Wearable Sensor-Based Sleep Quality Recommendation Based on Association Rules

Aleksander Romanha Santos
School of Electrical and Computer Engineering
University of Campinas, Brazil
a212271@dac.unicamp.br

Emely Pujólli da Silva
Institute of Computing
University of Campinas, Brazil
emelypujolli@gmail.com

Rosana Veroneze
School of Electrical and Computer Engineering
University of Campinas, Brazil
rveroneze@gmail.com

Fernando J. Von Zuben
School of Electrical and Computer Engineering
University of Campinas, Brazil
vonzuben@unicamp.br

***Abstract*—Frequent pattern mining and the subsequent proposition of association rules are used here to promote sleep quality from wearable data sensors. The idea is to automatically provide users with relevant and personalized information regarding possible contextual factors that impact their sleep, including physiological monitoring data. This objective was achieved by revealing existing associations between contextual factors and sleep characteristics to the user. Based on data provided by users of Samsumg Galaxy Watch4 over 21 days, the FPGrowth algorithm was used to mine frequent itemsets, followed by the generation of association rules. Due to the potentially large number of generated rules, a summarization is performed by the FPMax variant. FPMax mines the maximal frequent itemsets, thus generating a condensed version of the results capable of preserving as much information as possible. The obtained result is a set of association rules capable of disclosing the user's sleep patterns, thus indicating what to improve in his/her routine to get better sleep.**

***Index Terms*—Wearable Sensoring; Recommender System; Frequent Pattern Mining; Association Rules.**

I. Introduction

Several association mining problems are based on the detection of frequent co-occurrence of items of interest in a dataset. The appearance of items in the dataset in question, both individually or in conjunction with other items, may follow some patterns. Such patterns are not always obvious, intuitive, or easily detectable by simple inspection. Therefore, systematic methods for searching for such frequent patterns or, as they are commonly called, algorithms for mining frequent patterns, have become a focus of study in artificial intelligence [1]. Frequent pattern mining is therefore a technique capable of finding hidden patterns from data, thus revealing association rules used to establish a co-occurrence relationship of events that present a certain statistical interdependence [2].

It can be said that the rising of pattern mining was due to the emergence of the frequent itemset mining problem and the proposition of the classic Apriori algorithm developed in [3]. The algorithm was proposed having as motivation the resolution of the problem of shopping cart analysis, where the information was organized in a binary database and its operation was given by an exploration in amplitude of the candidate itemsets [3]. Other proposals inspired by Apriori were then conceived to search more efficiently for itemsets or frequent patterns.

A more significant performance gain was achieved with the emergence of algorithms that brought new approaches to the problem, as is the case of the FP-Growth algorithm [4]. The authors proposed "a new pattern tree structure (FP-tree), which is an extended prefix-tree structure for storing compressed and crucial information about frequent patterns, and develop an efficient FP-tree-based mining method, FP-growth, to mine the full set of frequent patterns by pattern fragment growth". With this, it was possible to obtain an execution speed approximately one order of magnitude faster than Apriori.

Another interesting approach was developed in [2], changing the way of indexing the dataset to make the support counting process more efficient, employing a vertical representation of the binary dataset. For this, the concept of tidset was proposed, which corresponds to the set of transactions that contain a certain itemset. This facilitates support computation because "the support of a candidate itemset can be calculated by the intersection of tidsets of subsets properly chosen". The ECLAT algorithm was conceived based on this concept. An update based on the use of diffsets (difference of tidsets) guided to improvements, making it possible to reduce the size of intermediate tidsets. This updated algorithm was named dECLAT [2].

Although these techniques have emerged in the context of shopping cart analysis, several other problems have similar structure, where frequent pattern mining algorithms fit appropriately. Therefore, several areas could benefit from extracting valuable information from large datasets by developing pattern mining tools. These include Web mining, software bug detec-

Part of the results presented in this article were carried out as part of the project "Hub de Inteligência Artificial em Saúde e Bem-Estar – VIVA BEM", funded by Samsung Eletrônica da Amazônia Ltda, under the terms of IT Law (Federal Law 8.248/91).

tion, image and multimedia data mining, chemical applications in toxicological evaluation, biological applications in RNA sequence analysis, and medical applications such as diagnosis from data [5].

In this paper, the focus will be on personalized recommender systems devoted to quality of sleep. Sleep analysis is an important field of medicine that studies the structure, duration and quality of an individual's sleep. With the growing amount of data generated by wearable devices, data mining techniques have been increasingly applied to identify specific patterns that can lead to a better understanding of the causes and effects of different sleep disorders. Those revealing patterns can be properly identified by frequent pattern mining techniques, thus allowing the extraction of relevant information for sleep analysis. Our approach will collect individual data from wearable data sensors, involving physiological and behavioral features considered relevant for the development of a personalized sleep recommender system. The aim is to disclose the most applicable factors and which actions should be taken to improve the quality of sleep.

Given the popularization and greater accessibility of wearables to the general public, especially smartwatches, several new solutions and features are being developed and even requested by users. With the aim of making sleep monitoring and analysis more accessible, something previously only done in rather fearsome clinics where the patient needs to use robust and not very comfortable devices, wearable technologies are sought to play a similar role and deliver equivalent results. Perhaps a wearable device is not yet capable of competing with the data acquisition procedures performed in a specialized clinic, but wearable technology is improving quickly to make such performance discrepancies as small as possible. Add to that the advantage and comfort of collecting data privately and during the patient's daily routine, indoors or outdoors. In fact, the developed applications do not have the role of replacing the already consolidated techniques that are the gold standard of diagnosis, monitoring, analysis, and evaluation of an individual's sleep health. Proposals such as the one in this paper may be interpreted as an auxiliary and non-invasive technology devoted to personalized usage.

