Prospección de Datos aplicada a problemas sociales *

Rafael Morales-Bueno, Jorge Wallace-Ruiz y Manuel Baena-García

Departamento de Lenguajes y Ciencias de la Computación
E.T.S. Ingeniería Informática. Universidad de Málaga
Málaga, 29071, España

Resumen En este artículo mostramos aplicaciones de la prospección de datos a problemas sociales. La intención es aumentar la calidad de los procesos de atención social, realizar estudios de cuáles de los datos obtenidos en dichos procesos representan mayor cantidad de información útil para la valoración de la situación social de una persona y desarrollar sistemas de ayuda a la valoración para dicha situación. Concretamente se realizan estudios sobre incapacidad permanente y violencia de género. Estos estudios siguen la metodología de prospección CRISP-DM.

1. Introducción

La automatización de procesos se integra cada día más en nuestra sociedad. La tendencia actual es el desarrollo de "sistemas de información" para la adquisición y transferencia de información. Del resultado de esta automatización se obtienen datos "informatizados" que en ocasiones son almacenados y en ocasiones descartados. La prospección de datos [FPSM91] pretende extraer información interesante desde datos almacenados en grandes bases de datos, combina técnicas de estadística clásica y técnicas de inteligencia artificial como aprendizaje estadístico y computacional.

Desde el grupo "Investigación y Aplicaciones en Inteligencia Artificial" (IA^2) de la Universidad de Málaga apostamos por la evolución de los sistemas de información hacia "sistemas de información inteligentes". Con el propósito de potenciar dicha evolución desarrollamos proyectos de prospección de datos orientados a generar conocimiento y desarrollar aplicaciones para tratar problemas reales. Nuestra intención es dar a conocer lo que representa la prospección de datos y cómo puede aplicarse para resolver eficientemente problemas de la sociedad.

En un intento por acercarse y comprender la realidad social estamos realizando estudios de temática social relevante. En este artículo presentamos dos de ellos: estudio de la Incapacidad Permanente y estudio para la prevención de Violencia de Género.

En la siguiente sección se presenta la guia para la prospección de datos CRISP-DM [CCK+99]. A continuación se muestra la aplicación de una iteración

* Este trabajo ha sido parcialmente financiado por el proyecto MOISES, número TIC-2002-04019-C03-02 del Ministerio de Ciencia y Tecnología.

del proceso CRISP-DM a datos de Incapacidad Permanente. Por último exponemos las fases iniciales de un proyecto de prospección a datos de Violencia de Género.

2. El Proceso CRISP-DM

Al hablar de prospección de datos debemos hablar del proceso de prospección. Existen diferentes metodologías, la más aceptada y por la que nos hemos decantado es la guía CRISP-DM. En ella se define un modelo jerárquico de tareas a cuatro niveles de abstracción: En el nivel más alto el proceso se organiza en un número de fases; cada fase consiste en varias tareas genéricas de segundo nivel. Las tareas específicas solucionan tareas genéricas en situaciones específicas. Y los procesos son cada una de las aplicaciones de las tareas específicas a datos.

El modelo de proceso provee una visión global del ciclo de vida de este tipo de proyectos. Este ciclo de vida contiene las fases del proyecto, sus respectivas tareas y las relaciones entre estas tareas. El interés de esta sección es describir, a nivel de fases, el ciclo de vida de la guía CRISP-DM.

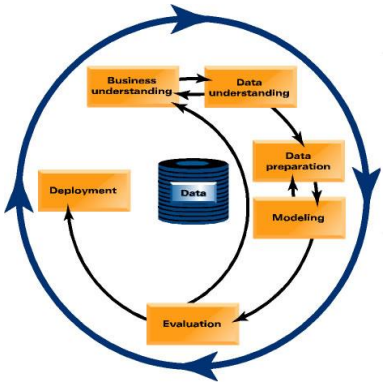

Figura 1. Proceso CRISP-DM. http://www.crisp-dm.org/

El ciclo de vida de un proyecto de prospección de datos consta de seis fases (figura 1). La secuencia de realización de estas fases no es estricta, se requieren movimientos de una fase a otra en base a los resultados obtenidos durante el proceso. Las flechas indican las dependencias más frecuentes entre las diferentes fases. El círculo exterior simboliza la naturaleza cíclica de la prospección de datos. La prospección de datos no concluye cuando se encuentra una solución, pues esta solución puede abrir nuevas puertas a explorar. La iteración en el proceso beneficia la experiencia adquirida en la iteración anterior.

