Explicaciones en Argumentación Rebatible

Claudio J. Giulietti **Alejandro J. García †**

Laboratorio de Investigación y Desarrollo en Inteligencia Artificial
Departamento de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur,
† Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)
Av. Alem 1253, (B8000CPB) Bahía Blanca, Argentina
Tel: (0291) 459-5135 / Fax: (0291) 459-5136
e-mail: cgiulietti@gmail.com ajg@cs.uns.edu.ar

Resumen

En la literatura es usual encontrar acuerdo, en que un sistema experto o de recomendación debe responder con un nivel comparable a los humanos expertos, pero tambiéen debe tener la capacidad de explicar, en una forma entendible por el usuario, el proceso de razonamiento que empleó para resolver el problema y generar una recomendación. En general, cuando un usuario es responsable por la toma de decisiones, normalmente tiende a descartar una recomendación que no entiende completamente. Por ejemplo, aunque una recomendación efectuada por el sistema sea técnicamente correcta, la pérdida de confianza en ella puede deberse simplemente a que esté basada en alternativas que el usuario no había previsto. El objetivo principal de esta línea de investigación es intentar avanzar sobre el desarrollo de explicaciones que resulten aceptables para el usuario que recibe una recomendación de un sistema.

Palabras claves: Explicaciones, argumentación, DeLP.

Contexto

Esta línea de investigación se realizará dentro del ámbito del Laboratorio de Investigación y Desarrollo en Inteligencia Artificial (LIDIA) del Departamento de Ciencias e Ingeniería de la Computación de la Universidad Nacional del Sur. Está asociada con el proyecto de investigación: "Formalismos Argumentativos aplicados a Sistemas Inteligentes para Toma de Decisiones". Código: PGI 24/ZN18 financiado por la Universidad Nacional del Sur, y con el Proyecto de Investigación Plurianual, (PIP): "Sistemas de Apoyo a la Decisión Basados en Argumentación: formalización y aplicaciones", financiado por CONICET.

1. Introducción

Una explicación según [9] debe ser "*entendible*", debe permitir "*mejorar el conocimiento*", y debe ser "*satisfactoria*" en el sentido de cumplimentar las expectativas del interlocutor. En [14] se define una explicación como "*una transferencia de conocimiento de un interlocutor a otro dentro del contexto de un diálogo*". De acuerdo a lo expresado en [12], desde el punto de vista del diseño de las explicaciones, una explicación "*tiene que ser planeada y luego transmitida de una forma apropiada*", es decir una explicación es un "*objeto para ser diseñado*" y un "*acto comunicativo*" a ser logrado.

Dentro de la literatura encontramos una clasificación generalizada de las explicaciones [1, 14, 12, 15]. Esta tipología responde a que el concepto de explicación ha sido usado ampliamente para referirse a cualquier tipo de requerimiento de información, por ejemplo, requerimiento de instrucciones de operación, datos, explicación de términos, realimentación o jus-

tificacón de los métodos de razonamiento o del asesoramiento brindado. En [15] se expresa que intentar agrupar bajo el mismo concepto todos estos distintos requerimientos de información no resulta útil.

En términos generales se especifican 4 tipos de explicaciones, las cuales se mencionan a continuación:

1. "*Trace Explanation*": en [12] y [1] se define como una cadena de inferencias que muestra como fue realizada la derivación, es decir la secuencia de razonamiento ejecutada. En [15], se refiere al registro de los pasos inferenciales tomados por un sistema experto para alcanzar una conclusión.

2. "*Strategic Explanation*": [1, 14, 12, 15], revela la estrategia de resolución de problemas que posee el sistema para ejecutar una tarea, es decir como arriba el sistema a una determinada conclusión.

3. "*Deep Explanation*": [1, 14], en este tipo de explicaciones la respuesta ante una consulta se realiza utilizando el conocimiento del interlocutor, y no sólo del sistema. El sistema debe poseer el conocimiento del interlocutor. Este tipo de explicación se ajusta a la definición, de *transferencia de conocimiento*.

