ESTUDIO DE INTEGRACIÓN DE MÉTODOS DE DESCUBRIMIENTO DE CONOCIMIENTO EN WEB

Hernán Merlino, Eduardo Diez, Juan Manuel Rodríguez, Santiago Bianco, Ramón García-Martínez

Laboratorio de Investigación y Desarrollo en Arquitecturas Complejas
Grupo Investigación en Sistemas de Información
Departamento de Desarrollo Productivo y Tecnológico. Universidad Nacional de Lanús
29 de Septiembre 3901 (1826) Remedios de Escalada, Lanús. Argentina. Tel +54 11 5533 5600 Ext. 5194
hmerlino@gmail.com, rgm1960@yahoo.com

RESUMEN

La extracción de conocimiento a partir de fuentes heterogéneas de información embebida y en volúmenes de datos demasiado grandes, como podría ser la Web, tradicionalmente ha requerido de participación humana en la forma de reglas de extracción o bien de ejemplos de entrenamiento etiquetados de forma manual. Desde hace mas de una becada se han desarrollado una conjunto de algoritmos como OIE, TEXT RUNNER, WOE-parse, WOE-pos, SRL-Lund, SRLUIUC, ReVerb, TRIPLEX. OLLIE, entre otros, dedicados a la tarea de extracción de conocimiento. En este contexto, este proyecto busca desarrollar un proceso capaz de integrar diversos algoritmos de extracción de conocimiento de forma inteligente, que dado una estructura de información inicial como entrada, que contienen conocimiento embebido, genere un conjunto de piezas de conocimiento (reglas de producción, subgrafos de una red semántica, entre otros).

Palabras clave: Descubrimiento de Conocimiento, Extracción de Conocimiento, Algoritmos de Extracción de Conocimiento.

CONTEXTO

El proyecto: [a] inicia una línea de trabajo en el campo de extracción de conocimiento a partir de fuentes heterogéneas de información embebida dentro del campo de la Informática; [b] articula la Línea de Investigación Prioritaria “3. Desarrollos Informáticos” del Instituto de Economía, Producción y Trabajo, aprobada por Resolución Consejo Superior UNLa N° 113/14, promoviendo la mejora de los sistemas productivos, en particular los utilizados en la producción de tecnologías especiales; y [c] responde a los lineamientos estratégicos de la CADENA DE VALOR DEL SOFTWARE Y SERVICIOS INFORMÁTICOS que establece el Plan Industrial 2020 del Ministerio Industria [MI, 2014], promoviendo la mejora del sistema productivo de la industria del Software con foco en sistemas software especiales.

INTRODUCCIÓN

El desafío de la extracción de conocimientos comienza a fines de la década de 1970 como es señalado en [Cowie & Lehnert, 1996]. Más tarde en los años 90 la investigación fue alentada y financiada por la Agencia de Proyectos Avanzados de Defensa (DARPA) [Konstantinova, 2014].

Los métodos de extracción de conocimiento comenzaron trabajando en la detección y clasificación de nombres propios, utilizando como entrada fuentes de conocimiento embebido, como son las fuentes de información no estructurada, este tipo de extracción de conocimiento es llamado Reconocimiento de Nombres de Entidades (NER según sus siglas en inglés). En general estos sistemas de extracción de conocimiento buscan nombres de personas, compañías, organizaciones y lugares geográficos [Konstantinova, 2014]. El siguiente paso que dieron los métodos de extracción de conocimiento fue el de resolver correferencias y el de extraer relaciones entre nombres de entidades [Jurafsky & Martin , 2009].

Hacía fines de la década de 2000 los métodos de extracción de conocimiento se habían diversificado y especializado. En [Jurafsky & Martin , 2009] se reconocen distintos tipos de piezas de conocimiento susceptibles de ser extraídas: nombres de entidades, expresiones temporales, valores numéricos, relaciones entre entidades y expresiones previamente identificadas, eventos , entre otras.

