UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE FILOLOGÍA
Departamento de Filología Inglesa I

TERMINOLOGÍA DE ESTADÍSTICA Y MINERÍA DE DATOS EN LENGUA INGLESA.

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

José Luis Llavona Arregui

Bajo la dirección de la doctora

Paloma Tejada Caller

Madrid, 2010

ISBN: 978-84-693-7755-0

PARTE I. MARCO TEÓRICO

PARTE II. CREACIÓN DE UN GLOSARIO TERMINOLÓGICO DE MINERÍA DE DATOS

PREFACIO

"Para las lenguas receptoras, pues, el problema no es de política lingüística, sino de estímulos a la inventiva o de capacidad tecnológica reforzada por una buena estrategia comercial. Quien fabrica y vende la mercancía, como quien concibe y titula una obra literaria, tiene sin duda derecho a ponerle el nombre que juzgue conveniente. Que este nombre se difunda depende de la aceptación, si no de la calidad, del nombre y del producto"

Emilio Lorenzo Criado

En su discurso de entrada en la Real Academia Española (1981:41) Emilio Lorenzo ya comentaba una cuestión que, casi treinta años después, se mantiene plenamente vigente: ¿Qué actitud deben mantener las autoridades lingüísticas de las lenguas de destino ante la continua llegada de neologismos? Como veremos más adelante, tales neologismos, dependiendo de su ámbito de uso, inciden o bien en el habla cotidiana, dando lugar a la incorporación de nuevas palabras, o bien en el vocabulario técnico y científico, en cuyo caso lo que aumenta es el número de términos propios del campo de especialidad de que se trate.

En el presente trabajo trataremos de establecer claramente las diferencias entre los conceptos de *palabra* y *término*, con el propósito de demostrar que, mientras las palabras importadas sí pueden alterar el vocabulario de las lenguas receptoras en mayor o menor medida (dependiendo de la incidencia de los factores que resalta Lorenzo), la terminología específica de un campo de conocimiento permanece recluida y a salvo en los entornos de especialidad que son su *habitat* natural. Es en estos entornos donde nacen, evolucionan y permanecen o acaban desapareciendo, ajenas al hablante común.

Sirvan estas palabras de homenaje a quien, inicialmente como profesor en los seminarios de doctorado y después como primer director de esta tesis, hizo nacer en este autor el gusto por la pureza del lenguaje, sin exclusiones ni radicalismos, aprendiendo que las lenguas por sí mismas nunca deben ser consideradas una amenaza: La influencia que puedan ejercer unas en las otras dependerá del buen criterio y formación de sus hablantes, que por lo común aceptan de buen grado aquello que les enriquece.

INTRODUCCIÓN

Esta tesis tiene como finalidad mostrar los resultados de mi investigación en el campo de la Terminología a través de una aplicación práctica de los conocimientos adquiridos. Es una aproximación que nació sin ideas preconcebidas, si tal cosa es posible. La motivación fundamental de este autor ha sido el deseo de descubrir un campo nuevo que se mostraba afín y atractivo por las razones que más adelante se detallan, pero que podemos resumir en dos palabras que definen a mi modo de ver esta ciencia: precisión y pragmatismo.

El punto de partida había de ser necesariamente el estudio de las distintas corrientes teóricas que en este campo existen y las diferentes propuestas que surgen de las mismas. Estas propuestas se ven reflejadas tanto en las obras de los autores que las sustentan como en los diferentes cursos de formación en terminología para lingüistas que se ofertan en la actualidad. De todos ellos he procurado recoger los más representativos. Como consecuencia de lo anterior, este trabajo continúa con una reflexión propia sobre el campo de estudio, que he denominado “la terminología como problema sociolingüístico”.

Un estudio terminológico que se limitara a dar una idea de la situación en que se encuentra esta ciencia hoy en día, por más que aporte impresiones del autor, estaría incompleto si no incorporara una aplicación práctica de los

fundamentos teóricos estudiados. Por eso, una vez revisados los aspectos metodológicos, la investigación se centró en la creación de un glosario terminológico sobre los términos propios de la minería de datos, actividad que surge en los últimos doce años como resultado de los avances experimentados en los campos de la estadística y la informática. Era este un campo prácticamente virgen, en el que nunca antes se había realizado una recopilación de su terminología particular. Finalmente se incluyen los apartados de conclusiones, apéndices y bibliografía.

Resulta imprescindible en esta introducción hacer una consideración previa que, sin duda, facilitará la comprensión del contenido del presente trabajo: al igual que no hay obra independiente de su autor, no hay autor independiente de su formación académica y entorno profesional. Le podremos llamar circunstancias, contexto, o como más nos guste, pero en cualquier caso, es un factor que condiciona nuestra perspectiva del campo de estudio, nuestros métodos de trabajo y la manera en que plasmamos los resultados de nuestra investigación.

En mi caso la formación en filología encontró su desarrollo profesional en un entorno sumamente determinante, como es la Universidad Politécnica de Madrid. Como consecuencia, el objetivo de mi actividad docente e investigadora ha sido siempre la formación de ingenieros. La pregunta que surge entonces es: ¿en qué se diferencia la formación de un ingeniero de la de cualquier otro universitario y cómo influye esta actividad en quien la practica?

Podríamos intentar contestar haciendo una relación de aquello a lo que, aun a nuestro pesar, debemos renunciar (por motivos de diversa índole que sin duda no serán desconocidos para quien, procediendo de un entorno humanístico, desarrolla su vida profesional entre técnicos) al impartir docencia en lenguas para entornos de ingeniería: a despertar en el discente el disfrute del conocimiento por el mero placer de poseerlo y transmitirlo, a la estética del

texto bien escrito... y podría seguir, pero entonces incurriría en una contradicción con el objetivo de este punto aclaratorio, por lo que paso a responder de forma directa: en el mundo de la ingeniería prima por encima de todo la aplicación práctica del conocimiento. Dar soluciones reales a problemas reales. Interesa que la estructura proyectada cumpla su fin, que la máquina sea exacta hasta la infalibilidad o que el programa informático funcione. ¿Qué hace entonces un lingüista en este mundo de números y fórmulas?; ¿cómo conjugar lo aparentemente incompatible? (una ciencia marcadamente teórica con el empirismo de la ingeniería). Buscando puntos en común, si hay una rama de la lingüística en contacto directo con las ciencias aplicadas, esa es sin duda la terminología. Terminología es precisión, terminología es, como veremos posteriormente, transmisión de conocimiento (con frecuencia técnico) especializado. Parece pues perfectamente adecuada a nuestro entorno y propósito.

Consecuentemente, este estudio se ve impregnado de ese pragmatismo que marcan las circunstancias de su autor. El método de trabajo a seguir es el mismo que se daría en un entorno de ingeniería: una vez constatada la existencia de un problema (la falta de un glosario terminológico completo en un campo de conocimiento), se estudian las distintas soluciones teóricas (llámense corrientes, escuelas o tendencias actuales en terminología) y se propone una solución práctica adecuada al caso (creación de un glosario terminológico de utilidad para estadísticos, informáticos y, fundamentalmente, expertos en minería de datos en el caso que nos ocupa).

En consecuencia con lo dicho anteriormente, vamos a tratar de analizar como punto de partida el proceso de creación, aplicación y decantación final (o, de no producirse ésta, desaparición) de lo que venimos a llamar término. Para ello comenzaremos por definir término:

Según la Norma ISO (International Standardization Organization) (1) 1087-1, término es

> "La designación verbal de un concepto general en un campo de conocimiento específico"

Por lo tanto los términos no pertenecen al léxico del hablante general, sino que nacen, viven y permanecen o terminan por caer en desuso en entornos de lengua especializada. Esto es lo que distingue fundamentalmente a los términos de las palabras comunes: pertenecen al léxico especializado.

En este sentido podríamos llegar a afirmar que los términos "huyen de la luz": desde el momento en que cualquier hablante de la lengua lo usa, un término ya no es un término estrictamente hablando, puesto que deja de pertenecer al ámbito –restringido- de la lengua especializada (2) y comienza a estar connotado; desde ese instante pasaría a convertirse en una palabra más del léxico común. Bajo este supuesto de partida, la función del terminólogo ha de consistir en facilitar la comprensión entre especialistas (que puede estar limitada por barreras idiomáticas) pero no entre especialistas y el público en general.

En función de sus usuarios, podemos analizar los términos desde múltiples puntos de vista:

(1) ISO es una red formada por los institutos de estandarización –o normalización- de 157 países, con sede en Ginebra. Es una organización no gubernamental, si bien gran número de sus institutos miembros sí pertenecen a los gobiernos de sus países de origen. Tiene un fuerte entronque con los sectores industriales punteros. Su nombre procede de la palabra griega "isos", igual. Recibe el mismo nombre independientemente del idioma en que se cite. (Fuente:www.iso.org-2007)

(2) Los especialistas de un ámbito determinado del conocimiento comparten un dominio cognitivo idealizado que no es compartido por el hablante general ni por especialistas de otros ámbitos del saber. Es en este contexto donde los términos existen como tales. (Fuente: TSS 2006)

- según quien lo crea (técnico, investigador, etc.)
- según el usuario técnico que comparte el idioma de creación
- según el usuario técnico que no comparte el idioma de creación
- según el terminólogo (lingüista) que debe decidir cómo se incorpora a la lengua de destino.

En los dos primeros casos se trata de un proceso natural: el término surge de la necesidad de dar nombre a lo que no lo tenía, y se acepta como algo propio por el destinatario. En algunos campos, por ejemplo en informática "overflow", no plantean mayores problemas, fundamentalmente por la fácil asociación mental entre nombre y objeto. En otros se puede dar la circunstancia de que varios términos que designan el mismo objeto coexistan en el idioma de origen ("USB memory", "pendrive", "memory stick", "flash memory") sin que ninguno se imponga sobre los otros, al menos a corto plazo.

Son los casos de los usuarios no nativos los que sí dan lugar a situaciones complejas. El usuario especialista cuya lengua materna no coincide con la que ha originado el término se encuentra ante una triple opción:

-aceptar el término tal como le llega –*overflow*

-traducirlo literalmente a su propia lengua- "desbordamiento"

-crear un término nuevo en la lengua de destino- "saturación del sistema o de la memoria".

Recordemos que estamos hablando de campos de investigación de materias en constante evolución, y el tiempo no es habitualmente un factor menor; podemos hablar incluso de una cierta urgencia en esta situación, lo que lleva a que los resultados finales no sean muy deseables en ocasiones (en España resulta llamativo el uso de los más variados términos en las décadas de los 80 y 90 por parte de la emergente "casta de los informáticos", que dio lugar a creaciones de tipo "lincar" por "link", "alimentar datos" por "feed data", "comandos" por "commands" y otros ejemplos de lo que podríamos denominar

seudo-jerga técnica; soluciones inmediatas que, a la larga, terminan por desaparecer puesto que la lengua de destino tiene recursos suficientes para dar respuesta a la necesidad planteada (“conectar”, “introducir datos”, o “instrucciones”, en este caso).

El lingüista tiene un problema de partida: no es especialista (habitualmente) en el campo de origen del término, por lo que precisa de asesoría técnica que variará dependiendo del campo en que esté ocupado en cada circunstancia. En el proyecto que nos ocupa trataremos de abordar el análisis del glosario terminológico desde el punto de vista del técnico exclusivamente: el especialista que trabaja habitualmente en este caso con terminología de estadística e informática y, más concretamente, de minería de datos. La razón de este proceder está en la estricta coherencia con lo expresado anteriormente: dado que es un campo sumamente especializado, los usuarios del glosario final no serán lingüistas, sino técnicos, y la información que habrán de precisar será aquella referente a conceptos básicos de su área de conocimiento.

Trataremos de recopilar un glosario trilingüe inglés-español-alemán; inglés por ser la fuente originaria de la práctica totalidad de los términos que se han creado en el campo de la minería de datos, y español y alemán por tratarse de lenguas productivas en esta materia, con distintas soluciones lingüísticas (por su propia naturaleza estructural) para la producción y/o adaptación de términos y, en ambos casos, con un gran número potencial de usuarios del glosario resultante. Conviene llegados a este punto hacer una precisión respecto al título de esta tesis: según fuimos profundizando en el conocimiento de los principios que rigen los estudios terminológicos, nuestra intención inicial de abarcar los campos de la estadística y la minería de datos se mostró desacertada. A lo largo del estudio práctico, la minería de datos se reveló como una ciencia independiente de la estadística, y no parte de ella según era nuestra impresión inicial.

Por otro lado, la creación de un glosario sobre estadística habría supuesto una mera revisión y actualización de lo ya existente. No era éste el caso de la minería de datos. El glosario resultante sería el primero que tratara sobre esta materia, y no sólo en español y alemán, sino que, según pudimos constatar a lo largo del proyecto, sería el primer glosario terminológico creado siguiendo métodos normalizados sobre terminología de minería de datos en inglés.

2- LA TERMINOLOGÍA COMO CIENCIA CON ENTIDAD PROPIA

Ha habido varios intentos de encontrar una ubicación exacta para la terminología dentro de la lingüística general. En unos casos, como el de Cabré, se llega a hablar de ciencia claramente diferenciada de la lexicología. En otros, y es corriente mayoritaria, la relación es de pertenencia: terminología como parte integrante de la lexicología.

A continuación revisaremos algunos de los enfoques teóricos comúnmente aceptados como más destacables sobre este campo, representados por María Teresa Cabré, Sonneveld y Loening, Reiner Arntz, Silvia Pavel y Kio Kageura. La mención a las teorías de estos terminólogos no está sujeta a filtro alguno; aparecen tal cual las sustentan sus autores y son objeto de comentario por parte de este autor en cada apartado individual y posteriormente en el apartado 2.6 Conclusiones.

2.1 M.T. Cabré

Según María Teresa Cabré (1996:15) la terminología está adquiriendo entidad propia como disciplina, si bien se mantienen las reticencias de los lingüistas en torno a esta cuestión. De acuerdo con la aproximación al tema

que preconiza esta autora, la terminología ha de estudiarse desde una triple óptica:

- su conceptualización
- tendencias actuales
- sus múltiples y diversas aplicaciones

La terminología representa sobre todo diversidad. Diversidad de conceptos sobre su propia esencia, diversidad de campos y de funciones que permite desempeñar. Diversidad de usuarios y de organizaciones que se benefician de ella.

No es este concepto de terminología el rígido centroeuropeo, sino más bien uno flexible, adaptable a cada medio y adecuado a metas específicas. Terminología es pluralidad, diversidad y multifuncionalidad.
Tras esta diversidad Cabré (1996:17) señala no obstante la existencia de una unidad de bases, de fundamentos científicos y de campo de investigación: unidad de disciplina.

> "Es una única disciplina con una forma poliédrica de fundamentos (conceptos) enfoques (orientaciones) y práctica (aplicaciones)".

Al definir los conceptos, Cabré distingue tres significados de terminología según consideremos la materia en sí, su práctica o el producto resultante de la misma.

Como materia, es la disciplina que se ocupa de términos especializados. Como práctica, el conjunto de principios orientados a la recopilación de términos. Finalmente, como resultado es el conjunto de términos de un determinado campo temático.

La autora se plantea entonces una serie de cuestiones sobre estas definiciones, comenzando por el concepto de "término especializado". Según la

definición de término proceda de la lingüística, la filosofía o las disciplinas científico-técnicas encontraremos diferencias en cuanto a las perspectivas del marco de referencia, la perspectiva desde la que se enfocan los términos y la función principal atribuída al término. Pero podemos concluir que la terminología es el mismo objeto para las tres disciplinas, puesto que en cualquier caso las tres coinciden en considerarla como el conjunto de términos -que se concibe como una unidad de significado triple (cosa-nombre-significado)- que hacen referencia a la realidad especializada.

Para Cabré la terminología es una disciplina en tanto que, al igual que cualquier otra materia, presenta aspectos teóricos y prácticos y genera aplicaciones definidas.

Su modelo teórico coincide con el de la lingüística; en este sentido, no es original, pero sí lo es en otros dos:

- selecciona de los temas objeto de su estudio algunas bases específicas y descarta otras. No toma todos los aspectos de la lingüística, ni siquiera de la lexicología –por no decir de la morfología o semántica léxica-
- reconfigura estos fundamentos teóricos para crear su propio espacio original diferenciado en cuanto a estructura, objeto, método y objetivos.

Siempre desde la perspectiva de Cabré, la terminología difiere de la lingüística en teoría y práctica. En cuanto a teoría, difieren en aspectos tan cruciales como la conceptualización del lenguaje, la conceptualización del objeto de estudio por parte de la disciplina, la perspectiva de estudio de la materia e, incluso, las metas principales.

Como práctica, o más bien como disciplina aplicada, la terminología difiere de la lexicología en cuanto a la metodología de los siguientes puntos:

- los métodos de recopilación
- el tratamiento de la información

- la presentación de dicha información en forma de glosario
- el concepto de lenguaje

La lingüística se centra en el lenguaje desde una muestra idealizada de hablante y así intenta explicar la competencia. La terminología se centra en el lenguaje real –toma su información de la documentación- y da cuenta de las denominaciones especializadas.

Como disciplina aplicada, la terminología y la lexicología son claramente distintas (aquí cabré sigue a Wüster -véase pie de página 5 en página 23-) con respecto a la formación de términos y a su idea de lenguaje. En cuanto al lenguaje, la lexicología –como parte de la lingüística que se ocupa del léxico- no concibe el significado independientemente de la palabra. La terminología, por el contrario, se fija en el concepto –su principal objeto de estudio- como fundamental y prioritario, y lo concibe de forma independiente del término que lo designa.

Para la lexicología los términos son relevantes en cuanto que son parte del discurso, mientras que para la terminología los términos son relevantes por sí mismos, sin referencias a su inflexión (que les dá la morfología adecuada en el contexto) o a su sintaxis (que los sitúa en el contexto gramatical adecuado).

La lingüística distingue e incluye los aspectos sincrónicos y diacrónicos de las palabras, mientras que la terminología convencional sólo se preocupa de la sincronía de sus unidades.

La lingüística general –y dentro de ella la lexicología- apoyan una evolución libre del lenguaje y por tanto descartan la intervención. Se oponen a la normalización –selección de un término como más adecuado que otros-. La

terminología no evita la intervención, dado que una de sus aplicaciones consiste en el establecimiento de formas normalizadas.

La terminología concibe los términos en un sentido internacional y da por tanto prioridad a aquellas maneras de formación (Cabré 1996:21) que acercan las lenguas históricas unas a otras. Este fenómeno lleva a adoptar criterios internacionales para la formación de palabras, y sistemas de trabajo cuya validez trasciende cada lenguaje particular. Debido a esto, la terminología atiende a la formación de palabras fundamentalmente basándose en elementos tomados del latín y el griego.

En este contexto de internacionalización, la terminología interviene exclusivamente en la formación escrita de las palabras –tanto en su forma completa como en sus variantes y abreviaturas- y no tiene en cuenta la pronunciación, que es una de las principales áreas de interés de la lingüística.

En cuanto al objeto de estudio, la lexicología se encarga del estudio de las palabras; la terminología, de los términos.

La lexicología se ocupa de la competencia léxica del hablante, el vocabulario que domina, cómo crear palabras y cómo emplearlas adecuadamente. La terminología se ocupa sólo de los términos o palabras de un campo especializado (física, antropología, etc) o dominio profesional (industria, comercio, deporte, etc).

La terminología –continúa Cabré (1996:22)- no es parte de la lexicología dado que palabras y términos no son lo mismo. Desde un punto de vista lingüístico, los términos pertenecen al campo de las palabras, pero desde un punto de vista terminológico esto no es así. Las diferencias se pueden apreciar realizando un estudio que compare un glosario terminológico con las palabras de un diccionario.

Los métodos de formación (antes mencionados por Cabré como maneras de formación) de términos no tienen la misma frecuencia que los de formación de palabras. En terminología las unidades compuestas por formas de expresión aprendidas y estructuras de frase fija se producen con más frecuencia que en lexicología. Las formas de expresión morfológicas y las reglas de formación léxica son, no obstante, las mismas. Pese a ello, la presencia de elementos de palabras latinas y griegas y la frecuencia de estructuras frasales en terminología sugieren una cierta diferenciación.

En un estudio terminológico abundan los sustantivos. Los verbos, adjetivos, adverbios, determinantes, pronombres, etc, son raros. Las palabras son, además de lo dicho anteriormente, unidades de comunicación (pragmática) que identifican a los hablantes por la forma en que las utilizan en determinadas situaciones comunicativas. Los aspectos pragmáticos son los que nos permiten establecer una diferencia más clara entre términos y palabras:

-Difieren en sus usuarios en cuanto que los términos son usados por profesionales de los distintos campos, mientras que las palabras son usadas por los hablantes en general.

-Difieren en cuanto al tema que tratan puesto que en terminología hablamos de glosarios de términos especializados y no del vocabulario general de los glosarios léxicos.

-Difieren en su finalidad. El propósito de la terminología es meramente referencial. No está interesada en los usos expresivos, comunicativos, poéticos o metalingüísticos del lenguaje general.

-Y difieren finalmente en las situaciones comunicativas en que se pueden encontrar y en los tipos de discurso: general frente a profesional y científico.

En cuanto a los objetivos teóricos y descriptivos también encontramos diferencias sustanciales. La terminología usa los términos para establecer una forma de referencia. La lexicología se centra en la competencia léxica del hablante.

La terminología busca proporcionar elementos teóricos y principios prácticos que puedan regir la búsqueda, selección y orden de los términos de un campo determinado a fin de normalizar su forma y contenidos. La terminología busca identificar claramente segmentos de una realidad profesional especializada. La labor de la terminología se orienta a denominar conceptos que pertenecen a una materia determinada.

En cuanto a los objetivos aplicados podemos reseñar también importantes diferencias. La lexicografía –rama aplicada de la lexicología- se ocupa de crear diccionarios; la terminografía –rama aplicada de la terminología- se ocupa de crear glosarios terminológicos o diccionarios especializados. El proceso de creación de ambos difiere en cuanto a método de trabajo y finalidad del mismo.

El lexicógrafo establece una lista de palabras que constituyen las entradas del diccionario. A continuación las describe semánticamente vía definición. Es un proceso basado en la semasiología: de forma a significado. El proceso terminográfico es inverso: va del concepto a su denominación.

En cuanto a la finalidad del trabajo, la terminografía busca la estandarización de los términos de un dominio especializado determinado. No tiene finalidad divulgativa o descriptiva únicamente, sino que pretende fijar las

unidades terminológicas como formas estándar. Crea la referencia que descarta otras variantes que denominan al mismo concepto. La meta última es el logro de una comunicación profesional precisa, moderna y carente de ambigüedades. El resultado final de un proceso de normalización terminológica es el fruto del acuerdo de una comisión de especialistas.

Terminología y lexicología difieren también en aspectos lingüísticos. Los terminólogos toman la documentación especializada como única fuente de materiales. Sus entradas son siempre lexémicas –de una o más palabras- y esta información se presenta de acuerdo con normas internacionales. Parte del concepto para llegar a su denominación. Por lo tanto, da una descripción detallada del objeto vía su definición, siendo ésta una definición de base descriptiva que a menudo muestra relaciones entre distintos conceptos. La terminología ordena sistemáticamente sus entradas, no alfabéticamente. Es un orden por conceptos, un enfoque más internacional.

Por lo que respecta a sus aplicaciones, la terminología tiene como principales metas la representación y la transferencia de conocimiento dentro del campo de la realidad especializada. Su función de representación sirve a tres tipos de disciplinas o actividades:

documentación,
ingeniería y lingüística ocupacional
y especialidades científico-técnicas.

Sirve a estas materias como herramienta y a la vez depende de ellas para constituir sus propios objetivos de trabajo. Los tesauros y sistemas de clasificación son básicamente inventarios temáticamente organizados y formalmente controlados. En ingeniería del lenguaje, terminología es la simulación del conocimiento. Cada término constituye una unidad conceptual, y el conjunto de términos de un dominio representa la organización conceptual de la realidad de ese dominio.

La terminología sirve a las distintas especialidades representando el conocimiento de forma organizada –en manuales y glosarios- y unificando el conocimiento –en estándares y normas-.

En cuanto a su función de transferencia, sirve como herramienta clave a los especialistas, quienes, sin los términos, no serían capaces de expresar y comunicar su conocimiento. Para los especialistas, la terminología está en la raíz del conocimiento especializado. La terminología es un apoyo fundamental a la estandarización de un lenguaje técnico que enriquece y moderniza las lenguas, convirtiéndolas en lenguas de cultura. (Aquí Cabré incurre a mi modo de ver en una contradicción flagrante; el lenguaje técnico no enriquece las lenguas en que se expresa, puesto que, en palabras de la propia autora: "los términos son usados por profesionales de los distintos campos, mientras que las palabras son usadas por los hablantes en general"; si la terminología pertenece al ámbito de la lengua especializada, no trasciende sus límites, y ese proceso de enriquecimiento y modernización es muy relativo, por no mencionar el hecho de las nuevas connotaciones que adquiere un término al ser empleado por un hablante no-especialista).

Estos intentos de Cabré de desligar la terminología de la lexicología quedaron, no obstante, sumamente matizados, o más bien superados, posteriormente por la propia autora. De este modo, no podríamos finalizar esta referencia al enfoque teórico de la terminología que aporta María Teresa Cabré sin hacer referencia a su Teoría de las Puertas (3). En ella propone una aproximación multidisciplinar a la terminología "que se ocupe de los términos y que integre los aspectos cognitivos, lingüísticos, semióticos y comunicativos de las unidades terminológicas", en la cual "el objeto término es una unidad

(3) Publicada con el título «Terminologie et linguistique: la théorie des portes», en la revista *Terminologies nouvelles (2000). Terminologie et diversité culturelle*, 21, p. 10-15.

formada por tres vertientes diferentes: una vertiente semiótica y lingüística, una vertiente cognitiva y una vertiente comunicativa". Es un enfoque que permite ver los términos desde una perspectiva más amplia que la que propone Wüster (véase nota 5), puesto que admite su uso por los no-especialistas al extremo de llevar la terminología al ámbito de lo cotidiano. De este modo, en el año 2005, en el Curso de Verano de terminología de la Universidad Pompeu Fabra (4), Cabré ya adelantaba una aceptación en consonancia con la corriente teórica predominante del concepto de terminología como parte (diferenciada, pero parte al fin) de la lexicología, acorde con su Teoría de las Puertas.

En el momento de cerrar estas líneas el último posicionamiento conocido de la autora que nos ocupa tuvo lugar en el XI Simposio Iberoamericano de Terminología RITerm 2008, celebrado en Lima en Octubre de este año y al que tuve la oportunidad de acudir como ponente. En dicho foro surgió la cuestión, (en forma de pregunta dirigida al panel de expertos que cerraban la sesión del día en una mesa redonda sobre buenas prácticas terminológicas), de cuál sería entonces la diferencia entre lexicología y terminología. Estando presente la Doctora Cabré, los expertos de la mesa en cuestión le remitieron la pregunta a ella en su calidad de autoridad máxima en la materia. La respuesta fue sin duda reveladora, pues vino a decir que las diferencias tienden a ser imperceptibles. La tesis que sostiene el presente trabajo difiere sustancialmente del enfoque de la Dra. Cabré en este punto, como tendremos ocasión de comentar posteriormente.

(4) IULATERM (2005) Materiales de la V Escuela Internacional de Verano de Terminología. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

2.2 Sonneveld y Loening

Sonneveld y Loening (1993) en su *Terminology; Applications in Inter-disciplinary Communication*, consideran la accesibilidad de la información como un factor clave en las empresas y centros de aprendizaje modernos. La rapidez en el acceso a la información ha de ir pareja con la facilidad de manejo de la misma. La precisión y la falta de ambigüedad en los términos científicos y tecnológicos es fundamental para el intercambio de información y comprensión mutua entre especialistas.

Es preciso que exista un vocabulario "controlado" en las bases de datos de las disciplinas que se estudien. Los avances en las diversas disciplinas han de ir acompañados por desarrollos en la terminología que emplean. La terminología y la terminología computacional ayudan a estos fines.

Terminología es, según la Norma ISO 1087:

> "Cualquier actividad que se ocupe de la sistematización y representación de conceptos o de la presentación de terminologías –conjuntos de términos que representan el sistema de conceptos de un campo de investigación determinado- basadas en principios y métodos establecidos".

Sonneveld y Loening nos ofrecen un estudio diacrónico de la terminología. Desde finales del siglo XIX en adelante se desarrollaron principios de denominación en campos tales como la química, la zoología, botánica, la medicina y las matemáticas. La industrialización dio como resultado la necesidad de comunicación entre países y, por ende, los primeros esfuerzos en el campo de la terminología. En particular, comenzó la estandarización de la terminología técnica y la regulación de la terminología científica.

Eugen Wüster (5) puso gran empeño en el desarrollo de los principios y métodos terminológicos. Estos principios y métodos aún forman parte de la base de la teoría y práctica terminológica actual. En un principio, los esfuerzos se encaminaban al desarrollo de vocabularios multilingües basados en teorías lingüísticas. Se buscaba presentar la equivalencia de los términos en dos o más idiomas, ya fuera con su definición o sin ella.

En los años setenta la investigación terminológica dio como resultado que la investigación ya no se podía basar únicamente en principios lingüísticos, sino que era de naturaleza esencialmente multidisciplinar. Como resultado, la terminología no se había de limitar a una mera recopilación de listas alfabéticas de términos, sino al estudio fundamental de conceptos y la ordenación de conocimientos, a la transferencia de conocimientos, la intermediación lingüística, la formulación de conocimientos científicos y técnicos, el almacenamiento y recuperación de información y la ingeniería del conocimiento.

Sonneveld y Loening concluyen que el campo de la terminología es aún joven; la discusión sobre sus características, metas y desarrollos permanece abierta. Su principal aportación, no obstante, es clara: proporcionar al especialista, ya sea técnico o científico, una herramienta de comunicación y transmisión de los conocimientos de su área de especialidad por medio de los términos que le son propios. Es un enfoque pragmático compartido por este autor: no es una ciencia teórica, sino una aportación de la lingüística a la

(5) Eugen Wüster (Wieselburg 1898 - Viena 1977) Lingüista austriaco considerado el padre de la terminología. Suya es la TGT, Teoría General de la Terminología, que sentó las bases teóricas de esta disciplina. Eugen Wüster comenzó su vida profesional como industrial en el campo de la electrónica. A principios de los años treinta creó un centro de estudios terminológicos como parte de su propia empresa. En 1951, se hizo cargo del Comité ISO/TC 37 “Terminology (principles and co-ordination)”. Infoterm fue creado en 1971 en base a un contrato entre la UNESCO y la empresa de Wüster, y, tras la muerte de este, continuó con su legado, que sería el Archivo Eugen Wüster de la Universidad de Viena. En 1996 Infoterm se convirtió en una asociación internacional independiente no lucrativa. (Fuente: www.infoterm.info)

transmisión de conocimiento especializado, cuantificable, con métodos de trabajo normalizados y por lo tanto sujetos a controles objetivos de calidad del producto final.

2.3 Reiner Arntz

Reiner Arntz en su artículo de 1993 *Terminological Equivalence and Translation,* afirma que un concepto sólo puede ser entendido en el contexto al que pertenece. Por lo tanto, antes de comparar dos idiomas, es preciso establecer o descubrir los sistemas independientes de conceptos que existen en cada idioma individual.

Para determinar las equivalencias de significado de términos de distintas lenguas es bueno comprobar si la relación jerárquica dentro de su campo semántico es la misma. Arntz propone el ejemplo siguiente: en alemán existe la relación jerárquica de población activa, con su subdivisión en población empleada y desempleados, dividiéndose estos a su vez en desempleados inscritos y desempleados no inscritos. Una vez que comprobamos que esta estructura se repite también en francés, queda así probada la equivalencia de los términos de ambos idiomas en este caso.

Cuando los términos de dos idiomas difieren considerablemente o cuando un término existe únicamente en un idioma, hay tres técnicas básicas para reproducir el término en el otro idioma:

- préstamos o “traducciones préstamo” del idioma fuente o de origen.

Es una solución indicada cuando el contenido del término es especialmente característico del área temática original y resulta por tanto difícil

de traducir. Así, tenemos palabras como *test, know-how, joint venture*... que llegaron inalteradas al alemán y a otros idiomas. La traducción préstamo (*contact lenses* – "lentes de contacto") puede facilitar la comprensión de un término desconocido en el área donde se habla el idioma destino.

- Acuñar un término en el idioma destino. Por ejemplo, de *nonproliferation treaty* surge en alemán *Atomwaffensperrvertrag* en vez de *Nonproliferationsvertrag*.

Es el mismo caso de "Tratado de no proliferación nuclear" en español. Tanto en alemán como en español, la palabra "nuclear" es un añadido que viene a clarificar un término que, de otro modo, habría podido resultar confuso.

- Crear una paráfrasis equivalente.

Es ésta una solución especialmente relevante en la práctica profesional del traductor técnico. Pone por ejemplo el término "brinkmanship": política arriesgada, suicida, que pone al país al borde de una guerra.

Arntz recoge la tradición de Wüster en el sentido de valorar la terminología fundamentalmente como medio de transmisión de conocimiento especializado (Arntz y Picht 1989:102-190). Habla de lexicografía terminológica, cuyo objetivo no es otro que reunir los resultados de la investigación terminológica para ponerlos al servicio de los usuarios por medio de la creación de diccionarios técnicos. Estamos nuevamente ante un enfoque pragmático similar al de Sonneveld y Loening que no entra a valorar interferencias de la lengua común en el lenguaje especializado o viceversa, sino que entiende ambos como pertenecientes a distintos niveles de habla que coexisten en paralelo sin llegar a cruzarse nunca.

No habla de una única lengua de especialidad, sino de varias que, no obstante comparten una misma terminología fundamental de referencia. En este sentido, introduce el tema de lo que podríamos llamar "globalización de la terminología": es el enfoque basado en el concepto como núcleo fundamental

de la terminología. La lengua de expresión no es obstáculo si el sistema de conceptos es compartido.

2.4 Silvia Pavel

Silvia Pavel (1993) en su artículo *Neology and Phraseology as Terminology-in-the-making* basa su estudio sobre la neología en las lenguas para fines específicos en la distinción de Holton (1988:405) entre *science in the making* y *science as institution*. Básicamente, se trata de lo primero, más concretamente del paso de variación semántica a estabilidad conceptual.

El conocimiento científico heredado se caracteriza por formar redes semánticas cuyos nodos representan conceptos conectados por enlaces estables. La evolución en el conocimiento implica ajustes en estos nodos y enlaces, y los cambios conceptuales importantes suponen la reestructuración y sustitución de redes conceptuales enteras. La aparición del término “oxígeno” es una muestra de ello. Lavoisier lo acuñó tras varias fases de investigación y sucesivas redenominaciones.

Pavel habla de encontrar los principios que rigen la elección de unos términos frente a otros como una de las principales tareas de los terminólogos. Esta tarea se explica mediante tres tipos de procesos, siguiendo el enfoque sociológico de Turner (1988): procesos motivacionales, procesos interactuales y procesos de estructuración.

En una comunidad científica o profesional, los procesos motivacionales se pueden contemplar en términos de necesidades y valores: la necesidad de expresar conocimiento emergente por analogía con hechos o experiencia

comúnmente aceptados, la necesidad de nombrar y renombrar objetos para un uso específico o un interés común, la necesidad de identificarse y ser reconocido por un grupo a través de actividades culturales y valores compartidos. La primera preocupación de la investigación terminológica en la fase de identificación de conceptos debería ser localizar los temas centrales que atraen la atención del especialista, las tradiciones culturales responsables de sus patrones de pensamiento; los modelos, analogías y mecanismos que usan para captar atributos conceptuales. Cada ciencia se distingue de las demás no sólo por el conjunto de fenómenos objeto de su estudio, sino también por el enfoque que adopta. De cada enfoque podemos obtener un modelo del cual extraer consecuencias prácticas y usos. De este modo, si vemos al hombre como un actor cuyo proceso de pensamiento interno no puede ser analizado, entonces seremos considerados "psicólogos del comportamiento", y estudiaremos el comportamento humano. Si vemos al hombre como un cerebro, una máquina conectada por neuronas, entonces nos llamarán "biólogos", y estudiaremos las respuestas neuro-fisiológicas, etc.

Los cambios conceptuales y terminológicos –sigue Pavel citando a Hayles (1991)- se producen por medio de:

> "negociaciones en múltiples lugares entre aquellos que generan la información, la interpretan y la extrapolan para llevarla a alcanzar un significado cultural y filosófico más amplio."

Estas negociaciones se producen, por parte del creador de los nuevos términos, a través de simposios, charlas informales y publicaciones, para llegar así a los otros actores, (colegas, traductores, editores, credores de vocabulario y usuarios), de forma que éstos puedan interpretarlos y reaccionar ante ellos. El creador del nuevo término intenta convencerlos y que su creación prospere. La acuñación del término "boojum" en física para reflejar la cualidad de lo que se desvanece suave y repentinamente ejemplifica lo anteriormente dicho. El autor del término se basó en "The Hunting of the Snark", de L. Carrol, en el que

una variedad de Snark –el Boojum- hacía precisamente eso ante cualquiera que lo encontrara.

El proceso de creación de nuevos términos puede también seguir el modelo del RINT (6) francés; todos los nuevos términos se envían al "Termium", banco de datos terminológico canadiense. Allí se sistematizan en registros terminológicos y se hacen llegar a diversos especialistas interesados en la materia, los cuales a su vez incorporan y retornan sus comentarios al Termium. Luego se integran y preparan para ser publicados y puestos a disposición de los hablantes de francés. (Se entendería entonces que aquí Pavel hace referencia a los hablantes especializados, no al hablante común).

Los terminólogos y traductores han de ser creativos; no se trata de crear algo a partir de nada, sino de estructurar la mente con elementos nuevos y los que ya disponemos para llegar a una conclusión nunca antes formulada. Para Pointcaré el proceso de creatividad suponía cuatro fases: preparación, incubación, retrospección y verificación. Para Koestler (1964) consiste en asociar dos matrices conceptuales que no son a menudo asociadas, y que incluso podrían parecer incompatibles. Cuanto más infrecuente la asociación, más creativo el resultado. Se trata de "ver una analogía que nadie antes vió", o generarla vía imaginación. La creatividad del terminólogo es más propia del artista que transfiere, interpreta, y adapta, que del inventor.