La fase inicial del proceso de prospección de datos (comprensión del problema) está enfocada en la comprensión de los objetivos y requisitos del proyecto desde una perspectiva del cliente. En la comprensión de datos se recopilan los datos necesarios para el desarrollo del proyecto, se ofrece información de la estructura de estos datos, de su calidad y subconjuntos de interés. En la preparación de datos se cubren todas las actividades para construir los conjuntos de datos finales (datos que serán utilizados en las herramientas de modelado) a partir de los datos en bruto iniciales. En el modelado se seleccionan y aplican diferentes técnicas y sus parámetros son calibrados a los valores óptimos. Antes de proceder con el desarrollo final del modelo, es importante una evaluación más profunda y una revisión de los pasos seguidos en su construcción para asegurar que sus propiedades están dentro de los objetivos del proyecto.

3. Proceso de Prospección de Incapacidad Permanente (IP)

En esta sección se exponen los resultados de la aplicación de la metodología CRISP-DM a datos de Incapacidad Permanente. Se presenta un resumen del informe final (trabajo completo en [Gar03]), en el que quedan recogidos los resultados de cada una de las fases. Con esto se pretende dar una visión general de lo que sería el desarrollo de un proceso de prospección. No pretende ser un estudio exhaustivo de la Incapacidad Permanente, pues el conjunto de datos usado es limitado (978 expedientes). En los siguientes puntos se presenta el resultado de cada una de las fases del proceso:

- Comprensión del Problema: El Real Decreto 1300/95 establece las competencias que en materia de incapacidad laboral permanente le corresponde al Instituto Nacional de la Seguridad Social (INSS). Dicho texto legal crea en las Direcciones Provinciales del INSS los Equipos de Valoración de Incapacidades (EVI). La incapacidad permanente (IP), en su modalidad contributiva, tiene en cuenta la alteración continuada de la salud y, fundamentalmente, la incidencia que dicha alteración tiene en la realización de la actividad profesional. Tiene un perfil exclusivamente profesional y su calificación debe obviar toda referencia a otras circunstancias (socio-económicas, de edad, familiares, etc.). Se clasifica conforme a los siguientes grados: IP Parcial, IP Total, IP Absoluta, Gran Invalidez y Lesiones Permanentes No Invalidantes. En base a lo expuesto, los objetivos del problema son dos: aportar al INSS una serie de datos que permitan a los Médicos Inspectores tener una aproximación, previa a la evaluación individualizada, del resultado que cabe esperar en cada uno de los expedientes de IP y permitir satisfacer las expectativas de los beneficiarios de la Seguridad Social. Se pretende que los asegurados conozcan en cada momento, si su situación clínico-laboral es susceptible de generar una IP con el menor margen de error posible.
- Comprensión de datos: La tarea principal de esta fase es la Descripción de datos. En esta tarea se examinan, a groso modo, las propiedades de los datos capturados. En el informe se describen el tamaño de los datos (análisis volumétrico) y su formato, se realizan medidas estadísticas básicas y se analiza el significado de los resultados, por ejemplo la relevancia de un atributo para los objetivos y su comparación con la opinión del experto del dominio. La Unidad Médica del EVI elabora los Informes Médicos de Síntesis (IMS) como documento preceptivo para evaluar la discapacidad laboral. Por otro lado, las actas resultantes de las sesiones celebradas por el EVI constan de la propuesta del grado de invalidez permanente, la contingencia determinante y si hay que realizar revisión o no y en que fecha. Los datos se han obtenido, por personal autorizado, tanto de los IMS como de las actas de las sesiones garantizando la privacidad. Algunos de los datos, como la edad o el sexo, han sido extraídos directamente de estos documentos, otros, como la repercusión laboral, son datos calculados, capturados por personal cualificado.