La extracción de conocimiento tradicionalmente ha requerido de participación humana en la forma de reglas de extracción o bien de ejemplos de entrenamiento etiquetados de forma manual. En particular para los casos de extracción de relaciones entre entidades, es el usuario quien debe explícitamente, especificar cada relación que le interese, tarea ardua, sobre todo cuando se trabaja con fuentes de información embebida heterogéneas y con volúmenes de datos demasiado grandes, como

podría ser la Web. Debido a ello en general los sistemas de extracción de conocimiento fueron utilizados sobre fuentes de información embebida más bien pequeñas y homogéneas [Banko et al., 2007].

En el año 2007 Michele Banko introduce un nuevo concepto en materia de extracción de conocimiento, al que llama en inglés: Open IE (OIE). Se trata de un paradigma de extracción de conocimiento en donde un sistema informático realiza una sola pasada sobre el total de las fuentes de conocimiento embebido dadas como entrada y extrae un gran conjunto de tuplas relacionales sin requerir ningún tipo de participación humana. En el mismo trabajo Banko presenta un método llamado TEXT RUNNER, el cual es el primer método que trabaja dentro de este nuevo paradigma [Banko et al., 2007].

A partir de este trabajo se propusieron otros métodos de extracción de conocimiento bajo el paradigma que Banko llamó Open IE. En [Wu & Weld, 2010] se propusieron dos métodos WOE-parse y WOE-pos, el primero WOE-parse utiliza un enfoque ligeramente distinto, utiliza un árbol de dependencias, realizando un análisis sintáctico en cada oración, para extraer las relaciones. Y si bien logra un mayor número de extracciones que TEXT RUNNER (1.42 tuplas por oración frente a 0.75) es 30 veces más lento que su predecesor. WOE-pos por el contrario es igual de rápido que TEXT RUNNER y ligeramente mejor (1.05 tuplas extraídas por oración). Si bien WOE-parse y WOE-pos, son métodos de propósito general su base de entrenamiento fue Wikipedia.

En [Mesquita et al., 2010] se presentó un método de extracción de conocimiento llamado SONEX pensado para extraer relaciones de redes sociales y de la blogosfera.

En [Christensen et al., 2011] se presentó un nuevo enfoque bajo las mismas consignas de extracción de conocimiento planteadas por Banko, se buscó utilizar la técnica de etiquetamiento secuencial, basado en la función semántica (en inglés Semantic Role Labeling) para la extracción de relaciones entre entidades. Se crearon don métodos nuevos SRL-Lund y SRL-UIUC, se los comparó con TEXT RUNNER en dos conjuntos de fuentes de conocimiento embebido, uno pequeño y otro grande. Ambos demostraron ser más precisos que TEXT RUNNER, SRL-Lund obtuvo una precisión de 0.7 y una medida F1 de 0.59 , SRL-UIUC obtuvo una precisión de 0.63 y una medida F1 de 0.68 mientras que TEXT RUNNER obtuvo una precisión de 0.55 y una medida F1 de 0.35. Sin embargo, al trabajar con el conjunto más grande de datos de entrada, TEXT RUNNER demostró tener una ventaja adicional, era 20 veces más rápido que SRL-LUND y 500 veces más rápido que SRL-UIUC.

Ese mismo año, en [Fader et al., 2011] se propone un nuevo método de extracción de conocimiento que logra un área bajo la curva ROC mayor que WOE-parse, WOE-pos y que TEXT RUNNER, se trata de ReVerb. ReVerb fue puesto a prueba utilizando un conjunto de conocimiento embebido que constaba de 500 millones de sentencias web, demostró ser más rápido incluso que TEXT RUNNER. En un subconjunto de 100 000 sentencias se obtuvieron los siguientes tiempos: WOE-parse tardó 11 horas, WOE-pos y TEXT RUNNER tardaron 21 minutos cada uno y ReVerb 16 minutos. La mejora introducida por Fader constó en agregar restricciones a TEXT RUNNER y centrarlo en la extracción de relaciones basadas en verbos.