Desde nuestro punto de vista esta es una visión idealizada del terminólogo, al presentarlo como un individuo aislado que se ve en la necesidad de dar nombre a lo que antes no lo tenía. Es esta una visión más propia de siglos pasados que de la actualidad, puesto que las nuevas creaciones técnico-científicas actuales suelen más bien ser fruto de complejos

(6) RINT : **Réseau International de Néologie et de Terminologie.** Organización de países francófonos dedicada al desarrollo terminológico y la cooperación en materia lingüística. Su sede está en Québec.

proyectos de investigación en los que están integrados equipos de técnicos –con frecuencia multidisciplinares- a los que se les asignan diferentes funciones; la aparición de neologismos en estos equipos de trabajo es fruto de negociaciones internas forzadas por la necesidad de entendimiento, y el fruto de esas negociaciones suelen ser términos que, una vez recogidos en un informe final, se verán posiblemente modificados atendiendo a criterios de diversa índole, pero fundamentalmente culturales y comerciales.

2.5 Kio Kageura

Si los autores mencionados hasta ahora trataban de ubicar la terminología en una posición de mayor o menor proximidad con la lexicología (inmediata para Cabré y en gran medida Pavel, y distante e independiente para Sonneveld y Loenning, Arntz y Picht), Kio Kageura en su obra del año 2002 *The Dynamics of Terminology. A Descriptive Theory of Term Formation and Terminological Growth*. Hace una de las más encendidas defensas de la visión de la terminología como ciencia con entidad propia, como campo de investigación independiente.

Citando a Bessé, Nkwenti-azeh y Sager (1997), Kageura define término como “unidad léxica que consta de una o más palabras que representan un concepto dentro de un dominio”. A continuación establece que la terminología sería entonces “el vocabulario de un *subject field*” (área/campo de conocimiento). Kageura lamenta que estos autores usen “palabra” en vez de “unidad léxica” al definir término, puesto que la finalidad única de un glosario terminológico es distinguir términos de palabras. Propone también usar “conjunto de términos” en vez de vocabulario. Resalta, no obstante, que

"término" designa a un elemento individual, y "terminología" hace referencia al objeto colectivo.

Continúa estableciendo las definiciones de otros términos que va a emplear constantemente, como son:

Concepto.

> "Unidad abstracta que comprende las características de una serie de objetos concretos o abstractos que son seleccionados siguiendo unos criterios específicos, científicos o convencionales, adecuados a un dominio".

Estructura conceptual.

> "Representación de la estructura de los conceptos que pertenecen a un área específica del conocimiento o dominio"

Característica.

> "El elemento semántico que junto con otros constituye la intension de un concepto"

Dominio.

> "Área específica del conocimiento, disciplina, proceso de producción o método en el cual se usa un concepto".

Subject field (Área específica del conocimiento).

> "Campo del saber que se establece con el fin de agrupar en categorías convencionales los conocimientos relacionados entre sí".

Kageura admite que la definición de término siempre ha sido polémica. Fue objeto de un número monográfico de "Terminology" (Vol. 5 Nº1, 2000). Cualquier definición fuera de un contexto específico es siempre controvertida y provisional. (Es por este motivo que en nuestro trabajo recurrimos a las definiciones que marcan las normas ISO).

Al definir terminología como el conjunto de términos de un área específica del conocimiento, los términos quedan situados en el reino del habla, la expresión del lenguage, como opuesto a lengua, el sistema del lenguaje. Citando a Sager afirma que, como signos lingüísticos, los términos son una clase funcional de elementos léxicos. Aquí usa “funcional” en el sentido de función comunicativa, no sintáctica.

A partir de aquí podemos establecer una clara distinción entre términos y palabras, puesto que una palabra puede ser reconocida y definida a muchos niveles ligüísticos, incluidos el fonológico, ortográfico, gramatical y semántico. Así pues, pasa a definir palabra:

> Semánticamente: una de las unidades más pequeñas de significado aislado totalmente satisfactorio (Sapir 1921:34)
>
> Ortográficamente: una unidad que, al ser impresa, está limitada por espacios en blanco a ambos lado. (Bauer 1983:7)
>
> Fonológicamente: Unidad limitada por puntos sucesivos en los que es posible establecer una pausa (Hockett 1958:166)
>
> Gramaticalmente: Una mínima forma libre (Bloomfield [1933] 1984:178).

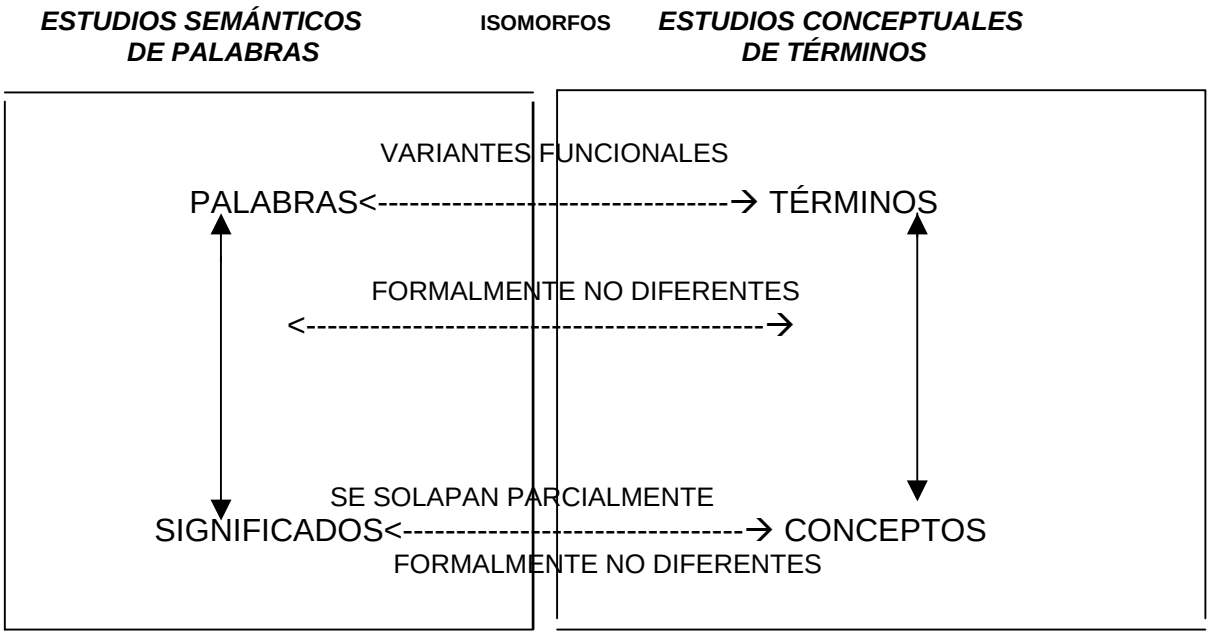
Hay una serie de conceptos que ya resultan tradicionales en el análisis de la terminología, como la independencia de la terminología. Desde Wüster (1959-60) pasando por Felber y Pitch (1984), muchos autores han defendido la existencia de la terminología como una ciencia con entidad propia, con su propia teoría y métodos.

Felber y Pitch ya hablan de la relevancia del concepto, de los términos de conceptos y la perspectiva siempre sincrónica del lenguaje. El concepto es

la “piedra angular” de la teoría general de terminología (Felber 84:102) y el punto de partida del trabajo terminológico.

Kageura se distancia del enfoque de la lingüística tradicional por no considerar específicamente el estatus de concepto dentro de la teoría de la terminología; lo que es crucial para Kageura es incorporar “concepto” al marco teórico de la terminología.

A partir de la evolución de estas premisas, Kageura establece un isomorfismo entre los estudios semánticos de las palabras y los estudios conceptuales de los términos, como se ilustra en la siguiente figura:



A continuación Kageura se centra en investigar las relaciones que rigen los patrones de especificación conceptual y las categorías conceptuales. El autor establece cuatro grandes categorías conceptuales: entidad, actividad, cualidad y relación, relevantes todas ellas en el estudio de la terminología de documentación japonesa. A cada categoría le atribuye varias subcategorías, creando un sistema conceptual con estructura jerárquica en el que se

distribuyen los diversos términos y morfemas extraídos de las bases de datos sobre términos de documentación.

Al asignar morfemas y términos a un sistema conceptual, crea las bases para describir combinaciones conceptualmente motivadas en la formación de términos. No se trata de dar explicaciones detalladas de los conceptos representados por los términos o morfemas, sino de proporcionar un marco estructural en el cual poder describir las estructuras de formación de términos. Kageura ofrece datos en cuanto a las tendencias cuantitativas de las categorías y subcategorías, identifica las relaciones intra-término y los patrones de especificación conceptual. La elaboración de relaciones intra-término se fundamenta en las relaciones propuestas por Pugh (1984) y en un análisis refinado basado en la aplicación a varios ejemplos. Estas relaciones son reinterpretadas y agrupadas en diferentes estructuras de especificación conceptual. Tanto las categorías conceptuales, como las relaciones intra-término y las estructuras de especificación conceptual aparecen reflejadas por un sistema de abreviaturas.

Kageura describe y evalúa las estructuras conceptuales en la formación de términos de documentación. Cada categoría conceptual está caracterizada por la frecuencia de los términos a ella asignados (distingue entre términos simples, de dos y de tres elementos), por las estructuras conceptuales de esos términos (término núcleo, determinante, relación intra-término) y por las estructuras de especificación dominantes.

Las distintas categorías conceptuales son agrupadas en función de que tengan estructuras comunes de especificación; de este modo, pertenecen al mismo grupo las categorías: “gente”, “organizaciones”, “máquinas” y “software”, puesto que comparten la característica de “ser activos”. Partiendo de la base de que la formación de nuevos términos está determinada por factores sistemáticos dentro de una terminología determinada, el autor establece que se

pueden hacer predicciones respecto a los esquemas de formación de nuevos términos en un dominio en particular, en función de las estructuras de formación de términos ya conocidas.

Kageura llega a describir un modelo matemático, llamado interpolación y extrapolación binomial, mediante el cual se pueden generar predicciones de las estructuras/patrones de crecimiento de una terminología determinada. Aplicando este modelo a la distribución de morfemas dentro de distintas categorías y subcategorías, el autor obtiene curvas de crecimiento de morfemas como consecuencia de la aparición de términos. De este modo, el crecimiento de morfemas de una categoría o subcategoría se puede predecir en base al incremento de los términos de esa categoría.

Luego sigue un análisis por secciones de cada categoría conceptual, explorando los patrones de crecimiento de los distintos elementos que intervienen en la formación de términos (núcleo, determinantes, categorías determinantes y estructuras de especificación). En cada caso se obtienen tablas que muestran las cantidades básicas de cada uno de estos elementos que se obtienen para los términos de una categoría determinada. Las curvas de desarrollo de los distintos tipos (núcleo, determinantes, etc.), y la tasa de crecimiento se muestran en diferentes gráficas en las que cada curva muestra el número de términos de una determinada categoría conceptual.

El inconveniente que plantea la obra de Kageura es que en muchos casos sus conclusiones se centran en la morfología del japonés, no siempre exportable a otras lenguas. Resulta a nuestro modo de ver utópico pretender excluir los factores subjetivos (culturales, afectivos, comerciales, etc.) que intervienen en el proceso de creación de nuevos términos.

2.6 Conclusiones

Podríamos seguir dando referencias bibliográficas que ilustren la situación actual de la teoría terminológica, y alargar este capítulo citando autores como (7) Dubuc (1985), Felber (1984), Rondeau (1984), que, sin duda, más de un terminólogo considerará imprescindibles, pero no es tal el objeto de este trabajo.

Consideramos suficiente la visión que aportan los anteriormente citados, puesto que representan lo que podríamos llamar las dos grandes corrientes actuales en la ciencia que nos ocupa: la que considera la terminología como una vía de entrada de nuevo vocabulario en la lengua de destino (Cabré, Pavel) y aquella que estima que terminología y lexicología cubren necesidades diferentes de diferentes destinatarios, sin llegar nunca a interferir la una con la otra (Sonneveld y Loening, Picht, Arntz y Kageura). Ambos enfoques han dado lugar a la aparición de dos grandes corrientes o escuelas teóricas dentro del panorama de la terminología contemporánea, la corriente francófona y la corriente centroeuropea, dotando de una dimensión interlingüística e internacional a lo que antes era meramente intralingüístico. Pese a coincidir en gran medida en los aspectos técnicos (sistemas de tratamiento informatizado de la información y plasmación de resultados en glosarios), difieren fundamentalmente en el propósito y la metodología; la escuela francófona es prescriptiva y teorizante, mientras que la centroeuropea es descriptiva y pragmática.

(7) Robert Dubuc (1985) *Manuel Pratique de Terminologie.* Helmuth Felber (1984) *Terminology Manual.* Guy Rondeau (1984) *Introduction à la Terminologie.*

En el caso francófono el método de trabajo implica la existencia de un "comité de sabios" que recopila los nuevos términos, los analiza y propone sus equivalentes en la lengua de destino. Una vez superados todos los pasos del proceso, los glosarios resultantes son de obligado uso por los especialistas del área de conocimiento de que se trate. Es una perspectiva "defensiva", podríamos decir; busca conservar la pureza de la lengua de destino, evitando que se vea contaminada por barbarismos, calcos, préstamos, etc. El terminólogo abre camino al técnico en cuanto que le indica qué término usar en cada caso.

El enfoque centroeuropeo se basa en la recopilación *a posteriori* de los términos en el idioma de destino, una vez que se constata su existencia en los medios de divulgación científica habituales (ponencias en congresos, artículos, libros, etc.). Recoge la tradición pragmática de Wüster y Felber, y su labor al frente del ISO/TC 37(8). La labor del terminólogo es reunir datos sobre lo que ya existe, y cerciorarse de la exactitud del glosario resultante para que pueda servir de referencia a los que se inician en el estudio del área de conocimiento de que se trate. No hay una preocupación manifiesta por la pureza del lenguaje: prevalece el criterio de precisión, aunque ello implique la aceptación de préstamos, calcos, y barbarismos. El terminólogo sigue al técnico.

La corriente francófona y de las lenguas periféricas (o minorizadas, calificativo frecuentemente empleado en foros tales como la Asociación Europea de Terminología o en los propios cursos de la Escuela de Verano de

(8)The objective of ISO/TC 37/SC 4 is to prepare various standards by specifying principles and methods for creating, coding, processing and managing language resources, such as written corpora, lexical corpora, speech corpora, dictionary compiling and classification schemes. These standards will also cover the information produced by natural language processing components in these various domains. Standards produced by ISO/TC 37/SC 4 should particularly address the needs of industry and international trade as well as the global economy regarding multi-lingual information retrieval, cross-cultural technical communication and information management. Fuente: http://www.tc37sc4.org/what_is_Objective.htm

la Universidad Pompeu Fabra) ha puesto gran empeño en ampliar el alcance de la palabra terminología. Y digo “palabra” con plena conciencia de ello, pues los constantes intentos de flexibilización del objeto de estudio de la terminología hacen, a mi modo de ver, que esta corriente teórica caiga más bien en lo que se conoce tradicionalmente como lexicología.

De este modo, podemos encontrar numerosos ejemplos de casos en los que se considera glosario terminológico la recopilación de expresiones que se emplean en ámbitos no necesariamente científicos o técnicos, o de dudosa especialización, como puede ser el vocabulario relacionado con las instituciones penitenciarias (ejemplo del Curso de Verano U. P. F. 2005).

Así, se crean palabras nuevas en la lengua de destino para cubrir carencias existentes, y a estas creaciones se les denomina “términos” cuando, a lo sumo, se podría hablar de léxico con un ámbito de aplicación limitado, pero nunca lenguaje especializado. Sus usuarios no están, según establece la propia norma ISO 1087 en un campo de conocimiento específico, ni en la academia, ni en un ámbito técnico especializado.

La mejor forma de ilustrar cuán alejados están ambos enfoques entre sí no consiste, curiosamente, en analizar sus planteamientos teóricos. La clave no está en la lingüística, sino en la economía: en el modo de financiar los estudios terminológicos en cada caso. El modelo francófono está basado en la subvención pública, y, de este modo, existe una infraestructura estable de instituciones y personas (RINT, Termcat, etc.) cuya función consiste en proporcionar nuevo vocabulario al idioma local. El producto se crea para satisfacer una potencial demanda. Los estudios que preconiza el modelo centroeuropeo se financian con iniciativas puntuales de editoriales o empresas privadas. La certeza de la demanda crea el producto.

De este modo, y en consecuencia con sus propios planteamientos teóricos, vemos que las vidas profesionales de algunos terminólogos aquí citados son reflejo de sus distintas aproximaciones a la terminología: A parte de sus actividades académicas, M. T. Cabré dirigió el Termcat durante varios años, y Sonneveld y Loening crearon su propia empresa de estudios terminológicos (Topterm).

Si bien la finalidad última en ambos casos debería ser facilitar la comprensión entre especialistas, determinadas aproximaciones a la terminología, por demasiado amplias, corren el riesgo de diluirse para acabar cayendo en la lexicología general. Un ejemplo que ilustra esta situación: la consulta del mes (diciembre de 2007, fuente www.termcat.cat el día 13/12/2007) al sistema en línea de consultas del Termcat era, y traduzco literalmente:

- PREGUNTA: "cómo hemos de llamar en catalán al espumillón?"
- RESPUESTA: "la forma aprobada por el TERMCAT como alternativa catalana al término castellano espumillón es serrellet".

El hecho de considerar "espumillón" término hace que el ejemplo se comente por sí mismo. Es una palabra que se encuadra en un campo semántico concreto, y cubre un vacío de la lengua de destino, pero la acuñación de "serrellet" difícilmente podría considerarse como medio de transmisión de conocimiento especializado.

En palabras de la propia M. T. Cabré (1999:244):

> "Si bien es cierto que un esquema de globalización económica y cultural requiere necesariamente una uniformización en las formas de pensamiento y de expresión, la ampliación del conocimiento más allá de los círculos restringidos de los especialistas a través de la enseñanza y los medios de comunicación ha descontrolado el contexto en el que la normalización terminológica se desenvolvía. A este factor hay que añadir

> además la tendencia defensiva que ante la situación de uniformización han desarrollado las sociedades actuales, reivindicando el derecho a preservar su identidad. Esta paradoja ha producido una situación inicialmente contradictoria entre dos fuerzas opuestas (la unificación y la diversificación), pero cada vez más saludable, por cuanto he hecho nacer una diferenciación de situaciones de comunicación especializada ante las que los grupos deben definir los usos lingüísticos".

Es aquí donde, a mi modo de ver, difieren ambas corrientes de un modo más ostensible. Esta aproximación a la terminología que propone Cabré implica ampliar el objeto de estudio más allá del lenguaje especializado, llegando al ámbito de lo cotidiano. Las posibles causas de esta divergencia son múltiples (fundamentalmente histórico-políticas y, por ende, sociológicas) y su análisis y estudio no son el objetivo de este trabajo.

La situación en España es peculiar. Dado que coexisten varias lenguas oficiales junto con el español (9) en nuestro territorio nacional, se da la circunstancia de que las instituciones responsables de la lengua autonómica en Cataluña, Galicia y País Vasco siguen fielmente el modelo francófono, mientras que la Real Academia Española ha seguido siempre la vía del no-intervencionismo. Sobre este punto volveremos más adelante, en el apartado 4.

En consecuencia con lo expuesto en la introducción, este trabajo se decanta claramente por el enfoque centroeuropeo. Analizaremos la terminología desde el punto de vista del análisis y la recopilación de los términos propios de un campo específico del saber, como medio de

(9) Optamos por hablar de español y no de castellano por coherencia con el espíritu de este trabajo; nuestro idioma es universalmente conocido como español. No es intención de este autor entrar en polémicas que no existen más allá de nuestras fronteras. Para tener una visión general sobre la situación de la terminología en España véase: *La Terminología: Panorama Actual y Cooperación internacional* de las Dras. Rodríguez Ortega Y Schnell, de la Universidad Pontificia de Comillas, en www.acta.es/articulos_mf/37009.pdf

comunicación y transmisión de conocimiento entre los especialistas de ese campo. Será un proyecto de terminografía, cuyo objetivo consistirá en reunir los resultados de la investigación terminológica para ponerlos al servicio de los usuarios. Trataremos de recopilar lo existente y completar los vacíos, donde los hubiera, con las recomendaciones de los técnicos en minería de datos.

3. PLANIFICACIÓN Y REALIZACIÓN DE UN PROYECTO TERMINOLÓGICO

En este apartado comenzaremos por analizar cómo ha evolucionado la forma de presentar los datos terminológicos en los últimos veinte años, a través del análisis de algunos modelos característicos. Se trata de propuestas realizadas por terminólogos a partir de lo que parecía ser la corriente de investigación en este campo en su momento. Estos modelos dan como resultado final una ficha o entrada fraseológica distinta en cada década. Posteriormente plantearemos el método a seguir para realizar nuestro proyecto.

3.1. El modelo de Picht. Años 80.

Heribert Picht (1984) propuso un método de planificación y realización de proyectos terminológicos que seguía la norma ISO 919. Esta norma lleva por título “Guía para la elaboración de vocabularios sistemáticos”

El documento tiene la siguiente estructura:

1. comentarios preliminares
2. decisiones preliminares
3. documentación
4. datos terminológicos
5. extracción de datos terminológicos
6. desarrollo del sistema de conceptos
7. el trabajo terminográfico

3.1.1. Comentarios preliminares.

Es una norma pensada como una guía. Su finalidad es facilitar la organización del trabajo terminológico propiamente dicho, así como contribuir a aumentar el grado de homogeneidad y la calidad de los productos terminológicos en distintas organizaciones.

Se trata de un método diseñado para la elaboración de de vocabularios multilingües estructurados sistemáticamente. Limita el campo de aplicación de la norma a las ciencias naturales y a la técnica.

3.1.2. Decisiones preliminares.

Es necesario definir y delimitar con precisión el campo de estudio. Se debe decidir de antemano si se han de incluir o no conceptos de campos del saber colindantes. De no hacer esta delimitación se corre el riesgo de que el proyecto se desborde y sea necesario hacer ajustes a posteriori, con el consiguiente coste añadido por la pérdida de tiempo.
Como medios auxiliares de definición del campo se sugieren las clasificaciones especializadas (p.ej. la Clasificación Decimal Universal) y bibliografías clasificadas sobre vocabularios especializados, como las de la Unesco.

También recomienda decidir el número de conceptos que se pretende incluir en el vocabulario. La inclusión de un número muy elevado de términos en un único sistema no facilita su uso ni a colaboradores ni a usuarios, por lo cual es más aconsejable subdividirlo. En cuanto a las lenguas a incorporar, plantea la cuestión de cuáles deben ser las lenguas de las definiciones y en cuántas otras lenguas se deben indicar los términos equivalentes.

En el caso de las ciencias naturales y sus aplicaciones basta con una lengua de definición, normalmente una de las principales, como el inglés. Esto implica que no haya muchas incongruencias conceptuales en las lenguas de las que se incorporan términos equivalentes. Si se trata de ciencias sociales, la

situación es bien distinta, puesto que en estos campos los sistemas de conceptos muestran poca congruencia entre las distintas lenguas e incluso entre los distintos países del mismo idioma. El número de lenguas en que se indican sólo los términos equivalentes puede variar considerablemente, dependiendo de razones puntuales como el destinatario final (p.ej. los glosarios de la Unión Europea, que deben incluir todas las lenguas oficiales de la Unión) o razones de tipo económico.Los datos terminológicos a incluir en el vocabulario seguirán el siguiente modelo de ficha:

<table>
<tr><th colspan="3">DATOS INCLUÍDOS EN UN REGISTRO TERMINOLÓGICO</th></tr>
<tr><th>DATOS ASOCIADOS</th><th colspan="2">DATOS TERMINOLÓGICOS

RELACIONADOS CON EL CONCEPTO
DATOS DEPENDIENTES DEL LENGUAJE</th></tr>
<tr><td>DATOS BÁSICOS:

IDENTIFICADOR DEL REGISTRO
FECHA DE REGISTRO ORIGEN DEL REGISTRO CÓDIGO DE IDIOMA CÓDIGO DE AUTORIDAD FUENTE(S)
DESCRIPCIÓN BIBLIOGRÁFICA DE LA FUENTE</td><td>DATOS BÁSICOS:

CÓDIGO DE MATERIA DEFINICIÓN
EXPLICACIÓN
CONTEXTO
CÓDIGO DE CONCEPTO

DATOS NUMÉRICOS</td><td>DATOS BÁSICOS:

TÉRMINO
FRASE
SINÓNIMO
FORMA ABREVIADA
FORMA COMPLETA
(A MENOS QUE COINCIDA CON TÉRMINO)

FÓRMULAS
HIPERÓNIMO
HIPERÓNIMO GENÉRICO
HIPERÓNIMO PARTITIVO
TÉRMINO DE SIGNIFICADO MÁS RESTRINGIDO (TSR)
TSR GENÉRICO
TSR PARTITIVO
TÉRMINO COORDINADO(s)
PARTES CONSTITUTIVAS</td></tr>
<tr><td>OTROS DATOS:

NOTAS DE USO
CÓDIGO DE FIABILIDAD
RESTRICCIONES</td><td>OTROS DATOS:

ANTÓNIMO
GRADO DE EQUIVALENCIA ENTRE CONCEPTOS EN DISTINTOS IDIOMAS</td><td>OTROS DATOS:

TÉRMINO PREFERIDO
TÉRMINO PERMITIDO
TÉRMINO RECHAZADO</td></tr>
</table>

Este sería pues el modelo de ficha propuesto en 1984, en el que Picht considera indispensable incluir:

1. Fecha de elaboración y nombre del autor.
2. Números de serie. Indican la secuencia de los conceptos en el vocabulario. Pretenden simplificar la referencia tanto entre el índice alfabético y la parte sistemática del vocabulario como entre los distintos datos terminológicos entre sí, por ejemplo entre un término y la definición de otro concepto en que aparece dicho término.
3. Símbolo de clasificación. El símbolo contiene informaciones exactas sobre la relación entre el concepto en cuestión y el resto de los conceptos del vocabulario. El símbolo de clasificación permite la transformación de un sistema de conceptos expuestos gráficamente (por ejemplo mediante un árbol) en una relación clasificada. De esta manera constituye un suplemento a la definición. El símbolo puede constar de letras, números, signos o combinaciones de ellos.
4. Los términos que designan el concepto. Todos los términos que figuran en esta categoría deben corresponder exactamente a la definición, independientemente de que se trate de una, dos o más lenguas. Es el concepto y no el término lo que constituye la base de la unidad terminológica.
5. Los sinónimos. Si son verdaderos sinónimos, deben figurar en una categoría aparte de la anterior.
6. La explicación del concepto. Por lo general es una definición que puede ser ampliada mediante ilustraciones y ejemplos.
7. El término en el contexto. La función de esta categoría consiste en ofrecer al usuario las informaciones lingüísticas necesarias para poder situar *correctamente* el término en un texto. Correctamente significa aquí según las normas de la lengua profesional de un campo del saber. Por ejemplo, al indicar el verbo, adjetivo y preposición que se utilizan junto con el término: Girar/librar_una letra de cambio_sobre/a /a cargo de_persona/lugar

Apretar/aflojar_una tuerca.

8. Las notas. Pueden referirse a cualquier parte de la unidad terminológica, como su definición, sinónimos, información gramatical, etc.
9. Indicación de conceptos colindantes y sus términos. Esta información permite una orientación rápida sobre la vecindad conceptual sin tener que consultar otras partes del vocabulario. También recomienda indicar el antónimo.
10. Las fuentes.
11. El empleo de símbolos. Se recomienda el uso de los símbolos normalizados por la ISO. El autor nos remite a la ISO 636 sobre símbolos para lenguas y autoridades y la ISO 1951 sobre símbolos lexicográficos.

3.1.3. La documentación.

La norma ISO 919 distingue entre los siguientes tipos de documentación:

- Publicaciones terminológicas, como diccionarios técnicos.
- Tratados sobre problemas terminológicos.
- Manuales, enciclopedias, artículos, catálogos, etc.
- Clasificaciones.

Al elegir la documentación se deben tener en cuenta otros criterios aparte del tipo o clase de documentos. Estos criterios pueden ser:

- Año de edición del documento.
- Reputación profesional del autor.
- Pertenencia o no del mismo a una escuela determinada.
- Para quién está elaborado el documento.
- Si son originales o traducciones.

Aparte de estas fuentes escritas existe lo que se puede llamar la "documentación oral", es decir, los expertos a quienes se debe consultar para aclarar problemas de los que la documentación escrita no ofrece soluciones

satisfactorias. Una vez reunida la documentación accesible hace falta evaluar la misma, tarea para la cual también se debe consultar a los expertos.

3.1.4. La gestión de los datos terminológicos.

En este apartado Picht reflexiona sobre cómo el desarrollo de los ordenadores ha cambiado el trabajo lexicográfico y especialmente el terminográfico, terminando con las “operaciones a mano”. La última propuesta de ISO 919 con que trabajó en 1984 “hace referencia a este medio”.

3.1.5. La extracción de los datos terminológicos.

En esta fase de trabajo el terminólogo colecciona todas las informaciones posibles acerca de un concepto y sus términos. Para unir todos los datos y asegurar la homogeneidad de los mismos se utilizan fichas en las que se han indicado todas las categorías deseadas. La norma ISO 1149 regulaba entonces cómo elaborar una ficha que asegurara un mínimo de datos comparables entre sí. Se recomienda establecer una ficha para cada término en cada una de las lenguas a incluir. De este modo los presuntos sinónimos no figuran en la misma ficha.

3.1.6. La elaboración del sistema de conceptos.

Se elabora un sistema de conceptos provisional partiendo de la descripción y delimitación del proyecto y basándonos tanto en los conocimientos profesionales de los colaboradores como en las estructuras de conceptos inherentes al campo del saber.

Una vez elaborado el sistema provisional de conceptos, se enumeran los conceptos con números de serie. La enumeración de las fichas hace posible ordenarlas sistemáticamente. A continuación se reunen las fichas que lleven números idénticos, aunque estén redactadas en distintas lenguas. Tras este trabajo de reunir información sobre un concepto, empieza el análisis y la elaboración de la unidad. Los puntos centrales son:

- Equivalencia

- Sinonimia
- Formulación de una definición nueva o reformulación de una ya existente
- Comentarios acerca de la sinonimia,equivalencia, estilística, etc.
- Elaboración o elección de gráficos,ilustraciones, fotografías, etc.
- Términos propuestos para designar conceptos que no existen en una de las lenguas en cuestión.
- Modificación del sistema de conceptos como consecuencia de factores y conocimientos que hayan surgido a lo largo de la elaboración.

3.1.7. El trabajo terminográfico.

Una vez terminado el análisis terminológico y establecido el sistema de conceptos definitivo, empieza la redacción de manuscritos. En este trabajo se reúnen todos los datos en el orden y de la forma en que tendrán que aparecer en el manuscrito final que se entrega a la editorial.

A esta altura del trabajo se añaden a la ficha el número de serie definitivo y la anotación. El primer documento reunido de esta manera servirá de versión de discusión. Todos los colaboradores y expertos repasan el manuscrito y añaden sus comentarios y suplementos, a partir de los cuales se elabora el manuscrito final.

3.1.8. Modelos de colaboración.

3.1.8.1 Colaboración entre profesión y lengua.

Tres combinaciones fundamentales con variantes:

1. Terminólogo (lingüista) + uno o más asesores expertos.
2. Grupos de terminólogos (pluralidad de lenguas) + asesores expertos.
3. Grupo de expertos + un terminólogo como asesor en cuestiones lingüísticas.

3.1.8.2. Tipos de proyecto.

- Proyecto nuevo. Se elabora por primera vez un campo del saber.
- Ampliación del número de lenguas. Se añaden una o más lenguas a una terminología ya existente sin cambiar su estructura ni su volumen.
- Ampliación del campo elaborado. Se añaden nuevas partes todavía no elaboradas a una terminología ya existente.
- Proyecto de revisión. Se lleva a cabo una revisión terminológica de un campo del saber. (10)

Es importante determinar el número de lenguas que deben elaborarse simultáneamente, por sus implicaciones en la composición del grupo de trabajo y en el número de lenguas que debe dominar el terminólogo.

En cuanto a si el proyecto debe ser normativo o descriptivo, Picht se decanta por el segundo método, (distanciándose de la tendencia centroeuropea actual predominante, que establece un proceso descriptivo inicial que se torna prescriptito una vez fijada la terminología de que se trate en glosarios normalizados) salvo excepciones que no precisa. Al tratarse el trabajo que nos ocupa de una tesis doctoral, el modelo habrá de ser forzosamente descriptivo.

(10) Esta tesis estaría considerada entonces como un proyecto de revisión, si bien sucede que en ocasiones habrá que crear nuevos términos en español y alemán para designar conceptos que sólo se expresan hasta el momento en inglés. En cuanto al modelo de colaboración, estaríamos encuadrados en el modelo 1 dado que el autor actúa como lingüista terminólogo y cuenta con la colaboración de estadísticos y expertos en bases de datos.

3.2. Modelos previos a la norma ISO 12620. Años 90.

3.2.1.Maria Teresa Cabré (1993-282) propone su modelo de ficha:

"Una ficha terminológica estándar suele contener las informaciones siguientes:

- identificación del término
- término de entrada
- fuente del término
- categoría gramatical
- área(s) temática(s)
- definición
- fuente de la definición
- contexto(s)
- fuente del contexto
- remisión a términos sinónimos
- concepto de la remisión
- otros tipos de remisión
- concepto de cada tipo de remisión
- autor de la ficha y fecha de redacción
- notas para informaciones no previstas
- equivalencias en otras lenguas con indicación de la lengua
- fuente de cada equivalencia

...y si se trata de una ficha plurilingüe, deberá incluir las equivalencias denominativas del término en todas las lenguas de trabajo."

3.2.2. Otros modelos de los 90

Podemos tomar como ejemplo de modelo de finales de los años 90 el trabajo de Mª Belén Tercedor Sánchez (1999) propuesto en su Tesis Doctoral titulada *La Fraseología En El Lenguaje Biomédico: Análisis Desde Las Necesidades Del Traductor*.

Partiendo de una combinación de fichas propuestas por Gouadec (1994) y Battaner (1994) propone el siguiente ejemplo de entrada fraseológica en la ficha provisional:

UNIDAD FRASEOLÓGICA
Palabra clave
Variables

CAMPOS A LOS QUE PERTENECE
Campo 1
Campo 2
Campo n

FUNCIÓN/NOCIÓN

SIGNIFICADO METAFÓRICO: TOTAL/PARCIAL

CARACTERÍSTICAS DEL ENTORNO DE LA UNIDAD FRASEOLÓGICA
Localización

FUENTE
Especializada/semiespecializada/no especializada
Grado de ponderación de la fuente

FINALIDAD COMUNICATIVA
Registro

Tono
Intención
Connotaciones

LECTOR ESTEREOTIPO

VARIABLES ENCONTRADAS

UNIDADES FRASEOLÓGICAS RELACIONADAS
Sinonimia
Antonimia

Genérica
Específica

GESTIÓN DE LA FICHA
Autor:
Validación del experto
Fecha

Esta ficha contiene información sistemática y exhaustiva, no obstante lo cual presenta un problema según Tercedor, y es que no está ubicada dentro de ningún modelo de base de datos específica para la información fraseológica. La misma crítica podría hacerse extensiva a las propuestas de Cabré.

En la década de los 90 la implantación de la informática en todas las áreas del conocimiento, que ya avanzaba Picht (1984), es un hecho indudable que supera incluso sus previsiones.

Ya a finales de los ochenta un grupo de trabajo de la ISO bajo la dirección de A. Melby [Comité técnico 37 (TC37), grupo de trabajo 3 (WG3) subcomité 3] inició una investigación cuya finalidad era elaborar una norma con una propuesta de modelo de datos para información terminológica, con la particularidad de que el modelo había de ser compatible con el formato

MARTIF (11). Este formato consiste en una aplicación XML (Extensible Markup Language) para el intercambio de datos terminológicos entre distintas bases de datos.

La investigación terminológica auspiciada por la ISO apostaba entonces por la informatización de los archivos terminológicos existentes y futuros. El concepto de glosario terminológico pasa desde este momento a ser inseparable del de base de datos, teoría defendida con empeño por Tercedor.

Esta autora sostiene en su tesis (1999-59) que:

"Aunque ninguna gran compañía informática ha desarrollado aún un protocolo para el uso de MARTIF, pensamos que esto ocurrirá pronto, dada la exhaustividad del trabajo de investigación de este grupo (el de Melby) y su importancia en el campo de la normalización y la estandarización".

El problema reside en que, hasta la fecha y pese a las enormes ventajas que se le suponen, este protocolo continúa sin ser desarrollado. No obstante, la norma ISO 12620 apareció por fin en el año 2000, y mantiene la representación en formato MARTIF para todos los elementos de su menú de categorías de datos.

(11)Machine-Readable Terminology Interchange Format (MARTIF), also known as ISO (FDIS) 12200. 150 data categories are described for MARTIF in ISO (FDIS) 12620. Fuente: http://coral.lili.uni-bielefeld.de/~ttrippel/terminology/node82.html (28 MAYO 1999)

3.3. El modelo de la norma ISO 12620 (1999)

La norma ISO 12620 establece una clasificación de la información en diez secciones, que se agrupan en cuatro clases:

A- Clase término

1-Comprende los propios términos

Término: Forma de designar un concepto concreto en un lenguaje especial por medio de una expresión lingüística.

Los términos pueden ser palabras sueltas o estar formados por cadenas de varias palabras. No se deben confundir con las unidades fraseológicas, que combinan más de un concepto de un modo lexicalizado para expresar situaciones complejas. Así,

"Control de calidad" es un término, mientras que

"Cumplir los requisitos de calidad" es una unidad fraseológica,

no obstante lo cual, en algunas bases de datos las unidades fraseológicas son tratadas como términos.

B- Clase información relativa al término

2-Categorías de datos relativos al término

Tipo de término. Atributo signado al término.
Posibilidades admitidas:
Definición del Término.
Sinónimo.
Cuasi-sinónimo o sinónimo próximo. Incorpora una nota aclaratoria admitiendo la gran relatividad y subjetividad de la sinonimia.
Término científico internacional. Ej.: *Homo sapiens.*
Nombre común.
Internacionalismo. Término que tiene la misma representación fonética en muchos idiomas.
Forma completa.
Forma abreviada. Puede ser:
Abreviatura: de adjetivo, *adj.*
Forma corta: de grupo de los siete países más industrializados del mundo, a *grupo de los siete* .

Iniciales: De Encelofatía Espongiforme Bovina, a EEB.
Acrónimo. De Radio Detecting And Ranging a *Radar.*
Término truncado (clipped form). De *influenza* a *flu.*
Variante. *Catalogue* (BE) *Catalog* (AmE).
Forma transliterada.
Forma transcrita.
Forma romanizada.
Símbolo.
Fórmula.
Ecuación.
Expresión lógica
Categorías de administración de materiales.
Unidad fraseológica.
2.1.18.1 Colocación.
2.1.18.2 Frase lexicalizada.
2.1.18.3 Frase sinónima.
Texto estándar.

