Aplicación de técnicas de agrupamientos sobre datos generados por una red educativa en línea

Juan Francisco Rodríguez Saredo[1], Regina Motz[2]

[1] Pedeciba-Informática, Montevideo, Uruguay

[2]Universidad de la República, Montevideo, Uruguay

jfrodriguez@fing.edu.uy, rmotz@fing.edu.uy

Abstract. *The work presented in this paper seeks to identify trends and groupings in the use of an online educational network. This network has an extensive coverage throughout the national territory and provides Internet accessibility to all the students (children and adolescents). The challenge is that the data does not only belong to a particular study platform, but also includes access records to any other type of website through the network, making the selection of appropriate data quite difficult. Clustering techniques are proposed to perform the analysis. The clustering analysis is based on a space defined by the number of connections in a time interval. The knowledge obtained can be used to support educational decision making.*

Resumo. *El trabajo presentado en este artículo busca identificar tendencias y agrupaciones en el uso de una red educativa en línea, la cual tiene una amplia cobertura en el territorio nacional y ofrece acceso a Internet a todos los estudiantes (niños y adolescentes). El desafío radica en que los datos no pertenecen a una plataforma de estudio en particular sino que también abarcan registros de acceso a cualquier otro tipo de sitio web, dificultándose la selección adecuada de datos. Se proponen técnicas de agrupamiento para realizar el análisis, el cual se basa en un espacio definido por el número de conexiones en un intervalo de tiempo. El conocimiento obtenido puede ser útil como apoyo para la toma de decisiones.*

1. Antecedentes

La masiva generación de datos puede ser percibida en nuestras vidas cotidianas. Su recolección, procesamiento, gestión y análisis son utilizados con el objetivo de generar conocimiento o información destinada a contribuir en el proceso de la toma de decisiones. Las tareas nombradas no pueden llevarse a cabo por los métodos habituales, los cuales están sustentados sobre bases de datos relacionales. El área de la educación, no es ajena a este fenómeno. Un ejemplo lo constituyen los datos originados a través de cursos en línea de alcance masivo. En nuestro país, existe una red con una amplia cobertura que posibilita el acceso a Internet desde todos los centros educacionales (Primaria, Secundaria y Universidad del Trabajo). Se trata de un plan educacional que diariamente genera un tráfico de varios terabytes, además de un registro de navegación, el cual es almacenado. El presente estudio aplica técnicas de clusterización a los efectos de obtener información de utilidad en cuanto al uso, previsiones y tendencias de dicha red. 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 data mining*) [Ashish, Saeed, Maizatul and Mahroeian

DOI: 10.5753/cbie.wcbie.2017.714

2015] y otros que aplican técnicas clásicas o adaptan algoritmos a situaciones concretas [Xu, Recker, Qi, Flann and Ye 2013]. La mayoría de dichas publicaciones son efectuadas considerando datasets generados sobre aplicaciones específicas de estudio. El desafío que presenta este trabajo, es que los datos no pertenecen a una plataforma de estudio en particular sino que también abarcan registros de acceso a cualquier otro tipo de sitio web.

2. Alcance del Trabajo

Cada local de estudio tiene asignado un conjunto de direcciones *IP,* éstas son fijas, conocidas y están asociadas a su ubicación geográfica y al tipo de local de estudio. No se dispone de ningún medio de reconocimiento de los usuarios asegurándose la protección de su identidad. 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. Los restantes campos los crean los propios sistemas: firewalls, Internet, entre otros (los cuales no son de interés para el estudio). A pesar de que la cantidad de dimensiones del problema es reducida y que no es posible obtener ninguna característica personal de los usuarios, se espera poder relacionarlos o buscar patrones en la utilización de los recursos.

En general se pretende encontrar patrones o características que permitan describir las siguientes situaciones:

- Mediante el análisis de la red, probar la existencia de horarios en el que se observan aproximadamente la misma cantidad de conexiones en centros de estudio en determinados días.
- Variación de la cantidad de conexiones en el correr del año para los distintos centros de estudio (observados en su totalidad o relacionados a visitas a sitios de estudio).
- Existencia de aglomeraciones de conexiones en horario de clases o de usuarios que se conectan a un sitio web de determinado tipo en relación a sus lugares de residencia (esta característica habitualmente está asociada a la situación económica y social).

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, proporcionando

indicadores destinados al soporte para la toma de decisiones. A continuación se enumeran los objetivos generales.

