Discovering Water Use Activities for Smart Metering

Rachel Cardell-Oliver #1

# *School of Computer Science & Software Engineering*
The University of Western Australia, Perth, Australia
1 rachel.cardell-oliver@uwa.edu.au

***Abstract*—Smart water meter systems are large scale wireless sensor networks: water meters installed in thousands of households, collect hourly measurements that are reported over a wireless network to a central database. This paper introduces a new method for activity discovery in real-world, hourly water meter readings. The method addresses the following constraints: 1) observations are unlabelled and so unsupervised learning of activity types is required, 2) only automatically collected readings are used, and 3) coarse-grained hourly readings mask sub-hourly concurrent and sequential activities. Automatic rule-based labelling is combined with hierarchical clustering. New criteria are introduced for evaluating the quality of discovered activity clusters. We demonstrate the utility of our activity discovery and evaluation methods using a real-world case study of over 35,000 example days from a smart water meter trial in the inland Western Australian town of Kalgoorlie Boulder. The results show that the new method is able to discover meaningful and significant activity patterns from coarse-grained hourly readings.**

I. Introduction

Water utility companies have recently begun large scale deployments of smart water meters to monitor household water use. Smart water meter systems are wireless sensor networks: water meters installed in thousands of households collect hourly measurements that are reported in real-time over a wireless network to a central database. The term “smart metering” refers to the analysis of gathered data to to inform decision making, as apposed to ”smart meters”, which are equipment assets [10]. For individual householders, smart metering can provide early warnings of unusual events such as continuous flow of water which may be “leaks”. For the water provider, smart metering informs decision making on pricing strategies, intervention policies and setting water use reduction targets.

If householders are to reduce their water usage, then they need to change their current behaviours. Behavioural science tells us that in order to achieve change, feedback to users needs to draw “a close link between specific actions and their effects” [6]. Therefore, an important aspect of smart metering is to recognise relevant activity types and to identify the occurrence of those activities in time series of smart meter readings.

The problem addressed in this paper is how to discover water use activities from water meter time series. This problem is one of activity discovery, the task of identifying the common actions of one or more agents from a series of observations. In the context of smart water metering, the actions are household behaviours that contribute to water use, the agents are households (occupants and their appliances), and the sensed observations are water meter readings together with limited contextual information about those readings.

This paper introduces a new method for discovering meaningful activities from real-world smart water meter observations. There are several constraints that distinguish our work from previous studies. We identify activities using only automatically collected observations: hourly meter readings. This constraint enables our approach to be scaled to thousands of meters running over periods of years. However, it also means that the richer contextual information provided by multiple sensors, fine grained sensing, and user demographics, is not available for the interpretation of water use activities. We also assume the data is unlabelled, and that labelled training data sets are not available. Therefore, unsupervised learning must be used to discover and label activities. Another challenge for activity recognition from a single data stream of hourly readings per household is that many interesting details are hidden. Concurrent activities such as taking a shower while the washing machine is running, and sequential activities such as showering and breakfasting while getting ready for work in the morning, are all aggregated into single, hourly water volume measurements. The main contribution of this paper is a new unsupervised learning method for the discovery of water use activities under these constraints.

The activity discovery method presented in this paper has three steps: feature selection, unsupervised learning and cluster evaluation. First, feature selection is used to convert raw timer-series data into a set of instances, each representing a day of water use for a single metered property. The next step is to partition those instances into activity clusters, which is done in two parts. Common sense rules are used to identify sub-classes such as days of peak use or days with continuous flows. Then large clusters are further broken down into sub-activities using the k-means unsupervised clustering algorithm. Finally, evaluation criteria are introduced for assessing the interestingness of discovered activities, in terms of their usefulness in solving the original problem of providing information to support behaviour change.

To test the utility of our approach we have a case study of over 35,000 days of meter reading examples selected from network of over 11,000 smart meters in Kalgoorlie Boulder, a large inland town in Western Australia. The scope of this paper is the task of activity discovery for a user population.

978-1-4673-5501-8/13/$31.00 © 2013 IEEE

The problems of recognising activities on the fly, visualising activities for feedback to users, and dynamic analysis of individual users' patterns will be reported elsewhere.

