocessed and transformed into a convenient format using an analogy with text processing techniques, where documents are typically represented by bags of words (i.e. the count of the occurring words in the document). We propose to use a bag-of-applications model which uses the frequency of use of each application as the basic representation. Since we are interested in understanding how the user behavior changes with respect to the time of the day, the counts are distinguished for 4 timeslots: *Night (0am-6am)*, *Morning (6am-12am)* , *Afternoon (12am-6pm)*, and *Evening (6pm-0am)*. In the bag-of-application model, these timeslots are denoted by their initial letters, i.e. *n-m-a-e*.

The bag-of-apps model is described in Figure 2. Firstly, a histogram of application events for all applications and timeslots is built for each day in the life of each user. Secondly, these counts are quantized in order to obtain the corresponding level of usage within each timeslot. We used 4 levels:

level 1 (no-use)	: 0 times
level 2 (low-use)	: 1-2 times
level 3 (middle-use)	: 3-4 times
level 4 (high-use)	: more than 4 times

This quantization step maps specific counts into a small number of semantic classes that reflect typical patterns of usage (e.g. a couple of times, a few times or a lot).

Finally, we build directly a bag-of-apps representation for a day of a user from the level-of-usage vector, where each unit or *application-word* is a triplet *application-timeslot-usagelevel*. Since there are 5 applications, 4 timeslots, and 4 usage levels, a total of 80 app-words can possibly occur in a day. This bag representation is amenable for probabilistic topic modeling which will be used to discover usage patterns. This is described in the next section.

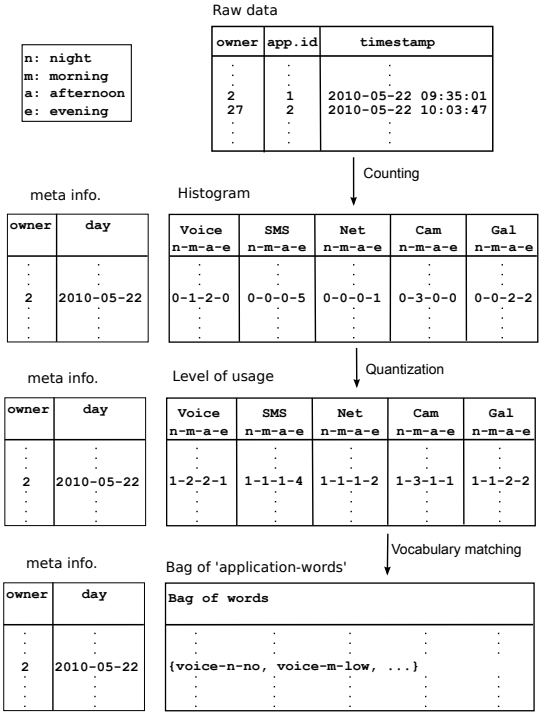

Figure 2: The bag-of-applications model, from raw data to high level representation. See text for details.

6. AUTHOR TOPIC MODEL

There are several statistical methods that can be used for discovering some emergent factors of the data such as principal component analysis (PCA) or clustering. However, PCA is a global factor analysis which does not allow discovering of emergent patterns which occur on a subset of the data. Applying clustering methods on our data can discover some clusters of "usage days", but can not necessarily discover basic patterns that can explain what specific parts of some days are similar (when they are supposed to be the same pattern). We consider a more complex machine learning technique that can discover particular patterns from the data in probabilistic terms.

From an applicative viewpoint, topic model is a tool for extracting emergent hidden patterns from a collection of data [2, 11]. Since our data consists of a collection of "user-specific days", we consider the Author-Topic model (ATM) in order to exploit efficiently the data ownership information [21]. Using the owner information we can distinguish between common behaviors of a user and rare behaviors (can be viewed as noise), and focus on relevant patterns.

While widely used in text modeling, topic models have applications to other domains such as image retrieval and bioinformatics [13, 23]. Note that topic models work with discrete data, and so we need a quantifying step that transforms application data into discrete data as explained in Figure 2.

As explained in Section 5, an *application-word*, denoted here by w, is a basic unit of the model, and it correspond to usage level of an application during a timeslot (e.g. *voice-m-low*). Since we have 5 applications, 4 timeslots and 4 usage levels, there are 80 possible app- words. A day of a user's life, denoted by $\mathbf{w}_d = \{w_d^i\}_{i=1..N_d}$, is represented as the set of N_d app-words, corresponding to the usage levels of the 5 applications at the 4 different timeslots (in our setting, by construction we always have $N_d = 20$ for every day). Finally, we denote by $\mathbf{a}_d$ the set of owners of the data for a given day d . Note that, we could merge data from multiple users into a bag (e.g. to represent who called who) to represent daily usage. However, in our setting, this set consists of a single owner of the considered day of data, i.e, the user.

