Learning Routines Over Long-Term Sensor Data Using Topic Models

Federico Castanedo & Diego López-de-Ipiña
Deusto Institute of Technology, DeustoTech. University of Deusto
{fcastanedo,dipina}@deusto.es

Hamid K. Aghajan
Stanford University
aghajan@stanford.edu

Richard Kleihorst
Xetal NV and Ghent University
richard@xetal.eu

Abstract

Recent advances on sensor network technology provide the infrastructure to create intelligent environments on physical places. One of the main issues of sensor networks is the large amount of data they generate. Therefore, it is necessary to have good data analysis techniques with the aim of learning and discovering what is happening on the monitored environment. The problem becomes even more challenging if this process is done following an unsupervised way (without having any a priori information) and applied over a long-term timeline with many sensors. In this work, topic models are employed to learn the latent structure and the dynamics of sensor network data. Experimental results using two realistic datasets, having over fifty weeks of data, have shown the ability to find routines of activity over sensor network data in office environments.

Keywords: Unsupervised Learning, Sensor Networks, Topic Models.

1 Introduction

Sensor networks are a key element in future Ambient Intelligence (AmI) and smart environments applications since they provide a mechanism to gather data from the monitored environment. As an example of AmI and smart environments scenarios we can consider current modern buildings equipped with a huge number of cheap sensors, such as Passive Infra-Red (PIR) sensors. An important research question in this scenario is how to transform the obtained raw data from the sensors into valuable knowledge? As an example of valuable knowledge we can think about the benefits of smart buildings which reduce energy consumption by learning the behaviour patterns of their users.

A large number of works have been published in the past years about activity recognition. Experiments and obtained conclusions of most previous works are based only on short-term real data or simulation results. By short-term real data we refer to experiments carried out using only some hours or days of real data. This means that, in some cases, the obtained conclusions from short-term data may not apply when the system employs long-term data. Because the model is learnt without enough data. However, in this work, we follow a novel approach and provide an insight on discovering routines over long-term, using more than fifty weeks of real sensor data. The proposed method is completely unsupervised and only uses binary motion sensors to build a user behaviour model.

Several data analysis methods are available to process data streams generated by sensor networks (Pauwels, Salah & Tavenard 2007) (Sigg, Haseloff & David 2010). However, it becomes clear that is necessary to have a good technique to mine the generated data and infer the latent structure or the underlying common activities/routines (Bettini, Brdiczka, Henricksen, Indulska, Nicklas, Ranganathan & Riboni 2010).

From our point of view, this technique should have the following three requirements: (i) to follow

an unsupervised way of learning, that is, learning data without a priori associated target values, (ii) since the amount of data generated in the network is usually large it must represent high dimensional data in a low dimensional space (dimensionality reduction) and (iii) have the power to express the obtained latent routines in terms of probability distributions over activities, giving a more expressive power than a classical hard clustering.

To support the previous requirements, in this work we made an analogy between a text document and the obtained sensor data over a period of time. Therefore, we first cast the problem of detecting activities into a topic model inference algorithm where a topic will be a concept similar to a user routine composed of several activities and second we solve it using well-known topic models techniques. Topic models are generative probabilistic models for discovering the underlying semantic structure of a collection of documents using a hierarchical Bayesian model. These models are based on the idea that documents are mixtures of topics, where a topic is a probability distribution over words.

One of the first statistical models employed to obtain semantic information from a word-document co-occurrence matrix was Latent Semantic Analysis (LSA) (Landauer & Dumais 1997). In LSA, words and documents are represented as points in Euclidean space. In contrast, in topic models the relations between words and documents are expressed in terms of probabilistic topics. Hofmann introduced the probabilistic Latent Semantic Indexing method (pLSI) (Hofmann 1999) (Hofmann 2001), which does not make any assumption about how the mixture weights are generated, making difficult to classify new documents when the model is trained. Blei, Ng & Jordan (2003) extended the pLSI model by introducing Dirichlet priors as the underlying statistical distribution of topics and words, resulting on the Latent Dirichlet Allocation (LDA) model.

We can summarise the main contributions of this work as follows:

1. We propose an unsupervised method to learn user routines in office environments. This method has been usually employed for extracting topics in text documents and here is applied to sensor data. The nature of this scenario allows us to obtain behaviour patterns because these types of working activities usually follow similar patterns.

2. The proposed method is evaluated in realistic scenarios, using two different real datasets having data over fifty weeks.

3. Different algorithms and implementations for training the model are employed and their results are reported.

The work presented here follows a novel approach, since the sources of information are sensor measurements which are grouped to form artificial words (low level activities) and documents (a set of activities over time). The set of documents generates a corpus which represents the dynamic behaviour of the monitored environment.

The remainder of this paper is structured as follows. The next section provides a review of related works. Then in section 3 presented the applied topic model method over sensor network data. Section 4 presents experimental results carried out using two realistic data sets, the Innotek dataset (section 4.1) and the Mitsubishi Electric Research Labs (MERL) dataset (section 4.2). Finally the conclusions of this work are given in section 5.

2 Related works

Since the beginning, topic models have been primarily applied over text corpora. Griffiths & Steyvers (2004) used the LDA algorithm to analyse abstracts from the Proceedings of the National Academy of Sciences (PNAS) and automatically extract the topics. In contrast to the original LDA model which uses variational inference they employed a Markov Chain Monte Carlo (MCMC) inference model.

Beyond their natural language processing roots, the LDA model and their variants have also been applied to other domains, such as the following.

2.1 Computer vision

In the computer vision domain, the LDA model has been used to automatically infer the images' categories and understand the content of the images (Sivic, Russell, Efros, Zisserman & Freeman 2005) (Horster, Lienhart & Slaney 2007). It has also been employed for learning human action categories in

videos (Niebles, Wang & Fei-Fei 2008). An extension of the LDA model taking into account interactions between the activities was presented by (Wang, Ma & Grimson 2009). They use low-level features (moving pixels) extracted from video sequences to learn the proposed LDA mixture model. Hospedales, Li, Gong & Xiang (2011) use topic models to identify rare behaviours in video. They represent video features by computing an optical flow vector for each pixel and generating a codebook. The problem is very challenging since the number of rare behaviours examples is really low.

2.2 Activity modelling

Activity modelling in general has been considered in the work of Georgeon, Mille, Bellet, Mathern & Ritter (2011). They constructed a symbolic abstract representation of an activity from an activity trace. Their work is mostly focused on the visualisation issues and is based on ontology modelling instead of using probabilistic techniques.

Huynh, Fritz & Schiele (2008) considered daily routines as a probabilistic combination of activity patterns and used LDA to detect those patterns. To evaluate their work they collected data from one subject over 16 weekdays (28 hours total due to sensor failure) using wearable sensors. They manually annotated the activities and, after filtering out rare activities, got 34 classes. Their sensor data were acquired using wearable sensors instead of PIR sensor which we consider to be less intrusive. Moreover their approach was tested on a short-term scenario, that is, using only seven days of real-world activity data. In contrast, the presented work is tested over long-term sensor data.

Farrahi & Gatica-Perez (2008) (Farrahi & Gatica-Perez 2011) used LDA and a modification of the basic LDA model, the Author-Topic Model (Rosen-Zvi, Griffiths, Steyvers & Smyth 2004) to discover daily location-driven routines from a massive dataset of mobile phones user's location. In this work, we have followed a similar approach to build the *bag-of-words* representation of the documents. Their experiments focused on data which have different locations from each user, instead this work models different patterns of sensor activations at fixed locations. Their results also support the advantage of using LDA for discovering routines on large amount of data.