2.2. Gramática.
2.2.1 Parte del habla: sustantivo/adjetivo/verbo/otro.
2.2.2 Género gramatical: masculino/femenino/neutro/otro.
2.2.3. Número gramatical: singular/plural/dual/incontables/otros
2.2.4. Naturaleza animada.
2.2.5. Clase de sustantivo.
2.2.6. Clase de adjetivo.

2.3. Uso.
2.3.1. Nota de uso. Nota que contiene información sobre el uso habitual del término.
2.3.2. Uso geográfico.
2.3.3. Registro. Neutro/técnico/"de la casa" (terminología que se crea dentro de una empresa y raramente se usa fuera de ese contexto)/"bench level"-shop term/registro slang/registro vulgar.
2.3.4. Frecuencia de uso.
2.3.5. Cualificador temporal.Arcaico/desfasado/obsoleto.
2.3.6. Restricción temporal.
2.3.7. Restricción de propiedad. Trademark y trade name.

2.3.8. Formación del término.
2.3.9. Origen. Préstamos y neologismos.
2.3.10. Etimología.

2.4. Pronunciación.

2.5. Silabificación.

2.6. Separación por medio de guiones. (Hyphenation)

2.7. Morfología.

2.8.1 Elemento morfológico. La unidad que resulta de dividir las palabras en sus más pequeñas unidades con significado.

2.8.2 Elemento témino. Cualquier parte con significado lógico de un término más amplio. Ej.: *Inmuno depresor* .

2.9. Estatus del término.

2.9.1 Autorización normativa: Término estándar/preferido/admitido/rechazado/descartado/legal/regulado.

2.9.2 Cualificador de planificación del lenguaje: Término recomendado/no-estandarizado/propuesto/nuevo

2.9.3 Estatus administrativo.

2.9.4 Estatus de proceso: sin procesar/en proceso/finalizado.

2.10. Grado de sinonimia.

3-Equivalencia

3.1. Grado de equivalencia.
3.2. Falso amigo.
3.3. Direccionalidad.
3.4. Código de fiabilidad.
3.5. Comentario de transferencia.

4-Campo del saber

También dominio. Puede admitir varios niveles:

Nivel 1 de campo: enfermedad
Nivel 2 de campo: cáncer
Nivel 3 de campo: linfoma no- Hodgkins

4.1. Sistema de clasificación: La disposición de los conceptos en clases y su subdivisión para expresar las relaciones entre ellos.
4.2. Número de clasificación: Conjunto de símbolos, con las normas para su aplicación, utilizados para expresar clases y sus interrelaciones.

5-Descripción relativa al concepto.

5.1. Definición. Una expresión (statement) que define un concepto permitiendo diferenciarlo de otros dentro de un sistema.
5.2. Explicación.
5.3. Contexto.
5.4. Ejemplo.
5.5. Medios de ilustración no-textuales: figura/audio/vídeo/tabla/otros datos binarios.

5.6. Unidad. Relación con un valor de referencia según definición de una institución autorizada; una cantidad medida.
5.7. Rango.
5.8. Característica.

6-Relación de concepto

Conexión semántica entre conceptos.

6.1. Relación genérica.
6.2. Relación partitiva.
6.3. Relación secuencial: Temporal/espacial
6.4. Relación asociativa.(También llamada temática y pragmática)
Ej: Maestro-escuela/carretera-coche.

7-Estructuras conceptuales

7.1. Sistema de conceptos. Se utiliza conjuntamente con 7.2
7.2. Posición del concepto. La ubicación de un concepto en un sistema de conceptos.
7.2.1. Concepto más amplio genérico/partitivo.
7.2.2. Concepto superordinado genérico/partitivo.
7.2.3. Concepto subordinado genérico/partitivo.
7.2.4. Concepto coordinado genérico/partitivo.
7.2.5. Concepto relacionado.

8-Notas

También denominadas comentarios y observaciones.

9-Lenguaje de documentación

9.1. Nombre de tesauro.
9.2. Descriptor de tesauro.
9.2.1. Término superior: el descriptor de tesauro que representa el concepto de nivel superior en una relación jerárquica.
9.2.2. Término más amplio: el descriptor de tesauro que representa un término superordinado en una relación jerárquica.
9.2.3. Término más limitado: el descriptor de tesauro que representa un término subordinado en una relación jerárquica.
9.2.4. Término relacionado: Un término conectado con otro término mediante una relación coordinativa o asociativa.
9.3. No-descriptor: Término de un tesauro que no debe ser utilizado para representar un concepto, pero que hace referencia a uno o más descriptores que se deben usar en su lugar.
9.4. Palabra clave.
9.5. Encabezamiento de índice.

10-Datos administrativos

3.4 Conclusión

¿Qué modelo de ficha deberemos seguir entonces?

Teniendo en cuenta que existe ya una ficha estándar, la que fija la norma ISO 12620, deberemos elegirla por coherencia. Será una ficha adaptada a las necesidades de los usuarios del glosario final, es decir los futuros (y/o presentes) técnicos en minería de datos. De este modo, y según se recoge en el apartado 6.5, incorporará los datos siguientes campos:

- Término en inglés, español y alemán.
- Definición en inglés y español
- Sinonimia
- Origen del término (SF1) y ubicación en la jerarquía conceptual del mismo.
- Posición del concepto en minería de datos (SF2)
- Ejemplo de uso del término en un contexto real.

En cuanto al soporte informático, tradicionalmente se habla del uso de MARTIF (véase nota 11) que permite el intercambio de grandes ficheros entre ordenadores. Sin duda es muy útil cuando se trata de intercambiar ficheros informáticos muy voluminosos, pero no es éste el caso que nos ocupa. Teniendo en cuenta que el número de fichas que esperamos obtener no excede de las 150, optaremos por utilizar un formato estándar de documento Microsoft Word para cada ficha, por su facilidad de uso y gran exportabilidad. (12)

(12) Exportabilidad: término informático que designa la cualidad que tienen algunos lenguajes de ser entendidos por diferentes sistemas operativos. (N. Del Aut.)

4. LA TERMINOLOGÍA COMO PROBLEMA SOCIOLINGÜÍSTICO

A partir de lo tratado anteriormente, volvemos sobre el hecho constatado de la existencia de dos grandes corrientes teóricas en el mundo de la terminología actual:

- La francófona y de las “lenguas minoritarias” (o minorizadas)
- La centroeuropea o “global”

Aquí resulta particularmente relevante establecer la importancia del “estado de salud” de la lengua de destino: ¿se trata de una lengua minorizada? (aquellas en situación de defender su propia existencia como tal lengua que forma parte de una identidad cultural que, por razones de diversa índole, se ha podido ver amenazada) En este caso no es siempre la exigencia de precisión lo que mueve al especialista a usar un término determinado en la lengua de destino, sino la existencia de instituciones (como por ejemplo el RINT francófono o el Termcat en Cataluña) dedicadas expresamente a adaptar la terminología “exterior” de modo que no contamine o degrade su patrimonio cultural. Así, el especialista que trabaja en estas lenguas usará, al menos en su ámbito doméstico, el término que le viene dado ya por parte de dichas instituciones. Son modelos de creación de terminología dirigidos.

¿Es una lengua productiva de términos técnicos? El inglés es en la actualidad y con diferencia la lengua más productiva en lo que a nuevos términos se refiere. La hegemonía técnico-económica de los EEUU y el hecho de que sea el inglés la *lingua franca* de las ultimas décadas –lo cual lleva a que

incluso los investigadores cuya lengua materna no es el inglés presenten sus resultados en congresos donde ésta es la lengua empleada- son las causas fundamentales de esta situación. En la Unión Europea, con ser lenguas oficiales todas las de los países miembros, el inglés es la lengua de referencia en instituciones como el BCE (Banco Central Europeo). Es, en cualquier caso, la tercera lengua del mundo por número de hablantes nativos (fuente: http://es.wikipedia.org).

Si la lengua receptora del término es la de un país con una comunidad científica productiva y una industria tecnológica potente, con una larga tradición como productora de términos técnicos –el alemán, por ejemplo- admitirá con más naturalidad los préstamos y traducciones préstamo, puesto que no existe una postura "defensiva" previa. En este caso, que podemos denominar modelo centroeuropeo/anglosajón, la incorporación de los términos es el resultado de un proceso de decantación natural. La base –comunidad científica- crea –o asimila- nuevos términos que luego son recogidos en glosarios propios de cada especialidad.

No es este el caso de lenguas como el catalán, gallego o vascuence, que, pese a contar con una dilatada tradición literaria, no incorporan terminología técnico-científica con la misma naturalidad. Para ilustrar esta afirmación podemos traer aquí el caso del representante del gobierno de la Comunidad Autónoma de Galicia en su ponencia en la reunión de noviembre del 2006 de la AEFT celebrada en Bruselas (13).

Tal ponente, cuyo nombre no es preciso mencionar aquí, afirmó que una forma eficaz de potenciar y defender el uso de las lenguas minoritarias –el gallego en este caso- sería hacer obligatoria por ley la traducción de los

(13) EAFT III Terminology Summit. Bruselas, 13-14 noviembre de 2006. Sección 2: Major problems with minor languages.

interfaces de usuario de productos informáticos como el Windows de Microsoft. Su propuesta fue contestada por un asistente en el turno de preguntas haciéndole ver lo poco factibles que resultan dichas imposiciones gubernamentales a empresas privadas en países democráticos, donde rige una economía libre de mercado. Ante la insistencia del ponente, su interlocutor le recordó lo que ya había ocurrido con en el intento del Gobierno Vasco de traducir al vascuence el interfaz de Windows: tras una fuerte inversión económica resultó que el producto final apenas encontró respaldo entre sus destinatarios (los funcionarios vascos), que optaron por seguir usando el interfaz en español o en inglés. Puede parecer una anécdota, pero ilustra hasta qué punto el intervencionismo de algunas instituciones puede alterar el proceso natural de creación terminológica.

Por más que se someta la solución final al juicio de un “comité de expertos”, la creación de listados de obligado uso en el ámbito local no nos parece la mejor solución: las imposiciones no son compatibles con el principio de libertad de expresión del pensamiento científico, y nunca deben estar por encima del objetivo final del terminólogo: facilitar la comprensión entre especialistas en aras a una mayor difusión del conocimiento. La terminología prescriptiva sólo es aceptable tras un exhaustivo proceso descriptivo previo. La precisión y la exactitud del término, en cuanto que remita más claramente al concepto que designa, han de estar por encima de los recelos culturales.

Si pensamos en las consecuencias de una terminología imprecisa en campos como la medicina o la aeronáutica, por citar dos de los ejemplos más evidentes, nos resulta más fácil establecer nuestras prioridades. Donde existen tuercas y tornillos es obvio que no hacen falta “nuts and bolts”, pero un stent (14) (por más que el cirujano hispanohablante lo pueda pronunciar estén) es

(14) Stent: a slender tube inserted inside a tubular body part (as a blood vessel) to provide support during and after surgical anastomosis. Fuente: www.dict.org

siempre un stent.

En ocasiones será mejor aceptar un préstamo (WWW, software, pendrive, etc.) que recurrir a soluciones incomprensibles (Multi-Malla Mundial, soporte blando, memoria portátil –por imprecisa-), por citar algunas propuestas

en español (extraídas del foro “Spanglish”, dedicado a buscar propuestas alternativas y traducciones para términos informáticos de origen inglés).

¿Debe ser entonces un proceso dirigido o libre? No hay una respuesta absoluta, pero el dirigismo corre el riesgo de producir resultados que no sean aceptados por los miembros de la comunidad técnico-científica, que, a la postre, son quienes van a usar el término que mejor se acomode a sus necesidades. Ambas visiones son compatibles, pero dado que el lenguaje técnico procede del conocimiento especializado compartido por los expertos en el área de que se trate, la decisión final sobre qué término usar se basará en criterios pragmáticos.

Tal era el enfoque preconizado por Amelia de Irazazabal (químico y doctora en ciencias), precursora de la terminología en España, que siempre defendió su uso como medio de transmisión de conocimiento y comprensión entre comunidades científicas de distintas lenguas. La experiencia ha demostrado que el término que va a prevalecer será aquel que cuente con una mayor certeza de comprensión en el ámbito final de uso, en la propia comunidad técnico-científica, cada vez menos local y más internacional.

La terminología, y esto es una convicción profunda de este autor, debe ser siempre la solución y nunca el problema.

5. METODOLOGÍA

Vamos a analizar como punto de partida los planteamientos previos a todo estudio terminológico, analizando para ello las recomendaciones de las dos escuelas; por una parte, las del Instituto Universitario de Lenguas Aplicadas (IULA) -Universidad Pompeu Fabra, (que sigue el modelo francófono) y, por otra, el planteamiento de proyecto terminológico que plantea TERMNET. Aquí estaríamos ante el modelo centroeuropeo.

5.1 Modelo francófono

Los procedimientos que a se describen a continuación recogen la presentación llevada a cabo por las doctoras Estopà y Feliu , en la Escuela de Verano de Terminología IULATERM 2005.

Parámetros que condicionan el trabajo terminológico

. Si ha de ser un estudio monolingüe o plurilingüe

. Si será un estudio sistemático o puntual

. Si vamos a realizar un estudio descriptivo o prescriptivo (o si han de ser consecutivos)

. La temática y el punto de vista

Las fases del trabajo terminológico:

A.- Definición y delimitación del trabajo

- Presentación del tema y adquisición de conocimiento
- Delimitación del trabajo:

Tema
Destinatarios
Finalidad
Dimensiones

B.- Preparación del trabajo
- Ampliación y selección de la información
- Estructuración del conocimiento
- Redacción del plan de trabajo

C.- Elaboración de la terminología
- Confección del corpus de trabajo
- Vaciado terminológico
- Fichero terminológico

D.- Supervisión del trabajo
- Análisis y revisión del fichero terminológico
- Resolución de casos problemáticos

E.-Presentación del trabajo
- Presentación del glosario terminológico
- Edición

La exposición de las doctoras Estopà y Feliu continuaba planteándose cuáles son las competencias que el terminólogo debe tener a fin de poder realizar su trabajo; estas serían las siguentes:

A.- Competencia cognitiva. Conocimiento del ámbito especializado objeto del trabajo.

B.- Competencia lingüística. Conocimiento de la lengua o lenguas objeto del trabajo.

C.- Competencia socio-funcional. Conocimiento de los fines que persigue y los destinatarios a los que se dirige el trabajo.

D.- Competencia metodológica. Conocimiento de los principios a respetar y del proceso a seguir.

En cuanto a los recursos complementarios para el trabajo terminológico, estos serían: documentación y tecnología.

Documentación general

- Documentación sobre documentación: bases de datos documentales, centros de gestión terminológica, bibliografías, etc.
- Documentación sobre la especialidad: sobre terminología, sobre la temática y las lenguas de trabajo.
- Documentación sobre los términos: diccionarios, generales y especializados, enciclopedias, bases de datos terminológicas, bases de conocimiento, léxicos, etc.
- Documentación sobre el método: normas sobre metodología y normas sobre términos.

Documentación específica

- Los textos especializados: almacén de información especializada.

En cuanto a los recursos tecnológicos, éstos consisten básicamente en las diversas herramientas informáticas al servicio del terminólogo.

- Confección del corpus de trabajo a través de la búsqueda y captura de textos digitalizados e Internet.
- Marcaje estructural, morfológico y sintáctico del corpus textual; desambiguación.
- Extracción automática/asistida de terminología.
- Trasvase semiautomático de la información a una base de datos.
- Depuración semiasistida de los registros.

Elementos de partida

Elementos condicionantes de la toma de decisiones:

1-Teóricos: toda aplicación se fundamenta en unos principios teóricos

2-Metodológicos: Toda aplicación debe respetar los supuestos metodológicos en que se sustenta

3-Comunicativos

3.1 Actividad profesional:

-Es necesario acotar el objeto de trabajo pertinente para una actividad profesional concreta.

- -Se deben definir las necesidades terminológicas de la actividad para la que se realizará la aplicación.
- Las necesidades terminológicas no son uniformes.

3.2 Contexto sociocultural; establecer el contexto sociocultural (sociolingüístico) donde se utilizará la aplicación:

- características lingüísticas
- características culturales
- características laborales (lugar, tiempo, recursos)
- características económicas (nivel de desarrollo de los recursos informáticos)

3.3 El tema

- ámbito versus materia versus tema versus objeto temático, etc
- disciplinas científicas versus ámbitos profesionales
- ciencia versus técnica
- ciencias básicas versus ciencias aplicadas
- ciencias básicas versus ciencias humanas versus ciencias sociales
- disciplinas versus interdisciplinas versus transdisciplinas

4- Prácticos

4.1 La organización del trabajo terminológico

- perfiles de los miembros del equipo
- distribución de tareas
- calendario y fases del trabajo
- tareas de control y revisión
- presupuesto

4.2 Los aspectos de infraestructura

- acceso a documentación
- procesamiento de los textos
- gestión de los datos terminológicos
- edición de la aplicación

Elementos condicionados de la toma de decisiones:

La definición del tipo de trabajo y la concreción de cada uno de los parámetros condicionantes anteriores (comunicativos y prácticos) condicionan los siguientes aspectos:

- selección de asesores
- delimitación temática
- selección de la información
- representación de la información
- selección del soporte

El paso siguiente consistirá en la redacción del plan de trabajo.

La presentación del diseño

- maqueta de la aplicación
- protocolos de funcionamiento
- informe de los aspectos logísticos
- prototipo- modelo construido para comprobar la eficacia de un diseño antes de la fabricación del producto comercial. Implica innovación y aplicabilidad real.
- pruebas del prototipo

La implementación

Implementar significa realizar una idea, método, esquema, algoritmo, etc., hasta convertirlo en un objeto concreto.

En el caso de las aplicaciones terminológicas, la implementación se refiere tanto al proceso de construcción y prueba del prototipo como a la producción final de productos acabados comercializables.

Será necesario seleccionar las herramientas informáticas y documentales necesarias para completar y gestionar una base de datos terminológica.

- Un sistema de creación y gestión de una BDT.
- Herramientas de creación y etiquetaje de un corpus textual informatizado
- Extractor de terminología.
- Recurso a una ontología para completar la información conceptual de la unidad léxica especializada (ULE) y mejorar la posterior recuperación de información.
- Acceso en línea a diccionarios y otras bases de datos terminológicas.
- Recurso a los especialistas, ya sea directamente o a través de listas de discusión y contacto electrónico.

5.2 Modelo centroeuropeo del TERMNET (15)

El planteamiento que proponen para la realización de proyectos terminológicos es el siguiente: en primer lugar, definen proyecto terminológico siguiendo la norma ISO 15188, página 2:

> "Proyecto encaminado a reunir, desarrollar, analizar y guardar la terminología de uno o más campos de investigación o áreas de conocimiento".

(15)TERMNET es una asociación no-lucrativa internacional fundada a iniciativa de la UNESCO para promover el mercado terminológico. Es una red empresarial y de cooperación que reúne a más de cincuenta miembros (varias universidades, el Banco Central Europeo, la Asociación Austriaca de Informática, y empresas como SAP, entre otros) de veinte países. Fomenta la realización conjunta de proyectos, la promoción de productos y servicios, el intercambio de información y, en general, la potenciación de la terminología como medio de intercambio de información en un entorno de buena praxis. (Fuente:linux.termnet.org 2007)

Con este fin, se proponen cuatro fases:

1.- Preparación

2.- Diseño

3.- Implementación

4.- Revisión, evaluación y verificación.

1.- Preparación

. Estudio del marco legal: aspectos legales del trabajo terminológico en el sentido de averiguar si la información con la que vamos a trabajar está sujeta a derechos de copyright, y en ese caso quién los posee y cómo podremos acceder a dicha información.

. Factibilidad:

- propósito y objetivos
- necesidades de los usuarios
- posibles problemas

. Especificaciones: requisitos y criterios de selección

2.- Diseño

Existen cuatro modelos de organización:

Los llamados “committee”

A.- Con un terminólogo que actúa de consultor externo al grupo de trabajo

B.- Con un terminólogo integrante del grupo de trabajo

Los “terminology centred”

C.- Con un terminólogo que crea un vocabulario con un especialista

D.- Con uno o varios terminólogos que trabajan con especialistas consultores

En todos los casos resulta fundamental contar con un director de proyecto que esté familiarizado con el área de conocimiento y con los principios y métodos del trabajo terminológico.

3.- Fase de implementación

- ajustarse a los estándares
- respetar las especificaciones
- evaluar continuamente el trabajo
- documentar las dificultades encontradas
- anotar las decisiones

4.-Fase de revisión, evaluación y verificación.

La fase de revisión, evaluación y verificación del producto final y de todo el proyecto debe ajustarse a la Norma ISO 10006:1997 sobre directrices para la calidad en la gestión de proyectos.

A modo de resumen, lo más destacado de ambos planteamientos está en lo complejo del esquema de IULATERM, más teorizante, frente al enfoque pragmático de TERMNET.

De hecho, los tres primeros puntos del esquema IULA son obviados en el planteamiento de TERMNET. La razón: los aspectos teóricos, metodológicos y comunicativos del proyecto se dan por supuestos al tratar con terminólogos profesionales, cuya actividad primordial reside en solucionar problemas de empresas.

A este respecto, destacamos el apartado referente a “Implementación” en el enfoque del IULA. La mención a “producción final de productos acabados comercializables” implícitamente presupone la existencia de una actividad terminológica potencialmente gratuita impensable en el enfoque TERMNET.

A la luz de lo recogido en las páginas anteriores, la pregunta que surge es: ¿cuál de los dos enfoques seguir?

Como decíamos anteriormente, en este trabajo trataremos de ver el enfoque desde el punto de vista del técnico, exclusivamente. El especialista que trabaja habitualmente en este caso con terminología de estadística y, más concretamente, de minería de datos. Sin desdeñar un análisis lingüístico más profundo, consideramos que daría lugar a una sobredimensión del proyecto que no lo haría práctico ni atractivo a sus usuarios finales, informáticos, estadísticos y profesionales de la minería de datos. Por lo tanto, optaremos por elegir el MODELO D planteado por TERMNET, con un equipo formado por un terminólogo (el autor) más dos especialistas (español y alemán). Para la supervisión del producto final recurriremos a la ayuda de un terminólogo de reconocido prestigio (Prof. Dr. Klaus-Dirk Schmitz, de la Universidad de Colonia).

5.3 ¿Por qué centrarnos en la minería de datos?

Comenzaremos por definir “terminología”, “estadística” y “minería de datos”.

Terminología es, según la Norma ISO 1087 que volvemos a recoger:

> “Cualquier actividad que se ocupe de la sistematización y representación de conceptos o de la presentación de terminologías –conjuntos de términos que representan el sistema de conceptos de un campo de investigación determinado- basadas en principios y métodos establecidos”.

Parece entonces necesario establecer qué es entonces un término. Como ya mencionábamos en un principio, la propia Norma ISO 1087-1, define término como “La designación verbal de un concepto general en un campo de investigación concreto”.

Aquí trataremos de establecer un corpus terminológico específico de una rama de la estadística, la minería de datos. Son muchos los terminólogos (Picht, Cabré, Sauberer, etc) que recomiendan acotar el alcance del corpus a un área concreta, pues de lo contrario nuestro trabajo se vería desbordado. La estadística es una rama de las matemáticas demasiado amplia como para tratar de recopilar aquí un corpus trilingüe exhaustivo, por lo que parece recomendable ceñirse a la minería de datos, un área de reciente creación, más claramente acotada. De éste modo, cumplimos con lo que Sauberer (TSS 2006) denomina “fórmula SMART” para establecer nuestros objetivos:

S-pecific
M-easurable
A-chievable
R-ealistic/R-elevant
T-imely/T-ime specific

Objetivos específicos implica que sean claros y bien definidos, pudiendo incluir una mención al alcance del estudio con datos sobre lo que no se va a tratar. En nuestro caso, consiste en limitar el corpus a la minería de datos, excluyendo la estadística en general por demasiado amplia.

Medibles implica establecer un número determinado de términos que puedan ser definidos como tales; entre cien y ciento cincuenta parece ser un objetivo razonable a priori.

El que sea alcanzable presupone contar con los medios técnicos (herramientas informáticas) y humanos (especialistas informáticos,

matemáticos y lingüistas) precisos, así como tener acceso a las fuentes originales de los términos.

La relevancia del proyecto residirá en su utilidad para los usuarios finales; ha de ser un trabajo centrado en sus necesidades comunicativas, puesto que todo trabajo terminológico debe estar encaminado a facilitar la comunicación profesional (Picht 1996)

El factor temporal implica establecer una serie de fechas límite en las cuales deben estar concluidas cada una de las etapas de nuestro trabajo.

Pasamos entonces a centrar el área objeto de nuestro estudio.

Una definición de estadística la encontramos en la obra de Berson, Smith y Thearling (1999) , cuando dicen

> "Statistics is a branch of mathematics concerning the collection and the description of data."

Es decir, la rama de las matemáticas que se ocupa de reunir y describir datos.

La minería de datos procede de la estadística, de la informática y de otros campos, en función del estudio de que se trate (Menasalvas 2007); antes de dar una definición de "minería de datos" –traducción literal de "data mining"- puede resultar reveladora la opinión de Kurt Thearling (2005: 1252) cuando dice:

> "Data mining derives its name from the similarities between searching for valuable business information in a large database -for example, finding linked products in gigabytes of store scanner data- and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides."

Por lo tanto, la metáfora surge de la extracción de algo valioso, como es una veta de mineral en una montaña. En nuestro caso, la montaña son los datos corporativos, y la veta son aquellos datos que pudieran tener una especial relevancia para la propia empresa.

Una vez ubicado el término en su lengua de origen, resulta más fácil comprender la definición del mismo, que el propio autor nos facilita:

> "Data mining, *the extraction of hidden predictive information from large databases*, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analysees of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations."

Es decir, traduciendo literalmente: la extracción de información oculta, de valor predictivo, de grandes bases de datos.

La inclusión del párrafo que sigue a la definición no es casual en éste proyecto. Como tantas veces ocurre, el contar con un marco contextual nos ayuda a comprender mejor el significado del término.

Esta definición de Thearling viene a ampliar la que ya apareciera en "Discovering Data Mining From Concept to Implementation" (pp 12) que con el respaldo de IBM publicaran en 1997 Cabena, Hadjinian, Stadler, Verhees y Zanasi, donde leemos:

> "Although there is no one single definition of data mining that would meet with universal approval, the following definition is generally acceptable: data mining is the process of extracting previously unknown, valid and actionable information from large databases and then using the information to make crucial business decissions."

En primer lugar nos centraremos en el matiz imprescindible que establece que esa información ha de tener valor predictivo. Ahí reside la diferencia fundamental de la minería de datos frente a la estadística. La minería de datos se ocupa de información que, cuidadosamente seleccionada y procesada, permitirá a la empresa que disponga de ellos de una poderosa herramienta de previsión de acontecimientos futuros a partir de una mejor comprensión de los pasados, con los beneficios de toda índole que ello puede reportarle.

El conocimiento de esas futuras tendencias y comportamientos de que habla el autor permite a las empresas tomar decisiones de cara al futuro basándose no en la intuición, sino en el conocimiento. La minería de datos permite localizar información que, tradicionalmente, estaba fuera del alcance de los expertos.

Es en cualquier caso un campo de investigación de reciente implantación internacional y particularmente novedoso en España. No existen corpus referentes al mismo. La extensión del corpus resultante no se nos antoja inabarcable, por lo conciso del tema.

Cumplimos por lo tanto con dos requisitos imprescindibles de cualquier estudio terminológico, como son la limitación del trabajo en aras de su factibilidad, y la utilidad del producto final.

Para la realización del presente trabajo nos centramos originalmente, tras la fase de fundamentación teórica y selección de modelos metodológicos a seguir, en el siguiente objetivo:

Creación de un glosario terminológico trilingüe inglés-español-alemán sobre minería de datos.

5.4 Metodología seguida en esta Tesis

En el presente trabajo se describe el proceso de creación de un glosario terminológico sobre minería de datos en inglés, español y alemán. Según comentábamos en la introducción, la elección del tema se debe a la inexistencia constatada de glosarios terminológicos sobre la materia (ni en España, ni en Alemania) dado lo reciente de su aparición como especialidad informático-estadística de análisis de información oculta en grandes bases de datos. En cuanto a las lenguas de trabajo, el inglés es el origen de la totalidad de la literatura existente sobre la materia; el español es la lengua materna del autor y ofrece un número considerable de usuarios potenciales del glosario resultante, y, por último, el alemán permite contrastar los procesos de incorporación de nueva terminología a una lengua no-románica con tradición en la producción propia de términos técnicos.

Esta tesis nos da la oportunidad de analizar la capacidad de absorción y producción propia de términos de ambas lenguas, y, en nuestro caso, de comprobar los procesos que siguen los especialistas al enfrentarse por primera vez a la necesidad de incorporar a la docencia universitaria en sus respectivas lenguas maternas (español en el caso de la Dra. Menasalvas de la Universidad Politécnica de Madrid y Alemán en el del Dr. Lattner, de la Universidad J. W. Goethe de Frankfurt am Main) unos términos que han aprendido en su forma original inglesa. Por razones de índole práctica se optó por la elección de un equipo de trabajo formado por un terminólogo (el autor) y dos técnicos (16).

(16) Modelo organizativo “D” (terminology centred) según establece la Norma ISO 15188.

El proceso comenzó con la búsqueda de los glosarios monolingües en inglés ya existentes. La fuente de referencia obligada –una vez constatado que no existía un glosario semejante en el ámbito académico- eran las publicaciones de las principales empresas dedicadas a la minería de datos. De este modo se obtuvieron cuatro glosarios de partida, (por orden cronológico de publicación: IBM 1997, Clementine 2002, Two Crows Corporation 2006 y Kurt Thearling 2008. Véase sección de apéndices) que recogen una amplia selección de lo que estas cuatro empresas consideran terminología imprescindible de minería de datos. El hecho de que la publicación de estos glosarios se extendiera en un plazo de diez años permitiría, además, realizar un pequeño estudio diacrónico de alguno de los términos.

Estos cuatro glosarios fueron cruzados para evitar repeticiones, y el listado de candidatos a término resultante (17) fue sometido al análisis de los técnicos (los citados doctores Menasalvas y Lattner) de los países de las lenguas de destino, español y alemán. El listado incorporaba, además, una definición en inglés y una propuesta de traducción al español.

Como referencia para los especialistas, se les facilitó la siguiente orientación metodológica sobre cómo proceder ante las propuestas de entradas al glosario:

- Sólo deberían considerar como “términos” aquellas expresiones de una o más palabras cuyo uso y comprensión resultara imprescindible para transmitir los conocimientos teóricos propios de la minería de datos.
- Denominarían “semi-términos” a aquellas palabras y expresiones de otros campos paralelos (fundamentalmente matemáticas e informática)

(17) No podemos denominar a las palabras de este listado previo términos *stricto sensu*, puesto que la percepción del concepto “término” difiere sensiblemente en función de que lo considere como tal un académico o un ejecutivo. La empresa precisa comunicarse con clientes no-especialistas, con lo cual su enfoque es notablemente más laxo y carece en ocasiones de rigor científico.

cuyo uso fuera imprescindible para precisar conceptos relativos a procesos o técnicas propios de la minería de datos.

- Las demás propuestas del listado que no cumplieran con los dos requisitos mencionados anteriormente debían ser descartadas.

Los términos aceptados fueron recogidos en fichas terminológicas normalizadas (18) en las que, una vez incorporadas las indicaciones y sugerencias de los doctores Menasalvas y Lattner, se incluyeron los siguientes campos:

.Término en inglés, español y alemán.
.Definición en inglés y en español, con referencia a temas afines.
.Sinónimos en las tres lenguas.
.Área de conocimiento de la que procede el término.
.Apartado de minería de datos a que pertenecen(19).
.Ejemplo de uso del término en un texto del corpus.
.Posición del concepto en el área de conocimiento.

Evidentemente, esta relación no comprende la totalidad de los apartados recogidos en la Norma ISO 12620. Como ya indicábamos anteriormente, el espíritu práctico que rige esta tesis nos obliga a centrarnos en aquellos apartados que consideramos más relevantes para los usuarios finales del glosario resultante: los técnicos en minería de datos. En otras palabras,

(18)Según Norma ISO 12620. Véase punto 3 de la sección “PLANIFICACIÓN Y REALIZACIÓN DE UN PROYECTO TERMINOLÓGICO”.

(19) Los especialistas distinguen hasta siete apartados distintos en el proceso de minería de datos: Datos, Proceso, Técnicas y Algoritmos, Tipos de problemas, Tipos de Resultados, Parámetros de Evaluación y Genérico.

se renuncia a un análisis lingüístico más profundo en aras a conseguir lo que en informática se llamaría un “entorno más amigable” para los destinatarios del producto final. Estimamos que la ficha resultante recoge la totalidad de la información que un técnico en minería de datos precisa a la hora de ubicar con precisión un término en su área de conocimiento.

El trabajo de supervisión del listado por parte de los especialistas se desarrolló en dos fases, primero en Madrid en el Campus de Montegancedo con la Doctora Menasalvas y una segunda en Frankfurt am Main, Alemania, en el Campus de Bockenheim con el Doctor Lattner. No podemos por menos que recordar aquí que ambos profesores están considerados entre las máximas autoridades en materia de minería de datos en el mundo académico de sus respectivos países (véase reseña en el apartado correspondiente de apéndices) y sin su colaboración y ayuda este trabajo no habría sido posible.

En la primera fase le fue presentado a la Dra. Menasalvas un listado inicial de términos con su denominación original en inglés, la correspondiente definición en inglés, y una propuesta de traducción al español del término y su definición. A lo largo de una serie de reuniones (aproximadamente 12) de entre una y dos horas de duración en su despacho habitual de trabajo fueron revisadas las fichas una por una, obteniendo los siguientes resultados:

-La propuesta de traducción (literal) del término fue modificada en la mayoría de los casos.

-La propuesta de traducción (también literal) de la definición del término fue alterada todos los casos bien parcial o totalmente, siendo añadidas referencias a otros sub-apartados o precisiones aclaratorias.

-la definición original en inglés fue mejorada.

-El campo correspondiente a sinónimos quedó desierto en aproximadamente el 75% de los casos.

-A los campos inicialmente propuestos se incorporó uno nuevo, el correspondiente a la ubicación del término en su correspondiente sub-apartado dentro del campo general de la minería de datos.
-Se completó el campo referente al origen del término.
-Se completó el campo correspondiente a la posición del concepto en el área de conocimiento de procedencia.

La Doctora Menasalvas constató la ausencia de algunos términos imprescindibles en el listado propuesto, (fundamentalmente los relativos al proyecto CRISP DM (20)), razón por la cual y siguiendo sus orientaciones de documentación se procedió a incorporar el glosario Clementine.

El mismo método, basado en reuniones de trabajo periódicas en el despacho habitual del técnico, se siguió en Alemania con el Doctor Lattner a lo largo de cuatro meses. En esta fase del proyecto el volumen de información de cada ficha era mucho mayor que el presentado a la Dra. Menasalvas, lo que nos permitiría añadir, a parte de la información ligüística, una constatación técnica de los contenidos teóricos que presentaban los campos originales y los recién incorporados.

Los resultados de estas sesiones de trabajo con el Dr. Lattner son los siguientes:

-Se completó el campo correspondiente a equivalencia de cada término en alemán. En este caso no había sugerencia previa de traducción.
-Se revisaron y mejoraron las definiciones originales en inglés, y
-a sugerencia del Dr. Lattner se incorporó en algunos casos un nuevo campo, el correspondiente a referencias imprescindibles a temas relacionados.

(20) CRISP DM.- CRoss Industry Standard Process for Data Mining. El grupo, aparte de SPSS –matriz de Clementine-, también incluye a Two Crows Corp.

-La sinonimia en alemán subió por encima del 30%. En algunos casos se trataba del original en inglés, en otros era una expresión equivalente en alemán al término preferido, el original inglés, y en un número significativo de términos se planteaba una opción de término “en proceso de asimilación”, con una grafía alemana que representaba el original en inglés (sobre este punto se incide más extensamente en el apartado de conclusiones).

-Hubo una coincidencia próxima al 100% con los planteamientos de la Dra. Menasalvas en cuanto a ubicación de cada término en su campo correspondiente de minería de datos.
-Hubo una coincidencia próxima al 95% en cuanto al origen de cada término.
-al contar con más datos (ejemplos en contexto, número de glosarios de partida en los que contaba cada término y frecuencia de aparición de cada término en los textos del corpus) se pudo realizar una profunda revisión técnica de las fichas en su conjunto que permitió realizar algunos descartes (véase apartado de conclusiones).

Simultáneamente al trabajo con los técnicos se realizó un análisis estadístico de frecuencias de uso de los términos en textos propios de la especialidad que avalara objetivamente su opinión. Si se trataba efectivamente de términos propios del campo, deberían aparecer frecuentemente y en posiciones preponderantes en dichos textos.

Con este fin, se seleccionó un corpus sobre minería de datos de una extensión aproximada de 200.000 palabras que comprendiera las diversas temáticas propias de la materia. Los textos que forman el corpus proceden tanto de fuentes académicas (documentación de referencia para alumnos

universitarios) como de guías de empresas que proporcionan servicios de minería de datos a sus clientes. La totalidad de los textos que forman el corpus de referencia se puede consultar en la sección de Apéndices.

La localización, recuento y consiguiente validación de los términos se realizó por medio del análisis con herramientas informáticas (21) de los textos del corpus. Estos programas fueron utilizados para conseguir los siguientes objetivos:

.Localizar el término buscado en su contexto de uso, para obtener ejemplos que ilustraran el campo correspondiente en cada ficha.
.Obtener los porcentajes aparición del término en la totalidad del corpus. A mayor porcentaje, mayor certeza de pertenencia al campo de la minería de datos.

La razón por la que se emplearon dos analizadores, uno de textos españoles y otro de textos ingleses está en que ambos se complementan para conseguir un análisis más fidedigno dadas las características de este tipo de programas. El empleo de uno sólo mostró señales evidentes de error en los datos porcentuales de frecuencias de aparición por término, dado que sólo reconoce nombres o grupos nominales; ello daba lugar a que términos como *sampling* no fueran localizados por la versión en inglés (22). Los datos estadísticos de cada término se pueden consultar también en la Sección 8 APÉNDICES.

De este modo, podemos resumir el proceso diciendo que los datos obtenidos a partir de los glosarios de partida fueron filtrados en primer lugar por

(21) Analizadores de Textos Trobes Version 6.2.4 (build 0013) de Acetic (español) y Zoom Version 7.0.0 (build 0041) de Acetic (inglés)

(22) Véase el apartado correspondiente al analizador de textos en la sección 7 CONCLUSIONES.

el criterio de los técnicos y que el producto resultante de esta etapa fue finalmente validado mediante el análisis estadístico de la frecuencia de aparición de cada término en el corpus de referencia. El glosario así obtenido fue nuevamente sometido a la consideración de los técnicos, quienes dieron su visto bueno al resultado final.