In text processing, given a collection of documents, the topic models aim at finding the underlying structure of the data. By considering a document as a mixture of topics (i.e, the main themes the document is about), the model summarizes a document by a vector of mixture coefficients. Topics are latent variables that correspond to particular patterns. Of course, the latent topics are not explicit but are defined as a model parameter. Discovering topics is the process of fitting the model parameters to the observed data, and then visualizing the topics based on the model parameter.

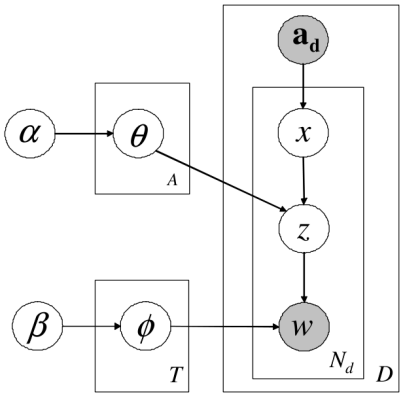

Figure 3: Graphical representation of the author topic model. Observed variables are represented by shadowed nodes.

The ATM graphical model is illustrated in Figure 3. The model has one latent variable z for topics, and two sets of latent parameters θ and ϕ, and two hyper-parameters α and β. The conditional probability of an observed day $\mathbf{w}_d$ given a user a_d is defined as:

$$P(\mathbf{w}_d|\mathbf{a}_d) = \prod_{i=1}^{N_d} \sum_{a \in \mathbf{a}_d} \frac{1}{|\mathbf{a}_d|} \sum_{z=1}^{T} P(w_d^i|z)P(z|a) \quad (1)$$

where T is the number of latent topics, $P(z|a)$ is a multinomial distribution with parameters θ_a and $P(w_d^i|z)$ is a multinomial distribution with parameters ϕ_z. The generative process is as follows:

For each of A users a: Sample $\theta_a \sim Dir(\alpha)$
For each of T topics z: Sample $\phi_a \sim Dir(\beta)$
For each day d in the collection of data (whose authors are $\mathbf{a}_d$)
For each of N_d words w_d^i:
Sample $a \sim Uniform(\mathbf{a}_d)$
Sample $z \sim Multinomial(\theta_a)$
Sample $w_d^i \sim Multinomial(\phi_z)$

where $Dir(.)$ denotes the Dirichlet distribution, that is the conjugate prior of the multinomial distribution.

In our setting, since there is only one owner for a given observed day, we consider a modified version of ATM where the random variable assigning a word to the author can be ignored (i.e. the x node in Figure 3 can be removed). Then, the generative probability can be simplified as

$$P(\mathbf{w}_d|a_d) = \prod_{i=1}^{N_d} \sum_{z=1}^{T} P(w_d^i|z)P(z|a_d) \quad (2)$$

where a_d is the single owner of the data day d. At the end, each user a is characterized by a multinomial distribution over topics with parameters (θ_a), and each topic z is characterized by a multinomial distribution over words (ϕ_z). Learning the model (parameter estimation) correspond to find the relation between users and topics, and the relation between topics and words. The problem of finding optimum model parameters is intractable in general. However, a wide variety of approximation techniques can be used, including Laplace approximation, variation approximation, and Markov chain Monte Carlo (MCMC). In our experiment, we use Gibbs sampling which is a special instance of MCMC [8].

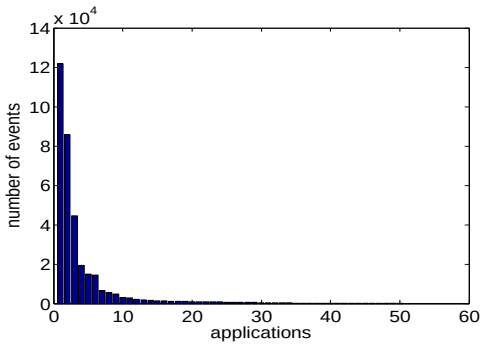

Figure 4: Number of events vs applications.

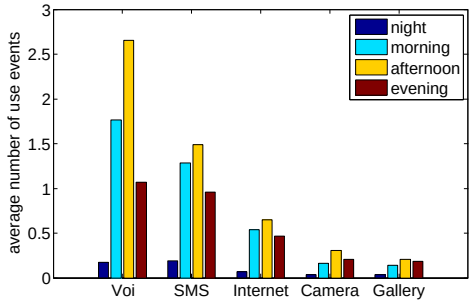

Figure 5: Average number of uses for each application.

