Drawing the traffic map in school networks

Juan Francisco Rodriguez Saredo[1] and Regina Motz[2]

[1] Programa de Desarrollo de las Ciencias Básicas, Informática, Uruguay
[2] Instituto de Computación, Facultad de Ingeniería, UdelaR, Uruguay
{jfrodriguez,rmotz}@fing.edu.uy

Abstract. This work shows an application designed to identify groups in the use of an online educational network associated with the socioeconomic characteristics of the neighborhoods of Montevideo where their users live. The network has a wide coverage throughout the national territory and offers Internet access to all students (children and adolescents). The knowledge obtained from the study can be applied as support for decision making.

Keywords: Clustering, Hopkins, Clara.

1 Introducción

La gestión y análisis de los datos obtenidos en los ambientes educacionales habitualmente presentan dificultades para su conducción. La aplicación de técnicas de *clustering* para obtener conocimiento útil de ellos, posee dificultades adicionales tales como la definición del espacio por sus atributos, la medida de distancia a utilizar y las muestras con las cuales generar el modelo, entre otras.

La originalidad y dificultad de este trabajo radica en que se debe trabajar con dos *datasets* de muy diferente origen: por un lado se dispone de datos de acceso a los recursos provenientes de una red extendida en todo el territorio nacional, con una amplia cobertura que posibilita el acceso a Internet desde todos los centros educacionales (Primaria, Secundaria y UTU) [6] y, por otro lado, datos que se recabaron de la *Web* en carácter de datos abiertos disponibles y provenientes de organismos oficiales [3].

Resulta interesante determinar si existe algún tipo de relación entre el uso de los recursos educacionales en línea y el lugar donde habita el estudiante. Los barrios de la ciudad poseen características únicas que genera una idiosincrasia que, se conjetura, influye en sus hábitos de estudio. Un factor importante para analizar es el constituido por los aspectos relacionados con el nivel socio económico de cada uno de los barrios.

Existen abundantes estudios relativos a usos de técnicas de *clustering* en ambientes de educación. Algunos de ellos consolidan los diferentes tipos de algoritmos de agrupación aplicados en el contexto de la minería de datos educativos, para abordar diferentes problemas que se presentan en *EDM (educational*

[3] http://municipioe.montevideo.gub.uy/sites/municipioe/files/censo_2011_-_informe_im.pdf

data mining) [1] y otros aplican las técnicas a datos que no pertenecen a una plataforma de estudio en particular sino que también abarcan registros de acceso a cualquier otro tipo de sitio *Web* [6].

A los efectos de aplicar las técnicas descriptivas se clasifican las localidades de la ciudad en cuatro categorías, fundamentados en los datos abiertos mencionados anteriormente (éstas se describen en la *Sección 4, Análisis del Problema*).

2 Alcance del Trabajo

Los datos son originados a través de dos fuentes. Una de ellas la constituye la red educacional donde cada local de estudio tiene asignado un conjunto de direcciones *IP*, las cuales son fijas, conocidas y están asociadas a su ubicación geográfica y al tipo de local de estudio. Los registros de navegación disponibles son almacenados en formato de texto plano y contienen ciertos atributos de las visitas a los sitios *Web*. De cada observación generada, se utilizan los atributos: fecha, hora, *IP* y *URL* solicitada. La segunda fuente proporciona datos demográficos provenientes de organismos oficiales y aportan importante información sobre los municipios que forman la capital del país y las características socio-económicas de los barrios que lo integran. Estos datos pueden ser accedidos en la *Web*.

El trabajo intenta obtener relaciones entre las conductas de empleo de los recursos educacionales utilizables en la red y la situación socio-económica del estudiante basado en su lugar de residencia.

3 Objetivo

El objetivo principal es aplicar técnicas de analítica sobre los datos obtenidos de los archivos que se generan diariamente por la navegación de los usuarios de la red, asociando las *ip* de los centros de estudio con las características de la localidad donde habita. Los resultados obtenidos serán destinados al soporte para la toma de decisiones de las autoridades del Plan, facilitándose su interpretación por medio de una adecuada visualización de los resultados a través de mapas "inteligentes".