La restricción de ReVerb no le permite encontrar relaciones basadas en otro tipo de palabras, de categorías gramaticales, que no sean verbos, es por eso que algunos autores han propuesto diversos métodos para extraer relaciones basadas en otro tipo de categorías gramaticales, en particular sustantivos, es el caso de [Schmitz & Soderland, 2012] con OLLIE, de [Yahya et al., 2014] con ReNoun y el de [Mirrezaei et al., 2015] con TRIPLEX. OLLIE fue planteado directamente como una mejora a ReVerb, siendo su objetivo encontrar relaciones basadas no solo en verbos, sino también en sustantivos y adjetivos. Además plantea la posibilidad de hacer un análisis del contexto para encontrar relaciones no explicitas. OLLIE logró obtener 2.7 veces más área sobre la curva ROC que ReVerb y 1.9 veces más área bajo la curva ROC que WOE-parse, además OLLIE encontró 4.4 veces más extracciones correctas que ReVerb y 4.8 veces más que WOE-parse [Schmitz & Soderland, 2012]. El enfoque de TRIPLEX es ligeramente distinto ya que funciona como un complemento a ReVerb o a OLLIE, en el estudio realizado en [Mirrezaei et al., 2015], TRIPLEX por sí solo no logra superar a OLLIE (se comparó utilizando la medida F1 en este caso) y es el uso conjunto de OLLIE más TRIPLEX el que arroja un mejor resultado, aunque no muy lejano al que arroja OLLIE solo.

Para concluir en el trabajo de [Del Corro & Gemulla, 2013] se presenta un nuevo método de extracción de conocimiento llamado ClauseIE (respetando el paradigma propuesto por Banko); en dicho trabajo ClauseIE es comparado contra ReVerb, OLLIE, TEXT RUNNER y WOE utilizando distintas fuentes de conocimiento embebido: 500 oraciones extraídas del conjunto de

datos de prueba utilizado con ReVerb en [Fader et al., 2011], 200 oraciones aleatorias extraídas de Wikipedia y 200 oraciones aleatorias extraídas del New York Times. El resultado en todos los casos fue favorable a ClauseIE quien obtuvo una mejor precisión que los demás métodos.

CONVENCIONES, PREGUNTAS PROBLEMA, HIPÓTESIS Y OBJETIVOS GENERALES Y ESPECÍFICOS

Convenciones

Término: "pieza de conocimiento"

Estructura de información a la que se le puede asignar un significado y que tiene la propiedad de ser automáticamente manipulable en procesos de razonamiento automático. Podría ser una regla de producción o un subgrafo en una red semántica [García-Martínez & Britos, 2004; Gómez et al., 1997]. Dicha estructura representa por lo general relaciones entre conceptos.

Término: "conocimiento embebido"

Conocimiento disponible en estructuras de información entendible por el ser humano pero no por un sistema informático [García-Martínez & Britos, 2004; Gómez et al., 1997]. Un ejemplo de esta clase de conocimiento es el que se encuentra embebido en expresiones del lenguaje natural.

Término: "extracción de conocimiento"

Es el proceso que hace explicito el conocimiento embebido en una estructura de información. Este proceso no requiere especificar las relaciones dentro de las piezas de conocimiento, eventualmente descubre relaciones mientras hace una sola pasada por las estructuras de información que tienen el conocimiento embebido [Banko et al., 2007]. Si se piensa a la extracción de conocimiento como una transformación algebraica podría plantearse:

extracción_de_conocimiento (estructuras_de_información) = piezas_de_conocimiento.

Preguntas Problema:

¿Es posible crear procesos capaces de integrar diversos métodos de extracción de conocimiento de forma inteligente, teniendo en cuenta las fortalezas y debilidades de cada uno a partir un un conjunto de estructuras de información de dado?