7. RESULTS

In this section, we present our findings from the usage data. In Section 7.1, we start with some global basic statistics of the data which show, on average, how each application is used. Next, Section 7.2 presents usage patterns that were discovered by the ATM model. In Section 7.3 we present a user retrieval task which provides an objective evaluation of our framework.

7.1 Basic statistics

Hundreds of applications were found in the data but most of them are system applications. Considering only user-related application (that is, both user-downloadable or pre-installed in the phone), there are roughly 50 different applications.

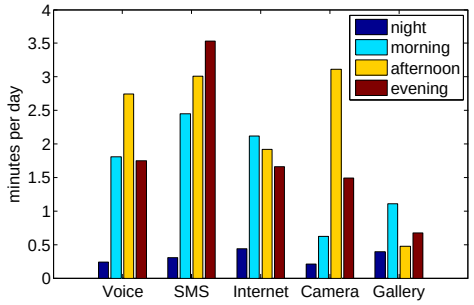

Figure 6: Average usage time per day for each application.

Figure 4 shows the histogram of usage events over these applications. As can be seen, the number of events drops quickly and a few top applications dominate the distribution of events. For this reason, as described in Section 4, we only selected the 5 most popular applications for the analysis.

On average, considering all timeslots, the population makes 5.7 phone calls a day, sends/receives 3.9 SMS, surfs the net less than twice a day, and uses the camera or the gallery less than once a day. On the other hand, although having low numbers in term of events, Internet and Camera have very competitive usage duration compared to Voice and SMS. In term of duration, the users spend 6.5 minutes making voice calls, 9.2 minutes messaging (read and writing SMS), 6.1 minutes surfing the net and checking email, 5.4 minutes using the Camera, and 2.7 minutes using the Gallery. Compared to the published results by Nokia [19] (Voice: 5.8min, SMS:17.8min, Browsing: 3.8min), users in our population spend more time on voice calls, and much less time on SMS. Interestingly, our results on Internet (Browser and Email) usage time is very comparable to the Nokia study, which was based on "advanced smartphone users" [19], which is not the case for many of the users in our population.

Figure 5 and 6 show the average frequency and usage time per user of each application in each timeslot. As can be seen, the statistics varies with respect to the specific timeslot. Afternoon is the most active timeslot. Note that while users send or receive less number of SMS in the evening than in the afternoon (Fig. 5), evening is still the timeslot where users spend most time for messaging (Fig. 6). An explanation might be people usually send long SMS in the evening„ which take more time for writing and reading.

Looking at Camera and Gallery, we see that both of them have similar number of use events for some timeslots, but the usage statistics are much different in the Afternoon slot. Although these two apps are mostly used in the afternoon (in term of frequency), users spend much more time on shooting a photo than looking at the resulting outcome.

We conclude this initial analysis by showing in Figure 7 the correlation matrix between pairs of application-timeslot based on the histogram matrix. In addition to some self-correlations between the same application with different timeslots, we can observe the correlations of 2 pairs of applications that tend to occurs at the same timeslots: Voice-SMS and Camera-Gallery. The Voice-SMS correlation is easy to understand, as people use both modalities to communicate often exchangeably. Interestingly, Gallery does not correlate with itself in another timeslot, but it correlates with Camera at the same timeslot, as users are very likely to open Gallery after using Camera in order to look quickly at the result. Note that this correlation, in general, could be likely increased by creating easy links between the two applications, which is something that many smartphones have already built-in.

The analysis in this section has highlight the key apps being used by our population, and contrasted a few basic findings about phone usage against existing studies. We now show how the data can be further mined to discover daily patterns automatically.

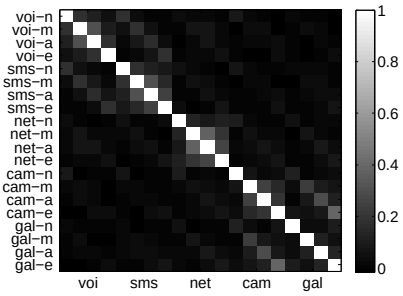

Figure 7: Correlation matrix between application-timeslot.

7.2 Emergent usage patterns

An important parameter of topic modeling is the number of topics (i.e. the number of discrete values that the latent variable can take). A small number of topics (e.g. 10) will provide a broad overview of the contents of the collection of data. Setting this parameter to a large number allows to have a more accurate model and produce fine-grain results. In all experiments, the (symmetric) Dirichlet parameter are set $\alpha = 1$ and $\beta = 0.01$.