II. Main scope of the research

It was observed that the existing mobile apps devoted to sleep monitoring and recommendation do not deliver functionalities with the scope covered in this work. They generally exhibit qualitative statistics regarding sleep, bringing relevant information about sleep stages and overall quality of the sleep experience. However, the recommendations to help the user sleep better, when the application provides such a service, are somewhat generic and almost without individual customization. Aiming at improving the functionalities of smartwatch sleep applications and making them more user-specific, we automatically extract multiple datasheets produced by the wearable device when monitoring events along the daily activities and during sleep. Those datasheets feed frequent pattern mining techniques capable of extracting association rules of high confidence and high coverage, which can be used to support personalized recommendations toward sleep quality improvement.

It is known that sleep quality is strongly influenced by several contextual factors, such as consumption of alcohol [6] [7] or caffeine [8], fulfillment of physical activity [9] [10], eating snacks late at night, skipping the breakfast meal [11] [12], taking naps during the day, among others [13]. Furthermore, such factors and their impact on a person's sleep quality tends to be user-specific: a given contextual factor can influence the sleep of different individuals at different levels and even at different directions. With that in mind, this work seeks to relate contextual factors to sleep metrics through association rules, to provide the user with possible inducement for having slept well or poorly. Such results will help the user in a personalized way to identify what he/she can change in his/her routine or lifestyle to get better sleep.

III. Problem Definition, Methods and Objectives

The problem can be characterized by a set of data collected daily, consisting of contextual variables and sleep variables. We can define a set D that contains the data records collected daily from a user:

$$D = \{D^1,\ D^2,\ D^3,\ ...,\ D^t\}$$

Such daily data correspond to a set of contextual variables X and a set of sleep variables Y:

$$\begin{aligned} D = \{D^1,\ D^2,\ D^3,\ ...,\ D^t\} = \\ \{\{X^1,\ Y^1\},\ \{X^2,\ Y^2\},\ \{X^3,\ Y^3\},\ ...,\ \{X^t,\ Y^t\}\}\end{aligned}$$

The daily contextual variables used were: the time the person starts to sleep, if the person ate a late-night snack, if the person skipped breakfast, if the person consumed caffeine after 6PM, if the person ingested alcohol, if the person took a nap longer than 30 minutes, the daily step count and the average speed of these steps. The sleep variables used were: the number of awakenings during sleep, sleep duration, and sleep efficiency. So, the contextual and sleep data can be represented as follows given a day t:

$$\begin{aligned} X^t = \{X^t_{BedTime},\ X^t_{NightSnack},\ X^t_{SkippedBreakfast}, \\ X^t_{Caffeine},\ X^t_{Alcohol},\ X^t_{Nap},\ X^t_{Steps},\ X^t_{Speed}\}\end{aligned}$$

$$Y^t = \{Y^t_{NumberAwakenings},\ Y^t_{SleepDuration},\ Y^t_{SleepEfficiency}\}$$

The objective, therefore, is to automatically find association rules from the user data, which relate contextual variables $X_{Context} \subseteq X$ with some sleep variable $Y_{SleepMetric} \subseteq Y$ and $|Y_{SleepMetric}| = 1$ [14] [15]:

$$X_{Context} \rightarrow Y_{SleepMetric}$$

The values of sleep variables will be classified as positive or negative w.r.t. sleep quality. For this, a threshold chosen based on health recommendations will be defined. Thus, the mined association rules will have the role of relating contextual factors with a good or bad aspect of the user's sleep, thus indicating the possible associated factors responsible for a good or bad sleep experience, according to the sleep metrics:

$$X_{Context} \rightarrow Y^{+}_{SleepMetric} \quad ; \quad X_{Context} \rightarrow Y^{-}_{SleepMetric}$$

IV. Case Study

An experimental design is proposed with data collected from the Samsung Galaxy Watch4 of several participants (duly approved terms of ethics: CAAE 55532622.0.0000.5404), using the Samsung Health and Samsung Sleep Coaching applications. We will present results from two participants, to illustrate the user-specific nature of the approach. Participant 1 is a healthy individual and Participant 2 is an individual with somnambulism. In case of Patient 2, we could not verify if it occurred a somnambulism event in the period of data capture.

For both participants, data capture was performed during a complete period of Sleep Coaching, which lasts four weeks. Among these weeks, the first is used to calibrate and define the user's profile. From the second week onwards, the application begins to collect relevant contextual information from the user's daily life, which is essential for the proposed methodology.

The necessary information was extracted from four datasheets provided by the Samsung Health and Samsung Sleep Coaching applications, namely, *Sleep_Stage.csv, Sleep_Data.csv, Sleep_Coaching_Mission.csv* and *Step_Daily_Trend.csv*. Next, there will be a brief description of each datasheet, which variables were extracted, and how they were obtained. For illustrative purposes, the data fragments of Tables I to VII are from Participant 1.

A. Samsung Galaxy Watch4 Data

The *Sleep_Stage.csv* spreadsheet (Table I) shows each user's sleep stage associated with a given sleep experience, that is, a given *sleep_id*, as well as the start and end time and date of each stage. The sleep stages are represented by the codes 40001, 40002, 40003 and 40004, which correspond to the awake, light sleep, deep sleep and REM sleep stages, respectively. From these data, the variables $X_{BedTime}$ and $Y_{NumberAwakenings}$ are extracted. $X_{BedTime}$ is given by the start time of the first sleep stage in temporal order for each *sleep_id*. $Y_{NumberAwakenings}$ is calculated for each *sleep_id* by counting how many times the awakened stage occurs, that is, how many times the code 40001 appears.

The *Sleep_Data.csv* worksheet (Table II) is based on the previous worksheet and calculates some sleep metrics, such as sleep efficiency, sleep duration and time in the awake stage. Such information is extracted directly from the spreadsheet in the columns "sleep_duration", "efficiency" and "factor_05" for each corresponding sleep ("sleep_id"), providing the variables $Y_{SleepDuration}$, $Y_{SleepEfficiency}$ and $Y_{AwakenDuration}$. The latter is used only as an auxiliary variable to make some adjustments to the data, which will be explained later.

