FACULTAD DE CIENCIAS EXACTAS
UNIVERSIDAD NACIONAL DEL CENTRO DE LA PROVINCIA DE BUENOS AIRES

Plan de Trabajo Final

Carrera Ingeniería de Sistemas

Facultad de Ciencias Exactas - UNICEN

Tema: Predicción de comportamiento sedentario a partir del análisis de datos de lifelogging

Alumno: Sr. Martín Santillán Cooper

Director: Prof. Dr. Marcelo G. Armentano
Co-director: Dra. Antonela Tommasel

1. Introducción

Los dispositivos móviles modernos están equipados con varias características (WiFi, GPS y Bluetooth, entre otros) capaces de registrar pasivamente las actividades del usuario y la información contextual, como la ubicación, el uso de las aplicaciones e incluso llamadas y mensajes enviados y recibidos. Esto abre nuevas oportunidades de investigación para la minería y el análisis del comportamiento del usuario. Particularmente, los sensores y dispositivos móviles permiten capturar de múltiples formas complementarias la totalidad de las experiencias de los individuos, las cuales son almacenadas en la forma de un registro multimedial. En particular, un área que podría obtener grandes beneficios de dicho análisis es la investigación en la salud pública. Por ejemplo, la medición y monitoreo de las actividades físicas podrían proporcionar un feedback importante a las personas para lograr un estilo de vida saludable, o podrían ayudar a elaborar programas gubernamentales para prevenir sedentarismo y enfermedades derivadas. En este contexto, el objetivo de este plan de trabajo es la definición de técnicas para la extracción de información a partir de los datos de comportamiento de los usuarios con el objetivo de revelar patrones de comportamiento que permitan identificar situaciones de sedentarismo.

2. Motivación

Es sabido que la actividad física tiene un importante impacto sobre la salud. Recientemente se ha demostrado que el comportamiento sedentario, como permanecer sentado durante un tiempo prolongado e ininterrumpido, es un importante factor de riesgo para la salud, independiente del nivel de actividad física realizado [Owen et al., 2010]. Tanto la actividad física insuficiente, como un exceso de comportamiento sedentario están asociados con un incremento en el riesgo de obesidad, diabetes de tipo 2, enfermedades cardiovasculares y depresión.

Medir el sedentarismo no es una tarea sencilla. Los auto-reportes de los usuarios generalmente poseen un importante sesgo y por lo tanto no son confiables. Hoy en día, los dispositivos móviles (por ejemplo, *smartphones* y *smartwatches*) pueden utilizarse para la recopilación de información acerca del comportamiento humano. La recolección de datos en tiempo real y el análisis de datos de dispositivos móviles pueden revelar información acerca del estado de salud de un usuario, pudiendo ser usados para diagnosticar si un paciente se convierte en sintomático y solicitar un tratamiento temprano, o para generar recomendaciones de actividades saludables cuando se detecten patrones de comportamiento que puedan tener un impacto negativo sobre algún aspecto de la salud del usuario. Algunos síntomas que pueden detectarse a partir de los sensores de los dispositivos móviles son la ansiedad, el estrés, la propagación de una enfermedad y la obesidad [Madan et al., 2012].

El concepto de *lifelogging* puede definirse como una forma de computación omnipresente, que consiste en un registro digital unificado de la totalidad de las experiencias de un individuo, capturado de forma multimodal a través de sensores digitales, y almacenado permanentemente como un archivo multimedia personal [Dodge and Kitchin, 2007]. En otras palabras, el *lifelogging* permite capturar, a partir de una variedad de fuentes de datos, información sobre nuestro entorno y nosotros mismos [Dobbins et al., 2012], abriendo la posibilidad diversas aplicaciones de descubrimiento de conocimiento a partir de estos datos.