Phung, Adams, Tran, Venkatesh & Kumar (2009) investigate the use of a LDA model to infer, following an unsupervised way, the sequences of places which the user visits periodically. Experimenting with the discovery of routines for a user over the course of one month period and taking into account the user context, they cast the motion state detection problem as an unsupervised incremental clustering problem.

Ferrari & Mamei (2011) applied the LDA topic model to obtain routine behaviours from mobility data collected from Google Latitude. They recorded daily whereabouts of two persons over a period of almost a year. The most common places from the acquired data are manually labelled by the users, resulting in 10 and 12 relevant places for each user. Information of each day is divided into 48 time slots of 30 minutes duration each. Their *bag-of-words* representation is composed by a sliding window of three symbols (30 minutes duration each) plus one of eight different time labels. In their experiments they employed a fixed number $K = 30$ of topics or routines.

Rashidi, Cook, Holder & Schmitter-Edgecombe (2011) use an unsupervised method and identify frequent activities in home environments to detect changes in patterns and lifestyle. They combine sequence mining and clustering algorithms to identify frequent activities and cluster similar patterns together. Their method uses the Levenshtein distance to compute the similarity measure between two patterns. Finally, to cluster sequences into group of activities the method uses a standard K-Means algorithm and compute the number of edit operations that are required to make two activities equal.

In the context of office buildings, the work of Castanedo, López-de-Ipiña, Aghajan & Kleihorst (2011) and Castanedo, Aghajan & Kleihorst (2011) presents a LDA topic model to infer routines over real world sensor network data (Innotek dataset). In those works, instead of having activations of continuous sensors as in the MERL dataset, authors dealt with minutes of occupancy at indoor rooms. This paper, extends previous works by using different words length and showing more experimental results with an extended version of the previous Innotek dataset plus the use of MERL dataset.

2.3 MERL dataset

Existing works employed the MERL dataset but with different techniques. Connolly, Burns & Bui (2008), used the MERL dataset with the aim of modelling three different social features: (1) visiting another person, (2) attending meetings with another person and (3) travelling with another person. Their work is based on information theoretic measures (entropy) and graph cuts to obtain the previous patterns on the MERL dataset. To extract relationships on the occupancy data they model pairwise statistics over the dataset and their goal was to identify potentially important individuals within the organisation.

Salah, Pauwels, Tavenard & Gevers (2010), reviewed the existing techniques for the discovery of temporal patterns in sensor data and proposed a modified T-Pattern algorithm (Magnusson 2000). This algorithm was tested using the MERL dataset and the presented results outperformed the Lempel-Ziv compression based methods.

3 Employed method

In contrast to other mixture of unigrams models the LDA model allows documents to represent multiple topics with different degrees and each topic is also represented using a distribution of words. This probabilistic multinomial representation provides an enrichment activity model.

Our proposed method is based on mapping the obtained discrete sensor data into artificial words, combining them into documents and therefore generating a corpus from the obtained signals of the environment. An important question here is to define what a document will be and what kind of information gathered from sensor measurements corresponds to the words of the documents. That is, how the process to build the vocabulary and the corpus is followed.

In this work, the observed data refer to sensor activations over time which are then transformed into artificial words (see Figure 2). A set of words of one particular sensor, that is a set of activations from that sensor over a period of time (i.e. one day) is established as a document. Finally, the set of generated documents over the complete history are known as the corpus.

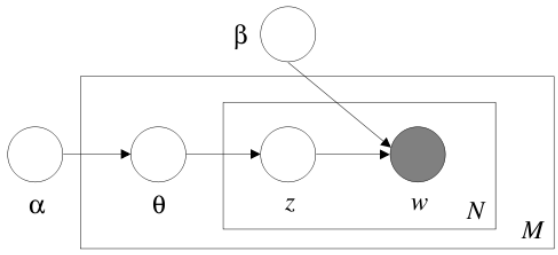

Figure 1: Graphical representation of the LDA model. Each node represents a random variable, edges represent possible dependence between variables and plates denote replicated structure. The LDA model only observes words (shaded w random variable) of each document. Variables β and θ are drawn from a Dirichlet distribution. N represents the number of words and M the number of documents.

The task is to learn both what the topics are and which documents employ them in which proportions. It is important to note that the only observed information are the generated words, thus this is a classical unsupervised machine learning problem.

All topic models presume that a document is a mixture of topics, but they differ on the statistical assumptions over the observed data.

3.1 Overview of the LDA topic model

The LDA topic model is a generative and unsupervised probabilistic machine learning algorithm, that assumes Dirichlet priors as the underlying statistical distribution of topics and words. The Dirichlet distribution is the conjugate prior of the multinomial distribution. Let α be the prior distribution and θ the posterior one, this means that the posterior distribution $prob(\theta|\alpha)$ (in this case the multinomial distribution) and the prior distribution $Dir(\alpha)$ are in the same family which is also known as conjugate distributions. This issue provides an easy numerical integration when applying the Bayes theorem to perform an inference in the model.

A graphical representation of the LDA topic model is given in Figure 1, where words are the only observed random variables (shaded circle) and the basic unit of the model. Each word w comes

from a predefined vocabulary V and is generated by the state of the sensor over a short time period (section 3.3 gives more details).

As we mention before a document is defined as a distribution over topics $prob(Document|Topics)$, so each document is represented as a random mixture over latent topics. Each topic is also defined as a distribution over words $prob(Topic|Words)$. Being N the number of words in V and Z the topic variable with K number of topics, LDA assumes the following generative process for each document in a corpus:

1. Choose $N \sim Poisson(\xi)$.
2. Select $\theta \sim Dirichlet(\alpha)$.
3. For each of the N words W_n:
   (a) Choose a topic $Z_n \sim Multinomial(\theta)$.
   (b) Sample a word W_n from the probability distribution $p(W_n|Z_n, \beta)$, a multinomial probability conditioned on the topic Z_n.

The dimensionality of the topic variable Z and therefore the dimensionality of the Dirichlet distribution defined as K is assumed known and fixed. Also, the word probabilities are parameterised by the $K \times |V|$ matrix β where $\beta_{ij} = p(w^j = 1|z^i = 1)$. This matrix β represent a set of quantities to be estimaed by the model during the training process.

A posterior distribution $(\theta_1, ..., \theta_k)$ is known to be Dirichlet distributed:

$$(\theta_1, ..., \theta_k) \sim Dirichlet(\alpha_1, ..., \alpha_k) \tag{1}$$

with parameters $(\alpha_1, ..., \alpha_k)$, if has the following probability density:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} ... \theta_k{}^{\alpha_{k-1}} \tag{2}$$

where the parameter α is a k-vector (positive vectors that sum to one) with components $\alpha_i > 0$ and $\Gamma(x)$ is the Gamma function.

