Agrupamientos en Glosarios del Universo de Discurso

Marcela Ridao[1], Jorge Doorn[2,3]

[1]INTIA, Facultad de Ciencias Exactas, UNCPBA - Campus Universitario, Paraje Arroyo Seco S/N, Tandil(7000)

[2]Escuela de Informática, Universidad Nacional del Oeste

[3]Departamento de Ingeniería. Universidad Nacional de La Matanza - mridao@exa.unicen.edu.ar, 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. Éstas han sido concebidas para maximizar la expresividad del modelo en relación con su propósito. Pese a esto (y muy posiblemente debido a esto) puede ocurrir que dichos modelos contengan información no perceptible durante su uso rutinario. En el caso de algunos modelos de IR basados en lenguaje natural se ha observado que una segunda lectura permite adquirir al menos parte de esa información oculta. Se propone una estrategia para visualizar la existencia de agrupamientos de términos en el Léxico Extendido del Lenguaje que, efectivamente, se corresponden con núcleos semánticos del Universo del Discurso que está siendo estudiado. Esta estrategia se basa en la construcción automática de grafos utilizando los vínculos hipertextuales empotrados en el modelo.*

Palabras Claves: *Ingeniería de Requisitos, Agrupamientos Semánticos, LEL, métodos dirigidos por fuerzas.*

Abstract: *During software development, and especially during Requirements Engineering process, models are created with well-defined purposes and structures. These structures have been thought to maximize the model expressiveness in relation to its purpose. Despite this and quite possibly because of this, it may happen that such models contain information not seen during its regular use. In some Requirements Engineering models, written in natural language, it has been noticed that a second reading can allow to recover at least part of that hidden information. In this paper, a strategy is proposed to visualize the existence of clusters of terms in the Language Extended Lexicon. These clusters are deeply linked to semantic groups in the Universe of Discourse being studied. The strategy is based on automatic construction of graphs using the model hypertext links.*

Keywords: *Requirements Engineering, Semantic Clusters, LEL, force-directed methods.*

INTRODUCCIÓN

Últimamente 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 (Barabasi, 2002; Dorogovtsev and Mendes, 2003). Existen numerosos ejemplos, en diversas disciplinas, donde la detección de agrupamientos representa una contribución significativa a la mejor comprensión del fenómeno que está siendo estudiado (Sarmah et al., 2011 , Mo et al., 2012, Zimmermann et al., 2012). En particular, el método de priorización Pirogov, propuesto por Duan (Duan et al., 2009) analiza documentos del proceso de requisitos mediante la detección de agrupamientos. En dicho método los requisitos se ubican en clusters mediante algoritmos automáticos iterativos, se priorizan los clusters negociando entre los involucrados obteniendo así un ranking de requisitos de acuerdo al cluster al que pertenece cada uno.

En este tipo de problemas se destacan algunos tales como las redes organizacionales o las sociales organizacionales, las redes de referencias bibliográficas o las 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, los mismos 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 1más promisorios es el Léxico Extendido del Lenguaje (LEL) (Leite and Franco, 1990). Este modelo registra el vocabulario del Universo del Discurso1 mediante la descripción de los términos utilizados por el cliente-usuario postergando la comprensión del problema (Ver Sección 2).

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 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. Epistemológicamente este enfoque es muy similar al utilizado en minería de datos (Artz, 2009), en el sentido que se hace visible información oculta mediante el uso de una técnica notoriamente diferente a la utilizada rutinariamente.

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 no siempre visibles en su navegación interactiva.

Estos agrupamientos o sub-agrupamientos, cuando existen, 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.

La propuesta de este trabajo consiste en analizar el grafo construido a partir del LEL con el fin de detectar los agrupamientos mencionados.

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 (Arango and Prieto, 1991; Schäfer et al., 1993; Leite and Franco, 1990). De hecho varias experiencias han mostrado que un glosario del vocabulario de los clientes-usuarios es, en sí mismo, una fuente de información para elicitar información del Universo de Discurso (UdeD) (Ben Achour et al., 1999; Rolland and Ben Achour, 1998; Oberg et al., 1998; Regnell, 1999; Prakash et al., 2004).