II. Background

State of the art studies of end-user water or electricity usage typically take a mixed-method approach [2], [3], [11], [5]. Information such as social and behavioural aspects of the household, an audit of water (or electricity) appliances and fixtures, water use diaries, land use surveys, and weather conditions are used to triangulate observations about patterns of water use. These studies also utilise fine-grained water meter readings for accurate flow-trace analysis. For example, 10 second water use data, can be analysed with the TraceWizard software tool to identify activities such as indoor and outdoor use, toilet flushes, washing machine use, leaks, and irrigation [1]. Large scale water use studies of this type include a study of 200 homes in South East Queensland [2] and 700 single-family homes in California [3]. While mixed method studies provide a detailed insight into the water use activities of individual households, such studies are costly to run and require a high level of user involvement. In this paper we introduce a scalable method using data mining techniques to recognise significant activities given only automatically gathered observations.

Human activity discovery is a significant field of data mining and machine learning, with many different approaches. Activity recognition in "smart homes" uses a collection of sensors placed throughout the home. In most cases labelled training data is used, in which observations from the sensors are annotated with activity labels, either added by expert analysis or by users keeping a diary of their activities. Rashidi et. al. [9] propose the DVSM algorithm, for discovering frequent activities from unlabelled sensor data. Their approach combines sequence mining and clustering algorithms to distinguish multiple activities interleaved in time, and activity sequences that occur in different order. We also combine sequence mining and clustering techniques. However, since hourly water meter data is too coarse grained to distinguish activities of multiple agents or different sequences of activities, DVSM is not directly applicable to our problem.

Activity recognition is a special class of problems in the more general domain of time series mining. Approaches for time series clustering include those based on raw time series and techniques which first extract features from the data and then cluster on that feature sets using standard clustering techniques [8]. Keogh and Lin argue against the former approach claiming that "clustering of time series subsequences is meaningless" since many patterns in time series are "averaged out" when sliding windows are used to select time series for clustering [7]. Xing et. al. also argue that "there are no explicit features in time series data" [12] and so meaningful features must be identified before learning can take place. On the other hand, Ye and Keogh argue that the simple nearest neighbour algorithm for measuring the distance between two timed sequences is usually very effective [13]. They argue, however, that clustering by time series differences lacks the ability to explain the classification results, and they propose sub-sequence primitives, called shapelets, for characterising time series. An experimental comparison of distance measures by Ding et. al. found that simple Euclidean distance between two timed sequences of values performed just as well as more complex measures [4]. In this paper we also use Euclidean distance between sequences of meter readings. Our study exposes both the strengths and weaknesses of this approach and techniques for addressing the weaknesses are proposed.

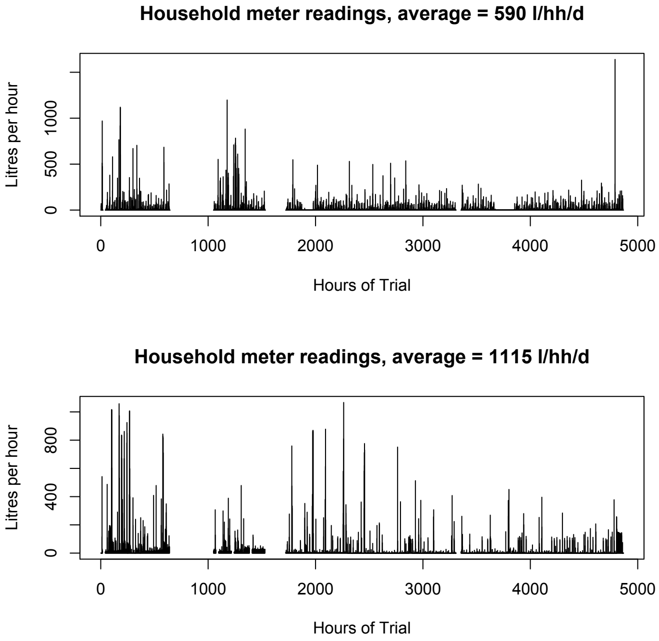

Fig. 1. Raw Sensor Streams from Smart Water Meters

III. Features of Water Use Activities

This section describes the data model and feature selection techniques we use to identify water use activities. Figure 1 shows the raw meter readings per hour from January to August from two households in the case study. The top household is a low water user (590 l/hh/d) whilst the lower household is a high water user (1115 l/hh/d). It is difficult to identify patterns in this raw data. The task of activity discovery is to encode this raw data using higher level, more meaningful, activity labels. In order to do that, we must first identify meaningful features of the raw data, and describe the structure of the data in terms of those features.