3. Hipótesis y Objetivos

La hipótesis del plan de trabajo a desarrollar es que los datos que pueden obtenerse mediante *lifelogging* pueden ayudar a revelar patrones de comportamiento que permitan la predicción de sedentarismo. Esta información de los usuarios, extrapolada al conjunto de habitantes de una comunidad, puede permitir la elaboración de nuevas políticas tendientes a mejorar la salud pública. En particular, el proyecto se centra en dos objetivos específicos: 1) el análisis de datos de *lifelogging* para la definición de perfiles dinámicos de usuario capaces de capturar y representar el comportamiento del usuario, basados en fuentes de datos heterogéneas y 2) la utilización de los perfiles definidos para identificar patrones de comportamiento sedentario.

4. Solución Propuesta

Para lograr los objetivos planteados, el proyecto de tesis se estructurará en dos fases.

Fase 1: Modelado de usuarios en el contexto de datasets heterogéneos.

Con el objetivo de identificar los desafíos que surgen al modelar las actividades realizadas por los usuarios utilizando datasets con datos enriquecidos provenientes de diversas fuentes (por ejemplo, mapas, fotografías o registro de actividades físicas, entre otros), se analizarán las características particulares de los datasets de lifelogging existentes.

En particular, se partirá del análisis de los siguientes datasets:

- NTCIR Lifelog [Gurrin et al., 2016] incluye datos multimedia, señales biométricas y actividades que fueron creadas automáticamente durante un período de 40 días por dos usuarios activos.
- UbiqLog [Rawassizadeh et al., 2013] [Rawassizadeh et al., 2015] contiene un registro de información de *lifelogging* de 35 usuarios, durante 2 meses. Incluye información acerca de las llamadas telefónicas realizadas, encabezados de SMS (no el contenido), uso de aplicaciones, redes WiFi y dispositivos Bluetooth cercanos al usuario, ubicación geográfica y actividades físicas tomadas de la API de Google Play.
- DeviceAnalyzer [Wagner et al., 2014] es quizás el dataset más grande disponible con estados de hardware y configuraciones de dispositivos Android. Contiene más de 10 billones de registros de datos crudos de sensores de smartphones de 23.000 usuarios.
- StudentLife [Wang et al., 2014] es un dataset con registros de los sensores de los teléfonos de 48 estudiantes de la universidad de Dartmouth. El dataset incluye datos sobre la actividad, las conversaciones, las horas de sueño y la localización de cada estudiante a lo largo de 10 semanas.

Considerando la naturaleza heterogénea de los dispositivos inteligentes, es necesario idear nuevas alternativas para modelar con precisión a los usuarios y su comportamiento. Por otro lado, el alto volumen de datos personales generados dificulta la gestión eficiente de los mismos.

Como resultado de esta fase se obtendrá un modelo de usuario que permitirá la representación integral de los usuarios en un único perfil.

Fase 2: Identificación de patrones de comportamiento sedentario

Existen diferentes patrones de comportamiento que podrían ser considerados como indicadores de sedentarismo, por ejemplo, el tiempo utilizado para mirar TV, conducir en auto, trabajar sentado y utilizar la computadora. En esta etapa, se hará un relevamiento acerca del estado del arte en cuanto a la predicción de patrones de comportamiento sedentario y se estudiarán técnicas para la identificación de dichos patrones en los perfiles construidos a partir de los datasets analizados. Para ello, se evaluará la performance de

diferentes clasificadores, capaces de identificar categorías de actividades (por ejemplo, caminar, correr, subir escaleras, permanecer quieto, conducir un automóvil) a partir de la información disponible en el perfil de usuario (trayectorias, mediciones del acelerómetro, mediciones del sensor de ritmo cardíaco, etc.).

Para evaluar la eficiencia de la técnica definida, se utilizarán las clásicas métricas de precisión y área bajo la curva ROC (AUC). La precisión es una métrica empleada para medir el rendimiento de los sistemas de búsqueda y recuperación de información y reconocimiento de patrones, que puede ser también entendida como una medida de la relevancia. En este contexto se denomina precisión (o *valor positivo predicho*) a la fracción de instancias recuperadas (o predichas) que son relevantes o correctas. Una curva ROC, por otra parte, es una representación gráfica de la *sensibilidad* frente a la *especificidad* para un sistema clasificador binario según se varía el umbral de discriminación [Fawcett, 2004]. En otras palabras, es la representación de la razón o ratio de verdaderos positivos frente a la razón o ratio de falsos positivos también según se varía el umbral de discriminación (valor a partir del cual se predice un caso es un positivo). El análisis de la curva ROC proporciona herramientas para seleccionar los modelos posiblemente óptimos y descartar modelos subóptimos independientemente de la distribución de las clases sobre las que se predice. La elección se realiza mediante la comparación del área bajo la curva (AUC) de ambas pruebas. Esta área posee un valor comprendido entre 0,5 y 1, donde 1 representa un valor diagnóstico perfecto y 0,5 es una prueba sin capacidad discriminatoria diagnóstica (equivalente al azar).