Given the parameters α and β. The joint distribution of a topic mixture θ, a set of K topics Z, and a set of N words w, is given by:

$$p(\theta, Z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(Z_n|\theta) p(w_n|Z_n, \beta) \tag{3}$$

Replacing $p(\theta|\alpha)$ from equation (2) into equation (3) we get the following joint distribution of the model:

$$p(\theta, Z, w|\alpha, \beta) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} ... \theta_k{}^{\alpha_{k-1}} \prod_{n=1}^{N} p(Z_n|\theta) p(w_n|Z_n, \beta) \tag{4}$$

Given D documents that could be expressed with K topics over W words, it could be possible to represent $Prob(Words|Topic)$ with a set of K multinomial distributions over the W words. $Prob(Words|Topic)$ represents the probability distribution over words W given a $Topic$. These distributions are encoded into the $\beta_{1:k}$ matrix for each one of the K topics, where each topic is a multinomial distribution ϕ_k over the word vocabulary V. LDA assumes a Dirichlet prior distribution on the parameters ϕ_k.

Each document d is also represented as a topic mixture θ_d over K latent topics. Parameter θ_d is a $|D|$xK matrix of document-specific mixture weights for the K topics each of them drawn from a Dir(α) prior distribution with hyper-parameter α.

3.2 Training and building the LDA model

When a generative model is training, the goal is to find the best set of parameters and latent variables that can explain the observed data, assuming that the model actually generates the observed data. When a LDA model is already trained it could be used to infer the topic distribution of new documents given the *Topic-Words* learned distribution. Therefore, trained distributions could be used for classifying new documents or measuring similarities between new and existing documents. Thus, in the LDA model the classification of new documents is performed by using a model already trained.

Parameters α, β and θ from equation (3) provide information about the underlying data (corpus). Parameter α indicates how semantically diverse documents are (in our case daily sensor activations), with lower α values indicating increase diversity. Hyper-parameter $\beta_{1:k}$ provides information about how similar the different topics are, it

gives $prob(Words|Topics)$ for each one of the K topics.

The computation of the parameters α, $\beta_{1:k}$ and θ_d from equation (3) is intractable for exact inference, due to the coupling between θ_d and $\beta_{1:k}$. Therefore in practice, approximate inference algorithms, such as Gibbs sampling, Laplace approximation, Variational Bayes (VB) approximation and Markov Chain Monte Carlo (MCMC) must be employed.

We briefly describe some differences of two common training algorithms. For a good comparison between different inference algorithms that could be used in LDA, we refer the reader to (Asuncion, Welling, Smyth & Teh 2009) and (Mukherjee & Blei 2009).

3.2.1 Variational Bayes (VB)

Variational Bayes inference consists on defining a parametric family of distributions that forms a tractable approximation to an intractable true joint distribution. Blei's original LDA model (Blei et al. 2003) proposed a variational Bayes *Expectation-Maximization* inference algorithm in order to estimate the parameters in the training phase of the algorithm. They suggest a variational Dirichlet distribution γ_i for each document and a variational multinomial ϕ_{ni} over topics for each word position in the document.

Since the number of iterations required for a single document is on the order of the number of words in the document, the algorithm requires $O((N+1)K)$ operations which yield a total number of operations roughly on the order of N^2K, being K the number of topics and N the number of words in the document. To alleviate these computational requirements, parallel versions of Variational EM have been published (Mariote, Medeiros & da Torres 2007).

3.2.2 Distributed Gibbs sampling

Gibbs sampling is a MCMC method that involves iterating over a set of variables $z_1, z_2, .., z_n$ sampling each z_i from $Prob(z_i|z_{\setminus i}, w)$. A complete software architecture to perform inference using distributed Gibbs sampling and the simplifications proposed in (Yao, Mimno & McCallum 2009) was developed by Smola & Narayanamurthy (2010). They made LDA scalable by using a stochastic learning approach (for parameters α and β) and a good technical architecture design. Their architecture scales very easily and makes possible to add more machines in a cluster running a LDA model. This framework was recently released under the Mozilla Public License[1]. We use their code to perform some experiments with a big number of topics and documents.

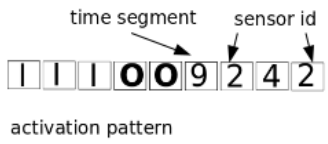

Figure 2: Example of word (IIIOO9242) which indicates an activation pattern of the sensor 242 at time segment 9.

3.3 Generating the corpus

The LDA model is based on the *bag-of-words* assumption, that is, the only information relevant to the model is the word frequencies in each document. It also follows the ex-changeability assumption of each word in the document. This means that words are conditionally independent and identically distributed with respect to the latent parameters (topics). The input to the model is the *bag-of-words* representation of a collection of text documents, where documents D are represented as sparse vectors of $|W|$ non-negative counts from the words of a vocabulary V.

In order to generate the artificial words from the sensor data, we followed a similar approach as (Farrahi & Gatica-Perez 2011) (Castanedo, Aghajan & Kleihorst 2011) (Castanedo, López-de-Ipiña, Aghajan & Kleihorst 2011). Each word will correspond to a five minute activation pattern of one specific sensor. Words are also classified into nine different time intervals according to the time of the day which generates the activity and this information is also encoded in the word. Since each time interval is encoded into the generated word we are providing a meaningful segmentation of daily activities into the model. Therefore each possible word w_n (see Figure 2) of the vocabulary V is composed of

[1] http://github.com/shravanmn/Yahoo_LDA

the following nine digits:

- 5 digits for the sensor activation status. ie. `IIIOO` means that the sensor have been activated for 3 minutes and without being activated for 2 minutes. Each digit corresponds to one minute information.

- 1 digit referring to the the slot of the day, with the following intervals: (1) from 00:00 to 6:00, (2) from 6:00 to 7:00, (3) from 7:00 to 9:00, (4) from 9:00 to 11:00, (5) from 11:00 to 14:00, (6) from 14:00 to 17:00, (7) from 17:00 to 19:00, (8) from 19:00 to 21:00 and (9) from 21:00 to 00:00.

- 3 digits for the sensor identification number.

Table 1: Main characteristics of the employed datasets.

	Innotek	MERL
Start date	March 2010	March 2006
End date	March 2011	Dec. 2007
Total num. of sensors	135	290
Total activations (weekdays)	3.507.357	31.469.913
Num. of weeks	50	90
Vocabulary Size $\|V\|$	38.880	83.520
Num. of documents $\|D\|$	9.140	88.795

4 Experimental results

We performed experiments using two realistic long-term datasets: (i) the Innotek building dataset in Geel Belgium and (ii) the MERL dataset in Cambridge, USA. Table 1 gives a summary of their main characteristics. All the experiments were carried out using a single machine and following batch execution mode. Since LDA is a unsupervised technique we first trained the model and computed the obtained results with held-out data using the trained model without changing the parameters. Thus we are measuring the predictive power of the trained model over unseen data. We followed a 10-fold cross-validation methodology. Perplexity is employed as the measure to evaluate the quality of the trained model and is defined as the reciprocal geometric mean of the likelihood of a test/validation corpus given a model. Formally for M documents, standard *perplexity* measurement is defined as:

$$perplexity = exp(-\frac{\sum_{m=1}^{M} \log p(W_m|LDA)}{\sum_{m=1}^{M} N_m}) \quad (5)$$

where $p(W_m|LDA)$ is the probability of the unseen set of words in document m given the LDA trained model and N_m the number of words in each document.