En este trabajo se analizará 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 que intenta

[1]UdeD: todo el contexto en el cual el software se desarrolla e incluye todas las fuentes de información y todas las personas relacionadas con el software: usuarios, ingenieros de software, expertos del dominio, etc. Es la realidad acotada por el conjunto de objetivos establecidos por quienes demandan una solución de software (Leite et al., 2004).

capturar el vocabulario de una aplicación. Su objetivo principal es que el ingeniero de requisitos conozca el lenguaje que habla el usuario entendiendo los términos que usa, sin preocuparse por entender el problema (Leite and Franco., 1993; Leite et al., 1997). Involucra la denotación y la connotación de cada símbolo descubierto como una palabra o frase relevante al dominio de la aplicación. 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, en primera instancia, por símbolos que pueden ser objetos activos o Sujetos (realizan acciones), objetos pasivos u Objetos (las acciones se realizan sobre ellos), Verbos (acciones del sistema) y Estados significativos del sistema (Leite et al., 2004).

Cada símbolo en el léxico tiene uno o más nombres o frases que lo identifican y dos tipos de descripciones, la noción y el impacto. La primera describe la denotación de la palabra o frase. Indica quién, cuándo ocurre, qué procesos involucra, qué significado tiene el símbolo, etc. El segundo describe la connotación del símbolo, es decir, su repercusión en el sistema. Esta descripción puede contener acciones que se ejecutan sobre otros objetos o que se aplican al que se está describiendo, situaciones derivadas de la que se está definiendo, etc. 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 (Leite and Franco, 1990):

- Principio de circularidad: en la descripción de la noción o impacto de los símbolos se debe maximizar el uso de otros símbolos del léxico.

De esta manera el conjunto de símbolos determina una red que permite representar al LEL mediante un hipertexto que puede ser navegado para conocer todo el vocabulario del problema.

- Principio del vocabulario mínimo: se debe minimizar el uso de símbolos externos al lenguaje de la aplicación. De este modo se acota el lenguaje al menor conjunto de símbolos posible. Si se utilizan símbolos externos, éstos deben pertenecer al vocabulario básico del lenguaje natural que se está utilizando.

En la Figura 1 se presenta el modelo utilizado para representar los símbolos.

LEL: REPRESENTACIÓN DE LOS SÍMBOLOS EN EL LENGUAJE DEL DOMINIO DE LA APLICACIÓN.
SINTÁXIS: {SÍMBOLO},^{N}

SÍMBOLO: ENTRADA DEL LÉXICO QUE TIENE UN SIGNIFICADO ESPECIAL EN EL DOMINIO DE LA APLICACIÓN.
SINTÁXIS: {NOMBRE},^{N} + {NOCIÓN}.^{N} + {IMPACTO},^{N}

NOMBRE: IDENTIFICACIÓN DEL SÍMBOLO. MÁS DE UNO REPRESENTA SINÓNIMOS.
SINTAXIS: PALABRA | FRASE.

NOCIÓN: DENOTACIÓN DEL SÍMBOLO. DEBE SEREXPRESADO USANDO REFERENCIAS A OTROS SÍMBOLOS Y USANDO UN VOCABULARIO MÍNIMO.
SINTAXIS: SENTENCIA.

IMPACTO: CONNOTACIÓN DEL SÍMBOLO. DEBE SER EXPRESADO USANDO REFERENCIAS A OTROS SÍMBOLOS Y USANDO UN VOCABULARIO MÍNIMO.
SINTAXIS: SENTENCIA.

Figura 1 - Modelo del Léxico Extendido del Lenguaje. Sentencia está compuesta por Símbolos y No-Símbolos pertenecientes al vocabulario mínimo, + significa composición, | significa or, y {x}M N significa de M a N ocurrencias de x.

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, etc. También se utilizan los grafos en la representación de interacciones menos tangibles como puede ocurrir en ecosistemas, relaciones sociológicas, bases

de datos o en el flujo de control de un programa computacional (Gross and Yellen, 2003; Gross and Yellen, 2006; Brandes et al., 2003).

El Trazado de Grafos, como una rama de la Teoría de Grafos, aplica topología y geometría para derivar representaciones de los mismos en dos dimensiones. El trazado de un grafo es básicamente una representación gráfica de él en el plano, usualmente destinada a una visualización conveniente de ciertas propiedades del grafo en cuestión o del o los objetos modelados por el mismo (Di Battista et al., 1999).

Un grafo G = (N,A) es un conjunto N de nodos y un conjunto A de arcos en el cual un arco une un par de nodos. Normalmente los grafos se dibujan con los nodos como puntos en un plano y sus arcos como líneas o segmentos curvos conectando esos puntos. Existen diferentes estilos de representación adecuados a diferentes tipos de grafos o diferentes propósitos de representación.