Sleep_Coaching_Mission.csv (Table III) shows, among other missions, the daily questions about contextual factors of the user's routine, in which he/she marks in the application which factors occurred that day. These questions are identified by the code "SP_MSN_01" in the column "mission_id" and each question has an associated answer, given by the column "answer". Responses are related to whether the person ate a late night snack, skipped breakfast, consumed caffeine after 6 pm, ingested alcohol or took a nap longer than 30 minutes. These answers are equivalent, respectively, to the codes "SP_SVY1_ANS_01", "SP_SVY1_ANS_02", "SP_SVY1_ANS_03", "SP_SVY1_ANS_04" and "SP _SVY1_ANS_05". Then, from this worksheet, it was extracted the variables $X_{NightSnack}$, $X_{SkippedBreakfast}$, $X_{Caffeine}$, $X_{Alcohol}$ and X_{Nap}.

Finally, the *Step_Daily_Trend.csv* spreadsheet (Table IV) presents a data entry for each physical activity event, thus indicating the date that occurred, represented by the column "day_time", the count of steps, indicated by the column "count", the distance covered, indicated by the column "distance" and the average speed of these steps, indicated by the column "speed". The variables X_{Steps} are then extracted, given by the sum of the entries in column "count" for the same day, X_{Speed}, given by the calculation of the average speed for the same day and $X_{Distance}$, given by the sum of entries in column "distance" for the same day ("day_time"). This last variable will also be used only to help in later adjustments.

B. Data Consolidation

After extracting the variables, it will be necessary to aggregate them in the same worksheet. Initially, data will be cleaned and synchronized, where the contextual variables must be related to the corresponding sleep experience, under the same index. However, data from different spreadsheets may have different indexing. Sleep data may show more than one entry corresponding to the same day, indicating that the person slept more than once that day, or no entries on a given day, indicating that the person did not sleep that day or that the smartwatch could not collect the data. Bearing in mind that contextual factors are related to sleep data on a daily scale, indexing will be based on the date the sleep experience occurred and the contextual factors. Thus, if the person has slept more than once during a given day, the contextual data referring to that day will be associated with such sleep data entries. The dataframe will have a row of sleep data for each sleep experience, indicating the date it occurred, if there is more than one sleep event on the same date, they will have the same associated contextual data, which refers to the date in question. When there is no sleep data on a given day, the contextual data for that day will be disregarded, as there is no corresponding sleep experience entry.

Another situation to be treated is when a sleep experience occurs at unconventional times, for example, when the person goes to sleep at 9AM. A question arises: What contextual data should this sleep experience be associated with? The contextual data of the previous or the current day? It would be reasonable to consider that the contextual data referring to this sleep experience are the contextual data of the previous day since they were the data that occurred immediately before the referred sleep experience. In addition, if the contextual data of the day on which such sleep experience occurred were considered, a large part of the contextual data would have been collected in the rest of the day, after the sleep experience, with no influence, therefore, on the 9AM event, but on the

start_time	sleep_id	stage	time_offset	end_time
2022-05-08 03:31:00.000	b46ad7e7-959a-4af6-9932-ec73527a53ff	40002	UTC-0300	2022-05-08 03:54:00.000
2022-05-08 03:54:00.000	b46ad7e7-959a-4af6-9932-ec73527a53ff	40003	UTC-0300	2022-05-08 04:01:00.000
2022-05-08 04:01:00.000	b46ad7e7-959a-4af6-9932-ec73527a53ff	40001	UTC-0300	2022-05-08 04:03:00.000
2022-05-08 04:03:00.000	b46ad7e7-959a-4af6-9932-ec73527a53ff	40002	UTC-0300	2022-05-08 04:12:00.000
2022-05-08 04:12:00.000	b46ad7e7-959a-4af6-9932-ec73527a53ff	40004	UTC-0300	2022-05-08 04:30:00.000

TABLE I: Fragment of the Data Sheet *Sleep_Stage.csv*

factor_05 [min]	efficiency	sleep_duration [min]	com.samsung.health.sleep.datauuid
29.0	93.0	448.0	b46ad7e7-959a-4af6-9932-ec73527a53ff
43.0	91.0	485.0	4aa144d3-e031-45fc-9298-9f2025ad8032
29.0	93.0	483.0	0b58993a-5cd8-4608-b68a-4f1785f57725
32.0	93.0	490.0	61764b32-a4e7-4dba-8dca-d834174d8dde
31.0	92.0	397.0	88391f03-4a41-487b-b8bc-e773d902399f

TABLE II: Fragment of the Data Sheet *Sleep_Data.csv*

answer	day_index	mission_id
[“SP_SVY1_ANS_01”]	7	SP_MSN_01
[“SP_SVY1_ANS_01”]	8	SP_MSN_01
[“SP_SVY1_ANS_01”]	9	SP_MSN_01
[“SP_SVY1_ANS_01”]	10	SP_MSN_01
[“SP_SVY1_ANS_01”,“SP_SVY1_ANS_03”]	11	SP_MSN_01

TABLE III: Fragment of the Data Sheet *Sleep_Coaching_Mission.csv*

count	speed [m/s]	distance [m]	time [s]	day_time
108	1.601810	74.979996	46.809547	2018-04-12
108	1.071053	74.979996	70.005876	2018-04-12
17	1.166667	12.050000	10.328572	2018-04-13
17	1.070036	12.050000	11.261299	2018-04-13
42	1.782949	31.530000	17.684182	2018-04-14

TABLE IV: Fragment of the Data Sheet *Step_Daily_Trend.csv*

next sleep experience instead. Taking this into account, it was defined that if the sleep experience occurs before 12:00, it will be associated with the contextual data of the previous day and if the sleep experience occurs after 12:00, it will be associated with the contextual data of the current day. It is worth mentioning that an extreme case was considered here as an example, and such a situation is not expected to occur frequently.