A. Data Model

The task of activity discovery in smart meter data is an unsupervised learning, data mining task. The data set is a set of instance examples, e, each a tuple comprising a meter identifier m, a date stamp t, and a sequence of hourly meter readings over a 24 hour period, v_0 to v_{23}.

$$e = \langle m, t, v_0, v_1, ..., v_{23} \rangle$$

The meter readings, in litres, for each hour of the day (0..23) are natural numbers v_i. Each attribute a_i of e can be accessed using the denotation $e.a_i$, such as $e.m$ and $e.v_0$.

Our aim is to discover water use behaviours from a data set D of examples. We partition D into subsets (clusters) $D_1, ..., D_k$ each of which characterises a different type of activity. The subsets may overlap (they need not be mutually exclusive) and subsets can be created hierarchically.

B. Activity Length

In order to discover activities that lie within a long sequence of observations, it is necessary to identify a temporal window in which the activities can occur. There are many ways of defining windows within time series [7]. The size of windows may be fixed or variable size, and the position of the window may be slid across the whole set, or determined by particular events in the sequence. For the hourly data of this study we use fixed, whole day, rather than variable windows, since traditional variable length water use activities such as showering or toilet flushes, are not visible in hourly data. Each window captures 24 hours of meter readings. Daily activity windows do not overlap and any incomplete days in the data set are discarded.

Given windows of a single day, it is possible to zoom in and out of the data set to identify certain features.

YEAR Annual per-meter summaries can be used to link water usage to billing bands and to distinguish between water use norms in different regions (e.g. inland vs coastal towns). However, from a behaviour change point of view, providing only an annual use summary, as in current bills, has the disadvantage that the effect (final bill amount) is not linked to the specific actions that caused it.

WEEK Weekly summaries of water use can be used to visualise seasonal variation and to hide some irrelevant details such as which days of the week are rostered garden watering days.

DAY Daily usage can be linked to contextual information such as daily temperature, wind, humidity or rainfall, as well as calendar-based comparisons of weekday, weekend or holiday use.

HOUR Hourly usage is the finest-grained information available in this study, and so the closest link to human activities. However, hourly aggregates of water use do not expose concurrent use by individuals or machines in the same household, nor sequences of events within an hour.

QUARTER DAYS Aggregating hourly meter readings into 6 hour buckets, or quarter-days is a compromise between daily and hourly models of water use. This reduces noise that arises in hourly data from variations of the underlying multi-user, multi-activities.

SHAPELETS are recurring subsequences of any length that can occur at any position within a larger sequence [13]. Shapelets have proved effective for discovering patterns within noisy time series data. However, for hourly meter data, irregular and unpredictable activities, such as running a washing machine, are below the threshold to stand out from other activities, and so shapelet subsequences are unable to detect such patterns.

C. Common-Sense Rules

Household water usage is an established research area, with a taxonomy of common activities. Many identified activities relate to the use of individual appliances (e.g. washing machine or toilet flushes) but these activity types are hidden in hourly time series. However, larger-scale activities such as continuous flows (leaks), extreme use (peaks), and empty days, are visible in hourly time series. Each of these classes can be specified by a first order logic rule that characterises the set of days on which an activity occurs.

Recall that activity classes are defined as subsets of a training set D of daily water uses for a population of different meters and days given by $e = \langle m, t, v_0, v_1, ..., v_{23} \rangle$.

1) Empty Days: Empty days are those where no water is drawn, presumably because the householders are absent from the property. Although empty days account for zero water use, it is important to identify such days in order to exclude them from skewing other calculations.

$$\text{Empty} =_{def} \{ \ e \mid e \in D \wedge \Sigma_{i=0}^{23} e.v_i = 0 \ \}$$

2) Continuous Flow: Continuous flows over 24 hours indicate a possible water leak. A leak is an unintended water loss, usually unknown to the householder. However, there are also intentional reasons for continuous flows such as the continuous use of evaporative air conditioners. Thus the term “continuous flow” is used to characterise such days, rather than the term “leak”. We follow the water industry convention that a continuous flow is defined as at least 2 litres of flow in every hour of a 24 hour calendar day.