1. Relación de visitas a sitios con material educativo comparados con: visitas a sitios de diverso contenido; entre turnos a lo largo del día y en distintas épocas del año. Por ejemplo, detectar la utilización de sitios web de esparcimiento comparado con sitios de estudio, por hora, día o periodos de tiempo.
2. Detectar horarios de uso de la red que son de utilización máxima o intermedia. Asimismo determinar los horarios que son de escasa o nula utilización a nivel nacional.

Para el presente estudio y siguiendo los lineamientos de los objetivos generales, se definen los siguientes objetivos específicos:

1. Búsqueda de patrones en el uso de la red según los horarios.
2. Exploración de algunos clústeres que respondan a los puntos mencionados en los objetivos generales.

3. Detectar *outliers* (centros de estudio que difieren en la utilización habitual de la red, en relación a los horarios).

Los resultados alcanzados serán de utilidad para las autoridades del Plan Educacional ya que posibilitará un mejor aprovechamiento de los recursos y, en base a la detección de patrones de conducta sobre una base del anonimato de sus usuarios, poder escalarlos en forma genérica para la mejora del sistema educativo *online*.

Los objetivos planteados, serán abordados por medio del análisis de agrupamientos identificando tendencias en el uso de la red según los horarios, ubicaciones y tipo de sitios visitados. Se dispone de un conjunto de registros de navegación de usuarios del Plan, identificados por su centro de estudio, hora y URL accedida, entre otros atributos. Este estudio presenta el desafío de que los registros son anónimos dificultando la determinación de franjas por edades entre los usuarios, su género o el grado al que asiste.

4. Análisis y metodología empleada

El trabajo se organiza llevando a cabo las siguientes tareas:

Etapa 1: Colección de los datos, sanitación, estudio de los datos faltantes y medidas para su solución.
Etapa 2: Análisis exploratorio de los datos
Etapa 3: Extracción de información por medio de herramientas estadísticas.
Etapa 4: Obtención de conocimiento que apoye a la toma de decisiones.

4.1 Etapa 1: Colección de los datos, sanitación, estudio de los datos faltantes y medidas para su solución.

Los datos son registros que se almacenan diariamente en archivos de texto plano (formato .log). Por día se generan entre 200 MB y 1 GB de registros (aproximadamente un total de 1,5 millones de entradas en promedio).

Por medio de una rutina desarrollada en Python y aplicando expresiones regulares al atributo “URL solicitada”, se eliminan registros que son irrelevantes como, por ejemplo, actualizaciones de sistemas operativos, auditorías, notificaciones de antivirus, entre otros. Un ejemplo de las expresiones regulares empleadas para este caso es:
'\S+clients3\S+', '\S+windowsupdate\S+', '\S+mcafee\S+',
'\S+216.239.32.20/generate_204\S+', '\S+clients1\S+', '\S+connectivitycheck\S+',
'\S+URLMOD\S+', '\S+msftncsi\S+'

La información contenida en las URLs correspondientes a los s*trings* utilizados, se refiere a procesos automatizados y fueron proporcionadas por los técnicos del Plan. Los datos resultantes, se almacenan en una base de datos *MongoDB*. Luego de finalizado el proceso, se ejecutan rutinas (programadas en *Javascript*) que, empleando nuevas expresiones regulares (las cuales definen las categorías para las URLs accedidas), asocian los registros de navegación con los locales a los que pertenecen. Los locales son identificados por un código y. adicionalmente, se dispone de una tabla que indica sus características: nombre y ubicación entre otros atributos. Este proceso genera archivos en formato *json* (uno para cada día y para cada expresión regular detectada en la búsqueda en la base de datos). Debido a que los archivos *json* obtenidos tienen anidaciones, los datos son "aplanados" pasándolos a una base de datos relacional.

Se dispone de datos correspondientes a 214 días desde febrero 2016 a mayo 2017, irregularmente distribuidos, presentando considerables periodos de tiempo sin datos. Los locales de estudio son aproximadamente 1100. Luego del procesamiento, se dispone de 3:300.006 registros, que suman 122:205.324 conexiones. Una primera sumarización de los datos permite

clasificarlos en categorías, obteniéndose cierta cantidad de registros para cada una de ellas, presentadas en la *Tabla 1, Categorías de los sitios y cantidades*.