In order to have a general view of the method's performance, we run firstly the ATM model with $T = 10$ topics. As discussed, these topics correspond to coarse usage patterns of the data. In Figure 8, we illustrate the 10 discovered topics by showing the top 50 most likely days for each topic, ranked by $P(\mathbf{w}_d|z) = \prod_{i=1}^{N_d} \phi_z^{\mathbf{w}_d^i}$. Note that for each application, the 4 columns correspond to each of Night-Morning-Afternoon-Evening timeslots. Usage levels are represented by the color: blue for no-use, cyan for low-use, yellow for middle-use, and red for high-use. Table 2 also shows the corresponding top app-words for each topic.

From the 10 discovered topics, we see that there are some co-occurrences both between applications and between different timeslots for one application. For instance, topic 1 in Figure 8 and Table 2 corresponds to days of people who use both Voice and SMS, and who at the same time are not likely to use other applications. This corresponds to days (and users) where the phone is predominately used for communication. Interestingly, while both Voice and SMS are almost never used in the evening (6pm-midnight), they are sometimes used late at night (after mid-night). Similar to topic 1 is topic 7, where Voice and SMS are both used in the morning and afternoon again. But here, the co-occurrence of the two basic phone applications is more solid. Furthermore, we do not observe any occurrence of Voice or SMS during the period 6pm-6am (evening-night) in the top days.

Since Voice and SMS are the most common applications in our data (see Figure 5), it is not surprising that ATM discovered many topics that describe the way of using these two applications. Topic 3 is strongly dominated by SMS use with high-level usage in most timeslots (except the night). This reflects the behavior of texter users (e.g. young people), who communicate mainly by SMS. Complementary to topic 3, topic 10 represents the habit of making direct phone calls

Topic 1		Topic 2		Topic 3		Topic 4		Topic 5	
w	$P(w\|z)$	w	$P(w\|z)$	w	$P(w\|z)$	w	$P(w\|z)$	w	$P(w\|z)$
voi-m-low	0.046	gal-e-low	0.046	sms-a-high	0.068	net-m-low	0.032	voi-a-high	0.028
sms-n-low	0.042	cam-a-low	0.046	sms-m-high	0.060	net-e-low	0.032	voi-m-low	0.024
voi-n-low	0.030	cam-e-low	0.039	sms-e-high	0.041	sms-m-low	0.029	voi-e-high	0.023
sms-a-low	0.028	gal-a-low	0.036	sms-a-med	0.029	voi-e-low	0.029	sms-a-low	0.020
voi-a-med	0.027	cam-m-low	0.031	voi-a-high	0.023	net-a-low	0.029	voi-e-med	0.020
net-a-no	0.092	net-m-no	0.069	net-n-no	0.056	cam-e-no	0.059	net-m-no	0.059
net-e-no	0.067	net-e-no	0.069	cam-m-no	0.053	gal-a-no	0.056	gal-m-no	0.056
gal-n-no	0.064	net-a-no	0.060	voi-n-no	0.051	cam-a-no	0.055	cam-m-no	0.054
net-m-no	0.059	voi-n-no	0.054	cam-n-no	0.046	cam-m-no	0.053	net-a-no	0.052
gal-e-no	0.054	sms-n-no	0.052	gal-n-no	0.045	gal-n-no	0.052	cam-n-no	0.051
Topic 6		Topic 7		Topic 8		Topic 9		Topic 10	
w	$P(w\|z)$	w	$P(w\|z)$	w	$P(w\|z)$	w	$P(w\|z)$	w	$P(w\|z)$
net-m-low	0.031	voi-m-low	0.034	voi-e-low	0.046	sms-e-low	0.042	voi-a-high	0.074
net-a-low	0.025	voi-a-low	0.034	sms-e-low	0.035	sms-a-low	0.038	voi-m-high	0.067
net-e-low	0.022	sms-m-low	0.028	voi-a-med	0.026	voi-a-low	0.031	voi-n-low	0.021
voi-a-low	0.022	sms-a-low	0.023	sms-m-low	0.023	sms-m-low	0.031	voi-m-med	0.019
net-m-med	0.011	cam-a-low	0.013	sms-a-low	0.023	gal-e-low	0.015	gal-m-low	0.011
sms-e-no	0.069	voi-e-no	0.069	net-e-no	0.067	voi-m-no	0.075	sms-e-no	0.092
sms-a-no	0.053	net-m-no	0.059	gal-a-no	0.061	cam-n-no	0.070	cam-n-no	0.065
sms-m-no	0.053	net-a-no	0.058	gal-e-no	0.061	net-n-no	0.056	sms-n-no	0.058
sms-n-no	0.052	cam-n-no	0.056	voi-n-no	0.059	cam-m-no	0.048	sms-a-no	0.057
voi-e-no	0.049	sms-n-no	0.054	net-a-no	0.057	gal-n-no	0.048	gal-e-no	0.055

Table 2: Top app-words for the 10 discovered topics. Each app-words correspond to a triplet app-timeslot-usagelevel (n: night, m: morning, a: afternoon, e: evening). For each topic, we show the top 5 words whos level of usage is non-zero, and the the top 5 words with null usage level.

rather than sending SMS. Interestingly, we see that when Voice and SMS do not co-occur, the usage level of the single chosen communication method is high. In other words, users who have a preference between Voice and SMS are also very active phone users, using the phone many times to communicate.