4 Análisis del Problema

El análisis del problema permite identificar los siguientes desafíos:

- De los datos generados en la red, extraer aquellos relacionados con accesos a sitios de estudio.
- A partir de los datos disponibles en la *Web*, construir un *ranking* de los barrios que represente en forma fidedigna la situación socio-económica de cada uno (En este ítem se considerarán los datos demográficos oficiales disponibles en la *Web*).
- Asociar ambos grupos de datos.

– Determinar los grupos de datos a los cuales aplicar las técnicas de *clustering*.
– Técnicas de Agrupamiento a ser empleada.

Los cinco puntos antes nombrados son desarrollados a continuación.

Datos generados en la red La extracción de los datos de tráfico en Internet a sitios de estudio se efectuó a través de rutinas desarrolladas en *Python*, aplicando expresiones regulares al atributo "*URL* solicitada". Luego de una adecuada sanitación de los datos se extraen los registros que contienen dichas expresiones. Un ejemplo de las expresiones regulares empleadas para este caso es:

```
'\setminus S+clients3\setminus S+',
'\setminus S+windowsupdate\setminus S+',
'\setminus S+mcafee\setminus S+',
'\setminus S+216.239.32.20/generate_204\S+',
'\setminus S+clients1\setminus S+',
'\setminus S+connectivitycheck\setminus S+',
'\setminus S+URLMOD\setminus S+',
'\setminus S+msftncsi\setminus S+'
```

Clasificación de los barrios Los datos disponibles en la *Web* sobre la situación socio-económica de los barrios componentes de cada municipio se fundamentan en los registros obtenidos en el último Censo Nacional. Esta información está acompañada de valoraciones generales sobre los municipios que dividen a Montevideo. Se utilizó un procedimiento *ad hoc* para la construcción del *ranking*: se observaron las valoraciones de cada municipio y a cada característica positiva detectada en la valoración, se le asignó un coeficiente positivo y a las negativas uno negativo. Este procedimiento permitió clasificar a los municipios en cuatro categorías: Muy Favorable (MF), Favorable (F), Desfavorable (D) y Muy Desfavorable (MD). Luego se revisaron los barrios de cada municipio y puntualmente algunos fueron cambiados de categoría en base al conocimiento de su realidad particular (4 barrios en un total de 75).

Asociación de los registros de tráfico con los barrios Para logar la asociación de los dos grupos de datos fue necesario efectuar un procesamiento apropiado de los atributos comunes entre ellos, que posibilitara un posterior relacionamiento. El atributo común es el nombre del barrio y se debió realizar un relevamiento de cada uno de los registros de los centros de estudio (alrededor de *1200* en Montevideo) que utilizan el sistema.

Grupos de datos a los cuales aplicar las técnicas de clustering A partir del análisis exploratorio de los datos (desarrollado en la Sección 5) se puede definir tres grandes grupos de usuarios: Primaria (Escuela Pública), Secundaria (Liceo Público), UTU (Centros de estudio públicos donde se enseñan mayormente oficios, de nivel enseñanza media).

Técnicas de Agrupamiento a ser empleada Se considera que dos locales de estudio están próximos, si la cantidad de conexiones en un período de tiempo son parecidas con un umbral de tolerancia. El espacio así definido no es euclídeo, lo cual condiciona la elección del algoritmo de *clustering* a ser utilizado. Un espacio es euclidiano si el promedio de cualquier conjunto de sus puntos pertenece al espacio [5].

5 Análisis exploratorio de los datos

Para el análisis exploratorio de los datos, se empleó el lenguaje *R* (versión 3.6.1) conectado a la base de datos *MongoDB* (*versión 3.2.10*). A los efectos de evaluar si existe una tendencia al agrupamiento de los datos se utiliza el estadístico de *Hopkins* [3]. La función *Hopkins* del paquete *clustertend* disponible en *R* permitió obtener en todos los estudios valores cercanos a *0,004*, lo que es indicativo de presencia de agrupamientos.