Tabla 1: Categorías de los sitios y cantidades

Categoría	Cantidad
Sitios de Búsqueda	32636343
Apoyo educación	1445627
Utilitarios	7752560
Redes sociales	69526210
Juegos	5044463
Periódicos	1083627
Estudio	4716237

En la *Figura 1, Sumarizaciones iniciales*, se puede visualizar, en la primer imagen, las proporciones de las categorías exploradas. En la misma figura, la segunda imagen indica las proporciones de las expresiones regulares para la educación y la tercer imagen para las redes sociales. En esta última es interesante observar que *Instagram* tiene un mayor uso que *Facebook* (para la muestra considerada y considerando las características de los usuarios del sistema)

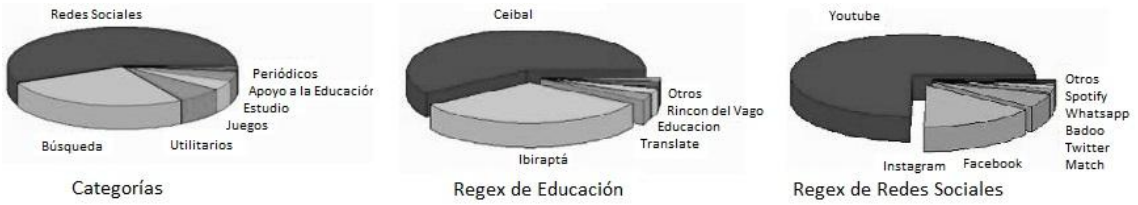

Figura 1: Sumarizaciones iniciales

4.2 Etapa 2 Análisis exploratorio de los datos

Para la etapa de análisis exploratorio de los datos, es empleado 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 [Han, Kamber and Pei 2012] el cual indica qué tan alejados se encuentran datos de la muestra, seleccionados aleatoriamente, de presentar una distribución uniforme (cuanto más alejado de tal distribución mayor es la probabilidad de que los datos se agrupen). Para ello se utiliza la función *hopkins* del paquete *clustertend* disponible en *R* obteniéndose valores cercanos a *0,003,* lo que es indicativo de presencia de agrupamientos (el estadístico varia de 0 a 1 y 0,5 indica distribución uniforme)

Con el propósito de obtener cualitativamente alguna información de los datos, se considera una muestra de ellos. La determinación del tamaño de una muestra en estudios cualitativos generalmente es un juicio subjetivo, tomado a medida que la investigación avanza, recomendándose el concepto de saturación para lograr un tamaño de muestra apropiado (saturación es denominada la situación en la cual, agregar nuevas observaciones, no mejora las perspectivas de obtener nueva información) [Glaser and Strauss 1967]. A tales efectos, se consideran los registros generados en el acceso al Plan, en un determinado día y se contabilizan las conexiones a internet en cada hora y en cada centro de estudio, analizando su distribución empírica. Observando que los datos disponibles cubren alrededor de un año, se consideró conveniente, a los efectos de obtener muestras representativas, repetir los estudios eligiendo días aleatoriamente, generando nuevas muestras, hasta lograr la saturación. En esta etapa, también se pretende identificar candidatos a ser posibles outliers durante el proceso.

Para los estudios, se utilizan vistas de los datos en las que cada centro de estudio dispone de campos para cada hora del día donde se indica la cantidad de conexiones establecidas por dicho local. La *Tabla 2*, *Conexiones por local y por hora en una fecha determinada,* es un ejemplo de cómo se dispondrán los datos: las filas representan los centros de estudio, las columnas los rangos de horas y en las celdas se encuentran la cantidad de conexiones de cada centro de estudio por rango de horas.

Tabla 2: Conexiones por local y por hora en una fecha determinada

Horas / Local	0000 - 0059	0100 - 0159	0200 - 0259	....	2200 - 2259	2300 - 2359
23456	2	1	5		500	475
32345	3	5	3		651	597
13246	1	3	0		321	384

4.3 Etapa 3: Extracción de información utilizando herramientas estadísticas.

Con el objetivo de proporcionar claridad al plantear los casos de estudio, se define un procedimiento que proporciona una notación adaptada al problema denominado *clusteres,* el cual tiene *n+1* argumentos. El cabezal del procedimiento *clusteres(x; x_1,x_2,...,x_n)* indica la aplicación de cualquier método de agrupamiento para determinar los clústeres *x*, a partir de los atributos x_1,x_2,...,x_n de los datos. Dichos atributos no son necesariamente independientes entre sí (puede ocurrir que algún campo se genere a partir de otros). En caso de que los datos sean dinámicos (variables con el tiempo), se agrega al tiempo como un nuevo argumento, utilizándose la notación *clusteres(x;x_1,x_2,...,x_n,t).*

En esta etapa se utilizó el lenguaje *R* como herramienta de analítica con el apoyo de los lenguajes *PHP 7.0* y un motor de base de datos relacional para el procesamiento de datos ya sumarizados

Casos de estudio

***clusteres(horas de uso; cantidad de conexiones)*:** Indica la búsqueda de horarios en que la cantidad de conexiones es similar (corresponde a uno de los objetivos específicos planteados, determinación de patrones en el uso de la red según los horarios).