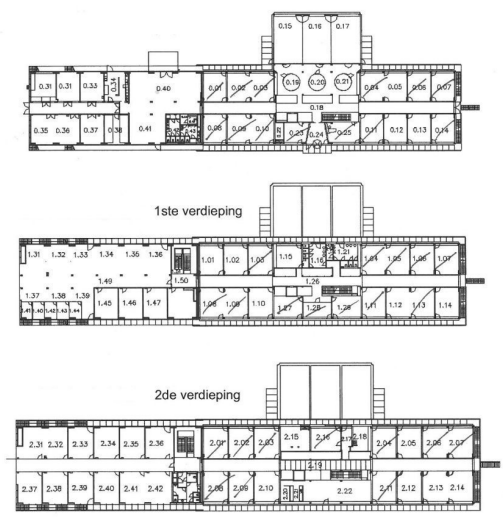

Figure 3: Ground plan of the three floors of Innotek building where the dataset were obtained.

4.1 Innotek dataset

The Innotek dataset was obtained from monitoring the Innotek building in Geel, Belgium (see Figure 3) from March 2010 till March 2011. This three floor building facilitates start-up companies. It contains offices with different uses. In the building we will find Innotek's own administrative offices, such as receptionist, secretarial office, management office, building maintenance office, cantine and restroom facilities. Also, several rooms are rented out to start-up companies, usually 1 to 4 rooms per company. These rooms are mainly functioning as office space to meet prospective customers or guests, so they show a different character of use compared to regularly occupied offices. The top-floor of the Innotek building houses some laboratories where people have to be present to follow up experiments. These spaces also show occupancy at night, during weekends and in holidays.

Data are captured by centrally mounted high-quality PIR-based sensors that are part of the Philips system to install the automatic lighting. Detection of motion will trigger the lights which will remain switched on for at least 10 minutes. If no motion has been detected in this period, the lights will be turn-off again. Innotek has arranged for connecting these sensors to the bus of their Johnson Controls climate management system, which enables their use to also control the room temperature. The climate management system will set the room to *hibernation mode* if no occupancy has been detected for a certain time interval. Once occupancy is detected, the room climate will go to the comfort level as set locally per room by the inhabitants. New office buildings in Flanders will have to meet smart building controls like this in the very near future.

As the Johnson Controls sensor bus can be read out at a central point we have been able to log the data of most rooms of the building on a 1 minute accurate time-scale. The events are based on integration of detections over the 1 minute interval and indicate the occupancy with a high level of confidence. The log files follow this structure:

1. 0 01 0 10 3 12 6 13 37
2. 0 01 1 10 3 12 6 14 33
3. 0 01 0 10 3 12 6 14 48
4. 0 01 1 10 3 12 6 15 5
5. ...

The first line means that sensor 1 of floor 0 provides no occupancy information (0 at third column) the day 12 of march 2010, which has a type day of 6 (meaning it was Friday) at the time 13:37. In the second line, the same sensor provides occupancy information (value 1 at third column) at time 14:33, line three shows that the room was empty again at time 14:48 and so on.

Using only binary occupancy with 1 minute frequency we construct the documents, one document for each room and day, which are sequences of words over 5 minutes interval. Different word lengths could also be employed, for instance in previous work (Castanedo, Aghajan & Kleihorst 2011) we used a word-length of 30 minutes. The vocabulary is composed of 38800 possible words ($2^5 * 9$ time segments * 135 sensors) and the corpus contains 9140 documents. Employed data are publicly available for research purposes and can be retrieved from `http://fcastanedo.com/?page_id=126`.

4.1.1 Distribution of sensor activations over time

As a first step we perform an exploratory data analysis to support our assumption that people follow behaviour patterns in office environments. We show a high level grey-scale occupancy picture of all the rooms during the 50 weeks in Figure 4. In the picture black cells indicates a higher frequency occupancy than the white ones and vertical lines indicate the end of each week day. It can be shown that patterns appear in the middle times of each day. These patterns can be seen more clearly in a video, which plots the obtained frequencies on the building ground-plane, posted online at `http://bit.ly/URTuYx`.

In Figure 5 the distribution of sensor activations in each time interval for the Innotek (50 weeks) and MERL (90 weeks) datasets is shown. It can be noted that for time interval 5 and 6 (from 11:00 to 17:00) this distribution provides the most density, as it was expected. Despite the distributions are very similar some slightly differences can be appreciated. There are more MERL sensor activations at intervals 7,8, and 9, in contrast, Innotek provides higher probabilities at intervals 3 and 4. This may

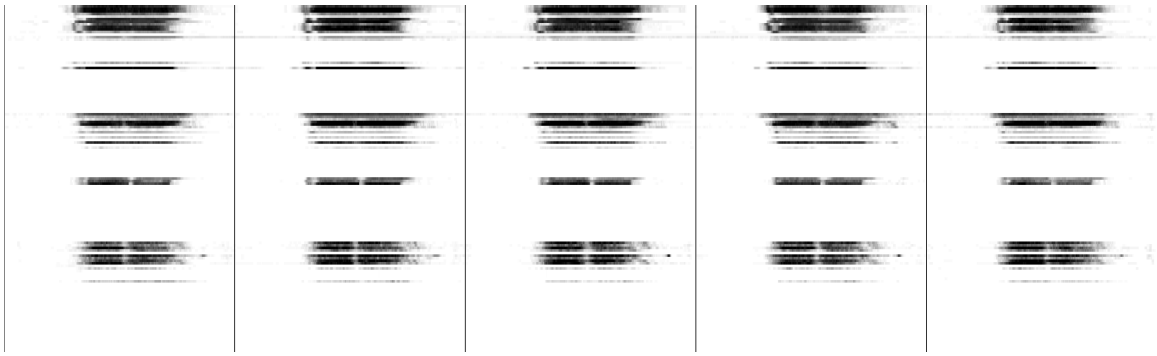

Figure 4: Grey-scale image that represents the distribution of sensor activities over the complete 50 weeks, that is the probability of each sensor being activated at each of the 7200 minutes of the week (with black being most likely). Rows represent sensors and columns time evolution in one minute frequency. The hight-lighted vertical lines represent the end of each day.

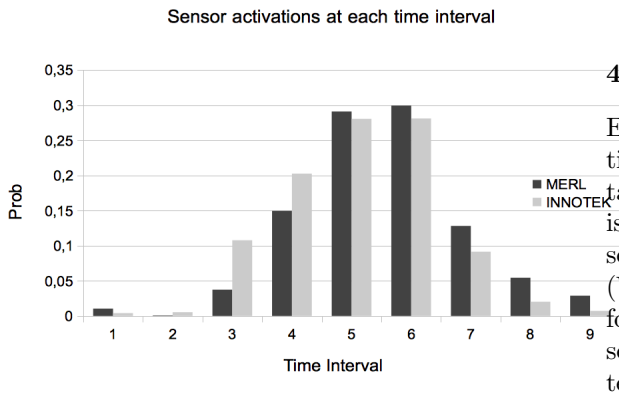

Figure 5: Distribution of sensor activations at each time interval for Innotek (50 weeks) and MERL (90 weeks) dataset. Time Intervals correspond to: (1) from 00:00 to 6:00, (2) from 6:00 to 7:00, (3) from 7:00 to 9:00, (4) from 9:00 to 11:00, (5) from 11:00 to 14:00, (6) from 14:00 to 17:00, (7) from 17:00 to 19:00, (8) from 19:00 to 21:00 and (9) from 21:00 to 00:00.

indicate that in MERL building employees usually spend more time working at the end of the day (maybe for international meetings) than Innotek building users. On the other hand the odds of having people from 7:00 to 9:00 in the building are higher in Innotek than MERL.