Evaluación de la existencia de agrupamientos Los estudios que se realizaron en esta etapa consistieron en la apreciación de la existencia de agrupamientos y en la determinación del tamaño de la muestra para cada estudio.

El Lenguaje *R* permite al menos dos formas de evaluar la tendencia a la clusterización de los datos. Una de ellas es el estadístico de *Hopkins* que indica qué tan alejado está una muestra aleatoria de presentar una distribución uniforme y, en consecuencia, de presentar agrupaciones (cuanto más alejado de tal distribución mayor es la probabilidad de que los datos se agrupen). La otra opción es el uso de otra funcionalidad consistente en la evaluación visual de la tendencia al agrupamiento de los datos (*VAT: Visual Assessment of cluster Tendency*). El procedimiento se fundamenta en el cálculo de la matriz de disimilaridad entre los objetos de la muestra considerando que la distancia es euclídea [4].

Debido a que la medida de distancia no es euclídea, el método *VAT* es descartado, empleándose el estadístico de *Hopkins* para evaluar la existencia de clústeres.

Tamaño de la muestra A los efectos de determinar el tamaño de la muestra más apropiado, se empleó el concepto de saturación (saturación es denominada la situación en la cual, agregar nuevas observaciones, no mejora las perspectivas de obtener nueva información) [2]. En esta etapa, también se identificaron posibles *outliers*.

Procedimiento Luego de asociados los registros se exportaron a una base de datos relacional y se condujeron consultas para observar las tendencias.

A los efectos de normalizar los datos de entrada se dividió la cantidad de conexiones entre la población de cada barrio, evitando que localidades con gran cantidad de habitantes distorsionen los resultados.

Algunos resultados de las consultas efectuadas se presentan en la Figura 1.

De este resultado se destaca que los únicos barrios cuyos centros educativos utilizaban en ese día los recursos eran los considerados Muy Desfavorables (*MD*).

Cantidad de conexiones	Barrio	Tipo
11802	Casavalle	MD
4242	Cerro	MD
70	Ituzaingó	MD
56	La Teja	MD
126	Nuevo París	MD
504	Peñarol	MD

Fig. 1. Cuadro con cantidad de accesos por barrio

En base a esta información, el estudio se orientó a investigar la existencia de clústeres de centros de estudio en determinados barrios que utilizaban con mayor intensidad la red educacional en función de su realidad socio-económica.

6 Técnicas aplicadas y resultados obtenidos

Debido a que el espacio no es euclídeo se empleó el paquete de *datamining* disponible en el lenguaje *R: CLARA (Clústering Large Applications)*. Este algoritmo efectúa la búsqueda de clústeres en base a medoides (puntos ya existentes del clúster), prescindiendo de la búsqueda de centroides.

A continuación se presentan algunos de los resultados que evidencian el mayor uso de los recursos por parte de los sectores socio-económicos más desamparados.

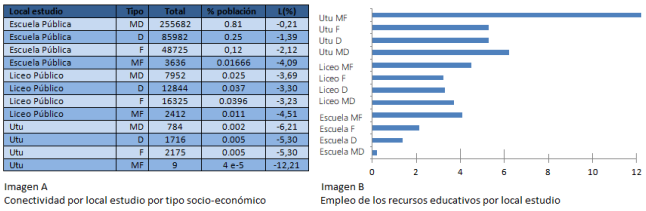
Sumarizaciones Iniciales En la tabla de la Figura 2 se presentan la cantidad de habitantes de acuerdo al tipo socio-económico categorizado en este trabajo y la cantidad de barrios de acuerdo a los tipos. Ambos resultados fueron importantes en el transcurso del estudio tanto para efectuar los cálculos como para la interpretación de los resultados.