To address these two issues, a global index “day_index” will be created, under which all sleep data and contextual data will be consolidated into a single dataframe, respecting the considerations made earlier. Therefore, the dataframe will have an appearance similar to the one shown in Table V.

Hereafter, it was observed that some consecutive data entries had a small time difference between the end of the first entry and the beginning of the second data entry. Probably the smartwatch subdivided the data corresponding to a single sleep experience into more than one entry. This could be due to the person moving a lot in their sleep for a relatively long period and the watch identifying that the person is awake, or that the person could have gone to the bathroom. Anyway, such entries should not be subdivided, as they correspond to the same sleep experience. Therefore, data processing was carried out to detect where such a situation occurs and aggregate such subdivided entries into a single sleep experience entry.

The histogram in Figure 1 shows the time intervals between consecutive sleep experience entries and the number of times they occurred for both participants. Similar behavior was observed for both participants individually, motivating the joint presentation. A majority set of occurrences can be seen around 1000 minutes, which corresponds to a normal spacing time between sleep (about 16 hours), and separate from this core there are cases farther apart. Some of them have higher time values, reaching up to 2000 or 6000 minutes, which are probably caused by the absence of a sleep entry on one or more consecutive days, probably due to the user not having used the smartwatch on that day. On the other hand, there is a set of occurrences close to zero. When zooming the figure in this region (Figure 2), it is verified that most time values are below 15 minutes. Apparently, such occurrences are possible misinterpretations of the smartwatch, and that is the time it takes to recognize again that the person is asleep. Therefore, data processing will be done with the aim of correcting such occurrences, in which the time interval between sleep experiences is very low and it is assumed that it is a misinterpretation of the data made by the mobile device.

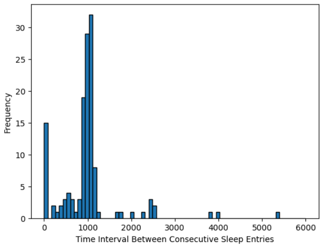

Fig. 1: Histogram of time intervals between consecutive sleep entries.

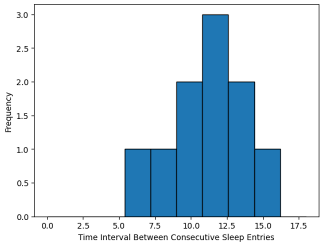

Fig. 2: Zoom histogram of time intervals between consecutive sleep entries.

To detect these cases, information on the time of onset of sleep is adopted, using the variable $X_{BedTime}$, and the duration of sleep, using the variable $Y_{SleepDuration}$, to identify the time at which that sleep experience entry ended. A time criterion was adopted for the difference between the time at which a sleep entry ends and the start time of the consecutive sleep entry, in order to define whether such entries will be aggregated or not. If this time difference is up to 15 minutes between consecutive entries, they will be considered referring to a single sleep and will be aggregated. To merge the entries, it will be necessary to recalculate the variables $Y_{NumberAwakenings}$, $Y_{AwakeDuration}$, $Y_{SleepDuration}$ and $Y_{SleepEfficiency}$.

bed_time	number_of_ awakenings	awake_ duration [min]	sleep_ duration [min]	efficiency	day_index	count	distance [m]	time [s]	speed [m/s]	daily_answer
2022-07-02 04:04:00	7	10.0	178.0	94.0	2022-07-01	6607.0	4862.430700	3575.013244	1.582654	[“SP_SVY1_ANS_01”]
2022-07-03 01:29:00	16	40.0	526.0	92.0	2022-07-02	9474.0	6564.296500	5471.145669	1.722222	[“SP_SVY1_ANS_01”, “SP_SVY1_ANS_03”]
2022-07-03 14:08:00	2	11.0	138.0	92.0	2022-07-03	13470.0	9318.070000	7956.124353	1.722222	[“SP_SVY1_ANS_01”]
2022-07-04 01:58:00	23	45.0	457.0	90.0	2022-07-03	13470.0	9318.070000	7956.124353	1.722222	[“SP_SVY1_ANS_01”]
2022-07-05 00:59:00	3	6.0	116.0	94.0	2022-07-04	14401.0	10255.270300	7756.301847	1.750000	[“SP_SVY1_ANS_01”]
2022-07-05 03:04:00	18	47.0	410.0	88.0	2022-07-04	14401.0	10255.270300	7756.301847	1.750000	[“SP_SVY1_ANS_01”]
2022-07-06 02:18:00	19	32.0	488.0	93.0	2022-07-05	18604.0	13142.572702	9873.594097	1.248937	[“SP_SVY1_ANS_01”]

TABLE V: Dataframe fragment with consolidated data

The $Y_{NumberAwakenings}$ values of the subdivided entries will be summed and added by 1, to also consider the agreed period between the two subdivided entries. For $Y_{AwakeDuration}$, the sum of their values corresponding to the subdivided entries plus the time interval between these sleep entries, which the person spent in the awake stage, will be made. Likewise, $Y_{SleepDuration}$ will be the sum of the sleep times of the subdivided entries. And $Y_{SleepEfficiency}$ will be recalculated using the new values of $Y_{AwakeDuration}$ and $Y_{SleepDuration}$, according to the following equation to obtain sleep efficiency:

$$Y_{SleepEfficiency} = \frac{Y_{SleepDuration} - Y_{AwakeDuration}}{Y_{SleepDuration}}$$

Then, the data is consolidated into a single dataframe, with the necessary indexing corrections and data processing. The same dataframe snippet from Table V is shown, after treatment, in Table VI. Note that the rows corresponding to “day index” equal to “2022-07-04” in Table V were summed, which resulted in Table VI.

C. Preparing for Mining Frequent Patterns and Generating Association Rules

Before supplying the data to the algorithm, the variables will need to be discretized and, subsequently, the database must be converted to a binary database. With the exception of the variables $X_{NightSnack}$, $X_{SkippedBreakfast}$, $X_{Caffeine}$, $X_{Alcohol}$ and X_{Nap}, which are already discrete variables, the others will be discretized according to the subsequent description.