Hipótesis:

Hipótesis I:

Existen diversos y efectivos métodos de extracción de conocimiento en forma de subgrafos. Desde el año 2007 se han propuesto, entre otros: TextRunner [Banko et al., 2007], WOE-parse y WOE-pos [Wu & Weld, 2010], ReVerb [Fader, 2011], OLLIE [Schmitz & Soderland, 2012], TRIPLEX [Mirrezaei et al., 2015]. Si bien recientemente se han publicado estudios que formulan algunas comparaciones [Del Corro & Gemulla, 2013], no existen estudios comparativos exhaustivos de la calidad de las piezas de conocimiento extraídas por los distintos métodos que refiere la literatura.

Hipótesis II:

Los métodos de extracción de conocimiento están pensados para trabajar con grandes estructuras de información bajo el supuesto de que la calidad de las piezas de conocimiento obtenidas son independientes del dominio. Sin embargo hay indicios experimentales [Schmitz & Soderland, 2012; Mirrezaei et al., 2015] de que esta independencia no es tal [Lopez-Nocera, 2012].

Hipótesis III:

En trabajos de investigación recientes [Del Corro & Gemulla, 2013; Mirrezaei et al. 2015], se plantea la posibilidad de combinar métodos para obtener una sinergia entre los mismos que redunde en una mayor calidad de las piezas de conocimiento obtenidas. Sin embargo no hay estudios comparativos sobre integración de métodos de extracción de conocimiento en general.

Objetivo General:

El objetivo de este proyecto es construir una familia de métodos de extracción de conocimiento tal que dado una estructura de información inicial como entrada, que contienen conocimiento embebido, genere un conjunto de piezas de conocimiento (reglas de producción, subgrafos de una red semántica, entre otros).

Objetivos Específicos:

Objetivo específico vinculado a la Hipótesis I (OE1):

Realizar una comparación entre los distintos métodos de extracción de conocimiento relevados en la literatura, indicando, para diversas estructuras de información, la calidad y los tiempos de ejecución asociados a cada uno.

Objetivo específico vinculado a la Hipótesis II (OE2):

Identificar fortalezas y debilidades en los diversos métodos evaluados en OE1 con el fin de detectar las condiciones particulares bajo las cuales es conveniente utilizar un método por sobre otro.

Objetivo específico vinculado a la Hipótesis III (OE3):

Desarrollar una familia de métodos integrados de extracción de conocimiento que exhiban un mejor

comportamiento que los métodos individuales integrados.

METODOLOGÍA DE TRABAJO

Para construir el conocimiento asociado al presente proyecto de investigación, se seguirá un enfoque de investigación clásico [Riveros y Rosas, 1985; Creswell, 2002] con énfasis en la producción de tecnologías [Sábato y Mackenzie, 1982]; identificando métodos, materiales y abordaje metodológico necesarios para desarrollar el proyecto:

Métodos:

Revisiones Sistemáticas:

Las revisiones sistemáticas [Argimón, 2004] de artículos científicos siguen un método explícito para resumir la información sobre determinado tema o problema. Se diferencia de las revisiones narrativas en que provienen de una pregunta estructurada y de un protocolo previamente realizado.

Prototipado Evolutivo Experimental (Método de la Ingeniería):

El prototipado evolutivo experimental [Basili, 1993] consiste en desarrollar una solución inicial para un determinado problema, generando su refinamiento de manera evolutiva por prueba de aplicación de dicha solución a casos de estudio (problemáticas) de complejidad creciente. El proceso de refinamiento concluye al estabilizarse el prototipo en evolución.

Materiales:

Para el desarrollo de los formalismos y procesos propuestos se utilizarán:

- Formalismos de modelado conceptual usuales en la Ingeniería de Software [Rumbaugh et al., 1999; Jacobson et al., 2013] y en la Ingeniería del Conocimiento [García-Martínez y Britos, 2004].
- Modelos de Proceso usuales en Ingeniería de Software [IEEE, 1997; ANSI/IEEE, 2007; Oktaba et al., 2007].

