Análisis del conjunto de datos PISA 2009 mediante técnicas de agrupamiento

Yuridiana Alemán, David Pinto, and Nahun Loya y Helena Gómez Adorno

Facultad de Ciencias de la Computación,
Benemérita Universidad Autónoma de Puebla, Puebla, México
yuridiana.aleman@gmail.com, dpinto@cs.buap.mx, israel_loya@hotmail.com,
helena.adorno@gmail.com

Resumen PISA es una evaluación internacional promovida por la OCDE con la finalidad de determinar el nivel de conocimientos y habilidades de estudiantes que están por concluir su educación obligatoria (con una edad alrededor de los 15 años). La última prueba aplicada a nivel mundial fue en el año de 2009 y la próxima será en el 2013. De ahí la importancia de analizar las características que han impactado en el hecho de que en México estemos por debajo de la media. En este trabajo, se atiende esta necesidad desde el punto de vista de la inteligencia artificial, usando algoritmos de agrupamiento (*clustering*). Se aplican dos metodos y se evalúan los experimentos usando la medida F-measure, una medida armónica entre *precision* y *recall*. Los valores obtenidos muestran que el conjunto de características seleccionado para representar la colección de datos es adecuado.

1. Introducción

La Organización para la Cooperación y el Desarrollo Económico (OCDE) coordina la aplicación de PISA, una evaluación internacional desarrollada de manera conjunta por todos los países participantes de la OCDE. Esta evaluación es aplicada con el objetivo de determinar si los estudiantes han adquirido los conocimientos y habilidades relevantes para participar activa y plenamente en la sociedad moderna. PISA se aplica cada tres años, evaluando tres áreas de conocimiento: Matemáticas, Ciencias y Lectura, además; en cada periodo se enfatiza una de esta ellas, por lo que se aplican mas reactivos del área seleccionada para su análisis. En la última edición (2009), el área de énfasis fue lectura[1], participando 34 países de la OCDE y 31 asociados [1].

La evaluación de competencias es una parte integral de PISA, y se hace hincapié en el dominio de los procesos, comprensión de conceptos y capacidad para resolver diversas situaciones dentro de cada área [2]. PISA no mide qué tanto pueden reproducir lo que han aprendido, sino que indaga lo que se denomina competencia *(literacy)*; es decir, la capacidad de extrapolar lo que se ha aprendido a lo largo de la vida y su aplicación en situaciones del mundo real, así como la

[1] Instituto Nacional para la Evaluación de la Educación: http://www.inee.edu.mx/

capacidad de analizar, razonar y comunicar con eficacia al plantear, interpretar y resolver problemas en una amplia variedad de situaciones.

El proceso de evaluación pretende constituirse en la base para la investigación y análisis destinados a mejores políticas en el campo de la educación [3].PISA ofrece un conjunto de datos del cual se puede obtener imformación importante, que puede ser procesada por diferentes técnicas (estadísticas, económicas, sociológicas y econométricas). Dentro del campo de la minería de datos se pueden aplicar varias técnicas para el análisis de conjuntos de datos. Un ejemplo sería el uso de técnicas de clasificación con la finalidad de descubrir patrones, reglas de predicción, etc. En este artículo, se utilizan técnicas de clasificación para el análisis de PISA, específicamente clasificación no supervizada.

El objetivo de esta investigación es agrupar a los estudiantes de acuerdo a las respuestas de su examen. Para este propósito se analizan los datos usando dos métodos de clustering: simple k-means y el algoritmo basado en densidad. Se comparán los resultados con el nivel obtenido por los alumnos, a fin de determinar relaciones entre los atributos o características seleccionadas.

El trabajo está estructurado de la siguiente manera: en la Sección 2 se analizan los trabajos relacionados al tema de clasificación y análisis estadístico en el examen PISA. En la Sección 3 se describe la prueba PISA en México, usada en los experimentos de este trabajo. En la Sección 4 se presenta el conjunto de datos usado para llevar a cabo los experimentos de clustering, los metodos de clustering usados, la descripción de las variables y el análisis de los resultados obtenidos. Finalmente se presentan las conclusiones y trabajo futuro.