Although there are only a few discovered topics related to Internet, Camera, and Gallery, the patterns from these topics are quite interesting. We can see the occurrence of Camera and Gallery are grouped in topic 2, which reflects the fact the people usually use these two applications jointly (i.e. taking a picture with Camera then looking it via the Gallery). Although it is not entirely clear why, Camera and Gallery active users seem to use more Voice than SMS. This suggests a user type who takes pictures and calls (in other words, who seems to rely on the use of more traditional media), but who does not SMS or surfs the net (which can be seen as more modern media types). Among people who use Internet, the ATM model finds two different behaviors:(i) joint use of Voice, SMS and Internet in Topic 4, and (ii) use of Internet in topic 6.

Topic models assume that observations are a mixture of topics. In other words, one day may be generated from various topics and a user may have different behaviors. Figure 9 illustrates the relation between user and topic by showing the probability of observing topic given user $P(z|a)$ for all users in our dataset. From this plot, we see that most users are dominated by a few topics, which suggests that user behavior is to some degree predictable. The figure also shows that topic 5 is the most popular topic, while topic 2 occurs infrequently (which reflects the fact that there is less frequent use of Camera/Gallery). The topic distribution for each user can also be used to estimate the relation between topics. For instance, looking at the correlation matrix between topics over the set of users, we found a correlation of 0.3 (the highest one) between topic 1 and 3.In Table 3, we show the top 5 users of each topic, ranked by $P(z|a)$. One can see this feature of ATM as producing a "soft" segmentation of the population based on their single most likely topic. Interestingly, the lists of highest rank users in 10 topics are mainly disjoint sets which suggests that the topics are actually capturing trends existing in different segments of the population. On the other hand, the small set of users that were highly ranked in more than one topic are marked in bold (users ids 67,68,76,111). For instance, user number 67 has strong probabilities for both topic 1 and 10. Since the total probability of having topic 1 or 10 is 0.75, this user is quite probably a classical phone user who use the phone mainly for communication.

Discovering more topics. The 10 discovered topics shown before are meaningful and easy to visualize. To discriminate between finer usage patterns, one could need a larger number of topics to be able to accommodate more patterns. In order have a closer view, we run the ATM model with $T = 100$ topics.

In Figure 10, we illustrate some topics by again showing the top 50 days in the collection of data. Generally, the discovered topics in this setting are specifications of the more global topics in Figure 8. For instance, the three topics 58, 67, and 82 are all characterized by the use of SMS in

Topic 1		Topic 2		Topic 3		Topic 4		Topic 5		Topic 6		Topic 7		Topic 8		Topic 9		Topic 10	
a	$P(z\|a)$	a	$P(z\|a)$	a	$P(z\|a)$	a	$P(z\|a)$	a	$P(z\|a)$	a	$P(z\|a)$	a	$P(z\|a)$	a	$P(z\|a)$	a	$P(z\|a)$	a	$P(z\|a)$
67	**0.34**	31	0.31	109	0.48	7	0.45	72	0.54	85	0.62	37	0.48	14	0.38	18	0.33	**67**	**0.41**
103	0.29	54	0.28	**111**	**0.41**	81	0.41	95	0.53	34	0.59	2	0.45	20	0.37	28	0.33	105	0.40
111	**0.26**	**76**	**0.25**	100	0.36	69	0.40	50	0.52	39	0.46	12	0.43	46	0.36	73	0.30	**68**	**0.36**
104	0.25	108	0.25	106	0.31	78	0.33	55	0.49	**76**	**0.40**	30	0.40	15	0.31	65	0.30	87	0.27
101	0.24	**68**	**0.21**	74	0.31	48	0.30	60	0.47	35	0.34	36	0.38	45	0.29	9	0.29	93	0.25

Table 3: Top 5 users of the 10 representative topics. Users that have high ranks for multiple topics are marked in bold. The probability that user a has a behavior described by topic z is also shown.