Abordaje Metodológico:

Para el desarrollo de esta propuesta técnica se han previsto utilizar las siguientes metodologías de investigación y desarrollo:

Para el Objetivo OE1 se propone: (i) realizar una investigación documental exploratoria buscando identificar los métodos de extracción de conocimiento en la literatura reciente; (ii) en base a los resultados del punto precedente identificar las distintas estructuras de información utilizadas para la evaluación de los distintos métodos; (iii) implementar los métodos relevados y ejecutarlos en las estructuras de información características identificadas; (iv) utilizar la metodología propuesta en [Bronzi et al., 2012] para la cuantificación de la calidad de los distintos métodos.

Para el Objetivo OE2 se propone: (i) realizar una investigación documental exploratoria buscando identificar diferencias en la calidad de las piezas de conocimiento obtenidas por los diversos métodos; (ii) reproducir los experimentos que evidencian las diferencias observadas; (iii) identificar las causas de las diferencias observadas; (iv) identificar que otros métodos son afectados por las mismas causas; (v) proponer nuevos experimentos basados en las características intrínsecas de los métodos de extracción de conocimiento relevados; (vi) identificar nuevas causas que evidencien diferencias entre los algoritmos; (vii) repetir los puntos anteriores hasta que se tenga un conjunto razonable de causas o situaciones que evidencien

Para el Objetivo OE3 se propone: (i) realizar una investigación documental exploratoria buscando identificar que métodos de extracción de conocimiento han sido integrados o utilizados de forma conjunta para mejorar la calidad de las piezas extraídas; (ii) identificar que otros métodos podrían ser integrados para ampliar la calidad o bien la cantidad de las piezas extraídas; (iii) identificar las circunstancias puntuales bajo las cuales la integración de métodos logra extraer más piezas de conocimiento o de mejor calidad que otro grupo de métodos integrados; (iv) desarrollar por el método de prototipado evolutivo un proceso automatizable, que bajo los supuestos anteriores, integre diversos métodos de extracción de conocimiento intentando maximizar la calidad y los tiempos totales según las necesidades del usuario.

RESULTADOS OBTENIDOS/ESPERADOS

El proyecto prevé formular aportaciones conceptuales en el área de métodos de descubrimiento de conocimiento en web. Contar con herramientas de este tipo permite explorar aplicaciones web en metabuscadores y en dispositivos adaptativos de explotación de información.

FORMACIÓN DE RECURSOS HUMANOS

El grupo de trabajo se encuentra formado por dos investigadores formados, dos investigadores en formación, y un asesor en metodología de la investigación. En su marco se desarrollan una Tesis de Doctorado en Ciencias Informáticas y una Tesis de Maestría en Tecnología Informática.

FINANCIAMIENTO

Las investigaciones que se proponen en esta comunicación cuentan con financiamiento como Proyecto 80020150200065LA de la Secretaria de Ciencia y Técnica de la Universidad Nacional de Lanús (Argentina).

REFERENCIAS

ANSI/IEEE, 2007. Draft IEEE Standard for software and system test documentation. ANSI/IEEE Std P829-2007.

Argimón, J. 2004. Métodos de Investigación Clínica y Epidemiológica. Elsevier España. 84-8174-709-2.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., & Etzioni, O. (2007, January). Open information extraction for the web. In IJCAI (Vol. 7, pp. 2670-2676).

Basili 1993. The Experimental Paradigm in Software Engineering. En Experimental Software Engineering Issues: Critical Assessment and Future Directions (Ed. Rombach, H., Basili, V., Selby, R.). Lecture Notes in Computer Science, Vol. 706. ISBN 978-3-540-57092-9.

Britos, P. 2008. Procesos de Explotación de Información Basados en Sistemas Inteligentes. Tesis de Doctorado en Ciencias Informáticas. Facultad de Informática. Universidad Nacional de La Plata.

Bronzi, M., Guo, Z., Mesquita, F., Barbosa, D., & Merialdo, P. (2012, June). Automatic evaluation of relation extraction systems on large-scale. In Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (pp. 19-24). Association for Computational Linguistics.

Christensen, J., Soderland, S., & Etzioni, O. (2011, June). An analysis of open information extraction based on semantic role labeling. In Proceedings of the sixth international conference on Knowledge capture (pp. 113-120). ACM.