El proyecto fue sometido en su etapa final a la supervisión de un terminólogo de constatada experiencia, el Profesor Dr. Klaus-Dirk Schmitz, de la Universidad de Colonia. El Dr. Schmitz, además de su dilatada experiencia en la dirección de proyectos terminológicos, es Presidente del Consejo de Terminología Alemana (RaDT), Presidente del Centro Internacional de Información para la Terminología (Infoterm) y posee otros cargos institucionales y académicos que se pueden consultar en la sección correspondiente de Apéndices.

Una vez presentado el proyecto al Dr. Schmitz se incorporaron las siguientes mejoras:

-En el apartado de definiciones de los términos sugirió que se hiciera una depuración y mejora por parte de los técnicos.

-La denominación “semi-término” debía ser eliminada; cualquier expresión cuyo uso sea imprescindible para explicar algún apartado de minería de datos debe ser considerada término, aunque proceda de otro campo.

-En la incorporación de nuevos términos al alemán indicó que debería establecerse el género correspondiente de los sustantivos.

-En la sección correspondiente a origen del término, indicó la conveniencia de introducir la nota “subject field 1”, dado que es la forma estandarizada de referirse a este campo en las fichas terminológicas.

-La ubicación del término en el apartado correspondiente de minería de datos debería ser referida como “subject field 2”, por las mismas razones indicadas en el punto anterior.

-Estableció como suficiente la extensión del corpus de referencia, (200.000 palabras), dadas las características del campo a investigar. La calidad de los mismos debería primar sobre su número.
-No estimó necesario incluir la frecuencia de aparición por texto de cada término.
-Como punto final, dio su visto bueno a los procedimientos de documentación, análisis técnico y tratamiento informático de los términos que constituyen los pilares de esta proyecto.

El glosario resultante quedaba así certificado.

6- GLOSARIO TERMINOLÓGICO DE MINERÍA DE DATOS

6.1 Listado de términos en inglés

Accuracy	
Activation function	
Antecedent	
Association rule	
Associations discovery	
Associations	
Attribute	
Back propagation	
Binning	
Boosting	
Business intelligence	
CART	
Case	
Categorical	Véase discrete
Categorical data	
Categorical variable	
CHAID	
Classification	
Classification model	
Cleaning	
Cleansing	Véase cleaning
Cluster	
Clustering	
Column	
Confidence	Véase predictability
Confidence factor	
Confusion matrix	
Continuous	
Continuous variable	
CRISP DM	

Cross validation	
Data item	Véase variable
Data mining	
Data preparation	
Data understanding	
Data warehouse	
Decission tree	
Degree of fit	
Dependent variable	Véase target
Deployment	
Deviation/Outlier detection	
Dimension	
Discrete	
Discretization	
Discriminant analysis	
Evaluation	
Exploratory data analysis	
Feature	
Feed forward	
Field	
Genetic algorithms	
Hidden layer	
Induction	
Interval	Véase range
Item	
K-means	
K-nearest neighbor	
Kohonen feature map	
Layer	
Leaf	
Learning	
Learning algorithm	
Left hand side	Véase antecedent
Linear regression	
Link analysis	
Logistic discriminant analysis	Véase logistic regression
Logistic regression	
Machine learning	
Market basket analysis	
Misclassification matrix	Véase confusion matrix
Missing data	
Model	
Modeling	
Neural net	Véase neural network
Neural network	
Node	

Noisy data	
Nominal variable	
Normalize	
Outliers	
Overfitting	
Overtraining	Véase overfitting
Pattern	
Precision	
Predictability	
Prediction	
Predictive modeling	
Predictor	
PCA Principal Component Analysis	
Prevalence	
Processing element	Véase processing unit
Processing unit	
Pruning	
Radial basis function	
Range	
Record	
Regression tree	
Right-hand side	Véase consequent
Rule	
Rule body	Véase antecedent
Rule head	Véase consequent
Rule induction	
Sample	
Sampling	
Scoring	
Segment	
Segmentation	
Sensitivity analysis	
Sequence discovery	
Sequential patterns	
Significance	
SOM	Véase Kohonen feature map
Standarize	Véase normalize
Supervised learning	
Support factor	
Target	
Taxonomy	
Test data	
Test error	
Test&training	Véase training
Time series	
Time series model	

Training
Training data
Transaction
Transformation
Unsupervised learning
Validation
Value prediction
Variable
Visualization
Windowing

6.2 Listado de Términos en Español

Accuracy	accuracy
Algoritmo de aprendizaje	learning algorithm
Algoritmos genéticos	genetic algorithms
Alimentación hacia delante	feed forward
Almacén de datos	data warehouse
Análisis de cesta de la compra	Market basket analysis
Análisis de componentes principales	PCA Principal Component Analysis
Análisis de datos exploratorio	Exploratory data analysis
Análisis de sensibilidad	sensitivity analysis
Análisis discriminante	Discriminant analysis
Análisis logístico discriminante	logistic regression
Antecedente	antecedent
Aprendizaje automático	machine learning
Aprendizaje no supervisado	unsupervised learning
Aprendizaje supervisado	supervised learning
Aprendizaje	learning
Árbol de decisión	decission tree
Árbol de regresión	regression tree
Asociaciones	associations
Atributo	attribute
Bining	binning
Boosting	boosting
Business intelligence	business intelligence
Cabeza de regla	consequent
Campo	field
Capa	layer
Característica	feature
CART	CART
Caso	case
Categórico	discrete
CHAID	CHAID
Clasificación	classification
Cluster	cluster
Clustering	clustering
Columna	column
Comprensión de datos	data understanding
Confianza	predictability
Consecuente	consequent
Contínuo	continuous
CRISP DM	CRISP DM
Cuerpo de regla	antecedent
Data mining	data mining

Data warehouse	data warehouse
Dato	variable
Datos atípicos	outliers
Datos categóricos	categorical data
Datos con ruido	noisy data
Datos de entrenamiento	training data
Datos de prueba	test data
Datos fuera de rango	outliers
Datos vacíos	missing data
Deployment	deployment
Descubrimiento de secuencias	sequence discovery
Detección de anomalías	deviation/outlier detection
Detección de desviaciones	deviation/outlier detection
Dimensión	dimension
Discretización	discretization
Discreto	discrete
Entrenamiento	training
Error del modelo	test error
Exactitud	accuracy
Evaluación	evaluation
Factor de confianza	confidence factor
Factor de soporte	support factor
Fila	case
Función basada en el radio	radial basis function
Función de activación	activation function
Grado de ajuste	degree of fit
Hoja	leaf
Inducción de regla	rule induction
Inducción	induction
Intervalo	range
Item	item
KDD	data mining
K-medias	k-means
Lado derecho	consequent
Lado izquierdo	antecedent
Limpieza	cleaning
Link análisis	link analysis
Mapa de Kohonen	Kohonen feature map
Matriz de confusión	confusion matrix
Missing data	missing data
Modelado predictivo	predictive modeling
Modeling	modeling
Modelo de clasificación	classification model
Modelo de series temporales	time series model
Modelo predictor	predictor
Modelo	model

Muestra	sample
Muestreo	sampling
Nivel oculto	hidden layer
Nodo	node
Normalizar	normalize
Objetivo	target
Overfitting	overfitting
Patrón	pattern
Patrones secuenciales	sequential patterns
Poda	pruning
Precisión	precision
Predecibilidad	predictability
Predicción de valores	value prediction
Predicción	prediction
Preparación de datos	data preparation
Prevalencia	prevalence
Proceso de descubrimiento de asociaciones	associations discovery
Rango	range
Red neuronal	neural network
Registro	record
Regla de asociación	association rule
Regla	rule
Regresión lineal	linear regression
Regresión logística	logistic regression
Repositorio de datos	data warehouse
Retroalimentación	back propagation
Scoring	scoring
Segmentación	segmentation
Segmento	segment
Series temporales	time series
Significancia	significance
Sobreaprendizaje	overfitting
SOM	Kohonen feature map
Taxonomía	taxonomy
Transacción	transaction
Transformación	transformation
Tupla	case
Unidad de proceso	processing unit
Validación cruzada	cross validation
Validación	validation
Variable	variable
Variable a predecir	target
Variable categórica	categorical variable
Variable continua	continuous variable
Variable de salida	target
Variable decisión	target

Variable dependiente	target
Variable nominal	nominal variable
Vecino K-cercano	k-nearest neighbor
Vecino más cercano	k-nearest neighbor
Ventana temporal	windowing
Visualización	visualization

6.3 Listado de términos en alemán

Abweichungserkennung/Entdeken von Ausreissern	deviation/outlier detection
accuracy	accuracy
activation function	activation function
Aktivierungsfunktion	activation function
Anwendungsphase	deployment
Artikel	item
association rule mining	asociations discovery
Assoziationen	associations
Assoziationsregel	association rule
Assoziationsregellernen	associations discovery
Attribut	attribute
Ausreisser	outliers
Back Propagation	back propagation
Begriffshierarchie	taxonomie
Beispiel	case
Beispielmenge	sample
Binning	binning
Blatt	leaf
Boosting	boosting
Business Intelligence	business intelligence
CART	CART
CHAID	CHAID
cleaning	cleaning
Cluster	cluster
Clustering	clustering
Confidence	confidence factor
confusion Matrix	confusion matrix
consequent	conbsequent
CRISP DM	CRISP DM
cross Validation	cross validation
Data Mining	data mining
Data Understanding	data understanding
Data Warehouse	data warehouse
Datenbereinigung	cleaning
Datervorberarbeitung	data preparation
Datenvorbereitung	data preparation
Dimension	dimension
Decision Tree	decision tree
diskret	discrete
Diskretisierung	discretization
Diskriminanzanalyse	discriminant analysis

Einheit	processing unit
Eintrag	record
Entscheidungsbaum	decision tree
Evaluation	evaluation
explorative Datenanalyse	exploratory data analysis
Feature	feature
feed forward	feed forward
fehlende Daten	missing data
Feld	field
Genauigkeit	accuracy
generieren einesnVorhersagemodell	predictive modeling
Genetische Algorithmen	genetic algorithms
grad der Anpassung	degree of fit
Häufigkeit	prevalence
Hidden Layer	hidden layer
Hidden-Schicht	hidden layer
Induktion	induction
Induktives Lernen	induction
interne Daten	internal data
Intervall	range
Klassifikation	classification
Klassifikationsmodell	classification model
K-Means	K-means
k-nächste Nachbarn	k-nearest neighbor
k-nearest Neighbors	k-nearest neighbor
Knoten	node
Knowledge Discovery	data mining
Kohonen-Karte	Kohonen feature map
Konfidenz	predictability
Konfidentzwert	confidence factor
Konsequentz	consequent
kontinuierlich	continuous
kontinuierliche Variable	continuous variable
Layer	layer
Lernen	learning
Lernenalgoritmus	learning algorithm
lineare Regression	linear regression
logistische Regression (logistische discriminantz Analyse)	logistic regression
maschinelles Lernen	machine learning
Merkmal	feature
Modell	model
Modellbildung	modeling
Modellierung	modeling
Muster	pattern
Neuron	processing unit
neuronales Netz	neural network

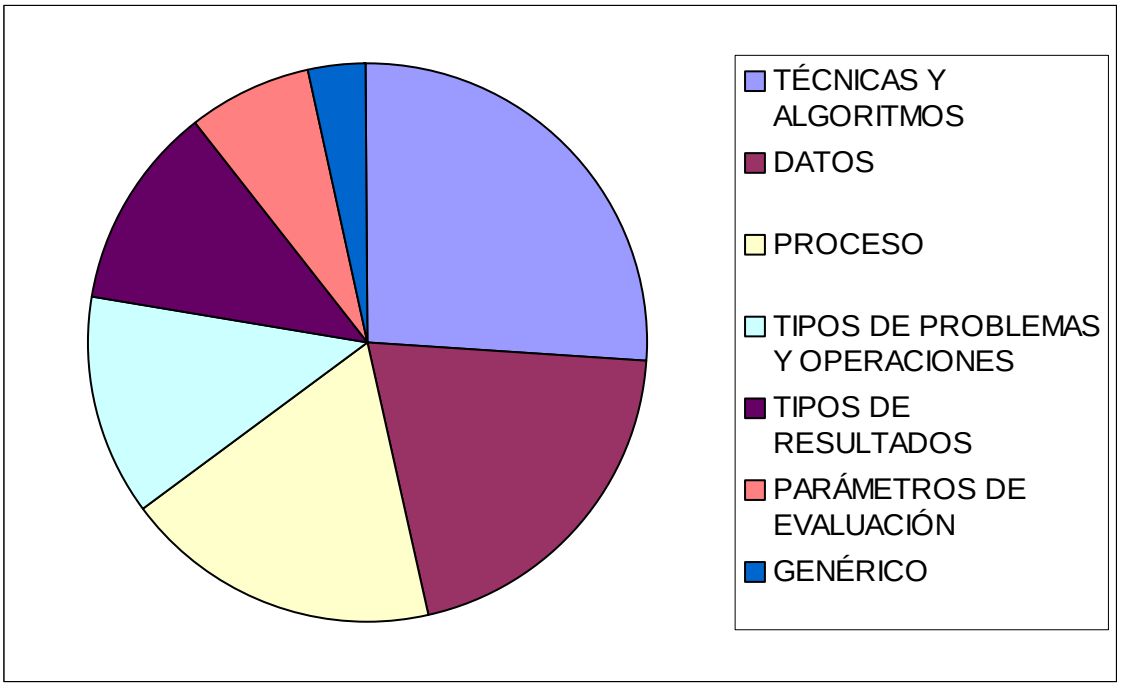
nominal variable	nominal variable
nominale variable	nominal variable
normalisieren	normalize
Overfitting	overfitting
outlier detection	deviation
outliers	outliers
Output	target
Pattern	pattern
Prädiktion	prediction
Präzision	precision
pre condition	antecedent
predictive modeling	predictive modeling
Predictor	predictor
Principal Component Analysis	Principal Component Analysis
Prognose	prediction
Pruning	pruning
radiale Basisfunktion	radial basis funktion
range	range
Record	record
reellwertig	continuous
reellwertige variable	continuous variable
Regel	rule
Regelinduktion	rule induction
Regression Tree	regression tree
Sampling	sampling
Schicht	layer
Scoring	scoring
Segment	segment
Segmentierung	segmentation
Sensitivitätsanalyse	sensitivity analysis
Sequence Discovery	sequence discovery
sequentielles Muster	sequential patterns
Signifikanz	significance
SOM	Kohonen feature map
Spalte	column
Supervised Learning	classification/supervised learning
Support	support factor
symbolische Daten	categorical data
symbolische Variable	categorical variable
Taxonomie	taxonomy
Testdaten	test data
Testfehler	test error
Training	training
Trainingsdaten	training data
Transaktion	transaction
Transformation	transformation

überwachtes Lernen	supervised learning
unsupervised learning	unsupervised learning
unüberwachtes Lernen	unsupervised learning
Validierung	validation
Variable	variable
verrauschste Daten	noisy data
versteckte Schicht	hidden layer
Visualisierung	visualization
vorbedingung	antecedent
Vorhersage	prediction
Vorhersagemodell	predictor
Vorhersagbarkeit	predictability
Warenhorbanalyse	maket basket analysis
Wertvorhersage(Vorhersage)	value prediction
Windowing	windowing
Wissensentdeckung	data mining
Zeitreihe	time series
Zeitreihenmodell	time series model
Zielattribut	target
Zusammenhänge	associations
Zusammenhangsanalyse	link analysis

6.4 Principales sub-apartados en minería de datos y porcentajes de términos que comprenden

Este apartado se corresponde con el campo denominado "POSICIÓN DEL CONCEPTO" en la ficha. Equivale al *subject field 2.*

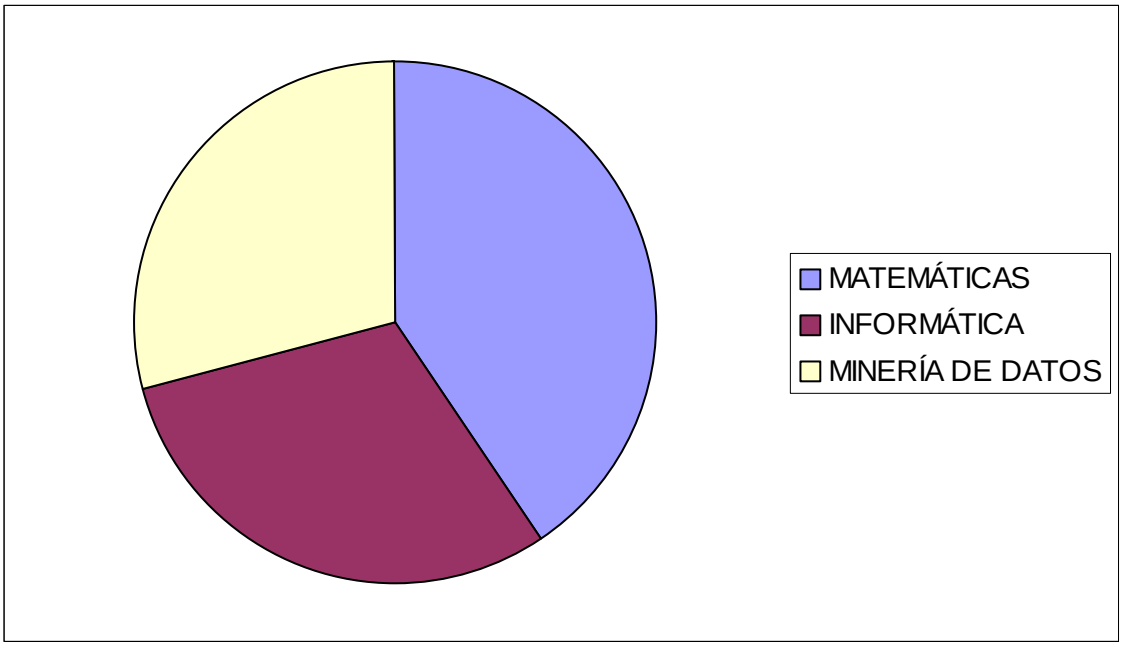
- Técnicas y algoritmos 26,22%
- Datos 20,16%
- Proceso 18,35%
- Tipos de problemas y operaciones 12,70%
- Tipos de resultados 12,09%
- Parámetros de evaluación 7,25%
- Genérico 3,22%



6.5 Origen de los términos que constituyen el campo de la minería de datos.

Este apartado se corresponde con el campo denominado "ORIGEN" en la ficha. Equivale al *subject field 1.*

INFORMÁTICA	**40,5%**
Inteligencia artificial	30,7%
Bases de datos	7,9%
Otros	1,9%
MATEMÁTICAS	**30,2%**
Estadística	21,4%
Economía	4,4%
Otros	4,4%
MINERÍA DE DATOS	**29,3%**



6.6 Ficha modelo

Descripción de cada uno de los campos que forman la ficha.

término en inglés **(% de textos del corpus en que se encuentra)** **Es término en español** **DE término en alemán**
DEFINICIÓN
ES definición en español proporcionada por los técnicos EN definición original en inglés supervisada y modificada por los técnicos
SINÓNIMO
ES Sinónimos en las tres lenguas si los hubiera o Ø EN DE
ORIGEN (subject field 1) ciencia que procede el término ubicación jerárquica del término en la SF1
POSICIÓN DEL CONCEPTO sub-apartado de minería de datos donde se ubica el término.
CONTEXTO Fragmento de texto original en el que se ubica el texto y fuente de la que procede.

6.7 Glosario

accuracy (70,8%)

ES accuracy DE Accuracy

DEFINICIÓN

ES accuracy. La exactitud es un factor importante al evaluar la calidad de las reglas obtenidas en un modelo de predicción. Hace referencia al grado de ajuste entre el modelo y los datos. Ver *precision.*

EN accuracy is an important factor in assessing the success of data mining. When applied to data, accuracy refers to the rate of correct values in the data. When applied to models, accuracy refers to the degree of fit between the model and the data. This measures how error-free the model's predictions are. Since accuracy does not include cost information, it is possible for a less accurate model to be more cost-effective. Also see *precision.*

SINÓNIMO

ES exactitud
EN Ø
DE Genauigkeit

ORIGEN Informática

Informática > inteligencia artificial > modelos de predicción > técnicas > accuracy

POSICIÓN DEL CONCEPTO Parámetros de evaluación

CONTEXTO

"A model that can be understood is a model that can be trusted. While statistical methods build some trust in a model by assessing its **accuracy**, they cannot assess the model's semantic validity — its applicability to the real world. A data mining algorithm that uses a human-understandable model can be checked easily by domain experts, providing much needed semantic validity to the model. Unfortunately, users are often forced to trade off accuracy of a model for understandability". (Fuente: *Visualizing Data Mining Models* by Kurt Thearling et al).

activation function (12,5%)

ES función de activación DE Aktivierungsfunktion

DEFINICIÓN

ES función de activación. Función que usa un nodo en una red neuronal para transformar los datos introducidos a partir de cualquier dominio de valores en una gama finita de valores. La idea original era aproximar el modo en que las neuronas se activaban, y la función de activación adoptaba el valor 0 hasta que el valor de entrada crecía y el valor saltaba a 1. La discontinuidad de esta función 0 ó 1 causaba problemas matemáticos, y ahora se utilizan funciones de tipo sigmoide (por ejemplo, la función logística).

EN A function used by a node in a neural net to transform input data from any domain of values into a finite range of values. The original idea was to approximate the way neurons fired, and the activation function took on the value 0 until the input became large and the value jumped to 1. The discontinuity of this 0-or-1 function caused mathematical problems, and sigmoid-shaped functions (e.g., the logistic function) are now used.

SINÓNIMO

ES Ø
EN Ø
DE activation function

ORIGEN Informática

Informática > inteligencia artificial > clasificación > clustering> redes neuronales > función de activación

POSICIÓN DEL CONCEPTO Técnicas y algoritmos

CONTEXTO

"Kohonen networks are feedforward neural networks generally with no hidden layer. The networks generally contain only an input layer and an output layer but the nodes in the output layer compete amongst themselves to display the strongest **activation** to a given record. What is sometimes called "winner take all". (Fuente: *An Overview of Data Mining Techniques* Excerpted from the book Building Data Mining Applications for CRM by Alex Berson, Stephen Smith, and Kurt Thearling)

antecedent (12,5%)

ES antecedente DE Vorbedingung

DEFINICIÓN

ES antecedente. Elemento izquierdo en una regla. Por ejemplo, en la relación “cuando un buscador compra un pico, también compra una pala en el 14% de los casos”, “compra un pico” es el antecedente.

EN When an association between two variables is defined, the first item (or left-hand side) is called the antecedent. For example, in the relationship "When a prospector buys a pick, he buys a shovel 14% of the time," "buys a pick" is the antecedent.

SINÓNIMO

ES cuerpo de regla, lado izquierdo
EN rule body, left-hand side
DE pre condition

ORIGEN Matemáticas/Informática

Matemáticas > lógica > antecedente
Informática > inteligencia artificial > representación de conocimiento > antecedente

POSICIÓN DEL CONCEPTO Tipos de resultados

CONTEXTO

“Associations are written as A Þ B, where A is called the **antecedent** or left-hand side (LHS), and B is called the consequent or right-hand side (RHS)”. (Fuente: Two Crows *Data Mining in Brief*)

association rule (58,3%)

ES regla de asociación DE Assoziationsregel

DEFINICIÓN

ES regla de asociación. Regla que expresa la afinidad entre la presencia de ciertos valores de atributos en un registro y la presencia de otros valores de atributos en el mismo registro. La forma general de una regla de asociación es X1, X2,... => Y1, Y2... . Un ejemplo sería:
A=>B, es decir, si A existe en una transacción, entonces B también existe. (Véanse también *support factor* y *confidence factor*. Y agrawal 1994).

EN a rule that expreses the affinity between the presence of certain items in a transaction and the presence of other items in the same transaction.

SINÓNIMO

ES Ø
EN Ø
DE association rule

ORIGEN Minería de datos

Minería de datos > patrones > tipos de resultado >regla de asociación

POSICIÓN DEL CONCEPTO Tipos de resultados

CONTEXTO

"It is often difficult to decide what to do with **association rules** you've discovered. In store planning, for example, putting associated items physically close together may reduce the total value of market baskets — customers may buy less overall because they no longer pick up unplanned items while walking through the store in search of the desired items. Insight, analysis and experimentation are usually required to achieve any benefit from association rules". (Fuente: Two Crows *Data Mining in Brief*)

associations (66,6%)

ES asociaciones DE Assoziationen

DEFINICIÓN

ES asociaciones. Afinidades entre elementos en transacciones que se descubren como resultado del *descubrimiento de asociaciones.* Las asociaciones se expresan como *normas de asociación*.

EN affinities between items in transactions that are discovered as the result of associations discovery. Associations are expressed as association rules.

SINÓNIMO

ES Ø
EN Ø
DE Zusammenhänge

ORIGEN Minería de datos

Minería de datos > patrones > descriptivo > asociaciones

POSICIÓN DEL CONCEPTO Tipos de resultados - Tipos de problema

CONTEXTO

"For example, more meaningful **associations** on products in typical market baskets may be obtained when different runs of the association operator are done on data that represents a given quarter, or a given week, as opposed to data that spans an entire year. The reason for this is that rules that exist for a given support level are less likely to appear when mining association data that represents long periods of time. Likewise, it may be more meaningful to discover associations among products in typical market baskets when the search is done using data from one store at a time". (Fuente: IBM *Data Mining: Extending the Information Warehouse Framework*)

associations discovery (50%)

ES descubrimiento de asociaciones DE Assoziationsregellernen

DEFINICIÓN

ES proceso de descubrimiento de asociaciones. Proceso de minería de datos que se utiliza para descubrir asociaciones. Su finalidad es buscar valores de atributos en un registro que impliquen la presencia de otros valores de atributos en el mismo registro. (Véase *a priori algorithm* y *basket analysis*).

EN data mining technique for discovering associations. Its aim is to find items in a transaction that imply the presence of other items in the same transaction. The process of finding associations.

SINÓNIMO

ES Ø
EN Ø
DE association rule mining

ORIGEN Minería de datos

Minería de datos > tipos de problemas > problemas descriptivos > proceso de descubrimiento de asociaciones

POSICIÓN DEL CONCEPTO Tipos de problema

CONTEXTO

"The two most common approaches to link analysis are ***association discovery*** and *sequence discovery*. Association discovery finds rules about items that appear together in an event such as a purchase transaction. Market-basket analysis is a well-known example of association discovery". (Fuente: Two Crows *Data Mining in Brief*)

attribute (62,5%)

ES atributo DE Attribut

DEFINICIÓN

ES atributo . Término que se emplea a veces en lugar de *variable*. Cada una de las características que definen una determinada entidad. (Véase entidad, modelo entidad-relación y definición de Chen).

EN an alternative name for variable

SINÓNIMO

ES variable
EN variable
DE Variable

ORIGEN Informática

Informática > bases de datos > tabla > atributo

POSICIÓN DEL CONCEPTO Datos

CONTEXTO

"Specific **attribut**e values are selected in the Evidence Visualizer in order to predict income for people with those characteristics". (Fuente*: Visualizing Data Mining Models* by Kurt Thearling *et al*)

back propagation (20,8%)

ES retroalimentación DE Back Propagation

DEFINICIÓN

ES Retroalimentación. Algoritmo de aprendizaje supervisado de propósito general que muchas redes neuronales usan. Las redes de propagación regresiva son redes multinivel de alimentación progresiva que utilizan un algoritmo de aprendizaje supervisado para ajustar los pesos de conexión interna. (Véase back propagation neural network)

EN a general purpose, supervised learning algorithm used by many neural networks. Back propagation networks are feed-forward, multi-layer networks that use a supervised learning algorithm to adjust the internal connection weights.

SINÓNIMO Ø

ORIGEN Informática

Informática > inteligencia artificial > redes neuronales > red neuronal con retroalimentación

POSICIÓN DEL CONCEPTO Técnicas y algoritmos

CONTEXTO

Back propagation training is simply a version of gradient descent, a type of algorithm that tries to reduce a target value (error, in the case of neural nets) at each step. (Fuente: Two Crows *Data Mining in Brief*)

binning (25%)

ES bining DE Binning

DEFINICIÓN

ES bining [bainin]. Técnica de discretización de datos que convierte los datos continuos en datos discretos al sustituir un valor de un rango continuo por un identificador de intervalo, de modo que cada intervalo representa un rango de valores. Por ejemplo, la edad se puede convertir en intervalos tales como 0 a 20 , 21 a 40, 41 a 65 y mayor de 65.

EN a data preparation activity that converts continuous data to discrete data by replacing a value from a continuous range with a bin identifier, where each bin represents a range of values. For example, age could be converted to bins such as 20 or under, 21-40, 41-65 and over 65.

SINÓNIMO Ø

ORIGEN Estadística

Matemáticas > estadística > discretización > bining

POSICIÓN DEL CONCEPTO Proceso

CONTEXTO

"In many cases, it would be better to drop certain variables rather than to include them due to the proportion of missing values, the proportion of values in the spike (if one exists), and/or the excessive number of levels. Several modeling methods have the ability to group some of these levels and can address some of the problems associated with extremely skewed predictors (that is, **binning** transformations). It is far better to transform problematic variables or remove them from consideration". (Fuente: *Identifying and Overcoming Common Data Mining Mistakes* Doug Wielenga, SAS Institute Inc., Cary, NC).

boosting (16,6%)

ES boosting DE Boosting

DEFINICIÓN

ES técnica empleada para incrementar la precisión del modelo, en base a la construcción secuencial de varios modelos. Una vez construido el primer modelo, se comprueban los datos para localizar los registros en los que el modelo da error. A partir de la información obtenida se construye el segundo modelo, se comprueban nuevamente los datos en busca de errores y se genera un nuevo modelo. Se procede así sucesivamente hasta que se ha construido un número específico de modelos. El modelo boosted es el conjunto de modelos construidos, y las predicciones finales se obtienen a partir de la combinación de las predicciones de los modelos individuales.

EN technique used to increase the accuracy of the model. The technique uses multiple models built sequentially. The first model is built normally. The data are then weighted to emphasize the records for which the first model generated errors and the second model is built. The data are then weighted again based on the second model's errors and another model is built, and so on until the specified number of models has been built. The boosted model consists of the entire set of models, with final predictions determined by combining the individual model predictions.

SINÓNIMO Ø

ORIGEN informática

Informática > inteligencia artificial > clasificación > boosting

POSICIÓN DEL CONCEPTO técnicas y algoritmos

CONTEXTO

"If you were to build a model using one sample of data, and then build a new model using the same algorithm but on a different sample, you might get a different result. After validating the two models, you could choose the one that best met your objectives. Even better results might be achieved if you built several models and let them vote, making a prediction based on what the majority recommended. Of course, any interpretability of the prediction would be lost, but the improved results might be worth it. This is exactly the approach taken by **boosting**, a technique first published by Freund and Schapire in 1996. Basically, boosting takes multiple random samples from the data and builds a classification model for each. The training set is changed based on the result of the previous models. The final classification is the class assigned most often by the models. The exact algorithms for boosting have evolved from the original, but the underlying idea is the same. Boosting has become a very popular addition to data mining packages". (Fuente: Two Crows Data Mining in Brief)

business intelligence (41,6%)

ES business intelligence DE Business Intelligence

DEFINICIÓN

ES business intelligence. Término general que abarca todos los procesos, técnicas y herramientas de apoyo a la toma de decisiones empresariales basadas en la tecnología de la información.

EN general term covering all proceses, techniques and tools that support business decision making based on information technology.

SINÓNIMO Ø

ORIGEN Economía

Economía > marketing > business intelligence

POSICIÓN DEL CONCEPTO Genérico

CONTEXTO

"**Business intelligence** (BI) is a broad category of applications and technologies for gathering, storing, analyzing, and providing access to data to help enterprise users make better business decisions. BI applications include the activities of decision support systems, query and reporting, online analytical processing (OLAP), statistical analysis, forecasting, and data mining. Business intelligence applications can be: Mission-critical and integral to an enterprise's operations or occasional to meet a special requirement - Enterprise-wide or local to one division, department, or project -Centrally initiated or driven by user demand.This term was used as early as September, 1996, when a Gartner Group report said: By 2000, Information Democracy will emerge in forward-thinking enterprises, with Business Intelligence information and applications available broadly to employees, consultants, customers, suppliers, and the public. The key to thriving in a competitive marketplace is staying ahead of the competition. Making sound business decisions based on accurate and current information takes more than intuition. Data analysis, reporting, and query tools can help business users wade through a sea of data to synthesize valuable information from it - today these tools collectively fall into a category called "Business Intelligence." (Fuente: Luca Rosseti 2006)

CART (33,3%)

ES CART DE CART

DEFINICIÓN

ES CART. (Siglas correspondientes a Classification And Regression Trees, árboles de clasificación y regresión). CART consiste en dividir las variables independientes en grupos reducidos y ajustar una función constante a conjuntos de datos pequeños. Esta función constante dependerá del tipo de variable (categórica o numérica). En árboles categóricos la función constante es aquella que se ajusta a un grupo reducido de valores finitos (por ejemplo, Sí o No, bajo o medio o alto). En árboles de regresión, el valor medio de la respuesta se ajusta a conjuntos pequeños de datos conectados.

EN Classification And Regression Trees. CART is a method of splitting the independent variables into small groups and fitting a constant function to the small data sets. In categorical trees, the constant function is one that takes on a finite small set of values (e.g., Y or N, low or medium or high). In regression trees, the mean value of the response is fit to small connected data sets.

SINÓNIMO Ø

ORIGEN Matemáticas

Matemáticas > estadística > técnicas > CART

POSICIÓN DEL CONCEPTO Técnicas y algoritmos

CONTEXTO

"Some decision tree algorithms may use heuristics in order to pick the questions or even pick them at random. **CART** picks the questions in a very unsophisticated way: It tries them all. After it has tried them all CART picks the best one uses it to split the data into two more organized segments and then again asks all possible questions on each of those new segments individually". (Fuente: *An Overview of Data Mining Techniques.* Excerpted from the book Building Data Mining Applications for CRM by Alex Berson, Stephen Smith, and Kurt Thearling).

case (70,8%*)

ES caso DE Beispiel

DEFINICIÓN

ES un registro en una base de datos. Conjunto de datos relativos a un caso concreto. Un caso puede ser un cliente, o una transacción. (Un registro podría ser una caso y todo caso se puede representar con un registro). Ver *registro.*

EN a single object or element of interest in the data set. A case can represent a customer, a transaction or other basic units of analysis.

SINÓNIMO

ES registro, fila, tupla
EN record
DE Ø

ORIGEN Informática

Informática > ficheros > bases de datos > registro

POSICIÓN DEL CONCEPTO Datos

CONTEXTO

"The next step is deciding on the type of prediction that's most appropriate: (1) *classification*: predicting into what category or class a **case** falls, or (2) *regression*: predicting what number value a variable will have (if it's a variable that varies with time, it's called *time series* prediction). (Fuente: Two Crows *Data Mining In Brief*)

*NOTA: El dato de frecuencia se corresponde tanto a la palabra case como al término.

categorical data (54,2%)

ES datos categóricos DE symbolische Daten

DEFINICIÓN

ES datos categóricos. Aquellos definidos mediante variables categóricas y, consecuentemente, con un rango finito de valores (por ejemplo, sueldo alto – medio – bajo). La información sobre categoría puede aparecer o bien no-ordenada (nominal), del tipo "género" o "ciudad", o bien ordenada (ordinal) del tipo "temperatura alta - media o baja".

EN categorical data fits into a small number of discrete categories (as opposed to continuous). Categorical data is either non-ordered (nominal) such as gender or city, or ordered (ordinal) such as high, medium, or low temperatures.

SINÓNIMO

Ø

ORIGEN

Minería de datos

Minería de datos > preparación > datos

POSICIÓN DEL CONCEPTO

Datos

CONTEXTO

"Data can be *continuous,* having any numerical value (e.g., quantity sold) or *categorical,* fitting into discrete classes (e.g., red, blue, green). **Categorical data** can be further defined as either *ordinal,* having a meaningful order (e.g., high/medium/low), or *nominal,* that is unordered (e.g., postal codes)". (Fuente: Two Crows *Data Mining in Brief*)

categorical variable (54,16%*)

ES variable categórica DE Symbolische Variable

DEFINICIÓN

ES una variable cuyos valores no guardan ninguna relación entre sí. Los valores sólo son útiles como etiquetas. Por ejemplo, *tipo de coche*, donde los valores posibles son Ford, Nissan y Lincoln. A veces se le llama *variable nominal*.

EN a variable whose values do not have any relationship among them. The values are useful only as labels. For example, car-type, where the posible values are Ford, Nissan and Lincoln. Sometimes called a nominal variable.

SINÓNIMO

ES en ocasiones, variable nominal.
EN nominal variable.
DE nominal variable

ORIGEN Minería de datos

Minería de datos > bases de datos > entrada de datos > variables > variable categórica

POSICIÓN DEL CONCEPTO Datos

CONTEXTO

"logistic regression: A linear regression that predicts the proportions of a **categorical** target **variable**, such as type of customer, in a population". (Fuente: Data Mining Page *An Introduction to Data Mining*)

CHAID (29,2%)

ES CHAID DE CHAID

DEFINICIÓN

ES Chi-Square Automatic Interaction Detector. Algoritmo para clasificación en el que se utiliza una función de ajuste basada en chi-cuadrado.

EN an algorithm for fitting categorical trees. It relies on the chi-squared statistic to split the data into small connected data sets.

SINÓNIMO Ø

ORIGEN Matemáticas

Matemáticas > estadística > técnicas > chaid

POSICIÓN DEL CONCEPTO Técnicas y algoritmos

CONTEXTO

"**CHAID** is similar to CART in that it builds a decision tree but it differs in the way that it chooses its splits. Instead of the entropy or Gini metrics for choosing optimal splits the technique relies on the chi square test used in contingency tables to determine which categorical predictor is furthest from independence with the prediction values". (Fuente: *An Overview of Data Mining Techniques.* Excerpted from the book Building Data Mining Applications for CRM by Alex Berson, Stephen Smith, and Kurt Thearling)

classification (75%)

ES clasificación DE Klassifikation

DEFINICIÓN

ES un tipo de problema predictivo que se usa para designar aquellos problemas en los que la variable a predecir es categórica. (Véase *predicción de valor)*.

EN a specialization of predictive modelling for assigning a class identity to a record in a database. The assigned class identity is one from a set of previously known class identities and is based on variables within the record. For example, classification could be used to assign a class identity of “Stay” or “Leave” to a database of credit card client records. See value prediction.