Tipo	Población	# barrios
MUY DESFAVORABLE	317476	27
DESFAVORABLE	342952	14
FAVORABLE	412019	27
MUY FAVORABLE	218173	8

Fig. 2. Datos de la población.

En la Imagen 1 de la Figura 3 se presentan los datos de una muestra correspondiente a la conectividad de acuerdo al tipo socio-económico y al tipo de local de enseñanza correspondiente a *14* días seleccionados en forma aleatoria. La última columna de la mencionada tabla consiste en los logaritmos naturales del porcentaje de la población de cada barrio (ya que son número muy pequeños) y se presentan en la gráfica de barras de la Imagen 2 de la misma Figura. En ella, el opuesto del logaritmo indica el uso que cada estrato de la población (una

longitud de la barra pequeña es indicativo que hay un uso intenso de los recursos). Por ejemplo, los alumnos de *UTU* del tipo *MF* serían los que menos utilizan los recursos (para la muestra seleccionada) y los alumnos de *Escuela Pública* del tipo *MD* (contexto económico-social muy desfavorable) quienes más los emplean.



Local estudio	Tipo	Total	% población	L(%)
Escuela Pública	MD	255682	0.81	-0,21
Escuela Pública	D	85982	0.25	-1,39
Escuela Pública	F	48725	0,12	-2,12
Escuela Pública	MF	3636	0.01666	-4,09
Liceo Público	MD	7952	0.025	-3,69
Liceo Público	D	12844	0.037	-3,30
Liceo Público	F	16325	0.0396	-3,23
Liceo Público	MF	2412	0.011	-4,51
Utu	MD	784	0.002	-6,21
Utu	D	1716	0.005	-5,30
Utu	F	2175	0.005	-5,30
Utu	MF	9	4 e-5	-12,21

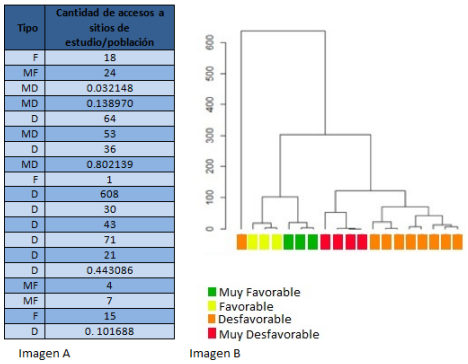
Imagen A
Conectividad por local estudio por tipo socio-económico

Imagen B
Empleo de los recursos educativos por local estudio

Fig. 3. Empleo de los recursos educativos.

Resultados obtenidos para los estudiantes de enseñanza primaria (Escuela Pública) Para este sector se presentan 3 estudios realizados en marzo 2016, noviembre 2016 y mayo 2017.

Estudio 1 En la primer imagen de la Figura 4 se observan los datos empleados para la construcción del dendograma presentado en la segunda imagen.



Tipo	Cantidad de accesos a sitios de estudio/población
F	18
MF	24
MD	0.032148
MD	0.138970
D	64
MD	53
D	36
MD	0.802139
F	1
D	608
D	30
D	43
D	71
D	21
D	0.443086
MF	4
MF	7
F	15
D	0. 101688

Imagen A

Imagen B

Fig. 4. Datos y dendograma para escuela pública

Del análisis del dendograma se observan cuatro clústeres bien definidos (sin ningún dato mal clasificado de acuerdo al algoritmo). Se aprecian un agrupamiento de barrios *Desfavorables, D* (color anaranjado) y de *Muy Desfavorables, MD* (color rojo). Luego se aprecian dos clústeres de *Faborables, F* (color amarillo) y *Muy Favorables, MF* (color verde). Se concluye que para estos datos

existe una mayor cantidad de barrios que utilizan los recursos pertenecientes a contextos Desfavorables. El siguiente agrupamiento con mayor cantidad de barrios corresponde a contextos Muy Desfavorables.

