Agrupamientos Semánticos en Glosarios del Universo de Discurso

Marcela Ridao [(1)] Jorge H. Doorn [(1) (2)]

(1) INTIA, Universidad Nacional del Centro de la Provincia de Buenos Aires, Argentina
(2) Universidad Nacional de La Matanza, Argentina
e-mail: {mridao, jdoorn}@exa.unicen.edu.ar

Resumen

Los modelos construidos a lo largo del proceso de desarrollo de software en general, y en las actividades de la Ingeniería de Requisitos en particular, son creados con propósitos y estructuras bien definidas. Estas estructuras han sido concebidas para maximizar la expresividad del modelo en relación con su propósito. En particular, el presente proyecto se enmarca en una estrategia de Ingeniería de Requisitos cuyos principales modelos son el LEL (Léxico Extendido del Lenguaje) y los Escenarios, a partir de los cuales se obtiene la especificación de los requisitos del sistema de software. Los modelos mencionados han sido ampliamente utilizados en el proceso de especificación de los requisitos de diversos casos de estudio. Pese a esto, se están realizado nuevas lecturas de algunos de estos modelos con el objetivo de la extracción de información no inmediatamente legible en los mismos. En el caso de los modelos de Ingeniería de Requisitos basados en lenguaje natural, parece ser que una segunda lectura de algunos de los modelos permite una nueva elicitación de conocimiento. En el presente artículo, se propone una estrategia para visualizar y comprender la estructura semántica de glosarios del proceso del negocio, utilizando grafos asociados a los mismos.

Palabras clave: Ingeniería de Requisitos, Glosarios, Trazado de Grafos, Fuerzas Dirigidas

Contexto

Esta línea de I/D se enmarca en el proyecto acreditado Bases de Datos y Procesamiento de Señales, que se desarrolla en la Facultad de Ciencias Exactas de la Universidad Nacional del Centro de la Provincia de Buenos Aires, INTIA (INvestigación en Tecnología Informática Avanzada).

Introducción

Ultimamente han ido adquiriendo importancia las disciplinas dedicadas al estudio de fenómenos donde el aspecto dominante es la complejidad estructural y no la complejidad esencial de los elementos involucrados en la estructura [1] [2]. En este tipo de problemas se destacan algunos tales como las redes organizacionales o redes sociales organizacionales, las redes de referencias bibliográficas o redes de grupos de interés entre muchas otras.

La representación visual clásica de estas redes se realiza mediante grafos. Pero ocurre que apenas superado un número moderado de nodos, estos grafos resultan inapropiados para distinguir los aspectos relevantes de la estructura.

Algunos de los modelos de la Ingeniería de Requisitos pueden ser estudiados desde el punto de vista estructural. En particular, uno de los más promisorios es el Léxico Extendido del Lenguaje (LEL), que se analiza en la siguiente sección. Si se observa un LEL bajo la óptica estructural se puede construir un grafo donde los símbolos sean los nodos y las menciones a otros símbolos sean arcos dirigidos. Desde este punto de vista, el LEL puede visualizarse como una suerte de red lingüística con una estructura claramente compleja. Es así que, además de la información explícita almacenada en cada nodo, existe una información implícita empotrada en la estructura de las relaciones entre los nodos.

Construir y analizar el grafo de los símbolos del LEL se constituye así en una suerte de minería de información y de conocimiento, ya que se obtiene información sintáctica acerca de la estructura del LEL y al mismo tiempo se adquiere parte del conocimiento subyacente bajo esa estructura.

El conocimiento más elemental deducible de la estructura del LEL está relacionado con la existencia o no de agrupamientos de símbolos por alguna razón no siempre visibles en su visualización plana o en su navegación interactiva. Estos agrupamientos o sub-agrupamientos, de existir, permiten la visualización de componentes de posibles taxonomías del proceso del negocio.

Intuitivamente, se puede suponer que si el Universo de Discurso contiene distintas áreas de interés, fragmentos de la organización o subprocesos diferenciados, entonces se debería esperar cierto grado de acoplamiento mayor entre términos que describen sujetos, objetos, verbos o estados de un fragmento determinado que entre los términos correspondientes a áreas diferentes.

Por lo anterior, la propuesta de este trabajo consiste en analizar el grafo construido a partir del LEL, con el fin de analizar los mencionados agrupammientos.