SINÓNIMO

ES aprendizaje supervisado
EN supervised learning
DE Überwachtes lernen, Supervised Learning

ORIGEN Minería de datos

Minería de datos > tipos de problema > problemas predictivos > clasificación

POSICIÓN DEL CONCEPTO

Tipos de problema/Técnicas yalgoritmos/Tipos de resultados

CONTEXTO

“The predictions from different classifiers can be used as input into a meta-learner, which will attempt to combine the predictions to create a final best predicted **classification**. So, for example, the predicted classifications from the tree classifiers, linear model, and the neural network classifier(s) can be used as input variables into a neural network meta-classifier, which will attempt to "learn" from the data how to combine the predictions from the different models to yield maximum classification accuracy”. (Fuente: Statsoft *Data Mining Techniques*)

classification model (75%*)

ES modelo de clasificación DE Klassifikationsmodell

DEFINICIÓN

ES el modelo resultante de efectuar la operación de minería de datos llamada *clasificación*. (Véase modelado *predictivo*, clasificación, error).

EN a model produced by the classification data mining operation. (See predictive modeling, classification, error).

SINÓNIMO Ø

ORIGEN Minería de datos

Minería de datos > tipos de patrones > patrón predictivo > modelo de clasificación

POSICIÓN DEL CONCEPTO Tipos de resultados

CONTEXTO

"This transformation results in a distribution with three modes: one normal mode below zero, one mode at zero, and one normal mode above zero. This transformation may not be valid for a prediction or **classification model**, but the neural clustering algorithm can distinguish between negative, positive, zero, and differing levels of positive and negative values better than with the original distribution". (Fuente: Gary Saarenvita *Mining Customer Data*)

*NOTA: la frecuencia es la de *classification.*

cleaning (25%)

ES limpieza DE Cleaning

DEFINICIÓN

ES denomina un paso en la preparación de datos para minería de datos. Los datos erróneos son detectados y corregidos (por ejemplo, fechas improbables) y se reponen los datos que faltan.

EN refers to a step in preparing data for a data mining activity. Obvious data errors are detected and corrected (e.g., improbable dates) and missing data is replaced.

SINÓNIMO

ES Ø
EN cleansing
DE Datenbereinigung

ORIGEN Minería de datos

Minería de datos > proceso > preparación > limpieza

POSICIÓN DEL CONCEPTO Proceso

CONTEXTO

"Data **cleansing**... the process of ensuring that all values in a dataset are consistent and correctly recorded". (Fuente: Data Mining Page *An Introduction to Data Mining*)

cluster (75%)

ES cluster DE Cluster

DEFINICIÓN

ES nombre que se da alternativamente a *segmento* en una base de datos. Tipo de problema descriptivo que se usa para descubrir asociaciones de registros. (Véase aprendizaje no supervisado).

EN an alternative name for a database segment.

SINÓNIMO

ES segmento
EN segment
DE Ø

ORIGEN Informática - Minería de datos

Minería de datos > tipos de problema > descriptivos > cluster
Informática > inteligencia artificial > aprendizaje no-supervisado > cluster

POSICIÓN DEL CONCEPTO

Técnicas/Tipos de problema/Tipos de resultado

CONTEXTO

"the goal of a **cluster** function is to produce a reasonable segmentation of the set of input records according to some criteria. The criteria itself is defined by the clustering tool. Thus, different clustering functions may produce different segmentations of the set of input records". (Fuente: IBM *Data Mining: Extending the Information Warehouse Framework*)

clustering (70,8%)

ES clustering DE Clustering

DEFINICIÓN

ES término que se usa alternativamente a *segmentación* de una base de datos.

EN an alternative name for database segmentation.

SINÓNIMO

ES segmentación.
EN segmentation.
DE Ø

ORIGEN Informática - Minería de datos

Minería de datos > tipos de problemas > descriptivo > clustering
Informática > inteligencia artificial > aprendizaje no-supervisado > clustering

POSICIÓN DEL CONCEPTO

Técnicas/Tipos de problema/Tipos de resultado

CONTEXTO

"A simple example of **clustering** would be the clustering that most people perform when they do the laundry - grouping the permanent press, dry cleaning, whites and brightly colored clothes is important because they have similar characteristics. And it turns out they have important attributes in common about the way they behave (and can be ruined) in the wash. To "cluster" your laundry most of your decisions are relatively straightforward. There are of course difficult decisions to be made about which cluster your white shirt with red stripes goes into (since it is mostly white but has some color and is permanent press). When clustering is used in business the clusters are often much more dynamic - even changing weekly to monthly and many more of the decisions concerning which cluster a record falls into can be difficult." (Fuente: *An Overview of Data Mining Techniques.* Excerpted from the book Building Data Mining Applications for CRM by Alex Berson, Stephen Smith, and Kurt Thearling)

column (50%)

ES columna DE Spalte

DEFINICIÓN

ES término que se usa en ocasiones para designar *variable*. Cada una de las características que definen una *entidad*.

EN an alternative name for variable.

SINÓNIMO

ES variable.
EN variable.
DE Ø

ORIGEN Informática

Informática > bases de datos > tabla > columna

POSICIÓN DEL CONCEPTO Datos

CONTEXTO

"By looking at this second histogram the viewer is in many ways looking at all of the data in the database for a particular predictor or data **column**. By looking at this histogram it is also possible to build an intuition about other important factors. Such as the average age of the population, the maximum and minimum age". (Fuente: *An Overview of Data Mining Techniques*. Excerpted from the book Building Data Mining Applicationsfor CRM by Alex Berson, Stephen Smith, and Kurt Thearling).

confidence factor (41,7%)

ES factor de confianza DE Konfidentzwert

DEFINICIÓN

ES dada una *regla* A=>B, el número de *registros* en los cuales B aparece junto con A expresado como porcentaje de todos los registros en que A aparece, ya sea con o sin B. El factor indica la fuerza de la afinidad entre los dos *elementos*. (Véase *factor de apoyo, error.* Es una medida de calidad de la regla).

EN given an association rule A=>B, the number of records in which B occurs along with A as a percentage of all records in which A occurs, with or without B. The factor indicates the strength of the affinity between the two items. See support factor.

SINÓNIMO

ES Ø
EN Ø
DE confidence

ORIGEN Informática

Informática > inteligencia artificial > factor de confianza

POSICIÓN DEL CONCEPTO Parámetros de evaluación

CONTEXTO

FÓRMULA $\frac{P(A \cap B)}{P(A)}$

"Some algorithms will create a database of rules, **confidence factors**, and support that can be queried (for example, "Show me all associations in which ice cream is the consequent, that have a confidence factor of over 80% and a support of 2% or more"). (Fuente: Two Crows *Data Mining in Brief*)

continuous (45,8%)

ES contínuo DE kontinuierlich

DEFINICIÓN

ES una variable es contínua cuando toma cualquier valor en un intervalo de números reales. Es decir, el valor no tiene porqué ser un número entero. Continuo es lo contrario de discreto o categórico.

EN Continuous data can have any value in an interval of real numbers. That is, the value does not have to be an integer. Continuous is the opposite of discrete or categorical.

SINÓNIMO

ES variable continua
EN continuous
DE reellwertig

ORIGEN Minería de datos

Minería de datos > proceso > pre-proceso > tipos de variables > continuo
Minería de datos > datos > tipos de datos > continuo

POSICIÓN DEL CONCEPTO Proceso/Datos

CONTEXTO

"Data can be ***continuous,*** having any numerical value (e.g., quantity sold) or *categorical,* fitting into discrete classes (e.g., red, blue, green)." (Fuente: Two Crows *Data Mining in Brief*)

confusion matrix (12,5)

ES matriz de confusión DE Confusion Matrix

DEFINICIÓN

ES matriz de confusión. Se usan en los modelos de clasificación para medir la calidad del modelo. Muestran para cada valor de la clase el número de instancias con ese valor en los datos reales frente a los que predice el modelo. (Véase *error, accuracy, precision*)

EN shows the counts of the actual versus predicted class values. It shows not only how well the model predicts, but also presents the details needed to see exactly where things may have gone wrong.

SINÓNIMO

ES Ø
EN misclassification matrix
DE Ø

ORIGEN Informática

Informática > inteligencia artificial > técnicas > matriz de confusión

POSICIÓN DEL CONCEPTO Resultados

CONTEXTO

"For classification problems, a **confusion matrix** is a very useful tool for understanding results. A confusion matrix [...] shows the counts of the actual versus predicted class values. It shows not only how well the model predicts, but also presents the details needed to see exactly where things may have gone wrong". (Fuente: Two Crows *Data Mining in Brief*)

consequent (12,5%)

ES consecuente DE Konsequenz

DEFINICIÓN

ES consecuente. Si la regla define una asociación entre dos variables, el segundo elemento (el de la derecha) recibe el nombre de consecuente. Por ejemplo, en la relación “cuando un buscador compra un pico, el 14% de las veces compre una pala”, “compra una pala” es el consecuente.

EN when an association between two variables is defined, the second item (or right-hand side) is called the consequent. For example, in the relationship "When a prospector buys a pick, he buys a shovel 14% of the time," "buys a shovel" is the consequent.

SINÓNIMO

ES lado derecho, cabeza de regla
EN right-hand side, rule head
DE consequent

ORIGEN Matemáticas/Informática

Matemáticas > lógica > consecuente
Informática > inteligencia artificial > representación de conocimiento > consecuente

POSICIÓN DEL CONCEPTO Tipos de resultado

CONTEXTO

“Associations are written as A Þ B, where A is called the antecedent or left-hand side (LHS), and B is called the **consequent** or right-hand side (RHS)”. (Fuente: Two Crows *Data Mining in Brief*)

continuous variable (45,83)

ES variable contínua DE kontinuierliche Variable

DEFINICIÓN

ES aquella variable cuyos valores pueden estar formados por un subconjunto de números reales. Por ejemplo, temperatura o ingresos. (Es una característica de la variable; véase también variable discreta y variable nominal).

EN a variable whose values can take on a subset of real numbers. For example, income or temperature.

SINÓNIMO

ES Ø
EN Ø
DE reellwertige Variable

ORIGEN Matemáticas

Matemáticas > estadística > tipos de datos > variable contínua

POSICIÓN DEL CONCEPTO Datos

CONTEXTO

"The concept of bagging (voting for classification, averaging for regression-type problems with **continuous** dependent **variables** of interest) applies to the area of predictive data mining, to combine the predicted classifications (prediction) from multiple models, or from the same type of model for different learning data". (Fuente: Statsoft *Data Mining Techniques*).

CRISP DM (8,33%*)

ES CRISP DM DE CRISP- DM

DEFINICIÓN

ES Cross-industry Standard Process for Data Mining. Modelo de proceso general para minería de datos.

EN Cross-industry Standard Process for Data Mining. CRISP-DM was conceived in late 1996 by DaimlerChrysler (then Daimler-Benz) SPSS (then ISL) and NCR.

SINÓNIMO Ø

ORIGEN Minería de datos

Minería de datos > proceso > CRISP DM

POSICIÓN DEL CONCEPTO Proceso

CONTEXTO

"Horizontally, the **CRISP-DM** methodology distinguishes between the reference model and the user guide. The reference model presents a quick overview of phases, tasks and their outputs and describes what to do in a data mining project. The user guide gives more detailed tips and hints for each phase and each task within a phase and depicts how to do a data mining project". (Fuente: CRISP DM 1.0 Step-by-Step Data Mining Guide)

NOTA*: La baja frecuencia de aparición en el corpus se debe a que es terminología propia de Clementine.

cross validation (50%)

ES validación cruzada DE cross validation

DEFINICIÓN

ES validación cruzada. Método para estimar la precisión de una clasificación o modelo de regresión. Los datos se dividen en varios sub-conjuntos, cada uno de los cuales se usa sucesivamente para probar un modelo ajustado a las partes restantes.

EN A method of estimating the accuracy of a classification or regression model. The data set is divided into several parts, with each part in turn used to test a model fitted to the remaining parts.

SINÓNIMO Ø

ORIGEN Informática

Informática > inteligencia artificial > clasificación

POSICIÓN DEL CONCEPTO Técnicas y algoritmos

CONTEXTO

"The exploration of data can only serve as the first stage of data analysis and its results can be treated as tentative at best as long as they are not confirmed, e.g., **crossvalidated**, using a different data set (or and independent subset). If the result of the exploratory stage suggests a particular model, then its validity can be verified by applying it to a new data set and testing its fit (e.g., testing its *predictive validity*)".

data mining (100%)

ES minería de datos/data mining* DE Data Mining

DEFINICIÓN

ES minería de datos (también data mining [data mainin]. Proceso consistente en extraer de grandes bases de datos información aplicable, válida y previamente desconocida, para luego emplearla en el apoyo a toma de decisiones empresariales de gran trascendencia. (Véase definición de Piatetsky – Shapiro)

EN the process of extracting previously unknown, valid and actionable information from large databases and then using the information to inform crucial business decissions.

SINÓNIMO

ES En el mundo académico, KDD (Knowledge Discovery in Databases

EN Ø

DE Wissensentdeckung. Knowledge Discovery

ORIGEN Informática/Estadística

Originalmente es una parte del proceso de KDD -la que ahora se denomina modeling- y posteriormente pasa a designar todo el proceso.

POSICIÓN DEL CONCEPTO Minería de datos

En un punto de intersección de la informática, la estadística y otros campos.

CONTEXTO

"**Data mining**, *the extraction of hidden predictive information from large databases*, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses". (Fuente: Data Mining Page *An Introduction to Data Mining*)

*NOTA: En español se usan indistintamente el término original inglés y la traducción préstamo.

data preparation (50%)

ES preparación de datos DE Datenvorbereitung

DEFINICIÓN

ES fase del modelo de proceso de CRISP DM que comprende la selección de los datos así como su limpieza, construcción, integración y asignación de formato.

EN a phase in the CRISP DM process which implies the selection, cleaning, construction, integration and formatting of data.

SINÓNIMO

ES Ø
EN Ø
DE Datenvorberarbeitung

ORIGEN Minería de datos

Minería de datos > proceso > preparación de datos

POSICIÓN DEL CONCEPTO Proceso

CONTEXTO

"The reports in the **data preparation** phase focus on the pre-processing steps that produce the data to be mined". (Fuente: CRISP DM 1.0 Step-by-Step Data Mining Guide)

data understanding (62,5%*)

ES comprensión de datos DE Data Understanding

DEFINICIÓN

ES fase del modelo de proceso de CRISP DM que comprende reunir los primeros datos, describirlos, explorarlos y verificar su calidad.

EN a phase in the CRISP DM process model which includes collecting initial data, describing, exploring and verifying their quality.

SINÓNIMO Ø

ORIGEN Minería de datos

Minería de datos > proceso > comprensión de datos

POSICIÓN DEL CONCEPTO Proceso

CONTEXTO

"The results of the Data Understanding phase are usually documented in several reports. Ideally, these reports should the written while performing the respective tasks. The reports describe the datasets that are explored during data understanding. For the final report, a summary of the most relevant parts is sufficient". (Fuente: CRISP DM 1.0 *Step-by-Step Data Mining Guide*)

Nota*: porcentaje elevado dudoso por peculiaridades del analizador de textos.

data warehouse* (41,7%)

ES data warehouse DE Data Warehouse

DEFINICIÓN

ES conjunto integrado de datos, documentado, con relevancia temporal, referentes a un tema concreto, que se utilizan para ayudar en la toma de decisiones empresariales relevantes. (Véase definición de Bill Immon).

EN a subject-oriented, documented, integrated and time-dimensional collection of data that is used to inform crucial business decissions.

SINÓNIMO

ES repositorio de datos, almacén de datos.
EN Ø
DE Ø

ORIGEN Informática

Informática > Bases de datos > data warehouse

POSICIÓN DEL CONCEPTO Datos

CONTEXTO

"The ideal starting point is a **data warehouse** containing a combination of internal data tracking all customer contact coupled with external market data about competitor activity. Background information on potential customers also provides an excellent basis for prospecting. This warehouse can be implemented in a variety of relational database systems: Sybase, Oracle, Redbrick, and so on, and should be optimized for flexible and fast data access". (Fuente: Data Mining Page *An Introduction to Data Mining*)

*Nota: La elevada presencia porcentual de este término en el corpus de textos original llevó a su inclusión en el glosario, pese a que los técnicos participantes en el proyecto no fueran favorables a considerarlo término propio de minería de datos.

decision tree (83,3%)

ES árbol de decisión DE Entscheidungsbaum

DEFINICIÓN

ES árbol de decisión. Grupo de *nodos* y *hojas* conectados para representar la salida de un algoritmo de *clasificación.*

EN a group of nodes and leaves linked together to represent the output of a classification algorithm.

SINÓNIMO

ES Ø
EN Ø
DE Decision Tree

ORIGEN Informática

Informática > inteligencia artificial > árbol de decisión

POSICIÓN DEL CONCEPTO

Tipos de respuesta/Técnicas y algoritmos

CONTEXTO

"The increased power of computers and their lower cost, coupled with the need to analyze enormous data sets with millions of rows, have allowed the development of new techniques based on a brute-force exploration of possible solutions. New techniques include relatively recent algorithms like neural nets and **decision trees**, and new approaches to older algorithms such as discriminant analysis". (Fuente: Two Crows *Data Mining in Brief*)

degree of fit (25%)

ES grado de ajuste DE Grad der Anpassung

DEFINICIÓN

ES una medida de la proximidad con que el modelo se ajusta a los datos entrenados. Una medida común es r-cuadrado.

EN A measure of how closely the model fits the training data. A common measure is r-square.

SINÓNIMO Ø

ORIGEN Informática

Informática > inteligencia artificial > clasificación > grado de ajuste

POSICIÓN DEL CONCEPTO Parámetros de evaluación

CONTEXTO

If the result of the exploratory stage suggests a particular model, then its validity can be verified by applying it to a new data set and **testing its fit** (e.g., testing its *predictive validity*). (Fuente: Statsoft *Data Mining Techniques*).

deployment (29,2%)

ES deployment DE Anwendungsphase

DEFINICIÓN

ES una vez que el modelo ha sido entrenado y evaluado, se utiliza para analizar nuevos datos y hacer predicciones. Es el nombre que recibe este uso del modelo.

EN After the model is trained and validated, it is used to analyze new data and make predictions. This use of the model is called deployment.

SINÓNIMO

ES implantación
EN Ø
DE Ø

ORIGEN Informática

POSICIÓN DEL CONCEPTO Proceso

CONTEXTO

"The concept of **deployment** in predictive data mining refers to the application of a model for prediction or classification to new data. After a satisfactory model or set of models has been identified (trained) for a particular application, one usually wants to deploy those models so that predictions or predicted classifications can quickly be obtained for new data. For example, a credit card company may want to deploy a trained model or set of models (e.g., neural networks, meta-learner) to quickly identify transactions which have a high probability of being fraudulent". (Fuente: Statsoft *Data Mining Techniques*)

deviation/outlier detection (41,7%)

ES detección de desviaciones
DE Abweichungserkennung – Entdeken von Ausreissern

DEFINICIÓN

ES operación de minería de datos encaminada a la detección de comportamientos anómalos en las bases de datos y determinar su origen y naturaleza.

EN a data mining operation to detect outliers in databases and determine their nature and cause.

SINÓNIMO

ES detección de anomalías.
EN Ø
DE outlier detection

ORIGEN Matemáticas

Matemáticas > estadística > detección de desviaciones

POSICIÓN DEL CONCEPTO Tipos de problema

CONTEXTO

"Domain knowledge is also critical for **outlier detection** needed to clean data and avoid classic problems such as a juvenile crime committed by a 80-year-old "child". If a data mining model were build using the data in Figure 1, it is possible that outliers (most likely caused by incorrect data entry) will skew the resulting model (especially the zero-year-old children, which are more reasonable than eighty-year-old children)". (Fuente: Kurt Thearling et al *Visualizing Data Mining Models*).

dimension (58,3%)

ES dimensión DE Dimension

DEFINICIÓN

ES término que se usa a veces en lugar de *variable. (Nota: en data warehousing adquiere un significado diferente – modelo multidimensional)*

EN an alternative name for variable.

SINÓNIMO

ES variable
EN variable
DE Ø

ORIGEN Informática

Informática > bases de datos > dimensión

POSICIÓN DEL CONCEPTO Datos

CONTEXTO

"The problem in using visualization stems from the fact that models have many **dimensions** or variables, but we are restricted to showing these dimensions on a two-dimensional computer screen or paper. For example, we may wish to view the relationship between credit risk and age, sex, marital status, own-or-rent, years in job, etc. Consequently, visualization tools must use clever representations to collapse *n* dimensions into two". (Fuente: Two Crows Data Mining in Brief")

discrete (33,3%)

ES discreto DE discret

DEFINICIÓN

ES dato que tiene un conjunto finito de valores. Discreto es lo contrario a continuo.

EN data item that has a finite set of values. Discrete is the opposite of continuous.

SINÓNIMO

ES categórico
EN categorical
DE Ø

ORIGEN Minería de datos

Minería de datos > proceso > pre-proceso > tipos de variables > discreto
minería de datos > datos > tipos de datos > discreto

POSICIÓN DEL CONCEPTO Proceso/Datos

CONTEXTO

"Data can be *continuous,* having any numerical value (e.g., quantity sold) or *categorical,* fitting into **discrete** classes (e.g., red, blue, green). Categorical data can be further defined as either *ordinal,*having a meaningful order (e.g., high/medium/low), or *nominal,* that is unordered (e.g., postal codes). (Fuente: Two Crows *Data Mining in Brief*)

discretization (4,2%)

ES discretización DE Diskretisierung

DEFINICIÓN

ES el acto de asignar un conjunto de valores discretos a una *variable continua*. Por ejemplo, la variable continua *ingresos* podría ser discretizada en una nueva variable llamada *grupo de ingresos* con valores 1, 2, etc. que correspondería a valores de ingresos reales de 10.000$, 20.000$, etc. Véase *intervalos*.

EN the act of assigning a set of discrete values to a continuous variable. For example, the continuous variable income could be discretized into a new variable called income-group with values of 1, 2, and so on to correspond to original income values of $10,000, $20,000 and so on.

SINÓNIMO Ø

ORIGEN Matemáticas

Matemáticas > discretización

POSICIÓN DEL CONCEPTO Variables/Proceso

CONTEXTO

“Sometimes it's necessary to manually prescribe the ranges for a particular **discretization** for comparison purposes. Existing data or external data -- such as government census, survey, or list-brokered data -- may contain variables that are collected in buckets. In order to compare your internal data to an external source, you must discretize the data using the same ranges. You can do this discretization using Intelligent Miner's data processing functions”. (Fuente: Gary Saarenvirta *Mining Customer Data*).

discriminant analysis (16,7%)

ES análisis discriminante DE Diskriminanzanalyse

DEFINICIÓN

ES método estadístico basado en la probabilidad máxima para determinar los límites que separan los datos en categorías.

EN A statistical method based on maximum likelihood for determining boundaries that separate the data into categories.

SINÓNIMO Ø

ORIGEN Matemáticas

Matemáticas > estadística > análisis discriminante

POSICIÓN DEL CONCEPTO Técnicas y algoritmos

CONTEXTO

"Suppose your data mining project includes tree classifiers, such as C&RT or CHAID, linear **discriminant analysis** (e.g., see GDA), and *Neural Networks*. Each computes predicted classifications for a crossvalidation sample, from which overall goodness-of-fit statistics (e.g., misclassification rates) can be computed. Experience has shown that combining the predictions from multiple methods often yields more accurate predictions than can be derived from any one method". (Fuente: Statsoft Data Mining Techniques).

evaluation (33,3%)

ES evaluación DE Evaluation

DEFINICIÓN

ES fase del modelo de proceso de CRISP DM que comprende la evaluación de los resultados obtenidos, la revisión del proceso de minería realizado y el establecimiento de los pasos siguientes.

EN a phase in the CRISP DM process model which comprises the evaluation of the results as well as the revision of the data mining process and determining the next steps.

SINÓNIMO Ø

ORIGEN Minería de datos

Minería de datos > proceso > evaluación

POSICIÓN DEL CONCEPTO Proceso

CONTEXTO

"**Evaluation**: assessment of data mining results with respect to business success criteria. This report compares the data mining results with the business objectives and the business success criteria. (Fuente: CRISP DM 1.0 *Step-by-Step Data Mining Guide*)

exploratory data analisis (37,5%)

ES análisis de datos exploratorio DE explorative Datenanalyse

DEFINICIÓN

ES término general que designa todo aquel análisis de datos que investigue estructuras y/o contenidos de datos de implantación profunda usando técnicas de OLAP (OnLine Analitical Processing) y estadística, pero no operaciones y técnicas de minería de datos.

EN a general term covering any data analysis to investigate deep-seated data contents and/or structure, typically using techniques from statistics and OLAP but excluding data mining operations and techniques.

SINÓNIMO Ø

ORIGEN Matemáticas/Minería de datos/Data warehouse

Matemáticas > estadística > análisis de datos exploratorio

Minería de datos > proceso > análisis de datos exploratorio

POSICIÓN DEL CONCEPTO Proceso

CONTEXTO

"Then, depending on the nature of the analytic problem, this first stage of the process of data mining may involve anywhere between a simple choice of straightforward predictors for a regression model, to elaborate **exploratory analyses** using a wide variety of graphical and statistical methods (see *Exploratory Data Analysis (EDA)*) in order to identify the most relevant variables and determine the complexity and/or the general nature of models that can be taken into account in the next stage". (Fuente: Statsoft *Data Mining Techniques*)

feature (66,6%)

ES característica DE Merkmal

DEFINICIÓN

ES término que se usa a veces en lugar de *variable.*

EN an alternative name for variable.

SINÓNIMO

ES variable
EN variable
DE Attribut, Feature

ORIGEN Informática

Informática > bases de datos > característica

POSICIÓN DEL CONCEPTO Datos

CONTEXTO

"Using decision tables as a model representation generates a simple but large model. A full decision table theoretically contains the entire dataset, which may be very large. Therefore simplification is essential. The MineSet decision table arranges the model into levels based on the importance of each **feature** in the table". (Fuente: Kurt Thearling et al *Visualizing Data Mining Models*)

feed forward (20,8%)

ES alimentación hacia delante DE feed forward

DEFINICIÓN

ES una red neuronal en la que los datos fluyen sólo en una dirección, desde las entradas hacia las salidas.

EN a neural network in which the signals only flow in one direction, from the inputs to the outputs.

SINÓNIMO Ø

ORIGEN Informática

Informática > inteligencia artificial > redes neuronales > alimentación hacia delante

POSICIÓN DEL CONCEPTO Técnicas y algoritmos

CONTEXTO

"***Feed forward:*** The value of the output node is calculated based on the input node values and a set of initial weights. The values from the input nodes are combined in the hidden layers, and the values of those nodes are combined to calculate the output value". (Fuente: Two Crows *Data Mining in Brief*)

field (50%)

ES campo DE Feld

DEFINICIÓN

ES término que se usa a veces en lugar de *variable*.

EN an alternative name for variable.

SINÓNIMO

ES variable
EN variable
DE Ø

ORIGEN Informática

Informática > bases de datos > campo

POSICIÓN DEL CONCEPTO Datos

CONTEXTO

"Many times, while performing a query, a request is made to compute functions related to the records being inspected during the query (e.g., count the number of records, find the average of a given **field** of the records, etc.) All these operations result in additional information being returned together with the query". (Fuente: IBM *Data Mining: Extending the Information Warehouse Framework*)

genetic algorithms (37,5%)

ES algoritmos genéticos DE Genetischе Algoritmen

DEFINICIÓN

ES método generado por ordenador para crear y probar combinaciones de posibles parámetros de entrada para encontrar el resultado* óptimo. Usa procesos basados en conceptos de la evolución natural tales como combinación genética, mutación y selección natural.

EN a computer-based method of generating and testing combinations of possible input parameters to find the optimal output. It uses processes based on natural evolution concepts such as genetic combination, mutation and natural selection.

SINÓNIMO Ø

ORIGEN Informática

Informática > inteligencia artificial > algoritmos genéticos

POSICIÓN DEL CONTEXTO Técnicas y algoritmos

CONTEXTO

"The most commonly used techniques in data mining are: [...]**Genetic algorithms**: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution". (Fuente: Data Mining Page *An Introduction to Data Mining*)

*Nota: La traducción habitual de "*output*" en minería de datos es "resultado", en contraste con el habitual "salida" de informática.

hidden layer (37,5%)

ES nivel oculto DE Hidden Layer

DEFINICIÓN

ES conjunto de *unidades de procesamiento* en una red neuronal situadas entre los niveles de entrada y salida, que se utilizan para calcular la salida de la red.

EN a set of processing units in a neural network, positioned between the input and output layers and used to calculate the network output.

SINÓNIMO

ES Ø
EN Ø
DE versteckte Schicht, Hidden-Schicht

ORIGEN Informática

Informática > inteligencia artificial > clasificación/aprendizaje > redes neuronales > hidden layer

POSICIÓN DEL CONCEPTO Técnicas y algoritmos

CONTEXTO

"The structure of a neural network looks something like the following: [graph] The bottom layer represents the input layer, in this case with 5 inputs labels X1 through X5. In the middle is something called the **hidden layer**, with a variable number of nodes. It is the hidden layer that performs much of the work of the network. The output layer in this case has two nodes, Z1 and Z2 representing output values we are trying to determine from the inputs. For example, predict sales (output) based on past sales, price and season (input). Each node in the hidden layer is fully connected to the inputs which means that what is learned in a hidden node is based on all the inputs taken together". (Fuente: QUB *Data Mining techniques*)

induction (50%)

ES inducción DE Induktion

DEFINICIÓN

ES técnica que infiere generalizaciones a partir de la información que aportan los datos. En minería de datos todo el proceso es siempre inductivo.

EN a technique that infers generalizations from the information in the data.

SINÓNIMO

ES Ø
EN Ø
DE induktives Lernen

ORIGEN Informática

Informática > inteligencia artificial > técnicas de descubrimiento de conocimiento > inducción

POSICIÓN DEL CONCEPTO Técnicas y algoritmos

CONTEXTO

"There are two main inference techniques available ie deduction and induction [...] **Induction** has been described earlier as the technique to infer information that is generalised from the database as in the example mentioned above to infer that each employee has a manager. This is higher level information or knowledge in that it is a general statement about objects in the database. The database is searched for patterns or regularities". (Fuente: QUB Parallel Computer Centre *Data Mining Techniques*)

item (66,7%)

ES item DE Artikel

DEFINICIÓN

ES subapartado en una transacción que resulta identificable como único, típicamente por medio de una clave de registro de tipo código universal de producto (UPC) o número de cliente. Íntimamente ligado a algoritmo a priori. Algoritmo creado expresamente para minería de datos. (Véase Agrawal 1996)

EN a subpart of a transaction that is uniquely identifiable, typically by a record key such as a universal product code (UPC) or customer number.

SINÓNIMO Ø

ORIGEN Minería de datos

Minería de datos > técnicas y algoritmos > asociación > item

POSICIÓN DEL CONCEPTO Técnicas y algoritmos

CONTEXTO

"It's easy to determine the proportion of transactions that contain a particular **item** or item set: simply count them. The frequency with which a particular association (e.g., the item set "hammers and nails") appears in the database is called its *support* or *prevalence"*. (Fuente: Two Crows *Data Mining in Brief*)

k-means (25%)

ES K-medias DE K-Means

DEFINICIÓN

ES técnica de clustering que define k clusters y les asigna registros sucesivamente en función a las distancias desde la media de cada cluster hasta que se encuentra una solución estable.

EN a approach to clustering that defines k clusters and iteratively assigns records to clusters based on distances from the mean of each cluster until a stable solution is found.

SINÓNIMO

ES k-means
EN Ø
DE Ø

ORIGEN Estadística

Estadística > técnicas de clustering > K-means

POSICIÓN DEL CONCEPTO Técnicas y algoritmos

CONTEXTO

"After you have found clusters that reasonably segment your database, these clusters may then be used to classify new data. Some of the common algorithms used to perform clustering include Kohonen feature maps and **K-means**." (Fuente: Two Crows Data Mining in Brief")

k-nearest neighbor (29,2%)

ES vecino k-cercano DE k-nächste Nachbarn

DEFINICIÓN

ES método de clasificación que clasifica (redundancia inevitable N. del A.) un punto por medio del cálculo de las distancias entre el punto y otros puntos en el conjunto de datos de entrenamiento. Entonces asigna el punto a la clase que sea más común entre sus vecinos K-próximos (donde K es un número entero).

EN a classification method that classifies a point by calculating the distances between the point and points in the training data set. Then it assigns the point to the class that is most common among its k-nearest neighbors (where k is an integer).

SINÓNIMO

ES vecino más cercano
EN Ø
DE k-nearest Neighbors

ORIGEN Matemáticas

Matemáticas > estadística > técnicas de clasificación > vecino k-cercano

POSICIÓN DEL CONCEPTO Técnicas y algoritmos

CONTEXTO

"One of the improvements that is usually made to the basic nearest neighbor algorithm is to take a vote from the **"K" nearest neighbors** rather than just relying on the sole nearest neighbor to the unclassified record. In Figure 1.4 we can see that unclassified example C has a nearest neighbor that is a defaulter and yet is surrounded almost exclusively by records that are good credit risks. In this case the nearest neighbor to record C is probably an outlier - which may be incorrect data or some non-repeatable idiosyncrasy. In either case it is more than likely that C is a non-defaulter yet would be predicted to be a defaulter if the sole nearest neighbor were used for the prediction. In cases like these a vote of the 9 or 15 nearest neighbors would provide a better prediction accuracy for the system than would just the single nearest neighbor. Usually this is accomplished by simply taking the majority or plurality of predictions from the K nearest neighbors if the prediction column is a binary or categorical or taking the average value of the prediction column from the K nearest neighbors". (Fuente: An Overview of Data Mining Techniques Excerpted from the book *Building Data Mining Applications for CRM* by Alex Berson, Stephen Smith, and Kurt Thearling)

Kohonen feature map (25%)

ES mapa de Kohonen DE Kohonen-carte

DEFINICIÓN

ES modelo de *red neuronal* formado por *neuronas* dispuestas en un nivel de entrada y un nivel de salida. Todos los procesadores del nivel de entrada están conectados a cada procesador del nivel de salida. El *algoritmo de aprendizaje* que se emplea implica que haya competencia entre unidades para cada entrada, y la declaración de una unidad ganadora. Se utiliza en segmentación neuronal para particionar una base de datos en *segmentos*.

EN a neural network model composed of processing units arranged in an input layer and an output layer. All processors in the input layer are connected to each processor in the output layer. The learning algorithm used involves competition between units for each input pattern and the declaration of a winning unit. Used in neural segmentation to partition a database into segments.

SINÓNIMO

ES Ø
EN SOM
DE SOM

ORIGEN Informática

Informática > inteligencia artificial > aprendizaje no-supervisado > mapa de Kohonen

POSICIÓN DEL CONCEPTO Técnicas y algoritmos

CONTEXTO

"After you have found clusters that reasonably segment your database, these clusters may then be used to classify new data. Some of the common algorithms used to perform clustering include **Kohonen feature maps** and K-means". (Fuente: Two Crows Data Mining in Brief)

layer (45,8%)

ES capa DE Schicht

DEFINICIÓN

ES los nodos en una red neuronal se suelen agrupar en capas, y cada capa está descrita como de entrada, de salida u oculta. Hay tantos nodos de entrada como variables de entrada (independientes) haya y tantos nodos de salida como variables de salida (dependientes). Suele haber uno o dos nodos ocultos.

EN nodes in a neural net are usually grouped into layers, with each layer described as input, output or hidden. There are as many input nodes as there are input (independent) variables and as many output nodes as there are output (dependent) variables. Typically, there are one or two hidden layers.

SINÓNIMO

ES Ø
EN Ø
DE Layer

ORIGEN Informática

Informática > inteligencia artificial > clasificación > redes neuronales > capa

POSICIÓN DEL CONCEPTO Técnicas y algoritmos

CONTEXTO

"A neural network (Figure 4) starts with an *input **layer***, where each node corresponds to a predictor variable. These input nodes are connected to a number of nodes in a *hidden **layer***. Each input node is connected to every node in the hidden layer. The nodes in the hidden layer may be connected to nodes in another hidden layer, or to an *output layer*. The output **laye**r consists of one or more response variables". (Fuente: Two Crows *Data Mining in Brief*)

leaf (8,3%)

ES hoja DE Blatt

DEFINICIÓN

ES la parte de *un árbol de decisión* que representa los extremos.

EN the part of a decision tree that represents an end-point at which records are collected.

SINÓNIMO Ø

ORIGEN Informática

Informática > inteligencia artificial > clasificación > árboles de decisión > hoja

POSICIÓN DEL CONCEPTO Técnicas y algoritmos

CONTEXTO

"Depending on the algorithm, each node may have two or more branches. For example, CART generates trees with only two branches at each node. Such a tree is called a binary tree. When more than two branches are allowed it is called a multiway tree. Each branch will lead either to another decision node or to the bottom of the tree, called a **leaf** node. By navigating the decision tree you can assign a value or class to a case by deciding which branch to take, starting at the root node and moving to each subsequent node until a leaf node is reached. Each node uses the data from the case to choose the appropriate branch". (Fuente: Two Crows *Data Mining in Brief*)

learning (70,8%)

ES aprendizaje DE Lernen

DEFINICIÓN

ES entrenar modelos (estimando sus parámetros) a partir de datos existentes.

EN training models (estimating their parameters) based on existing data.

SINÓNIMO Ø

ORIGEN Informática

POSICIÓN DEL CONCEPTO Proceso

CONTEXTO

"A simple algorithm for boosting works like this: Start by applying some method (e.g., a tree classifier such as C&RT or CHAID) to the **learning** data, where each observation is assigned an equal weight. Compute the predicted classifications, and apply weights to the observations in the learning sample that are inversely proportional to the accuracy of the classification. In other words, assign greater weight to those observations that were difficult to classify (where the misclassification rate was high), and lower weights to those that were easy to classify (where the misclassification rate was low"). (Fuente: *Statsoft Data Mining Techniques*)

learning algorithm (66,7%)

ES algoritmo de aprendizaje DE Lernalgorithmus

DEFINICIÓN

ES conjunto preestablecido de normas que se utilizan durante el proceso de creación de un *modelo predictivo*.

EN a set of well-defined rules used during the training process to build a predictive model.

SINÓNIMO Ø

ORIGEN Informática

Informática > inteligencia artificial > aprendizaje > algoritmo de aprendizaje

POSICIÓN DEL CONCEPTO Proceso

CONTEXTO

"Note that some weighted combination of predictions (weighted vote, weighted average) is also possible, and commonly used. A sophisticated machine **learning algorithm** for generating weights for weighted prediction or voting is the boosting procedure" (Fuente: Statsoft *Data Mining Techniques*)

linear regression (45,8%)

ES regresión lineal DE Lineare Regression

DEFINICIÓN

ES técnica matemática de estimación de un modelo lineal para conseguir un campo de salida continuo.