The variable $X_{BedTime}$ will be categorized into four-time ranges: *bed times between 0 and 6 hours*, *6 and 12 hours*, *12 and 18 hours* or *18 and 24 hours*. The variables X_{Steps} and X_{Speed} will be discretized into three equally spaced categories, namely *low steps*, *medium steps* or *high steps* for the variable X_{Steps}, and *low speed*, *medium speed* or *high speed* for the variable X_{Speed}.

As for the sleep variables, they will be categorized into two classes, one to indicate a good aspect of sleep and the other a bad aspect of sleep. For sleep efficiency, a threshold of 90% was adopted, in which $Y_{SleepEfficiency}$ will be categorized as *low sleep efficiency*, if it is below 90%, or *high sleep efficiency*, if it is greater than or equal to 90%:

$$If\ Y_{SleepEfficiency} < 90\% \longrightarrow low\ sleep\ efficiency$$

$$If\ Y_{SleepEfficiency} \geq 90\% \longrightarrow high\ sleep\ efficiency$$

In $Y_{NumberAwakenings}$ a threshold of 10 was used. This apparently high value for the number of awakenings during sleep is consistent with the accuracy presented by the smartwatch in measuring this variable, most of the time overestimating this value. This overestimation can be verified by the histograms of Figures 3 and 4, generated from the data of Participants 1 and 2, respectively. Even in the case of the healthy individual, it is noted that the values are exacerbated, in which the region with the highest frequency of the number of awakenings is around 15 to 20. In addition, the average of the values of this variable for both participants was calculated and a value of approximately 14 was obtained. It is known that a healthy number of times an adult wakes up during sleep is up to 4 times, on average [16]. So, we chose to use the average value between the last two values (the average number of awakenings observed in the smartwatch collection and the average number of awakenings considered healthy), resulting in 9.

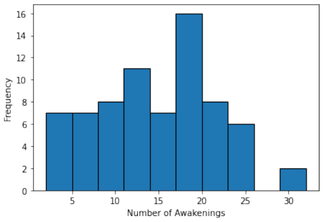

Fig. 3: Histogram of the variable $Y_{NumberAwakenings}$ for Participant 1.

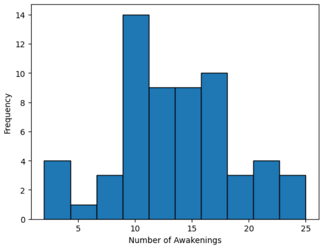

Fig. 4: Histogram of the variable $Y_{NumberAwakenings}$ for Participant 2.

Therefore, if the variable $Y_{NumberAwakenings}$ has a value less than 10, it will be classified as *low number of awakenings* and if it is greater than or equal to 10, it will be classified as *high number of awakenings*:

$$If\ Y_{NumberAwakenings} < 10 \longrightarrow low\ number\ of\ awakenings$$

$$If\ Y_{NumberAwakenings} \geq 10 \longrightarrow high\ number\ of\ awakenings$$

For variable $Y_{SleepDuration}$ it was defined that if the sleep duration is between 360 minutes (6 hours) and 540 minutes (9 hours), it will be classified as *good sleep duration* and otherwise it will be classified as *bad sleep duration*:

$$If\ 360min \leq Y_{NumberAwakenings} \leq 940min \longrightarrow good\ sleep\ duration$$

$$If\ Y_{NumberAwakenings} < 360min\ or\ Y_{NumberAwakenings} > 940min \longrightarrow bad\ sleep\ duration$$

Then, the discrete database is transformed into a binary database, in which each of the discretization ranges will be considered as an item, given that we want to use the frequent itemset mining approach. Thus, the binary database will have

bed_time	**number_of_ awakenings**	**awake_ duration [min]**	**sleep_ duration [min]**	**efficiency**	**day_index**	**count**	**distance [m]**	**time [s]**	**speed [m/s]**	**daily_answer**
2022-06-30 01:59:00	16	52.0	480.0	89.000000	2022-06-29	5827.0	4087.689910	3262.274476	4.250000	[“SP_SVY1_ANS_01”]
2022-07-01 01:52:00	17	24.0	482.0	95.000000	2022-06-30	3102.0	2143.139994	1825.603501	3.361111	[“SP_SVY1_ANS_01”]
2022-07-02 04:04:00	7	10.0	178.0	94.000000	2022-07-01	6607.0	4862.430700	3575.013244	1.582654	[“SP_SVY1_ANS_01”]
2022-07-03 01:29:00	16	40.0	526.0	92.000000	2022-07-02	9474.0	6564.296500	5471.145669	1.722222	[“SP_SVY1_ANS_01”, “SP_SVY1_ANS_03”]
2022-07-03 14:08:00	2	11.0	138.0	92.000000	2022-07-03	13470.0	9318.070000	7956.124353	1.722222	[“SP_SVY1_ANS_01”]
2022-07-04 01:58:00	23	45.0	457.0	90.000000	2022-07-03	13470.0	9318.070000	7956.124353	1.722222	[“SP_SVY1_ANS_01”]
2022-07-05 00:59:00	21	62.0	526.0	88.212928	2022-07-04	14401.0	10255.270300	7756.301847	1.750000	[“SP_SVY1_ANS_01”]

TABLE VI: Dataframe fragment with consolidated data after data processing.

the items represented by columns and the rows will be the tids (transaction identifiers), which correspond to each of the sleep data entries. A snippet from this database is shown in Table VII, where the presence of an item, that is, the presence of some contextual or sleep factor in a given data entry is indicated by bit 1 positioned in the row and column corresponding to the data entry.

Now we can provide the binary database to the frequent itemset mining algorithm, which will return the set of frequent itemsets and, with these itemsets, generate the association rules.