$$\text{ContinuousFlow} =_{def} \{ \ e \mid e \in D \wedge \forall \ i \in [0, 23]. \ e.v_i \geq 2 \ \}$$

3) Peak Days: A peak day is defined relative to the whole population’s normal daily use. as $ub = me + 1.5 * iqr$ for population median me and inter-quartile range iqr of all daily totals for all meters.

$$\text{Peak} =_{def} \{ \ e \mid e \in D \wedge (\Sigma_{i=0}^{23} e.v_i) > ub \ \}$$

Since exceptional (peak) behaviours account for a large proportion of overall water use, targeting exceptional use with personalised feedback has the potential to be a highly effective water reduction strategy. A per-meter definition of peak days can be specified in the same way.

4) Normal Days: Normal days are defined by exclusion as those days not in the above categories.

$$\text{Normal} =_{def} D - \text{Empty} - \text{ContinuousFlow} - \text{Peak}$$

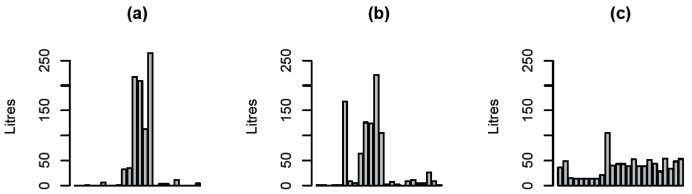

Fig. 2. Hour of day vs litres per hour for three trial days

D. Clustering

The activity rules defined above capture activities that can be used to identify water use reduction strategies. However, the category of "Normal" days is a large one, and so a further breakdown of normal days into sub-activities is desirable. The k-means algorithm will be used to discover activity patterns within normal days. But doing this requires a suitable metric to capture the "distance" between water use days.

Figure 2 illustrates some of the problems in selecting a distance metric. It shows litres of water used per hour of day for three days, Days (a) and (b) are similar in that they both have their highest water use in the middle of the day (between 10:00 to 15:00 hours), although there are significant differences for some hours of the day. Day (c) has no peak hours, but instead has a continuous flow throughout the 24 hours. In fact we have two normal days and one continuous flow day. One similarity, which is not obvious from this view of the data, is that exactly 900 litres of water is used on all three days. The Euclidean distance metric can represent each of these views, using different time granularities such as hourly, daily or 6-hourly.

IV. Evaluation Metrics

Evaluating the quality of discovered clusters is a challenge for unsupervised learning algorithms since there is no ground truth available to decide whether the label assigned to a given day is correct or not. Therefore we propose both quantitative and qualitative criteria for evaluating water use activity clusters. Whether a discovered activity is "interesting" or not can be defined in terms of these criteria.

A. Qualitative Metrics

Meaningful: The activity cluster corresponds to a description of water use that makes sense in the real world. Evaluation of this criteria is necessarily subjective, but the idea is to favour clusters that can be understood by water users, so that information about the type of activity can be used to manage water use. For example, the cluster "leaks" describes days in which a continuous flow of water has occurred throughout the day. The concept of continuous flow is easy to understand, and there are straightforward actions that can be taken to reduce water use from leaks.

General: An activity description should be sufficiently general to apply to more than one case study. In machine learning terms, we require that activity clusters should not be over-fitted to a particular training set.

Activity	Peak	Normal	C. Flow	Empty
Significance (to nearest megalitre)	15	13	3*	0
Significance (% volume)	44.1	38.7	9.6*	0.0
Frequency (% days N=170)	12.7	63.1	26.6	3.9
Prevalence (% meters N=226)	90.7	95.6	79.6	28.8

TABLE I
Water Use Activities defined by Rules (* is continuous flow volumes only)

B. Quantitative Metrics

Significance (% volume): The proportion of all water use accounted for by an activity. Activities that account for a high proportion all water used can be targeted for intervention policies to reduce this type of use. Thresholds for global significance are set for the whole population. Alternatively, context dependent thresholds can be defined locally.

Frequency (% days): The proportion of days across the whole population that are covered by an activity. This criteria does not take into account skew, either to particular days or to a particular group of meters, but simply identifies an activity that occurs sufficiently often to justify naming it as an activity.