EN a mathematical technique for estimating a linear model for a continuous output field.

SINÓNIMO Ø

ORIGEN Matemáticas

Matemáticas > estadística > regresión lineal

POSICIÓN DEL CONCEPTO Técnicas y algoritmos

CONTEXTO

"In the simplest case, regression uses standard statistical techniques such as **linear regression.** Unfortunately, many real-world problems are not simply linear projections of previous values. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. (Fuente: Two Crows *Data Mining in Brief*)

link analysis (41,7%)

ES link análisis De Zusammenhangsanlyse

DEFINICIÓN

ES operación de minería de datos encaminada a detectar afinidades entre *elementos* tanto dentro de como entre *transacciones*. El análisis de enlaces incluye las técnicas de *minería de datos* conocidas como *localización de asociaciones*, *localización de patrones secuenciales* y *localización de secuencias cuasi-simultáneas*.

EN a data mining operation to detect affinities between items both within and between transactions, possibly over time. Link analysis includes the associations discovery, sequential pattern discovery and similar time sequence discovery data mining techniques.

SINÓNIMO

ES Ø
EN associations discovery/finding
DE Ø

ORIGEN Minería de datos

Minería de datos > tipos de problemas > link análisis

POSICIÓN DEL CONCEPTO Tipos de problemas

CONTEXTO

"**Link analysis** is a descriptive approach to exploring data that can help identify relationships among values in a database. The two most common approaches to link analysis are *association discovery* and *sequence discovery*. Association discovery finds rules about items that appear together in an event such as a purchase transaction". (Fuente: Two Crows: *Data Mining in Brief*)

logistic regression (25%)(logistic discriminant analysis)

ES regresión logística (análisis discriminante logístico)
DE logistische Regression (logistische Diskriminantz-analyse)

DEFINICIÓN

ES regresión logística (análisis discriminante logístico). Una generalización de la regresión lineal. Se utiliza para predecir una variable binaria (con valores tales como sí/no o 0/1). Un ejemplo de su uso es modelar la probabilidad de que el tomador de un préstamo acabará siendo un cliente moroso en base a sus ingresos, deudas y edad.

EN a generalization of linear regression. It is used for predicting a binary variable (with values such as yes/no or 0/1). An example of its use is modeling the odds that a borrower will default on a loan based on the borrower's income, debt and age.

SINÓNIMO Ø

ORIGEN Matemáticas

Matemáticas > estadística > técnicas de regresión > regresión logística

POSICIÓN DEL CONCEPTO Técnicas y algoritmos

CONTEXTO

"When trying to predict a customer response that is just yes or no (e.g. they bought the product or they didn't or they defaulted or they didn't) the standard form of a line doesn't work. Since there are only two possible values to be predicted it is relatively easy to fit a line through them. However, that model would be the same no matter what predictors were being used or what particular data was being used. Typically in these situations a transformation of the prediction values is made in order to provide a better predictive model. This type of regression is called **logistic regression** and because so many business problems are response problems, logistic regression is one of the most widely used statistical techniques for creating predictive models". (Fuente: *An Overview of Data Mining Techniques* Excerpted from the book Building Data Mining Applications for CRM by Alex Berson, Stephen Smith, and Kurt Thearling)

machine learning (50%)

ES aprendizaje automático DE Maschinelles Lernen

DEFINICIÓN

ES conjunto de métodos que permiten a un ordenador aprender a realizar una tarea determinada, como puede ser la toma de decisiones, realizar estimaciones, clasificaciones, predicciones... etc, sin tener que programarlo previamente para ello. También designa el proceso de aplicar tales métodos a los datos.

EN a set of methods for allowing a computer to learn a specific task -such as decission making, estimation, classification, prediction...- without having to be programmed to do so. Also the process of applying such methods to data.

SINÓNIMO Ø

ORIGEN Informática

Informática > inteligencia artificial > aprendizaje automático

POSICIÓN DEL CONCEPTO Técnicas y algoritmos

CONTEXTO

"**Machine learning**, computational learning theory, and similar terms are often used in the context of data mining to denote the application of generic model-fitting or classification algorithms for predictive data mining. Unlike traditional statistical data analysis, which is usually concerned with the estimation of population parameters by statistical inference, the emphasis in data mining (and machine learning) is usually on the accuracy of prediction (predicted classification), regardless of whether or not the "models" or techniques that are used to generate the prediction is interpretable or open to simple explanation". (Fuente: Statsoft *Data Mining Techniques*)

market basket anaysis (41,6%)

ES análisis de cesta de la compra DE Warenhorbanalyse

DEFINICIÓN	
ES	aplicación de modelos basados en asociación que busca describir pares de clusters de artículos que tienden a ser adquiridos por el mismo cliente al mismo tiempo.
EN	an application of association-based models that attempts to describe pairs of clusters of items that tend to be purchased by the same customer at the same time.
SINÓNIMO	Ø
ORIGEN	Economía
POSICIÓN DEL CONCEPTO	Tipos de problema

Economía > marketing > análisis de cesta de la compra

CONTEXTO

"Another notable marketing application is market-basket analysis (Agrawal et al. 1996) systems, which find patterns such as, "If customer bought X, he/she is also likely to buy Y and Z." Such patterns are valuable to retailers". (Fuente: *Fayyad et al From data Mining to KDD*)

missing data (45,8%)

ES missing data DE fehlende Daten

DEFINICIÓN

ES los datos pueden estar ausentes por no haber sido medidos, o contestados, por ser desconocidos o por haberse perdido. Los métodos de minería de datos varían en la forma de tratar los valores ausentes. Típicamente, ignoran los valores ausentes, u omiten cualquier registro que los contenga, o sustituyen los valores ausentes con la moda o la media, o infieren los datos ausentes a partir de los existentes. (Véase *non aplicable data*).

EN data values can be missing because they were not measured, not answered, were unknown or were lost. Data mining methods vary in the way they treat missing values. Typically, they ignore the missing values, or omit any records containing missing values, or replace missing values with the mode or mean, or infer missing values from existing values.

SINÓNIMO

ES datos vacíos
EN Ø
DE Ø

ORIGEN Informática

Informática > bases de datos > missing data

POSICIÓN DEL CONCEPTO Datos

CONTEXTO

”Sometimes the value for a field is missing. Inconsistencies must be identified and removed when consolidating data from multiple sources. **Missing data** can be a particularly pernicious problem. If you have to throw out every record with a field missing, you may wind up with a very small database or an inaccurate picture of the whole database. The fact that a value is missing may be significant in itself. Perhaps only wealthy customers regularly leave the “income” field blank, for instance. It can be worthwhile to create a new variable to identify missing values, build a model using it, and compare the results with those achieved by substituting for the missing value to see which leads to better predictions”. (Fuente: Two Crows *Data Mining in Brief*)

model (95,8%)

ES modelo DE Modell

DEFINICIÓN

ES la función de los proyectos de minería de datos es la creación de un modelo. Un modelo puede ser descriptivo o predictivo. Un modelo descriptivo ayuda en la comprensión de comportamientos o procesos subyacientes. Por ejemplo, un modelo de asociación describe el comportamiento del consumidor. Un modelo predictivo es una ecuación o conjunto de reglas/normas que hacen posible predecir un valor no-visto o no-medido (la variable dependiente o salida) a partir de otros valores conocidos (variables independientes o entrada).

EN the goal of data mining projects is the production of a model. A model can be descriptive or predictive. A descriptive model helps in understanding underlying processes or behavior. For example, an association model describes consumer behavior. A predictive model is an equation or set of rules that makes it possible to predict an unseen or unmeasured value (the dependent variable or output) from other, known values (independent variables or input).

SINÓNIMO Ø

ORIGEN Minería de datos

Minería de datos > modelo

POSICIÓN DEL CONCEPTO Genérico

CONTEXTO

"If someone told you that he had a **model** that could predict customer usage how would you know if he really had a good model? The first thing you might try would be to ask him to apply his model to your customer base - where you already knew the answer. With data mining, the best way to accomplish this is by setting aside some of your data in a vault to isolate it from the mining process. Once the mining is complete, the results can be tested against the data held in the vault to confirm the model's validity. If the model works, its observations should hold for the vaulted data". (Fuente: Data Mining Page *An Introduction to Data Mining*).

modeling (58,3%)

ES modeling DE Modellierung

DEFINICIÓN

ES fase del modelo de proceso de CRISP DM que comprende la selección de técnicas de modelado, la generación de diseños de tests y la construcción y evaluación de modelos.

EN a phase in the CRISP DM process model which involves selecting modeling techniques, generating test designs and building and assessing models.

SINÓNIMO

ES Ø
EN Ø
DE Modellbildung

ORIGEN Minería de datos

Minería de datos > proceso > modeling

POSICIÓN DEL CONCEPTO Proceso

CONTEXTO

"Most of the models and algorithms discussed in this section can be thought of as generalizations of the standard workhorse of **modeling**, the linear regression model". (Fuente: Two Crows Data Mining in Brief)

neural network (83,3%)

ES red neuronal DE neuronales Netz

DEFINICIÓN

ES modelo informático basado en la arquitectura del cerebro consistente en múltiples *unidades de procesamiento* (neuronas) simples conectadas por uniones a las que se asocia un peso. Conjunto de *unidades de procesamiento* y conexiones adaptativo que se diseña para realizar una función de procesamiento específica. Las redes neuronales se utilizan para el aprendizaje.

EN a computer model based on the architecture of the brain consisting of multiple simple processing units connected by adaptive weights. A collection of processing units and adptive connections that is designed to perform a specific processing function. A neural network is used for pattern recognition, particularly for classification, but also for other tasks that involve approximation such as predictive modelling.

SINÓNIMO

ES Ø
EN neural net
DE Ø

ORIGEN Informática

Informática > inteligencia artificial > aprendizaje > aprendizaje supervisado > redes neuronales

POSICIÓN DEL CONCEPTO Técnicas y algoritmos

CONTEXTO

"One of the major advantages of **neural networks** is that, theoretically, they are capable of approximating any continuous function, and thus the researcher does not need to have any hypotheses about the underlying model, or even to some extent, which variables matter". (Fuente: Satsoft *Data Mining Techniques*).

node (50%)

ES nodo DE Knoten

DEFINICIÓN

ES cada una de las partes de un árbol interconectadas por arcos. Aquella parte de un *árbol de decisión* que representa una prueba del valor de una *variable* asociada. Dependiendo de los distintos valores de la *variable*, el árbol podría conectarse a otros nodos y así sucesivamente. Los nodos finales se denominan *hojas.*

EN the part of a decision tree that represents a testing of the value of an associated variable. On the basis of the different values of the variable, the tree will potentially branch out to other nodes and so on.

SINÓNIMO Ø

ORIGEN Informática

Informática > inteligencia artificial > aprendizaje supervisado > árboles de decisión > nodo

POSICIÓN DEL CONCEPTO Técnicas y algoritmos

CONTEXTO

"In order to make a prediction the neural network accepts the values for the predictors on what are called the input **nodes**. These become the values for those nodes those values are then multiplied by values that are stored in the links (sometimes called links and in some ways similar to the weights that were applied to predictors in the nearest neighbor method). These values are then added together at the node at the far right (the output node) a special thresholding function is applied and the resulting number is the prediction. In this case if the resulting number is 0 the record is considered to be a good credit risk (no default) if the number is 1 the record is considered to be a bad credit risk (likely default)." (Fuente: *An Overview of Data Mining Techniques.* Excerpted from the book Building Data Mining Applications for CRM by Alex Berson, Stephen Smith, and Kurt Thearling)

noisy data (16,7%)

ES datos con ruido DE verrauschte Daten

DEFINICIÓN

ES datos que, o bien carecen de valores, o, de tenerlos, no son válidos. Véase *valores fuera de rango.*

ENG data with missing or invalid values. See outliers.

SINÓNIMO Ø

ORIGEN Informática

Informática > bases de datos > datos con ruido

POSICIÓN DEL CONTEXTO Datos

CONTEXTO

"Missing and **noisy data:** This problem is especially acute in business databases. U.S. census data reportedly have error rates as great as 20 percent in some fields". (Fuente: Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth *From Data Mining to Knowledge Discovery in Databases*)

nominal variable (*)

ES variable nominal DE nominale Variable

DEFINICIÓN

ES término que a veces se utiliza en vez de *variable categórica*.

EN an alternative name for a categorical variable.

SINÓNIMO

ES variable categórica
EN Ø
DE Ø

ORIGEN Minería de datos

Minería de datos > datos > variables > variable nominal

POSICIÓN DEL CONCEPTO Datos

CONTEXTO

"Some common strategies for calculating missing values include using the modal value (for **nominal variables**), the median (for ordinal variables), or the mean (for continuous variables)". (Fuente: Two Crows *Data Mining in Brief*)

*Nota: el analizador de textos no permite obtener un porcentaje diferente al de variable

normalize (*)

ES normalizar DE normalisieren

DEFINICIÓN

ES un grupo de datos numéricos se normaliza restando el valor mínimo de todos los valores y dividiendo por el rango de los datos. Esto proporciona datos cuyo histograma tiene una forma similar, pero con todos los valores entre 0 y 1. Es útil hacer esto con todas las entradas a las redes neuronales y también con las entradas en otros modelos de regresión.

EN a collection of numeric data is normalized by subtracting the minimum value from all values and dividing by the range of the data. This yields data with a similarly shaped histogram but with all values between 0 and 1. It is useful to do this for all inputs into neural nets and also for inputs into other regression models.

SINÓNIMO

ES Ø
EN standarize
DE Ø

ORIGEN Matemáticas

Matemáticas > álgebra> normalizar

POSICIÓN DEL CONCEPTO Proceso

CONTEXTO

"There are also many important steps required for preprocessing the data that goes into a neural network - most often there is a requirement to **normalize** numeric data between 0.0 and 1.0 and categorical predictors may need to be broken up into virtual predictors that are 0 or 1 for each value of the original categorical predictor". (Fuente: *An Overview of Data Mining Techniques* Excerpted from the book Building Data Mining Applications for CRM by Alex Berson, Stephen Smith, and Kurt Thearling)

*Nota: el analizador de textos no permite dar porcentajes sobre verbos

outliers (50%)

ES datos atípicos DE Ausreisser

DEFINICIÓN

ES valores de registros que no se adaptan a la norma esperada. Pueden indicar la existencia de información potencialmente valiosa (una transacción fraudulenta o que recibe una valoración inusualmente alta), o ruido.

EN variables of records which have values that do not conform to some expected norm. Outliers can be good or bad: good outliers indicate potentially valuable information (an unusually high-valued or fraudulent transaction), and bad outliers indicate noisy data.

SINÓNIMO

ES datos fuera de rango
EN Ø
DE outliers

ORIGEN Matemáticas

Matemáticas > estadística > datos atípicos

POSICIÓN DEL CONCEPTO Tipos de problema

CONTEXTO

"Domain knowledge is also critical for outlier detection needed to clean data and avoid classic problems such as a juvenile crime committed by a 80-year-old "child". If a data mining model were build using the data in Figure 1, it is possible that **outliers** (most likely caused by incorrect data entry) will skew the resulting model (especially the zero-year-old children, which are more reasonable than eighty-year-old children). The common role of visualization here is mostly in terms of annotating model structures with domain knowledge that they violate". (Fuente: Kurt Thearling et al *Visualizing Data Mining Models*)

overfitting (33,3%)

ES overfitting DE Overfitting

DEFINICIÓN

ES fenómeno por el cual los modelos predictivos tienden a ajustarse demasiado a los datos del conjunto de entrenamiento, y por lo tanto no pueden clasificar bien las nuevas instancias. También llamado sobreaprendizaje.

EN the phenomenon by which predictive models learn too well the detailed patterns in the input data during the training process and are therefore unable to make good generalizations about new input data. Also called overtraining

SINÓNIMO

ES sobreaprendizaje
EN overtraining*
DE Ø

ORIGEN Informática

Informática > inteligencia artificial > aprendizaje supervisado > overfitting

POSICIÓN DEL CONCEPTO Técnicas y algoritmos

CONTEXTO

"The techniques used in data mining, when successful, are successful for precisely the same reasons that statistical techniques are successful (e.g. clean data, a well defined target to predict and good validation to avoid **overfitting**). And for the most part the techniques are used in the same places for the same types of problems (prediction, classification discovery). In fact some of the techniques that are classical defined as "data mining" such as CART and CHAID arose from statisticians". (Fuente: An Overview of Data Mining Techniques Excerpted from the book Building Data Mining Applications for CRM by Alex Berson, Stephen Smith, and Kurt Thearling)

*Nota: el término original que designaba el concepto era "overtraining", usado por IBM. Su bajo porcentaje de aparición en el corpus (4,2%) sugiere su sustitución por el término preferido.

pattern (87,5%)

ES patrón DE Muster

DEFINICIÓN

ES definición de alto nivel de los datos. Los analistas y los estadísticos pasan mucho tiempo buscando patrones en los datos. Un patrón puede ser una relación entre dos variables. Las técnicas de minería de datos incluyen el descubrimiento automático de patrones que hace posible detectar complicadas relaciones n-lineales en los datos. Los patrones no son lo mismo que la causalidad. (Véase KDD: Piatesky-Sapiro)

EN high level definition of data. Analysts and statisticians spend much of their time looking for patterns in data. A pattern can be a relationship between two variables. Data mining techniques include automatic pattern discovery that makes it possible to detect complicated non-linear relationships in data. Patterns are not the same as causality.

SINÓNIMO

ES Ø
EN Ø
DE Pattern

ORIGEN Minería de datos

POSICIÓN DEL CONCEPTO Genérico

CONTEXTO

"An example of **pattern** discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors". (Fuente: Data Mining Page: *An Introduction to Data Mining*")

precision (12,5%)

ES precisión DE Präzision

DEFINICIÓN

ES la precisión de una estimación de un parámetro en un modelo es una medida de lo variable/mutable que sería la estimación al ser aplicada a otros conjuntos de datos. La precisión no mide la exactitud. La exactitud es una medida de lo cerca que está la estimación con respecto a el valor real del parámetro. La exactitud se mide a partir de la distancia media de la estimación en distintos conjuntos de datos respecto al valor real. Las estimaciones pueden exactas pero no precisas, o precisas pero no exactas. Un estimación precisa pero inexacta suele estar desviada, siendo el desvío igual a la distancia media respecto al valor real del parámetro.

EN the precision of an estimate of a parameter in a model is a measure of how variable the estimate would be over other similar data sets. A very precise estimate would be one that did not vary much over different data sets. Precision does not measure accuracy. Accuracy is a measure of how close the estimate is to the real value of the parameter. Accuracy is measured by the average distance over different data sets of the estimate from the real value. Estimates can be accurate but not precise, or precise but not accurate. A precise but inaccurate estimate is usually biased, with the bias equal to the average distance from the real value of the parameter.

SINÓNIMO

ES Ø
EN Ø
DE Genauiggeit

ORIGEN Matemáticas

Matemáticas > estadística > precisión

POSICIÓN DEL CONCEPTO Parámetros de evaluación

CONTEXTO

"Imagine the complexity of a decision tree derived from a database of hundreds of attributes and a response variable with a dozen output classes. Such a tree would be extremely difficult to understand, although each path to a leaf is usually understandable. In that sense a decision tree can explain its predictions, which is an important advantage. However, this clarity can be somewhat misleading. For example, the hard splits of decision trees imply a **precision** that is rarely reflected in reality. (Why would someone whose salary was $40,001 be a good credit risk whereas someone whose salary was $40,000 not be?) Furthermore, since several trees can often represent the same data with equal accuracy, what interpretation should be placed on the rules? Decision trees make few passes through the data (no more than one pass for each level of the tree) and they work well with many predictor variables. As a consequence, models can be built very quickly, making them suitable for large data sets". (Fuente: Two Crows *Data Mining in Brief*)

predictability (4,2%)

ES predecibilidad DE Vorhersagbarkeit

DEFINICIÓN

ES factor de confianza. Dada una *regla* A => B, el número de *registros* en los cuales B aparece junto con A expresado como porcentaje de todos los registros en que A aparece, ya sea con o sin B. El factor indica la fuerza de la afinidad entre los dos *elementos*. (Véase *factor de apoyo, error.* Es una medida de calidad de la regla).

EN a measure of the likelihood that an event will occur. Sometimes used to mean the same as confidence.

SINÓNIMO

ES confianza
EN confidence
DE Konfidenz

ORIGEN Matemáticas

Matemáticas > estadística > predecibilidad

POSICIÓN DEL CONCEPTO Parámetros de evaluación

CONTEXTO

"To discover meaningful rules, however, we must also look at the *relative* frequency of occurrence of the items and their combinations. Given the occurrence of item A (the antecedent), how often does item B (the consequent) occur? That is, what is the conditional **predictability** of B, given A? Using the above example, this would mean asking "When people buy a hammer, how often do they also buy nails?" Another term for this conditional predictability is *confidence.* Confidence is calculated as a ratio: (frequency of A and B)/(frequency of A)". Fuente: Two Crows Corporation data Mining in Brief)

prediction (70,8%)

ES predicción DE Prädiktion

DEFINICIÓN

ES estimación del valor de un resultado para un caso desconocido en base a un modelo y a los valores de otros resultados para ese caso.

EN an estimate of the value of some output field for an unknown case based on a model and the values of other fields for that case.

SINÓNIMO

ES Ø
EN Ø
DE Prognose, Vorhersage

ORIGEN Minería de datos

Minería de datos > tipos de problema > predicción

POSICIÓN DEL CONCEPTO Tipos de problema

CONTEXTO

"In predictive models, the values or classes we are predicting are called the *response*, *dependent* or *target variables.* The values used to make the **prediction** are called the *predictor* or *independent* variables". (Fuente: Two Crows *Data Mining in Brief*)

predictive modeling (50%)

ES modelado predictivo DE Generieren eines Vorhersagemodell

DEFINICIÓN

ES operación de minería de datos que utiliza el contenido de una base de datos, o casos conocidos, para generar un modelo que pueda ayudar a predecir una clase o valor asociado a nuevos casos no etiquetados. Tiene dos variedades: clasificación y predicción de valor.

EN a data mining operation that uses the contents of a database of known cases to generate a model that can help to predict a class or value associated with new, unseen cases. It has two specializations: classification and value prediction.

SINÓNIMO

ES Ø
EN Ø
DE predictive modeling

ORIGEN Minería de datos

Minería de datos > tipos de problemas > modelado predictivo

POSICIÓN DEL CONCEPTO Tipos de problemas

NOTA: CONTEXTO EN PÁGINA SUGUIENTE

CONTEXTO

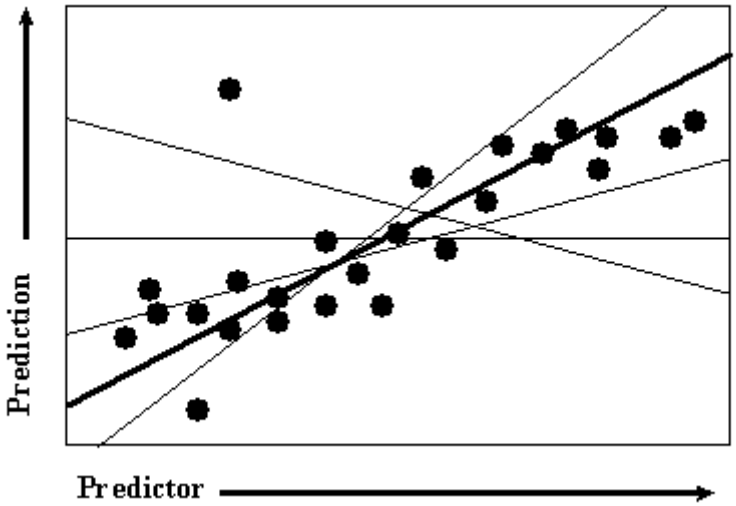


Figure 1.3 Linear regression is similar to the task of finding the line that minimizes the total distance to a set of data.

The predictive model is the line shown in Figure 1.3. The line will take a given value for a predictor and map it into a given value for a prediction. The actual equation would look something like: Prediction = a + b * Predictor. Which is just the equation for a line Y = a + bX. As an example for a bank the predicted average consumer bank balance might equal $1,000 + 0.01 * customer's annual income. The trick, as always with **predictive modeling**, is to find the model that best minimizes the error. The most common way to calculate the error is the square of the difference between the predicted value and the actual value. Calculated this way points that are very far from the line will have a great effect on moving the choice of line towards themselves in order to reduce the error. The values of a and b in the regression equation that minimize this error can be calculated directly from the data relatively quickly. (Fuente: *An Overview of Data Mining Techniques* Excerpted from the book Building Data Mining Applications for CRM by Alex Berson, Stephen Smith, and Kurt Thearling)

predictor (33,3%)

ES modelo predictor DE Predictor

DEFINICIÓN

ES el resultado de un problema de predicción es un modelo predictor. Véase *predicción*.

EN the sesult of a prediction problem is a predictor model. See prediction.

SINÓNIMO

ES Ø
EN Ø
DE Vorhersagemodell

ORIGEN Minería de datos

Minería de datos > tipos de resultado > modelo predictor

POSICIÓN DEL CONCEPTO Tipos de resultado

CONTEXTO

"In predictive models, the values or classes we are predicting are called the *response*, *dependent* or *target variables.* The values used to make the prediction are called the ***predictor*** or *independent variables*". (Fuente: Two Crows Data Mining in Brief)

prevalence (4,2%)

ES prevalencia DE Häufigkeit

DEFINICIÓN

ES la medida de la frecuencia con la que el grupo de elementos en una asociación aparece conjuntamente expresado como porcentaje de todas las transacciones. Por ejemplo, “entre el total de las compras en una ferretería, en el 2% de los casos fueron adquiridos un pico y una pala”. (Véase *support, confidence factor*).

EN the measure of how often the collection of items in an association occur together as a percentage of all the transactions. For example, ”in 2% of the purchases at the hardware store, both a pick and a shovel were bought”.

SINÓNIMO Ø

ORIGEN Minería de datos

Minería de datos > parámetros de evaluación > prevalencia

POSICIÓN DEL CONCEPTO Parámetros de evaluación

CONTEXTO

“Thus we can see that the likelihood that a hammer buyer will also purchase nails (30%) is greater than the likelihood that someone buying nails will also purchase a hammer (19%). The **prevalence** of this hammer-and-nails association (i.e., the support is 1.5%) is high enough to suggest a meaningful rule”. (Fuente: Two Crows Data Mining in Brief)

principal component analysis (*)

ES análisis de componentes principales
DE Principal Component Analysis

DEFINICIÓN

ES método de reducción de datos que se basa en resumir la varianza total de muchos campos relacionados utilizando unos pocos campos de referencia.

EN a method od data reduction that works by summarizing the total variance in a large number of related fields using a small number of derived fields.

SINÓNIMO Ø

ORIGEN Matemáticas

Matemáticas > estadística > análisis de componentes principales

POSICIÓN DEL CONCEPTO Técnicas y algoritmos

CONTEXTO

"The second aspect is typically addressed by reducing the number of features, by either selection of a subset based on a suitable criteria, or by transforming the original set of attributes into a smaller one using linear projections (e.g., principal component analysis (**PCA**)) or through non-linear (Chang and Ghosh 2001) means". (Fuente: Strehl & Ghosh Relationship-Based Clustering and Visualization for High-Dimensional Data Mining)

Nota*: no se proporcionan frecuencias de aparición en el corpus dada la particular estructura del término, con un adverbio como elemento fundamental del grupo nominal, lo cual no permite su tratamiento con el analizador de textos.

processing unit (41,7%)

ES unidad de proceso DE Einheit

DEFINICIÓN

ES unidad de una red neuronal que se utiliza para calcular un valor de salida mediante la suma de todos los valores de entrada multiplicados por sus respectivos pesos.

EN a unit in a neural network used to calculate an output value by summing all incoming values multiplied by their respective adaptive connection weights.

SINÓNIMO

ES Ø
EN processing element
DE Neuron

ORIGEN Informática

Informática > inteligencia artificial > aprendizaje supervisado > redes neuronales > unidad de proceso

POSICIÓN DEL CONCEPTO Técnicas y algoritmos

CONTEXTO

"Neural networks use a set of processing elements (or nodes) analogous to neurons in the brain. These **processing elements** are interconnected in a network that can then identify patterns in data once it is exposed to the data, i.e the network learns from experience just as people do. This distinguishes neural networks from traditional computing programs, that simply follow instructions in a fixed sequential order". (Fuente: QUB *Data Mining techniques*)

pruning (25%)

ES poda DE Pruning

DEFINICIÓN

ES eliminar las particiones de nivel inferior o sub-árboles enteros en un árbol de decisión. Este término se usa también para describir algoritmos que ajustan la topología de una red neuronal eliminando (por ejemplo, podando) nodos ocultos.

EN eliminating lower level splits or entire sub-trees in a decision tree. This term is also used to describe algorithms that adjust the topology of a neural net by removing (i.e., pruning) hidden nodes.

SINÓNIMO Ø

ORIGEN Informática

Informática > inteligencia artificial > técnicas de clasificación > técnicas de predicción > técnicas de optimización > poda

POSICIÓN DEL CONCEPTO Técnicas y algoritmos

CONTEXTO

"Beyond interactive classification, interactively guiding the model-building process provides additional control and understanding to the user. Angoss [4] provides a decision tree tool that gives the user full control over when and how the tree is built. The user may suggest splits, perform **pruning**, or manually construct sections of the tree. This facility can boost understanding greatly". (Fuente: Kurt Thearling el al *Visualizing Data Mining Models*)

radial basis function (29,2%)

ES función basada en el radio DE radiale Basisfunktion

DEFINICIÓN

ES una de las funciones utilizadas en las técnicas de modelado predictivo. Se basa en calcular la función de la distancia o el radio desde un punto en concreto. Se utiliza para construir aproximaciones a funciones más complejas.

EN used in the data mining technique that predicts values. Represents a function of the distance or the radius from a particular point. Used to build up approximations to more complex functions.

SINÓNIMO Ø

ORIGEN Minería de datos

Minería de datos > técnicas y algoritmos >modelado predictivo

POSICIÓN DEL CONCEPTO Técnicas y algoritmos

CONTEXTO

"Many algorithms are available to build your models. You might build the neural net using backpropagation or **radial basis functions**. For the decision tree, you might choose among CART, C5.0, Quest, or CHAID". (Fuente: Two Crows *Data Mining in Brief*).

range* (66,7%)

ES rango DE Intervall

DEFINICIÓN

ES el rango de los datos es la diferencia entre el valor máximo y el valor mínimo. Alternativamente, el rango puede incluir el mínimo y el máximo, como en la expresión “el valor tiene un rango de 2 a 8”.

ENG the range of the data is the difference between the maximum value and the minimum value. Alternatively, range can include the minimum and maximum, as in "The value ranges from 2 to 8."

SINÓNIMO

ES intervalo
EN interval
DE range

ORIGEN Matemáticas

Matemáticas > estadística > datos > rango

POSICIÓN DEL CONCEPTO Proceso/Datos

CONTEXTO

“Often, the method by which the data where gathered was not tightly controlled, and so the data may contain out-of-**range** values (e.g., Income: -100), impossible data combinations (e.g., Gender: Male, Pregnant: Yes), and the like”. (Fuente: *Statsoft Data Mining Techniques*)

*Nota: Pese a ser un término indudablemente propio de la estadística, el hecho de que sea imprescindible usarlo en la fase de pre-proceso de preparación hace que debamos incluirlo en el glosario. La frecuencia de aparición en el corpus lo avala, aunque al ser además de término una palabra, (gama), este dato no es totalmente fiable.

record (62,5%)

ES registro DE Eintrag

DEFINICIÓN

ES conjunto de valores de datos relativos a un ejemplo o caso concreto, como, por ejemplo, a la misma transacción o al mismo individuo. Véase variable.

EN a collection of data values all belonging to a particular instance or occurrence, as for instance, to the same transaction or to the same individual. See variable.

SINÓNIMO

ES fila, tupla, atributo, variable
EN attribute, variable
DE Record

ORIGEN Informática

Informática > ficheros > bases de datos > registro

POSICIÓN DEL CONCEPTO Datos

CONTEXTO

"the episode database contained 6.8 million records, one for each patient visit. Each **record** contained up to 20 pathology tests, which the GP ordered as a result of the visit". Fuente: Cabena, Hadjinian, Stadler, Verhees, Zanasi (1997-107)

regression tree (62,5%)

ES árbol de regresión DE Regression Tree

DEFINICIÓN

ES árbol de decisión que predice valores de variables continuas.

EN a decision tree that predicts values of continuous variables.

SINÓNIMO Ø

ORIGEN Matemáticas

Matemáticas > estadística > técnicas de regresión > árbol de regresión

POSICIÓN DEL CONCEPTO Técnicas y algoritmos

CONTEXTO

"Decision trees which are used to predict categorical variables are called *classification trees* because they place instances in categories or classes. Decision trees used to predict continuous variables are called ***regression trees***". (Fuente: Two Crows *Data Mining in Brief*).

rule (83,3%)

ES regla DE Regel

DEFINICIÓN

ES representación de un patrón de conocimiento en la forma *antecedente => consecuente.*

EN a knowledge pattern expressed as *antecedent => consequent.*

SINÓNIMO Ø

ORIGEN Informática

Informática > inteligencia artificial > ingeniería del conocimiento > representación de conocimiento > regla

POSICIÓN DEL CONCEPTO Tipos de resultado

CONTEXTO

"A data mine system has to infer a model from the database that is it may define classes such that the database contains one or more attributes that denote the class of a tuple ie the predicted attributes while the remaining attributes are the predicting attributes. Class can then be defined by condition on the attributes. When the classes are defined the system should be able to infer the **rules** that govern classification, in other words the system should find the description of each class. Production rules have been widely used to represent knowledge in expert systems and they have the advantage of being easily interpreted by human experts because of their modularity i.e. a single rule can be understood in isolation and doesn't need reference to other rules". (Fuente: QUB *Data Mining Techniques*)

rule induction (50%)

ES inducción de reglas DE Regelinduktion

DEFINICIÓN

ES el proceso de obtener automáticamente reglas de toma de decisiones a partir de casos de muestra.

EN the process of automatically deriving decision-making rules from example cases.

SINÓNIMO Ø

ORIGEN Minería de datos

Minería de datos > técnicas y algoritmos > inducción de reglas

POSICIÓN DEL CONCEPTO Técnicas y algoritmos

CONTEXTO

"**Rule induction** is a method for deriving a set of rules to classify cases. Although decision trees can produce a set of rules, rule induction methods generate a set of independent rules which do not necessarily (and are unlikely to) form a tree". (Fuente: Two Crows *Data Mining in Brief*)

sample (66,6%)

ES muestra DE Beispielmenge

DEFINICIÓN

ES muestra de datos. Subconjunto de casos seleccionados entre un conjunto mayor de casos posibles (llamado población). Las conclusiones que se extraen del análisis de esta muestra se aplican a la totalidad de la población.
Véase también *sampling*.

EN a subset of cases selected from a larger subset of possible cases (called population). The conclusions drawn from the analysis of the sample are then applied to the population. It also designs the process of selecting the subset.

SINÓNIMO Ø

ORIGEN Matemáticas

Matemáticas > estadística > muestra

POSICIÓN DEL CONCEPTO Datos/Proceso

CONTEXTO

"For example, a sample of a mailing list would be sent an offer, and the results of the mailing used to develop a classification model to be applied to the entire database. Sometimes an expert classifies a **sample** of the database, and this classification is then used to create the model which will be applied to the entire database". (Fuente: Two Crows Data Mining in Brief)

sampling (41,67%)

ES muestreo DE Sampling

DEFINICIÓN	
ES	crear un subconjunto de datos a partir del total. El muestreo aleatorio pretende representar el todo eligiendo la muestra a partir de un mecanismo aleatorio.
EN	creating a subset of data from the whole. Random sampling attempts to represent the whole by choosing the sample through a random mechanism.

SINÓNIMO	Ø

ORIGEN	Matemáticas

Matemáticas > estadística > muestreo

POSICIÓN DEL CONCEPTO	Proceso

CONTEXTO

"As in cross validation, the model is built on the entire dataset. Then numerous data sets called bootstrap samples are created by **sampling** from the original data set. After each case is sampled, it is replaced and a case is selected again until the entire bootstrap sample is created." (Fuente: Two Crows *Data Mining in Brief*)

scoring (20,8)

ES scoring DE Scoring

DEFINICIÓN

ES proceso mediante el cual se etiquetan registros en base a un modelo de clasificación previamente obtenido. Un caso típico es el credit scoring, en el que se clasifica la tasa o nivel de riesgo de una solicitud de crédito en función de varios aspectos relativos al solicitante y al crédito solicitado.

EN the process of producing a classification or prediction for a new, untested case. An example is credit scoring, where a credit application is rated for risk based on various aspects of the applicant and the loan in question.

SINÓNIMO Ø

ORIGEN Economía

Economía > marketing > scoring

POSICIÓN DEL CONCEPTO Tipos de problema

CONTEXTO

"Many data mining problems can be transformed to classification problems. For example, credit **scoring** tries to assess the credit risk of a new customer. This can be transformed to a classification problem by creating two classes, good and bad customers. A classification model can be generated from existing customer data and their credit behavior. This classification model can then be used to assign a new potential customer to one of the two classes and hence accept or reject him". (Fuente: CRISP DM 1.0 *Step-by-Step Data Mining Guide*)

segment (62,5%)

ES segmento DE Segment

DEFINICIÓN

ES conjunto de registros de una base de datos los cuales tienen todos características similares que se basan en el parecido entre los valores de sus *variables.*

EN a sub-population of records within a database all having similar characteristics based on similarity between the values of their variables.

SINÓNIMO

ES cluster
EN cluster
DE Ø

ORIGEN Informática

Informática > inteligencia artificial > aprendizaje no supervisado > segmento

POSICIÓN DEL CONCEPTO Tipos de resultados

CONTEXTO

"If we started off with our population being half churners and half non-churners then we would expect that a question that didn't organize the data to some degree into one **segment** that was more likely to churn than the other then it wouldn't be a very useful question to ask. On the other hand if we asked a question that was very good at distinguishing between churners and non-churners - say that split 100 customers into one segment of 50 churners and another segment of 50 non-churners then this would be considered to be a good question. In fact it had decreased the "disorder" of the original segment as much as was possible". (Fuente: An Overview of Data Mining Techniques Excerpted from the book Building Data Mining Applications for CRM by Alex Berson, Stephen Smith, and Kurt Thearling)

segmentation (54,2%)

ES segmentacion DE Segmentierung

DEFINICIÓN

ES operación de minería de datos cuya meta es obtener conjuntos de elementos parecidos entre sí (clustering).

ENG an alternative name for database segmentation.