LEL: Léxico Extendido del Lenguaje

La construcción de un vocabulario que capture la jerga usada por los expertos del dominio ha sido propuesta por distintos autores [3] [4]. De hecho, varias experiencias han mostrado que un glosario del vocabulario de los clientes-usuarios es, en si mismo, una fuente de información para elicitar información del Universo de Discurso [5] [6] [7] [8] [9].

En este trabajo, se estudiará un modelo de glosario en particular: el Léxico Extendido del Lenguaje (LEL). El LEL es una representación de los símbolos del lenguaje del dominio del problema, cuyo objetivo principal es que el ingeniero de requisitos conozca el lenguaje que habla el usuario, sin preocuparse por entender el problema [10] [11].

El propósito de la construcción del léxico no sólo es habilitar una buena comunicación y acuerdo entre los clientes/usuarios y el equipo de ingeniería sino también facilitar la construcción de escenarios y ayudar a su descripción, facilitando la validación.

Este léxico se construye utilizando lenguaje natural y está compuesto por símbolos que pueden ser Sujetos (realizan acciones), Objetos (las acciones se realizan sobre ellos), Verbos (acciones del sistema) y Estados significativos del sistema [12].

Cada símbolo tiene uno o más nombres o frases que lo identifican y dos tipos de descripciones, la noción y el impacto. La noción describe la denotación de la palabra o frase. Indica quién, cuándo ocurre, qué procesos involucra, qué significado tiene el símbolo, etc. El impacto describe la connotación del símbolo, es decir, su repercusión en el sistema. Cada entrada puede contener una o más nociones y uno o más impactos.

En la descripción de los símbolos deben cumplirse simultáneamente dos reglas básicas:

> ***Principio de circularidad***: en la descripción de la noción o impacto de los símbolos se debe maximizar el uso de otros símbolos del léxico.
> ***Principio del vocabulario mínimo:*** se debe minimizar el uso de símbolos externos al lenguaje de la aplicación.

Trazado de Grafos: Métodos dirigidos por fuerzas

La Teoría de Grafos tiene diversidad de aplicaciones. La representación mediante nodos y conexiones es usada para representar redes físicas como circuitos eléctricos, carreteras, moléculas orgánicas, y también interacciones menos tangibles como relaciones sociológicas, bases de datos, o el flujo de control de un programa computacional [13] [14].

El Trazado de Grafos aplica topología y geometría para derivar representaciones de grafos en dos dimensiones. Básicamente, consiste en una representación gráfica del grafo en el plano, usualmente destinada a una visualización conveniente de ciertas propiedades del grafo en cuestión o de los objetos modelados [15] [16].

En este trabajo, al representar el grafo de un LEL, el énfasis se pone en la estructura de la red, y no en los criterios estéticos utilizados generalmente en el trazado de grafos [17], como distribución uniforme de los nodos, minimización de cruces de arcos, uniformidad en la longitud de los arcos, simetría, etc.

Existen diferentes estilos de representación, adecuados a diferentes tipos de grafos o diferentes propósitos de representación [18]. Entre una gran variedad de algoritmos se destaca una familia de métodos conocidos como “dirigidos por fuerzas”. Estos métodos son muy usados hoy en día para dibujar grafos, porque dan buenos resultados, son sencillos de implementar, y son muy flexibles, por lo que pueden ser fácilmente adaptados a aplicaciones concretas con requerimientos de visualización específicos [19] [20]. Estos algoritmos usan analogías físicas para dibujar el grafo. Tienen como denominador común las siguientes características:

- Modelan al grafo como un sistema físico.
- El trazado del grafo es obtenido buscando el equilibrio del sistema físico.

Los modelos físicos más comunes son los que consisten de un sistema de fuerzas (donde generalmente se definen fuerzas que actúan entre los vértices del grafo), en cuyo caso el objetivo del algoritmo es encontrar una posición para cada vértice, de manera que el total de la fuerza ejercida en cada vértice sea cero.

Entre los primeros autores aplicando analogías con sistemas físicos para el trazado de grafos, se destaca el “Spring Embedder” propuesto por Eades [21], que se basa en reemplazar los nodos por anillos de acero y cada arco con un resorte para formar un sistema físico. Los nodos son ubicados en alguna disposición inicial, y se dejan actuar las fuerzas de los resortes hasta lograr un estado de energía mínima. La implementación de Eades, sin embargo, no siguió al pie de la letra la ley de Hooke, sino que incorporó al cálculo de las fuerzas resultantes, fuerzas repulsivas calculadas entre los nodos no conectados.