***clusteres(horas de acceso a la red; locales de estudio)*:** Con datos similares a los de la *Tabla 1,* se busca identificar candidatos a ser agrupamientos indicando los clústeres formados por las observaciones en las cuales se contabiliza, aproximadamente, la misma cantidad de conexiones en un determinado día, en varios locales de estudio. Se busca identificar locales que presentan un comportamiento parecido en lo que se refiere a la conectividad en determinados horarios durante el día, basándose en la cantidad de conexiones por hora. En este sentido dos locales se consideran próximos si la cantidad de conexiones en cada intervalo de tiempo son parecidas con un umbral de tolerancia. Se debe considerar que el espacio definido no es euclídeo, al momento de definir el algoritmo de clustering a ser utilizado. Se considera que un espacio es euclidiano si el promedio de cualquier conjunto de sus puntos pertenece al espacio [Leskovec, Rajaraman and Ullman 2014].

La discretización del tiempo en intervalos definidos en horas es justificada considerando que las horas del día indican las actividades de las personas, Por ejemplo, el turno matutino de los alumnos de primaria comienza 8 a.m. y finaliza a las 12 del mediodía y el turno vespertino comienza 13 p.m. y finaliza 17 p.m., condicionando el resto de las actividades. Aunque no debe descartarse realizar el estudio con otros extremos de intervalos o considerar una mayor granularidad.

Se considera que un punto *x* pertenece a un cluster cuyo clustroide (un punto, existente o no, que oficia como centroide en clusteres en espacios no euclídeos) es *y*, si $d(x,y) \leq \delta$, es decir si la diferencia entre la cantidad de sus conexiones es menor que un umbral a ajustar.

Método de clustering a ser utilizado

Se debe utilizar un método que permita la adaptación de los nuevos datos generados continuamente. Por lo tanto, se descartan los métodos jerárquicos que son de orden n^3 [Leskovec, Rajaraman and Ullman 2014]. En los métodos por asignación de puntos, el algoritmo *k medias* se descarta porque es recomendado para espacios euclídeos [Leskovec, Rajaraman and Ullman 2014]. Cuando los datos aumenten en gran cantidad, para los espacios que no son euclídeos, tratándose de grandes datasets, se sugiere utilizar el algoritmo *GRGPF* (que se adapta a este tipo de problemas), o métodos de procesamiento en paralelo utilizando *MapReduce* [Leskovec, Rajaraman and Ullman 2014].

En esta etapa se procedió a utilizar paquetes de *datamining* disponibles en *R*. Dicho lenguaje tiene implementados los algoritmos *PAM (Partitioning Around Medoids)* y *CLARA (Clustering Large Applications)*. Entre ambos se decide la utilización de *CLARA* debido a que se adecúa al procesamiento de grandes conjuntos de datos. *CLARA* efectúa la búsqueda de clústeres en base a medoides (puntos ya existentes del clúster), prescindiendo de la búsqueda de centroides, que no tendría sentido en el espacio considerado.

4.4 Etapa 4: Obtención de conocimiento de apoyo a la toma de decisiones.

De los indicadores planteados en 3. Objetivo, se mencionan algunos elementos que pueden brindar información de utilidad:

- Tendencias en el acceso a la red: búsqueda de patrones en su utilización según los horarios y variación de la cantidad de conexiones o tráfico de sitios relacionados con estudio a lo largo del año.
- Outliers: centros de estudio que difieren en la utilización de la red en relación a los horarios, lo que permitiría identificar centros que utilizan mínimamente el servicio o centros que lo sobre utilizan, lo cual podría estar asociado a la presencia de intrusos en la red.
- Horarios comunes a todos los centros de estudio donde la utilización es mínima, lo cual serviría para programar tareas de mantenimiento en el servicio sin afectar a los usuarios, entre otras posibilidades

5. Resultados Obtenidos

5.1 Resultados de la Etapa 2 (Análisis Exploratorio de los datos)

Los resultados comprenden el análisis cualitativo de los datos y la búsqueda de *outliers.*