4.1.2 Setting the number of latent topics

Establishing the number of latent topics (or routines) that the model must learn is one important decision when training a topic model. This issue has been attracting the interest of many researches in the machine learning community, see (Wallach, Murray, Salakhutdinov & Mimno 2009) for a good overview. In this work, we performed several experiments increasing the number of latent topics and evaluated the obtained results with the aim of obtaining a good model.

Innotek dataset is composed of words from a vocabulary of 38800 different words which comes from 2^5 combinations times 9 time segments times 135 sensors. We used 50 complete weeks from the Innotek dataset which gives a total amount of 9.140 documents and split them into training and testing sets. A 10-fold cross validation scheme was followed. In total ≈8.200 documents were used for training and ≈900 documents for testing.

We performed experiments measuring perplexity while increasing the number of topics with $K = \{7, 8, 9, 10, 20, 30, 50, 100, 150, 200, 300, 400\}$ and reported the average and standard deviation

of ten different executions in Figure 6. A lower perplexity value indicates a better prediction over held-out data. It can be shown that perplexity values decrease while we increase the number of topics till 200, then with 300 topics start increasing and when the model is trained with 400 topics a big difference is observed. These results suggest that with the employed corpus the underlying Innotek data are better represented with a number of latent topics around 100-200 than using higher numbers. It can be also due to the fact that VB can not converge (with the 50 max iterations established) when so many topics are used. Obtained α results also support these differences on perplexity since for values $K \leq 300$ they are always less than 1 and for $K = 400$ are around 20. Please note that α values indicate how semantically diverse documents are with lower α values indicating a higher diversity.

Experiments were executed using Blei's LDA implementation with the following parameters:

- 50 EM iteration (max) for training.
- Estimation of α parameter from Dir(α) equal for all k: $\alpha_1 = \alpha_2 = ... = \alpha_k$
- Full variational inference with convergence error of 0.000001 for training and inference.
- 1000 EM iterations (max) for inference.

4.1.3 Discovered routines

Some obtained topics and the associated probabilities of the four most likely words are shown in table 2. These topics can be interpreted as follows.

Topic 4 indicates that room 1.02 in the first floor (sensor 42) is occupied with high probability (≈0.53) from 11:00 to 19:00.

We discovered a relation between room 1.02 in the first floor (sensor 42) and room 2.04 in the second floor (sensor 95). Topic 10 is just an example of this relation but other obtained topics modelled the relation between these rooms. This topic represents a common pattern of occupancy followed by 1 minute of no occupancy and then an occupancy again in room 1.02 (word `IIOII6042`) from 14:00 to 17:00 which is related with a pattern (word `IOOII5095`) of two-minute empty room of 2.04 from 11:00 to 14:00. From the obtained

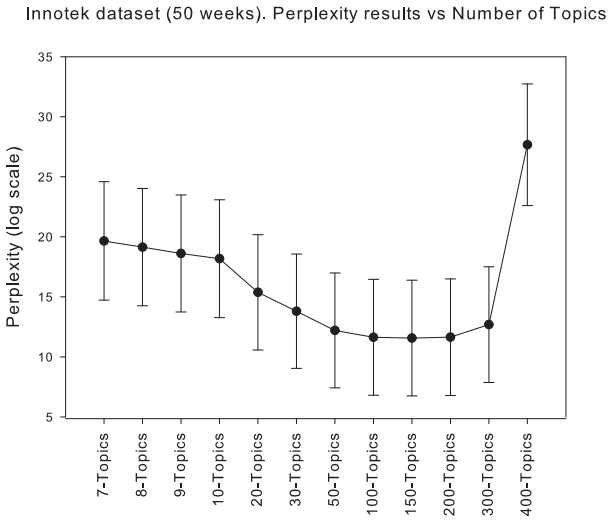

Figure 6: Perplexity results of Innotek dataset trained with VB. The average and standard deviation of 10 different executions are plotted. Lower perplexity indicates better prediction over held-out data.

topics, it seems that these rooms had very dynamic occupancy behaviours.

Topic 27 models room 0.06 (sensor 5) of the ground floor as a room most likely occupied between 9:00 and 14:00 (words `IIIII4005`, `IIIII5005`) and empty between 17:00 and 21:00 (words `OOOOO7005`, `OOOOO8005`).

It has been discovered in topic 62 that room 1.27 (sensor 67) is occupied from 9:00 to 17:00 with high probability (≈ 0.44) while is empty from 19:00 to 21:00 (word `OOOOO8067`). Similar occupation pattern has been represented in Topic 100 for room 0.09 (sensor 8). This topic indicates an occupation from 9:00 to 14:00 with high probability (≈ 0.46) while the room is empty for 19:00 to 21:00.

Topic 33 represents a dynamic pattern of room 1.06 (sensor 46) which is characterised to be occupied, empty during one minute from 7:00 to 9:00 (word `IIIOI3046`) and occupied again. This pattern takes place with a high probability (0.81) and is jointly modelled with an occupation of the same room after some time being empty from 14:00 to 17:00 (word `OOOOI5046`). This last behaviour may be the time when the room is occupied again after lunch.

Room 1.01 located at first floor (sensor 41) provides a very dynamic occupation pattern as can be shown in Topic 66. Between 14:00 and 17:00 it is empty at least for 4 minutes with high probability ($\approx$ 0.49) and it also presents similar occupancy pattern from 8:00 to 11:00 and from 19:00 to 21:00.

In the case of topic 94, it indicates that room 2.06 (sensor 97) has been used also at the end of the day from 17:00 to 21:00 (words `OOOOI8097` and `IOOOI7097`).

Topic 18 indicates that room 0.01 (sensor 0) is mostly often occupied from 9:00 to 19:00, indicates that this room follows a constant behaviour of occupancy every day.

Finally, topic 35 represents the occupancy pattern of room 2.02 (sensor 93) which is most likely empty from 9:00 to 11:00 and occupied at the end of each day.

4.2 MERL dataset

Mitsubishi Electric Research Labs (MERL) has collected motion sensor data from a network of almost 300 sensors for more than two years since March 2006 (Wren, Ivanov, Leigh & Westhues 2007). Thess data are the residual trace from the people working in their research laboratory in Cambridge, Mass. (US), so it consists of purely real data. The release of this dataset together with some a priori information about meetings schedules and social/working activities provides an incredible resource to test the application of topic models over sensor network data. The dataset was released in March 2009, it contains over 50 million raw motion records, spanning two calendar years and two floors of their research laboratory (see figure 7 for the ground-floor overview). The sensors were placed at approximate two meters intervals along hallways, meeting rooms and lobbies. In contrast to the Innotek dataset (Castanedo, López-de-Ipiña, Aghajan & Kleihorst 2011) (Castanedo, Aghajan & Kleihorst 2011) there were no sensors placed in individual offices. That is, sensors basically measure movements of people through the corridor and meeting rooms (with some exceptions for shared rooms). The employed protocol in the sensor network does not contain checksum information and packets are sometimes lost, duplicated or confused. This issue arose some uncertainty in the provided data due to missing information. Therefore employing a probabilistic technique is a good approach to model the latent structure.