- Preparación de datos: Un subconjunto de los datos adquiridos en fases previas es seleccionado, basándose en características recalcadas en dichas fases, y se crean conjuntos de datos válidos para la aplicación de las técnicas de prospección en la fase de modelado.
- Modelado: En esta fase se seleccionan las técnicas de modelado, se aplican sobre los conjuntos de datos y se calibran sus parámetros a valores óptimos. Particularmente, en el proceso de prospección de IP, hemos utilizado árboles de decisión y reglas de asociación.
- Evaluación: En las fases previas (principalmente en la fase de modelado) se completa la evaluación de la precisión de los modelos construidos. En esta fase se evalúan los modelos respecto a los objetivos a solventar del problema. Es cuando se tiene que decidir si existen motivos por los que los modelos generados sean deficientes.
  De acuerdo a la evaluación de los resultados y la revisión del proceso, se decide como proceder en el futuro. Se necesita decidir cuando finalizar el proyecto y, en caso de ser apropiado, realizar el desarrollo de software o cuando iniciar una nueva iteración del proceso o preparar nuevos proyectos de prospección de datos.
  Con el fin de completar los objetivos de la prospección de datos, y dado que los resultados obtenidos en los modelos lo permiten, desarrollamos el sistema SAVI, un sistema para la ayuda a la valoración de incapacidad permanente.
- Desarrollo: La creación del modelo no supone, generalmente, la finalización del proyecto. Así, si el propósito del modelo es incrementar el conocimiento de los datos, este conocimiento obtenido necesita organizarse y representarse de manera que pueda ser usado. Con el sistema SAVI mostramos un ejemplo de lo que se podría conseguir mediante la aplicación de procesos de prospección a datos sanitarios (ver figura 2).

4. Prospección de Datos de Violencia de Género

Continuando el intento por parte de la Universidad de Málaga de acercarse a la realidad social y continuando a su vez la colaboración establecida entre el departamento de Lenguajes y Ciencias de la Computación y el de Medicina Legal y Forense, surge de nuevo la necesidad de realizar un estudio en profundidad de temática social relevante. En este caso la violencia de género (o doméstica), este primer término, que ha sido rechazado por parte de la Real Academia Española de la Lengua, era el que estaba vigente a la hora de comenzar nuestro trabajo. Por violencia de género entendemos cualquier acto de violencia sobre personas que estén totalmente desamparadas debido a cuestiones de sexo, edad o algún otro tipo de relación existente entre agresor-víctima.

Este estudio es en realidad parte del Proyecto Fin de Carrera de Jorge Wallace Ruiz, cuyos tutores son Rafael Morales (Dpto. LCC) e Ignacio Santos (Dpto. Medicina Legal). Este proyecto (aún en desarrollo) tiene por título "Técnicas de Prospección de Datos aplicadas a la prevención de la Violencia de Género". El objetivo final del proyecto es prevenir la violencia de género mediante la

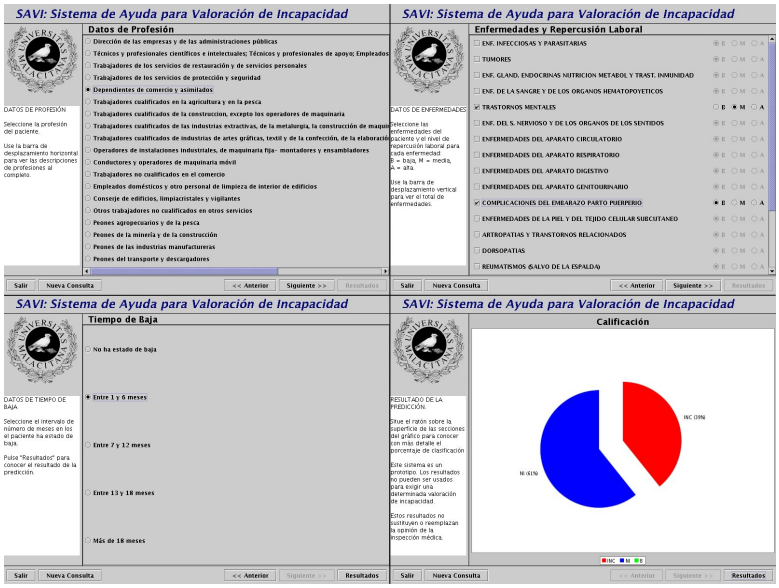

Figura 2. Capturas de pantalla del sistema SAVI

extracción de conocimiento relevante a partir de los datos. Es importante recalcar que este estudio engloba tanto a mujeres maltratadas como a relaciones familiares (p.ej : Padre-Madre/ Hijo).