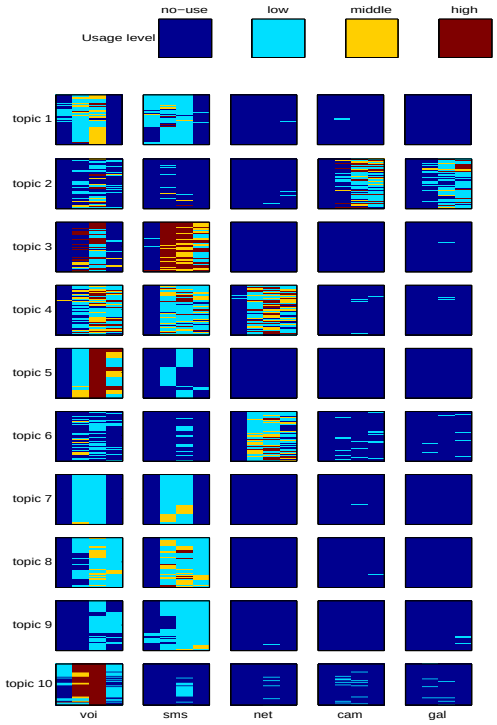

Figure 8: The 10 representative usage patterns of the considered population. Each topic is illustrated by the top 50 likely days. Days (rows) are represented as usage levels (no-use:blue, low-use:cyan,middle-use:yellow, high-use:red) of the 5 applications at the 4 timeslots. See text for more detail.

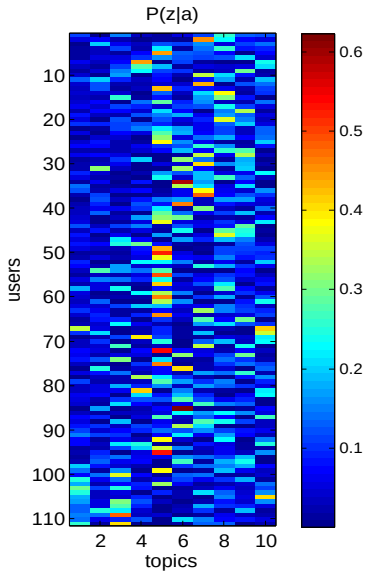

Figure 9: Topic distribution for each user $P(z|a)$. The x-axis indicates topic number, the y-axis is the user index.

the morning and in the afternoon, with 3 different usage patterns: using of Voice in the morning and in the afternoon, using Voice more in the morning and using voice more in the afternoon. Similarly, active internet usage throughout the day is now shown in two topics 2 and 81, with different usage levels (low-use and high-use).

The ATM with large number of topics also discovered new patterns. An example is the joint use of all the 5 applications in topics 64. This refers to very active users who exploit many of the available functionality of the phone. These experiments highlight a current open issue in topic modeling: the selection of the optimal number of topics. One solution is by cross-validation as an objective task (e.g. Section 7.3) or by conducting perplexity experiments in hold-out data.

7.3 User retrieval

The learned ATM was used for discovery and visualization of latent topics of the data, but it can also be used for inferring relevant users given one or more days of usage. The idea is to rank users based on the posterior probability $P(a|\mathbf{w}_q)$ (where $\mathbf{w}_q = \{\mathbf{w}_d\}_{d \in q}$ stands for the union of days of data in the query q) in order to obtain the top relevant users who are most likely to be the owner of these days of data. Note that our main goal is to find relevant users (forming a category) rather than identifying users from query.

Using Bayes' theorem, the posterior probability of a user given a day (or some days) of observation is $P(a|\mathbf{w}_q) = P(\mathbf{w}_q|a)P(a)/P(\mathbf{w}_q)$. Assuming that the prior distribution $P(a)$ is uniform, the posterior probability is proportional with $P(\mathbf{w}_q|a)$ which can be estimated easily using the learned ATM (Cf. Equation 2). Note that inferring a user from multiple days of usage data should be easier than from only one day. The more days we have in the query, the more accurate the retrieval result could be.

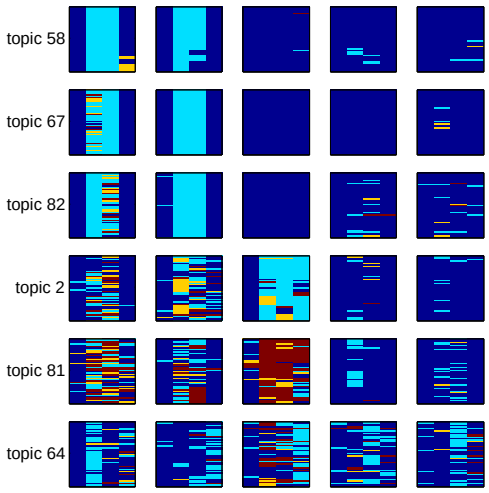

Figure 10: Some fine usage patterns discovered by ATM with $T = 100$ topics.