Prevalence (% meters): The proportion of households covered by an activity cluster. In most studies, leaks are restricted to a small proportion of households and so are not prevalent, but in our data study continuous flows were prevalent, although their significance (overall water use) and frequency (proportion of days) were relatively low.

Coherence: The coherence of an activity cluster is the spread of the examples in the cluster. Coherence can be visualised by a box plot in which a box shows the median and upper and lower quartiles of the cluster, with whiskers showing the full range of the data. The more compact the box and whiskers, the more coherent the cluster.

V. Evaluation

This section analyses water use activities in a historical data selected from the 2010-12 Kalgoorlie Smart Meter Trial by the Water Corporation of Western Australia [10]. The data set comprises over 35,000 days of hourly water meter readings for 239 meters selected from the 13,800 properties metered in the trial. We analyse 226 meters associated with houses, units and duplexes. Other types of land use, such as sports parks and clubs, are not considered in this paper.

The following algorithm is used to discover and evaluate water use activity discovery clusters:

1) Cluster the set of sample days according to given water use *rules* defined in Section III-C.
2) Choose the largest sized cluster from step 1 and choose a *distance metric* between water-use days and the *number of clusters* to be found. Use k-means clustering to partition the cluster with these parameters.

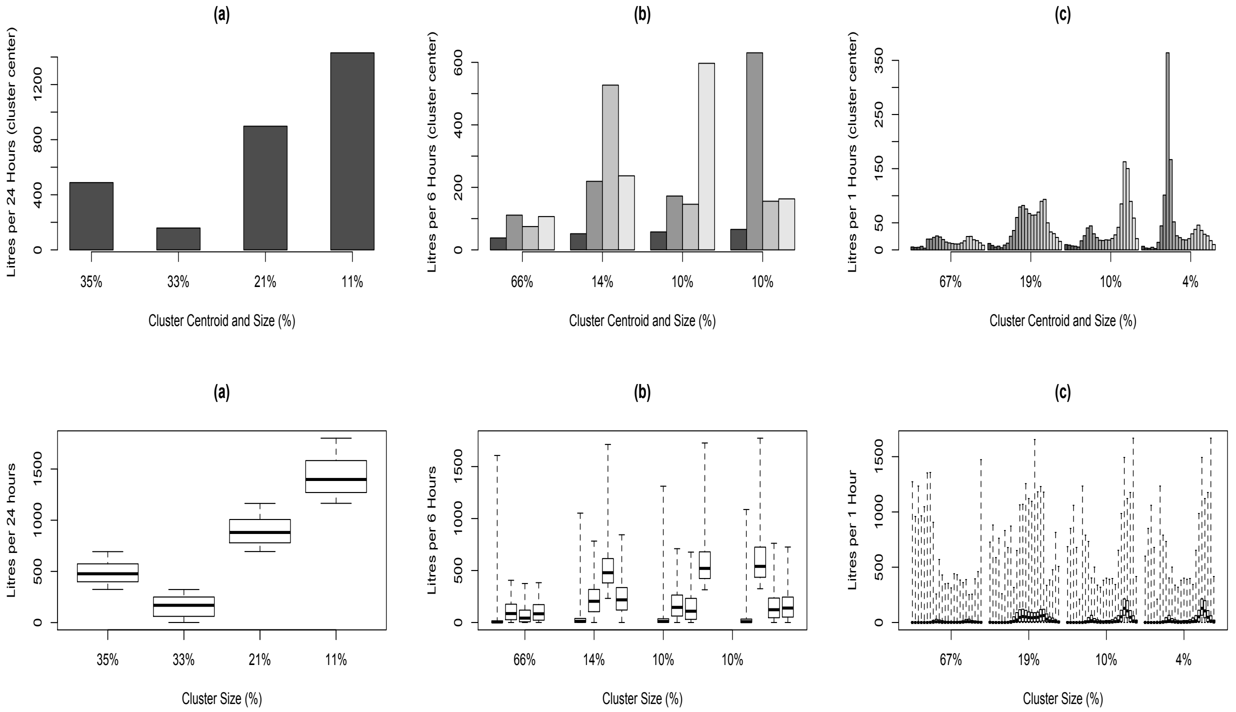

Fig. 3. Centroids (top row) and Coherence (bottom row) of clusters by aggregates of (a) 24 hours (b) 6 hours (c) 1 hour.