Las categorías *F* y *MF*, si bien se agrupan en dos clústeres diferentes, se puede comprobar que el uso que hacen de los recursos es mínimo. Por último la hoja que no pertenece a ningún grupo presenta 608 conexiones, quedando fuera de los grupos y se trata de un barrio del tipo *D*.

Estudio 2 El segundo estudio corresponde a una muestra de noviembre de 2016 de Escuela Pública. El dendograma correspondiente se presenta en la primer imagen de la Figura 5. Se utiliza el mismo sistema de referencia de colores para las categorías de los barrios que en el Estudio 1. Se observan cuatro clústeres. El primero y el tercero (empezando por izquierda) corresponden a localidades *D*, el segundo presenta una prevalencia de barrios también *D* y el cuarto, no permite decidir ya que estan mezcladas varias categorías.

Estudio 3 El estudio correspondiente a mayo de 2017 de las escuelas públicas, se aprecia en la segunda imagen de la Figura 5. De los cinco agrupamientos identificados, los tres primeros presentan individuos de todas las poblaciones no permitiendo desarrollar una conclusión y el cuarto y quinto presentan una mayor presencia de *D* y *MD* si se consideran conjuntamente.

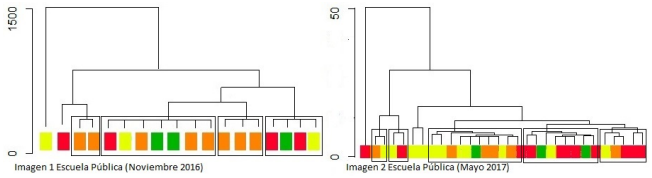


Fig. 5. Dendogramas para la actividad en escuelas públicas.

Resultados obtenidos para los estudiantes de enseñanza media (Liceo Público) Del estudio para los liceos públicos, en marzo de 2016, se obtiene el dendograma de la Figura 6, donde se aprecian cuatro agrupamientos. Se observa una tendencia diferente a la presentada en las muestras correspondientes a la Escuela Pública. El primer clúster es mayoritariamente *F* y *MF*, el segundo es *MF* y el tercero es mayormente D. El cuarto no presenta características relevantes.

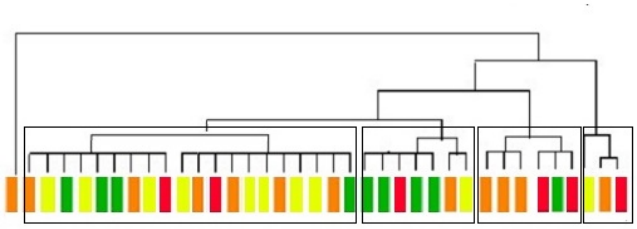


Fig. 6. Dendograma para la actividad en liceo público

Resultados obtenidos para los estudiantes de enseñanza media (*UTU*) Al igual que en los otros centros de estudio se observan los agrupamientos de acuerdo a los barrios (Figura 7). Empleando el mismo sistema de referencias de colores, se observan cuatro agrupamientos, tres de los cuales son mayoritariamente F y uno de ellos MD y D.

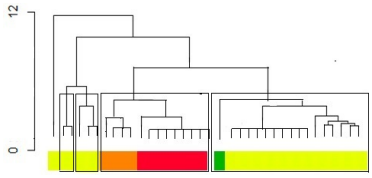


Fig. 7. Dendograma para la actividad en UTU.

Bosquejo de un mapa de tráfico En base a los datos disponibles se presenta una distribución del uso de la red por barrio en Montevideo, la cual presenta una idea primaria de la visualización del empleo de los recursos en relación a la distribución geográfica. El suministro de datos en tiempo real permitirá la activación de indicadores que podrían cruzarse con datos provenientes de diversas fuentes, como se estableció en la *Sección 3, Objetivo* (Figura 8).