The system employed to gather the information generates $\approx$2 million sensor activations per month, giving an amount of more than 50 million activations in the complete dataset. The first sensor activation provided by the dataset came from sensor 420 and took place on Tuesday 21 March 2006 23:00:24 EST GMT-4 time, and the last one was reported by sensor 265 on Monday 29 September 2008 10:53:18 EST GMT-4 time. In the following experiments weekends are skipped and only weekdays are considered, we noticed that in the published dataset some weeks do not provide information. At the end, a total amount of 90 weeks were employed in our experiments, from Monday 27 March 2006 00:000:00 EST GMT-4 till Friday 14 December 2007 24:00:00 EST GMT-4 giving more than 31 million sensor activations.

The vocabulary for this dataset is composed of 83520 different words (2^5 * 9 day segments * 290 sensors). The corpus of the 90 employed weeks contains 88795 documents.

4.2.1 Distribution of sensor activations over time

Figure 8 shows the distribution of activations in grey-scale (black corresponds to a more likely activation) over the weekdays, with rows representing specific sensors and columns time evolution with one minute frequency. Similar to Innotek data, working patterns from the common working hours clearly arise, as it was supposed.

On other hand, a video file was recorded plotting sensor activations over time, which can be accessed at `http://bit.ly/Qfu7iX`. The frequencies of sensor activations every half and hour can be observed in the video file and clearly shows a higher amount of activations during usual working hours. Finally the distribution of the sensor activations at each time interval is shown in Figure 5.

4.2.2 Setting the number of latent topics

Due to the amount of time required to train the LDA model with MERL data using VB inference algorithm (setting more than 50 topics), we decided to employ Yahoo LDA implementation (Smola & Narayanamurthy 2010). This implemen-

Table 2: Some examples of 4 most probable words for some topics obtained with a LDA model trained with VB ($T = 100$) using the Innotek dataset

Topic 4		Topic 10		Topic 27		Topic 62		Topic 100	
word	p(w|t)	word	p(w|t)	word	p(w|t)	word	p(w|t)	word	p(w|t)
IIIII5042	0.203722	IIOII6042	0.592740	IIIII5005	0.248329	IIIII5067	0.184889	IIIII5008	0.189949
IIIII6042	0.198493	IOOII5095	0.308819	OOOOO7005	0.193592	IIIII6067	0.153970	IIIII6008	0.188247
IIIII7042	0.142416	OIIII3095	0.096328	IIIII4005	0.170163	OOOOO8067	0.133721	OOOOO8008	0.112029
IIIII4042	0.096907	IOIII5095	0.000452	OOOOO8005	0.161540	IIIII4067	0.116834	IIIII4008	0.105928
Topic 33		**Topic 66**		**Topic 94**		**Topic 18**		**Topic 35**	
word	p(w|t)	word	p(w|t)	word	p(w|t)	word	p(w|t)	word	p(w|t)
IIIOI3046	0.817912	OIIII4041	0.333204	OOOOI8097	0.572353	IIIII5000	0.191475	IIIII5093	0.239209
OOOOI5046	0.151072	IOOOO6041	0.333204	IOOOI7097	0.252334	IIIII6000	0.190710	OOOOO4093	0.228550
OOOII5046	0.029688	OOOOI6041	0.166627	OOOOO4097	0.034805	IIIII4000	0.140858	IIIII8093	0.226729
OOOII8046	0.001213	OOOII8041	0.166207	IIIIO6097	0.024649	IIIII7000	0.105399	IIIII7093	0.044512

Figure 7: Ground plan of the two floors of MERL building where the dataset were obtained.

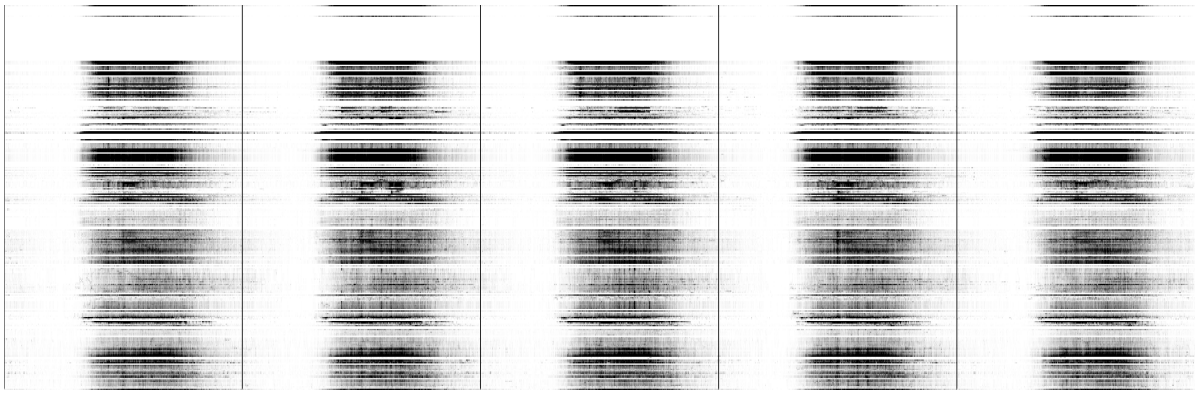

Figure 8: High level perspective. Grey-scale image that represents the distribution of sensor activities over the complete 90 weeks during weekdays, that is the probability of each sensor being activated at each of the 7200 minutes of the week (with black being most likely). Rows represent sensors and columns time evolution in one minute frequency. The hight-lighted vertical lines represents the end of each day. It could be shown that activation patterns are similar between weekdays.

tation could run over thousand of topics and use Gibbs sampling to learn and adjust the parameters of the model in the training phase. Yahoo implementation could also be employed for on-line operation or streaming mode which could be of interest in order to use the model over deployed sensor networks.

In this particular work and since the size of the employed dataset is not too big, the framework was executed just using a single machine. The model has been trained using $K = 1000$ topics.

4.2.3 Discovered routines

Some obtained Topics from a trained LDA model using Gibbs sampling and $K = 1000$ topics are shown in Table 3. It could be noted that for some of the topics the first 3 or 4 words provide enough discriminative information to classify a document into one particular topic.

For instance topic 0 is a mixture of an activation of sensor 222 (`OOOOI8222`) which is placed near the staircase between 19:00 and 21:00 with a probability of 0.24, an activation of sensor 291 (`OIIIO7291`) which is in the north conference room of 8th floor between 17:00 and 19:00 with a probability of 0.24, an activation of sensor 303 (`OIIOO2303`) which is placed before entering the kitchen between 6:00 and 7:00 with a probability of 0.24 and an activation of sensor 340 (`IOIIO6340`) which is located in the lobby of 8th floor between 14:00 and 17:00. This topic provides information about equally probable common behaviours: leaving the building between 19:00 and 21:00, having occupied the conference room between 17:00 and 19:00, having someone in the kitchen between 6:00 and 7:00 and occupancy in the lobby between 14:00 and 17:00.

Topic 11 discovers a routine which involves the activation of sensor 485 (`OIIOO2485`) located near the stairs in the 7th floor between 6:00 and 7:00; sensors 329, 373, 370 (`OOIII2329`,`IOIIO3373`,`OIOII3370`) located at corridor places in the 8th floor, early in the morning between 6:00 and 9:00; and sensor 418 (`OOIIO7418`) in the corner of the 8th floor corridor between 17:00 and 19:00.