En el trazado de grafos existen criterios estéticos mayormente aceptados (Fruchterman and Reingold, 1991):

- Distribuir los nodos uniformemente en el marco de trabajo.
- Minimizar los cruces de arcos.
- Hacer que las longitudes de los arcos sean uniformes.
- Reflejar simetría.
- Adecuarse al marco de trabajo.

La mayoría de estos criterios, sin embargo, atentan contra el objetivo de visualización de grafos que se desarrollará en esta propuesta donde el énfasis se pone en la estructura de la red y no en aspectos estéticos.

La generación automática del trazado de grafos tiene importantes aplicaciones en muchas áreas de las ciencias de la computación tales como compiladores, bases de datos, ingeniería de software, VLSI, diseño de redes, interfaces gráficas, etc. La aplicación en otras áreas incluye análisis gráfico de datos (por ejemplo en todos los campos de la ingeniería, biología o ciencias sociales) y la visualización de información en general (por ejemplo diagramas de flujo, mapas esquemáticos o toda clase de diagramas) (Kaufmann and Wagner, 2001).

La amplia variedad de familias de grafos ha hecho que los algoritmos de trazado desarrollados varíen según el tipo de ellos que permiten visualizar.

Existen algoritmos específicos para dibujar árboles, para grafos dirigidos acíclicos, para grafos planares, etc. Entre esta gran variedad existen algoritmos para trazado de grafos generales. Y entre ellos 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 (Aiello and Silveira, 2004; Walshaw, 2003).

Los métodos de trazado de grafos dirigidos por fuerzas son una familia de algoritmos que usan analogías físicas para dibujarlos. 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.

Una tercera característica que se puede agregar, si bien no forma parte de la definición estándar pero que se verifica en la mayoría de los casos, es que se aplican a grafos generales, es decir, los algoritmos no se basan en ninguna propiedad estructural del grafo (por ej., su planaridad), sino que se aplican a cualquier tipo de grafo (Aiello and Silveira, 2004).

En general este tipo de algoritmos tiene dos componentes principales:

- Un modelo físico del grafo que representa a los elementos del mismo (nodos y arcos) junto a los criterios estéticos que se desean obtener del dibujo.
- Un algoritmo para encontrar un equilibrio del sistema físico que se corresponderá, en principio,

con un trazado del grafo estéticamente agradable.

Los modelos físicos más comunes son los que consisten en 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 un equilibrio para este sistema de fuerzas, es decir, 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 (Eades, 1984) que se basa en reemplazar los nodos por anillos de acero y cada arco con un resorte para formar un sistema físico, como se ve en la Figura 2.

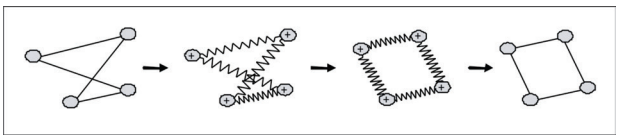

Figura 2 - Spring Embedder.

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.

Otra forma de definir el modelo del grafo, en lugar de usando un sistema de fuerzas, es a través de un sistema de energía. En este caso el propósito del algoritmo es encontrar una disposición para los vértices del grafo tal que la energía total del sistema sea mínima. Diferentes autores han optado por esta variante para el trazado de grafos, entre ellos Kamada y Kawai (Kamada and Kawai, 1989) y Davidson y Harel (Davidson and Harel, 1996).

Fruchterman y Reingold (Fruchterman and Reingold, 1991) proponen un método derivado, principalmente, del Spring Embedder de Eades, basado 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.

Las guías para este algoritmo son obtenidas de la física de partículas:

A una distancia de cerca de 1 fm (femtómetro) la fuerza nuclear es atractiva y equivalente a cerca de 10 veces la fuerza eléctrica entre 2 protones. Decrece rápidamente a medida que crece la distancia haciéndose completamente insignificante cuando se llega a cerca de 15 veces esta separación. Cuando dos núcleos están a una distancia de entre 0 y 4 fm uno de otro las fuerzas nucleares se convierten en repulsivas.

Si los nodos se comportan como partículas atómicas o cuerpos celestes ejerciendo fuerzas atractivas y repulsivas sobre los demás, las fuerzas inducen movimiento. El algoritmo de Fruchterman y Reingold se basa en simulaciones moleculares o planetarias. 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 sí, mientras todos los vértices se repelen unos a otros. Esto es consistente con la asimetría propuesta por los dos principios antes enunciados.