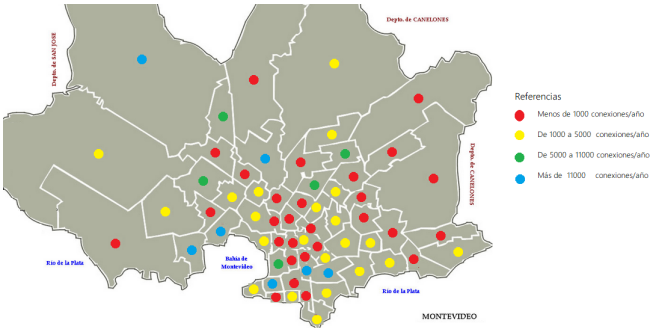


Fig. 8. Mapa de tráfico.

7 Conclusiones y trabajos futuros

Los clústeres obtenidos en cada uno de los estudios indica la presencia de agrupamientos relacionados con la situación socio-económica de cada barrio.

A nivel de primaria (escuelas públicas) se observan agrupamientos más acentuados y definidos que en los otros tipos de locales de estudio en los datos correspondientes a los primeros meses. En los sucesivos meses, se podría considerar una tendencia a la uniformización en el uso.

De la observación de los datos se puede afirmar que en el año 2016 los barrios más desamparados eran los que más utilizaban los recursos (casi en exclusividad). En abril 2017 aunque el uso era mayor por parte de todos, los barrios con más pobreza son los que más lo emplean.

En cambio, para Liceos Públicos y *UTU* se observa una tendencia a un mayor uso en las categorías favorables y muy favorables.

La presencia de estos agrupamientos permite formular algunas conjeturas que relacionan el nivel social y económico con el uso de la red educacional. En caso de que efectivamente sea ésta la realidad, se podrá llevar adelante políticas integradas (sociales y educativas) para aprovechar las características relevadas en el estudio.

En lo que concierne a la visualización de los resultados, si las autoridades educativas entienden que es de interés, se suministrará a los mapas inteligentes datos obtenidos en tiempo real. Esto permitirá, enmarcado en un adecuado y más ambicioso proyecto de ciencia de datos, un análisis temporal complementado con otras fuentes de información con el objetivo de prevención de inasistencias y mejoras educativas (relativas al uso de los recursos). Por ejemplo si se dispone de información proveniente del Ministerio del Interior de un incremento de actividades delictivas en determinados lugares, puede detectarse la incidencia en la asistencia a clases (efectuando el cruzamiento de datos provenientes del ministerio nombrado y los recibidos por el uso de la red educativa).

Una aplicación interesante podría consistir en el cruce de datos provenientes del Ministerio de Salud Pública relativos a enfermedades estacionales (gripes y

resfríos, por ejemplo) e inasistencias a clase detectadas por el poco empleo de la red.

En relación a las expresiones regulares utilizadas para identificar los sitios de estudio, es posible ampliar el *corpus* a los efectos de mejorar la calidad de los agrupamientos.

References

1. Dutt, A., Aghabozrgi, S., Ismail, M., Mahroeian, H.: *Clustering algorithms applied in educational data mining.* International Journal of Information and Electronics Engineering **5**(2), 112 (2015)
2. Glaser, B., Strauss, A.: *Discovery of grounded theory: Strategies for qualitative research.* Routledge (2017)
3. Han, J., Pei, J., Kamber, M.: *Data mining: concepts and techniques.* Elsevier (2011)
4. Hu, Y., Hathaway, R.J.: An algorithm for clustering tendency assessment. WSEAS Trans. Math. **7**(7), 441–450 (Jul 2008), http://dl.acm.org/citation.cfm?id=1466906.1466908
5. Leskovec, J., Rajaraman, A., Ullman, J.: *Mining of Massive Dataset.* Cambridge University Press (2014)
6. Saredo, J., Motz, R.: *Application of clustering techniques on data generated by an Online Educational Network.* In: Anais dos Workshops do Congresso Brasileiro de Informática na Educação. vol. 6, p. 714 (2017)