Author Topic Model (ATM)				
query days	Top 1	Top 5	Top 10	Avg. rank
1	11.7±0.7	31.1±1.1	45.1±1.2	21.3±0.3
2	19.6±1.2	44.3±1.4	60.2±1.9	14.4±0.4
4	31.0±1.8	62.3±1.8	76.8±1.5	8.6±0.3
8	49.3±4.4	77.7±2.6	88.4±2.3	4.5±0.3
Multinomial distribution (MULT)				
query days	Top 1	Top 5	Top 10	Avg. rank
1	11.4	31.9	46.6	22.3
2	19.6	46.6	59.0	16.6
4	34.7	61.3	72.5	11.5
8	47.7	70.3	82.0	8.1
Dummy model: Random ranking				
	0.9	4.5	9.0	56

Table 4: Precision (Top 1,Top 5,Top 10) and average rank of the true author of the querying days.

The above principle can be applied to any model that defines a distribution $P(\mathbf{w}_q|a)$. We use the Multinomial distribution as a elegant baseline, called MULT in the following discussion, where we have one multinomial distribution $P(w|a)$ per user. Since MULT defines $P(w|a)$ directly, it would be a very competitive method for a user retrieval task, although it is not able to discover patterns as ATM does. Recall that, while the MULT method defines $P(w|a)$ directly based on the counts of app-words of a given user, the ATM relies on an additional latent variable z, and $P(w|a)$ is estimated based on $P(z|a)$ and $P(w|z)$. Both models use the independent assumption, i.e. $P(\mathbf{w}_q|a) = \prod_{w \in \mathbf{w}_q} P(w|a)$.

For the evaluation, we separated the data set into a training set and a holdout test set. The test set consists of 8 days (randomly selected) for each user, which results 888 days in total. The rest of the data is used as training set. We perform the user retrieval task with various size of the query, from 1 to 8 days. For instance, we have 888 queries of 1 day, 444 queries of 2 days and 111 queries of 8 days.

	Group 1	Group 2
Training days per user	22-73	73-226
Average number of days	43.9	120.5

Table 5: The two groups of users used for experiments about the effect of size of training data.

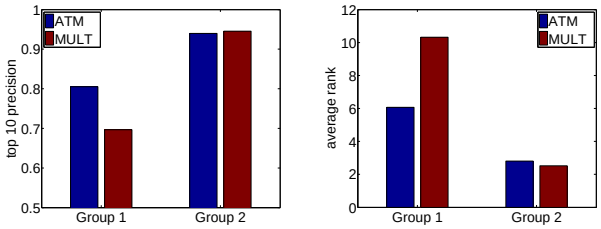

Figure 11: Results of different groups of users. Left: Precision of top 10 user prediction task. Right: Average rank of the correct user.

Table 4 reports the results of the retrieval task with two criteria: (i) the precision that the true owner of the query is in the top 1, top 5 and top 10 list, and (ii) the average rank of the true owner for the set of queries. These are standard measures in information retrieval [1]. The ATM results are reported as mean and standard deviation over 10 runs with different initializations. This is needed as we use MCMC sampling to learn the ATM model, and so statistical fluctuations can be expected.

As expected, we get better results with larger queries (higher precision and lower rank) for both ATM and MULT. Using 8 days of usage data for querying, ATM has a precision of 88% that the true author is on the top ten, which is relatively high given that the performance of random ranking is 9%.

The ATM outperforms the MULT method in most cases for precision and all cases for average rank, and the difference is large when size of query is large. This suggests that the ATM model efficiently exploit the rich information of user behavior in order to perform the user retrieval task, while the MULT model is not as good. Note also that the difference of performance between ATM and MULT also varies between top 1, top 5 and top 10 precision results. The top 1 results between ATM and MULT are quite similar, while ATM outperforms significantly MULT for top 10 results, which suggests that MULT and ATM generate different ranking strategies.

Influence of sample size in training data. To investigate this issue, we divide the set of 111 users into 2 groups based on the number of available training days. The first group consists of people who have lower number of training days, and the second group consists of people who have larger number of training days. Table 5 gives more details on the 2 groups of users.

Figure 11 shows results of user-retrieval for ATM and MULT on each of the two groups. Again, we see that both models perform better when we have more data, but now it is the

size of training data rather than testing data. It is interesting to note that the performance of ATM and MULT are similar for the group with more training data, while ATM outperforms MULT significantly for the group with less training data. This suggests that ATM generalizes better than MULT. An explanation is that while MULT can be viewed as having one separate topic per user, ATM has a set of common topics between users allowing sharing data between users to learn latent topics. In other words, ATM generalizes better since it avoids over-fitting by taking into account the behavior of other users when learning model parameter for given user. In real application of user analysis, this advantage of ATM can bee very important when applying to new users who might have limited data available.