SINÓNIMO

ES clustering
EN clustering
DE Ø

ORIGEN Informática

Informática > inteligencia artificial > aprendizaje no supervisado > segmentación

POSICIÓN DEL CONCEPTO Tipos de problemas

CONTEXTO

"**Segmentation** of customers, products, and sales regions is something that marketing managers have been doing for many years. In the past this segmentation has been performed in order to get a high level view of a large amount of data - with no particular reason for creating the segmentation except that the records within each segmentation were somewhat similar to each other. In this case the segmentation is done for a particular reason - namely for the prediction of some important piece of information. The records that fall within each segment fall there because they have similarity with respect to the information being predicted - not just that they are similar - without similarity being well defined". (Fuente: *An Overview of Data Mining Techniques* Excerpted from the book Building Data Mining Applications for CRM by Alex Berson, Stephen Smith, and Kurt Thearling).

sensitivity analysis (25%)

ES análisis de sensibilidad DE Sensitivitänalyse

DEFINICIÓN

ES variar los parámetros de un modelo para evaluar los cambios en sus resultados.

EN varying the parameters of a model to assess the change in its output.

SINÓNIMO Ø

ORIGEN Informática

Informática > inteligencia artificial > modelos predictivos > análisis de sensibilidad

POSICIÓN DEL CONCEPTO Técnicas y algoritmos

CONTEXTO

"In this sense, assessing trust is also closely related to model comparison. In particular, it is very useful to understand the **sensitivity** of model predictions and quality to changes in parameters and/or structure of the given model. There are many ways to visualize such sensitivity, often in terms of local and global (conditional) probability densities — with special interest in determining whether multiple modes of high probability exist for some parameters and combinations". (Fuente: Kurt Thearling et al *Visualizing Data Mining Models*)

sequence discovery (50%)

ES descubrimiento de secuencias DE Sequence Discovery

DEFINICIÓN

ES lo mismo que asociación, pero incluyendo además la secuencia temporal de eventos. Por ejemplo, “el veinte por ciento de las personas que compra un reproductor de vídeo termina comprando una cámara en los cuatro meses siguientes”.

EN the same as association, except that the time sequence of events is also considered. For example, "Twenty percent of the people who buy a VCR buy a camcorder within four months."

SINÓNIMO

Ø

ORIGEN

Minería de datos

Minería de datos > tipos de problemas > descubrimiento de secuencias

POSICIÓN DEL CONCEPTO

Tipos de problemas

CONTEXTO

“The two most common approaches to link analysis are *association discovery* and ***sequence discovery***. Association discovery finds rules about items that appear together in an event such as a purchase transaction. Market-basket analysis is a well-known example of association discovery. Sequence discovery is very similar, in that a sequence is an association related over time”. (Fuente: Two Crows *Data Mining in Brief*)

sequential patterns (33,3%)

ES patrones secuenciales DE sequentielles Muster

DEFINICIÓN

ES asociaciones entre transacciones con semántica intrínseca de secuencialidad.

EN associations between transactions such that the presence of one set of items is followed by another set of items in a database of transactions over a period of time.

SINÓNIMO Ø

ORIGEN Minería de datos

Minería de datos > tipos de resultados > patrones secuenciales

POSICIÓN DEL CONCEPTO Tipos de resultados

CONTEXTO

"In the transaction log discussed above, the identity of the customer that did the purchase is not generally known. If this information exists, an analysis can be made of the collection of related records of the same structure as above (i.e., consisting of a number of items drawn from a given collection of items). The records are related by the identity of the customer that did the repeated purchases. Such a situation is typical of a Direct Mail application. In this case, a catalog merchant has the information, for each customer, of the sets of products that the customer buys in every purchase order. A **sequential pattern** function will analyze such collections of related records and will detect frequently occurring patterns of products bought over time. A sequential pattern operator could also have been used in one of the examples in the previous section to discover the set of purchases that frequently precede the purchase of a microwave oven". (Fuente: IBM *Data Mining: Extending the Information Warehouse Framework*)

significance (41,7%)

ES significancia DE Signifikanz

DEFINICIÓN

ES medida de probabilidad sobre la consistencia con que los datos apoyan un determinado resultado (normalmente de un test estadístico). Si la significancia de un resultado se dice que es del ,05, eso quiere decir que sólo hay una probabilidad del 0,05% de que el resultado pudiera ocurrido sólo por casualidad. Una significancia muy baja (menos del ,05) se suele tomar como la prueba de que el modelo de minería de datos debería ser aceptado, dado que los eventos con muy baja probabilidad apenas se producen. Así, si la estimación de un parámetro en un modelo mostrara una significancia del ,01 eso sería la evidencia de que el parámetro debe estar en el modelo.

EN a probability measure of how strongly the data support a certain result (usually of a statistical test). If the significance of a result is said to be .05, it means that there is only a .05 probability that the result could have happened by chance alone. Very low significance (less than .05) is usually taken as evidence that the data mining model should be accepted since events with very low probability seldom occur. So if the estimate of a parameter in a model showed a significance of .01 that would be evidence that the parameter must be in the model.

SINÓNIMO Ø

ORIGEN Matemáticas

Matemáticas > estadística > significancia

POSICIÓN DEL CONCEPTO Parámetros de evaluación

CONTEXTO

"Because CHAID relies on the contingency tables to form its test of **significance** for each predictor all predictors must either be categorical or be coerced into a categorical form via binning (e.g. break up possible people ages into 10 bins from 0-9, 10-19, 20-29 etc.)". (Fuente: *An Overview of Data Mining Techniques* Excerpted from the book Building Data Mining Applications for CRM by Alex Berson, Stephen Smith, and Kurt Thearling)

supervised learning (29,2%)

ES aprendizaje supervisado DE überwachtes Lernen

DEFINICIÓN

ES dado un conjunto de entradas con registros previamente clasificados, el aprendizaje supervisado define aquel problema consistente en aprender un modelo que sirva para clasificar en el futuro un registro no clasificado. La propagación regresiva, por ejemplo, usa aprendizaje supervisado y hace ajustes durante la el proceso de aprendizaje, de forma que el valor computado por la red neuronal se aproximará al valor real proporcionado a medida que la red aprende de los datos que se le proporcionan. Se usa en las técnicas de minería de datos que se emplean para clasificación y predicción de valores. Véase *aprendizaje no supervisado.*

EN a learning algorithm that requires input and resulting output pairs to be presented to the network during the training process. Back propagation, for instance, uses supervised learning and makes adjustments during training so that the value computed by the neural network will approach the actual supplied value as the network learns from the data presented. Used in the data mining techniques provided for classification and value prediction. See unsupervised learning.

SINÓNIMO

ES Ø
EN Ø
DE supervised learning

ORIGEN

Informática

Informática > inteligencia artificial > ingeniería del conocimiento > aprendizaje supervisado

POSICIÓN DEL CONCEPTO

Tipos de problemas

CONTEXTO

"Predictive models are built, or *trained,* using data for which the value of the response variable is already known. This kind of training is sometimes referred to as ***supervised learning***, because calculated or estimated values are compared with the known results. (By contrast, descriptive techniques such as clustering, described in the previous section, are sometimes referred to as *unsupervised learning* because there is no already-known result to guide the algorithms".)(Fuente: *An Overview of Data Mining Techniques.* Excerpted from the book Building Data Mining Applications for CRM by Alex Berson, Stephen Smith, and Kurt Thearling)

support factor (58,3%)

ES factor de soporte DE Support

DEFINICIÓN

ES dada una regla A => B, el número de registros en los cuales esta regla se produce, expresado como porcentaje de todos los registros en la base de datos. Este factor indica la frecuencia relativa con la que se produce la regla en los datos. Véase *factor de confianza*.

EN given an association rule A => B, the number of records in which this rule occurs as a percentage of all records in the database. The factor indicates the relative frequency with which the rule occurs in the data. See confidence factor.

SINÓNIMO

Ø

ORIGEN

Minería de datos

Minería de datos > parámetros de evaluación > factor de soporte

POSICIÓN DEL CONCEPTO

Parámetros de evaluación

CONTEXTO

"The coverage of the rule has to do with how much of the database the rule "covers" or applies to. Examples of these two measure for a variety of rules is shown in Table 2.2. In some cases accuracy is called the confidence of the rule and coverage is called the **support.** Accuracy and coverage appear to be the preferred ways of naming these two measurements". (Fuente: *An Overview of Data Mining Techniques* Excerpted from the book Building Data Mining Applications for CRM by Alex Berson, Stephen Smith, and Kurt Thearling)

target (70,8%)

ES objetivo DE Zielattribut

DEFINICIÓN

ES el campo que se quiere predecir, cuyo valor asumimos que se relaciona con los valores de otros campos (los predictores). También llamado campo de salida o variable dependiente.

EN the field to predict, whose value is assumed to be related to the values of other fields (the predictors). Also known as output field or dependent variable.

SINÓNIMO

ES variable dependiente, variable de salida, variable decisión, variable a predecir.
EN dependent variable
DE Output

ORIGEN Matemáticas

Matemáticas > estadística > modelos predictivos > objetivo

POSICIÓN DEL CONCEPTO Datos

CONTEXTO

"Linear regresion: statistical technique used to find the best-fitting linear relationship between a **target** (dependent) variable and its predictors (independent variables)". (Fuente: Data Mining Pages *An Introduction to Data Mining*)

taxonomy (20,8%)

ES taxonomía DE Taxonomie

DEFINICIÓN

ES clasificación que asigna jerarquías a elementos relacionados. La relación de taxonomía define categorías de elementos para cada nivel de la jerarquía. Por ejemplo, una jerarquía de productos, donde manzana sería la categoría la categoría de fruta, la cual a su vez es parte de la categoría de producto, etc.

EN a classification assigning hierarchies to related items. The taxonomy relation defines item categories for each level of the hierarchy. An example is a product hierarchy, where apple would be the fruit category, which in turn is part of the product category, and so on.

SINÓNIMO

ES clasificación
EN classification
DE Begriffshierarchie

ORIGEN Informática

Informática > ingeniería del conocimiento > representación del conocimiento > taxonomía

POSICIÓN DEL CONCEPTO Proceso (pre-proceso)

CONTEXTO

"You use the experimental Market Basket node to perform association rule mining over transaction data in conjunction with item taxonomy. This node is useful in retail marketing scenarios that involve tens of thousands of distinct items, where the items are grouped into subcategories, categories, departments, and so on, called *item* ***taxonomy***. The Market Basket node uses the taxonomy data and generates rules at multiple levels in the taxonomy". (Fuente: SAS *What is New in SAS Enterprise Miner 5.3*)

test data (75%)*

ES datos de prueba DE Testdaten

DEFINICIÓN

ES un conjunto de datos independiente del conjunto de datos de entrenamiento, que se usa para hacer un ajuste fino de las estimaciones de los parámetros del modelo (es decir, los pesos).

ENG a data set independent of the training data set, used to fine-tune the estimates of the model parameters (i.e., weights).

SINÓNIMO Ø

ORIGEN Informática

Informática > inteligencia artificial > clasificación > datos de prueba

POSICIÓN DEL CONCEPTO Datos

CONTEXTO

"One of the great advantages of CART is that the algorithm has the validation of the model and the discovery of the optimally general model built deeply into the algorithm. CART accomplishes this by building a very complex tree and then pruning it back to the optimally general tree based on the results of cross validation or test set validation. The tree is pruned back based on the performance of the various pruned version of the tree on the **test** set **data**. The most complex tree rarely fares the best on the held aside data as it has been overfitted to the training data". (Fuente: *An Overview of Data Mining Techniques* Excerpted from the book Building Data Mining Applications for CRM by Alex Berson, Stephen Smith, and Kurt Thearling)

NOTA: El dato de frecuencia no es fiable por las características del analizador de textos.

test error (58,3%)

ES error del modelo (tras la prueba) DE Testfehler

DEFINICIÓN

ES la estimación de error basada en la diferencia entre las predicciones de un modelo en un conjunto de datos de prueba y los valores observados en el conjunto de datos de prueba cuando éste no fue usado para entrenar el modelo.

EN the estimate of error based on the difference between the predictions of a model on a test data set and the observed values in the test data set when the test data set was not used to train the model.

SINÓNIMO Ø

ORIGEN Informática

Informática > inteligencia artificial > modelos predictivos > error del modelo (tras la prueba)

POSICIÓN DEL CONCEPTO Parámetros de evaluación

CONTEXTO

"PRWýs chart shows how the overall error and the test error are falling as training progresses. While the test error is fluctuating more rapidly than the average error, in general the spread between the red and blue lines is not increasing, suggesting that overfitting has not yet occurred. (Actually, there was a significant deviation earlier in training, but the error rates have converged since then.) (Fuente: Stelle Brand and Rob Gerritsen Neural Networks)

time series (50%)

ES series temporales DE Zeitreihe

DEFINICIÓN

ES una serie de mediciones tomadas en puntos consecutivos en el tiempo. Los productos de minería de datos que manejan series temporales incorporan operadores de relación temporal tales como ventanas deslizantes. (Véase también ventana temporal).

EN a series of measurements taken at consecutive points in time. Data mining products which handle time series incorporate time-related operators such as moving average. (Also see *windowing.*)

SINÓNIMO Ø

ORIGEN Minería de datos

Minería de datos > datos > tipos de datos > series temporales

POSICIÓN DEL CONCEPTO Datos

CONTEXTO

"Consider what it might be like in a **time series** problem - say for predicting the stock market. In this case the input data is just a long series of stock prices over time without any particular record that could be considered to be an object. The value to be predicted is just the next value of the stock price". (Fuente: *An Overview of Data Mining Techniques* Excerpted from the book Building Data Mining Applications for CRM by Alex Berson, Stephen Smith, and Kurt Thearling)

time series model (50%)

ES modelo de series temporales DE Zeitreihenmodell

DEFINICIÓN

ES modelo que pronostica valores futuros de una serie temporal basándose en los valores pasados. La forma del modelo y su entrenamiento suelen tomar en consideración la correlación entre valores como una función de su separación en el tiempo.

EN a model that forecasts future values of a time series based on past values. The model form and training of the model usually take into consideration the correlation between values as a function of their separation in time.

SINÓNIMO Ø

ORIGEN Minería de datos

Minería de datos > tipos de resultados > modelo de series temporales

POSICIÓN DEL CONCEPTO Tipos de resultados

CONTEXTO

"Recently,more general **models** have been developed for **time-series** applications, such as nonlinear basis functions, example-based models, and kernel methods. Furthermore, there has been significant interest in descriptive graphic and local data modeling of time series rather than purely predictive modeling (Weigend and Gershenfeld 1993). Thus, although different algorithms and applications might appear different on the surface, it is not uncommon to find that they share many common components". (Fuente: Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth *From Data Mining to Knowledge Discovery in Databases*)

training (test&training) (75%)

ES entrenamiento DE Training

DEFINICIÓN

ES otra forma de denominar al proceso de estimación de los parámetros de un modelo en base al conjunto de datos disponible.

EN another term for estimating a model's parameters based on the data set at hand.

SINÓNIMO Ø

ORIGEN Matemáticas/informática

Matemáticas > estadística > entrenamiento

Informática > inteligencia artificial > entrenamiento

POSICIÓN DEL CONCEPTO Proceso/Técnicas y algoritmos

CONTEXTO

"The new network is then subjected to the process of **"training"**. In that phase, neurons apply an iterative process to the number of inputs (variables) to adjust the weights of the network in order to optimally predict (in traditional terms one could say, find a "fit" to) the sample data on which the "training" is performed. After the phase of learning from an existing data set, the new network is ready and it can then be used to generate predictions". Fuente Statsoft *Data Mining Techniques*)

training data (66,7%)

ES datos de entrenamiento DE Trainingsdaten

DEFINICIÓN

ES conjunto de datos usado para estimar o entrenar un modelo.

EN a data set used to estimate or train a model.

SINÓNIMO Ø

ORIGEN Informática

Informática > inteligencia artificial > clasificación > datos de entrenamiento

POSICIÓN DEL CONCEPTO Datos

CONTEXTO

"The problem is that we all have an intuition that the name of the customer is not going to be a very good indicator of whether that customer churns or not. It might work well for this particular 2 record segment but it is unlikely that it will work for other customer databases or even the same customer database at a different time. This particular example has to do with overfitting the model - in this case fitting the model too closely to the idiosyncrasies of the **training data**. This can be fixed later on but clearly stopping the building of the tree short of either one record segments or very small segments in general is a good idea". (Fuente: *An Overview of Data Mining Techniques* Excerpted from the book Building Data Mining Applications for CRM by Alex Berson, Stephen Smith, and Kurt Thearling)

transaction (83,3%)

ES transacción DE Transaktion

DEFINICIÓN

ES conjunto de elementos o eventos unidos por un valor clave común, por ejemplo, el caso en el que la clave de tienda, el número de terminal y el número de secuencia de terminal definen el valor clave para una transacción de cliente en un establecimiento de ventas al por menor.

EN a set of items or events that are linked by a common key value, for example, where a store ID, EPOS terminal number, and transaction sequence number defines the key value for a customer transaction at a retail store.

SINÓNIMO Ø

ORIGEN Minería de datos

Minería de datos > datos > transacción

POSICIÓN DEL CONCEPTO Datos

CONTEXTO

"The AMRP is a frequent-shopper program; that is, the consumer can collect AIR MILES Travel Miles (AMTM) for making purchases at the coalition sponsors. Customers can then redeem the Travel Miles collected for rewards, which include not only air travel, but hotel accommodation, rental cars, theatre tickets, tickets for professional sporting events, a family night at the movies, and merchandise. The various coalition partners capture consumer **transactions** and transmit them to The Loyalty Group, which stores these transactions and uses the data for database marketing initiatives on behalf of the coalition partners. The Loyalty Group data warehouse currently contains more than 6.3 million household records and 1 billion transaction records". (Fuente: Gary Saarenvirta *Mining Customer Data*)

transformation (50%)

ES transformación DE Transformation

DEFINICIÓN

ES cualquier operación aplicada a los datos para que puedan ser tratados por los algoritmos de minería de datos.

EN a re-expression of the data such as aggregating it, normalizing it, changing its unit of measure, or taking the logarithm of each data item.

SINÓNIMO Ø

ORIGEN Minería de datos

Minería de datos > proceso > transformación

POSICIÓN DEL CONCEPTO Proceso

CONTEXTO

"After you've cleaned your data, treated all missing and invalid values, and made the known valid values consistent, you're ready to transform the data. The data in its original form is valuable, but transformations will maximize the information content that you can retrieve. There are two types of data **transformation**: *Data distribution transformation.* This type of transformation involves mathematically altering the distribution of the variable. *Data creation.* This type of transformation involves the creation of new variables by combining existing variables to form ratios, differences, and so forth". (Fuente: Gary Saarenvirta *Mining Customer Data*)

unsupervised learning (29,2%)

ES aprendizaje no supervisado DE unüberwachtes Lernen

DEFINICIÓN

ES algoritmo de aprendizaje que precisa únicamente de datos de entrada durante el proceso de aprendizaje. No se proporciona ningún resultado-objetivo; en su lugar, el resultado deseado se descubre durante la ejecución de la minería. Los mapas de Kohonen utilizan aprendizaje no supervisado. Véase aprendizaje supervisado.

EN a learning algorithm that requires only input data to be present in the data source during the training process. No target output is provided; instead, the desired output is discovered during the mining run. Kohonen feature maps use unsupervised learning. See supervised learning.

SINÓNIMO

ES Ø
EN Ø
DE unsupervised learning

ORIGEN Informática

Informática > inteligencia artificial > ingeniería del conocimiento > aprendizaje no supervisado

POSICIÓN DEL CONCEPTO Tipos de problemas/operaciones

CONTEXTO

"Predictive models are built, or *trained,* using data for which the value of the response variable is already known. This kind of training is sometimes referred to as *supervised learning,* because calculated or estimated values are compared with the known results. (By contrast, descriptive techniques such as clustering, described in the previous section, are sometimes referred to as ***unsupervised learning*** because there is no already-known result to guide the algorithms".) (Fuente: *Two Crows Data Mining in Brief*)

validation (54,2%)

ES validación DE Validierung

DEFINICIÓN

ES proceso de probar los modelos con un conjunto de datos distinto al conjunto de datos de entrenamiento.

EN the process of testing the models with a data set different from the training data set.

SINÓNIMO Ø

ORIGEN Matemáticas

Matemáticas > estadística > validación

POSICIÓN DEL CONCEPTO Proceso

CONTEXTO

"Sometimes a third data set, called the **validation** data set, is needed because the test data may be influencing features of the model, and the validation set acts as an independent measure of the model's accuracy. Training and testing the data mining model requires the data to be split into at least two groups: one for model training (i.e., estimation of the model parameters) and one for model testing. If you don't use different training and test data, the accuracy of the model will be overestimated". (Fuente: Two Crows *Data Mining in Brief*)

value prediction (79,2%)*

ES predicción de valores DE Wertvorhersage (Vorhersage)

DEFINICIÓN

ES especialización del modelado predictivo para asignar un valor o propensión a un registro en una base de datos. El valor asignado o propensión se basa en atributos dentro del registro. Por ejemplo, la predicción de valor podría usarse para asignar una propensión a responder a una campaña de buzoneo a registros en una base de datos prospectiva. Véase clasificación.

EN a specialization of predictive modelling for assigning a value or propensity to a record in a database. The assigned value or propensity is based on attributes within the record. For example, value prediction could be used to assign a propensity to respond to a mailing campaign to records in a prospects database. See classification.

SINÓNIMO

Ø

ORIGEN

Minería de datos

Minería de datos > tipos de problemas > predicción de valores

POSICIÓN DEL CONCEPTO

Tipos de problemas

CONTEXTO

"In statistics prediction is usually synonymous with regression of some form. There are a variety of different types of regression in statistics but the basic idea is that a model is created that maps **values** from **predictors** in such a way that the lowest error occurs in making a prediction. The simplest form of regression is simple linear regression that just contains one predictor and a prediction. The relationship between the two can be mapped on a two dimensional space and the records plotted for the prediction values along the Y axis and the predictor values along the X axis. The simple linear regression model then could be viewed as the line that minimized the error rate between the actual prediction value and the point on the line (the prediction from the model)". (Fuente: *An Overview of Data Mining Techniques* Excerpted from the book Building Data Mining Applications for CRM by Alex Berson, Stephen Smith, and Kurt Thearling)

variable (66,7%)

ES variable DE Variable

DEFINICIÓN

ES un dato que, dentro de un registro, representa alguna característica de la muestra descrita por el registro. Por ejemplo, ingresos y límite de crédito serían datos en un registro de cliente. Algunos nombres alternativos son atributo, columna, dimensión, característica y campo.

EN a data item within a record which represents some characteristics of the instance described by the record. For example, credit-limit and income could be items in a customer record. Some alternative names are attribute, column, dimension, feature an field.

SINÓNIMO

ES dato .
EN data item
DE Attribut

ORIGEN Matemáticas

Matemáticas > estadística > variable

POSICIÓN DEL CONCEPTO Datos

CONTEXTO

"Analysts must often limit the number of **variables** they examine when doing hands-on analysis due to time constraints. Yet variables that are discarded because they seem unimportant may carry information about unknown patterns. High performance data mining allows users to explore the full depth of a database, without preselecting a subset of variables". (Fuente: Data Mining Page *An Introduction to Data Mining*)

visualization (66,7%)

ES visualización DE Visualisierung

DEFINICIÓN

ES las herramientas de visualización muestran gráficamente los datos para facilitar una mejor comprensión de su significado. Las capacidades gráficas varían desde simples gráficas de puntos (scatter plots) a complejas representaciones multidimensionales.

EN visualization tools graphically display data to facilitate better understanding of its meaning. Graphical capabilities range from simple scatter plots to complex multi-dimensional representations.

SINÓNIMO Ø

ORIGEN Matemáticas

Matemáticas > estadíatica > visualización

POSICIÓN DEL CONCEPTO Proceso

CONTEXTO

“Data mining, on the other hand, extracts information from a database that the user did not already know about. Useful relationships between variables that are non-intuitive are the jewels that data mining hopes to locate. Since the user does not know beforehand what the data mining process has discovered, it is a much bigger leap to take the output of the system and translate it into an actionable solution to a business problem. Since there are usually many ways to graphically represent a model, the **visualizations** that are used should be chosen to maximize the value to the viewer. This requires that we understand the viewer's needs and design the visualization with that end-user in mind. If we assume that the viewer is an expert in the subject area but not data modeling, we must translate the model into a more natural representation for them. For this purpose we suggest the use of orienteering principles as a template for our visualizations”.(Fuente:*Visualizing Data Mining Models* by Kurt Thearling et al)

windowing (4,2%)

Es ventana temporal DE Windowing

DEFINICIÓN

ES se usa al entrenar un modelo con datos de series temporales. Una ventana es el período de tiempo usado para cada caso de entrenamiento. Por ejemplo, si tenemos datos semanales sobre precios de existencias que cubren cincuenta semanas, y fijamos la ventana en cinco semanas, entonces el primer caso de entrenamiento usará las semanas de la uno a la cinco, y comparará su predicción con la semana seis. El segundo caso usa las semanas de la dos a la seis para predecir la semana siete, y así sucesivamente.

EN used when training a model with time series data. A window is the period of time used for each training case. For example, if we have weekly stock price data that covers fifty weeks, and we set the window to five weeks, then the first training case uses weeks one through five and compares its prediction to week six. The second case uses weeks two through six to predict week seven, and so on.

SINÓNIMO Ø

ORIGEN Minería de datos

Minería de datos > proceso > preparación > ventana temporal

POSICIÓN DEL CONCEPTO Proceso

CONTEXTO

"Quinlan's **windowing** technique starts with a small random sample (called a *window*), and generates a classifier for the window. It then test the classifier on the remaining examples, and checks the quality (accuracy) of the classifier. If the quality is not sufficient, a set of mis-classified examples is to be added to the window and a new classifier generated. (Fuente: Xindong Wu *Data Mining from large Databases***)**

7- CONCLUSIONES

Las conclusiones que hemos podido extraer una vez terminado el trabajo son múltiples y deben necesariamente clasificarse en función del apartado correspondiente a que hacen referencia. Siguiendo el orden establecido en el índice, haremos una reflexión sobre cómo los planteamientos iniciales que se plasman en la introducción han visto cumplidos sus objetivos. A continuación, y dado que los apartados 2 (La terminología como ciencia con entidad propia) y 3 (Planificación y realización de un proyecto terminológico) ya cuentan con su apartado correspondiente de conclusiones, trataremos de analizar las causas últimas de la divergencia de enfoques que se ha podido constatar a lo largo del estudio. Finalmente revisaremos los descubrimientos más notables fruto del trabajo con los técnicos: los que surgen del análisis del propio proceso como método de trabajo y los resultados materiales del proyecto en sí, la obtención de un glosario terminológico sobre minería de datos.

7.1 SOBRE LOS PLANTEAMIENTOS ORIGINALES

En la introducción se señalaba la existencia de un problema: no existía un glosario que comprendiera la terminología propia de la minería de datos en español ni en alemán. Todo lo publicado hasta la fecha eran documentos en inglés de autores norteamericanos. Las búsquedas en la Biblioteca Nacional de Alemania en Frankfurt (si bien pasaron de un documento a diecinueve entre los

años 2006 y 2008) y en las fuentes de documentación españolas (bibliografía de los profesores universitarios que imparten la materia y técnicos consultados) lo confirmaron. Los académicos y profesionales de ambos países se nutren de bibliografía original en inglés para sus estudios. El trabajo realizado demuestra que no sólo esto es así, sino que al cruzar los glosarios existentes resultó que únicamente había ocho términos que se repitieran en todos ellos. El glosario resultante podría ser también, hasta donde podemos inferir, el primero que recopile exhaustivamente lo publicado en lengua inglesa. El objetivo de la originalidad y utilidad del glosario queda así logrado.

El planteamiento central de la introducción trataba, por otra parte, sobre la influencia del medio de trabajo en el autor de la obra. El lingüista en el mundo de la técnica. ¿Cómo se ha reflejado el hecho de ser el autor profesor en la Escuela Universitaria de Informática de la Universidad politécnica en el trabajo resultante? Una tarea fundamental ha sido la de depurar las definiciones originales en inglés de cada término y sugerir una traducción aparentemente adecuada. Los diecisiete años de experiencia en trabajo con textos informáticos parecían ser de gran ayuda; nos aportaban la familiaridad con el género y el que algunos términos y conceptos de procedencia informática no nos fueran totalmente desconocidos. Esto nos permitió en ocasiones arriesgar –en el caso del español- una propuesta de término y de su definición para poder presentarlas al juicio de los técnicos. Pero no podemos extraer a partir de aquí conclusiones erróneas: en la mayoría de los casos no fue más que eso, un experimento fallido, una propuesta arriesgada que a la postre precisó de profundas reformas una vez analizada por los expertos.

El terminólogo puede sugerir, documentar, establecer métodos y pautas de trabajo, hacer análisis estadísticos y recopilar la información obtenida para plasmarla finalmente en una ficha terminológica. Pero su labor termina ahí. El contenido central de esa ficha, es decir, el término, el concepto que encierra (su definición) y su ubicación en el área de conocimiento correspondiente son

tareas que competen exclusivamente a los técnicos. No es casual que algunos de los más grandes terminólogos comenzaran su formación académica en el mundo de la técnica (Wüster en la electrónica, Picht en maquinaria agrícola, Kurt Loening era químico -al igual que Irazazábal- Schmitz informático y matemático).

El técnico es el alma del trabajo terminológico y la transmisión de conocimiento especializado entre técnicos la razón de ser de la terminología. En el caso de este autor fue su ubicación profesional en un entorno de técnicos en informática lo que le permitió entrar en contacto con el campo de la minería de datos y a partir de ahí constatar la necesidad de crear un glosario terminológico sobre este área. No podemos dejar de mencionar aquí las facilidades dadas por la institución (Rector, Vicerrector de Investigación, Director del Departamento de Lingüística Aplicada y equipo de dirección de la Escuela Universitaria de Informática) para ayudar a la conclusión de esta tesis. El contexto constituyó una feliz circunstancia para el desarrollo de la obra.

7.2 SOBRE LA DIVERGENCIA DE ENFOQUES

Una de las conclusiones más evidentes a las que nos llevó el análisis de la situación actual de la terminología fue la constatación de la existencia de dos corrientes teóricas –la francófona y la centroeuropea- que, coincidiendo en el objeto de estudio, difieren ostensiblemente en el enfoque y las soluciones prácticas. No vamos a repetir aquí los planteamientos de cada una de ellas, que ya han sido analizados con anterioridad; nos limitaremos a expresar que esta circunstancia, lejos de ser una mera apreciación personal, constituye una realidad evidente para todo aquel que entre en contacto con el mundo de la terminología contemporánea. Añadiremos únicamente una última reflexión al respecto, fruto de la entrevista mantenida el 22 de abril de 2008 con el Dr. Schmitz de la Universidad de Colonia.

La reunión con el Profesor Schmitz supuso una ocasión única para intercambiar impresiones sobre el estado actual de la terminología. Al tratar sobre la disparidad de los enfoques de ambas corrientes teóricas, el Dr. Schmitz planteó un aspecto clave de esta divergencia con una sencillez y claridad de visión que sólo se consiguen por medio del conocimiento profundo de la materia: en la relación

Objeto – Concepto (unidad de pensamiento) – Término

que constituye la base de la Terminología como ciencia (23), el centro de nuestra atención debe estar necesariamente en el concepto: el estar tratando de un mismo concepto o unidad de pensamiento es lo que permite el entendimiento entre los científicos, ya sean técnicos o académicos. Si a partir de un objeto único (24), (realidad constatable extralingüística), llegamos a idear un concepto único (expresión mental también extraligüística) y somos capaces de crear finalmente un término que lo exprese en su totalidad, el término resultante no debe constituir nunca un obstáculo para la comprensión del concepto. Y no sólo eso, sino que el término puede perfectamente expresarse en diferentes lenguas siempre y cuando designe el mismo concepto.

El planteamiento centroeuropeo de la terminología difiere fundamentalmente en esto del francófono; éste hace del término el centro de su estudio, en un intento de crear nueva terminología que tenga cabida en la lengua de destino a fin de preservar su identidad, y es hostil a la incorporación de extranjerismos. En el enfoque centroeuropeo prima el concepto. Si compartimos la misma unidad de pensamiento, es irrelevante que el término que la designe esté en una lengua diferente a la nuestra.

(23) Picht TSS 2006 Viena.

(24) Los objetos pueden ser una realidad física, tangible, o inmateriales, como una teoría o un procedimiento, o incluso imaginarios, como un unicornio. (ISO 1087-1).

En esta tesis se ha seguido el modelo centroeuropeo en tanto que nunca ha sido la intención de su autor el hallar términos en español o alemán que equivalieran a los originales en inglés: a los técnicos se les presentó una definición, o representación del concepto, y el término inglés que habitualmente lo designa. A partir de ahí nuestra labor fue de mera recopilación del término que la comunidad científica usa habitualmente en cada una de las lenguas de destino para designar ese concepto. Tal y como establece la definición de terminografía (ISO 1087): recoger, procesar y presentar datos terminológicos adquiridos por medio de investigación terminológica.

Como resultado, y según veremos en el apartado siguiente, existen numerosos ejemplos de préstamos y traducciones préstamo en ambas lenguas. No estimamos que nuestra labor sea la de opinar sobre la salud de las lenguas de destino a partir de los resultados obtenidos; no es este un debate real. Recordemos nuevamente que estamos ante un ámbito de uso restringido: es la terminología que emplean los técnicos y como tal se recoge aquí.

7.3 SOBRE EL MÉTODO DE TRABAJO Y SUS RESULTADOS

En este proyecto terminológico hemos seguido, según mencionábamos en el apartado anterior, el modelo que establece la Norma ISO 15188, optando por un equipo de trabajo formado por un terminólogo y dos técnicos y sometiendo los procedimientos y resultados a la evaluación de un terminólogo de reconocida experiencia. Sobre el autor ya hemos hablado anteriormente, pero hay datos referentes a los técnicos que pueden resultar relevantes a la hora de interpretar los resultados: los doctores Menasalvas y Lattner se enfrentaban por primera vez a un proyecto semejante; no había por su parte ideas preconcebidas en cuanto a proyectos terminográficos. Su visión pragmática sobre lo que una ficha terminológica debía aportar a quien

consultara el glosario contribuyó a la mejora del mismo. Como consecuencia de sus comentarios y sugerencias se añadió a la ficha la entrada “subject field 2” (denominada en principio *posición del término en minería de datos*), en la que se ubica el término en su ámbito correspondiente dentro de la minería de datos. Del mismo modo se incluyó en los casos más relevantes la nota “Términos relacionados” como referencia necesaria para una mejor comprensión del concepto.

En ambos casos tienen un elevado dominio de la lengua inglesa, tanto oral como escrito; el uso del inglés es constante en su vida académica, como medio de acceso a las fuentes de documentación originales y por ser el la lengua de comunicación habitual con sus colegas de especialidad. El idioma de origen no supuso una barrera en absoluto; bien al contrario, es el idioma vehicular de su especialidad. Aun así, el recurso al préstamo se da sólo en un 11% de los casos en español y en un 17% en alemán (el mayor porcentaje es comprensible en este caso al tratarse de dos lenguas germánicas).

Si tomamos como referencia el listado de préstamos en español, vemos que coincide con el alemán en todos los casos menos tres (*modeling item*, y *deployment*). Esto no es casual. El recurso al préstamo se produce únicamente en aquellos casos en que una solución alternativa (habitualmente la traducción préstamo) podría suponer un obstáculo en la comprensión del concepto; consideramos que en una fase tan temprana de la evolución de la terminología de esta área de conocimiento no resulta extraño que algunos conceptos clave conserven su término original.

Pero la función de los técnicos no se limitó a proporcionar el término final y ubicarlo en su área de conocimiento: todas las definiciones procedentes de los glosarios de partida fueron depuradas con su concurso, dotando de una perspectiva académica a lo que eran mayoritariamente definiciones para profanos.

Pero al igual que el terminólogo debe permanecer en su puesto de fedatario de aquellos términos que se emplean en el área de conocimiento objeto de estudio (de la que no es por principio especialista), el técnico no debe invadir el territorio del lingüista. La apreciación del técnico está limitada fundamentalmente porque su aproximación al concepto es funcional: conoce el objeto y lo ubica en un sistema en el que los elementos interactúan en una dinámica problema-resultado. De este modo, en varias ocasiones (y con ambos técnicos) al intentar este autor conseguir una precisión mayor al definir un término, o su inclusión en un sistema jerárquico, la respuesta fue la misma: un diagrama en el que se mostraba la relación funcional de los términos. Al insistir sobre la conveniencia de establecer una taxonomía que facilitara la comprensión a los usuarios del glosario, se apreciaba una cierta reacción de perplejidad y el comentario era el mismo: nunca se lo habían planteado desde ese punto de vista. La definición por fórmulas y diagramas les resulta familiar y en ella se sienten cómodos. Trasladar esa información gráfica a texto escrito es una labor para la que el concurso del lingüista es a veces imprescindible.

El técnico se plantea la inclusión o no del término en el glosario a partir, fundamentalmente, de la relevancia del concepto en el campo de su especialidad y la frecuencia con que se emplea dicho término. Es la perspectiva que proporciona su propia experiencia. Mientras que para los términos clave esa opinión es más que autorizada, en aquellos casos en que el término es compartido por áreas colindantes (fundamentalmente estadística e informática en nuestro caso), es preciso contar con una información de la que el técnico carece: la que proporciona el tratamiento informático del corpus de referencia con el que sí cuenta el terminólogo.

Hay dos ejemplos que ilustran perfectamente lo que decimos: el caso de *data warehouse* y el de *overtraining*. En el primero de los casos había una fuerte resistencia por parte de ambos técnicos a incluirlo en el glosario final;

alegaban que era un error frecuente, y que la minería de datos era independiente del *data warehousing*. De este modo en principio fue descartado del glosario, pero el análisis del corpus reveló que era un termino presente en un 41,6% de los textos de referencia... ¿cómo prescindir de él? Acabó por ser aceptado.

En cuanto a *overtraining*, el recorrido del término fue el inverso: pasó de estar admitido originalmente como término "de pleno derecho" a ser considerado un sinónimo del verdadero término: *overfitting*. La razón estriba en dos datos de los que los técnicos, en principio, carecían: sólo uno de los glosarios de referencia (el más antiguo) lo incluía, y estaba presente en apenas uno de los textos del corpus. El término que designaba el mismo concepto -*overfitting*-, estaba presente en dos de los glosarios y se citaba en un tercio de los textos del corpus. La conclusión a la que se llegó fue que *overtraining* era el término más antiguo, el usado por IBM, pero que había terminado por caer en desuso. Las frecuencias de aparición de los términos en el corpus determinaron la inclusión o no en el glosario final de los casos dudosos: los que presentaban unos porcentajes más bajos (casos de *prevalence, resubstitution error* o *quantil*) quedaron descartados.