Topic 54 shows a high probability of activation from the sensor placed between the supply and mail room, sensor 304 (`IIIII3304`), and the sensor 498 (`IOIOI3498`) placed in the corridor of 7th floor close to some printers machines. This topic indicates a high probability of having occupation at those areas between 7:00 and 9:00.

Topic 133 detects a behaviour of sensor 502 which is located close to the fire exit of 7th floor. This routine indicates different activations (`IOIOI3502`, `OIOIO3502`, `OIOIO2502`, `IOIOI2502`, `IOIOI4502`) of this sensor at different time intervals: 3 (from 7:00 to 9:00), 2 (from 6:00 to 7:00) and time interval 4

(from 9:00 to 11:00). This topic may model those fire simulations that took place in the dataset on some days.

The activity of sensor 500, which is located near some printers on 7th floor, from 6:00 to 11:00 (`OIOIO3500`, `IOIOI3500`, `IOIOI4500`, `OIOIO2500`, `OIOIO4500`) is modeled in topic 375.

Topic 222 provides information about sensor 426 which is located close to an office in the 8th floor. This routine indicates a high probability (0.58) of having this sensor activated from 6:00 to 7:00 (`IOOOO2426`,`OIOOO2426`), which could means that the person that works in that office usually comes early to work.

Topic 358 gives information about the activations on the hardware lab of 8th floor, it basically indicates that this lab is occupied between 7:00 and 11:00 with a high probability.

Obtained topic number 643 provides a routine of activations (IIIII*448) from sensor 448, located in front of the male rest-rooms on 8th floor. It basically indicates that sensor is activated during 5 minutes with high probability (0.91) on time segments 7, 6, 5 and 4, that is from 9:00 to 19:00. This means that this area of the building is usually very busy.

In the case of topic 723, it indicates activations of sensor 425, which is placed on a corridor of 8th floor with high probability (0.75) from 7:00 to 11:00.

Finally, topic 738 indicates a high probability of occupation for the north conference room of 8th floor (`IIIII6291`, `IIIII4291`, `IIIII5291`) from 9:00 to 17:00. It also makes senses that the composed words of this topic indicate that sensor is measuring activity over the complete 5 minutes slot, since the room is employed for meetings.

4.2.4 Detecting significant changes

One of the benefits of having a model trained with the common behaviours or routines of their users is to automatically detect shifts in their common patterns. The pattern changes can be detected through evaluating the classification errors on the held-out data. For some days, MERL dataset provides calendar information of the total number of people absences. One particular week from August 21 2006 to August 25 2006 (Week 21 of our MERL-based dataset) indicates an absence of more than 30% each day. The total number of sensor activations for Week 21 are 175,459 in contrast Week 20 and Week 22 give 218,146 and 185,755 each one. Figure 9 shows the average and standard deviation sensor activations for those weeks at each time segment. Clearly the number of sensor activations of week 21 is less than week 20 and 22, having for instance around 10,000 activations less than week 20 and 22 in time segment 6. A higher number of activations is detected from week 20 at time interval 8 (from 19:00 to 21:00) which may be due to the necessity of finish the work before taken holidays the next week.

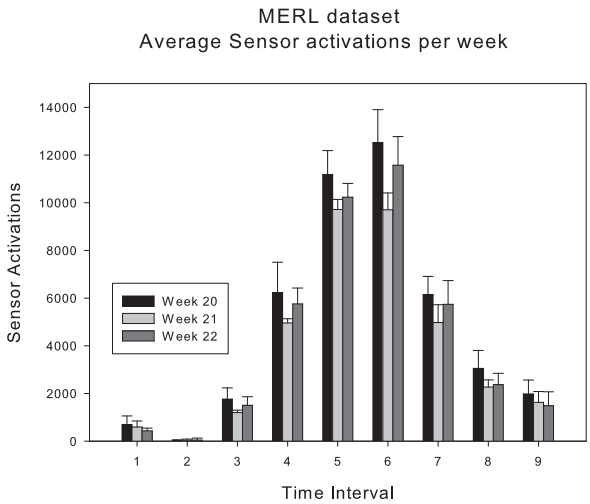

Figure 9: Average and standard deviation of all sensor activations at each time interval for weeks 20, 21 and 22 of the employed MERL data. Week 21 from August 21 2006 till August 25 2006 provides significantly less number of sensor activations due to absences.

5 Conclusions

Discovering patterns in sensor data can yield surprising results and provide insights on user behavioural patterns or working conditions.

In this work, we have introduced a novel method to model and infer the latent structure or routines in a sensor network environment deployed in office environments. It has been shown that topic models are able to detect known patterns and behaviours in sensor data just observing sensor activations over time. This method has been tested with two real

Table 3: Some examples of 5 most probable words for some topics obtained with a LDA model trained with Gibbs sampling ($T = 1000$) using the MERL dataset

Topic 0		Topic 11		Topic 54		Topic 133		Topic 375	
word	p(w|t)	word	p(w|t)	word	p(w|t)	word	p(w|t)	word	p(w|t)
OOOOI8222	0.2404	OIIOO2485	0.194231	IIIII3304	0.4590	IOIOI3502	0.26115	OIOIO3500	0.25566
OIIIO7291	0.2404	OOIII2329	0.194231	IOIOI3498	0.4590	OIOIO3502	0.26084	IOIOI3500	0.25443
OIIOO2303	0.2404	IOIIO3373	0.194231	OIOOO3214	0.0045	OIOIO2502	0.13026	IOIOI4500	0.12768
IOIIO6340	0.2404	OIOII3370	0.194231	OOOOI3214	0.0045	IOIOI2502	0.10887	OIOIO2500	0.12706
IIIOO4214	0.0023	OOIIO7418	0.194231	IOOOO3214	0.0045	IOIOI4502	0.10855	OIOIO4500	0.10681
Topic 222		Topic 358		Topic 643		Topic 723		Topic 738	
word	p(w|t)	word	p(w|t)	word	p(w|t)	word	p(w|t)	word	p(w|t)
IOOOO2426	0.395395	IIIII3424	0.29834	IIIII6448	0.35576	IIIII3425	0.31278	IIIII6291	0.37089
OIOOO2426	0.198026	IIIII4424	0.13278	IIIII5448	0.33800	IIIII4425	0.19914	IIIII4291	0.22692
OOOII2426	0.198026	OOOIO2424	0.06655	IIIII4448	0.13322	OIIII3425	0.14233	IIIII5291	0.21383
IIIOI5373	0.066447	IOOII3424	0.06655	IIIII7448	0.10657	OIIIO3425	0.11392	IIIIO4291	0.05239
IOIIO8266	0.066447	IIIIO3424	0.06655	IIIII8448	0.02246	OOOII2425	0.05710	IIOII4291	0.02622

datasets with more than 100 sensors and over 50 weeks of data. The obtained results indicate that the model is able to learn the underlying latent structure from the raw data. Some of the obtained topics (in this context, routines) show the ability to obtain routines that represent the common activities gathered from the sensor network. Having a good model that represents the underlying dynamics of a sensor network is useful for several tasks. For instance, opportunities arise in the context of smart buildings, optimising the future energy consumption of an office building using a long-term model.

We had chosen a LDA implementation that can scale to millions of documents and thousands of topics, in particular we performed some test experiments using 3000 topics with reasonable good scalability results. Another advantage of the LDA model is the unsupervised nature, which makes possible to recover the latent dynamic structure of a sensor network just by observing sensor activations.