Análisis cualitativo de los datos: Como ejemplo del estudio de una sola muestra de registros, se considera el día 02/03/2016 disponiéndose de los datos de 89 locales. Por medio de métodos estadísticos descriptivos se obtiene una distribución empírica en el uso de la red, reflejada en la gráfica de la *Figura 2, Utilización de la Red para un día específico.* En la imagen Histograma de cantidad conexiones/hora, se puede comprobar, como era de esperarse, que en los horarios de 8 a 17 (horario de clases) el uso es mayor y se reduce en el resto de las horas. Debido a que el histograma es un método pobre para determinar la forma de una distribución (dado que puede ser fuertemente afectado por las cubetas elegidas) y para suavizar la gráfica se utiliza la función densidad de núcleo (*Kernel Density Plot*), disponible en *R*. A los efectos prácticos se consideraron 20 muestras generadas aleatoriamente sin reposición con mil observaciones cada una. En la segunda imagen de la Figura 2, *Estimador Kernel de la*

Densidad, se puede observar qué tan alejado estarían las observaciones de una distribución normal, tendiendo a asemejarse a una mezcla de gausianas.

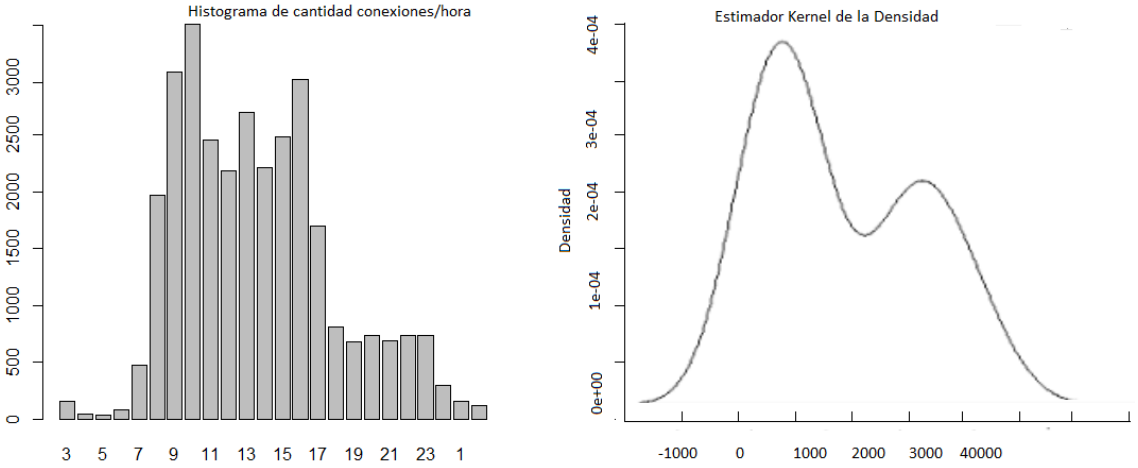

Figura 2: Utilización de la Red para un día específico

Búsqueda de outliers: Previo a cada procesamiento se eliminaron las observaciones que pudieran considerarse anómalas. Para la identificación de estos datos anómalos se consideró, de cada muestra, los locales, agrupados por fecha y hora. Esto es debido a que días de la semana diferentes tienen comportamientos diferentes. Por ejemplo, la actividad de un domingo difiere en sus características de otros días, pero posiblemente sean similares para otros centros de estudio. Este método se centra en una medida de distancia apropiada (diferencia del promedio de la cantidad de conexiones). Por ejemplo, si se considera la hora 4 a.m., generalmente se observa que la actividad es reducida en casi todos los centros de estudio (el máximo detectado en varias observaciones es de 26). La presencia de un local con una actividad muy alta indicaría la presencia de un *outlier*.

A partir de los datos se calculó su media y desviación típica y se considera anómalo todo dato que está fuera del intervalo *[mean(x) - 3*sd(x), mean(x) + 3*sd(x)]*. A tal fin, se utiliza las funciones disponibles en *R, mean(x)* y *sd(x)* en cada uno de los atributos (la hora del día en que se considera la cantidad de conexiones). Los outliers son retirados de la muestra y luego se normalizan los datos por medio de la función *scale* con una media de 0 y desviación estándar 1.

