Análisis del avance académico de alumnos universitarios. Un estudio comparativo entre la UTN-FRLP y la UNLP

Guillermo Baldino[1], Laura Lanzarini[2], María Emilia Charnelli[3]
[1] Laboratorio de Innovaciones en Sistemas de Información (LINSI). Dpto de Sistemas. UTN.
[2] Instituto de Investigación en Informática LIDI (III-LIDI).Facultad de Informática. UNLP.
[3] Laboratorio de Investigación en Nuevas Tecnologías Informáticas (LINTI). Facultad de Informática. UNLP.
gbaldino@linsi.edu.ar, laural@lidi.info.unlp.edu.ar, mcharnelli@linti.unlp.edu.ar

Resumen

La deserción y el desgranamiento universitarios son dos de los aspectos que más preocupan a las Universidades Nacionales. El avance académico es uno de los factores con mayor incidencia en estos temas.

Este trabajo propone utilizar técnicas de visualización y de Minería de Datos para identificar cuáles son los atributos más relevantes en lo que se refiere al rendimiento académico de los alumnos de la Facultad Regional La Plata, dependiente de la Universidad Tecnología Nacional.

A través de este estudio se completa la tarea ya realizada sobre la información de los alumnos de la Facultad de Informática de la UNLP. Con el objetivo de efectuar una comparativa entre ambas poblaciones el análisis inicia con los atributos que describen a los alumnos de la UNLP teniendo presentes las diferencias y similitudes existentes entre ambas poblaciones. Su aplicación a la información correspondiente a alumnos regulares y no regulares de la UTN-FRLP ha permitido reforzar algunas de las afirmaciones realizadas sobre las características de la UNLP e incorporar otros aspectos relacionados con la edad y la situación laboral de los estudiantes universitarios.

Conocer las razones que condicionan o favorecen el avance académico de los alumnos universitarios es fundamental a la hora de definir medidas tendientes a mejorarlo y este trabajo muestra una forma de hacerlo.

Palabras clave: Minería de Datos Educativa, Visualización, Avance Académico, Selección de atributos.

Introducción

El avance de la tecnología ha permitido generar volúmenes de datos cada vez más grandes y difíciles de comprender y analizar.
El área educativa no escapa a esta realidad. Por lo general, los establecimientos disponen de información sumamente detallada de cada alumno pero carecen de modelos que les permitan describir de manera objetiva a sus estudiantes. Caracterizar a los estudiantes de una institución académica aporta información no trivial y de utilidad para la toma de decisiones, como por ejemplo, establecer políticas tendientes a mejorar el desempeño académico de los alumnos lo cual redundará en la reducción de la deserción universitaria.
Distintas áreas han tratado de dar soluciones a este problema. Las técnicas de visualización a través de representaciones gráficas, algunas de las cuales son sumamente sofisticadas, han contribuido significativamente a la exploración y entendimiento de estos conjuntos de datos [1-2]. Por su parte, la Minería de Datos reúne un conjunto de técnicas capaces de modelizar y resumir la información, facilitando su comprensión y ayudando a la toma de decisiones en situaciones futuras [3-4].
El objeto de estudio presentado en este artículo es la Facultad Regional La Plata de la UTN. Esta Facultad fue creada el 24 de septiembre de 1954 y la carrera de Ingeniería en Sistemas de Información comenzó a dictarse en el año 1985. El Plan de Estudios fue modificado en el año 1995, reduciendo la

duración de la carrera de seis a cinco años. Si bien el Plan ha sufrido modificaciones en el periodo 1995 – 2014 la duración de la carrera permanece igual.

Anualmente ingresan alrededor de 600 estudiantes de las distintas ingenierías que se dictan, siendo las más numerosas las de Sistemas de Información e Industrial. También se dictan las Ingenierías Química, Eléctrica, Mecánica y Civil.