4. "*Justification Explanation*": [15], la cual es una descripcción explícita de la causalidad de los argumentos o de la racionalidad utilizada en la construcción de cada paso inferencial tomado por el sistema experto.

La generación de una "*explicación entendible*", en un sentido amplio, se refiere a una justificación o explicación coherente solicitada por el interlocutor. En [14] se indica que una propiedad o característica que deben poseer las explicaciones, es que deben ser "*explicaciones entendibles*". El agente que actua como interlocutor, sólo posee un conocimiento parcial del tema en cuestión, por lo cual solicita una explicación, con la esperanza que el agente que genera la explicación pueda cubrir sus vacíos o *gaps* de conocimiento.

En [12] se expresa que para que el usario acepte una explicación, el sistema debe convencerlo que la misma es "*relevante, justificada y útil*". Para esto el sistema debe explicar como alcanzó la conclusión y además indicar cuales son los argumentos sólidos (sound arguments) que soportan esta conclusión. Viendo las explicaciones como un acto comunicativo, para que sean aceptables, las explicaciones generadas por el sistema deben ser "*expresivas*", también deben estar "*adaptadas al conocimiento del usuario*", y deben ser "*sensibles*" a las necesidades del usuario.

En [12] se indica como crítica a los sistemas actuales de recomendación o apoyo a la decisión, que usualmente este tipo de sistemas no son usados eficientemente por los responsables en la toma de decisiones debido a una pérdida de confianza en la recomendación provista por el sistema. Además, en [7] se indica que, en general, cuando un usuario es responsable por la toma de decisiones, normalmente tiende a descartar la recomendación que no entiende completamente. La pérdida de confianza en el sistema, puede deberse a que la recomendación efectuada por el sistema, si bien puede ser técnicamente aceptable, puede ser inaceptable para la persona, ya que ésta puede estar basada en alternativas que el usuario no había previsto.

Varios autores [2, 10, 11, 13] concuerdan en que un sistema debe responder a un nivel comparable con los humanos expertos, pero también debe tener la capacidad de explicar, en una forma entendible por el usuario, el proceso de razonamiento que empleó para resolver el problema y generar una recomendación. Además [15] agrega que, si bien, la capacidad de brindar explicaciones es importante en todo producto de software, para el caso especial de los sistemas expertos, es aún mas importante, ya que sin ello, los usuarios no estarían en capacidad de rechazar una recomendación del sistema cuando esté equivocado y adicionalmente, serán renuentes a la aceptación de un consejo aún cuando el sistema esté acertado en su recomendación.

En [14], se identifican Explicaciones Fallidas

y Explicaciones Exitosas. Se expresa que el modelo conversacional, puede facilitar un contexto dentro del cual se puede arribar al éxito de una explicación. Los sistemas de explicaciones deben abordar o plantear distintos tipos de explicaciones o estrategias para lograr el éxito en su objetivo. Actualmente, tampoco se aborda el tema de calificar o juzgar cuan exitosa es una explicación ni tampoco el grado de entendimiento que pudo ser transferido al interlocutor. Dada una pareja de agentes que establecen un diálogo (en el contexto de las explicaciones) existe un *entendimiento compartido (sharing understanding)*, pero también existen *diferentes huecos o vacíos (gaps)* de entendimiento, los cuales determinan el éxito o el fracaso de la explicación. Es decir se establecen distintos niveles de conocimiento y de desconocimiento, que pueden provocar la falla en el intento de explicar una consulta.

2. Líneas de investigación y desarrollo

El objetivo principal de esta línea de investigación es intentar avanzar sobre el desarrollo de explicaciones que resulten aceptables para el usuario que recibe una recomendación de un sistema.

En esta línea de investigación nos centraremos, desde el punto de vista general en los Sistema Argumentativos Basados en Reglas *SABR*, y en particular trabajaremos sobre un Sistema Argumentativo concreto denominado DeLP (Defeasible Logic Programming) [6]. Los *SABR* son formalismos de argumentacíon en los cuales el conocimiento incluye un conjunto de reglas de inferencia que permiten construir argumentos a favor o en contra de una afirmación. Estos sistemas son de particular interés en el área de Inteligencia Artificial dado que este tipo de reglas de inferencia permite representar conocimiento de sentido común, posibilitando la construcción de argumentos de manera automática.