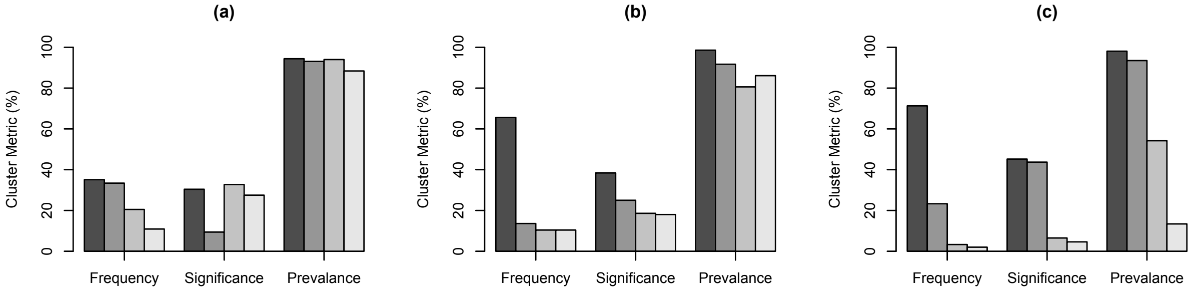

Fig. 4. Significance, Frequency and Prevalence of four clusters by aggregates of (a) 24 hours (b) 6 hours (c) 1 hour.

3) Use the *evaluation metrics* of Section IV to decide whether the discovered clusters are interesting. Repeat step 2 with different parameters if necessary.

A. Rule-based Activities

Table I summarises the activity groups defined by the water use rules of Section III-C, where we argued these categories are both meaningful and general. Evaluating quantitative metrics for the discovered clusters we note that the Peak activity is interesting because it has high significance (44.1%) relative to the frequency of peak days (12.7%). On the other hand, the Empty activity is interesting because it has low significance (0%) relative to frequency (3.9%). The continuous flow activity is notable in this study because of its high prevalence: 79.6% all metered houses had at least one continuous flow day. The Normal activity has a high frequency (63.1%) but relatively low significance (38.7%). This cluster is now further partitioned using the k-means algorithm.

B. Activity Clusters

This section shows the results of k-means clustering on the 22,609 Normal days of the Kalgoorlie data set. Four clusters are discovered in each case and three different distance metrics are used:

1) Daily water use: each day is represented by the total sum of each 24 hours of use.
2) Six-hourly aggregated use: each day is a 4-tuple of values.

3) Hourly water use: each day is a 24-tuple of values.

Figure 3 (top row) shows the centroids of the discovered clusters for each of the distance functions. The size of each cluster (% of all days) is given on the x-axis. The 6 hour centroids (top row (b)) show use during each quarter of the day: early morning (midnight to 6am), morning (6am to midday), afternoon (midday to 6pm) and evening (6pm to midnight). The discovered clusters show four clear pattern types: low use throughout the day (clusters 1), a peak in afternoon use (cluster 2), evening peak usage (cluster 3) and early morning peak use (cluster 4).

The 1 hour clusters (c) appear to give the most detailed information about daily usage patterns and the 24 hour clusters (a) the least. However, it is important to consider not only the cluster centroids, but also the coherence of the discovered clusters to avoid "meaningless" activities [7]. Figure 3 (bottom row) shows the spread of values in each cluster for each of the distance functions. The clusters for 24-hour aggregate (a) are strongly coherent in that each represents a narrow band of water use. However, these clusters have limited meaning for users because the same daily water total can be caused by very different water use patterns (as shown in Section III-D). The clusters of (b) 6 hour aggregates include outlier values (as shown by the whiskers) but have a compact interquartile range. The clusters of (c) 1 hour aggregate illustrate the "curse of dimensionality" problem. Whiskers indicate outlier values for each hour in each of the four clusters. For every hour, the outliers dominate the underlying patterns of use shown in the centres. Therefore, the 1 hour clusters are essentially meaningless, because they include an over-wide range of water use patterns.

Figure 4 compares the frequency, significance and prevalence of clusters discovered by each of the distance metrics. The clusters of (b) 6 hour aggregates have one large and three smaller clusters, all with reasonable significance and prevalence. The clusters of (c) 1 hour aggregate are skewed for all metrics while clusters of (a) 24 hour aggregates have high prevalence and similar frequencies.