Otros autores proponen algoritmos derivados del Spring Embedder de Eades, entre ellos Kamada y Kawai [22], Davidson y Harel [23], y Fruchterman y Reingold [17]. Este último se basa en los siguiente principios:

- Los nodos conectados por un arco deberían ser dibujados cerca.
- Los nodos no deberían ser dibujados *demasiado* cerca uno de otro.

Cuán cerca se deberían ubicar los nodos, depende de cuántos haya y cuánto sea el espacio disponible. El algoritmo se basa en simulaciones moleculares o planetarias. Si los nodos se comportan como partículas atómicas o cuerpos celestes, ejerciendo fuerzas atractivas y repulsivas sobre los demás, las fuerzas inducen movimiento. Sin embargo, no se propone una simulación exactamente fiel a la realidad. Del mismo modo que en el algoritmo de Eades, sólo los nodos que son vecinos se atraen entre si, mientras todos los vértices se repelen unos a otros. Esto es consistente con la asimetría propuesta por los dos principios antes enunciados.

Objetivos

Con el fin de detectar agrupamientos de símbolos, se propone aplicar el algoritmo propuesto por Fruchterman y Reingold en la visualización de los grafos correspondientes a los Léxicos de diferentes casos de estudio.

Para un LEL dado, entonces, se propone construir un grafo donde los símbolos son los nodos y las menciones a otros símbolos arcos dirigidos.

Para la configuración inicial, los nodos se ubican al azar en el marco de trabajo, y luego se aplica el algoritmo, modificando la ubicación de los nodos en forma iterativa. Cada iteración consta de tres pasos:

- calcular el efecto de las fuerzas atractivas sobre cada nodo,
- calcular el efecto de ls fuerzas repulsivas y
- limitar el desplazamiento total.

Las fuerzas

f_a y f_r son las fuerzas de atracción y de repulsión, respectivamente:

$$fa(d) = c1 * d^2/k \qquad k = C * \sqrt{\frac{Area}{NúmeroNodo \quad s}}$$

$$fr(d) = c2 * k^2/d$$

d es la distancia entre los vértices y k el radio vacío alrededor de un nodo

Las constantes C, c1 y c2 son obtenidas experimentalmente.

El marco de trabajo

El grafo debe ser confinado al marco especificado por el usuario. Para ello, el algoritmo considera la ubicación de nodos ficticios en el perímetro del marco de trabajo, que ejercen fuerzas repulsivas sobre los nodos del grafo, pero ellos mismos permanecen fijos. Por lo tanto, el marco se modela como cuatro paredes que contienen al grafo dentro de ellas.

Resultados obtenidos

Con el fin de verificar si el algoritmo propuesto permite detectar agrupamientos de símbolos, se lo aplicó a diferentes casos de estudio cuyos Léxicos habían sido verificados y validados previamente.

En primera instancia, se aplicó el algoritmo a dos casos de estudio para los cuales se prefirió no analizar en forma semántica la presencia o ausencia de agrupamientos de símbolos. Los casos analizados fueron:

Caso 1. Sistema de Planes de Ahorro Previo para la Adquisición de Vehículos 0Km [24]

Caso 2. Sistema de Depósito para una planta de Producción de Cerámicos [25]

En las figuras 1 y 2, se presentan los resultados obtenidos para cada uno. Cabe destacar que, tanto en éstas como en las siguientes figuras, no se muestran los arcos correspondientes a los vínculos entre los símbolos, para permitir una mejor visualización de la distribución de los nodos.

En los grafos presentados en las figuras no se observan agrupamientos. El análisis semántico de los Léxicos correspondientes a ambos casos de estudio, confirma la ausencia de agrupamientos destacados, validando el resultado arrojado por el algoritmo.