5.2 Resultados de la Etapa 3 (Extracción de información utilizando herramientas estadísticas)

Los resultados obtenidos en el análisis exploratorio sugieren que las observaciones se ajustan a un patrón al considerarse la cantidad de conexiones por hora en los centros de estudio, procediéndose a la búsqueda de agrupamientos en el dataset. De los agrupamientos presentados en los casos de estudio de 4.3 se obtienen los siguientes resultados:

clusteres(horas de uso; cantidad de conexiones); A los efectos de su identificación se utiliza la funcionalidad *clara* del paquete *CLARA* de *R.* La invocación del algoritmo se efectúa por medio de: *clara(dataset, k, samples = n)*, la cual aplica el algoritmo al dataset con *n* muestras de los datos buscando *k* clústeres y considerando a los medoides como sus centros. A los efectos de tener una idea del valor de *k,* se grafica un dendograma que proporciona una visualización de los posibles clústeres en la *Figura 3, Clústeres de conexiones por hora.* Se aprecia que la actividad de conexiones se puede agrupar en dos clústeres. Se ejecuta el algoritmo *CLARA* sobre el dataset pudiéndose visualizar en la segunda imagen de la misma figura, los dos clústeres claramente diferenciados: el de mayor densidad corresponde a los

horarios centrales del día (8 a 16 horas), que se corresponden con los horarios de actividad educativa (se utilizó la funcionalidad *fviz_cluster* disponible en el paquete *factoextra* de *R).*

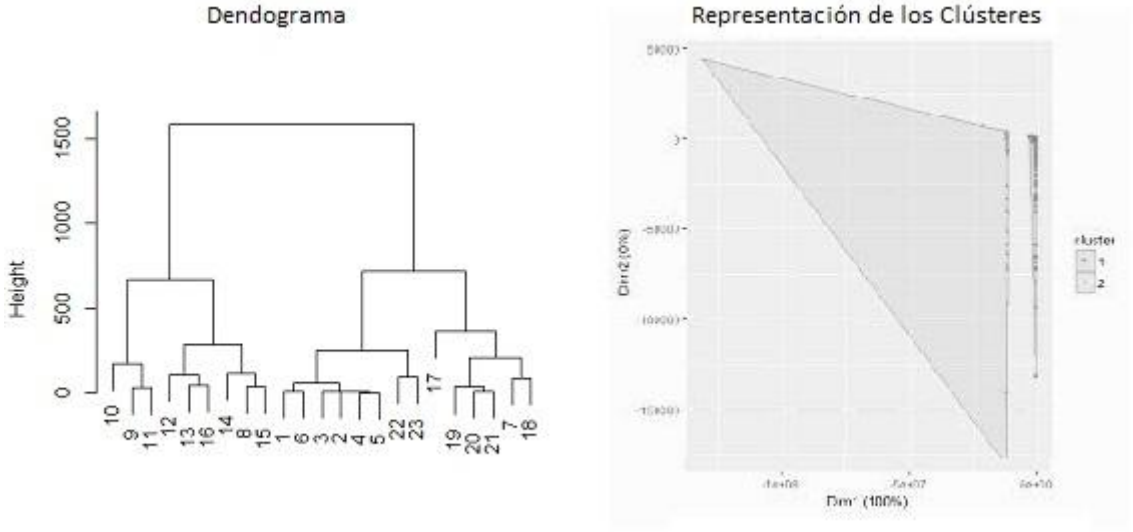

Figura 3: Clústeres de conexiones por hora

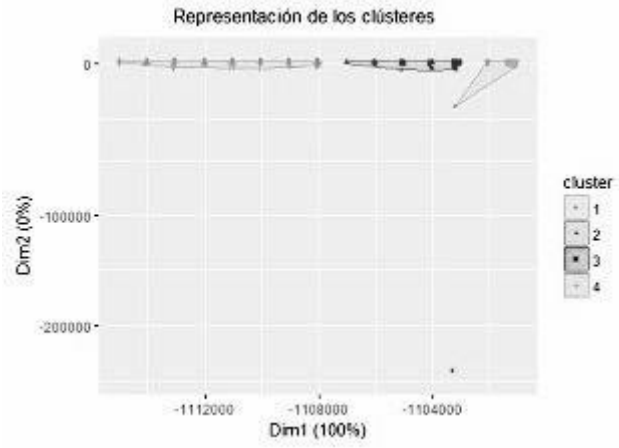

Figura 4: Clústeres de centros de estudio con tráfico de actividad educativa similar

***clusteres(horas de acceso a la red; locales de estudio)*:** Los resultados obtenidos utilizando *CLARA* sobre un conjunto correspondiente a 374 locales de estudio permiten establecer la presencia de tres clusteres y un outlier *(Figura 4: Clústeres de centros de estudio con tráfico de actividad educativa similar).* Los clústeres están compuestos por 97, 137 y 139 locales.

