Pervasive and Mobile Computing 15 (2014) 215–227

Contents lists available at ScienceDirect

Pervasive and Mobile Computing

journal homepage: www.elsevier.com/locate/pmc

Discovery of activity composites using topic models: An analysis of unsupervised methods

Julia Seiter [a,*], Oliver Amft [b], Mirco Rossi [a], Gerhard Tröster [a]

[a] *Wearable Computing Lab., ETH Zurich, Switzerland*

[b] *ACTLab, University of Passau, Germany*

ARTICLE INFO

Article history:
Available online 2 June 2014

Keywords:
Activity routines
Daily routines
Topic modeling
Hierarchical activity recognition
Activity discovery

ABSTRACT

In this work we investigate unsupervised activity discovery approaches using three topic model (TM) approaches, based on Latent Dirichlet Allocation (LDA), n-gram TM (NTM), and correlated TM (CTM). While LDA structures activity primitives, NTM adds primitive sequence information, and CTM exploits co-occurring topics. We use an activity composite/primitive abstraction and analyze three public datasets with different properties that affect the discovery, including primitive rate, activity composite specificity, primitive sequence similarity, and composite-instance ratio. We compare the activity composite discovery performance among the TM approaches and against a baseline using k-means clustering. We provide guidelines for method and optimal TM parameter selection, depending on data properties and activity primitive noise. Results indicate that TMs can outperform k-means clustering up to 17%, when composite specificity is low. LDA-based TMs showed higher robustness against noise compared to other TMs and k-means.

1. Introduction

Discovering activity composites in ubiquitous sensor data could provide insights into individual behavior with broad applications stretching from assisted living to medical diagnosis. Unsupervised discovery approaches enable users and analysts to detect and describe structures in activity sensor data without requiring annotations and supervised pattern learning.

A hierarchical abstraction has often been considered to partition human behavior into *activity primitives*, which can be recognized from on-body and ambient sensor data, and more abstract *activity composites*. Typically, activity primitives have a fine temporal granularity and must be suitable for recognition from sensor measurements. Subsequently, activity primitives could be composed into activity composites using discovery methods. Fig. 1 exemplarily illustrates the hierarchical abstraction. Some approaches towards activity discovery from ubiquitous sensor data have been proposed (see Section 2 for more details); however the required activity data properties and algorithm configurations are not established. In addition, datasets vary widely depending on the application, e.g. regarding activity primitive composition and composite specificity.

TMs are probabilistic graphical models and find their origin in the text processing community. TMs were initially used to discover hidden topics from a corpus of documents, each containing a bag-of-words from a predefined vocabulary [1]. In activity discovery, words correspond to activity primitives and topics to activity composites. An appropriate configuration of TMs depending on the data properties is essential to obtain meaningful composite discovery results. In particular, the selection of TM parameters, including primitive segment size and number of activity topics, has large impact on the

* Corresponding author. Tel.: +41 446322744.
E-mail address: julia.seiter@ife.ee.ethz.ch (J. Seiter).