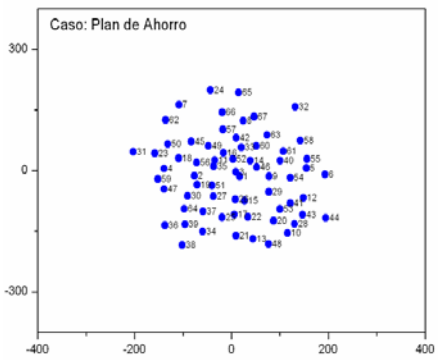

Fig. 1. Distribución de nodos caso 1

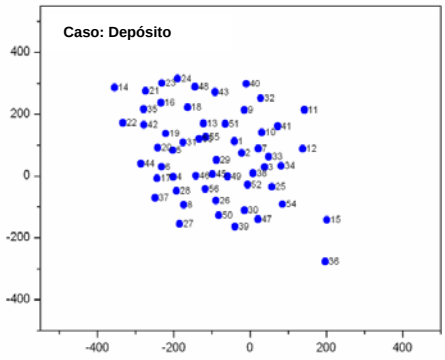

Fig. 2. Distribución de nodos caso 2

Luego, se aplicó el algoritmo a un caso de estudio para el cual se conocía de antemano la existencia de clusters.

Caso 3. LEL del proceso de Construcción de LEL y Escenarios [26].

Para este caso de estudio, existen al menos dos agrupamientos: uno constituido por los símbolos correspondientes a la Construcción del LEL, y otro constituido por los símbolos correspondientes a la construcción de Escenarios.

En la figura 3 se presenta la distribución de los nodos obtenida con el algoritmo.

Se verifica que todos los símbolos que corresponden a la construcción del LEL se agrupan a la izquierda del grafo, mientras los que corresponden a la construcción de Escenarios se agrupan a la derecha. En la zona central, se ubican aquellos símbolos relacionados tanto con un proceso como con el otro. Por ejemplo, el nodo 8 corresponde al símbolo Cliente-Usuario, y el nodo 49 corresponde al símbolo Contexto del Problema.

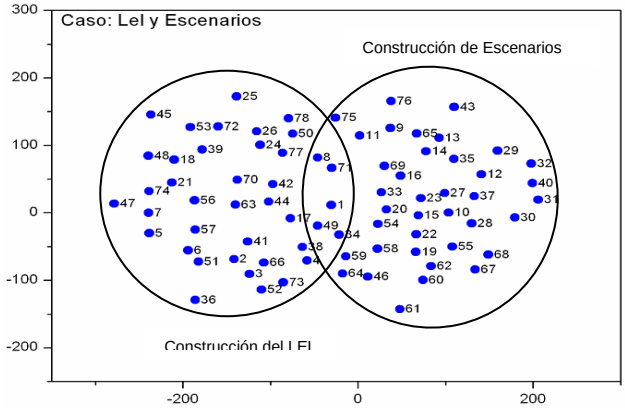

Fig. 3. Distribución de nodos para el caso 3

No se ha estudiado aún la forma de determinar, para los nodos ubicados entre clusters, cuáles pertenecen a uno de los grupos, y cuáles corresponden a la intersección entre ellos.

Por último, se aplicó el algoritmo a:

Caso 4. Relación entre productores de papa y una empresa acopiadora.

Un análisis semántico preliminar de este caso de estudio, determinó la existencia de dos agrupamientos de

símbolos; un grupo correspondientes a Producción y Entrega de Papas y otro a Canje de Semillas.

Sin embargo, en la visualización del grafo luego de la aplicación del algoritmo, se observa claramente la presencia de tres agrupamientos, como puede verse en la figura.

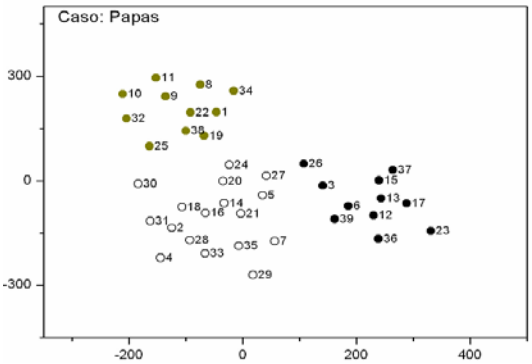

Fig. 4. Distribución de nodos para el caso 4

Con el fin de detectar los límites de los clusters visualizados, se aplicó el clasificador *K*-Means [27] a los datos obtenidos, indicando la presencia de 3 grupos. En la figura 4 se muestran los clusters detectados con diferentes colores, y en la tabla 1, se presentan algunos de los símbolos pertenecientes a cada agrupamiento.