2. Trabajo relacionado

Algunos autores como [4], [5], [6] y [7] analizan los resultados de la prueba PISA con distintas características, tales como zonas geográficas, atributos cualitativos de estudiantes, entre otros, a fin de proponer cambios en las políticas educativas de algún país o descripción de alguna zona geográfica determinada. Estos estudios son en su mayoría de carácter sociológico o psicológico, y su metodología es principalmente análisis estadístico. En [8], por ejemplo, se analizan los factores determinantes de la calidad educativa para Argentina, especialmente, el papel del tipo de gestión escolar. Para este propósito se utiliza un modelo de regresión multinivel y datos del Programa PISA del año 2006. Entre los principales resultados resalta que la correlacián entre el tipo de gestión escolar (pública o privada) y el rendimiento educativo se disipa al considerar el entorno socioeconómico escolar. En ese mismo país, periódicamente la OCDE publica un reporte oficial con todas las estadísticas obtenidas en el procesamiento de los datos [9]. En [1] se presenta una descripción detallada de todo el proceso y recopilación de las muestras de cada edición PISA, añadiendo también el análisis de los resultados en cada área, para obtener el ranking de los países. Cabe notar que las variaciones entre edición y edición no cambian drásticamente, estando los países mas desarrollados a la cabeza en todas las áreas. En estas mismas publicaciones, se añade el análisis estadístico de cada país destacando las tendencias

de sus resultados, así como comparaciones internacionales con algunas naciones con Producto Interno Bruto (PIB) similar.

3. PISA 2009 en México

En México, la institución responsable de coordinar la aplicación de la prueba PISA en todas sus fases desde 2003 es el Instituto Nacional para la Evaluación de la Educación (INEE). Este instituto realiza su propio análisis sobre México, y presenta comparaciones de los resultados nacionales con un conjunto de países seleccionados *ex profeso*, como el caso de los iberoamericanos y, sobre todo, centra su atención en el análisis al interior del país a fin de indagar sobre los avances nacionales logrados y los alcanzados por las entidades federativas [10]

En su última edición para México, PISA fue aplicado a una muestra representativa de 38,250 estudiantes (1,535 escuelas). Las preguntas de los exámenes son diseñadas para evaluar el nivel de competencia en cada área. El comité de la OCDE define las competencias como sigue:

Competencia lectora: Capacidad de comprender, utilizar y reflexionar sobre textos escritos, con el propósito de alcanzar sus objetivos personales, desarrollar su conocimiento y sus capacidades, para participar la sociedad [2].

Competencia matemática: Es un concepto que excede al mero conocimiento de la terminología y las operaciones matemáticas, e implica la capacidad de utilizar el razonamiento matemático en la solución de problemas de la vida cotidiana [3].

Competencia científica: Incluye los conocimientos científicos y el uso que de esos conocimientos haga un individuo para identificar preguntas, adquirir nuevos conocimientos, explicar los fenómenos científicos y sacar conclusiones basadas en evidencias, sobre asuntos relacionados con la ciencia [2].

Cada pregunta en el examen, tiene asignado un nivel de dificultad,el cual es calculado a partir del número de alumnos que respondieron bien el reactivo. Una vez calculado el puntaje final, se clasifica al alumno en un nivel de competencia por área.El puntaje mínimo para alcanzar cierto nivel se muestra en la Tabla 1. En el nivel 6, 5 y 4 el estudiante se encuentra en uno de los niveles más altos, por lo que tiene potencial para realizar actividades de alta complejidad cognitiva, científicas u otras; un nivel 3 es bastante bueno, aunque no del nivel deseable para la realización de las actividades cognitivas más complejas. El nivel 2 identifica el mínimo adecuado para desempeñarse en la sociedad contemporánea; por último, un nivel 1 ó 0 es insuficiente para acceder a estudios superiores y desarrollar las actividades que exige la vida en la sociedad del conocimiento.

Además del examen, se aplican dos cuestionarios extra, uno dedicado a los estudiantes,y otro a los directores de las escuelas, en donde se abarcan aspectos culturales, socioeconómicos y académicos. En el 2009, México, se posicionó por debajo de la media establecida por la OCDE, la Tabla 2 se da una referencia de la posición de México en el ranking internacional, con respecto a la media establecida por la OCDE, y los paises con mejor y peor puntuación.