DeLP es un formalismo que combina resultados de programación en lógica y argumentación rebatible. Utilizando DeLP es posible representar información tentativa de forma declarativa, mediante el uso de reglas "débiles". Adicionalmente, dado que es posible utilizar negación estricta en la cabeza de este tipo de reglas, es posible representar información contradictoria. En este formalismo se identifican aquellos elementos en contradicción, y posteriormente se lleva a cabo un proceso argumentativo de dialéctica para determinar cual de estos elementos prevalecerá.[5]. Este sistema se encuentra totalmente implementado y disponible online en [8].

Si bien, existe trabajo preliminar sobre explicaciones en DeLP [4, 3], el objetivo principal de esta línea de investigación, es reformular, adaptar y expandir este modelo, para lograr que el nuevo modelo, esté de común acuerdo a lo especificado en la literatura y a partir de allí, expandir varios aspectos que detallaremos mas adelante.

Nuestra línea de investigación está compuesta por dos ejes principales: el primer eje se basa en la adaptación del modelo actual con el fin de satisfacer y brindar todas las características mencionadas en la introducción y que a nuestro entender debe brindar una explicación; y el segundo eje de investigación se basa en expandir determinadas características del modelo de explicaciones.

2.1. Primer eje: adaptación

En los trabajos que anteceden a esta investigación [3], en general lo que se interpreta como explicación, es devolver al interlocutor el conjunto de árboles de dialéctica generados por el proceso de razonamiento. Este tipo de respuesta, desde el punto de vista del interlocutor, puede no ser satisfactoria, en el caso que el interlocutor reciba más información de la que realmente necesita (por ejemplo enviándole todo el conjunto de árboles de dialéctica). Esto haría que la explicación no fuese útil, y que tampoco pudiese adaptarse para la integración adecuada de la información (transferencia de conocimiento).

Como primer paso, se establecerán diferen-

tes niveles en lo que respecta a los tipos de explicaciones que brindará el modelo, las cuales se indican a continuación: *Conjunto de Arboles de Dialéctica*, *Arbol de Dialéctica*, *Línea de Argumentación* y *Argumento*. Cada uno de estos niveles permite generar distintos tipos de respuestas de acuerdo a las necesidades del interlocutor. Sobre cada uno de los niveles, se adaptarán las siguientes características o propiedades:

1. Explicación Concisa/Completa: concepto relacionado con la búsqueda de argumentos que permiten obtener una garantía para la consulta. En general, el método de razonamiento, una vez que encuentra un argumento que está garantizado (en caso de existir) no continua la búsqueda.

2. Explicación Satisfactoria: concepto relacionado con la capacidad de mostrar en la explicación generada, los argumentos que fueron construidos pero no fueron considerados por el algoritmo de razonamiento. Existen varias causas por la cual el método de razonamiento, no considera un argumento, una de las causas es que el argumento puede ser "*falaz*", en particular, es importante a nivel de explicación, poder tener una visión de las falacias que pueden producirse en la explicación y porque no fue considerado ese argumento.

3. Explicación Parcial/Exhaustiva: este concepto esta relacionado con la forma en la cual se construye un árbol de dialéctica o árbol de razonamiento, dentro del cual puede optimizarse realizando "podas", en las líneas de argumentación para optimizar la respuesta del sistema, lo cual brinda mayor rapidez. Esta optimización está ligada directamente al método de búsqueda de contra-argumentos para un argumento particular, una vez hallado un contra-argumento que no es derrotado (garantizado) se detiene la búsqueda. En la explicación exhaustiva, se obtendrían todos los posibles contra-argumentos, y esto permitiría poder presentar la misma explicación de distintas formas o soportadas por argumentos alternativos. Es decir, se dispondría de la capacidad de brindar la misma explicación de distintas formas.

4. Explicaciones Justificadas: concepto relacionado con la presentación de argumentos que justifican o sustentan la conclusión garantizada.