A efectos de facilitar el acceso a todos aquellos alumnos que estuvieran trabajando y siguiendo con el espíritu con el que fue creada la Universidad, la Facultad prevé tres bandas horarias en sus primeros 3 años de cursada. Luego, los dos últimos años se cursan solo en el turno noche.

Actualmente la problemática de deserción en las carreras de Ingeniería forma parte de una situación a la cual se enfrentan tanto autoridades como docentes. Existen distintas herramientas, desde becas, programas de tutorías y seguimientos por parte de los gabinetes pedagógicos para trabajar con alumnos que abandonan la carrera. Si bien se considera que estas herramientas son útiles e importantes, actúan en instancias en las cuales el alumno ya tomó la decisión de abandonar sus estudios.

La comunidad educativa coincide en la necesidad de hacer esfuerzos para revertir esta situación y cualquier tipo de medidas que se adopten deben estar basadas en información útil para la rápida toma de decisiones. Distintos autores han propuesto diferentes enfoques relacionados con la captación de estudiantes como en el análisis y detección de abandonos y también con la estimación de la duración de la carrera [5-11]. También hay autores que han estudiado cómo evoluciona el progreso de los alumnos durante sus estudios [12-13].

El presente trabajo se enmarca en lo que se conoce como proceso de Extracción de Conocimiento o KDD (Knowledge Discovery in Databases) y técnicas de Visualización aplicadas al análisis de la información disponible.

El proceso de KDD tiene como objetivo la detección automática de patrones sin necesidad de contar con una hipótesis especificada a priori. Sin embargo, su aplicación requiere identificar, en base al problema a resolver, cuál es la información sobre la que se va a trabajar y cuál es el tipo de modelo que se desea obtener.

En referencia a la información sobre la que se va a trabajar, este artículo propone una metodología de trabajo que utiliza visualizaciones de la información disponible para identificar los atributos que mejor caracterizan el avance académico de un alumno logrando reducir así la información a considerar. Esto permite enfocar el análisis en las características adecuadas y arribar a un perfil de alumno de fácil interpretación. Es sabido que el objetivo de una visualización es lograr una representación que ayude al usuario a interpretar un conjunto de datos y comunicar su significado [14]. Sin embargo, es común que la información que se desea representar no tenga una manifestación visual obvia. Ante esta situación, el proceso de mapeo del conjunto de datos originales a la vista minable o información a procesar a través del método seleccionado, puede llegar a ser no trivial [15].

La institución objeto de estudio de este trabajo es la Facultad Regional La Plata de la Universidad Tecnológica Nacional, UTN-FRLP. Este trabajo se enmarca como una continuación del artículo *Selección de atributos representativos del avance académico de los alumnos universitarios usando técnicas de visualización. Un caso de estudio* [16], por lo que se busca realizar una comparativa con los resultados obtenidos previamente al procesar la información de los alumnos de la Facultad de Informática de la Universidad Nacional de La Plata.

A diferencia del sistema SIU-Guaraní que la UNLP utiliza para la gestión académica de sus alumnos, en la UTN-FRLP se utiliza un desarrollo propio llamado Alumnos Web. Este sistema abarca la gestión de los datos de todos los alumnos en todas las etapas, desde la inscripción al curso de ingreso hasta la

culminación de su carrera. La primera versión del sistema fue desarrollada en el año 1996, siendo en aquella época la primera Facultad del país en poseer una aplicación Web de tales características.
Este trabajo está organizado de la siguiente forma: la sección 2 describe el preprocesamiento efectuado sobre los datos originales, la sección 3 muestra la selección de atributos relevantes a través de la generación de diferentes visualizaciones, la sección 4 muestra la construcción de un modelo a partir de los atributos seleccionados y los resultados obtenidos, mientras que en la sección 5 se presentan las conclusiones de este trabajo.