We noticed that the interpretation of detected topics is somehow hard with the employed codification. This could be easily solved using letters which provide semantic information instead of sensors identifiers (i.e. KTN for sensors at the Kitchen).

As a future work it will be very interesting to investigate the application of stream-based and online topic models (Hoffman, Blei & Bach 2010) in the sensor network domain. Stream-based can provide an output of the LDA classification in near real-time. Stream-based algorithms have been used in the past to mine sensor data but less work has been done under the topic models research area. This will be an interesting area to invest further research efforts.

Acknowledgments

We want to thank Chris Wren and Yuri Ivanov for making available the MERL dataset to the research community. Special thanks to Erik Degroof and Luc Peeters from Innotek, Belgium for their cooperation and the use of their dataset. Many thanks to David Blei, Alexander Smola and Shravan Narayanamurthy for release their LDA code.

The first and second authors' work is partially supported by projects TALIS+ENGINE (TIN2010-20510-C04-03) and DYNUI (PC2012-73A).

References

Asuncion, A., Welling, M., Smyth, P. & Teh, Y. W. (2009), On smoothing and inference for topic models, *in* Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence.

Bettini, C., Brdiczka, O., Henricksen, K., Indulska, J., Nicklas, D., Ranganathan, A. & Riboni, D. (2010), A survey of context modelling and reasoning techniques, *Pervasive and Mobile Computing* **6**(2), 161–180.

Blei, D., Ng, A. & Jordan, M. (2003), Latent dirichlet allocation, *The Journal of Machine Learning Research* **3**, 993–1022.

Castanedo, F., Aghajan, H. & Kleihorst, R. (2011), Modeling and discovering occupancy patterns in sensor networks using latent dirichlet allocation, *Foundations on Natural and Artificial Computation(IWINAC'11)* pp. 481–490.

Castanedo, F., López-de-Ipiña, D., Aghajan, H. & Kleihorst, R. (2011), Building an occupancy model from sensor networks in office environments, *5th International ACM/IEEE Conference on Distributed Smart Cameras(ICDSC'11)* .

Connolly, C., Burns, J. & Bui, H. (2008), Recovering social networks from massive track datasets, *IEEE workshop on Applications of Computer Vision, WACV* .

Farrahi, K. & Gatica-Perez, D. (2008), What did you do today?: discovering daily routines from large-scale mobile data, *in* Proceedings of the 16th ACM international conference on Multimedia, ACM, pp. 849–852.

Farrahi, K. & Gatica-Perez, D. (2011), Discovering routines from large-scale human locations using probabilistic topic models, *ACM Transactions on Intelligent Systems and Technology* **2**(1).

Ferrari, L. & Mamei, M. (2011), Discovering daily routines from google latitude with topic models, *in* International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), IEEE, pp. 432–437.

Georgeon, O., Mille, A., Bellet, T., Mathern, B. & Ritter, F. (2011), Supporting activity modelling from activity traces, *Expert Systems* **29**(3), 261–275.

Griffiths, T. & Steyvers, M. (2004), Finding scientific topics, *Proceedings of the National Academy of Sciences of the United States of America* **101**(Suppl 1), 5228–5235.

Hoffman, M., Blei, D. & Bach, F. (2010), Online learning for latent dirichlet allocation, *Advances in Neural Information Processing Systems* **23**, 856–864.

Hofmann, T. (1999), Probabilistic latent semantic indexing, *in* Proceedings of the 22nd ACM international conference on Research and developmentin information retrieval, ACM, pp. 50–57.

Hofmann, T. (2001), Unsupervised learning by probabilistic latent semantic analysis, *Machine Learning* **42**(1), 177–196.

Horster, E., Lienhart, R. & Slaney, M. (2007), Image retrieval on large-scale image databases, *in* Proceedings of the 6th ACM international conference on Image and video retrieval, ACM, pp. 17–24.

Hospedales, T., Li, J., Gong, S. & Xiang, T. (2011), Identifying rare and subtle behaviours: A weakly supervised joint topic model, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33**(12), 2451–2464.

Huynh, T., Fritz, M. & Schiele, B. (2008), Discovery of activity patterns using topic models, *in* Proceedings of the 10th international conference on Ubiquitous computing, ACM New York, NY, USA, p. 10–19.

Landauer, T. & Dumais, S. (1997), A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge., *Psychological review* **104**(2), 211.

Magnusson, M. (2000), Discovering hidden time patterns in behavior: T-patterns and their detection, *Behavior Research Methods* **32**(1), 93–110.

Mariote, L., Medeiros, C. & da Torres, R. (2007), Parallelized variational em for latent dirichlet allocation: An experimental evaluation of speed and scalability, *in* Seventh IEEE International Conference on Data Mining Workshops. ICDM Workshops, pp. 349–354.

Mukherjee, I. & Blei, D. (2009), Relative performance guarantees for approximate inference in

latent dirichlet allocation, *Advances in Neural Information Processing Systems* **21**, 1129–1136.

Niebles, J., Wang, H. & Fei-Fei, L. (2008), Unsupervised learning of human action categories using spatial-temporal words, *International Journal of Computer Vision* **79**(3), 299–318.

Pauwels, E., Salah, A. & Tavenard, R. (2007), Sensor networks for ambient intelligence, *in* IEEE 9th Workshop on multimedia Signal Processing, pp. 13–16.

Phung, D., Adams, B., Tran, K., Venkatesh, S. & Kumar, M. (2009), High accuracy context recovery using clustering mechanisms, *IEEE International Conference on Pervasive Computing and Communications* pp. 1–9.

Rashidi, P., Cook, D., Holder, L. & Schmitter-Edgecombe, M. (2011), Discovering activities to recognize and track in a smart environment, *IEEE Transactions on Knowledge and Data Engineering* (99).

Rosen-Zvi, M., Griffiths, T., Steyvers, M. & Smyth, P. (2004), The author-topic model for authors and documents, *in* Proceedings of the 20th conference on Uncertainty in artificial intelligence, AUAI Press, pp. 487–494.

Salah, A., Pauwels, E., Tavenard, R. & Gevers, T. (2010), T-Patterns Revisited: Mining for Temporal Patterns in Sensor Data, *Sensors* **10**(8), 7496–7513.

Sigg, S., Haseloff, S. & David, K. (2010), An alignment approach for context prediction tasks in ubicomp environments, *IEEE Pervasive Computing* **9**(4), 90–97.

Sivic, J., Russell, B., Efros, A., Zisserman, A. & Freeman, W. (2005), Discovering object categories in image collections, *ICCV* .

Smola, A. & Narayanamurthy, S. (2010), An architecture for parallel topic models, *Proceedings of the VLDB Endowment* **3**(1-2), 703–710.

Wallach, H., Murray, I., Salakhutdinov, R. & Mimno, D. (2009), Evaluation methods for topic models, *in* Proceedings of the 26th Annual International Conference on Machine Learning, ACM, pp. 1105–1112.

Wang, X., Ma, X. & Grimson, W. (2009), Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**(3), 539–555.

Wren, C. R., Ivanov, Y. A., Leigh, D. & Westhues, J. (2007), The MERL Motion Detector Dataset, *MERL TR2007-069* .

Yao, L., Mimno, D. & McCallum, A. (2009), Efficient methods for topic model inference on streaming document collections, *in* Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 937–946.