2.2. Segundo eje: expansión

El segundo eje está basado en investigar como lograr que un sistema que provee explicaciones satisfaga las características mencionadas en la literatura y resumidas en la introducción de este trabajo.

En particular se trabajará en lograr que el usario acepte una explicación. Como para ello el sistema debe convencerlo que la misma es relevante, justificada y útil. Entonces se debe trabajar en buscar que el sistema explique como alcanzó la conclusión y que indique cuales son los argumentos que soportan esta conclusión. Se buscará definir elementos para lograr identificar cuando las explicaciones generadas por el sistema son expresivas o adaptadas al conocimiento del usuario.

3. Resultados obtenidos y esperados

Los resultados obtenidos hasta el momento, debemos indicarlos dependiendo de los dos ejes principales en que se trabaja en esta investigación.

De la primera línea de investigación, se han identificado los distintos niveles de tipos de explicaciones, y para cada una de ellas, se han definido las características y formas de generación de las explicaciones, según sean *concisas/completas, satisfactoria, parcial/exhaustiva y/o justificada*.

Con respecto al segundo eje se espera poder definir explicaciones que resulten aceptables para el usuario que recibe una recomendación de un sistema.

Referencias

[1] B. Chandrasekaran. Generic task in knowledge based reasoning. *IEEE Experts*, 1(1):23–30, 1986.

[2] R.O. Duda and E.H. Shortliffe. Expert systems research. *Science*, (220):261–268, April 1983.

[3] A.J. Garcia, C.I. Chesñevar, N.D. Rostein, and G.R. Simari. Explaining why something is warranted in defeasible logic programming. In *IJCAI 2009 Workshop on Explanation aware Computing (Exact 2009)*, 2009.

[4] A.J. Garcia, N.D. Rostein, and G.R. Simari. Dialectical explanations in defeasible argumentaqtion. In *ECSQARU*, pages 295–307, 2007.

[5] Alejandro J. García. *Defeasible Logic Programming: Definition, Operational Semantics and Parallelism*. PhD thesis, Computer Science Department, Universidad Nacional del Sur, Bahía Blanca, Argentina, December 2000.

[6] Alejandro J. García and Guillermo R. Simari. Defeasible logic programming: An argumentative approach. *Theory and Practice of Logic Programming*, 4(1):95–138, 2004. http://xxx.lanl.gov/abs/cs.AI/0302029.

[7] E. Hollnagel. Commentary: Issues in knowledge-based decision support. In *International Journal of Man-Machine Studies*, pages 743–751, November/December 1987.

[8] DeLp home Page. web page: http://lidia.cs.uns.edu.ar/delp. 2007.

[9] C. Lacave and F. Diez. A review of explanation methods for heuristic expert systems. *The Knowledge Engineering Review*, 0(0):1–13, 2005. Cambridge University Press.

[10] J.D. Moore and W.R. Swartout. Explanation in expert systems: A survey. In Information Sciences Institute. ISI Research, editor, *Report, RR-88-228.*, pages 345–353. University of Southern California, Los Angeles, CA, 1988, 1988.

[11] J.D. Moore and W.R. Swartout. A reactive approach to explanation. In *Proceedings of the Eleventh Intemational Joint Conference on Artificial Intelligence*, Detroit, Ml, aug 1989.

[12] B. Moulin, H. Irandoust, M. Belnger, and G. Desbordes. Explanation and argumentation capabilities: Towards the creation of more persuasive agents. *Artificial Intelligence Review*, 17:169–222, 2002. ISSN 1386-9795 print/1741-5918 online/04/010071-19.

[13] R.L. Teach and E.H. Shortliffe. An analysis of physicians' attitudes. *Computers in Biomedical Research*, (14):542–558, December 1981.

[14] Douglas Walton. A new dialectical theory of explanation. *Philosophical Explorations*, 7(1), 2004. ISSN 1386-9795 print/1741-5918 online/04/010071-19.

[15] L. Richard Ye and Paul E. Johnson. The impact of explanation facilities on user aceptance of expert systems advice. *MIS Quarterly*, June 1995.