D. Mining Frequent Patterns and Generating Association Rules

The initial proposal would be to use some frequent itemset mining algorithm, such as FP-Growth, and from all frequent itemsets found generate the association rules. However, a relatively large number of mined rules was observed and many of them did not bring new information or were even redundant with the others. Therefore, the results were summarized, aiming at presenting to the user the essential information in a concise manner.

Some means of summarization were studied, such as the mining of association rules using closed frequent itemsets [17] and the mining of minimal and non-redundant association rules using closed frequent itemsets [18]. In both cases, the concept of closed itemsets is used, which is a method of summarizing itemsets in which itemsets that do not have supersets with the same support value are obtained. In the first case, this set of closed itemsets is used to generate the association rules, but this approach is not adequate for the problem studied in this work. The closed itemsets obtained mostly present, due to summarization, more than one sleep variable grouped in the same itemset. In this way, no generated rule will meet the requirements of presenting only one sleep variable in the consequent and no one in the antecedent (contextual variables only). The second approach also uses closed itemsets, but the purpose is to obtain rules whose antecedent is minimal and consequent is maximal. Such an approach is clearly inadequate for the problem of this work, in which contextual factors in the antecedent associated with only one aspect of sleep in the consequent are desired.

Therefore, a summarization of the rules using the concept of maximal itemsets was proposed. Maximal itemsets can be characterized as a set of frequent itemsets, which do not have supersets that are also frequent. Therefore, any subset of a maximal itemset will also be a frequent itemset [2]. Thus, the set of maximal itemsets is a reduced representation of the complete set of frequent itemsets.

This approach will be used to generate rules whose antecedents are maximal, that is, among the set of itemsets corresponding to the contextual variables, only those itemsets that are maximal will be kept [19]. It is noteworthy that the closed itemsets could also be used only in the antecedent, similarly to this proposal, to obtain only the rules whose antecedent parts were closed itemsets. However, the maximal itemset approach was chosen due to its greater ability to reduce or generalize frequent itemsets. This is because, if we consider for the same database its sets of maximal itemsets M, closed itemsets C and frequent itemsets F, we have that $M \subseteq C \subseteq F$.

The FP-Max algorithm was used to find the maximal frequent itemsets, considering only the contextual variables, that is, those that will compose the antecedent part of the rules. The chosen Support threshold was 0.1 ($minSup = 0.1$). Through the mined maximal itemsets, a database check was made to locate each transaction (data entry or row) in which such itemsets appear. For each maximal itemset and its corresponding transactions, it was verified for each sleep variable (itemset of one element) in how many of these specific transactions the variable is also present. With this, we have the support value of both the antecedent $Sup(X)$ (number of transactions in which the maximal itemset of the contextual variables appears in the database) and the antecedent-consequent set $Sup(X \cup Y)$ (number of transactions that jointly contain the maximal itemset of the antecedent and the sleep variable), and it is possible to calculate the confidence of the rule formed by the antecedent and consequent in question. A confidence threshold of 0.9 was defined ($minConf = 0.9$), so if the analyzed antecedent-consequent combination has confidence greater than 0.9, it will be mined and the association rule will be generated. The choice of these support and confidence thresholds was made with the intention that sporadic events, but which are possibly highly significant, could also be accepted.

At the end of this process, a set of association rules will be obtained, where the antecedent parts will be maximal and formed by a set of contextual variables and the consequent parts will be formed solely by a sleep variable. The rules generated from the data of each analyzed participant can be seen below. It is noted that the rules obtained manage to extract patterns from the data and inform the user of the possible factors that most influence their sleep behavior in a summarized and concise way.

ate a late-night snack	skipped breakfast	drank caffeine after 6PM	drank alcohol	took a nap longer than 30min	none of the above	bed time [0-6]	bed time [6-12]	bed time [12-18]	bed time [18-24]	high number of awakenings	low number of awakenings	good sleep duration	bad sleep duration	high sleep efficiency	low sleep efficiency	low number steps	medium number steps	high number steps	low speed	medium speed	high speed
1	0	0	0	0	0	1	0	0	0	1	0	1	0	1	0	1	0	0	1	0	0
1	0	0	0	0	0	0	0	1	0	0	1	0	1	0	1	1	0	0	0	0	1
1	0	0	0	0	0	1	0	0	0	1	0	1	0	0	1	1	0	0	0	0	1
1	0	0	0	0	0	1	0	0	0	1	0	1	0	1	0	1	0	0	0	0	1
1	0	0	0	0	0	1	0	0	0	0	1	0	1	1	0	1	0	0	1	0	0
1	0	1	0	0	0	1	0	0	0	1	0	1	0	1	0	0	1	0	1	0	0
1	0	0	0	0	0	0	0	1	0	0	1	0	1	1	0	0	0	1	1	0	0

TABLE VII: Fragment of database converted to binary matrix.

- Participant 1:

1) $\{bed\ time\ [0-6],\ low\ speed,\ medium\ number\ steps\} \longrightarrow \{high\ sleep\ efficiency\}$
2) $\{ate\ a\ late-night\ snack,\ bed\ time\ [0-6],\ medium\ number\ steps\} \longrightarrow \{high\ sleep\ efficiency\}$
3) $\{bed\ time\ [12-18],\ ate\ a\ late-night\ snack\} \longrightarrow \{low\ number\ of\ awakenings\}$
4) $\{bed\ time\ [12-18],\ ate\ a\ late-night\ snack\} \longrightarrow \{bad\ sleep\ duration\}$
5) $\{bed\ time\ [12-18],\ low\ speed\} \longrightarrow \{low\ number\ of\ awakenings\}$
6) $\{bed\ time\ [12-18],\ low\ speed\} \longrightarrow \{bad\ sleep\ duration\}$
7) $\{ate\ a\ late-night\ snack,\ high\ number\ steps,\ bed\ time\ [0-6],\ low\ speed\} \longrightarrow \{high\ number\ of\ awakenings\}$