Preparación de los datos

Las primeras etapas del proceso de KDD involucran la comprensión del dominio y la recopilación de los datos. La información de los alumnos de la UTN-FRLP registrada a través del Sistema Alumnos Web contiene datos personales, sociales, laborales y educativos, organizada como se observa en la Tabla 1.
Por las características del sistema, los alumnos de la UTN a diferencia de los de la UNLP, están obligados a completar el cuestionario. Esto permite contar con toda la información al momento de realizar este estudio mientras que para los alumnos de la UNLP la parte final del cuestionario se encuentra vacía en un alto porcentaje de los alumnos.
Una vez finalizada la adquisición de los datos, se continúa con la etapa de preparación y selección de atributos.

1. Datos Personales (estado civil, familiares a cargo, con quien vive, etc.).
2. Financiamiento de estudios (familia, beca, trabajo).
3. Situación laboral (si busca trabajo, cuántas horas trabaja, relación con la carrera.
4. Situación padres (si viven, nivel de estudios y su actividad profesional).
5. Otros estudios.
6. Tecnología (si dispone de PC, acceso a Internet, etc.)
7. Nivel de idiomas.

Tabla 1. Estructura de la encuesta en Alumnos Web

Análisis de atributos cualitativos

Una vez obtenida la información de los alumnos de la UTN-FRLP se procedió a analizar los atributos cualitativos mediante diagramas de barras y se observó que algunos de ellos poseían una moda con una frecuencia elevada. Por ejemplo, el 99,46% dijo no poseer Beca de estudio, el 94,38% tiene a su padre vivo y el 98,07% tiene a su madre viva. Se decidió descartar a los atributos cuya moda superara el 92%. De esta forma se logró reducir la dimensión de los datos de entrada.

Atributos con datos no generalizables

Se eliminaron atributos no generalizables como el nombre del estudiante, el número de documento, el número de legajo y el número de inscripción.
Se redujo la cardinalidad de algunos atributos utilizando categorías más genéricas incrementando así su capacidad predictiva.
Por ejemplo, se reemplazó la información correspondiente al nombre del colegio donde cada alumno cursó el nivel medio por el tipo de colegio, distinguiendo sólo si se trató de una escuela pública o privada.
La creación de características consiste en generar nuevos atributos con el objetivo de mejorar la calidad y comprensión del conocimiento extraído. En esta dirección, se transformó la fecha de nacimiento por la edad de los alumnos. Por otro lado, a partir del año de egreso del secundario y del año de ingreso a la facultad, se generó un nuevo atributo que calcula esta diferencia.
Otras transformaciones que se realizaron tienen que ver con si trabaja y busca trabajo, cómo costea sus estudios, con quién vive. Se realizó la numerización de algunos atributos y la posterior normalización de su rango, de

acuerdo a los requerimientos de las técnicas de Minería de Datos a utilizar. Se numerizó el máximo nivel de estudios de los padres, el tipo de actividad que realizan y la cantidad de horas semanales que trabaja el alumno.
Además de los datos personales de cada alumno, se dispone de toda la información académica de los estudiantes separados en información relacionada a las cursadas y los finales.
Con el objetivo de analizar el avance de los alumnos en sus estudios y por cuestiones de simplicidad sólo se trabajó con la cantidad de finales aprobados por alumno al finalizar cada año durante los primeros 5 años de su vida universitaria. Se consideró que esta cantidad de años resulta representativa y coincide con la duración de las carreras de la Facultad. Los valores de estos atributos se obtienen al calcular para cada alumno la proporción de finales aprobados desde el inicio de su carrera hasta el final de cada año lectivo en relación a la cantidad total de materias según cada carrera como se observa en la siguiente ecuación

$$avance_i = \frac{f_i}{F} \qquad i=1...5 \qquad (1)$$

donde

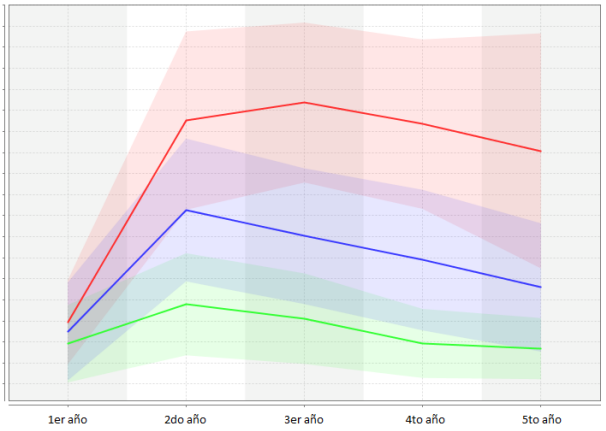
- f_i es la cantidad total de materias que el alumno registra como aprobadas al finalizar el i-ésimo año.
- F es la cantidad total de materias de la carrera.
- $avance_i$ es un valor entre 0 y 1 que representa el avance que el alumno tiene en su carrera al finalizar el i-ésimo año.

Por último, se creó un campo que resume el estado académico del alumno, cuyo valor indica si se trata de un alumno regular o no, para definir la regularidad se estableció el siguiente criterio: serán alumnos regulares todos aquellos alumnos que hayan aprobado o bien una cursada o bien un final durante el desarrollo completo de un año lectivo.

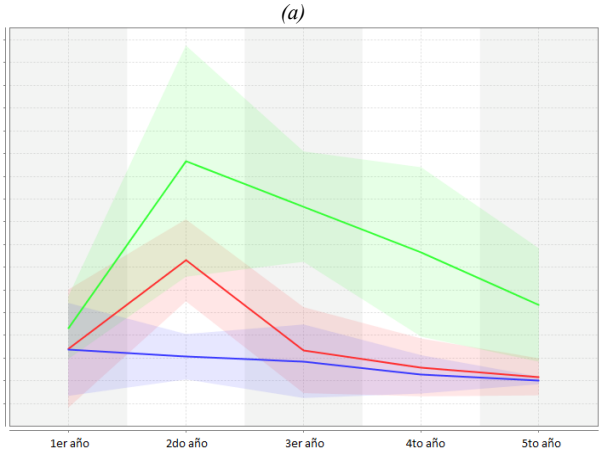
Selección de atributos

Las técnicas de Minería de Datos aplicadas sobre ejemplos de dimensión alta dan como resultado modelos complejos. Dependiendo de la técnica utilizada, datos con esta característica producen o bien árboles enormes o conjuntos de reglas con alta cardinalidad y antecedentes formados por un número importante de conjunciones [17] o funciones discriminantes difíciles de interpretar.
Para resolver este problema es preciso analizar, en forma previa a la construcción del modelo, cuáles son los atributos más representativos de la información disponible. Una vez seleccionados los atributos más relevantes, la técnica a utilizar verá simplificada su tarea y ofrecerá como resultado un modelo más sencillo y fácil de interpretar [18].
En el caso particular del problema a resolver en este artículo, la selección de características juega un rol fundamental ya que se espera poder identificar los atributos adecuados que permitan construir un modelo del avance académico de los alumnos por tratarse de una métrica estrechamente relacionada con la condición de regularidad.
Por lo tanto, luego de la etapa de preparación de los datos, se trabajó con técnicas de visualización para identificar dichos atributos. Tras evaluar diferentes metodologías, se consideró utilizar la técnica de *coordenadas paralelas* ya que resulta adecuada para visualizar conjuntos de datos multidimensionales [19]. Informalmente, esta técnica de coordenadas paralelas consiste en asignarle a cada dimensión un eje y disponer estos ejes paralelamente en el plano. Además de ser una técnica apta para datos multidimensionales, es también apropiada para grandes conjuntos de datos [20].
Con el objetivo de analizar el avance de los alumnos en cada uno de los primeros 5 años de su vida universitaria se utilizaron los atributos creados según la ecuación (1) y se construyeron las gráficas que se observan en la figura 2 separando los alumnos regulares (figura 2.a) de los no regulares (figura 2.b). En cada caso, se buscó identificar 3 grupos de

alumnos: los de mejor desempeño (línea con cruces), los de desempeño medio (línea con cuadrados) y los de bajo desempeño (línea sola).
En la figura 2, se optó por una visualización simplificada de la técnica coordenadas paralelas donde para cada atributo de cada grupo sólo se representa su valor promedio (línea central) y su desviación (zona sombreada que rodea a la línea central). De esta forma, se tiene una representación más conceptual de cada grupo.
Luego, observando la figura 2 puede advertirse la relación que existe entre el rendimiento de los alumnos en los primeros cinco años y su condición de regularidad. Independientemente de si se trata de alumnos regulares o no regulares, en las figuras 2 a) y b) se observa que existe un punto de inflexión en el segundo año y a partir de ese momento una gran cantidad de alumnos detienen su progreso en la carrera.



(a)



(b)

Figura 2 Gráfico de coordenadas paralelas simplificado donde sólo se representa la media y la desviación de los atributos avance$_i$ con i=1:5 para (a) Alumnos regulares (b) Alumnos no regulares

Esto coincide totalmente con lo observado en los alumnos de la Facultad de Informática de la UNLP. Es importante observar con mayor detalle las dificultades por las que atraviesan los alumnos al iniciar el tercer año.
Para realizar el estudio del resto de los atributos se utilizaron dos modelos: uno basado en un árbol de clasificación creado con el C4.5 y otro basado en reglas de asociación. En ambos casos el objetivo fue utilizar las características descriptivas de ambos modelos para identificar los atributos con mayor incidencia en la condición de regularidad de los alumnos.
En el primer caso, utilizando el método C4.5 con un umbral de poda de 0.25 se obtuvo el siguiente árbol

```
Edad <= 20
|   publico_privado = Pub: No
|   publico_privado = Pri: Si
Edad > 20: Si
```

Para medir el desempeño de este modelo se utilizó la tasa de acierto y la precisión de cada clase los cuales se calculan de la siguiente forma

$$tasa\_de\_acierto = \frac{t\_pos + t\_neg}{pos + neg} \quad (2)$$

$$precisión(pos) = \frac{t\_pos}{t\_pos + f\_pos} \quad (3)$$

$$precisión(neg) = \frac{t\_neg}{t\_neg + f\_neg} \quad (4)$$

donde

- *t_pos* y *t_neg* corresponden a la cantidad de casos positivos (alumnos regulares) y negativos (alumnos no regulares) correctamente clasificados por el método respectivamente.
- *f_pos* y *f_neg* representan la cantidad de casos positivos (alumnos regulares) y negativos (alumnos no regulares)

incorrectamente clasificados por el método respectivamente.

- *pos* y *neg* son la cantidad de casos positivos (alumnos regulares) y negativos (alumnos no regulares) reales del problema (las respuestas esperadas).

En este caso la *tasa_de_acierto* es del 82.63% mientras que los valores de *precisión(pos)* y *precisión(neg)* son 99.38% y 15.42%.
Es decir que el modelo representa de una manera adecuada a los regulares pero no ocurre lo mismo con quienes pierden la regularidad.
El atributo *publico_privado* toma valor *pri* o *pub* según si el alumno realizó el nivel medio en un colegio público o privado.

Con respecto a las reglas de asociación se probaron distintas alternativas considerando el conjunto de alumnos completo, y luego analizando los regulares y no regulares de manera independiente. Los resultados obtenidos a partir de la observación de los conjuntos de ítems frecuentes así como de las reglas de asociación se describen en la sección siguiente.

Resultados

La Tabla 2 muestra la lista de atributos seleccionados

Ritmo
Cant.de Hs. que trabaja
Edad
Tipo de escuela media
Nivel de estudios Padre
Nivel de estudios Madre
¿Con quién vive?
Publico_privado

Tabla 2. Atributos seleccionados

Analizando la información de los alumnos de UTN-FRLP puede afirmarse que:

- En base al árbol obtenido en la sección 3 puede afirmarse que para quienes ingresen directamente del nivel medio al universitario haber asistido a una escuela privada es un factor importante para obtener la regularidad.
- Hay una correlación lineal negativa con valor -0.615 entre la edad de los alumnos y el año de ingreso a la universidad lo que indica una demora en el inicio ya que la correlación es leve. A través de un diagrama de cajas realizado sobre el atributo Edad se observa que el rango de la variable es amplio ya que toma valores entre 19 y 60 años.
- Si se analizan sólo los no regulares: el 80% no trabaja, el 86% es de sexo masculino, el 75% fue a escuela pública y al 61% le costean los estudios la familia.
- Si se toman sólo los alumnos regulares: el 65% tiene madre con nivel secundario completo o universitario completo y al 63% le costean los estudios la familia.
- Si se consideran sólo los alumnos que trabajan se obtiene, entre otras, la siguiente regla de asociación
  Si (edad>29.5) y (materias<3) entonces (Cant. de Hs. que trabaja >=40).
  Siendo materias la cantidad total de materias aprobadas durante los primeros 5 años de la carrera.
  Esta regla tiene soporte 0.119, confianza 0.829 e interés 1.487. Es decir que la cumple casi el 12% de los alumnos que trabaja y tiene una precisión de aproximadamente el 83% con respecto a la cantidad de casos que cumplen con el antecedente.

Relacionando lo observado con los alumnos de la UNLP puede afirmarse que:

- Los alumnos de la UNLP son más jóvenes que los de la UTN-FRLP. En el primer caso la edad promedio es de 21 con una desviación estándar muestral de 4.4 años

mientras que en la UTN-FRLP es de 26.78 con un desvío de 4.76 años. Es importante destacar que la dispersión en el rango de edades es muy superior en la UTN.

- Con respecto al tiempo que transcurre desde que terminan el nivel medio e ingresan a la Universidad, en el caso de la UNLP los alumnos tienden a ingresar dentro de los 2 primeros años de finalizado el nivel medio mientras que en la UTN el valor promedio es de 3 años con un desvío aproximado de 3 años. Esto se relaciona con la edad promedio de los alumnos pero, a diferencia de la UNLP, no necesariamente indica que se trata de un aspecto negativo en relación a la regularidad.
- En cuanto a la situación laboral, se advierte una mayor incidencia del tiempo dedicado al trabajo con respecto al avance académico ya que en el caso de los alumnos UTN a cierta edad comienzan a trabajar y eso reduce su desempeño económico. Esto se observa con menos frecuencia en la UNLP.
-

Conclusiones

Se han utilizado técnicas de visualización y de Minería de Datos para identificar los atributos más relevantes en lo que se refiere al rendimiento académico de los alumnos de la Facultad Regional La Plata, dependiente de la Universidad Tecnología Nacional. También se han analizado las características principales de los alumnos que mantienen su condición de alumno regular.
Se ha incluido una comparación con un trabajo previo equivalente realizado con la información de los alumnos de la Facultad de Informática de la UNLP.

Queda pendiente el análisis de los 10 primeros años de permanencia en la universidad por parte de los alumnos UTN-FRLP ya que por motivos laborales suele ocurrir que los alumnos suspendan temporariamente sus estudios y los retomen cuando logran organizar sus respectivas situaciones personales o laborales. Esto es algo que no ocurre en la UNLP y por tratarse de un estudio comparativo entre estas dos poblaciones, en este artículo sólo se ha incluido el análisis de los primeros 5 años.

Referencias

[1] Koutek, M. Scientific Visualization in Virtual Reality: Interaction Techniques and Application Development. Computer Graphics & CAD/CAM group, Faculty of Information Technology and Systems (ITS), Delft University of Technology (TU Delft), 2003.