DESARROLLO

Fuerzas dirigidas en la visualización del LEL

Con el fin de detectar agrupamientos de símbolos se aplicó una modificación del algoritmo propuesto por Fruchterman y Reingold (Fruchterman and Reingold, 1991) en la visualización de los grafos correspondientes a los Léxicos de diferentes casos

de estudio. Para ello cada símbolo del LEL fue representado mediante un nodo y las menciones a otros símbolos incluidas en su definición se representaron como arcos dirigidos a los nodos respectivos.

```
area := W * L; { W y L representan el ancho y el
largo del marco }
G := (V , E); { las posiciones iniciales de los vértices
son determinadas al azar. }
k := √area/|V|;
function fa(z) := begin return z2/k end ;
function fr(z) := begin return k2/z end ;
for i := I to iterations do begin
    { ca v in V do begin cular fuerzas repulsivas }
    for
        { cada vértice tiene asociados dos vectores:
        .pos y .disp }
        v.disp := 0;
        for u in V do
            if (u ≠ v) then begin
            { Δ representa el vector diferencia
            entre las posiciones de los vértices}
            Δ := v.pos - u.pos;
            v.disp := v.disp + (Δ/|Δ|) *fr(|Δ|)
        end
    end
    { calcular fuerzas atractivas }
    for e in E do begin
        { cada arco es un par ordenado de vértices .v
        y .u }
        Δ := e.v.por - e.u.pos
        e.v.disp :=e.v.disp - (Δ/|Δ|) *fa(|Δ|);
        e.u.disp :=e.u.disp + (Δ/|Δ|) *fa(|Δ|)
    end
    { prevenir el desplazamiento fuera del amrco de
    trabajo }
    for v in V do begin
        v.pos :=v.pos + ( v. disp/ |v.disp|) *v.disp;
        v.pos.x := min(W/2, max(-W/2, v.pos.x));
        v.pos.y := min(L/2, max(-L/2, v.pos.y));
    end
end
```

Figura 3 - Algoritmo de visualización de grafos dirigido por fuerzas (Fruchterman and Reingold, 1991).

En la Figura 3 se presenta un pseudocódigo del algoritmo propuesto en (Fruchterman and Reingold, 1991).

Para la aplicación del algoritmo los nodos son ubicados al azar en el marco de trabajo y, posteriormente, se va modificando su ubicación en forma iterativa. Cada iteración tiene tres pasos:

- Calcular el efecto de las fuerzas atractivas sobre cada nodo.
- Calcular el efecto de las fuerzas repulsivas.
- Limitar el desplazamiento total.

LAS FUERZAS

fa y fr son las fuerzas de atracción y de repulsión, respectivamente. Como puede observarse en la Figura 3 las fuerzas propuestas en Fruchterman and Reingold, 1991 son:

$$fa(d) = \frac{d^2}{k} \quad (1)$$

$$fr(d) = \frac{k^2}{d} \quad (2)$$

$$k = C * \sqrt{\frac{Area}{NúmeroNodo \quad s}} \quad (3)$$

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

La aplicación de este sistema de fuerzas permitió detectar agrupamientos de símbolos para varios de los casos analizados. Sin embargo se experimentó con fuerzas diferentes con el fin de comprobar si se podía mejorar la visualización de los mismos.

En particular se trabajó con el par de fuerzas propuesto por Eades (Eades, 1984):

$$fa(d) = c1 * \log\left(\frac{d}{c2}\right) \quad (4)$$

$$fr(d) = c3 * \sqrt{d} \quad (5)$$

Las constantes c1, c2 y c3 se ajustaron experimentalmente de acuerdo a sus efectos sobre la visualización de los agrupamientos.

Con este sistema de fuerzas los agrupamientos que se habían detectado con las fuerzas originales se visualizaron de una forma mucho más clara. Por ello el algoritmo fue modificado considerando (4) y (5) como fuerzas de atracción y repulsión respectivamente (Ridao and Doorn, 2013).

EL MARCO DE TRABAJO

El grafo debe ser confinado al marco especificado por el usuario. El algoritmo propuesto por Fruchterman and Reingold, 1991 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. De este modo el marco se modela como cuatro paredes que contienen al grafo dentro de ellas.

Sin embargo las experiencias realizadas permitieron determinar que no era necesaria la ubicación de nodos en las paredes del marco de trabajo ya que, con la elección adecuada de las constantes que intervienen en el cálculo de las fuerzas, los nodos se mantienen dentro del marco definido.