Así pues, los cuatro glosarios de partida y el corpus de referencia supusieron la base objetiva del proyecto. Pero no debemos extraer conclusiones erróneas del hecho de contar con listados previos de vocabulario propio de la especialidad (cabría pensar que bastaba hacer un traslado de aquellas listas a la nuestra de destino): sólo hay ocho términos del glosario final que estén presentes en las cuatro listados de partida, y si contamos la totalidad de los que propone cada listado tendríamos aproximadamente doscientos cincuenta. Los 114 términos del glosario final son el resultado de un largo proceso de selección, depuración y análisis. Llegados a este punto conviene recordar que los términos no están sujetos a copyright, pero los listados y las definiciones sí lo están. En este proyecto no se han solicitado licencias de

reproducción porque para el glosario resultante se tomaron como referencia las definiciones de los listados de partida, pero las que han quedado finalmente son las aportadas por los técnicos, con la mejora considerable que ello supone. En cuanto a los textos del corpus, son todos de libre acceso y aparecen referenciados en la sección de Apéndices. Su selección obedece a criterios de calidad (en su mayor parte son obras de autoridades reconocidas como Piatetsky-Sapyro, Kurt Thearling...etc) y a su directa relación con la materia objeto del corpus. Se optó por un corpus de 200.000 palabras al comprobar que aumentar su número podría ir en detrimento de la calidad del contenido. Por otra parte, dado el alto nivel de especialización de la materia y su reciente aparición, la literatura existente no es muy abundante.

Pero este corpus por sí mismo no nos habría servido de mucho sin las herramientas informáticas necesarias para extraer la información que en él se guarda. El analizador de textos supone una ayuda imprescindible en el trabajo del terminólogo. Estimamos que las tareas de localización de términos y cálculo de frecuencias de aparición empleando estos programas supusieron más de 300 horas de trabajo en este proyecto; es difícil calcular el tiempo que hubiera sido necesario para realizar las mismas tareas “a mano”.

El analizador de textos empleado permite localizar términos de una o varias palabras, así como indexar una base de datos a partir de varios archivos de texto. De este modo fue posible establecer las frecuencias de aparición de cada término. La tabla de frecuencias se puede consultar en la sección de Apéndices.

Los resultados que aquí se presentan no son extrapolables a otros campos colindantes, como la estadística y la informática. Se deben analizar

siempre teniendo en cuenta que nos encontramos ante una especialidad extremadamente joven (25) y con un ámbito de difusión muy restringido.

De entre los datos que nos muestran las tablas de resultados obtenidas cabría destacar notablemente el hecho de que la práctica totalidad de los términos en español y alemán son préstamos o traducciones préstamo. Concretamente en español hay un 13% de préstamos, y el 87% restante lo constituyen traducciones préstamo (con la única excepción de "datos atípicos" por *outliers,* lo cual no deja de ser anecdótico: el dato atípico del glosario es el que designa este concepto en minería de datos). En alemán el porcentaje de préstamos sube al 17%, pero el porcentaje de traducciones préstamo, sin dejar de ser muy elevado –un 78%-, es algo menor, puesto que en un 5% de los casos crea un término propio (*Beispiel, Spalte, Anwendungsphase, Abweichungserkennung, Merkmal, Artikel, Muster, Zielattribut*) para designar el concepto correspondiente. En la mayoría de los casos esto se debe a que el alemán carece de la palabra de origen latino que dio lugar al término en inglés. En ambas lenguas y, según suele ser habitual en otros campos, los términos expresados por medio de siglas y acrónimos permanecen invariables (CART, CHAID y CRISP DM).

También hemos podido constatar el hecho de que los porcentajes de sinonimia son relativamente bajos. Sólo un 29% de los términos en español tienen un sinónimo (el propio término en inglés en dos casos), y este porcentaje sube apenas al 35% en el caso del alemán (el término en inglés en 18 casos). Al tratarse de una ciencia exacta este dato no es infrecuente.

Lo que sí resulta especialmente llamativo es el proceso de asimilación al

(25) La fecha comúnmente aceptada del aparición del término Data Mining es 1996, con la publicación en el número de otoño de Artificial Intelligence Magazine del artículo *From Data Mining to Knowledge Discovery in Databases*, de Usama Fayyad, Gregory Piatetsky-Shapiro y Padhraic Smyth.

alemán de algunos términos en inglés. Para centrar esta cuestión debemos comenzar por comentar que en los últimos años la lengua alemana ha adoptado como propias numerosas palabras de origen inglés. De este modo, al libros de ejercicios se le llama “das Trainer”, el teléfono móvil es “das Handy”, el ordenador “der Computer”, la camiseta “das T-shirt”, y un bebé “ein baby”, por citar algunos ejemplos de la vida cotidiana. Como podemos observar, en el proceso se da la correspondiente asignación de género y escritura con mayúscula de los sustantivos, indispensable en alemán; además, la pronunciación de la palabra se mantiene como en el original inglés. Según datos obtenidos por medio de una encuesta realizada por este autor (incluida en la sección Apéndices) entre varios profesores de la Volkshochschule y la Universidad J. W. de Frankfurt, la permeabilidad del alemán ante los anglicismos es elevada y el hablante común acepta con agrado su uso. No podemos establecer a ciencia cierta el que esto sea una moda pasajera similar a la que se dio en la España de los años 70, cuando multitud de cafeterías pasaron a llamarse “Pubs” y no era raro el uso del formas inglesas en la rotulación de comercios (como el que aun perdura en “VIPS”), y el castizo vale fue sustituído por “OK”, y que se vio reflejada en la incorporación literal de multitud de términos en una ciencia emergente como era entonces la informática (según comentamos ya anteriormente). Pero el caso es que esta tendencia pudiera afectar a la creación de nueva terminología en alemán.

En nuestro estudio este proceso se ve reflejado en que en 26 casos de sustantivos (según se puede apreciar en la tabla correspondiente) el término original inglés se ha incorporado al alemán manteniendo su pronunciación inicial pero con grafía en mayúsculas y género propio. Esta atribución de género no es aleatoria: los masculinos y femeninos se asignan en función de su equivalente en alemán, y el género neutro se utiliza en aquellos términos que designan proceso. De este modo, encontramos los neutros “das Cleaning” o “das Data Mining”, masculinos del tipo “der Decision Tree” o “der Hidden

Layer" (obsérvese que la palabra que modifica, sin ser nombre, se incorpora en mayúsculas) y el femenino "die Business Intelligence", por mencionar algunos ejemplos.

Es también especialmente llamativo el hecho de que los términos originales hayan variado tan poco al incorporarse a la lengua de destino: un caso en español y ocho en alemán. Esto viene a corroborar la teoría que hemos sostenido a lo largo de esta tesis: en la mente del técnico prima la precisión del término, el hecho de que refiera más claramente al concepto que designa, por encima de consideraciones estéticas, lingüísticas o de cualquier otro tipo.

Como cierre podríamos plantear una pregunta abierta: ¿Cuál es el ciclo de vida de un término? Históricamente en ocasiones fueron la creación de personajes geniales que trabajaban solos en pos de satisfacer su propia curiosidad científica y necesitaban dar un nombre a sus inventos o descubrimientos, como fue el caso del belga Jan Baptista van Helmont, autor en el s. XVII de la –hoy- palabra gas.

En la actualidad es más frecuente que sean equipos de investigación en universidades y empresas los que creen un nuevo término para designar sus innovaciones técnicas (como en las casos de airbag, ABS...etc de la industria automovilística) y estos terminan por incorporarse al vocabulario cotidiano. Hay casos llamativos en los que la palabra y el término coexisten con la misma grafía; tal es el caso de "celulitis". El término celulitis designa en entornos médicos la extensión de una infección bacteriana aguda bajo la superficie de la piel, caracterizada por la presencia de eritema, inflamación, calor y dolor (fuente MedicineNet.com). La palabra celulitis sin embargo se ha popularizado como la manera de denominar a la "piel de naranja" o acumulación de grasa localizada en determinadas partes del cuerpo de la mujer (generalmente). Según parece el uso erróneo del término que ha derivado en la palabra que

todos conocemos tiene su origen en Francia, en el momento en que algunas compañías de productos cosméticos comenzaron a publicitar los "poderes anticelulíticos" de sus cremas. Esto muestra cómo a veces las empresas recurren a la terminología para dotar de un aura de sofisticación a sus productos. Los términos, por definición, viven en entornos de especialidad; como decíamos al principio, "huyen de la luz". Su exposición al habla común los hace desaparecer porque desde ese momento connotan. En consecuencia, en el caso que nos ocupa el término médico que se emplea en lugar de celulitis es "fibroedema geloide mucoso subcutáneo".

Un ejemplo clásico del mismo concepto designado por una palabra y un término es el siguiente: la palabra (originalmente término y marca registrada de la empresa Bayer) "aspirina" es la forma común que empleamos para designar al compuesto químico expresado con el término "ácido acetil-salicílico". Es posible que la comercialización de este producto como medicamento genérico termine por hacer popular su nombre, con lo cual dejará de ser técnicamente un término para convertirse en palabra. En otros casos los términos desaparecen al quedar obsoleta la tecnología que designaban (como en el caso de "floppy disk" en informática: la aparición de las memorias portátiles les relegó al olvido).

En los procesos de creación de terminología intervienen factores culturales muy variados, como la formación académica o el país de origen del científico que crea el término. En el caso de la minería de datos hemos visto que se trata de una especialidad científica que nace y se desarrolla en Estados Unidos y como resultado de ello a veces se crean términos que desde una perspectiva científica europea resultan, como mínimo, peculiares. Baste citar el caso de la palabra "taxonomía" (del griego *taxis* –orden- y *nomos* –ciencia-), que se aplica a la clasificación de sectores de población en el contexto de minería de datos. En Estados Unidos cada vez resulta más frecuente el empleo del término *folksonomy* para designar ese concepto. Tal modificación del

compuesto griego puede parecer pintoresca, pero sin duda el nuevo término explica por sí mismo mejor el concepto que designa; y, ¿no es esa la finalidad de todos los términos?

Quizás la conclusión más evidente a la que nos ha llevado este trabajo de investigación sea que realmente no importa si estamos hablando de términos en correcto inglés, español o alemán: la lengua de destino y sus reglas pasan a un segundo plano. La preocupación principal de los técnicos participantes ha sido ubicar cada concepto con la mayor exactitud posible en su campo de conocimiento, y, en pos de lograr este objetivo, el hecho de que los términos que designan dichos conceptos estuvieran recogidos o no en su propia lengua materna en el glosario final resultó secundario. Mi función como terminólogo en el presente estudio no ha interferido con este propósito. Un ejemplo que podría ilustrar esta situación es el siguiente: en una de mis reuniones de trabajo con la Doctora Menasalvas surgió el debate sobre cuál sería el mejor equivalente en español para el término *data warehouse*. Desde mi perspectiva de lingüista sugerí lo que me pareció entonces una traducción perfecta: “repositorio de datos”. La equivalencia semántica parecía total, y el término tenía además un cierto aire “especializado” del que carecía “almacén de datos”. La Doctora Menasalvas escuchó atentamente mis razones, asintió, e hizo el siguiente comentario: “creo que hay un profesor en Valencia que lo llama así”. El término, naturalmente, se incorporó invariable como “data warehouse” al listado español, al igual que ocurrió posteriormente con el listado alemán al tratar el mismo caso con el Dr. Lattner, dado que esa es la forma internacional de designar el concepto.

Nuestra obligación como terminólogos no consiste, a mi modo de ver, en erigirnos en defensores de la pureza del propio idioma. No somos traductores literarios, sino que trabajamos por y para la transmisión de conocimiento especializado. Nuestro trabajo, consecuentemente, deberá consistir en averiguar la forma en que los expertos denominan los conceptos clave de sus

áreas de conocimiento, verificar objetivamente que la literatura existente confirma los datos obtenidos y, una vez hecho esto, plasmar dicha información en nuestros glosarios, a fin de allanar el camino a aquellos que quieran iniciarse en la materia. Entonces estaremos haciendo terminología.

8- APÉNDICES

8.1 Relación de Normas ISO citadas

ISO 636:2004 Símbolos para lenguas y autoridades.

ISO 919: 1969 Guía para la preparación de vocabularios sistemáticos.

ISO 1087-1: 2000 Vocabulario en trabajos terminológicos: teoría y aplicación.

ISO 1149:1969 Presentación de vocabularios multilingües.

ISO 1951:1997 Símbolos lexicográficos y convenciones tipográficas en terminografía.

ISO 10006:1997 Directrices para la calidad en la gestión de proyectos.

ISO 12620:1999 Grupos de categorías de datos relacionados con el término, el concepto y los datos administrativos.

ISO 15188:2001 Directrices en la gestión de proyectos para la estandarización terminológica.

8.2 Frecuencia de aparición de cada término en el corpus

Accuracy	70,83%
Activation function	12,5%
Antecedent	12,5%
Association rule	58,33%
Associations discovery	50%
Associations	66,66%
Attribute	62,5
Back propagation	20,83%
Binning	25%
Boosting	16,66%
Business intelligence	41,66%
CART	33,33%
Case*	70,83%
Categorical data	54,16%
Categorical variable*	54,16%
CHAID	29,16%
Classification model	75%
Classification	75%
Cleaning	20,83%
Cluster	75%
Clustering	70,83%
Column	50%
Confidence factor	41,66%
Confusion matrix	12,5%
Consequent	12,5%
Continuous variable*	45,83%
Continuous	45,83%
Crisp DM	8,33%
Cross validation	50%
Data mining	100%
Data preparation	50%
Data understanding	62,5%
Data warehouse	41,66%
Decission tree	83,33%
Degree of fit	33,33%
Deployment	29,16%
Deviation/outlier detection	20,83%
Dimension	58,33%
Discrete	33,33%
Discretization	4,16%
Discriminant analysis	16,66%

Evaluation	33,33%
Exploratory data analysis	37,5%
Feature	66,66%
Feed forward	20,83%
Field	50%
Genetic algorithms	37,5%
Hidden layer	37,5%
Induction	50%
Item	66,66%
K-means	25%
K-nearest neighbor	29,16%
Kohonen feature map	25%
Layer	45,83%
Leaf	8,33%
Learning	70,83%
Learning algorithm	66,66%
Linear regression	45,83%
Link analysis	41,66%
Logistic regression	25%
Machine learning	50%
Market basket analysis	41,66
Missing data	45,83%
Model	95,83%
Modeling	58,33%
Neural network	83,33%
Node	50%
Noisy data	16,66%
Nominal variable	No analizable
Normalize	No analizable
Outliers	50%
Overfitting	33,33%
Pattern	87,5%
Precision	12,5%
Predictability	4,16%
Prediction	70,83%
Predictive modeling	50%
Predictor	33%
PCA Principal Component analysis	4,16%
Prevalence	4,16%
Processing unit	41,66%
Pruning	25%
Radial basis function	29,16%
Range	66,66%
Record	62,5%

Regression tree	62,5%
Rule	83,33%
Rule induction	50%
Sample	66,66%
Sampling	41,66%
Scoring	20,83%
Segment	62,5%
Segmentation	54,16%
Sensitivity analysis	25%
Sequence discovery	50%
Sequential patterns	33,33%
Significance	41,66%
Supervised learning	29,16%
Support factor	58,33%
Target	70,83%
Taxonomy	20,83%
Test data	75%
Test error	58,33%
Time series	50%
Time series model	50%
Training	75%
Training data	66,66%
Transaction	83,33%
Transformation	50%
Unsupervised learning	29,16%
Validation	54,16%
Value prediction	79,16%
Variable	66,66%
Visualization	66,66%
Windowing	4,16%

8.3 Ficha técnica del analizador de textos

TROPES ZOOM

The documentation of this software is in the electronic form
(Acrobat PDF®), installed with the software on your computer.
Use [Start menu/Programs/Tropes.../Doc] to browse or print the
Reference manuals. If you do not have Adobe Reader installed,
download it (for free) on Adobe® website :
(http://www.adobe.com/products/acrobat).

NOTES ON SOFTWARE INSTALLATION:

- Please report all the functioning errors to our
  Technical support of (Email: support@semantic-knowledge.com).

- This free Special Edition of Tropes Zoom has limited
  capacities of treatement and analysis. Please consult
  our Website for more information.

- The conversion of PDF® documents requires installation
  of the Adobe® IFilter component in your system.
  You can download the last version of IFilter on
  the Adobe® site (free), at the following address:
  http://www.adobe.com/support/techdocs/12b42.htm.

- In the majority of the cases, we think that a PC endowed
  with 256 Mb of RAM memory will be sufficient. However,
  if you use Windows in network, or specially Windows® XP,
  with several other active applications or if you
  analyze regularly consequent texts, you must install
  at least 512 Mb of RAM memory on your computer.

LEGAL NOTICE (Please read)

THIS SOFTWARE AND DOCUMENTATION IS PROVIDED BY THE
COPYRIGHT
HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED
WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED
WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR
PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER
OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL,
SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT
NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES;
LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION)

8.4 Tabla comparativa de presencia de términos en los glosarios iniciales

TÉRMINO	IBM	TC	CLEMENTINE	KT
Accuracy	x	x		x
Activation function	x	x		
Antecedent	x	x	x	
Association rule	x			x
Associations	x	x	x	
Associations discovery	x			
Attribute	x			
Back propagation	x	x		x
Back propagation neural network SÓLO EN IBM				
Binning		x		x
Business intelligence	x		x	
CART		x	x	x
Categorical data		x		
Categorical variable	x			
CHAID		x		x
Chi-squared		x		
Classification	x	x	x	
Classification model	x			
Cleaning		x		
Cluster	x			
Clustering	x	x	x	x
Column	x			
Commoditization	x			
Competitive intelligence	x			
Confidence (factor)	x	x		
Confusion matrix		x		
Consequent		x	x	
Continuous		x		
Continuous variable	x			
Cross validation		x	x	
Data mining application/operation/service/technique/tool SÓLO EN IBM				
Data mining	x	x	x	x
Data warehouse	x		x	

TÉRMINO	IBM	TC	CLEMENTINE	KT
Database segmentation	x			
Decision tree	x	x	x	x
Degree of fit		x		
Demografic data	x			
Dependent variable		x	x	
Deployment		x	x	
Deviation/outlier detection	x	x		
Dimension	x	x		
Discrete	x	x		
Discretization	x			
Discriminant análisis	x	x		
Exploratory data análisis	x	x		x
External data		x		
Feature	x		x	
Feed forward		x		
Field	x		x	x
Fuzzy logic		x		x
Genetic algorithm		x		x
Heterogeneity	x			
Hidden layer	x			
Hidden node		x		
Homogeneity	x			
Independent variable		x		
Induction		x		
Intelligent system	x			x
Internal data		x		
Item	x			
K-nearest neighbor		x		
Kohonen feature map	x	x	x	x
Layer		x		
Leaf	x	x		
Learning algorithm	x			
Learning		x		
Least squares		x		
Left-hand side		x		
Link analysis	x			
Logistic regression		x	x	
Market basket analysis			x	

TÉRMINO	IBM	TC	CLEMENTINE	KT
MARS		x		
Maximum likelihood		x		
Median		x	x	
Metadata	x		x	
Missing data		x		
Model		x	x	x
Neural induction	x			
Neural network		x	x	x
Node	x	x	x	
Noisy data	x	x		
Nominal variable	x			
Non-applicable data		x		
Normalize		x		
OLAP	x	x		x
Optimization criterion	x			
Outliers	x	x	x	x
Overfitting	x	x	x	x
Overlay		x		
Overtraining	x			
Pattern		x		
Precision		x		
Predictability		x		
Predictive modelling	x			x
Prevalence		x		
Processing unit	x			
Pruning		x	x	
Psichographic data	x			
Quantile	x		x	
Quartile	x		x	
Radial basis function	x		x	x
Range		x		
Record	x		x	x
Regression tree		x	x	x
Resubstitution error		x		
Right-hand side		x		
R-squared		x		
Rule	x		x	
Rule body	x			

TÉRMINO	IBM	TC	CLEMENTINE	KT
Rule head	x			
Sampling		x	x	x
Scoring			x	
Segment	x		x	
Segmentation	x		x	x
SOM	x			
Sensitivity analysis		x	x	x
Sequence discovery		x	x	
Sequential pattern discovery	x			
Sequential patterns	x			
Significance		x	x	
Similar time sequence discovery	x			
Standarize		x		
Supervised learning	x	x	x	x
Support factor	x		x	x
Taxonomy	x			
Test data		x		
Test error		x		
Time series model		x		
Time series		x	x	x
Topology		x		
Training	x	x		
Training data		x		
Transaction	x			
Transformation		x	x	
Unsupervised learning	x	x	x	x
Validation		x		
Value prediction	x			
Variable	x		x	
Variance		x	x	
Visualization technique	x			
Visualization	x	x		x
Windowing		x		

8.5 Tabla trilingüe de términos

ENGLISH	ESPAÑOL	DEUTSCH
Accuracy	Accuracy	Accuracy
Activation function	Función de activación	Aktivierungsfunktion
Antecedent	Antecedente	Vorbedingung
Association rule	Regla de asociación	Assoziationsregel
Associations discovery	Proceso de descubrimiento de asociaciones	Assoziationsregellernen
Associations	Asociaciones	Assoziationen
Attribute	Atributo	Attribut
Back propagation	Retroalimentación	Back Propagation
Binning	Bining	Binning (das)
Boosting	Boosting	Boosting (das)
Business intelligence	Business intelligence	Business Intelligence (die)
CART	CART	CART
Case	Caso	Beispiel
Categorical data	Datos categóricos	Symbolische Daten
Categorical variable	Variable categórica	Symbolische Variable
CHAID	CHAID	CHAID
Classification model	Modelo de clasificación	Klassifikationsmodell
Classification	Clasificación	Klassifikation
Cleansing	Limpieza	Cleaning (das)
Cluster	Cluster	Cluster (der)
Clustering	Clustering	Clustering (der)
Column	Columna	Spalte
Confidence factor	Factor de confianza	Konfidentzwert
Confusion matrix	Matriz de confusión	Confusion Matrix (die)
Consequent	Consecuente	Konsequentz
Continuous variable	Variable continua	Kontinuierliche Variable
Continuous	Contínuo	Kontinuierlich
CRISP DM	CRISP DM	CRISP DM
Cross validation	Validación cruzada	Cross Validation (das)
Data mining	Data mining/minería de datos	Data Mining (das)
Data preparation	Preparación de datos	Datenvorbereitung
Data understanding	Comprensión de datos	Data Understanding (das)
Data warehouse	Data warehouse	Data Warehouse (das)
Decission tree	Árbol de decisión	Decision Tree (der)
Degree of fit	Grado de ajuste	Grad der Anpassung
Deployment	Deployment	Anwendungsphase
Deviation/outlier detection	detección de desviaciones	Abweichungserkennung/ Entdeken von Ausreissern
Dimension	Dimensión	Dimension
Discrete	Discreto	diskret
Discretization	Discretización	Diskretisierung

Discriminant analysis	Análisis discriminante	Diskriminanzanalyse
Exploratory data analysis	Análisis de datos exploratorio	explorative Datenanalyse
External data	Datos externos	externe Daten
Feature	Característica	Merkmal
Feed forward	Alimentación hacia delante	feed forward
Field	Campo	Feld
Fuzzy logic	Lógica difusa	Fuzzy Logic (die)
Genetic algorithms	Algoritmos genéticos	Genetische Algorithmen
Heterogeneity	Heterogeneidad	heterogeneität
Hidden layer	Nivel oculto	Hidden Layer (der)
Homogeneity	Homogeneidad	Homogeneität
Induction	Inducción	Induktion
Internal data	Datos internos	interne Daten
Item	Item	Artikel
K-means	K-means	K-Means
K-nearest neighbor	Vecino K-cercano	K-nächste Nachbarn
Kohonen feature map	Mapa de Kohonen	Kohonen-Karte
Layer	Capa	Schicht
Leaf	Hoja	Blatt
Learning	Aprendizaje	Lernen
Learning algorithm	Algoritmo de aprendizaje	Lernenalgoritmus
Linear regression	Regresión lineal	lineare Regression
Link analysis	Link análisis	Zusammenhangsanalyse
Logistic regression	Regresión logística (Análisis logístico discriminante)	logistische Regression (logistische discriminantz Analyse)
Machine learning	Aprendizaje automático	maschinelles Lernen
Market basket analysis	Análisis de cesta de la compra	Warenhorbanalyse
Missing data	Missing data	fehlende Daten
Model	Modelo	Modell
Modeling	Modeling	Modellierung
Neural network	Red neuronal	neuronales Netz
Node	Nodo	Knoten
Noisy data	datos con ruido	verrauschste Daten
Nominal variable	Variable nominal	nominale Variable
Normalize	Normalizar	normalisieren
Outliers	Datos atípicos	Ausreisser
Overfitting	Overfitting	Overfitting (das)
Pattern	Patrón	Muster
Precision	Precisión	Präzision
Predictability	Predecibilidad	Vorhersagbarkeit
Prediction	Predicción	Prädiktion
Predictive modeling	Modelado predictivo	generieren eines

		Vorhersagemodell
Predictor	Modelo predictor	Predictor
Prevalence	Prevalencia	Häufigkeit
Processing unit	Unidad de proceso	Einheit
Pruning	Poda	Pruning (das)
Radial basis function	Función basada en el radio	radiale Basisfunktion
Range	Rango	Intervall
Record	Registro	Eintrag
Regression tree	Árbol de regresión	Regression Tree (der)
Rule	Regla	Regel
Rule induction	Inducción de regla	Regelinduktion
Sample	Muestra	Beispielmenge
Sampling	Muestreo	Sampling (das)
Scoring	Scoring	Scoring (das)
Segment	Segmento	Segment
Segmentation	Segmentación	Segmentierung
Sensitivity analysis	Análisis de sensibilidad	Sensitivitätsanalyse
Sequence discovery	Descubrimiento de secuencias	Sequence discovery(das)
Sequential patterns	Patrones secuenciales	sequentielles Muster
Significance	Significancia	Signifikanz
Supervised learning	Aprendizaje supervisado	überwachtes Lernen
Support factor	Factor de soporte	Support (der)
Target	Objetivo	Zielattribut
Taxonomy	Taxonomía	Taxonomie
Test data	Datos de prueba	Testdaten
Test error	Error del modelo	Testfehler
Time series	Series temporales	Zeitreihe
Time series model	Modelo de series temporales	Zeitreihenmodell
Training	Entrenamiento	Training (das)
Training data	Datos de entrenamiento	Trainingsdaten
Transaction	Transaction	Transaktion
Transformation	Transformación	Transformation
Unsupervised learning	Aprendizaje no supervisado	unüberwachtes Lernen
Validation	Validación	Validierung
Value prediction	Predicción de Valores	Wertvorhersage / Vorhersage
Variable	Variable	Variable
Visualization	Visualización	Visualisierung
Windowing	Ventana Temporal	Windowing (das)

8.6. Textos empleados en la confección del corpus

Agrawal, R. et al 1996: *The Quest Data Mining System*. IBM.

Alex Berson, Stephen Smith, and Kurt Thearling 1999: *An Overview of Data Mining Techniques* (Excerpted from the book Building Data Mining Applications for CRM) Mc Graw Hill.

CRISP DM 1.0 *Data Mining Guide.* SPSS Inc. 2000.

Daniel S. Tkach 1998: Information Mining with the IBM Intelligent Miner Family. An IBM Software Solutions White Paper.

Fayyad, U., Piatetsky-Shapiro, G.and Smyth, P. 1996: *From Data Mining to Knowledge Discovery in Databases.* AI Magazine.

Llewellyn, Mark 2006: *Introduction to Data Mining*. School of Electrical Engineering and Computer Science. University of Central Florida.

Saarenvirta, G. 1998: *Mining Customer Data. A step-by-step look at a powerful clustering and segmentation methodology.*

SAS Institute inc 2008: *What's New in SAS Enterprise Miner 5.3*. Overview.

Stanton, J. 2005: *How Neural Networks Are Used in Data Mining.*

Strehl, A., Ghosh, J. 2002: *Relationship-Based Clustering and Visualization for High-Dimensional Data Mining*. Department of Electrical and Computer Engineering, The University of Texas at Austin,Austin, Texas.

Thearling, K. 2008: *An Introduction to Data Mining*. *Discovering hidden value in your data warehouse.* Kurt Thearling/ Kurt@thearling.com.

Thearling, K. et al 2001: *Visualizing Data Mining Models.* Published in Information Visualization in Data Mining Discovery, edited by Usama Fayyad, Georges Grinstein, and Andreas Wierse. Morgan Kaufman.

Wielenga, D. 2007: *Identifying and Overcoming Common Data Mining Mistakes.* SAS Institute Inc., Cary, NC.

www.almaden.ibm.com 2008: *Data Mining: Extending the Information Warehouse Framework.* Whitepapers IBM.

www.anderson.ucla.edu 2008: Data Mining: What is Data Mining?

www.bogelt.net 2008: Finding Association Rules/Hyperedges with the Apriori Algorithm.

www.crisp-dm.org 2008: CRISP-DM 1.0 *Step by Step Data Mining Guide.*

www.dbmsmag.com 2008: *Neural Networks*. DBMS Data Mining Solutions Suplement.

www.qub.ac.uk 2008: Parallel Computer Centre. *Data Mining Techniques.*

www.sas.com 2008: SAS Data Mining Software and Text Mining.

www.sas.com Wielenga, D. 2007: *Identifying and Overcoming Common Data Mining Mistakes*. SAS Institute Inc., Cary, NC.

www.statsoft.com 2008: *Data Mining Techniques*. Electronic Textbook. StatSoft Inc.

www.twocrows.com 2005: *Data Mining: in Brief*. Two Crows Corporation.

Xindong Wu 2008: *Data Mining from Large Databases.* Dept of Math and Computer Science*.* Colorado School of Mines.

8.7 Encuesta realizada entre profesores de la Universidad J. W. Goethe y de la Volkshochschule de Frankfurt

ENCUESTA SOBRE PERMEABILIDAD DEL ALEMÁN ANTE LOS ANGLICISMOS

Indique si en los siguientes campos existen palabras importadas del inglés, y, si así fuera, en qué porcentaje. Indique también si existe una palabra alemana de igual significado en ese campo y su grado de uso.

INFORMÁTICA

PALABRAS:

EQUIVALENTE **DE**

% INFLUENCIA MUY ALTO ALTO MEDIO BAJO MUY BAJO

TELECOMUNICACIONES

PALABRAS:

EQUIVALENTE **DE**

% INFLUENCIA MUY ALTO ALTO MEDIO BAJO MUY BAJO

COMERCIO

PALABRAS:

EQUIVALENTE **DE**

% INFLUENCIA MUY ALTO ALTO MEDIO BAJO MUY BAJO

INDUSTRIA

PALABRAS:

EQUIVALENTE **DE**

% INFLUENCIA MUY ALTO ALTO MEDIO BAJO MUY BAJO

DEPORTE

PALABRAS:

EQUIVALENTE **DE**

% INFLUENCIA MUY ALTO ALTO MEDIO BAJO MUY BAJO

8.8 Datos relativos a los técnicos participantes en el proyecto.

Klaus-Dirk Schmitz. (Supervisor del proyecto)

Doctor en Lingüística Aplicada y Licenciado en Matemáticas e Informática. Catedrático de Terminología en la Fachhochschule de Colonia, Alemania. Ostenta una posición de liderazgo en multitud de instituciones internacionales: Presidente del Consejo Alemán de Terminología (RaDT). Presidente del Centro Internacional de Información para la Terminología (Infoterm). Presidente del Comité Alemán de Estandarización en Aplicaciones Informáticas para Terminología. Sus actividades docentes e investigadoras se centran en la teoría y gestión de la Terminología, así como la aplicación de herramientas informáticas a la traducción. Autor, co-autor y editor de numerosos artículos y libros sobre la materia. Figura clave en la Terminología Centroeuropea contemporánea.

Ernestina Menasalvas. (Técnico para español)

Doctora en Informática. Profesora Titular en la Facultad de Informática de la Universidad Politécnica de Madrid, España. Su tesis doctoral se centró en la modelización del proceso de data mining, campo en el que centra su investigación. Forma parte del Grupo de Investigación DAME (DAta Mining Engineering) de la UPM. Ha participado en la Red de Excelencia Europea KD-Net y en la acción coordinada KD-ubiq. En la actualidad es Investigadora Principal de un Proyecto del Programa Nacional. Fruto de su participación en proyectos y de más de diez años de investigación en el área son sus publicaciones en *journals*, libros y conferencias relacionadas. Es *PC Member* de los principales congresos de data mining, en los cuales también actúa de revisora.

Andreas D. Lattner. (Técnico para alemán)

Doctor en Informática. Profesor-investigador en el Departamento de Informática y Matemáticas de la Universidad J. W. Goethe de Frankfurt am Main, Alemania. Su tesis doctoral versa sobre la minería de series temporales en entornos dinámicos. Su investigación se centra en la minería de datos, y es autor de numerosos artículos sobre la materia. PC Member y revisor en diversos talleres y conferencias sobre data mining.

9- BIBLIOGRAFÍA

Arntz, Reiner, Picht, Heribert 1995: *Introducción a la Terminología*. Madrid . Fundación Germán Sánchez Ruipérez . Pirámide. (Traducción de 1989 Einführung in die Terminologiearbeit, Hildesheim, Georg Olms).

Arntz, Reiner 1993: "Terminological Equivalence and Translation". Sonneveld, Helmi B. & Loening, Kurt L. eds. *Terminology: Applications in Interdisciplinary Communications*. Amsterdam/Philadelphia: John Benjamins Publishing Co. 5-19.

Arntz, Reiner 2006: "Minderheiten und ihre Sprachen im vielsprachigen Europa". *Uni Hildesheim - Das Magazin.* N 10: 43-48

Berson, Alex, Smith, Stephen and Thearling, Kurt 1999: *An Overview of Data Mining Techniques* (Excerpted from the book Building Data Mining Applications for CRM) Mc Graw Hill.

Bratko, I., Kubat, M. y Michalsky, R. (Eds) 1998: *Machine Learning and Data Mining: Methods and Applications*. Wiley.

Cabena, Hadjinian, Stadler, Verhees y Zanasi 1997: *Discovering Data Mining From Concept to Implementation*. Upper Saddle River, NJ. Prentice Hall.

Cabré, María Teresa 1993: *La Terminología. Teoría, Metodología, Aplicaciones*. Barcelona:Antártida/Empuries.

Cabré, María Teresa 1996: "Terminology Today". Harold Somers ed. *Terminology, LSP & Translation.* Manchester: John Benjamins Publishing Co. 15-34

Cabré, María Teresa 1999: *La terminología: Representación y Comunicación. Elementos para una Teoría de Base Comunicativa y otros Artículos.* Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra.

Cabré, María Teresa 2002: "Terminología y Lingüística: la Teoría de las Puertas Abiertas". Red Iris. *Estudios de Lingüística Española (EliEs)* Volúmen 16.

Fayad, U., Piatetsky-Shapiro, G., Smyth, P. y Uthurusamy, R. Eds 1996: *Advances in Knowledge Discovery and Data Mining*. MIT Press.

Felber, H. y Pitch, H. 1984: *Métodos de Terminografía y Principios de Investigación Terminológica.* Madrid: Instituto "Miguel de Cervantes".CSIC

Holton, G. 1988: *Thematic Origins of Scientific Thought*. Cambridge, Massachusetts: Harvard University Press.

Kageura, Kio 2002: *The dynamics of terminology. A descriptive theory of term formation and terminological growth*. John Benjamins.

Koestler, A. 1964: *The act of creation*. Nueva York, Dell.

Lorenzo Criado, Emilio 1981: *Utrum Lingua an Loquentes?(Sobre las Presuntas Dolencias y carencias de Nuestro Idioma)*. Discurso leído el 22 de noviembre de 1981 en su recepción pública por el Excelentísimo Señor Don Emilio Lorenzo Criado y contestación del Excelentísimo Señor Don Rafael Lapesa Melgar. Madrid: Real Academia Española.

Maimon & Rokach (Eds) 2005: *The dataming and knowledge discovery handbook*. Nueva York:Axel Springler.

Pavel, Silvia 1993: "Neology and Phraseology as Terminology-in-the-making".

Sonneveld, Helmi B. & Loening, Kurt L. eds. *Terminology: Applications in Interdisciplinary Communications*. Amsterdam/Philadelphia: John Benjamins Publishing Co. 21-34.

Rirdance, S. y Vasiljevs, A. 2006: *Towards Consolidation of European Terminology Resources. Experiences and Recommendations From EuroTerm Bank Project*. Riga: Tilde.

Ryszad S. Michalsky, Ivan Bratko, Miroslav Kubat eds. 1998: *Machine Learning and Data Mining: Methods and Applications.* Wiley

Sonneveld, Helmi B. y Loening, Kurt L. 1993: *Terminology; Applications in Inter-disciplinary Communication*. Amsterdam/Philadelphia. John Benjamins Publishing Co.

Tercedor Sánchez, M.B. 1999: *La Fraseología en el Lenguaje Biomédico: Análisis desde las Necesidades del Traductor*. Tesis doctoral. Dpto. de Traducción e Interpretación de la Universidad de Granada

Thearling, Kurt 2005: "Data Mining for CRM". Maimon, Oded & Rokach, Lior eds. *The Data Mining Discovery Handbook*. Nueva York: Axel Springer. 1249-1259

Turner, J. 1988: *A Theory of social interaction*. Stanford, California. Stanford University press.

DIRECCIONES EN INTERNET

http://www.elies.rediris.es/elies6/index.html (4 junio 2006)

http://www.ttt.org/clsframe/datcats.html CLS Framework: Listing of ISO 12620 Data Categories (18 febrero 1999)

http://coral.lili.uni-bielefeld.de/~ttrippel/terminology/node82.html (28 mayo 1999)

http://www.iso.org (23 mayo 2007)

http://de.wikipedia.org/wiki/Eugen_Wüster (29 abril 2007)

http://www.cfwb.be/franca/pg012.htm (27 de mayo de 2007)

http://www.dict.org (27 de mayo de 2007)

http://www.twocrows.com/glossary.htm (20 de octubre de 2007)

http://www.acta.es/articulos_mf/37009.pdf (16 de mayo de 2008)

CURSOS

IULATERM 2005 Materiales de la V Escuela Internacional de Verano de Terminología. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. ISBN: 84-89782-23-7 / DL: B-31.453-2005.

TSS 2006. Materiales de la Terminology Summer School. Escuela de traducción. Universidad de Viena.

EAFT SUMMIT 2006. Centre de Terminologie de Bruxelles. Institut Libre Marie Haps. Bruselas.