- Participant 2:

1) $\{drank\ caffeine\ after\ 6PM,\ bed\ time\ [0-6],\ ate\ a\ late-night\ snack\} \longrightarrow \{high\ number\ of\ awakenings\}$
2) $\{drank\ caffeine\ after\ 6PM,\ bed\ time\ [0-6],\ ate\ a\ late-night\ snack\} \longrightarrow \{good\ sleep\ duration\}$
3) $\{drank\ caffeine\ after\ 6PM,\ bed\ time\ [0-6],\ ate\ a\ late-night\ snack\} \longrightarrow \{low\ sleep\ efficiency\}$
4) $\{bed\ time\ [0-6],\ high\ speed,\ ate\ a\ late-night\ snack\} \longrightarrow \{high\ number\ of\ awakenings\}$
5) $\{bed\ time\ [0-6],\ high\ speed,\ ate\ a\ late-night\ snack\} \longrightarrow \{good\ sleep\ duration\}$
6) $\{bed\ time\ [0-6],\ high\ speed,\ ate\ a\ late-night\ snack\} \longrightarrow \{low\ sleep\ efficiency\}$
7) $\{bed\ time\ [0-6],\ low\ speed\} \longrightarrow \{high\ sleep\ efficiency\}$
8) $\{drank\ caffeine\ after\ 6PM,\ medium\ speed,\ bed\ time\ [0-6]\} \longrightarrow \{low\ sleep\ efficiency\}$

V. Results

The obtained results comprise seven association rules for Participant 1 and nine association rules for Participant 2, thus revealing a clear scenario involving contextual factors and sleep quality standards. Notice that the patterns mined for both participants are indeed personalized, with a very distinct composition in each case.

Participant 1 is a healthy individual, for whom sleep disorders are not expected. Therefore, association rules involving low quality sleep are not expected to be frequent for this user. The outcome was able to confirm this aspect, guiding mostly to rules involving positive sleep factors. Note a low number of awakenings in the first two rules and a high sleep efficiency in the fifth and sixth rules.

The interesting thing about this method is that the algorithm is able to identify patterns that are not so obvious. Even for a healthy individual, it is possible to capture events of low-quality sleep and associate them with co-occurring contextual factors. The last rule stands out, indicating that a high number of steps, a low speed of steps, eating a snack late at night and going to bed between 0AM and 6AM are associated with a high number of awakenings during sleep:

$$7)\ \{ate\ a\ late-night\ snack,\ high\ number\ steps,\ bed\ time\ [0-6],\ low\ speed\} \longrightarrow \{high\ number\ of\ awakenings\}$$

This rule lists a set of contextual factors that occurred with a significantly high frequency when the person woke up many times during sleep. This association may guide to adjustments in the daily routine or lifestyle. In the case of the rule in evidence, walking a high number of steps at low speed, possibly characterizing a physical activity to face sedentary lifestyle, is generally not a productive decision for this particular individual, given the immediate negative impact on sleep quality. So, for this user it might be interesting to increase the intensity of the exercises in order to reduce the number of awakenings during sleep. Also, the person should try to avoid late-night snacks and going to bed between 0AM and 6AM.

On the other hand, one finds a completely different set of association rules for Participant 2. In this case, it is observed that most of the rules are related to negative aspects of sleep, being mainly linked to a high number of awakenings and low sleep efficiency. Given the non-invasive nature of our methodology, episodes of sleepwalking during the experiment period were not monitored. Nonetheless, the high-frequency occurrence of sleeps with low efficiency and a high number of awakenings may be an indication of the occurrence of such a sleep disorder during the period of study. But regardless of that, the important thing is that the algorithm is able to extract patterns and deliver valuable information to the user. Participant 2 is fed with pertinent information to better understand his/her own sleep behavior, together with the association with contextual factors.

In fact, many of the results provided the relationship between bad aspects of sleep and contextual factors, thus creating the opportunity for Participant 2 to revise his/her daily routine, aiming at mitigating the factors that tend to impair sleep quality. Consider, for instance, rules number 1 and 3, which have the same antecedent:

$$1)\ \{drank\ caffeine\ after\ 6PM,\ bed\ time\ [0-6],\ ate\ a\ late-night\ snack\} \longrightarrow \{high\ number\ of\ awakenings\}$$

$$3)\ \{drank\ caffeine\ after\ 6PM,\ bed\ time\ [0-6],\ ate\ a\ late-night\ snack\} \longrightarrow \{low\ sleep\ efficiency\}$$

Such rules reveal that whenever this individual has ingested caffeine after 6PM, eaten a snack at night, and slept between 0AM and 6AM, there is a high correlation between waking up many times at night and having a low sleep efficiency. Such sleep measures may be linked to sleepwalking, which is characterized by “Repeated episodes of rising from bed during sleep and walking about, usually beginning during the first third of the major sleep episode, that typically are brief, lasting 1–10 minutes” [20]. And effectively the association rules show the possible causes that may be intensifying such harmful aspects of sleep.

Notice that the rules capture patterns regardless of whether they relate to positive or negative aspects of sleep. Therefore, rule number 2, having good sleep duration at the consequent part, and sharing the same antecedent part as the rules discussed above, shows the user that in terms of sleep duration, a good performance is presented and that such contextual factors, despite having a negative influence on the efficiency and number of awakenings, apparently does not influence the sleep duration of this specific individual.

The contrast between the automatically obtained association rules for the two individuals evidences the personalized character of the results. Therefore, one fundamental aspect of the proposed recommender system is its personal scope. This investigation confirms that the main objectives of the research were achieved with a non-invasive and personalized wearable sensor-based sleep quality recommendation system.

VI. Conclusions and Future Works

In this work, we start from multiple datasheets produced by the operation of wearable sensors available in the Samsung Galaxy Watch4, provided by the Samsung Health and Samsung Sleep Coaching applications. After straightforward preprocessing steps of the 21 days of data collection, frequent pattern mining and filtering techniques were applied to guide to a reduced set of association rules, capable of revealing the influence of contextual factors on an individual's sleep quality.