El objetivo final de esta fase consiste en obtener una visualización de los patrones de actividad de los usuarios a lo largo del día que facilite la predicción de comportamiento sedentario.

5. Cronograma de actividades

Actividad	Meses											
	1	2	3	4	5	6	7	8	9	10	11	12
Relevamiento de técnicas de modelado de usuarios en el contexto de datasets heterogéneos												
Análisis de características de datasets existentes de *lifelogging* que puedan ayudar a la identificación de sedentarismo												
Definición de un modelo de perfil de usuario a partir de la integración de los diferentes datos existentes en los datasets analizados												
Análisis de técnicas para la identificación de patrones de comportamiento sedentario a partir de los perfiles construidos												
Evaluación de la técnica de desarrollada												
Redacción del informe final												

6. Bibliografía

[Do and Gatica-Perez, 2014] Where and what: Using smartphones to predict next locations and applications in daily life (T.M.T. Do and D. Gatica-Perez). Pervasive and Mobile Computing, 12:79 91, 2014.

[Dobbins et al., 2012] Augmenting Human Digital Memories with Physiological Data (C. Dobbins, M. Merabti, P. Fergus, and D. Llewellyn-Jones). In Proceedings of the 3rd IEEE International Conference on Network Embedded Systems for Every Application (NESEA), Liverpool, UK, 13–14 December 2012; pp. 35–41. 2012.

[Dodge and Kitchin, 2007] Outlines of a world coming into existence: Pervasive computing and the ethics of forgetting (M. Dodge and R. Kitchin). Environ. Plan., 34, 431–445. 2007.

[Fawcett, 2004] ROC Graphs: Notes and Practical Considerations for Researchers. (T. Fawcett). Technical report. Palo Alto (USA): HP Laboratories; (2004)

[Gurrin et al., 2016] NTCIR Lifelog: The First Test Collection for Lifelog Research (C. Gurrin, H. Joho, F. Hopfgartner, L. Zhou, and R. Albatal). In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR '16). ACM, New York, NY, USA, 705-708. 2016.

[Madan et al., 2012] Sensing the “Health State” of a Community (A. Madan, M. Cebrian, S. Moturu, K. Farrahi, and A. Pentland). IEEE Pervasive Computing 11(4), 36–45. 2012.

[Owen et al., 2010] Too much sitting: the population-health science of sedentary behavior (N. Owen, G. N. Healy, C. E. Matthews, and D. W. Dunstan). Exercise and sport sciences reviews, vol. 38, no. 3, pp. 105, 2010.

[Rawassizadeh et al., 2013] UbiqLog: a generic mobile phone-based life-log framework (R. Rawassizadeh, M. Tomitsch, K. Wac and A.M. Tjoa). Personal and ubiquitous computing, 17(4), 621-637. 2013

[Rawassizadeh et al., 2015] Lesson Learned from Collecting Quantified Self Information via Mobile and Wearable Devices (R. Rawassizadeh, E. Momeni, C. Dobbins, P. Mirza-Babaei, and R. Rahanamoun). Journal of Sensor and Actuator Networks, 4(4), 315-335. 2015

[Wagner et al., 2014] Device Analyzer: Large-scale Mobile Data Collection (D. Wagner, A. Rice, and A. Beresford). SIGMETRICS Perform. Eval. Rev. ’14, 41(4):53–56. 2014.

[Wang et al., 2014] StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students using Smartphones (Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell) .In Proceedings of the ACM Conference on Ubiquitous Computing. 2014.