RESULTADOS Y DISCUSIÓN

Visualización de los Léxicos de diferentes casos de estudio.

Con el fin de verificar si la estrategia propuesta permite detectar agrupamientos de símbolos se la aplicó a diferentes casos de estudio cuyos Léxicos habían sido verificados y validados previamente.

Caso 1: Sin clusters conocidos a priori.

Se aplicó la estrategia a un caso de estudio para el cual se prefirió no analizar en forma semántica la presencia o ausencia de agrupamientos de símbolos.

Los casos analizados fueron:

- Sistema de Planes de Ahorro Previo para la Adquisición de Vehículos 0Km (Rivero et al., 1998)

En la Figura 4 se presentan los resultados obtenidos luego de aplicar la estrategia al grafo correspondiente al LEL de este caso de estudio.

Cabe destacar que, tanto en ésta 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 el grafo presentado en la figura no se observan agrupamientos claramente diferenciados. El análisis semántico del Léxico correspondiente al caso de estudio confirma la ausencia de agrupamientos destacados validando el resultado arrojado por la estrategia.

Podría considerarse que los símbolos constituyen un solo cluster del cual deberían descartarse aquellos que claramente se alejan del agrupamiento central. En este caso los símbolos 24, 62, 65, 32 y 44 parecen mantener escasa relación con el resto.

Caso 2: con 2 clusters conocidos.

A continuación se aplicó el algoritmo a un caso de estudio para el cual se conocía de antemano la presencia de clusters.

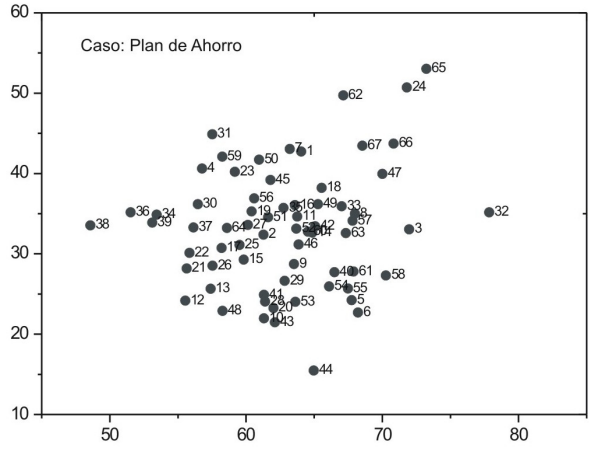

Figura 4 - Distribución de nodos para el caso 1.

El caso corresponde al LEL del proceso de Construcción de LEL y Escenarios (García and Gentile, 1999). 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 correspon-

dientes a la construcción de Escenarios.

En la figura 5 se presenta la distribución de los nodos obtenida con el algoritmo.

En la figura se observa que los símbolos correspondientes a la construcción del LEL tienden a agruparse hacia un lado, mientras los correspondientes a la construcción de escenarios, lo hacen al otro. En la zona central, se ubican aquellos símbolos relacionados tanto con el proceso de construcción del LEL como con el de Escenarios.

Por ejemplo, el nodo 11 corresponde al símbolo Componente dudas, considerado en la construcción de ambos modelos. En la misma zona, se observan otros símbolos que están asociados en igual medida con el modelo del LEL y el modelo de los Escenarios, pero además se agrupan entre sí. Por ejemplo, los símbolos 50, 75, 76, 77 y 78 están relacionados con la Vista de Hipertexto en los modelos de LEL y Escenarios.

Caso 3: con 2 clusters conocidos.

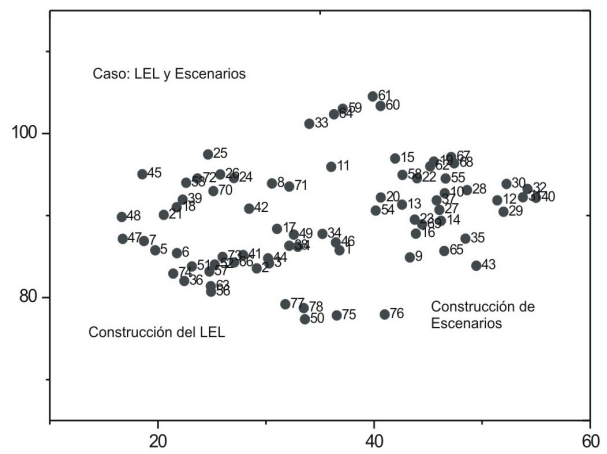

Figura 5 - Distribución de nodos para el caso 2.