Tabla 1. Clasificación por niveles de acuerdo a la OCDE

Nivel	Ciencias	Lectura	Matemáticas
6	707.93	669.3	698.32
5	633.33	606.99	625.61
4	558.73	544.68	552.89
3	484.14	482.38	480.18
2	409.54	420.07	407.47
1	334.94	357.77	334.73
0	-	-	-

Tabla 2. Países y puntajes de PISA 2009

País	**Lectura**	**Matemáticas**	**Ciencias**
Shanghai-China	556	600	575
Promedio OCDE	493	496	501
México	425	419	416
Kyrgyzstan	314	331	330

4. Experimentos realizados

En esta sección se presentan los experimentos de clustering realizados con los datos obtenidos de la prueba PISA 2009. Primeramente se describen los datos usados, así como las técnicas de procesamiento aplicadas. Posteriormente se describen los algoritmos de agrupamiento usados en los experimentos. La medida de evaluación de clusters es mostrada a continuación. Finalmente, se presentan y discuten los resultados obtenidos.

4.1. Preprocesamiento de los datos

Se trabajó con 3 bases de datos obtenidas del sitio oficial de la INEE[2]: Cuestionario escolar, cuestionario de estudiante y respuestas de los estudiantes. Dada la cantidad de reactivos repartidos entre los diferentes exámenes, el numero de atributos de esta base de datos es muy grande, y muy pocos tienen una respuesta en todas las tuplas. De estas bases de datos se utilizaron gráficas y tablas de contingencia para obtener una caracterización general de los datos, con respecto a las escuelas por ejemplo, las privadas son una minoría en todas las modalidades participantes (Secundaria General, Secundaria Técnica, Telesecundaria, Bachillerato General, Bachillerato Tecnológico y Profesional Técnico).

Como el análisis se centra en los estudiantes, se omite la base de datos de las escuelas dentro de las siguientes fases de preprocesamiento:

1. **Sustituir valores perdidos:** Esto se aplicó a algunos atributos, especialmente los referentes a las respuestas de cada una de las preguntas de los

[2] `http://www.inee.edu.mx/`

exámenes, para obtener los valores de promedio y nivel por área. Algunos atributos se llenaron de manera manual, analizando los elementos de a misma clase. En los demás atributos se trabajó como *valor perdido* (?).

2. **Selección de atributos:** Se eliminaron atributos no discriminantes; como los relacionados al país, algunos atributos de varianzas respecto a la media mundial, y donde los valores perdidos superaban el 50 % de las observaciones. En cuanto a las tuplas, no se eliminó ninguna. Además de crearon nuevos atributos a partir de los ya existentes. Por ejemplo, las categorías en cada área utilizando las métricas de la OCDE y clasificación del nivel de competencia, además de la obtención de los promedios generales por área, estos muestran un comportamiento estadístico normal (gaussiana) donde la mayoría se agrupa alrededor de los 400 puntos.
3. **Integración**: Se obtuvo una base de datos reducida, a las que se le aplicó una selección de atributos, rescatando algunas variables de carácter socioeconómico, académico y actividades de los alumnos fuera de clases. Las muestras se separaron por categoría (Alto, Bueno, Regular, Insuficiente) para su posterior análisis. Además, se obtuvieron los promedios por estados, tomando en cuenta el promedio general de los alumnos.Los estados con promedios generales más bajos son: Chiapas, Guerrero y Tabasco, mientras que los promedios más altos pertenecen a los estados de Chihuahua, Nuevo León y el Distrito Federal; sin embargo, la diferencia entre el mejor y peor promedio es de sólamente 90 puntos.

4.2. Algoritmos de agrupamiento

Para el caso que nos atañe, hemos usado las siguientes dos técnicas de agrupamiento implementadas en la herramienta WEKA[3]:

Make Density Based Clusters: El cluster se construye basado en las propiedades de densidad de la base de datos y bajo un enfoque de agrupación natural. Los grupos, y en consecuencia las clases, son rápidamente identificables por que tienen una densidad mayor con respecto a los otros puntos [11].

Simple K-means: Asigna cada punto con el grupo cuyo centro (también llamado centro de gravedad) es el más cercano. El centro es el promedio de todos los puntos del cluster, es decir, sus coordenadas son la media aritmética de cada dimensión de todos los puntos en el grupo [11]. Se puede utilizar la distancia euclídeana (por defecto) o la de Manhattan [12].

4.3. Medida de evaluación

Para evaluar la calidad del agrupamiento, se utiliza la medida *F-Measure*, la cual combina de manera armónica los conceptos de *Precisión* y *Recall* de la recuperación de información [13]. Esta métrica requiere por supuesto que los

[3] http://www.cs.waikato.ac.nz/ml/weka/

corpora usados estén previamente etiquetados. En las ecuaciones 1 y 2 se muestra el cálculo de *Precisión* y *Recall* para el cluster j y la clase i.

$$Pr(i,j) = \frac{n_{ij}}{n_j} \tag{1}$$

$$Re(i,j) = \frac{n_{ij}}{n_i} \tag{2}$$

donde n_{ij} es el número de miembros de la clase i en el cluster j, n_j es el número de miembros del cluster j y n_i es el número de miembros de la clase i.

Los niveles mas áltos de *Precisión* generalmente se obtienen con valores bajos de *Recall*. La ecuación 3 permite calcular el valor *F-Measure*, proporcionando un parámetro $\alpha(0 \leq \alpha \leq 1)$ que permite ponderar *Precisión* y *Recall*[14].

$$F_\alpha(i,j) = \frac{1}{\alpha\frac{1}{Pr(i,j)} + (1-\alpha)\frac{1}{Re(i,j)}} \tag{3}$$

Un valor de *F-Measure* es obtenido para toda la colección calculando un promedio pesado de todos los valores de las métricas *F-Measure*, según la ecuación 4.

$$F = \sum_i \frac{n_i}{n} max_j\{F_a(i,j)\} \tag{4}$$

4.4. Resultados obtenidos

Primeramente, se obtubieron los errores cuadráticos para 2, 3 y 4 clusters, los cuales ascendieron a 352,593.01, 325,685.05 y 311,204.40 respectivamente. Como se observa, los errores son elevados, sin embargo, esto se puede justificar por el número elevado de atributos (33) y la poca correlación que existe en tre estos. De hecho, el menor error se presenta para 4 clusters, pero las muestras se congregan en un sólo grupo.

Con la ayuda de la fórmula F-Measure, se compararon los resultados de los clusters con la clasificación realizada por la INEE, de acuerdo al puntaje obtenido. En la Tabla 3 se comparan los resultados por área y por número de clusters. Éstos indican que la mejor clasificación se obtiene usando sólamente dos clusters (cuando se manejan 4 categorías). Además, el área mejor clasificada fue ciencias, con una F-Measure de 0.67 (K-menas) y 0.73 (Density method). El área de matemáticas es la mas baja en cuanto a *F-Measure* en ambos algoritmos, a pesar de tener una correlación muy fuerte con el área de ciencias. En la Figura 1 se muestra la distribución de los clusters de acuerdo a la categoría en ciencias y del promedio general con el método *Make Density Based Clustered*. Se observa claramente la división de acuerdo a las categorías, donde prácticamente todas las observaciones de nivel *insuficiente* y algunas de nivel *mínimo* son parte del cluster 0, mientras que las categorías de *alto*, *bueno*, y gran porcentaje de *mínimo* son parte de cluster 1.

Tabla 3. Resultados usando F-measure con k-means y cluster basado en densidad

Área	**K-means**	**Densidad**
2 Clusters		
Matemáticas	0.613968	0.672767
Ciencias	0.674797	0.732246
Lectura	0.66827	0.716185
3 Clusters		
Matemáticas	0.574676	0.602189
Ciencias	0.622501	0.647574
Lectura	0.589295	0.595035
4 Clusters		
Matemáticas	0.496032	0.494610
Ciencias	0.518151	0.519770
Lectura	0.455953	0.559141

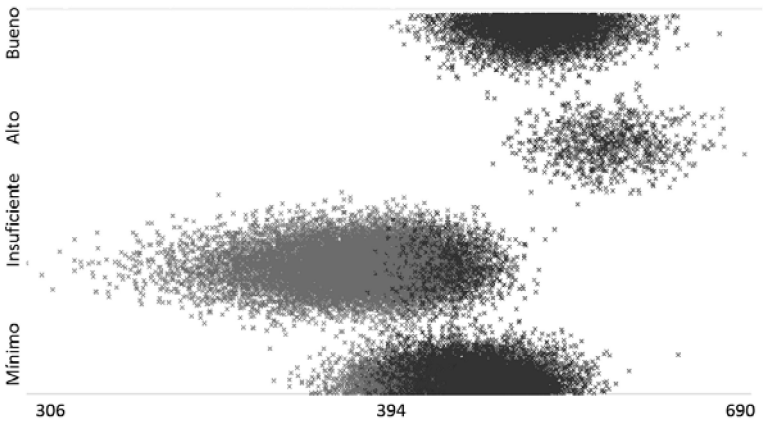

Figura 1. Cluster para categoría de ciencias

Para incrementar el valor de *F-measure*, de acuerdo a lo presentado en [15], donde se menciona que el análisis de clusters con muchos atributos puede ser difuso si se mezclan atributos de diferente origen, se propone dividirlos y hacer un análisis por separado de acuerdo a cada tópico. Se realizó un análisis con algunas características de los estudiantes como lugar de procedencia, promedios y tipo de escuela. La Tabla 4 muestra los resultados obtenidos. En este experimento sólo se utilizó el algoritmo *Make Density Based Clustered*, ya que en éste se obtuvieron mejores resultados de *F-measure*. La categoría de ciencias nuevamente obtuvo un mejor puntaje en todos los casos, llegando a 0.8 para dos clusters.

Tabla 4. F-measure para algoritmo por densidad por área

Clusters	**Matemáticas**	**Ciencias**	**Lectura**
2	0.740541	0.803383	0.719292
3	0.610018	0.667056	0.598345
4	0.63634	0.682440	0.606075

Analizando los promedios de los centroides en los clusters obtenidos con el algoritmo *Make Density Based Clustered*, el cluster 0 contiene el promedio general más bajo, y el valor de *insuficiente* en todas las áreas, mientras que el cluster 1 contiene categorías de *bueno* y *regular*, respecto a los demás atributos, los centroides son iguales para los dos clusters, concentrándose en los valores de atributos mas recurrentes, como son el tipo de sostenimiento (público), escolaridad de los padres (básica) y modalidad (bachillerato general).

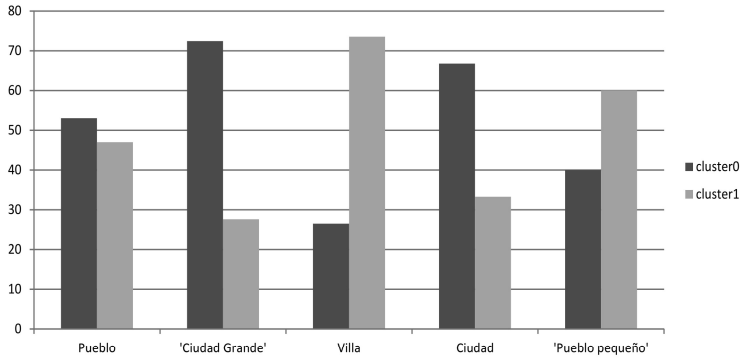

Figura 2. Porcentajes de tipos de localidad distribuidos en los clusters

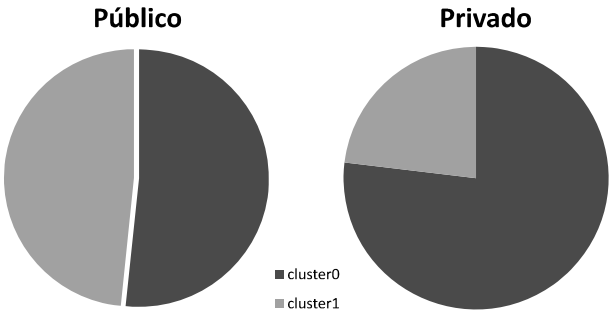

Figura 3. Porcentajes de acuerdo al tipo de sostenimiento de la escuela

En la Figura 2 se muestran los porcentajes de cada tipo de localidad divididas en clusters. En esta Figura se observa que en lugares donde la población es numerosa (*ciudad grande* y *ciudad*), el porcentaje de estudiantes clasificados en el cluster 0 es superior que en el 1 (en el cluster 0 se congregan los alumnos clasificados en las categorías mas altas), mientras que en lugares con pocos habitantes (*villa*, *pueblo pequeño*)la situación se invierte, ya que el porcentaje del cluster 1 es mayor, en el caso del valor de Pueblo, es mayor el porcentaje del cluster 0, aunque en una proporción muy pequeña. Dados los promedios de los clusters (tanto general como categorías en cada área), se observa que en las localidades grandes, el nivel de los alumnos es mayor que en pueblos pequeños.