5.3 Resultados de la Etapa 4

Las tendencias en el acceso a la red como ser la búsqueda de patrones en su utilización según los horarios fue analizado en la Etapa 4. La búsqueda de centros de estudio que difieren en la utilización de la red en relación a los horarios (*outliers*) fue tratado en la Etapa 2. Con respecto a la variación de la cantidad de conexiones o tráfico de sitios relacionados con estudio a lo largo del año, la primer imagen (Tráfico General)) de la *Figura 5, Tráfico en toda la muestra*, representa el comportamiento de la cantidad de conexiones para cada día de la muestra. Aproximadamente refleja el incremento del tráfico en el primer semestre del año lectivo 2016, el receso a partir de noviembre y en verano y la vuelta a la actividad en marzo del año 2017. La línea vertical negra indica el salto en los días disponibles

(de mayo a noviembre no existen registros). La segunda imagen, Tráfico para Sitios de Estudio, de la misma figura, muestra el tráfico en los días de la muestra para actividades de estudio. La línea vertical negra indica que de mayo a noviembre no se dispone de registros. Se destaca el hecho de que de noviembre de 2016 a marzo de 2017, se visualiza escasa actividad coincidente con las vacaciones lectivas. Finalmente, en la tercer imagen, Selección de las Expresiones Regulares, se grafica el comportamiento de las expresiones regulares utilizadas, mediante la representación de la cantidad de veces que la expresión regular empleada no fue detectada en el resumen de cada día. En ella se puede visualizar que, a lo largo del tiempo, existe una tendencia a la baja de dicha cantidad (posiblemente debido al aumento del uso del sistema).

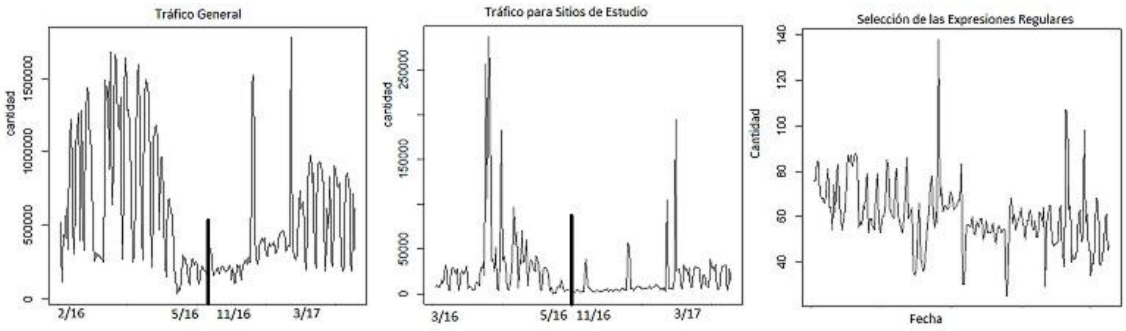

Figura 5: Tráfico en toda la muestra

6. Conclusiones y trabajos futuros

Las búsquedas de clústeres propuestas en los casos de estudio son, prácticamente, resultados directos de la aplicación de algoritmo *CLARA*. En cuanto a la evaluación de las herramientas utilizadas y considerando que fueron empleadas luego de haberse descartado otras, se considera que para las dimensiones que presentaban los datos, resultaron adecuadas. Al momento de la selección de la base de datos para el procesamiento de los datos, se descartó el uso de las relacionales y se optó por la utilización de *NoSQL (MongoDB)* debido a sus posibilidades de escalabilidad (pensando en futuros trabajos) y por el hecho de ser capaces de crecer, en número de máquinas, con pocos recursos. Otro factor a tener en cuenta es la posibilidad de optimizar consultas para la lectura de grandes cantidades de datos. Otra consideración, al momento de la selección de las herramientas, es que tanto *R* como *MongoDB* son de uso libre. Para el procesamiento inicial se descartó el uso de bases de datos relacionales debido a que sus posibilidades de escalabilidad se limitan a residir en grandes máquinas con costosas licencias o que directamente no se pueden adaptar a los casos en que los datos se generen a gran escala. A continuación se enumeran otras tareas relacionadas que pueden abordarse y que serían de utilidad de acuerdo a los objetivos generales.