Tabla 1. Fragmento de la lista de símbolos correspondientes a cada uno de los clusters

Producción y Entrega
2 Bonificación
4 Causa justificada
5 Contrato de adquisición y producción
14 El Productor
16 Recibir papa
18 Establecer Programa de Entregas
20 Inspeccionar papa
21 MC S.A.
24 Pagar por la papa
…
Canje de Semilla
3 Cancelar la operación de canje
6 Contrato de canje de semilla
12 División de Semillas MC S.A.
13 El Comprador
15 El Vendedor
17 Entregar semillas
23 Orden de carga
26 Papa Grado MC Consumo
37 Semilla de Papa
…
Calidad
1 Agrietadura
8 Corazón hueco
9 Defecto de calidad
10 Defecto externo
11 Defecto interno
19 Grado MC Consumo
22 Malformación
25 Papa chica
38 Tolerancias establecidas
…

Un análisis semántico más profundo del caso de estudio, confirmó la presencia de estos tres agrupamientos. En particular, el tercer grupo, correspondiente a Calidad de la Papa, no había sido detectado en el análisis inicial, ya que los símbolos incluidos en él, se habían considerado parte del cluster correspondiente a Producción y Entrega de Papa.

Es importante destacar que este resultado realza la capacidad predictiva de la estrategia, ya que, más allá de confirmar la presencia de agrupamientos conocidos, permite descubrir la presencia de otros.

Determinación del número de clusters

En los primeros dos casos, si se observan los grafos resultantes, se puede apreciar cierta simetría tanto en el eje X como en el eje Y. Por ello, podría asumirse que, si el cociente entre las variancias sobre el eje X y sobre el eje Y es próximo a 1, se está en presencia de un grafo con un solo cluster

Para el tercer caso, donde las clases son 2, se observa una forma elipsoidal en la distribución de los nodos. Por otro lado, se observa simetría tanto en X, como en Y. Se podría inferir que si el cociente entre las variancias es mayor que 1, se está en presencia de un grafo con dos clusters.

Para el último caso, en cambio, se puede ver que no es posible asociar esta forma con una elipse, como en el caso 2. Además, se observa una clara asimetría en los ejes X e Y. Por lo tanto, un coeficiente que indique la presencia de asimetría permitiría determinar la presencia de tres o más clusters, mientras que el valor de este coeficiente sería cercano a cero para los casos en que hay uno o 2 clusters (casos 1 al 3).

Conclusiones y trabajo futuro

Se ha propuesto un algoritmo que permite detectar agrupamientos de símbolos en un modelo del proceso de Requisitos: el Léxico Extendido del Lenguaje. El análisis de los resultados obtenidos hasta el momento, permite concluir que es posible efectuar una segunda lectura de algunos de los documentos de Ingeniería de Requisitos mediante el estudio de las estructuras semánticas subyacentes. El uso de grafos y el análisis de su estructura son herramientas apropiadas para dichos estudios.

Los resultados obtenidos se acoplan en forma muy natural con estrategias genéricas de clusterización como K-Means. El conocimiento previo del número de clusters es imprescindible para el uso de K-Means y la estrategia propuesta lo provee.

Queda pendiente el análisis de más casos de estudio para definir los límites inferior y superior para el cociente de variancias y coeficiente de asimetría que permitan determinar más precisamente la presencia de un número determinado de clusters.

Se planea estudiar, también, si la ubicación inicial de los nodos puede afectar la disposición final por el efecto que determinado nodo pudiera causar sobre sus vecinos, evitando su normal desplazamiento.

Además, se pretende analizar la aplicación del algoritmo a otras otras formas de representación del grafo. Por ejemplo, se planea estudiar la representación en tres dimensiones.

Formación de Recursos Humanos

El grupo de investigación que participa en el proyecto Bases de Datos y Procesamiento de Señales está compuesto por 7 integrantes, todos ellos docentes de la Fac. de Ciencias Exactas categorizados en el Programa de Incentivos.

La línea de I/D continúa con la investigación llevada a cabo para la tesis de Maestría en Ingeniería de Software de la integrante del grupo Marcela Ridao, y forma parte de su tesis doctoral.