Para realizar este estudio el primer paso fue la realización de un cuestionario de recogida de datos, para ello nos pusimos en contacto con los grupos de personas o entidades que de alguna forma tienen relación directa con la violencia doméstica, como pueden ser: Policía, Instituto Andaluz de la Mujer (IAM), Servicio de atención a la Víctima de Andalucía (SAVA) y Colectivos de Mujeres, como violencia cero, entre otros. De esta forma pudimos conocer de cerca la realidad social de estas mujeres y descubrimos también otro tipo de agresiones que también son importantes en nuestro estudio como las relaciones padre-hijo. A partir de estos encuentros se realizó un formulario preliminar el cual sirvió de base para el formulario actual. En este momento, cada caso de violencia doméstica lleva asociados 47 atributos entre víctima y agresor tales como edad, sexo, tiempo de relación,... Estos datos están siendo recogidos en la provincia de Málaga por una serie de cuerpos y fuerzas de seguridad del estado.

Al iniciar el proyecto nos propusimos unos objetivos principales a cumplir, estos son:

- Realizar un proceso de prospección de datos siguiendo la metodología CRISP-DM, resaltando la fase de comprensión de datos.
  Seguiremos el proceso de prospección de datos líder en el mercado Actualmente, CRISP-DM. El principal objetivo de este proyecto es realizar un

estudio en profundidad de la fase de comprensión de datos que realmente es de las más importantes de todo el proceso ya que suele irse en ella más del 80 % del tiempo de realización del proyecto. Suele ser de vital importancia ya que una mala ejecución de esta fase puede llevarnos al fracaso del proyecto.

- Obtención de perfiles conductuales de agresores y víctimas de violencia de género.
  Para ello utilizaremos técnicas de clustering sobre los datos. Estas técnicas consisten en realizar una partición sobre el conjunto de datos según una serie de criterios de distancias entre atributos y relaciones de los mismos.
- Obtención de reglas de asociación sobre el conjunto de datos.
  Mediante la aplicación de técnicas de obtención de reglas de asociación a priori, éstas nos permitirán obtener relaciones que se establecen entre distintos atributos presentes en nuestro conjunto de datos. También las utilizaremos para la demostración empírica o refutación de las distintas teorías establecidas sobre este problema concreto (Violencia de género) ya que los distintos colectivos implicados proponen teorías muchas veces totalmente enfrentadas.
- Clasificar a víctimas de violencia de género según un índice de peligrosidad.
  Las técnicas de árboles de decisión nos permiten realizar una clasificación "probabilística" en clases. En nuestro caso realizaremos una clasificación binaria homicida-no homicida, obteniendo de esta forma una clasificación con una probabilidad asociada.
- Generación de un programa basado en el modelo de conocimiento obtenido a partir de los datos iniciales.
  A partir de los datos iniciales (en principio esperamos que sean entorno a un millar) obtendremos un modelo de conocimiento. Este modelo nos serviría para clasificar nuevos casos a partir de los ya existentes en nuestro modelo de conocimiento.
  Se estudia también la posibilidad de crear un sistema de conocimiento que esté en continuo crecimiento, esto es, realizar un programa web mediante servlets de java de forma que expertos en la materia puedan introducir nuevos datos en nuestra base de datos y se actualice el modelo de conocimiento dinámicamente.

Actualmente estamos a la espera de recibir nuevos datos ya que actualmente estamos trabajando con un conjunto de datos que apenas llega a los 150 casos. Debido a la cadencia de los casos de violencia doméstica en Málaga y a su mayor incidencia de los mismos durante el verano, esperamos acercarnos a los 500 casos a finales de verano. Aún así con el conjunto inicial de datos ya hemos obtenido información muy interesante a la espera de realizar el contraste con el conjunto de datos completo que esperamos que alcance los 1000 casos a mediados del 2005.

Referencias

[BMCS04] M. Baena, R. Morales, S. Cabuchola, and I. Santos. Prospección de datos sanitarios: Estudio de la incapacidad permanente. In *Inforsalud04*, pages 127–130, 2004.

[CCK+99] P. Chapman, J. Clinton, T. Khabaza, T. Reinartz, and R. Wirth. The crisp-dm process model. Technical report, CRISPDM consortium, 1999.

[FPSM91] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus. Knowledge discovery in databases: an overview. In G. Piatetsky-Shapiro and W. J. Frawley, editors, *Knowledge discovery in databases*, pages 1–27, Menlo Park, CA/Cambridge, MA, 1991. AAAI Press/MIT Press.

[Gar03] Manuel Baena García. *Prospección de Datos: Estudio de diversas técnicas y su aplicación a datos sobre Incapacidad Laboral*. Pfc, Universidad de Málaga, 2003.