8. CONCLUSION

We presented a framework for mining large-scale patterns of mobile phone usage. We showed that the proposed bag-of-apps model can be integrated with the Author Topic model in order to discover meaningful usage patterns. The learned probabilistic model can also be used for user retrieval, for which we report an extensive objective evaluation. Our analysis confirmed with a huge dataset, intuitive usage patterns. Atypical patterns (in principle not so intuitive) could potentially be discovered by a similar methodology. We plan to investigate this in future work.

The proposed framework can be extended to analyze additional applications and exploit other features (e.g. event duration or the location on which the specific apps were used). Another direction is a comparative analysis of user populations based on the learned topic representation using well defined measures like entropy. This is also part of future work.

9. ACKNOWLEDGMENTS

This work was funded by Nokia Research Center Lausanne (NRC) through the LS-CONTEXT project. We thank Jan Blom and Juha K. Laurila (NRC) for insightful discussions.

10. REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. B. Srivastava. Participatory sensing. In *WSW at Sensys*, pages 117–134, 2006.

[4] F. D. Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3):319–339, September 1989.

[5] N. Eagle, A. S. Pentland, and D. Lazer. Mobile phone data for inferring social network structure. In H. Liu, J. J. Salerno, and M. J. Young, editors, *Social Computing, Behavioral Modeling, and Prediction*, pages 79–88. Springer US, 2008.

[6] N. Eagle and A. (Sandy) Pentland. Reality mining: sensing complex social systems. *Personal Ubiquitous Computing*, 10(4):255–268, 2006.

[7] K. Farrahi and D. Gatica-Perez. What did you do today?: discovering daily routines from large-scale mobile data. In *Proc. MM*, pages 849–852, New York, 2008.

[8] W. R. Gilks. *Markov Chain Monte Carlo In Practice*. Chapman and Hall/CRC, 1999.

[9] I. Guyon, V. Lemaire, M. Boullé, G. Dror, and D. Vogel. Design and analysis of the kdd cup 2009: fast scoring on a large orange customer database. *SIGKDD Explor. Newsl.*, 11(2):68–76, 2009.

[10] T. Hofmann. Probabilistic latent semantic indexing. In *Proc. SIGIR*, pages 50–57, New York, 1999. ACM.

[11] S.-J. Hong, K. Y. Tam, and J. Kim. Mobile data service fuels the desire for uniqueness. *Communications of the ACM*, 49(9):89–94, 2006.

[12] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proc. SIGIR*, pages 119–126, New York, 2003. ACM.

[13] N. Kiukkonen, J. Blom, O. Dousse, D. Gatica-Perez, and J. Laurila. Towards rich mobile phone datasets: Lausanne data collection campaign. In *Proc. ICPS*, Berlin, 2010.

[14] R. Kwok. Phoning in data. *Nature*, pages 959–961, 2009.

[15] D. Lazer, A. Pentland, L. Adamic, S. Aral, A. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Guttmann, T. Jebara, G. King, M. Macy, D. Roy, and M. V. Alstyne. Computational social science. *Science*, 323(5915):721–723, 2/6/2009 2009.

[16] R. Mika, O. Antti, and E. Nathan. Smartphones: An emerging tool for social scientists. *Sociological Methods Research*, 37(3):426–454, 2009.

[17] Nielsen. Critical mass: The worldwide state of the mobile web. Available at `http://www.nielsenmobile.com/documents/CriticalMass.pdf`, July 2008.

[18] Nokia. Smartphone 360 study: Responding to what the user wants and needs. Press Release, Oct. 2007 available at `http://www.nokia.com/NOKIA_COM_1/Press/twwln/press_kit/Nokia_Smartphone360_study_Press_Backgrounder_October_2007.pdf`.

[19] M. Raento, A. Oulasvirta, R. Petit, and H. Toivonen. Contextphone: A prototyping platform for context-aware mobile applications. *IEEE Pervasive Computing*, 4(2):51–59, 2005.

[20] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proc. UAI*, pages 487–494, Arlington, Virginia, United States, 2004. AUAI Press.

[21] H. Verkasalo. Handset-based analysis of mobile service usage. Doctoral dissertation, Helsinki University of Technology, 2009.

[22] H. Verkasalo, C. López-Nicolás, F. J. Molina-Castillo, and H. Bouwman. Analysis of users and non-users of smartphone applications. *Telematics and Informatics*, 27(3):242–255, 2010.

[23] B. Zheng, D. C. M. Jr, and X. Lu. Identifying biological concepts from a protein-related corpus with a probabilistic topic model. *BMC Bioinformatics*, 7:58, 2006.