El último caso de estudio al que se aplicó el algoritmo describe la 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 de la estrategia, se observa claramente la presencia de tres agrupamientos, como puede verse en la figura 6.

Con el fin de comprobar los límites de los clusters visualizados, se aplicó el clasificador K-Means (Peña et al., 1999) a los datos obtenidos, indicando la presencia de 3 grupos. Los agrupamientos detectados por el clasificador coinciden con los que se pueden observar a simple vista. En la figura 6 se ha rodeado con una elipse cada uno de los clusters identificados, y en la tabla 1, se presentan algunos de los símbolos pertenecientes a cada uno.

Revisando estos resultados se pudo confirmar la presencia de estos tres agrupamientos. En particular el tercer grupo, correspondiente a Calidad de la Papa, no había sido detectado en la lectura rutinaria, lo que confirma la capacidad de revelar información oculta por parte del método. Los

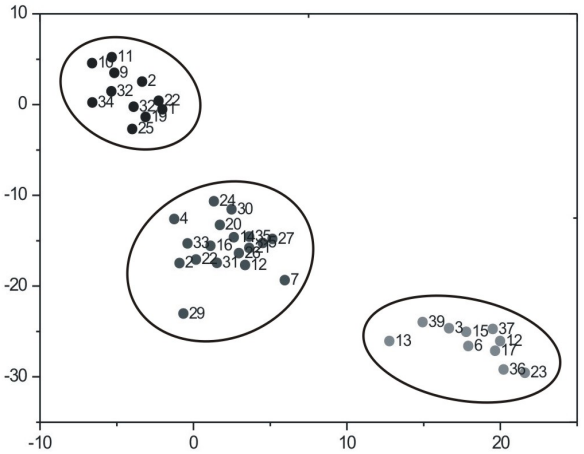

Figura 6 - Distribución de nodos para el caso 3.

símbolos incluidos en el agrupamiento calidad de la papa se habían considerado parte del cluster correspondiente a Producción y Entrega.

Como se indicó en la sección 4 la ubicación inicial de los nodos del grafo se hace al azar. Para todos los

casos estudiados se efectuaron múltiples pruebas con diferentes configuraciones iniciales produciéndose los mismos agrupamientos, o falta de ellos, que se puede observar en las Figuras 4, 5 y 6.

Se estudió también la posibilidad de ponderar los arcos entre símbolos según el número de referencias existentes entre ellos pero no se observaron efectos significativos en los agrupamientos detectados.

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

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

DETERMINACIÓN DEL NÚMERO DE CLUSTERS

Si se observa la forma geométrica de los grafos resultantes del análisis de los tres casos se notan algunas características que pueden ser consideradas para determinar el número de clusters detectados.

Para los casos 1 y 2 se aprecia cierta simetría tanto en el eje X como en el eje Y mientras que para el caso 3 esta simetría no está presente.

Se calcularon los coeficientes de asimetría en X y en Y para cada caso obteniendo los siguientes resultados:

Caso 1:

$$AS_x = 0,02 \quad (6)$$

$$AS_y = 0,14 \quad (7)$$

Sin embargo, si se descartan los nodos que se alejan del agrupamiento principal lo que hace suponer que no guardan una estrecha relación con el resto, los valores calculados son:

$$AS_x = -0,09 \quad (8)$$

$$AS_y = 0,04 \quad (9)$$

Caso 2:

$$AS_x = 0,05 \quad (10)$$

$$AS_y = 0,14 \quad (11)$$

Caso 3:

$$AS_x = 0,79 \quad (12)$$

$$AS_y = -0,48 \quad (13)$$

Se puede inferir que, cuando los coeficientes de asimetría se alejan de 0, es posible que se esté en presencia de un LEL con un número de agrupamientos mayor que dos. No se han definido los límites para los cuales es razonable ignorar la asimetría o considerar que la misma indica la presencia de tres o más clusters.

Por otra parte en el caso 2, donde las clases que se destacan son 2, se observa una forma elipsoidal en la distribución de los nodos donde los símbolos correspondientes a cada cluster se agrupan alrededor de cada uno de los focos de la elipse. El

cociente entre la variancia sobre el eje X y la variancia sobre el eie Y es:

$$CV = \frac{\sigma_X^2}{\sigma_Y^2} = 3{,}18 \qquad (14)$$

Para los casos 1 y 3, en cambio, no es posible asociar las formas resultantes con una elipse, como en el caso de LEL y Escenarios.

Entonces, para aquellos casos donde se observa cierta simetría en los ejes X e Y (coeficientes de asimetría cercanos a 0) y un cociente de varianzas mayor que 1 se puede inferir la presencia de dos clusters principales. No se ha estudiado aún cuál es el límite a partir del cual el valor de este cociente puede asegurar la presencia de dos agrupamientos.

Teniendo en cuenta los resultados analizados se proponen, como un primer enfoque para determinar el número de clusters en un LEL, las siguientes reglas:

- Si $AS_X \cong 0$ y $AS_Y \cong 0$ sin considerar el valor de CV se puede inferir que todos los símbolos constituyen un único cluster.

- Si $AS_X \cong 0$ y $AS_Y \cong 0$ y CV >1 se puede inferir la presencia de 2 clusters.

- Si $AS_X \neq 0$ y $AS_Y \neq 0$ sin importar el valor de CV se puede inferir la presencia de tres o más clusters.

Como se dijo anteriormente queda pendiente el análisis de más casos de estudio para definir los límites inferior y superior para CV y AS que permitan determinar más precisamente la presencia de un número determinado de clusters.

CONCLUSIONES

Se ha propuesto una estrategia que permite detectar agrupamientos de símbolos en un modelo del proceso de Requisitos: el Léxico Extendido del Lenguaje.

El análisis de los resultados obtenidos permite concluir que es posible efectuar una segunda lectura de algunos de los documentos de Ingeniería de Requisitos lo que brinda nueva información acerca del UdeD. Se comprobó que todos los clusters encontrados en los casos de estudio se corresponden con agrupamientos conceptuales observables en el mundo real.

El uso de grafos y el análisis de su estructura son herramientas apropiadas para el estudio de la información implícita empotrada en dichos documentos, en particular en el Léxico Extendido del Lenguaje (LEL).

Es posible determinar métricas de buena calidad que actúen como indicadores del número de clusters existentes.

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 en el caso de esta estrategia no es necesario.

En próximos trabajos se ampliará el número de dimensiones de la caja contenedora del grafo permitiendo construirlos en espacios de tres o más dimensiones. Se presume que esta forma de trabajar permitirá extraer aun más información del LEL y eventualmente poder encarar casos más complejos. Claramente esta ampliación de la cantidad de dimensiones obligará a utilizar técnicas de visualización más elaboradas.

Se ha considerado la posibilidad de ponderar los arcos entre símbolos según el número de referencias existentes entre ellos pero las experiencias realizadas hasta el momento no arrojaron diferencias significativas en los agrupamientos detectados. Se proyecta estudiar diferentes alternativas de ponderación que pudieran incidir en los resultados.

Con el fin de comenzar a analizar las posibles aplicaciones de la información recuperada mediante el análisis estructural de los documentos en etapas posteriores de la Ingeniería de Requisitos se planea comparar la técnica presentada en este artículo con otras de clusterización, como por ejemplo la propuesta por Duan (Duan et al., 2009).

REFERENCIAS

Barabasi, "Linked: The New Science of Networks", Perseus Publishing, USA, (2002).

Dorogovtsev and Mendes, "Evolution of networks: From biological nets to the Internet and WWW", Oxford University Press, UK, (2003).

Sarmah, Kalita and Bhattacharyya, "A pattern matching approach for clustering gene expression data", Int. Journal Data Mining, Modelling and Management, 3(2), 130-149, (2011).

Mo, Cao and Wang, "Occurrence-Based Fingerprint Clustering for Fast Pattern-Matching Location Determination", IEEE Communications Letters, 16(12), 2012 – 2015, (2012).

Zimmermann, Ntoutsi, Siddiqui, Spiliopoulou and Kriegel, H.P. "Discovering Global and Local Bursts in a Stream of News", Proceedings 27th Annual ACM Symposium on Applied Computing. (SAC '12), Italy, 807-812, (2012).

Duan, Laurent, Cleland-Huang and Kwiatkowski, "Towards automated requirements prioritization and triage", Requirements Engineering Journal, 14(2), 73-89, (2009).

Leite and Franco,"O Uso de Hipertexto na Elicitaçao de Linguagens da Aplicaçao", Proceedings IV Simpósio Brasilero de Engenharia de Software, SBC, Brazil, 134-149, (1990).

Leite, Doorn, Kaplan, Hadad and Ridao, "Defining System Context Using Scenarios", Perspectives on Software Requirements, Kluwer Academic Press, 169-199, (2004).

Artz, "Data Driven vs. Metric Driven Warehouse Design" Encyclopedia of Data Warehousing and Mining, IGI Global, 382-387, (2009).

Prieto Díaz and Arango, "Domain Analysis and Software Systems Modelling", IEEE Computer Society, (1991).

Schäfer, Prieto and Matsumoto, "Software Reusability", Ellis Horwood Ltd., (1993).

Ben Achour, Rolland, Maiden and Souveyet, "Guiding Use Case Authoring: Results of an Empirical Study", Proceedings International Symposium on Requirements Engineering, Limerick-Ireland, June 1999, 36-43, (1999).

Rolland and Ben Achour, C., "Guiding the construction of textual use case specifications", Data & Knowledge Engineering, 25, 125-160, (1998).

Oberg, Probasco and Ericsson, "Applying Requirements Management with Use Cases", Rational Software Corporation, Australia, (1998).

Regnell, "Requirements Engineering with Use Cases – a Basis for Software Development", Thesis (Ph.D). Department of Communication Systems, Lund University, (1999).

Prakash, Aurum and Kox, "Requirements Engineering Practice in Pharmaceutical and Healthcare Manufacturing", Proceedings 11th Asia-Pacific S.E. Conference, 402-409, (2004).

Leite and Franco, "A Strategy for Conceptual Model Acquisition", Proceedings IEEE International Symposium on RE, 243-246, (1993).

Leite, Rossi, Balaguer, Maiorana, Kaplan, Hadad and Oliveros, "Enhancing a Requirements Baseline with Scenarios", Requirements Engineeering Journal, 2(4), 184-198, (1997).

Gross and Yellen, Editors, "Handbook of Graph Theory", CRC Press, USA, (2003).

Gross and Yellen, Editors. Graph Theory and Its Applications, Second Edition (Discrete Mathematics and Its Applications), Chapman & Hall/CRC, USA, (2006).

Brandes, Kenis and Wagner, "Communicating centrality in policy network drawing", IEEE Transactions on visualization and computer graphics, 9(2), 241-253, (2003).

Di Battista, Eades, Tamassia and Tollis, "Graph Drawing: Algorithms for the Visualization of Graphs", Prentice Hall, (1999).

Fruchterman and Reingold, "Graph Drawing by Forcedirected Placement", Software-Practice and Experience, 21(11), 1129-1164, (1991).

Kaufmann and Wagner (eds.), "Drawing graphs: methods and models", LNCS, vol 2025, Springer-Verlag, Germany, (2001).

Aiello and Silveira, "Trazado de grafos mediante métodos dirigidos por fuerzas: revisión del estado del arte", Tesis (Licenciatura en Ciencias de la Computación), Departamento de Computación, Facultad de Ciencias Exactas y Naturales, UBA, (2004).

Walshaw, "A multilevel algorithm for force-directed graphdrawing", Journal of Graph Algorithms and Applications, 7(3), 253-285, (2003).

Eades, "A heuristic for graph drawing", Proceedings Congressus Numerantium, 42, 149–160, (1984).

Kamada and Kawai, "An algorithm for drawing general undirected graphs", Information Processing Letters 31, 7-15. (1989).

Davidson and Harel, "Drawing graphs nicely using simulated annealing", ACM Transactions on Graphics, 15(4), 301-331. (1996).

Ridao and Doorn, "Semántica Oculta en Modelos de Requisitos", XV Workshop de Investigadores en Ciencias de la Computación-WICC, Paraná, Argentina, Abril 2013, 471-475, (2013).

Rivero, Doorn, del Fresno, Mauco, Ridao and Leonardi, "Una Estrategia de Análisis Orientada a Objetos basada en Escenarios: Aplicación en un Caso Real", Proceedings WER'98 - Workshop en Engenharia de Requisitos, Maringá, Brasil, 1998, 79-90, (1998).

Garcia and Gentile, "Diseño de una herramienta para contrucción de LEL y Escenarios", Tesis (Ingeniería en Sistemas) Universidad Nacional del Centro de la Pcia. de Bs. As., (1999).

Peña, Lozano and Larrañaga, "An empirical comparison of four initialization methods for the K-Means algorithm", Pattern Recognition Letters, 20, 1027-1040, (1999).

Fecha de Presentación: 15/05/15
Fecha de Aprobación: 08/06/15