In conclusion, the clusters formed from 6 hour aggregates provide the most meaningful information for users about their patterns of water use on different days, whilst also having reasonable frequency, significance and prevalence for each activity. Combining common-sense rule clusters (Table 1) and 6 hour aggregates (Figures 3b and 4b) shows that the population of our case study has seven water use activities: peak (frequency 12.7%), continuous flow (26.6%), empty days (3.9%), with normal use (63.1%) partitioned into low use (41.6%), afternoon peak (8.8%), evening peak (6.3%) and morning peak (6.3%). Although low use days form the largest activity cluster, the smaller, peak, activity types we have identified with high significance offer good opportunities for informing behaviour change strategies to reduce water use.

VI. Conclusions and Future Work

This paper addresses an activity discovery problem that arises from the large scale deployment of smart water meters: how can relatively coarse-grained sensor streams be used for meaningful activity recognition of household water use? We present a new, unsupervised learning method that combines rule-based activity labels and k-means clustering. Using the evaluation criteria of meaningfulness, significance, frequency, prevalence and coherence of activity clusters, we demonstrate that the combined method produces good quality activity clusters for a real-world data set.

In future work we plan to extend the model presented here to take more account of contextual features such as seasonal correlations and land use types, and also to compare the findings of the Kalgoorlie Boulder case study with other communities. We shall investigate techniques to apply this method in a real-world setting with on-the-fly recognition of activities and visualisations to help water users and providers to explore patterns in their water use. This study focussed on a whole-population categorisation of water use. Future work will focus on dynamic models of water use patterns for individual households.

Acknowledgements

This research was supported by the Cooperative Research Centre for Water Sensitive Cities and the Water Corporation of Western Australia.

References

[1] U. Aquacraft Inc. Boulder, CO. Trace wizard software version 4.1. 1995-2010. www.aquacraft.com.

[2] C. Beal, R. Stewart, and T.-T. A. Huang. South East Queensland Residential End Use Study: Baseline Results - Winter 2010 (UWSRA-tr31), November 2010.

[3] W. B. DeOreo. California single-family water use efficiency study, July 2011. Aquacraft, Inc. Water Engineering and Management http://www.aquacraft.com/.

[4] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh. Querying and mining of time series data: experimental comparison of representations and distance measures. *Proc. VLDB Endow.*, 1(2):1542–1552, Aug. 2008.

[5] S. Firth, K. Lomas, A. Wright, and R. Wall. Identifying trends in the use of domestic appliances from household electricity consumption measurements. *Energy and Buildings*, 40(5):926 – 936, 2008.

[6] C. Fischer. Feedback on household electricity consumption: a tool for saving energy? *Energy Efficiency*, 1:79–104, 2008. 10.1007/s12053-008-9009-7.

[7] E. Keogh and J. Lin. Clustering of time-series subsequences is meaningless: implications for previous and future research. *Knowl. Inf. Syst.*, 8(2):154–177, Aug. 2005.

[8] T. W. Liao. Clustering of time series data: a survey. *Pattern Recognition*, 38(11):1857 – 1874, 2005.

[9] P. Rashidi, D. J. Cook, L. B. Holder, and M. Schmitter-Edgecombe. Discovering activities to recognize and track in a smart environment. *IEEE Transactions on Knowledge and Data Engineering*, 23:527–539, 2011.

[10] WCWA. Kalgoorlie smart metering trial frequently asked questions, Water Corporation of Western Australia, est. 2010. www.watercorporation.com.au.

[11] J. Widen, M. Lundh, I. Vassileva, E. Dahlquist, K. Ellegard, and E. Wackelgard. Constructing load profiles for household electricity and hot water from time-use data: Modelling approach and validation. *Energy and Buildings*, 41(7):753 – 768, 2009.

[12] Z. Xing, J. Pei, and E. Keogh. A brief survey on sequence classification. *SIGKDD Explor. Newsl.*, 12(1):40–48, Nov. 2010.

[13] L. Ye and E. Keogh. Time series shapelets: a novel technique that allows accurate, interpretable and fast classification. *Data Min. Knowl. Discov.*, 22(1-2):149–182, Jan. 2011.