Besides being a non-invasive methodology, many functionalities were added to the current state-of-the-art apps provided by smartwatches, not restricted to the Samsung device. The most prominent aspect is the intrinsic personalized nature of the results, helping the user to identify the most relevant associations between contextual factors and sleep quality. The main outcome are association rules, which can automatically capture patterns from data and establish co-occurrence relationships between events. Being a reduced set (only maximal itemsets were considered) and highly interpretable information, the obtained set of association rules turns to be a direct recommendation of what the user should or should not do to get a better sleep experience. In this way, those explainable results may sustain big changes or even little adjustments in the user daily routine.

The FP-Max algorithm was used to find the maximal frequent itemsets from contextual variables. From them, it was verified in the database which sleep variables, if added as a consequent, would generate strong association rules. Then, rules that met the minimum confidence requirements were generated. Datasets produced by two participants with distinct profiles were considered in the experiments: a healthy individual and an individual diagnosed with sleepwalking. The obtained results evidenced the unique aspect of the algorithm in delivering personalized recommendations for each user.

Future works are being conceived to further improve the results. The application of the developed methodology to a larger set of participants, mainly involving individuals with sleep disorders, would be interesting to verify the behavior in a greater diversity of data. In addition, refining the process of summarizing the association rules, adding new variables potentially conditioning sleep quality, and using a higher degree in the discretization of variables are aspects that can enrich the recommendation.

References

[1] J. M. Luna, P. Fournier-Viger, and S. Ventura, "Frequent itemset mining: A 25 years review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 6, p. e1329, 2019.

[2] M. J. Zaki and W. Meira, *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, 2014.

[3] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, 1993, pp. 207–216.

[4] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," *ACM Sigmod Record*, vol. 29, no. 2, pp. 1–12, 2000.

[5] C. C. Aggarwal, "Applications of frequent pattern mining," *Frequent Pattern Mining*, pp. 443–467, 2014.

[6] T. Roehrs and T. Roth, "Sleep, sleepiness, and alcohol use," *Alcohol Research & Health*, vol. 25, no. 2, p. 101, 2001.

[7] S.-Y. Park, M.-K. Oh, B.-S. Lee, H.-G. Kim, W.-J. Lee, J.-H. Lee, J.-T. Lim, and J.-Y. Kim, "The effects of alcohol on quality of sleep," *Korean Journal of Family Medicine*, vol. 36, no. 6, p. 294, 2015.

[8] C. Drake, T. Roehrs, J. Shambroom, and T. Roth, "Caffeine effects on sleep taken 0, 3, or 6 hours before going to bed," *Journal of Clinical Sleep Medicine*, vol. 9, no. 11, pp. 1195–1200, 2013.

[9] A. N. S. Bisson, S. A. Robinson, and M. E. Lachman, "Walk to a better night of sleep: testing the relationship between physical activity and sleep," *Sleep Health*, vol. 5, no. 5, pp. 487–494, 2019.

[10] M. A. Kredlow, M. C. Capozzoli, B. A. Hearon, A. W. Calkins, and M. W. Otto, "The effects of physical activity on sleep: a meta-analytic review," *Journal of behavioral medicine*, vol. 38, pp. 427–449, 2015.

[11] M. E. Faris, M. V. Vitiello, D. N. Abdelrahim, L. Cheikh Ismail, H. A. Jahrami, S. Khaleel, M. S. Khan, A. Z. Shakir, A. M. Yusuf, A. A. Masaad *et al.*, "Eating habits are associated with subjective sleep quality outcomes among university students: findings of a cross-sectional study," *Sleep and Breathing*, pp. 1–12, 2021.

[12] C. A. Crispim, I. Z. Zimberg, B. G. dos Reis, R. M. Diniz, S. Tufik, and M. T. de Mello, "Relationship between food intake and sleep pattern in healthy individuals," *Journal of clinical sleep medicine*, vol. 7, no. 6, pp. 659–664, 2011.

[13] L. A. Irish, C. E. Kline, H. E. Gunn, D. J. Buysse, and M. H. Hall, "The role of sleep hygiene in promoting public health: A review of empirical evidence," *Sleep Medicine Reviews*, vol. 22, pp. 23–36, 2015.

[14] Z. Liang, B. Ploderer, M. A. C. Martell, and T. Nishimura, "A cloud-based intelligent computing system for contextual exploration on personal sleep-tracking data using association rule mining," in *Intelligent Computing Systems: First International Symposium, ISICS 2016, Mérida, México, March 16-18, 2016, Proceedings 1*. Springer, 2016, pp. 83–96.

[15] Z. Liang, M. A. C. Martell, and T. Nishimura, "Mining hidden correlations between sleep and lifestyle factors from quantified-self data," in *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing: adjunct*, 2016, pp. 547–552.

[16] A. M. Berger *et al.*, "Sleep/wake disturbances in people with cancer and their caregivers: State of the science." in *Oncology Nursing Forum*, vol. 32, no. 6, 2005.

[17] N. Pasquier, "Mining assocation rules using frequent closed itemsets," in *Encyclopedia of Data Warehousing and Mining*. IGI Global, 2005, pp. 752–757.

[18] Y. Bastide, N. Pasquier, R. Taouil, G. Stumme, and L. Lakhal, "Mining minimal non-redundant association rules using frequent closed itemsets," in *International Conference on Computational Logic*. Springer, 2000, pp. 972–986.

[19] C. Ordonez, N. Ezquerra, and C. A. Santana, "Constraining and summarizing association rules in medical data," *Knowledge and information systems*, vol. 9, pp. 1–2, 2006.

[20] A. P. Association, *Diagnostic and statistical manual of mental disorders: DSM-IV*. American Psychiatric Association Washington, DC, 1994, vol. 4.