Cowie J, Lehnert W (1996) Information extraction. Communications of the ACM 39(1):80–91.

Creswell, J. 2002. Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research. Prentice Hall. ISBN 10: 01-3613-550-1.

Del Corro, L., & Gemulla, R. (2013, May). ClausIE: clause-based open information extraction. In Proceedings of the 22nd international conference on World Wide Web (pp. 355-366). International World Wide Web Conferences Steering Committee.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A. M., Shaked, T., Soderland, S., ... & Yates, A. (2005). Unsupervised named-entity extraction from the web: An experimental study. Artificial intelligence, 165(1), 91-134.

Fader, A., Soderland, S., & Etzioni, O. (2011, July). Identifying relations for open information extraction. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 1535-1545). Association for Computational Linguistics.

García-Martínez, R. & Britos, P. V. (2004). Ingeniería de sistemas expertos. Nueva Librería. ISBN 987-1104-15-4.

Gómez, A., Juristo, N., Montes, C., & Pazos, J. (1997). Ingeniería del conocimiento. Editorial Centro de Estudios Ramón Areces. ISBN 84-8004-269-9.

IEEE, 1997. IEEE Standard for Developing Software Life Cycle Processes. IEEE Std 1074-1997 (Revision of IEEE Std 1074-1995; Replaces IEEE Std 1074.1-1995)

Jurafsky D, Martin JH (2009) Speech and language processing an introduction to natural language processing, computational linguistics, and speech recognition, 2nd edn. Prentice-Hall, Inc.

Konstantinova, N. (2014). Review of Relation Extraction Methods: What Is New Out There?. In Analysis of Images, Social Networks and Texts (pp. 15-28). Springer International Publishing.

Lopez-Nocera, M., Britos, P. Rodriguez, D., Garcia-Martinez, R. 2012. Impacto de la Complejidad del Dominio en las Variaciones del Comportamiento de Procesos de Explotación de Información. Proceedings IX Jornadas Iberoamericanas de Ingeniería del Software e Ingeniería del Conocimiento. Pág. 55-62. Sello Editorial de la Pontificia Universidad Católica del Perú. ISBN 978-612-4057-85-4.

Mesquita, F., Merhav, Y., & Barbosa, D. (2010). Extracting information networks from the blogosphere: State-of-the-art and challenges. In Proceedings of the Fourth AAAI Conference on Weblogs and Social Media (ICWSM), Data Challenge Workshop.

Mirrezaei, S. I., Martins, B., & Cruz, I. F. (2015). The Triplex Approach for Recognizing Semantic Relations from Noun Phrases, Appositions, and Adjectives. In The Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data (Know@LOD) co-located with Extended Semantic Web Conference (ESWC), Portoroz, Slovenia.

Oktaba, H., Garcia, F., Piattini, M., Ruiz, F., Pino, F., Alquicira, C. 2007. Software Process Improvement: The Competisoft Project. IEEE Computer, 40(10): 21-28. ISSN 0018-9162.

Riveros, H. y Rosas, L. 1985. El Método Científico Aplicado a las Ciencias Experimentales. Editorial Trillas. México. ISBN 96-8243-893-4.

Rumbaugh, J., Jacobson, I., Booch, G. 1999. The Unified Modeling Language, Reference Manual. Addison Wesley, ISBN-10: 02-0130-998-X.

Sábato, J. y Mackenzie, M. 1982. La Producción de Tecnología. Editorial Nueva Imagen. México. ISBN 968-429-348-8.

Schmitz, M., Bart, R., Soderland, S., & Etzioni, O. (2012, July). Open language learning for information extraction. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 523-534). Association for Computational Linguistics.

Wu, F., & Weld, D. S. (2010, July). Open information extraction using Wikipedia. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 118-127). Association for Computational Linguistics.

Yahya, M., Whang, S. E., Gupta, R., & Halevy, A. (2014, October). Renoun: Fact extraction for nominal attributes. In Proc. 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar.