Construcción de grafos de glosarios guiada por el estilo del discurso

Marcela N. Ridao[1], Jorge H. Doorn[2,3], Gladys N. Kaplan[4]

[1] Facultad de Ciencias Exactas, Universidad Nacional del Centro de la Provincia de Buenos Aires
[2] Escuela de Informática, Universidad Nacional del Oeste
[3] Departamento de Ingeniería, Universidad Nacional de Tres de Febrero
[4] Departamento de Ingeniería e Investigaciones Tecnológicas, Universidad Nacional de La Matanza
mridao@exa.unicen.edu.ar, jdoorn@uno.edu.ar, gkaplan@unlam.edu.ar

RESUMEN

La visualización de grafos es un recurso cada vez más utilizado en diferentes dominios de las ciencias de la computación. En la Ingeniería de Requisitos (IR), esta estrategia permite detectar núcleos semánticos específicos en algunos de los modelos utilizados, en forma más eficaz y eficiente que el mero estudio de los documentos. El uso concreto de grafos en glosarios del Universo del Discurso (UdeD)[1] ha permitido detectar una influencia del estilo de la narrativa de la fuente de información, en los grafos resultantes. Aparentemente, si se ignora esa influencia, la misma se transforma en un factor perturbador que deteriora la calidad de los grafos obtenidos. En el presente proyecto se planifica aprovechar el conocimiento del estilo de la narrativa de la fuente de información, para construir grafos que permitan visualizar más eficazmente la información contenida en el glosario.

Palabras clave: Ingeniería de Requisitos. Complejidad estructural, Grafos, Ubicación de nodos, Estilo del discurso

CONTEXTO

La propuesta que se presenta es este trabajo forma parte del proyecto de investigación "Desarrollo y Aplicación de Técnicas de Extracción de Información en Data Science" de la Universidad Nacional del Centro de la Provincia de Buenos Aires (UNICEN), del proyecto de investigación "Incorporación Pragmática de Visiones Lingüístico-cognitivas en el Proceso de Requisitos" de la Universidad Nacional del Oeste (UNO) y del proyecto de investigación "Aspectos no funcionales en los procesos de requisitos" de la Universidad Nacional de La Matanza (UNLaM).

1. INTRODUCCIÓN

Las disciplinas dedicadas al estudio de fenómenos donde el aspecto dominante es la complejidad estructural y no la complejidad esencial de los elementos involucrados, han recurrido crecientemente al uso de grafos para visualizar, y de esa manera estudiar esta complejidad estructural [3] [5]. Existen numerosos ejemplos, en diversas áreas de la ciencia, donde la detección de agrupamientos y la relación entre los mismos, representan una contribución significativa a la mejor comprensión del fenómeno estudiado [19] [21] [22]. En particular, la detección de agrupamientos es utilizada con diversos fines en la Ingeniería de Requisitos. Por ejemplo, en [6] los requisitos se agrupan en clusters, con el objetivo de priorizarlos.

Algunos de los modelos de la Ingeniería de Requisitos pueden ser estudiados desde el punto de vista estructural. Este trabajo se basa en un proceso de IR, que utiliza modelos construidos en Lenguaje Natural. Estos modelos son: el Léxico Extendido del Lenguaje (LEL), los Escenarios Actuales y Futuros (EA y EF), el LEL de Requisitos (LELr), y la Especificación de Requisitos (ERS). Existen

[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. Es la realidad acotada por el conjunto de objetivos establecidos por quienes demandan una solución de software [14].

vínculos explícitos e implícitos dentro de los modelos y entre ellos [11]. Por esto, cada modelo podría ser visto como un grafo, donde los nodos son los elementos del modelo, y los arcos los vínculos entre ellos. Más aún, el conjunto formado por dos o más modelos podría ser representado mediante un único grafo, considerando las conexiones entre los elementos de los modelos como arcos.

Dentro de los modelos mencionados, el que se ha estudiado desde este punto de vista, hasta el momento, es el LEL [15]. Este modelo registra el vocabulario del UdeD, mediante la descripción de los términos utilizados por el cliente-usuario, postergando la comprensión del problema. En la descripción de los símbolos se debe maximizar el uso de otros símbolos del léxico. De esta manera, el conjunto de símbolos determina una red, que puede ser navegada para conocer todo el vocabulario del dominio.

Si se observa un LEL bajo la óptica estructural se puede construir un grafo donde los símbolos son los nodos y las menciones a otros símbolos, arcos dirigidos. Desde este punto de vista, el LEL puede visualizarse como una suerte de red lingüistica con una estructura compleja. Es así que, además de la información explícita almacenada en cada uno de los nodos, existe una información implícita empotrada en la estructura de las relaciones entre ellos. Epistemológicamente, este enfoque es muy similar al utilizado en minería de datos [2], en el sentido que se hace visible información oculta mediante el uso de una técnica notoriamente diferente a la utilizada comúnmente para la lectura del modelo.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

Las fuentes de información que se utilizan para construir el LEL tienen un estilo narrativo propio. Ese estilo es parte de la cultura organizacional y del propio proceso del negocio. Es así que el LEL resultante depende en cierta medida de ese estilo narrativo y todo estudio que se realice sobre el modelo debería considerarlo como un factor relevante.

Estilo de la Narrativa

Aun con una observación superficial, es posible percibir que el estilo del discurso de la información adquirida en diferentes casos y de diferentes orígenes, suele tener diferencias relevantes. Estas diferencias dependen de la fuente de información en si misma, pero también de la propia estructura del UdeD.

Sin ser los únicos, los principales estilos son: orientado a los procesos, orientado a las entidades y sus relaciones, orientado a los productos y orientado a los actores.

Obviamente, la importancia de los diferentes símbolos del LEL y de sus relaciones depende del estilo del discurso y por lo tanto el grafo que se construya debe procurar destacar los símbolos de mayor importancia. Por ejemplo, en un sistema de apoyo a la gestión de pacientes agudos, la narrativa estará centrada en los protocolos o procesos, y por lo tanto, la construcción de los grafos deberá centrarse en la visualización lo más clara posible de los verbos.

Trabajos previos [16] [17] [18] han permitido detectar la presencia de agrupamientos en el grafo construido a partir del LEL, mediante técnicas automáticas de trazado de grafos. Sin embargo, los agrupamientos detectados no muestran las relaciones entre actores, procesos u objetos. Se propone, entonces, modificar el algoritmo de detección de agrupamientos, poniendo énfasis en dichas relaciones.

Trazado de Grafos: Métodos dirigidos por fuerzas

La Teoría de Grafos tiene diversidad de aplicaciones, y es usada en la representación de circuitos eléctricos, carreteras, moléculas orgánicas, ecosistemas, relaciones sociológicas, y muchas áreas de las ciencias de la computación [4] [9] [10]. El Trazado de Grafos, como una rama de la Teoría de Grafos, aplica topología y geometría para derivar representaciones de grafos en dos dimensiones.

La generación automática del trazado de grafos tiene importantes aplicaciones en muchas de las ciencias antes mencionadas, y tiene gran impacto en la visualización de información en general (por ejemplo diagramas de flujo, mapas esquemáticos o toda clase de diagramas) [12].

La amplia variedad de familias de grafos ha hecho que los algoritmos de trazado desarrollados varíen según el tipo de grafos que permiten visualizar. Entre ellos, existen algoritmos para trazado de grafos generales, destacándose una familia de métodos conocidos como "dirigidos por fuerzas". Estos métodos 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 [1] [13] [20].

Los métodos de trazado de grafos dirigidos por fuerzas modelan al grafo como un sistema físico, y el trazado es obtenido buscando el equilibrio del mismo. Los modelos físicos más comunes son los que consisten en un sistema de fuerzas que actúan entre los vértices del grafo. El objetivo 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 [7], 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 una disposición inicial arbitraria, y se dejan actuar las fuerzas de los resortes hasta lograr un estado de equilibrio.

Fruchterman y Reingold [8] proponen un método derivado principalmente del Spring Embedder, basado en la física de partículas. Los nodos ejercen fuerzas de atracción y de rechazo sobre los demás, que inducen movimiento. El método propone que sólo los nodos conectados se atraigan entre sí, mientras todos los vértices se repelan unos a otros.

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 [8] 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.

En el algoritmo, los nodos son ubicados al azar en el marco de trabajo, y posteriormente se va modificando su ubicación en forma iterativa.

- Se calcula el efecto de las fuerzas de atracción sobre cada nodo.
- Se calcula el efecto de las fuerzas de rechazo.
- Se calculan las nuevas posiciones de los nodos.

La nueva posición se calcula desplazando los nodos una distancia proporcional a la fuerza neta resultante sobre cada uno de ellos.

En trabajos anteriores [16] [17] [18], se utilizaron tres conjuntos de fórmulas diferentes para fa (fuerza de atracción) y fr (fuerza de rechazo). En la Tabla 1 se presentan las fórmulas utilizadas en cada caso.

Fruchterman - Reingold [8]	Eades [7]	Ridao - Doorn [18]
$fa(d) = \frac{d^2}{k}$ (1) $fr(d) = \frac{k^2}{d}$ d: distancia entre los vértices k: radio vacío alrededor de un nodo.	$fa(d) = c1 * \log\left(\frac{d}{c2}\right)$ $fr(d) = c3 * \sqrt{d}$ (2) d: distancia entre los vértices c1, c2, c3: ajustadas experimentalmente, de acuerdo a sus efectos sobre la visualización de los agrupamientos.	$fa(d) = c1 * \log\left(\frac{d}{c2}\right)$ $fr(d) = d^{\alpha}$ (3) d: distancia entre los vértices c1, c2: ajustadas experimentalmente α: parámetro que permite modificar la relación entre fr y fa.

Tabla 1. Fórmulas utilizadas en [8], [7] y [18]

La aplicación del primer sistema de fuerzas permitió detectar agrupamientos de símbolos para algunos de los casos analizados. Con las fórmulas propuestas por Eades [7], los clusters que se habían detectado con (1) se visualizaron de una forma mucho más clara, y se comenzaron a detectar grupos para casos de

estudio donde las pruebas con el primer sistema de fuerzas no arrojaban resultados claros.

Con el fin de obtener mejores representaciones, se propusieron variaciones a las fórmulas originales [18]. Para ello, se consideró la relación entre las magnitudes de fr y fa. La idea central fue hacer crecer más lentamente la fuerza de rechazo con la distancia, suponiendo que esto traería como consecuencia una mayor cercanía entre los nodos de un cluster. Con estas fórmulas, y una adecuada relación entre las magnitudes de las fuerzas, la visualización de los agrupamientos se hizo mucho más clara.

En las figuras 1 y 2 se observa la evolución en la visualización de los agrupamientos para dos de los casos de estudio analizados.

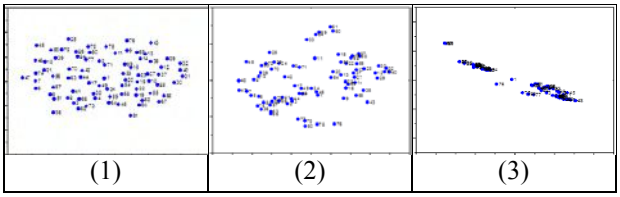

Fig. 1. Distribución de nodos del caso LEL y Escenarios, luego de aplicar los tres conjuntos de fórmulas

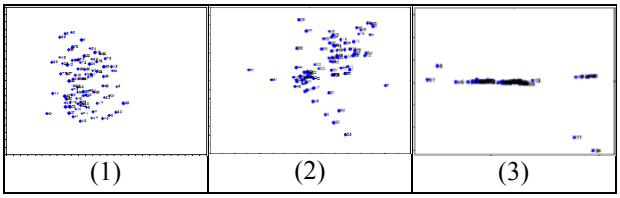

Fig. 2. Distribución de nodos del caso Biblioteca, luego de aplicar los tres conjuntos de fórmulas

En casos donde se observaba la presencia de grupos, pero con límites poco claros, la aplicación de las fuerzas (3) permitió visualizarlos con claridad. En otros, donde no se observaban agrupamientos, se pudo comprobar que estos existían.

La disposición de los grupos en el grafo resultante permite, además, mostrar la relación o falta de ella, entre los agrupamientos. Es así que, para algunos casos de estudio, se observó una alineación entre los clusters, formando una suerte de cadena que denota la ausencia de relación entre los grupos de los extremos, y una relación estrecha entre los grupos vecinos.

3. RESULTADOS ESPERADOS

Se espera mejorar la visualización del grafo asociado al LEL, incorporando la información del estilo de la narrativa como parámetro. Es así que, en forma preliminar se proponen dos tipos de grafos: a) orientados a sujetos u objetos y b) orientados a verbos.

Orientados a sujetos u objetos

A continuación, se presentan sólo las reglas que se proponen para visualización centrada en sujetos. Meramente intercambiando sujetos y objetos, se pueden obtener las reglas para la visualización centrada en objetos.

Regla 1: Ignorar los símbolos verbos.

Regla 2: Si existe una referencia de un sujeto a otro sujeto, partiendo desde la noción, incrementar por un factor M_1 la fuerza de atracción.

Regla 3: En toda otra relación entre sujetos, incrementar la fuerza de rechazo por un factor M_2.

Regla 4: Si existe una referencia de un objeto a otro objeto, partiendo desde la noción, incrementar por un factor M_3 la fuerza de atracción.

Regla 5: Si existe una referencia de un estado a un sujeto, partiendo de la noción, incrementar la fuerza de atracción por un factor M_4.

Regla 6: En toda otra relación entre estados y sujetos, incrementar la fuerza de rechazo por un factor M_5.

Regla 7: Si existe una referencia de un estado a un objeto, partiendo de la noción, incrementar la fuerza de atracción por un factor M_6.

Regla 8: En toda otra relación entre estados y objetos, incrementar la fuerza de rechazo por un factor M_7.

Regla 9: Si existe una referencia transitiva entre sujeto y objeto o viceversa, a través de un verbo considerarla como una referencia directa.

Orientados a verbos

Regla 1: Si existe una referencia de un sujeto a otro sujeto, partiendo desde la noción, incrementar por un factor M1 la fuerza de atracción.

Regla 2: Si existe una referencia de un objeto a otro objeto, partiendo desde la noción, incrementar por un factor M2 la fuerza de atracción.

Regla 3: Si existe una referencia de un verbo a otro verbo, partiendo desde la noción, incrementar por un factor M3 la fuerza de atracción.

Regla 4: En toda otra relación entre verbos, incrementar la fuerza de rechazo en un factor M4.

Se planifica determinar los mejores valores para los factores M_i. Las primeras pruebas se realizarán utilizando el mismo valor para todos los factores indicados.

4. FORMACIÓN DE RECURSOS HUMANOS

El presente proyecto es parte directa de la tesis doctoral de la Mag. Marcela Ridao y contribuye a la tesis doctoral de la Mag. Gladys N. Kaplan.

5. BIBLIOGRAFÍA

[1] 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. Departamento de Computación. Facultad de Ciencias Exactas y Naturales. UBA. (2004)

[2] Artz, J.M. "Data Driven vs. Metric Driven Warehouse Design" en John Wang "Encyclopedia of Data Warehousing and Mining, ISBN 1-59140-557-2, pag 223 a 227, tomo I, Idea group reference. (2006)

[3] Barabasi, A.: Linked, The New Science of Network. Perseus publishing. (2002)

[4] Brandes, U., Kenis, P., Wagner, D.: Communicating centrality in policy network drawing. IEEE Trans. on visualization and computer graphics, 9(2), 241-253. (2003)

[5] Dorogovtsev, S., Mendes, J.: Evolution of networks: From biological nets to the Internet and WWW. Oxford University Press, Oxford. (2003)

[6] Duan, C., Laurent, P., Cleland-Huang, J., y Kwiatkowski, C., Towards automated requirements prioritization and triage, Req. Engineering Journal, 14(2), 73-89. (2009)

[7] Eades, P.: A heuristic for graph drawing. Congressus Numerantium 42, 149–160. (1984)

[8] Fruchterman, T., Reingold, E.: Graph Drawing by Force-directed Placement. Software-Practice and Experience 21(11), 1129-1164. (1991)

[9] Gross, J., Yellen, J.: Editors. Handbook of Graph Theory. CRC Press. (2003)

[10] Gross, J., Yellen, J.: Editors. Graph Theory and Its Applications, Second Edition (Discrete Mathematics and Its Applications). Chapman& Hall/CRC. (2006)

[11] Kaplan, G., Doorn, J., Gigante, N. Evolución semántica de glosarios en los procesos de requisitos. XIX Congreso Argentino de Ciencias de la Computación, CACIC. Mar del Plata, Argentina. (2013)

[12] Kaufmann, M., Wagner, D.: (eds.) Drawing graphs: methods and models, LNCS, vol 2025. Springer-Verlag. (2001)

[13] Kobourov, Stephen G.: Spring Embedders and Force-Directed Graph Drawing Algorithms, arXiv:1201.3011. (2012)

[14] 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)

[15] Leite, J., Franco, A.: O Uso de Hipertexto na Elicitaçao de Linguagens da Aplicaçao. IV Simpósio Brasilero de Engenharia de Software, SBC, pp. 134-149. Brazil. (1990)

[16] Ridao, M., Doorn, J.: Semántica Oculta en Modelos de Requisitos. XV Workshop de Investigadores en Ciencias de la Computación, WICC. Paraná, Argentina. (2013)

[17] Ridao, M., Doorn, J.: Agrupamientos en Glosarios del Universo de Discurso. Tecnología y Ciencia - Revista de la Universidad Tecnológica Nacional - Edición Especial: CoNaIISI 2014, 13(27), 5-16. (2015)

[18] Ridao, M., Doorn, J., Visualización de Núcleos Semánticos en Glosarios del Universo de Discurso. IV Congreso Nacional de Ingeniería en Informática/sistemas de Información, CoNaIISI. Salta, Argentina (2016)

[19] Rosy Das, Jugal Kalita, Dhruba K. Bhattacharyya, A pattern matching approach for clustering gene expression data, Int. J. Data Mining, Modelling and Management, 3(2), 130-149 (2011)

[20] Walshaw, C.: A multilevel algorithm for force-directed graph-drawing. Journal of Graph Algorithms and Applications 7(3), 253-285 (2003)

[21] Yijun Mo, Zuo Cao, Bang Wang, Occurrence-Based Fingerprint Clustering for Fast Pattern-Matching Location Determination. Communications Letters, IEEE 16(12), 2012 - 2015 (2012)

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