- La búsqueda de sitios relacionados con temas educacionales a través de expresiones regulares, permite agruparlos según su frecuencia de visitas y de acuerdo al tipo de centro de estudio (primaria o secundaria). Esta información será de utilidad para las autoridades de la Educación al momento de evaluar los materiales de soporte de los cursos incluyendo sitios web de estudio. Adicionalmente, contextualizando dicha información con otros datos (ubicación geográfica u horario de acceso), será posible generar conocimiento más real del uso de los recursos en los diferentes departamentos del País, horarios y estudiando los comportamientos en los distintos años. Para ello se deberá disponer de más registros que se recopilarán en los años venideros.

- Relacionado con el punto anterior, se puede entrenar a un clasificador por medio de algoritmos de aprendizaje automático, para evaluar la performance de las expresiones regulares utilizadas (lo cual redundará en la calidad de los clusteres obtenidos).

- Investigación de la variabilidad de los clústeres al variar el tiempo, *clusteres(horas de acceso a la red; locales de estudio, tiempo).* El último argumento, tiempo, indica los registros de varios días. En este caso, la información obtenida puede indicar, para cada grupo de centros de estudio, si existen horarios fijos durante el día en que la cantidad de usuarios es máxima u horarios en que se mantiene constante la cantidad de conexiones, o, por lo contrario, se detectan fluctuaciones en el uso del servicio, al variar el tiempo. La identificación de tales clústeres permitirá determinar las horas de acceso a la red más utilizadas (horas pico) en el correr de los días a nivel nacional o en una región geográfica determinada. El estudio deberá ser conducido cuando se disponga de más días de registros y se podría emplear series de tiempo.

- Investigación de locales de estudio que acceden determinados sitios web de acuerdo a su localización geográfica, *clusteres(tipo sitio web; locales, horas de acceso a la red, ubicación, tiempo)*. Adicionalmente, se dispone de otros atributos en los datos: sitio web visitado y ubicación geográfica de los centros de estudio, que permitirán realizar otros análisis. Estos nuevos campos proporcionarán mayor riqueza a la información obtenida considerando que, en general, el lugar de residencia se relaciona con su situación socio económica. Por ejemplo se pueden obtener los tipos de sitio web accedidos en función de su ubicación. Asimismo, se puede determinar si existe una clusterización de conexiones en horario de clases o de usuarios que se conectan a determinados sitios en la web, de acuerdo al barrio. Actualmente se está recopilando la información necesaria para identificar correlaciones u otros patrones.

- Investigación de patrones que relacionen empleo de la red o determinados sitios de la web con locales considerados exitosos a nivel educacional. En este sentido, un indicador que sería interesante obtener es la identificación de alguna utilización especifica del Plan que se pueda correlacionar con locales de estudio exitosos en cuanto a las aprobaciones o calificaciones obtenidas por los alumnos. A tales efectos se debería disponer de datos adicionales, como ser la información de cantidad de alumnos de los centros de estudio, entre otras características, para evitar conclusiones que no contemplen la realidad completa de los estudiantes.

- Análisis del tráfico por medio de series de tiempo. Se trata de conducir una investigación, empleando series de tiempo, que permita realizar pronósticos de la utilización de la red y determinar características estacionales, a los efectos de extraer conclusiones relativas a las necesidades del servicio.

References

Ashish Dutt, Saeed Aghabozrgi, Maizatul Akmal Binti Ismail, Hamidreza Mahroeian (2015) "Clustering Algorithms Applied in Educational Data Mining" International Journal of information and Electronics Engineering, Vol. 5, No 2, http://www.ijiee.org/vol5/513-F1002.pdf, June, pages 112-115.

Beijie Xu, Mimi Recker, Xiaojun Qi, Nicholas Flann and Lei Ye (2013) "Clustering Educational Digital Library Usage Data: A Comparison of Latent Class Analysis and K-Means Algorithms" http://www.educationaldatamining.org/JEDM/index.php/JEDM/article/view/21/29, June.

Han J, Kamber M., Pei J. (2012) "Data Mining Concepts and Techniques". Edited by, Elsevier, USA, pages 484-485.

Glaser, B. G. & Strauss, A. L. (1967) "The discovery of grounded theory: Strategies for qualitative research" Edited by Aldine Transaction http://www.sxf.uevora.pt/wp-content/uploads/2013/03/Glaser_1967.pdf, June, pages 60-64.

Leskovec J, Rajaraman A, Ullman J (2014) "*Mining of Massive Datasets*" [versión electrónica] http://infolab.stanford.edu/~ullman/mmds/book.pdf, June, pages 95, 246, 269.