http://dx.doi.org/10.1016/j.pmcj.2014.05.007
1574-1192/

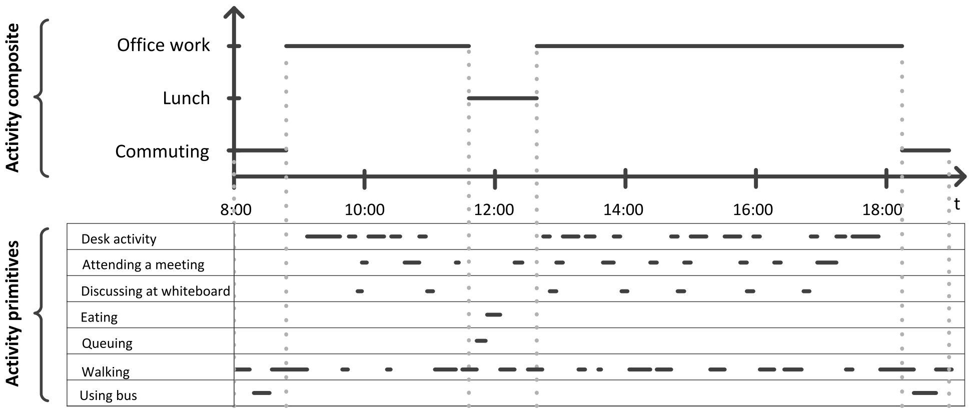

Fig. 1. Example of a hierarchical activity abstraction that is considered for activity discovery. In this example, three *activity composites* could be discovered based on various *activity primitives*. In this work, we analyze the performance of unsupervised discovery methods based on topic models using real-world datasets with different properties.

discovery performance. The most frequently applied TM based on Latent Dirichlet Allocation (LDA) uses activity primitive histograms, e.g. in [2]. These histograms represent time-independent statistics of primitives, and thus do not consider primitive sequences as they often occur in activity and behavior data. The sequence of activity primitives might provide important information on the data structure and could enhance discovery performance. Although discovery approaches exist that incorporate sequence information (e.g. [3–5]), the benefits of using sequence information are not yet established. Similarly, activity composites may co-occur, which could be captured using correlated TMs (CTM) [6].

In this paper, we investigate three topic modeling approaches based on LDA, the n-gram TM NTM and CTM, in three public activity datasets with different activity data properties that affect the topic modeling. The paper provides the following contributions:

1. We introduce LDA, NTM, and CTM approaches for activity discovery and compare performances in three datasets to a baseline method using k-means data clustering. Based on the results, we provide recommendations on optimal TM parameter choices. For this investigation, we consider three publically available datasets.
2. We investigate four essential dataset properties, including the activity primitive rate, composite specificity, primitive sequence similarity, and composite-instance ratio, which affect the TM operation to derive guidelines for TM parameter choice to achieve optimal discovery performance.
3. We analyze the effects of imperfect activity primitive recognition on discovery performance. Here, we consider primitive insertion and deletion errors to illustrate TM performance bounds.

In our previous work, we investigated LDA-based TMs and simulated activity data with varying data properties [7]. In this work, we extend the discovery analysis to include NTM, CTM, and a baseline method for comparisons on three structurally different datasets. We also refine the dataset properties considered to provide guidelines in parameter and method selection.

2. Related work

In activity discovery, histogram-based methods are frequently used to extract structural patterns. Gu et al. extracted characteristic object-use fingerprints applying web-mining and discovered contrast patterns for each activity using emerging patterns [8]. Each activity's fingerprint consisted of a histogram over object usage. Begole et al. applied a rhythm model to visualize daily rhythms from computer usage by clustering patterns of computer activity [9]. Besides the approaches using clustering-based methods, probabilistic models were also applied for activity discovery. Barger et al. used probabilistic mixture models to infer behavior patterns in daily life from clusters that were formed from occurrence statistics of senors events in a smart home [10].

TMs have been frequently applied to discover human activities and body postures from video data and image features such as [11–15]. Applications of TMs in activity discovery from wearable sensors are less frequent. Farrahi et al. inferred daily routines from proximity [16] and mobile phone data [17] using TMs. Huynh et al. discovered daily routine patterns from activity primitives by applying a TM [2] to a personal monitoring dataset obtained over several regular days. Subsequently, identified topics were mapped to daily life routines. In this work, we follow a similar approach for TM-based activity discovery. However, we aim at an in-depth analysis of TMs when being adapted for activity discovery across different datasets, investigate optimal parameter choices, and study TMs under primitive recognition noise with the aim to guide method and parameter selection.

Sequential information has been considered to infer activities of daily living. Aztiria et al. applied a sequential pattern mining algorithm and defined a descriptive language to infer user behavior in daily life from smart home sensor data [3].