Para analizar esta situación desde otro ángulo, en la Figura 3 se muestran los porcentajes de alumnos asignados a cada cluster de acuerdo al sostenimiento de la escuela (público o privado). Respecto al sostenimiento público, prácticamente la mitad de los alumnos es asignado a cada cluster, pero en el ámbito privado, mas de tres cuartas partes de los alumnos son clasificados en el cluster 0, y sólo una minoría en el cluster 1. Retomando los centroides, los alumnos de las escuelas privadas (a pesar de ser minoría) tuvieron un mejor puntaje.

5. Conclusiones

En este trabajo, se presentó un análisis sobre PISA 2009 usando clasificación no supervisada. Los algoritmos utilizados fueron *Simple K-means* y *Density Based Clustered* implementados en WEKA. El algoritmo basado en densidad obtuvo mejores resultados en todos los casos (*respecto a F-measure*), especialmente al comparar el área de ciencias, donde se logró obtener un 0.80 con uns selección manual de atributos. A pesar de que la mayoría de las preguntas del cuestionario de estudiantes fueron sobre lectura, esta área no obtuvo una buena medida respecto a la clasificación por clusters.

Los atributos de PISA 2009 son en su mayoría categóricos, y algunos de ellos binomiales. Con la clasificación no supervisada se pudieron obtener agrupaciones

naturales de los datos, sin restringirlos a una clase determinada. Para este caso, la agrupación en dos clusters fue mejor que en 3 y 4, independientemente del número de niveles por área.

Como trabajo futuro, se llevará a cabo un un análisis mas profundo y con enfoque pedagógico de los resultados obtenidos a fin de determinar posibles políticas o estrategias para mejorar a los estudiantes del cluster 1 (nivel insuficiente). También se utilizará esta metodología para el análisis de los datos de PISA 2012 México, que se presentarán en el 2013.

Agradecimientos. Este trabajo ha sido apoyado parcialmente por los proyectos CONACYT #106625 y VIEP #PIAD-ING11-II.

Referencias

1. OECD: El programa pisa de la ocde: Qué es y para qué sirve. Technical report, OECD (2006)
2. OECD: Measuring student knowledge and skills: A new framework for assessment. Technical report, OECD (1999)
3. OECD: Assessing scientific, reading and mathematical literacy: A framework for pisa 2006. Technical report, OECD (2006)
4. Cerveró, A.C., i Gràcia, J.V.: La competencia lectora en el estudio pisa. un análisis desde la alfabetización en información. Anales de Documentación **8**(8) (2008)
5. Ara, M.J.N., Durán, E.J.U.: Pisa y el triángulo de la evaluación. Psicothema **23**(4) (2011) 701–706
6. Rizo, F.M.: Pisa en américa latina: lecciones a partir de la experiencia de méxico de 2000 a 2006. Revista de Educación **1**(1) (2006) 153–167
7. Martínez, J.C., Sebastián Waisgrais, A.C.d.M.: Determinantes del riesgo de fracaso escolar en españa: una aproximación a través de un análisis logístico multinivel aplicado a pisa-2006. Revista de Educación **1**(1) (2010) 225–256
8. Formichella, M.M.: Se debe el mayor rendimiento de las escuelas de gestión privada en la argentina al tipo de administración? CEPAL (105) (2011) 151–166
9. OECD: Pisa 2006 technical report. Technical Report 56393, OECD (2009)
10. Gutiérrez, M.A.D., Vázquez, G.F.: México en pisa 2009. Technical report, INEE (2010)
11. P.Santhi, Bhaskaran, V.: Performance of clustering algorithms in healthcare database. International Journal for Advances in Computer Science **2** (2010) 26–31
12. Arthur, D., Vassilvitskii, S.: k-means++: The advantages of careful seeding. Technical Report 2006-13, Stanford InfoLab (June 2006)
13. Frakes, W.B., Baeza-Yates, R.A., eds.: Information Retrieval: Data Structures & Algorithms. Prentice-Hall (1992)
14. Arco, L., Bello, R., Mederos, J.M., Pérez, Y.: Agrupamiento de documentos textuales mediante métodos concatenados. Revista Iberoamericana de Inteligencia Artificial **10**(30) (2006) 43–53
15. Santana, O.F.: El análisis de cluster : aplicación, interpretación y validación. Papers : revista de sociologia (37) (1991) 65–76