Referencias

1. Barabasi, A.: Linked, The New Science of Network.. Perseus publishing (2002)
2. Dorogovtsev, S., Mendes, J.: Evolution of networks: From biological nets to the Internet and WWW. Oxford University Press, Oxford (2003)
3. Arango, G., Schafer, W., Prieto, R.: Domain Analysis Methods – Software Reusability. Ellis Horwood Ltd (1993)
4. Leite, J., Franco, A.: O Uso de Hipertexto na Elicitaçao de Linguagens da Aplicaçao. En: IV Simpósio Brasilero de Engenharia de Software, SBC, pp. 134-149. Brazil (1990)
5. Ben Achour, C., Rolland, C., Maiden, N., Souveyet, C.: Guiding Use Case Authoring: Results of an Empirical Study. In: International Symposium on Requirements Engineering, pp. 36-43. IEEE Computer Society Press, Limerick-Ireland (1999)
6. Rolland, C., Ben Achour, C.: Guiding the construction of textual use case specifications. Data & Knowledge Engineering 25, 125-160 (1998)
7. Oberg, R., Probasco, L, Ericsson, M.: Applying Requirements Management with Use Cases. Rational Software Corporation (1998)
8. Regnell, B.: Requirements Engineering with Use Cases – a Basis for Software Development, Doctoral Thesis. Department of Communication Systems. Lund University (1999)
9. Prakash, S., Aurum, A., Kox, K.: Requirements Engineering Practice in Pharmaceutical and Healthcare Manufacturing. In: 11th Asia-Pacific S.E. Conference, pp. 402-409 (2004)
10. Leite, J., Franco, A.: A Strategy for Conceptual Model Acquisition. In: IEEE International Symposium on RE, pp. 243-246. IEEE Computer Society Press (1993)
11. Leite, J., Rossi, G., Balaguer, F., Maiorana, V., Kaplan, G., Hadad, G., Oliveros, A.:. Enhancing a Requirements Baseline with Scenarios Requirements Engineeering Journal 2(4), 184-198 (1997)
12. Leite, J., Doorn, J., Kaplan, K., Hadad, G., Ridao, M.: Defining System Context Using Scenarios. In: Leite, J., Doorn, J (eds.) Perspectives on Software Requirements. Kluwer Academic Press, pp. 169-199 (2004)
13. Gross, J., Yellen, J.: Editors. Handbook of Graph Theory. CRC Press (2003)
14. Gross, J., Yellen, J.: Editors. Graph Theory and Its Applications, Second Edition (Discrete Mathematics and Its Applications). Chapman & Hall/CRC (2006)
15. Di Battista, G. , Eades, P. , Tamassia, R., Tollis, I.: Graph Drawing: Algorithms for the Visualization of Graphs. Prentice Hall (1999)
16. Brandes, U., Kenis, P., Wagner, D.: Communicating centrality in policy network drawing. IEEE Transactions on visualization and computer graphics 9(2), 241-253 (2003)
17. Fruchterman, T., Reingold, E.: Graph Drawing by Force-directed Placement. Software-Practice and Experience DOI 10.1002/spe.4380211102 Wiley Online Library (2006).
18. Kaufmann, M., Wagner, D.: (eds.) Drawing graphs: methods and models, LNCS, vol 2025. Springer-Verlag (2001)
19. Walshaw, C.: A multilevel algorithm for force-directed graph-drawing. Journal of Graph Algorithms and Applications 7(3), 253-285 (2003)
20. Aiello, A., Silveira, A.: Trazado de grafos mediante métodos dirigidos por fuerzas: revisión del estado del arte. Tesis de Licenciatura en Ciencias de la Computación (2004)
21. Eades, P.: A heuristic for graph drawing. Congressus Numerantium 42, 149–160 (1984)
22. Kamada, T., Kawai, S.: An algorithm for drawing general undirected graphs. Information Processing Letters 31, 7-15 (1989)
23. Davidson, R., Harel, D.: Drawing graphs nicely using simulated annealing, ACM Transactions on Graphics 15(4), 301-331 1996.
24. Rivero, L., Doorn, J., del Fresno, M., Mauco, V., Ridao, M., Leonardi, C.: Una Estrategia de Análisis Orientada a Objetos basada en Escenarios: Aplicación en un Caso Real. En: WER'98 - Workshop en Engenharia de Requisitos, pp. 79-90. Maringá, Brasil (1998)
25. Sánchez, M.: Léxico Extendido del Lenguaje y Escenarios para el Almacén de una Fábrica, Tesis de Grado. Universidad Nacional del Centro de la Pcia. de Bs.As. (2002)
26. Garcia, O., Gentile, C.: Diseño de una herramienta para contrucción de LEL y Escenarios, Graduation dissertation. Universidad Nacional del Centro de la Pcia. de Bs. As. (1999)
27. Peña, J., Lozano, J., Larrañaga P.: An empirical comparison of four initialization methods for the K-Means algorithm. Pattern Recognition Letters 20, 1027-1040 (1999)