[2] Nielson, G. M.; Shriver, B. Visualization in scientific computing. IEEE Computer Society Press.United States of America, 1990.

[3] Charnelli, E. Lanzarini, L. Baldino, G. Diaz, F. Determining the profiles of young people from Buenos Aires with a tendency to pursue computer science studies. XX Congreso Argentino de Ciencias de la Computación CACIC, 2014.

[4] Witten H., Frank E., Hall M. Data Mining: Practical Machine Learning Tools and Techniques (3er.edition). Morgan Kaufmann Series in Data Management Systems, Elsevier, 2011.

[5] La Red Martínez, D; Karanik M., Giovannini M., Pinto N. Perfiles de Rendimiento Académico: Un Modelo Basado en Minería de Datos. Revista Campos Virtuales, Vol. 4, Núm. 1, 2015.

[6] Zeng Li, Ling Li, Lian Duan, Kevin Lu, Zhongzhi Shi, Maoguang Wang, Wenjuan Wu, Ping Luo. Distributed data mining: a survey. Information Technology and Management 13, no. 4, p. 403-409, 2012.

[7] Valero S., Salvador A. Predicción de la deserción escolar usando técnicas de minería de datos. Simposio Internacional en Sistemas Telemáticos y Organizaciones Inteligentes SITOI, 2009.

[8] Rodallegas E., Torres A., Gaona B., Gastelloú E., Lezama R., Valero S.

Modelo predictivo para la determinación de causas de reprobación mediante minería de datos. II Conferencia Conjunta Iberoamericana sobre Tecnologías para el aprendizaje CcITA, 2010.

[9] Valero S., Salvador A., García M. Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. II Conferencia Conjunta Iberoamericana sobre Tecnologías para el aprendizaje CcITA, 2010.

[10]. Wang, J., Lu, Z., Wu, W., and Li, Y. The application of data mining technology based on teaching information. In Computer Science Education ICCSE, 2012.

[11] Formia S., Lanzarini L., Hasperué Waldo.

Caracterización de la deserción universitaria en la UNRN utilizando Minería de Datos. Un caso de estudio. Revista TE&ET; no.11, p. 92-98. ISSN: 1850-9959. 2013.

[12] Asif R., Merceron A., Pathan K. Investigating Performances's Progress of Students. CEUR Workshop Proceedings. ISSN 1613-0073, 2014

[13] Bower A.J. Analyzing the Longitudinal K-12 Grading Histories of Entire Cohorts of Students: Grades, Data Driven Decision Making, Dropping Out and Hierarchical Cluster Analysis, 2010.

[14] Lam H., Bertini E., Isenberg P., Plaisant C., Carpendale S. Empirical studies in information visualization: Seven scenarios. Visualization and Computer Graphics, IEEE Transactions on, 18(9), p. 1520-1536, 2012.

[15] Larrea M., Escarza L. et al. Ontologías y semántica en el proceso de visualización. XII Workshop de Investigadores en Ciencias de la Computación WICC, 2014.

[16] Lanzarini L, Charnelli E, Baldino G., Díaz J. Selección de atributos representativos del avance académico de los alumnos universitarios usando técnicas de visualización. Un caso de estudio. Revista TE&ET; no. 15. ISSN: 1850-9959. p. 42-50. 2015.

[17] Bernard M., Boyer L., Habrard A., Sebban, M. Learning probabilistic models of tree edit distance. Pattern Recognition, 41(8), p. 2611-2629, 2008.

[18] Thrun S.B., Bala J. et al. The monk's problems a performance comparison of diferent learning algorithms. Technical report, 1991.

[19] Inselberg, Dimsdale B. Parallel coordinates: A tool for visualizing ultidimensional geometry. IEEE Visualization, p. 361–378, 1990.

[20] Urribarri D. K., Castro S., Martig S.R. Escalabilidad visual en coordenadas paralelas . VIII Workshop de Investigadores en Ciencias de la Computación CACIC, 2006.